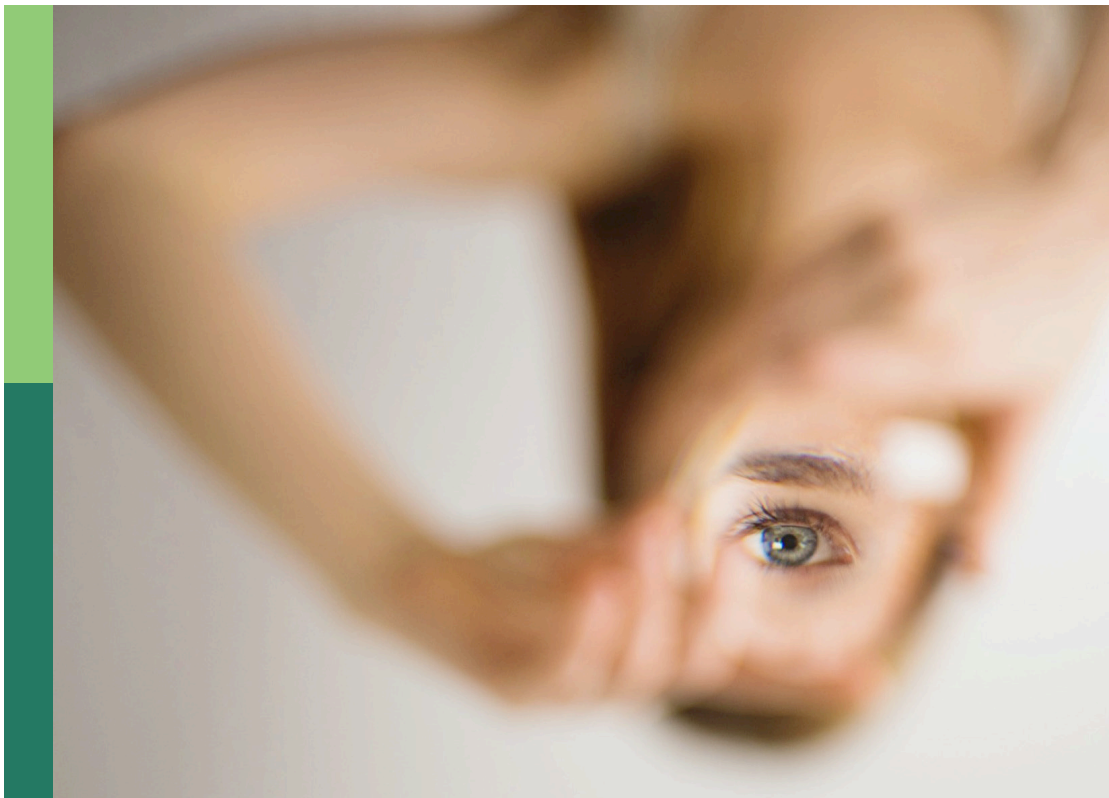
frontiers | Research Topics

Evaluating performance

Edited by
Michele Biasutti, George Waddell, Aaron Williamon and Roberta Antonini Philippe

Published in
Frontiers in Psychology
Frontiers in Education



ISSN 1664-8714
ISBN 978-2-8325-3626-1
DOI 10.3389/978-2-8325-3626-1

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Evaluating performance

Topic editors

Michele Biasutti — University of Padua, Italy
George Waddell — Royal College of Music, United Kingdom
Aaron Williamon — Royal College of Music, United Kingdom
Roberta Antonini Philippe — Université de Lausanne, Switzerland

Citation

Biasutti, M., Waddell, G., Williamon, A., Philippe, R. A., eds. (2023). *Evaluating performance*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-3626-1

Table of contents

frontiers in Psychology

BRIEF RESEARCH REPORT
published: 29 July 2021
doi: 10.3389/fpsyg.2021.691043

Can a Good Break Shot Determine the Game Outcome in 9-Ball?

*Jing Wen Pan[1], John Komar[1], Shawn Bing Kai Sng[1] and Pui Wah Kong[1,2]**

[1] Physical Education and Sports Science Academic Group, National Institute of Education, Nanyang Technological University, Singapore, Singapore, [2] Office of Graduate Studies and Professional Learning, National Institute of Education, Nanyang Technological University, Singapore, Singapore

This study aimed to quantify the break shot characteristics and identify their significance in predicting the game outcomes in 9-ball tournaments. The break shots of 275 frames (241 men's, 34 women's) of professional tournaments were analyzed from two aspects: (1) cue ball position, represented by the distance between the cue ball and the table center, and (2) ball distribution, indicated by the standard deviation of Voronoi cell areas determined from all remaining balls on the table. Spearman correlation and binary logistic regression were utilized to identify associations and to predict the frame outcomes, respectively. Results showed that the more balls falling into the pockets during the break, the more clustered the remaining balls ($r_s = 0.232$, $p < 0.001$). The closer the cue ball ending toward the table center, the more balls potted in the visit immediately after the break ($r_s = -0.144$, $p = 0.027$). Neither cue ball position nor ball distribution could predict table clearance or winning of a frame. In conclusion, pocketing more balls during the break is associated with more clustered balls remaining on the table. Parking the cue ball near the table center after the break can facilitate potting more balls immediately after.

Keywords: cue ball position, ball distribution, Voronoi diagram, pool, billiards

OPEN ACCESS

Edited by:
Roberta Antonini Philippe, University of Lausanne, Switzerland

Reviewed by:
Jonathan Douglas Connor, James Cook University, Australia
Scott Sinnett, University of Hawai'i at Mānoa, United States

***Correspondence:**
Pui Wah Kong
puiwah.kong@nie.edu.sg

Specialty section:
This article was submitted to Performance Science, a section of the journal Frontiers in Psychology

Received: *05 April 2021*
Accepted: *12 July 2021*
Published: *29 July 2021*

Citation:
Pan JW, Komar J, Sng SBK and Kong PW (2021) Can a Good Break Shot Determine the Game Outcome in 9-Ball? Front. Psychol. 12:691043. doi: 10.3389/fpsyg.2021.691043

INTRODUCTION

Nine-ball is a popular billiard game played with a cue stick on a rectangular table with six pockets. Players strike the white cue ball to pocket nine colored billiard balls in ascending numerical order (1-ball, 2-ball, . . . 9-ball). An individual frame is won by the player pocketing the last ball on the table which is the 9-ball. The player to win a predetermined number of frames first wins the game. Each frame of 9-ball games begins with a break shot. The purpose of the break shot is to separate the racked object balls and to pocket at least one ball so that the player can remain on the table. If the player misses, the visit (which consists of a series of consecutive successful shots) will be passed to the opponent.

It is generally believed that a powerful break shot can separate the object balls well such that the player can pocket the subsequent object balls easily and continue staying on the table until he/she wins by pocketing the 9-ball. Therefore, the break shot represents an important shot in 9-ball, and a good break shot can logically increase the chance of winning the game. Jeanette Lee, a professional 9-ball player ranking number 1 in 1990s, once shared that "*the break is the most important shot in nine-ball . . . because it can give you control of the table*" (Lee and Gershenson, 2007) (p. 102). It is further stated that "*the best breaks in nine-ball spread the rack*" (p. 102). Regarding how a player executes the break, it is advised that "*when you break, try to get the cue ball to hit the 1-ball as*

solidly as possible, so it will roll off about a foot and stop (stops near the center of the table)" (p. 103). Learning from her personal experience, a "good" break shot leads to the cue ball stopping near the center of the pool table and the object balls widely spread. In general, a player is expected to win the frame after a good break shot in professional tournaments. Furthermore, clearing the table is a specialized case of winning a frame where the player who takes the break shot pots all colored balls in sequence to win the frame without letting the opponent play at all. Players take advantages of a good break shot in some tournaments where the frame winner continues to break in the following frame (Shepard, 1997). In the Turning Stone Classic, for example, a player who always clears the table after a good break shot tends to keep winning without passing the play to the opponent. Despite some anecdotal evidence and experts' opinions, there is no scientific literature on how the break shot may impact game outcomes in cue sports.

Currently, there is no established method to quantify the characteristics of a break shot. Based on coaching expert's opinion, the position of the cue ball and the distribution of the remaining colored balls after the break are of key interest. To evaluate the ball positions and movements on the pool table, previous studies on cue sports (Haar and Faisal, 2020; Haar et al., 2020; Pan et al., 2021) applied 2D video analysis which involved a digital camera and analysis software. Based on video analysis, one could quantify the end position of the cue ball after the break to examine if parking the cue ball near the center of the pool table would indeed indicate a good break shot as suggested by the anecdotal evidence. Regarding how well the object balls are spread by the break shot, no previous studies have attempted to analyze the ball distribution pattern on the pool table after the break. One possible approach is to adopt the Voronoi diagrams which is a partitioning of a 2-D plane based on the "nearest-neighbor" rules (Aurenhammer, 1991) assigning each dot to a corresponding cell. As a computational geometry method, Voronoi diagrams have been widely applied in various disciplines (Fonseca et al., 2012; Lopes et al., 2017; Sun et al., 2018; Xiao et al., 2018). In civics and planning, Xiao et al. (2018) used Voronoi diagrams to describe the pedestrian motion patterns wherein each pedestrian was represented by a point in a Voronoi cell. In informatics, Voronoi diagrams were applied to partition the target region for further study of wireless local area network (Sun et al., 2018). In biology, Voronoi diagram was proposed as a new method to characterize the geometrical distribution of human cells (Lopes et al., 2017). In team sports, such as Futsal, Voronoi diagrams have been utilized to identify players' distribution patterns where each player on the pitch was treated as a dot with the coordinates (x, y) (Fonseca et al., 2012). In 9-ball, because the total dimension (the playing surface) is finite, an optimal spread of the balls on the table should lead to similar Voronoi cell areas. This would mean that a small standard deviation of the Voronoi cell areas among all the object balls would indicate that balls are spread evenly. Thus, Voronoi diagrams may be suitable tools for analyzing ball distribution pattern in cue sports by generating cells for the remaining balls on the table after the break.

This study aimed to quantify the characteristics of the break shot from two aspects: (1) the end position of the cue ball, and (2) the distribution of the balls remaining on the pool table after the break shot. Secondly, as the break shot is considered important in 9-ball, the relationship between the characteristics of the break shot and frame outcomes was also examined. Based on the professional player's experience, it was hypothesized that a good break shot would be characterized by the cue ball positioned close to the center of the table and the remaining balls widely spread across the table. A good break shot was also hypothesized to increase the chance of potting more balls immediately after, clearing the table, and winning a frame in professional 9-ball tournaments.

METHOD

Data

This study involved analyses of publicly available online videos and was approved by the Nanyang Technological University Institutional Review Board (Protocol Number: IRB-2019-05-42). Videos of the semifinals and finals of World Pool Billiard Association (WPA) ranked tournaments in the years of 2019 and 2020[1] were downloaded from the Internet. The finals and semifinals of six tournaments including 2019 Diamond Las Vegas Open (Men), 2019 International 9-ball Open (Men), 2019 World 9-Ball Championship (Men and Women), 2019 World 9-Ball Championship China Open (Men and Women), 2019 WPA Players Championship (Men), and 2020 Turning Stone Classic XXXIII (Men) were analyzed in this study. There were 275 frames in total (241 from men's tournaments and 34 from women's tournaments) after removing the frames where the cue ball fell into the pocket or jumped out of the pool table, because the relative positions of the cue ball could not be measured in those situations. Player rankings were within the top 50 for male players and top 5 for females according to the official rankings provided by WPA in January 2021 (see text footnote 1).

Data Processing

Top-view videos of professional 9-ball tournaments were downloaded and analyzed using Kinovea (version 0.8.27, available for download at: http://www.kinovea.org), a 2D motion analysis software which has shown good accuracy in measuring objects at an angle range of 45° to 90° (Puig-Diví et al., 2019). In a previous study on 9-ball test protocols, excellent inter- and intra-rater reliability was found in measuring ball movements from video recordings using Kinovea (Pan et al., 2021). After each break shot, the positions of all balls remaining on the pool table were digitized manually (**Figure 1**). Firstly, the "perspective grid" was applied to calibrate the pool table and was set as 127 cm × 254 cm (Pan et al., 2021) which was the dimension of a standardized pool table used in professional tournaments. Then, the "mark" tool was implemented to mark each of the ball remained on the table. The 2D coordinates (x, y) of all balls remaining on the table were obtained for analysis.

[1] https://wpapool.com/ranking/

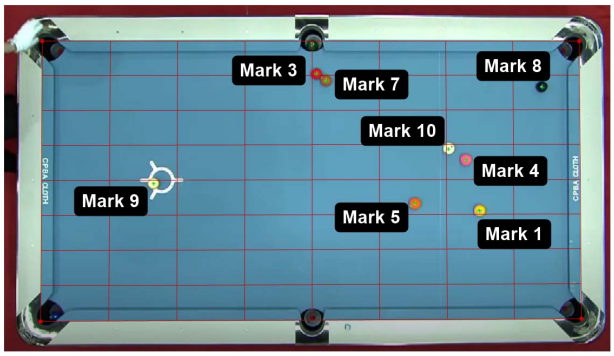


FIGURE 1 | Example of 2D video analysis in Kinovea, where a "perspective grid" (red) was applied to calibrate the pool table and the "mark" (green) was added in the center of each ball. Marks 1–9 indicates 1-ball to 9-ball; Mark 10 indicates the cue ball; 2-ball and 6-ball went into the pocket and therefore not shown. Video was extracted from publicly available source (https://www.youtube.com/watch?v=2sEqz5N0cug).

The first variable for the characteristics of the break shot is the cue ball distance (CBD, in cm), defined as distance between the end position of the cue ball and the center of the pool table. The CBD was calculated using the equation (1),

$$CBD = \sqrt{(x_{10} - x_0)^2 + (y_{10} - y_0)^2} \quad (1)$$

where x_{10} and y_{10} were the x and y coordinates of the cue ball (Mark 10), respective; x_0 and y_0 were the x and y coordinates of the center of the pool table. To examine the inter-rater and intra-rater reliability, a sub-sample of 20 out of the 275 frames were randomly chosen. To examine the inter-rater and intra-rater reliabilities, a sub-sample of 20 out of the 275 frames were randomly chosen. This sub-sample was independently digitized by two members of our research team, and repeated digitized twice by one team member, to obtain the CBD values.

Voronoi diagram was applied to analyze the ball distribution after the break shot using the MATLAB function (v2020b, MathWorks, Natick, MA, United States) proposed by Ong (2011). In a Voronoi diagram, each of the remaining balls on the table (blue dot in **Figure 2**) was assigned to a cell (polygon, shaped by the red edges and table boundaries in **Figure 2**). Once the area of each cell was obtained, the standard deviation of the areas of all cells (SDVD, in cm^2) was computed. A smaller SDVD (similar areas among the Voronoi cells, **Figure 2A**) would imply that the balls on the table were more evenly distributed as the areas occupied by each ball were similar. Likewise, a larger SDVD (different areas among the Voronoi cells, **Figure 2B**) would indicate that some balls were clustered, and others were far from the clusters.

During the break shot, the number of balls falling into pockets was counted. The player needs to pocket at least one ball during the break in order to stay on the table for the next visit. The following game outcomes were also obtained for each frame: (1) if the player who took the break shot cleared the table or not, (2) if the player won the frame or not, (3) the number of balls the player potted in the next visit immediately after the break shot. Table clearance is a specialized case of winning a frame where the player makes ball(s) pocketed during the break shot and then pots all object balls legally in one consecutive visit without give the opponent any chance to play. Theoretically, a desirable break shot should allow the player to pot more balls in the next visit and may even clear the table and win the frame.

Statistical Analyses

Data are expressed as mean (standard deviation). Inter-rater and intra-rater reliabilities were examined using intraclass correlation coefficient (ICC) on SPSS (version 26.0, IBM Corp., Armonk, United States). ICC was interpreted as *slight* (<0.20), *fair* (0.21–0.40), *moderate* (0.41–60), *substantial* (0.61–0.80), or *almost perfect* reliability (>0.80) (Altman, 1991; Heng et al., 2016). Standard error of measurement (SEM) was calculated from the ICC results using the formula: SEM = SD × $\sqrt{(1-ICC)}$. All other statistical analyses were performed on JASP (version 0.14.1; JASP Team, 2020) statistical software. The association between the number of balls potted into pockets in the break shot and SDVD was assessed using Spearman's rho (r_s) since the assumption of normality was violated. Spearman's rho (r_s) was also performed to test the associations between the characteristics of the break shot (CBD and SDVD) and the frame outcomes (i.e., the number of balls potted in the subsequent shots). Binary logistic regression was run to identify the significant characteristic indicators, with two dependent variables of match outcomes set as Clear = 1 and Not = 0, and Win = 1 and Not = 0. Odds ratios (OR) and corresponding 95% confidence intervals (95% CI) were presented (Robertson and Joyce, 2018). Statistical significance was set at the 0.05 level.

RESULTS

Results of all 275 digitized videos showed an average of 58.9 (30.7) cm for CBD, and 1975.6 (641.8) cm^2 for SDVD. In the break shot, 1.4 (0.8) balls were potted into the pockets. After the break shot, 3.1 (3.3) balls were potted in the subsequent visit. The results of ICC indicated *almost perfect* inter-rater ($ICC_{2,\ 2}$ = 0.972, SEM = 4.4 cm) and intra-rater ($ICC_{2,\ 1}$ = 0.999, SEM = 0.8 cm) reliabilities. Spearman correlation analysis indicated a significant positive association between the number of balls falling into pockets after the break shot and SDVD (r_s = 0.232, p < 0.001, **Figure 3A**). Also, a significant negative association between CBD and the number of balls potted in the subsequent visit was identified (r_s = −0.144, p = 0.027, **Figure 3B**). Conversely, no significant association between SDVD and the number of balls potted after the break shot was found (p = 0.129, **Figure 3C**).

Regarding the frame outcome, Clear the table or Not, the binary logistic regression model was statistically significant [$\chi^2(2)$ = 6.616, p = 0.037]. The model as a whole explained between 2.4% (Cox & Snell R^2) and 3.4% (Nagelkerke R^2) of the variance in the frame outcome for "Clear or Not," and correctly classified 70.5% of frames (0% for "Clear" and 100% for "Not"). Of the independent variables that were included in the model, only CBD made a unique statistically significant contribution

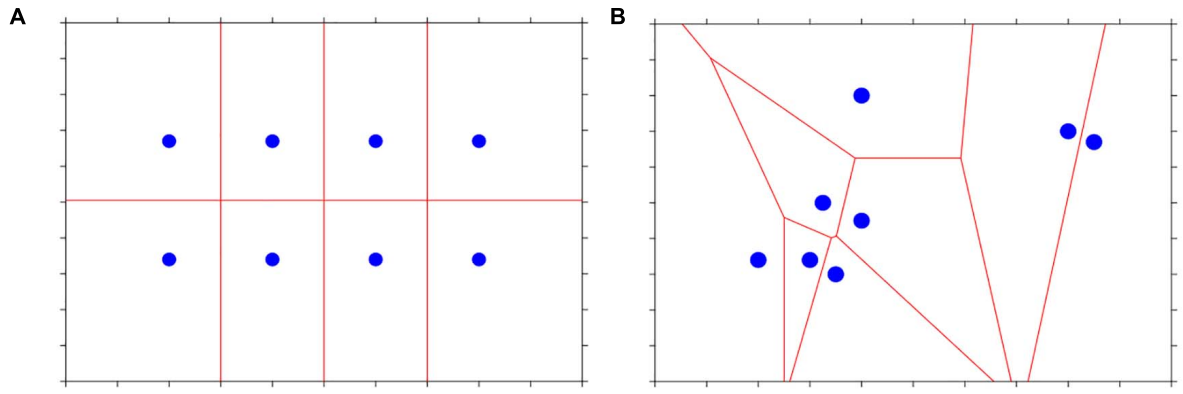


FIGURE 2 | Schematic representations of Voronoi diagram. **(A)** Dots (blue) are evenly distributed leading to a small standard deviation of cell areas. **(B)** Co-existence of both big and small cells causing a big standard deviation of cell areas.

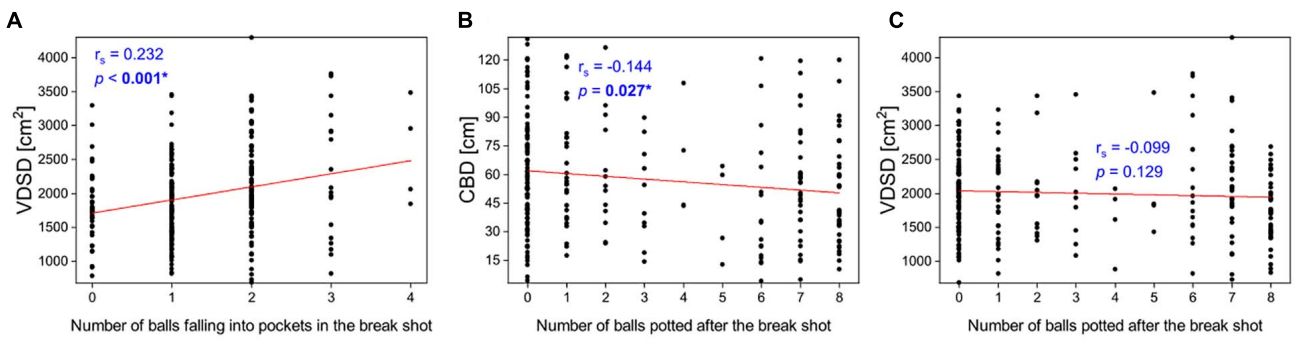


FIGURE 3 | (A) Correlations between the number of balls falling into pockets in the break shot and SDVD. Correlations between the number of balls potted after the break shot and **(B)** CBD and **(C)** SDVD. CBD denotes the distance between the end position of the cue ball and the center of the pool table; SDVD denotes the standard deviation of the areas of all the Voronoi diagram cells. Significant difference ($p < 0.05$) is shown in bold text and indicated by an asterisk.

TABLE 1 | Binary analysis of the two predictors of the break shot characteristics for the game outcomes in 9-ball.

	Clear or not			Win or not		
Predictors	χ^2	**OR (95% CI)**	***p***	χ^2	**OR (95% CI)**	***p***
CBD	6.268	0.989 (0.980, 0.998)	**0.012***	0.119	0.999 (0.991, 1.007)	0.730
SDVD	0.032	1.000 (1.000, 1.000)	0.859	1.588	1.000 (1.000, 1.001)	0.208

CBD denotes the distance between the end position of the cue ball and the center of the pool table; SDVD denotes the standard deviation of the areas of all the Voronoi diagram cells.
χ^2 denotes Wald's chi-square, OR denotes odds ratio, and CI denotes confidence interval.
Significant difference ($p < 0.05$) is shown in bold text and indicated by an asterisk.

to the model ($p = 0.012$, **Table 1**). Binary logistic regression identified no significant predictor of the game outcomes for "Win or Not" [$\chi^2(2) = 1.722, p = 0.423$].

DISCUSSION

This study aimed to quantify the characteristics of the break shot in 9-ball and its relationship with the outcomes of a frame. The novel contributions of the current study were that (1) we pioneered the application of Voronoi diagrams in cue sports to objectively quantify the characteristics of the break shot, and (2) we established the relationship between the break shot characteristics and the frame outcomes. Building on professional 9-ball player's experience (Lee and Gershenson, 2007), the cue ball position (indicated by CBD) and ball distribution on the table (indicated by SDVD) were proposed to represent important break shot characteristics. This study investigated only professional tournament and elite players to purposely avoid the situations where less-skilled players performed a good break shot but were unable to clear the table due to their relatively lower playing levels.

Balls Potted During the Break Shot

Having balls pocketed in the break shot allows the player to remain on the table and continue to play. Logically, it is desirable to have more balls falling into pockets during the break since having fewer balls remaining on the table will give the player a higher chance to clear the table and win the frame. Results of this present study showed a significant positive correlation between the number of balls potted during the break and SDVD. This indicates that the more balls falling into the pockets during the break, the more clustered the remaining object balls on the table. Although the association is weak ($r_s = 0.232$), having clustered balls on the table may not be a good situation for the next shots. If the player cannot clear the table due to the clustered balls blocking each other, the visit will be passed to the opponent. This finding suggests that it may not be a good strategy for players to pocket many balls in the break shot *per se* because they may face difficulties to continue pocketing the remaining balls that are clustered close to each other. Players should aim for potting at least one ball during the break to stay on the table and be aware that potting more balls is not necessarily better for subsequent shots.

Cue Ball Position

After the break shot, the current study found a significant correlation between CBD and the number of balls potted immediately after the break. This result reaffirmed the significance of the end position of the cue ball, revealing that when the cue ball was parked near the table center the player could pot more balls within the same visit. This finding is not surprising and confirms the anecdotal guidelines proposed by players and coaches. When planning for a break shot, players are advised to park the cue ball near the center of the pool table after the break to set up for the next shot.

Similar to studies on other sports in which important factors contributing to successful game outcomes were identified (Gómez et al., 2013; Rumpf et al., 2017; Robertson and Joyce, 2018), the current study examined the relationship between the break shot characteristics and the frame outcomes using binary logistic regression. For instance, in Australian Rules football, opposition rank and match location were two significant predictors found to be associated with the match outcomes (Win or Loss) (Robertson and Joyce, 2018). When investigating the matches in the 2014 FIFA World Cup Brazil, it was reported that shot accuracy was the best predictor for match success (Rumpf et al., 2017). In elite basketball, different predictors were identified for men and women for predicting ball possessions (Gómez et al., 2013). In the present study on cue sport, it was found that CBD was a significant predictor for the game outcome, "Clear or Not." Although the prediction accuracy was 70.5% overall, the model correctly predicted 100% of cases where the table was not cleared, but 0% of cases of table clearance (i.e., all object balls were pocketed within one visit following the breaks shot). This suggests that CBD is not successful in predicting table clearance following the break shot. Similarly, CBD also failed to predict the game results "Win or Not."

The complexity of 9-ball may explain why cue ball position alone could not predict table clearance or winning of a frame. It should the noted that the success of 9-ball games depends on many factors, such as the players' skills, strategies, experience, and even luck. An excellent break shot alone may not significantly contribute to the success of the remaining visits and outcome of the frame. For example, a player may have skillfully parked the cue ball at the table center in the break shot but the next object ball (lowest numbered ball on the table) is blocked by other balls. Failing to pot the next object ball would force the player to pass the play to the opponent and may then lose the frame eventually. In addition, players may choose to deliver a "safety shot" under situations when they are not confident in winning or making table clearance. A "safety shot" does not aim to pot the any object ball in the pocket but to place the cue ball in a challenging position for the opponent. The frequent use of "safety shot" in the profession tournaments was evident, as reflected by the high occurrence of "zero" in **Figure 3** which shows that no balls were potted after the break shot in 46.5% of the time (128 out of 275 frames). Thus, while CBD could provide the information about the relative position of the cue ball after the break shot, this variable alone is insufficient to indicate the likelihood of success in 9-ball games. More variables, such as the object balls distribution may be needed to comprehensively quantify the break shot characteristics.

Ball Distribution

Voronoi diagram as a computational geometry method (Aurenhammer, 1991) measures the surrounding areas of the balls left on the table after the break shot such that it shows if there are clusters of balls by comparing the areas among the cells. Similar to previous studies (Lopes et al., 2017; Sun et al., 2018; Xiao et al., 2018), the current study treated each ball as a simple dot regardless of its number when applying the Voronoi diagram to calculate the cell areas. It was hypothesized that a small SD of the areas (i.e., indicating that the balls are evenly distributed), which is a desirable game situation, would increase the chance of clearing the table and winning the frame. However, this hypothesis was not supported by the results of the present study. Ball distribution (indicated by SDVD) was not related to the number of balls potted after the break, nor did it predict whether the player could clear the table or win the frame. It is acknowledged that while Voronoi diagram describes the ball distribution, the relative positions of the cue ball and the next object ball to be potted are not taken into account. It is possible that the next ball to be potted is blocked by other balls despite the balls were sufficiently separated and showing a small value of SDVD. Future studies should therefore refine the current analysis to examine the balls distribution and at the same time considering also the relative positions of the cue ball to the next ball to be potted on the table (i.e., for the subsequent shot).

Although this study showed that Voronoi diagrams may be too simplistic for the application in 9-ball break shots, the method of quantifying object distribution may be useful in other situations. When applied to team sports, for example, it would be interesting to examine whether a more "spread" or "compact" players formation may lead to better performance. It is also

possible to investigate if factors, such as fatigue and playing level of the opponents would influence the position distribution of team sports players.

Limitations

There are a few limitations to this study that can be identified. Firstly, we only considered the number of balls potted in the visit immediately following the break. This approach does not reflect the game dynamics comprehensively because a frame may involve many visits played alternatively between the two players before one of them wins. Future work can extend the analysis to all visits played in a frame to better understand how a break shot can impact the playing characteristics of the entire frame. Secondly, the sample investigated in the current study was delimited to professional tournaments and elite players and therefore the established association between the break shot characteristics and the frame outcomes may not be directly applied to less-experienced players. Thirdly, the sample size of 275 frames may not have fully captured the diverse playing styles and break shot characteristics among professional players. Hence, an investigation into a large number of frames in 9-ball is warrant. Lastly, the present study included much fewer women's games (34 frames) compared with men's games (241 frames) and therefore all frames were treated in the same group. As the game characteristics and playing strategies may differ between sexes, future studies could compare the break shot and game outcomes between male and female players.

CONCLUSION

In professional 9-ball tournaments, pocketing more balls during the break is associated with more clustered balls remaining on the table. Parking the cue ball near the table center after the break can be a good strategy to facilitate pocketing more balls immediately after. Table clearance and winning of a frame are likely influenced by multiple factors and could not be predicted by the break shot characteristics alone. While Voronoi cell areas could provide an objective measure of the ball distribution on the table, this method did not reveal any association between ball distribution and game outcomes. Future work should consider more in-depth analysis of the object balls distribution after the break shot, taking into account the relative positions of the cue ball and the next object ball to be potted on the table.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Nanyang Technological University Institutional Review Board. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

PK, JK, and JP originated this project and performed statistical analysis. SS and JP processed and analyzed the data. All authors discussed the results and actively contributed to the final manuscript.

FUNDING

This study was funded by the National Institute of Education Academic Research Fund (NIE AcRF, RI 1/19 KPW). JP was supported by the China Scholarship Council (CSC).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.691043/full#supplementary-material

REFERENCES

Altman, D. G. (1991). *Practical Statistics for Medical Research*. Available online at: https://books.google.com.sg/books?id=Y5ebDwAAQBAJ (accessed March 18, 2021).

Aurenhammer, F. (1991). Voronoi diagrams—a survey of a fundamental geometric data structure. *ACM Comput. Surv.* 23, 345–405. doi: 10.1145/116873.116880

Fonseca, S., Milho, J., Travassos, B., and Araújo, D. (2012). Spatial dynamics of team sports exposed by Voronoi diagrams. *Hum. Mov. Sci.* 31, 1652–1659. doi: 10.1016/j.humov.2012.04.006

Gómez, M.-A., Lorenzo, A., Ibañez, S.-J., and Sampaio, J. (2013). Ball possession effectiveness in men's and women's elite basketball according to situational variables in different game periods. *J. Sports Sci.* 31, 1578–1587. doi: 10.1080/02640414.2013.792942

Haar, S., and Faisal, A. A. (2020). Brain activity reveals multiple motor-learning mechanisms in a real-world task. *Front. Hum. Neurosci.* 14:354. doi: 10.3389/fnhum.2020.00354

Haar, S., van Assel, C. M., and Faisal, A. A. (2020). Motor learning in real-world pool billiards. *Sci. Rep.* 10:20046. doi: 10.1038/s41598-020-76805-9

Heng, M. L., Chua, Y. K., Pek, H. K., Krishnasamy, P., and Kong, P. W. (2016). A novel method of measuring passive quasi-stiffness in the first metatarsophalangeal joint. *J. Foot Ankle Res.* 9:41. doi: 10.1186/s13047-016-0173-2

Lee, J., and Gershenson, A. (2007). *The Black Widow's Guide to Killer Pool: Become the Player to Beat. Crown/Archetype*. Available online at: https://books.google.com.sg/books?id=YP7n0LsdhiIC (accessed March 18, 2021).

Lopes, W., Vainstein, M. H., De Sousa Araujo, G. R., Frases, S., Staats, C. C., de Almeida, R. M. C., et al. (2017). Geometrical distribution of cryptococcus neoformans mediates flower-like biofilm development. *Front. Microbiol.* 8:2534. doi: 10.3389/fmicb.2017.02534

Ong, M. S. (2011). *Arbitrary Square Bounded Voronoi Diagram*. Available online at: https://ww2.mathworks.cn/matlabcentral/fileexchange/30353-arbitrary-square-bounded-voronoi-diagram (accessed March 18, 2021).

Pan, J. W., Komar, J., and Kong, P. W. (2021). Development of new 9-ball test protocols for assessing expertise in cue sports. *BMC Sports Sci. Med. Rehabil.* 13:9. doi: 10.1186/s13102-021-00237-9

Puig-Diví, A., Escalona-Marfil, C., Padullés-Riu, J. M., Busquets, A., Padullés-Chando, X., and Marcos-Ruiz, D. (2019). Validity and reliability of the Kinovea program in obtaining angles and distances using coordinates in 4 perspectives. *PLoS One* 14:e0216448. doi: 10.1371/journal.pone.0216448

Robertson, S., and Joyce, D. (2018). Evaluating strategic periodisation in team sport. *J. Sports Sci.* 36, 279–285. doi: 10.1080/02640414.2017.1300315

Rumpf, M. C., Silva, J. R., Hertzog, M., Farooq, A., and Nassis, G. (2017). Technical and physical analysis of the 2014 FIFA world cup Brazil: winners vs. losers. *J. Sports Med. Phys. Fit.* 57, 1338–1343. doi: 10.23736/S0022-4707.16.06440-9

Shepard, R. (1997). *Amateur Physics for the Amateur Pool Player*. Available online at: https://cdn.preterhuman.net/texts/science_and_technology/physics/Amateur%20Pool%20Physics%20for%20the%20Amateur%20Pool%20Player.pdf (accessed July 17, 2021).

Sun, Y., He, Y., Meng, W., and Zhang, X. (2018). Voronoi diagram and crowdsourcing-based radio map interpolation for GRNN fingerprinting localization using WLAN. *Sensors* 18:3579. doi: 10.3390/s18103579

Xiao, Y., Chraibi, M., Qu, Y., Tordeux, A., and Gao, Z. (2018). Investigation of Voronoi diagram based direction choices using uni- and bi-directional trajectory data. *Phys. Rev. E* 97:052127. doi: 10.1103/PhysRevE.97.052127

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 20 September 2021
doi: 10.3389/fpsyg.2021.689378

Skill Level in Tennis Serve Return Is Related to Adaptability in Visual Search Behavior

Jernej Rosker[1]* *and Ziva Majcen Rosker*[2]

[1] Faculty of Health Sciences, University of Primorska, Koper, Slovenia, [2] Faculty of Sport, University of Ljubljana, Ljubljana, Slovenia

Analyzing visual search strategies in tennis is primarily focused on studying relationships between visual behavior and tennis performance. However, diverse movement characteristics among different servers suggest the importance of adjusting the visual search strategies of an individual while playing against different opponents. The aim of this study was to analyze whether visual search strategies can be attributed to the individual server and the returning player during the tennis serve return or return performance. Seventeen tennis players were enrolled in this study (five international players and 12 national players) producing a sample of 1,020 returns measured with mobile eye trackers. The random forest machine learning model was used to analyze the ability to classify the returning player [area under the curve (AUC): 0.953], individual server (AUC: 0.686), and return performance category (AUC: 0.667) based on the location and duration of the focal vision fixation. In international tennis players, the higher predictability of the server was observed as compared with national level players (AUC: 0.901 and 0.834, respectively). More experienced tennis players presented with a higher ability to adjust their visual search strategies to different servers. International players also demonstrated anticipatory visual behavior during the tossing hand movement and superior information pickup during the final phases of the stroke of a server.

Keywords: expertise, racquet sports, interception tasks, focal vision, visual fixations

OPEN ACCESS

Edited by:
Roberta Antonini Philippe, University of Lausanne, Switzerland

Reviewed by:
Sean Müller, Federation University Australia, Australia
Matteo Bonato, University of Milan, Italy

***Correspondence:**
Jernej Rosker
jernej.rosker@fvz.upr.si

Specialty section:
This article was submitted to Performance Science, a section of the journal Frontiers in Psychology

Received: *31 March 2021*
Accepted: *16 August 2021*
Published: *20 September 2021*

Citation:
Rosker J and Majcen Rosker Z (2021) Skill Level in Tennis Serve Return Is Related to Adaptability in Visual Search Behavior. Front. Psychol. 12:689378. doi: 10.3389/fpsyg.2021.689378

INTRODUCTION

Efficient visual search strategies during interceptive precision tasks in highly dynamic sports have been associated with superior sports performance. Among others, they enable more accurate anticipation of forthcoming events, fast and accurate decision-making, and online movement adaptation (Triolet et al., 2013; Woolley et al., 2015; Connor and Knierim, 2017). However, studies suggest high interindividual and intraindividual variability in visual search patterns making it difficult to fully understand their function and adaptability (Dicks et al., 2017).

Several factors have been proposed to contribute to the characteristics of visual search patterns and their variability, such as temporal and spatial demands of the task, amount of information available during the task performed, and knowledge about visual properties and regularities of the environment (Paeye and Madelain, 2014; Dicks et al., 2017). In interceptive tasks, such as returning tennis serve, saving a penalty kick in soccer, or making a save in field hockey, intercepting a ball is performed under temporal and spatial constraints and requires movement preparation, execution,

and adaptation in a time window that can exceed the action of an opponent and the travel time of a ball (Jackson and Mogan, 2007; Müller and Abernethy, 2012; Morris-Binelli et al., 2021). Three main phases during interception tasks have been proposed to contribute significantly to interception performance (Müller and Abernethy, 2012; Mecheri et al., 2019). In the first preflight phase, the movement characteristics of an opponent are perceived, allowing the observer to initiate his movement toward the interception point. In the second phase, early ball flight characteristics are perceived and used to guide the interception movement, and in the third phase, late ball flight information is used to fine-tune the interception movement. Therefore, the role of specific visual information varies between different phases of interception tasks and contributes specifically to movement preparation and execution (Müller and Abernethy, 2012).

In addition, the movements of opponents are characterized by inherent movement variability (Latash, 2010), which poses a challenge for observers to extract relevant visual information. Opponents produce different visual cues by applying phase-specific synergies of limb and body movements (Maselli et al., 2017; Shafizadeh et al., 2019). These synergies change depending on the phase of the movement of an opponent. Such noisiness in stimulus presentation inherently leads to higher variability in the visual search strategies employed by the observer (Paeye and Madelain, 2014). For example, in the tennis serve return, the highest variability in the upper limb movements of a server was associated with the preparation phase of the serve and the ball contact phase. Between these two phases, the vertical ball toss was identified as one of the most critical phases related to anticipation of serve type and movement initiation (Jackson and Mogan, 2007; Mecheri et al., 2019). The movements of the arm and racquet during the backswing and upswing are thought to contribute to the anticipation of ball flight direction and guide the initial movements of the returning player (Jackson and Mogan, 2007; Button et al., 2011; Navia et al., 2017).

A considerable body of literature has demonstrated differences in visual behavior between more and less experienced athletes (Jackson and Mogan, 2007; Land, 2009; Button et al., 2011; Lebeau et al., 2016; Murray and Hunfalvay, 2017). More experienced athletes are able to use fewer saccades and longer fixations than less-skilled counterparts, as well as earlier fixations relative to the interception movement termination phase. In addition, their visual search strategies rely more on the top-down control of attention (Aglioti et al., 2008). This allows higher-skilled athletes to use more consistent visual search strategies while observing noisy opponent preflight actions compared with less skilled observers (Jones and Miles, 1978; Aglioti et al., 2008; Müller and Abernethy, 2012).

An additional factor affecting the characteristics of eye movements and consequently their variability is the type of research protocol used. Video-based observational studies are commonly used (Murray and Hunfalvay, 2017), which can be augmented with video occlusion techniques (Farrow et al., 2005; Mecheri et al., 2011; Giblin et al., 2017). However, these studies have been criticized as attending to visual cues and anticipating ball flight direction from a video recording are thought to differ in underlying neural processes as compared with *in vivo* measures of visual attention (Dicks et al., 2010). More specifically, the ventral visual information processing pathway is involved in video-based experiments, and the dorsal visual processing pathway is involved in real game situations (Müller and Abernethy, 2012). This has important implications for conducting ecologically valid studies. In other words, the dorsal pathway is responsible for rapid visual perception and its connection to movement, whereas the ventral pathway involves cognitive processes for recognizing objects of interest (Vaziri-Pashkam and Xu, 2017). Moreover, video-based studies usually present visual stimuli, which are deprived of at least some types of auditory stimuli. As reported in the literature, the presence of multisensory stimuli may reduce the variability of visual search strategies (Murray et al., 2019). In addition, the motor experience of an observer in performing the observed movement patterns contributes significantly to visual behaviors, such as a lower number of rapid saccadic eye movements and longer fixation times (Aglioti et al., 2008). Moreover, advanced anticipation in expert players has been shown to correlate with increased cortical activity, especially in areas within the medial and lateral frontal cortex which are critical for observing and interpreting actions performed by opponents (Wright et al., 2010). Overall, these factors may contribute to the noisiness of visual search patterns when searching for relevant visual information.

Although differences between athletes of different skill levels and successful and less successful task performance have been well-documented, no systematic attempts have been made to analyze other possible sources of variability in a visual search pattern, such as observing different opponents. This is particularly relevant given that in sports, opponents change regularly and present different subject-specific constraints. For example, kinematics related to ball flight direction have been shown to be specific to individual throwers, with high intra-thrower variability (Maselli et al., 2017). This in itself presents a particular challenge for the observer to extract opponent-specific relevant visual information.

The noisiness of visual search strategies poses a specific methodological challenge to study the different aspects of visual attention. To better understand the characteristics of visual perception in the context of sports performance, some authors (Dicks et al., 2017) have proposed the use of non-linear analysis methods. Such computational approaches allow the identification of other factors that contribute to sports performance, such as the adaptability of visual behavior to different opponents.

Therefore, the aim of this study was to apply a robust data mining method to investigate whether visual search strategies during the tennis serve return are related to the individual opponent (individual server), observer (returning player), and tennis serve return performance. We hypothesized that visual fixation duration and its position on the server during different phases of tennis serve return would be related to the individual returning player but to a lesser extent to the specific server and serve-return performance, regardless of the quality level of a player which was the primary goal of this study. The secondary goal was to investigate whether new characteristics in the aforementioned visual attention behavior could be identified that may contribute

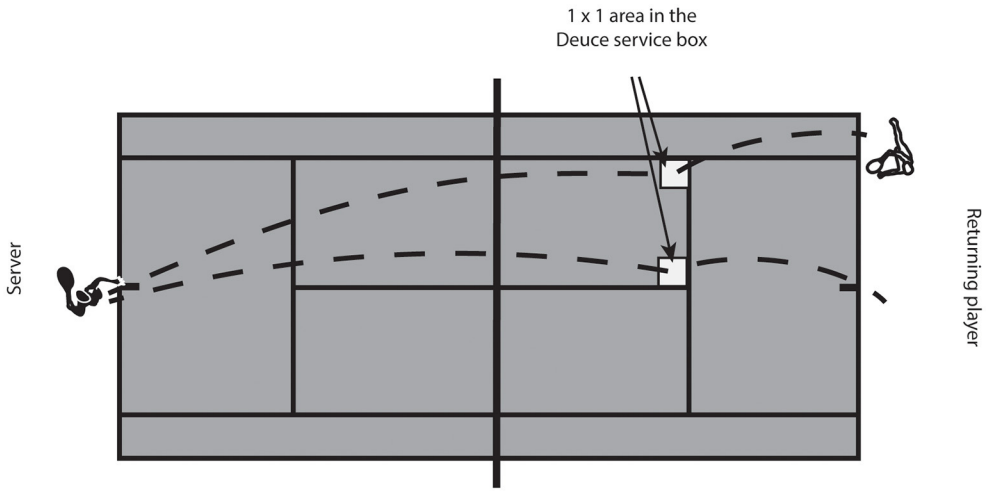


FIGURE 1 | Experimental setup. The position of the server and returning player are presented, as well as areas where the server was required to position his serve in the returning player service box.

to a better understanding of return performance and direct future research.

MATERIALS AND METHODS

Participants

Seventeen male tennis players who were enrolled in this study were divided into two groups. Twelve tennis players competing at a national level (mean age 20.0 ± 1.5 years, height 1.80 ± 0.06 m, weight 77.1 ± 6.7 kg, training experience 13 ± 0.8 years, all subjects were right-handed, placed in the top 20 players at the national junior rankings) were included in the first group. Four professional players competing at Davis Cup and one former professional international tennis player (mean age 30.5 ± 2.8 years, height 1.80 ± 0.08 m, weight 78.2 ± 6.3 kg, training experience 19 ± 4.7 years, one subject was left-handed) were included in the second group. Two players in the second group were ranked between the top 350 players and two players between the top 900 players in the ATP singles ranking. One player was retired, ranked between the top 300 players in the ATP singles ranking 1 year before the experiment but was still highly active as a tennis player. All players were contacted directly and invited to participate in this study. Subjects had to be enrolled in training that consisted of more than five training sessions per week, were free of musculoskeletal injuries 6 months before the experiment, and had a normal or corrected-to-normal vision. All participants were required to read and sign an informed consent form. This study was approved by the National Committee for Medical Ethics (No. 0120-47/2020/6) and was conducted according to the Declaration of Helsinki.

Design and Procedure

The experiment was conducted on a standard tennis hard court surface. Each participant performed ∼90 returns on the same side of the court with the goal of scoring a point. Three servers switched intermittently after performing the first three serves (flat serves) into the 1 × 1 m square located in the left and right corners of the Deuce service box (**Figure 1**). Serves were distributed equally to each side but were semi-randomized. Participants and servers were instructed to continue the play on a successful return until one of the players scored a point. Three servers were recruited for each of the two groups, corresponding to their quality level. Players used their own racquets and strings during the testing procedure. The task was completed after 15 ± 9.7 min.

Two synchronized 50-Hz video cameras (Logitech C920, Logitech, Lausanne, Switzerland) were positioned at the edge of the court and recorded the movements of the server and the returning player as well as the ball flight. During returns, each returning player wore a 50-Hz eye tracker (Tobii Pro Glasses 2, Tobii, Danderyd, Sweden). An additional strap was used to fix the eye tracker to the head of a participant to prevent slippage. The head unit was connected *via* a USB cable to the recording unit, which was attached to the hip of a participant. Before the experiment, a single-target calibration routine was performed using the Tobii Pro Glasses Controller (Tobii Pro Glasses Controller, Tobii, Danderyd, Sweden). After calibration, the participant was instructed to direct his gaze at a 0.1 m target placed at the position of a server while standing on the opposite side of the court behind the baseline. If the gaze position did not overlap with the target, the calibration routine was repeated until sufficient accuracy was achieved. Before the start of the experiment, each participant performed a warm-up routine consisting of 15 returns to get adjusted to the glasses. None of the participants reported any discomfort while wearing the eye tracker during the experiment.

Data Analysis

All returns were analyzed *post-hoc* by two nationally certified tennis coaches. They classified each return performance as

TABLE 1 | Definition of areas of interest in specific phases of the movement of servers.

	1st PSM	2nd PSM	3rd PSM	4th PSM	5th PSM
Area of interest	Area surrounding server	Tossing hand	Area of hand—racket movement	Area of hand—racket movement	Tossing hand
	Diagonal server—returner	Tossing hand movement area	Ball contact area	Head and upper body	Hand—racket
	Ball bounce area	Ball release	Head and upper body		Head
	Tossing hand and upper body	Ball upwards flight			Upper body
	Racket hand	Back leg			Lower body
	Racket	Front leg			
		Server head			
		Upper body			
		Hips			

Individual areas of interest are presented and the phases of the movement of a server they belong to; PSM, phases of server movement.

follows: (i) tactically superior return, meaning that the server could not continue the play or had significantly reduced tactical options (usually the ball landed just next to the side-line or under the server), (ii) tactically inferior return, meaning that the server had more tactical options to choose from (usually a slower or lob ball), or (iii) an error by the returning player, defined as an out or a ball lending in the net. All errors made by the serving player were excluded from this study. The agreement between the two coaches in classifying the return performance was 96%. The returns where the agreement between the two coaches was not reached were excluded from further analysis. Twenty returns were randomly selected for each of the three return performance categories. This resulted in 720 returns in the national group and 300 returns in the professional group.

Eye-tracker data were analyzed using Tobii Pro Lab software (Tobii Pro lab 1.145, Tobii, Danderyd, Sweden). Eye-tracking data were filtered using a raw filter without gap-fill interpolation and with noise reduction both available in the Tobii Pri Lab software. To assess the gaze behavior during the movement of a server, serve actions were categorized into the following five phases based on major serve events that also determine the important time frames for the emergence of movement synergies of servers (Shafizadeh et al., 2019, 2020), which are defined by the organization of different body parts that work together to achieve a specific goal of the movement task and provide stability and flexibility of the system (Latash et al., 2007): (i) the preparation phase, starting 500 ms before and ending at the first observable upward movement of the tossing hand (preparation), (ii) the ball toss, ending at the instance the ball left the hand (ball toss), (iii) the windup phase, ending at an instance when the racket began to move upward, (iv) the hitting phase, ending when the racket contacted the ball, and (v) the follow-through phase, ending at an instance when the gaze of the participant started following the flying ball. In addition, 25 areas of interest (AOI) were defined according to the movement synergies present during the movement of servers (Jackson and Mogan, 2007; Shafizadeh et al., 2019) (**Table 1**), six in the first, nine in the second, three in the third, two in the fourth, and five in the fifth phase of servers movement. Two of the AOIs in the second phase were determined to be the areas (i.e., the tossing hand movement area and the area of ball release) where the specific movement is going to occur.

The actual gaze positions (represented by a marker with a size of 0.73° of the visual angle, corresponding to 0.31 m at 25 m distance) were hand-mapped by an experimenter naïve with respect to the research question and the return performance category to a corresponding AOI in the corresponding phase. If the marker left the AOI with its edge, it was considered to have left the area where it was located in the previous sample. The procedure allowed the calculation of fixation durations in each AOI in milliseconds. These were defined as focal visual fixations that lasted longer than 100 ms and did not move outside the respective AOI (size between 0.72° and 2° of visual angle). The fixation duration to each AOI in each individual return was calculated and used for further analysis.

Statistical Analysis

Data analyzes were performed using Orange data mining software (Orange 3.26.0, University of Ljubljana, Ljubljana, Slovenia). To analyze the relationships between gaze fixation duration in individual AOIs and categories such as individual server, returning player, or return performance category, the non-linear random forest machine learning approach was used (Liu et al., 2012). The duration of fixations in individual AOI during corresponding serve phases for both groups of participants were used as predictor variables, and the return performance category, returning player, or individual server as predicted classes. First, using Naïve Bayes, seven predictor variables were identified that allowed the highest prediction probability for the individual predicted classes identified *via* a nomogram (Zhang and Su, 2004; Shariat et al., 2009). The seven predictor variables were fed into the random forest machine learning algorithm to classify the data into specific subgroups for each of the predicted classes separately. To develop the machine learning classifier, the predictor variables from 1,020 returns (all players combined) were randomly split into five folds. Four folds were used for model training and cross-validated with the remaining fold, repeating the procedure for all folds. For the random forest, the number of trees was varied and the set of 13 trees with the split subset limit set to five enabled the highest accuracy of the model (Liu et al., 2012; Rigatti, 2017). The performance of the machine

TABLE 2 | Performance of different classification models.

	Both groups				National tennis players				International tennis players			
	AUC	CA	Se	Sp	AUC	CA	Se	Sp	AUC	CA	Se	Sp
Server	0.686	0.215	0.201	0.214	0.834	0.453	0.577	0.526	0.901	0.575	0.575	0.575
Returning player	0.953	0.664	0.735	0.664	0.970	0.780	0.801	0.780	0.883	0.642	0.710	0.642
Return category	0.667	0.523	0.595	0.523	0.717	0.581	0.621	0.581	0.753	0.549	0.583	0.549

AUC, area under the curve; CA, classification accuracy; Se, sensitivity; Sp, specificity are presented for classifying individual server, returning player, and return category for all returns combined and for returns made by national and international tennis players separately.

learning classifier for each data set was described by the area under the curve (AUC), classification accuracy (CA), sensitivity (true-positive rate), and specificity (false positive rate).

In the second step, two separate data sets were created for each group, and the procedure described above was repeated for the prediction of return performance, returning player, and individual server. By splitting the data sets, the specifics of each group could be examined.

Finally, for all three data sets (both groups and combined data from both groups), seven predictor variables were evaluated according to their importance in the machine learning classifier that the random tree method created. These were classified by fast correlation-based filter (FCBF), taking into account redundancy due to pairwise correlations between predictor variables (Lei and Liu, 2003). The first seven predictor variables that were >0.001 were included.

RESULTS

Classification Accuracy for Both Groups Together

The performance of the random tree classifier is presented in **Table 2**. For the data set combining tennis players from both groups, the highest predictability was observed for the returning player followed by the classification of the return performance category, and the lowest predictability was observed for the individual server. In addition, the highest sensitivity was observed for the returning player and the lowest for the individual server classification. Similarly, the highest specificity was observed for the returning player classification and the lowest for the individual server classification. Examples of the visual behavior in two returning players are presented by heat maps in **Figure 2**.

Classification Accuracy for the Group of National Players

Similar trends were observed in the data set consisting of only national level tennis players. The highest predictability was observed for the returning player classification, followed by the return performance category classification, and the lowest predictability was observed for the individual server classification. The highest sensitivity and specificity were observed for the returning player classification, followed by the return performance category classification and individual server classification.

Classification Accuracy for the Group of International Players

For the international player group, the trend was somewhat different. The highest predictability was observed for the returning player followed by the individual server classification and the lowest for the return performance category classification. The highest sensitivity and specificity were observed for the returning player classification followed by the individual server classification and return performance category classification.

Most Important AOI Attributes

Scores for individual predictor variables with FCBF exceeding 0.001 are presented in **Table 3**. Tossing hand movement area in the second phase of the movement of a server proved to be one of the most crucial predictors for classifying individual servers in both groups studied, especially in the national level tennis players (**Table 3**). Other predictor variables differed between predicting classes in both groups, with the international players showing the most important predictor variables in the first four phases of the movement of a server, and the national level players showed differences in all five phases.

The most successful predictor variables for the return performance category classification differed in the two observed groups of tennis players. The tossing hand movement area located in the second phase of the movement of a server for the international level tennis players and the area of hand-racket movement for the national level tennis players proved to rank highest between predictor variables (**Table 3**). International level tennis players tended to be classified based on a smaller number of AOI as in the national level tennis players. In general, the return performance category in the international level tennis players differed more by the length of visual focus duration directed to the tossing hand movement area in the second phase of the movement of a server and specific AOI in the fourth and fifth phases of the movement of a server as compared with national level tennis players who differed more in the amount of visual focus duration directed to the area of hand-racket movement in the third phase of the movement of a server and specific AOI in the fourth and fifth phases of the movement of a server.

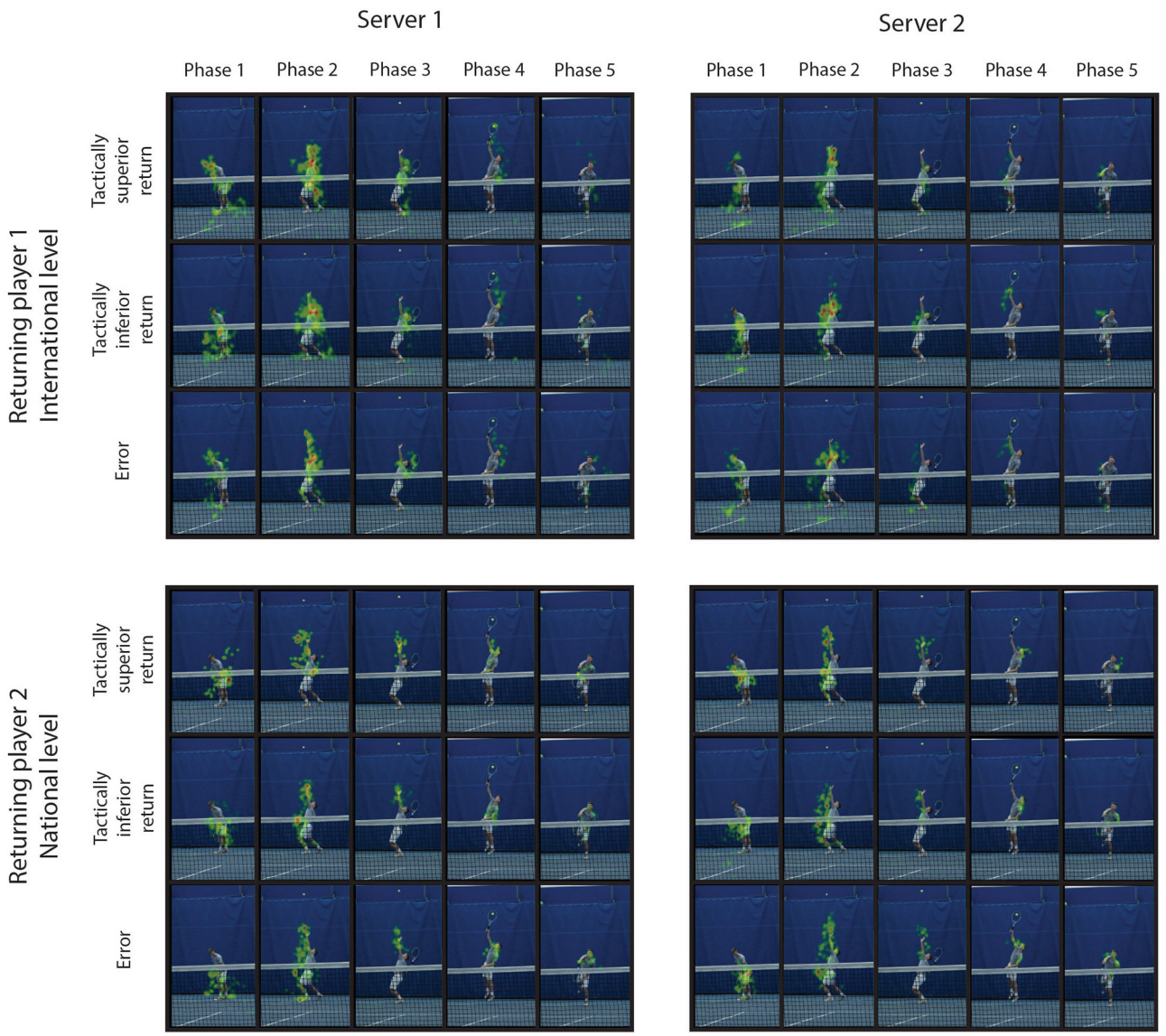


FIGURE 2 | Heat maps show examples of visual behavior in two returning players. Examples of heat maps for two returning players (upper and lower set of heat maps) each belonging to a different group are presented. The left and right sets of heat maps represent visual fixation duration and location while observing two different servers. The upper row of pictures in each individual set of heat maps depicts visual fixation duration in each phase of the movement of a server at tactically superior returns, the middle row depicts tactically inferior returns, and the bottom row represents the returns resulting in error.

As for classifying the returning player, the duration of visual focus of the international level tennis players differed more in observing the ball upward movement and tossing hand movement area in the second and third phases of the movement of a server (**Table 3**). In contrast, national level tennis players differed in visual focus duration in the tossing hand movement area during the second phase of the movement of a server and the area of hand and racket movement in the fourth phase of the movement of a server. In general, national level tennis players tend to differ from each other in their duration of focal visual attention in a higher number of AOI located in different phases of the movement of a server, which indicated higher variability between visual search strategies between national level tennis players.

DISCUSSION

In this study, the visual behavior during tennis serve return was investigated in relation to the returning player, individual server, and return performance category in two different groups of experts. The visual behavior was most strongly

TABLE 3 | Classifiers best predicting the individual server and return performance category.

	All		International tennis players		National tennis players	
	Classifiers	**FCBF**	**Classifiers**	**FCBF**	**Classifiers**	**FCBF**
Server	2—Tossing hand movement area	0.132	2—Tossing hand movement area	0,125	2—Tossing hand movement area	0.281
	2—Ball upwards movement	0.075	4—Area of hand-racket movement	0.097	3—Ball contact area	0.128
	5—Lower body	0.075	1—Area surrounding server	0.060	5—Lower body	0.127
	1—Area surrounding server	0.061	4—Head and upper body	0.059	1—Racket	0.108
	1—Racket	0.058	3—Area of hand-racket movement	0.044	2—Ball upwards movement	0.092
	2—Hips	0.057			3—Area of hand-racket movement	0.092
	2—Ball release	0.049			1—Area surrounding server	0.091
Return category	2—Tossing hand movement area	0.132	2—Tossing hand movement area	0.050	3—Area of hand-racket movement	0.068
	2—Ball upwards movement	0.075	5—Lower body	0.047	1—Racket	0.056
	5—Lower body	0.075	4—Area of hand-racket movement	0.042	5—Lower body	0.046
	1—Area surrounding server	0.061	4—Head and upper body	0.034	4—Head and upper body	0.039
	1—Racket	0.058	2—Ball release	0.012	2—Back leg	0.014
	2—Hips	0.057	1—Racket hand	0.009	2—Ball release	0.008
	2—Ball release	0.049			4—Head and upper body	0.008
Returning player	2—Tossing hand movement area	0.183	2—Ball upwards movement	0.129	2—Tossing hand movement area	0.324
	2—Ball upwards movement	0.130	2—Tossing hand movement area	0.112	4—Area of hand and racket movement	0.217
	4—Area of hand and racket movement	0.121	1—Area surrounding server	0.111	2—Ball upwards movement	0.186
	1—Area surrounding server	0.072	4—Head and upper body	0.105	2—Upper body	0.124
	5—Lower body	0.067	5—Lower body	0.078	5—Lower body	0.100
	2—Upper body	0.067			3—Area of hand-racket movement	0.089
	3—Area of hand-racket movement	0.057			1—Area surrounding server	0.066

Values of the fast correlation-based filter (FCB) are presented for seven most important predictor variables for classifying individual server, returning player, and return category for all returns combined and for returns made by national and international tennis players separately.

related to the returning player in both groups, indicating high interindividual variability as suggested by previous studies (Murray and Hunfalvay, 2017; Sáenz-Moncaleano et al., 2018). The second and third most successful classifications were for the return performance category and individual server, respectively, confirming our hypothesis that focal vision fixation duration at different AOIs was more related to the returning player than to other aspects such as individual server and return performance category. Interestingly, the individual server classification was low but more successful in the international group as compared with the national level tennis players. This suggests that the international players adapted visual search strategies to different servers more than the national level tennis players. The lowest ability to make predictions based on the duration and location of visual attention was observed for the return performance category. This was to be expected as serve return also depends on the observation of the ball flight phase (Sáenz-Moncaleano et al., 2018). In addition, seven or fewer AOIs from different phases of the movement of a server were identified that were best suited to differentiate between returning players, servers performing the tennis serve, and return performance categories.

The ability to classify an individual returning player suggests high interindividual variability and confirms findings from previous studies that participants use different visual search strategies (Button et al., 2011; Dicks et al., 2017). The availability of different sources of relevant visual cues and differences in the ability to interpret this information by individual performers could explain such differences between observers (Whiteside et al., 2013; Myers et al., 2017). This allows athletes to use different sources of visual information to perceive the movements of the same characteristics of opponents and consequently to vary their visual attention between repetitions and between individuals. This observation has been recently confirmed in field hockey goalkeepers, where different sources of visual information have been used by different goalkeepers all enabling successful saves, even after performing the same visual attention training intervention (Morris-Binelli et al., 2021).

Regardless of high interindividual variability, Alder et al. showed that visual search strategies in more skilled badminton players were highly related to the ability to attend to the most important visual information provided by the movements of an opponent (Alder et al., 2014). They showed that more skilled badminton players were able to focus on various kinematic cues that were better related to the trajectory of the ball and consequently were better able to make appropriate tactical solutions and shots. This was only partially observed in this study, with the visual behavior during the movement of a server being somewhat more strongly related to the return performance category in international players than in national level players. As suggested by other similar research, early information describing the movement of a server contributes significantly to online control of body movement and anticipation of ball flight direction and speed (van Soest et al., 2010; Müller and Abernethy, 2012; Triolet et al., 2013). This suggests that international tennis players in this study were more successful in using information from the movement of a server to predict the ball flight characteristics as their national counterparts. Our results add that a larger proportion of visual behavioral traits while observing movements of servers is related to intraindividual differences than to the return performance category, which is also in line with research performed by Morris-Binelli et al. (2021).

Interestingly, the observed classifications of the individual server were less accurate as for the returning players, especially among national-level tennis players. This may indicate that national-level tennis players in particular tend to use generic visual search strategies that are less attuned to the specifics of an individual server. In contrast, the visual search strategies of international level tennis players were more server-based, suggesting that the more experienced players are better able to adapt their visual search strategies to the specific visual cues of individual servers. However, these observations must be interpreted with caution. First, the sample size was small, particularly in the international tennis player group, as highly skilled international players are difficult to recruit. Second, an important limitation of this study was that differences between individual serving techniques were not examined. Therefore, future research should investigate whether the ability to adapt visual attention to the individual server is also related to other factors such as serve type, individual technique, and style and whether this adaptability affects interception performance.

The AOIs that were most important for classifying individual returning players when all participants were combined were in the first, second, fourth, and fifth phases of the movement of a server. Interestingly, the first two AOIs, namely, tossing hand movement area and ball upward movement, were also the most important predictors of return performance category and individual server. These results are consistent with the observations of Jackson and Mogan (2007), who pointed out the importance of ball toss for ball flight anticipation, as well as observations that important kinematic synergies are typical for these phases and can be specific for an individual server (Shafizadeh et al., 2019, 2020). Other AOIs located in the following phases of the movement of a server have been shown to contribute less to the classification of returning player and individual server. These observations are partially consistent with the results of studies performed in other interceptive tasks, where the importance of visual attention increases as the observer approaches closer to the movement initiation or interception (Button et al., 2011; Navia et al., 2017), such as movement initiation of returning players in this study. After the first two phases, the movement of a returning player is already initiated and begins to rely more on online control using information from ball flight characteristics.

The duration of focal vision fixation on movements of the tossing hand and the upward movement of the ball also differ significantly between the return performance categories. This suggests that players may interpret this phase of server action differently, which strongly influences the performance of the return, which is in line with the latest research (Morris-Binelli et al., 2021). As could be speculated on the above rationales, this phase is highly related to the first movement initiation of the returning player toward the interception point. If this movement initiation is delayed, the return performance might be compromised.

In the national level group, players also differed in visual attention duration focused on lower body parts such as hips and legs in the first three phases of the movement of a server. This can be better interpreted by including the results from the return performance category classification. The comparison of the two groups shows that in the international group, hand and racquet movements during all phases are the primary AOIs that relate to the return performance category. In contrast, in the national level group, visual attention duration to other AOIs, rather than hand and racquet movements, is related to the return performance category. Since more AOIs are important in the national level group for discriminating the return performance category, it could be speculated that their visual attention is directed to more different sources of visual information, which could reduce the time spend observing more important movements of the opponent. A limitation of this study was that the specific effect that each AOI had on classification was not examined. This would be of importance, as some AOIs that were found to be important classifiers could have a negative effect on return performance. As suggested by the studies on quiet

eye phenomena, more skilled athletes have fewer saccades and longer fixations to the most relevant AOI (Lebeau et al., 2016; Gonzalez et al., 2017). Applying these observations to this study, one could hypothesize that the longer duration of visual attention focused on the lower body of an opponent negatively affects the return performance.

Our results also show that the international players returning the serve also observe the area next to the tossing hand and the ball release area in the second phase of the movement of a server. These two areas were defined as the areas where the hand and the ball will be located in the upcoming moments. This visual behavior could be interpreted as anticipatory visual behavior or as more efficient buffering between visual information pickup and motor response. As suggested by Connor and Knierim (2017), the movement of focal vision to the areas of anticipated events suggests the anticipation of opponent movement and prompts recall of the anticipated action from visual memory. Being primarily present only in the international group, these tennis players might show better anticipation of the movement of a server, which could improve their movement response. Similar visual behavior has been reported by other studies (Land and Furneaux, 1997; Vickers and Adolphe, 1997; Furneaux and Land, 1999; Land and McLeod, 2000; Qian et al., 2019). These authors proposed that the earlier movement of focal vision relative to observer movement is enabled by visual working memory, which integrates information that cannot be observed simultaneously. As these studies show, more skilled performers have a higher buffering capacity, which allows for the integration of a greater amount of information. This could have important implications for responding to such complex visual stimuli as the tennis serve. Since noisiness of the movements of an opponent and a ball contains different relevant and irrelevant information, a higher visual memory capacity could be beneficial to include only the most relevant information. Based on the results of this study, it could be speculated that more skilled participants could have had superior visual memory capacity, besides more efficient information pickup strategies.

The classification of the return performance category suggests that the fourth and fifth phases of the movement of a server are important for success, but to a lesser extent. As our results suggest, the more experienced players are better able to extract information from the racquet upward movement, ball contact, and follow-through movement. According to van Soest et al. (2010), the final ball contact and the follow-through movement are characterized by funneling of the end effector (hand and racquet). Funneling represents a low degree of intertrial variability and is biomechanically highly representative of ball flight characteristics and therefore provides important visual cues for the tennis player returning the serve. As visual attention is more important at this stage for national level tennis players, it could be speculated that they rely more on online motion adaptation, possibly compensating for less efficient anticipation of ball flight characteristics during the movements of a server.

Overall, the results of this study suggest that a large interindividual difference in visual attention during serve return exists between tennis players. However, visual search strategies in more experienced tennis players may also adapt to the specific opponent. This is consistent with previous research showing more efficient anticipation based on the initial opponent movement. However, the results of this study add that more experienced athletes are better not only at predicting ball flight characteristics but also able to adapt to the specific constraints presented by an individual opponent.

These findings additionally present important implications for training tennis serve return. Bonato et al. (2020) reported on the positive effects of visual attention training on specific aspects of tennis serve return performance in junior tennis players. Findings from our research show that future training studies should additionally use tennis serve specific visual information, primarily focused on ball toss and ball vertical flight observation, which enable more efficient anticipation and movement initiation. This could be achieved *via* video or model-based observations or on-court training. Improved ability to recognize hand, ball, and racquet movement patterns could enable more efficient buffering of visual information, which could enable processing of other important visual cues that could altogether provide more reliable prediction of the ball flight trajectory. Moreover, such training routines should introduce variable servers and serve types to learn how to apply basic principles of visual search strategies to different opponents and serve types.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Slovenian Committee for Medical Ethics. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

JR and ZM: conceptualization, data analysis, investigation, resources, writing—review and editing, visualization, and supervision. JR: methodology and writing—original draft preparation. Both authors have read and agreed to the published version of the manuscript.

REFERENCES

Aglioti, S. M., Cesari, P., Romani, M., and Urgesi, C. (2008). Action anticipation and motor resonance in elite basketball players. *Nat. Neurosci.* 11, 1109–1102. doi: 10.1038/nn.2182

Alder, D., Ford, P., Causer, J., and Williams, A. (2014). The coupling between gaze behavior and opponent kinematics during anticipation of badminton shots. *Hum. Mov. Sci. Hum Mov Sci.* 37, 167–179. doi: 10.1016/j.humov.2014.07.002

Bonato, M., Gatti, C., Rossi, C., Merati, G., and La Torre, A. (2020). Effects of visual training in tennis performance in male junior tennis players: a randomized controlled trial. *J. Sports Med. Phys. Fit.* 60, 493–492. doi: 10.23736/S0022-4707.19.10218-6

Button, C., Dicks, M., Haines, R., Barker, R., and Davids, K. (2011). Statistical modelling of gaze behaviour as categorical time series: what you should watch to save soccer penalties. *Cogn. Process.* 12, 235–232. doi: 10.1007/s10339-010-0384-6

Connor, C. E., and Knierim, J. J. (2017). Integration of objects and space in perception and memory. *Nat. Neurosci.* 20, 1493–1492. doi: 10.1038/nn.4657

Dicks, M., Button, C., and Davids, K. (2010). Examination of gaze behaviors under *in situ* and video simulation task constraints reveals differences in information pickup for perception and action. *Attent. Percept. Psychophys.* 72, 706–702. doi: 10.3758/APP.72.3.706

Dicks, M., Button, C., Davids, K., Chow, J. Y., and van der Kamp, J. (2017). Keeping an eye on noisy movements: on different approaches to perceptual-motor skill research and training. *Sports Med.* 47, 575–572. doi: 10.1007/s40279-016-0600-3

Farrow, D., Abernethy, B., and Jackson, R. C. (2005). Probing expert anticipation with the temporal occlusion paradigm: experimental investigations of some methodological issues. *Motor Control* 9, 332–332. doi: 10.1123/mcj.9.3.330

Furneaux, S., and Land, M. F. (1999). The effects of skill on the eye-hand span during musical sight-reading. *Proc. Biol. Sci.* 266, 2435–2432. doi: 10.1098/rspb.1999.0943

Giblin, G., Whiteside, D., and Reid, M. (2017). Now you see, now you don't … the influence of visual occlusion on racket and ball kinematics in the tennis serve. *Sports Biomech.* 16, 23–22. doi: 10.1080/14763141.2016.1179337

Gonzalez, C. C., Causer, J., Miall, R. C., Grey, M. J., Humphreys, G., and Williams, A. M. (2017). Identifying the causal mechanisms of the quiet eye. *Eur. J. Sport Sci.* 17, 74–72. doi: 10.1080/17461391.2015.1075595

Jackson, R. C., and Mogan, P. (2007). Advance visual information, awareness, and anticipation skill. *J. Motor Behav.* 39, 341–342. doi: 10.3200/JMBR.39.5.341-352

Jones, C. M., and Miles, T. (1978). *Use of Advance Cues in Predicting the Flight of a Lawn Tennis Ball*. Available online at: https://paper/Use-of-advance-cues-in-predicting-the-flight-of-a-Jones-Miles/f34842637b152f0d2396428e488efa59b5d6145d (accessed March 18, 2021).

Land, M. F. (2009). Vision, eye movements, and natural behavior. *Vis. Neurosci.* 26, 51–52. doi: 10.1017/S0952523808080899

Land, M. F., and Furneaux, S. (1997). The knowledge base of the oculomotor system. *Philos. Trans. Royal Soc. Lond. Ser. B Biol. Sci.* 352, 1231–1232. doi: 10.1098/rstb.1997.0105

Land, M. F., and McLeod, P. (2000). From eye movements to actions: how batsmen hit the ball. *Nat. Neurosci.* 3, 1340–1342. doi: 10.1038/81887

Latash, M. (2010). Motor synergies and the equilibrium-point hypothesis. *Motor Control* 14, 294–292. doi: 10.1123/mcj.14.3.294

Latash, M., Scholz, J. P., and Schöner, G. (2007). Toward a new theory of motor synergies. *Motor Control* 11, 276–272. doi: 10.1123/mcj.11.3.276

Lebeau, J. C., Liu, S., Sáenz-Moncaleano, C., Sanduvete-Chaves, S., Chacón-Moscoso, S., Becker, J. B., et al. (2016). Quiet eye and performance in sport: a meta-analysis. *J. Sport Exerc. Psychol.* 38, 441–442. doi: 10.1123/jsep.2015-0123

Lei, Y., and Liu, H. (2003). ≫Feature selection for high-dimensional data: a fast correlation-based filter solution" in *Proceedings, Twentieth International Conference on Machine Learning, Vol. 2*. Washington, DC.

Liu, Y., Wang, Y., and Zhang, J. (2012). "New machine learning algorithm: random forest," in *Information Computing and Applications. Lecture Notes in Computer Science*, eds B. Liu, M. Ma, and J. Chang. (Berlin, Heidelberg: Springer), 246–52. doi: 10.1007/978-3-642-34062-8_32

Maselli, A., Dhawan, A., Cesqui, B., Russo, M., Lacquaniti, F., and d'Avella, A. (2017). Where are you throwing the ball? I better watch your body, not just your arm!' *Front. Hum. Neurosci.* 11:505. doi: 10.3389/fnhum.2017.00505

Mecheri, S., Gillet, E., Thouvarecq, R., and Leroy, D. (2011). Are visual cue masking and removal techniques equivalent for studying perceptual skills in sport? *Perception* 40, 474–472. doi: 10.1068/p6828

Mecheri, S., Laffaye, G., Triolet, C., Leroy, D., Dicks, M., Choukou, M. A., et al. (2019). Relationship between split-step timing and leg stiffness in world-class tennis players when returning fast serves. *J. Sports Sci.* 37, 1962–1962. doi: 10.1080/02640414.2019.1609392

Morris-Binelli, K., Müller, S., van Rens, F. E. C. A., Harbaugh, A. G., and Rosalie, S. M. (2021). Individual differences in performance and learning of visual anticipation in expert field hockey goalkeepers. *Psychol. Sport Exerc.* 52:101829. doi: 10.1016/j.psychsport.2020.101829

Müller, S., and Abernethy, B. (2012). Expert anticipatory skill in striking sports: a review and a model. *Res. Quart. Exerc. Sport* 83, 175–172. doi: 10.1080/02701367.2012.10599848

Murray, M. M., Thelen, A., Ionta, S., and Wallace, M. T. (2019). Contributions of intraindividual and interindividual differences to multisensory processes. *J. Cogn. Neurosci.* 31, 360–362. doi: 10.1162/jocn_a_01246

Murray, N. P., and Hunfalvay, M. (2017). A comparison of visual search strategies of elite and non-elite tennis players through cluster analysis. *J. Sports Sci.* 35, 241–242. doi: 10.1080/02640414.2016.1161215

Myers, N., Kibler, W., Lamborn, L., Smith, B., English, T., Jacobs, C., et al. (2017). Reliability and validity of a biomechanically based analysis method for the tennis serve. *Int. J. Sports Phys. Ther.* 12, 437–449.

Navia, J. A., Dicks, M., van der Kamp, J., and Ruiz, L. M. (2017). Gaze control during interceptive actions with different spatiotemporal demands. *J. Exp. Psychol. Hum. Percept. Perform.* 43, 783–782. doi: 10.1037/xhp0000347

Paeye, C., and Madelain, L. (2014). Reinforcing saccadic amplitude variability in a visual search task. *J. Vis.* 14:20. doi: 10.1167/14.13.20

Qian, J., Zhang, K., Liu, S., and Lei, Q. (2019). The transition from feature to object: storage unit in visual working memory depends on task difficulty. *Mem. Cogn.* 47, 1498–1492. doi: 10.3758/s13421-019-00956-y

Rigatti, S. J. (2017). Random forest. *J. Insur. Med.* 47, 31–32. doi: 10.17849/insm-47-01-31-39.1

Sáenz-Moncaleano, C., Basevitch, I., and Tenenbaum, G. (2018). Gaze behaviors during serve returns in tennis: a comparison between intermediate- and high-skill players. *J. Sport Exerc. Psychol.* 40, 49–42. doi: 10.1123/jsep.2017-0253

Shafizadeh, M., Bonner, S., Barnes, A., and Fraser, J. (2020). Effects of task and environmental constraints on axial kinematic synergies during the tennis service in expert players. *Eur. J. Sport Sci.* 20, 1178–1172. doi: 10.1080/17461391.2019.1701093

Shafizadeh, M., Bonner, S., Fraser, J., and Barnes, A. (2019). Effect of environmental constraints on multi-segment coordination patterns during the tennis service in expert performers. *J. Sports Sci.* 37, 1011–1012. doi: 10.1080/02640414.2018.1538691

Shariat, S. F., Karakiewicz, P. I., Godoy, G., and Lerner, S. P. (2009). Use of nomograms for predictions of outcome in patients with advanced bladder cancer. *Therapeut. Adv. Urol.* 1, 13–12. doi: 10.1177/1756287209103923

Triolet, C., Benguigui, N., Le Runigo, C., and Williams, A. M. (2013). Quantifying the nature of anticipation in professional tennis. *J. Sports Sci.* 31, 820–822. doi: 10.1080/02640414.2012.759658

van Soest, A. J. K., Casius, L. J. R., de Kok, W., Krijger, M., Meeder, M., and Beek, P. J. (2010). Are fast interceptive actions continuously guided by vision? revisiting Bootsma and van Wieringen (1990). *J. Exp. Psychol. Hum. Percept. Perform.* 36, 1040–2. doi: 10.1037/a0016890

Vaziri-Pashkam, M., and Xu, Y. (2017). Goal-directed visual processing differentially impacts human ventral and dorsal visual representations. *J. Neurosci. Off. J. Soc. Neurosci.* 37, 8767–8762. doi: 10.1523/JNEUROSCI.3392-16.2017

Vickers, J. N., and Adolphe, R. M. (1997). Gaze behaviour during a ball tracking and aiming skill. *Int. J. Sports Vis.* 4, 8–2.

Whiteside, D., Elliott, B., Lay, B., and Reid, M. (2013). A kinematic comparison of successful and unsuccessful tennis serves across the elite development pathway. *Hum. Mov. Sci.* 32, 822–835. doi: 10.1016/j.humov.2013.06.003

Woolley, T. L., Crowther, R. G., Doma, K., and Connor, J. D. (2015). The use of spatial manipulation to examine goalkeepers' anticipation. *J. Sports Sci.* 33, 1766–1762. doi: 10.1080/02640414.2015.1014830

Wright, M. J., Bishop, D. T., Jackson, R. C., and Abernethy, B. (2010). Functional MRI reveals expert-novice differences during sport-related anticipation. *Neuroreport* 21, 94–92. doi: 10.1097/WNR.0b013e328333dff2

Zhang, H., and Su, J. (2004). "Naive bayesian classifiers for ranking," in *Machine Learning: ECML 2004, Lecture Notes in Computer Science*, eds J. F. Boulicaut, F. Esposito, F. Giannotti, and D. Pedreschi (Berlin; Heidelberg: Springer), 501–512. doi: 10.1007/978-3-540-30115-8_46

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Psychology

ORIGINAL RESEARCH
published: 31 January 2022
doi: 10.3389/fpsyg.2021.788498

The Differences in the Performance Profiles Between Native and Foreign Players in the Chinese Basketball Association

Xing Wang[1,2†], Bin Han[3†], Shaoliang Zhang[4], Liqing Zhang[1], Alberto Lorenzo Calvo[2] and Miguel-Ángel Gomez[2]*

[1] Sport Coaching College, Beijing Sport University, Beijing, China, [2] Facultad de Ciencias de la Actividad Física y del Deporte, Universidad Politécnica de Madrid, Madrid, Spain, [3] College of General Education, Guangdong University of Science and Technology, Dongguan, China, [4] Division of Sport Science and Physical Education, Tsinghua University, Beijing, China

OPEN ACCESS

Edited by:
Roberta Antonini Philippe, University of Lausanne, Switzerland

Reviewed by:
Juan Pablo Morillo Baro, University of Malaga, Spain
Antonio Hernández-Mendo, University of Malaga, Spain

****Correspondence:***
Liqing Zhang
cherry4911@163.com

[†] These authors have contributed equally to this work and share first authorship

Specialty section:
This article was submitted to Performance Science, a section of the journal Frontiers in Psychology

Received: *02 October 2021*
Accepted: *16 December 2021*
Published: *31 January 2022*

Citation:
Wang X, Han B, Zhang S, Zhang L, Lorenzo Calvo A and Gomez M-Á (2022) The Differences in the Performance Profiles Between Native and Foreign Players in the Chinese Basketball Association. Front. Psychol. 12:788498. doi: 10.3389/fpsyg.2021.788498

The aim of the study was to (i) use an clustering analysis method to classify and identify native and foreign basketball players into similar groups based on game-related statistics; (ii) use the Pearson's Chi-square test to identify the key clusters that affect whether a team enters the playoffs; and (iii) use the classification tree analysis to stimulate the prediction of team ability and the construction of the team roster. The sample consisted of 422 foreign players and 1,775 native players across 9 seasons from 2011 to 2019. The clustering process allowed for the identification of nine native and six foreign player performance profiles. In addition, two clusters ($p < 0.001$, ES = 0.33; $p < 0.001$, ES = 0.28) of native players and one cluster ($p < 0.05$, ES = 0.16) of foreign players were identified that had a significant impact on team ability. These results provide alternative references for basketball staff concerning the process of evaluating native and foreign player performance in the Chinese Basketball Association.

Keywords: performance analysis, game statistic, cluster analysis, performance profiles, Chinese Basketball Association

INTRODUCTION

The process of player selection and team formation in basketball is regarded as a key factor to achieve successful game performances (Zhang et al., 2018). The selection of players in a team is a difficult decision-making task with many dimensions (Tavana et al., 2013). Coaches and managers are required to consider their technical and tactical performances, physical and physiological characteristics, or mental and psychological factors (Arnason et al., 2004). There is a huge gap between the best and worst players in terms of technical and tactical performances in the Turkish Basketball League (Özmen, 2019). Specifically, the shooting efficiency of foreign players was greater than native players, so the selection of core players may be the key to perform successfully in the league.

Success can be mainly dependent on the combination of players with complementary skills who are capable of performing according to the demands of the playing positions (Ige and Kleiner, 1998). Previously, the majority of studies were based on traditional player positions (guards, forward, and centers) to evaluate technical and physical performances (Page et al., 2007;

Sampaio et al., 2010; Pojskic et al., 2015; Gasperi et al., 2020). For example, Sampaio et al. (2006b) reported that forward were demonstrated to exhibit greater shooting efficacy inside the paint, which contributes more to game outcome than the efficacy of guards and centers. However, with the development of physical and technical performances of players, more players were able to play multiple roles on the court. Over the past few years, basketball has been considered more of a "position-less" team sport (Lutz, 2012; Samuel Kalman, 2020). Especially in the National Basketball Association (NBA), the "small ball" trend led by the Golden State Warriors promoted the revolution of modern basketball (Teramoto and Cross, 2017). The available research redefined nine playing positions of NBA players (Samuel Kalman, 2020) and predicted optimal lineups based on game-related statistics. Likewise, 13 positions were identified by the topological network in the NBA, which redefined a much finer stratification of NBA players such as "All star NBA," "All star NBA 2nd Team," "Paint Protectors," and "Role Players" (Lum et al., 2013). These algorithms provided a novel perspective to evaluate game performance. Similarly, Zhang et al. (2018) reported that players from different levels of teams in the NBA were distributed in five clusters according to the anthropometric attributes and playing experience. Most players from stronger teams were allocated to the low height and weight with middle experiences group while those from weaker teams were mainly distributed in the low height and weight with low experiences group. In addition, Mateus et al. (2020) used a two-step cluster model to identify three and five different performance profiles for Euroleague and national championships, and found that better performances of players may be attributed to more playing time on court, the age or playing position, as well as the competition level. However, to our knowledge, there is no study to identify this position-less phenomenon so far in the Asian basketball leagues. Therefore, it is necessary to assist coaches in understanding the detailed characteristics of different players from Asian basketball leagues in order to improve the recruitment and selection of the core players that make a huge contribution to team success.

Based on the above considerations, the aim of the present study was to (i) use an unsupervised clustering method to classify and identify native and foreign basketball players into similar groups based on game-related statistics in the Chinese Basketball Association (CBA); (ii) identify the key player clusters that affect whether a team enters the playoffs; and (iii) use classification tree analysis to stimulate the prediction of team ability and the construction of the team roster. Our study hypothesized that different levels of teams have different team characteristics according to the refined playing positions provided by cluster analysis.

MATERIALS AND METHODS

Data Collection and Pre-processing

The data were collected from RealGM[1] during the season period from 2011 to 2019. A total of 3,177 individual profiles were selected, including 577 foreign players and 2,600 native players (each sample represented each player's data in one season). Moreover, players who played less than 10 games in the whole season and had an average playing time of less than 5 min were excluded from the final sample because these players' transformed data were regarded as unreliable statistics (Kubatko et al., 2007). Then, the datasets were finally limited to 422 foreign players and 1,775 native players. The study was conducted in accordance with the Declaration of Helsinki (WMA, 2000; Bošnjak, 2001; Tyebkhan, 2003).

Variable Selection

The initial 39 variables were selected based on box-score and advanced statistics. The box-score statistics were transformed to per-minute statistics (original statistics/min × 40) according to players' game duration on the court (Kubatko et al., 2007). According to the available literature, a total of 20 variables were selected for analysis (see **Table 1**). The top four variables [height, weight, player efficacy rating (PER), points scored per 40 min (PTS)] were excluded from clustering analysis, and were only presented as descriptive analysis (Zhang et al., 2018).

In order to test the validity of datasets, a sub-sample of 50 games (at least five games in each season) was randomly selected and observed by two experienced analysts (basketball video coordinators with more than 5 years of experience in basketball performance analysis) by using Catapult Vision. The results were contrasted with the gathered data in the website in order to provide internal validity (ICC = 0.91) and external validity using generalizability analysis (generalizability coefficient, e2 = 0.96; and reliability coefficient, $\Phi = 0.65$) (Blanco Villaseñor et al., 2014; Hernández-Mendo et al., 2016; Royuela et al., 2017; Reigal et al., 2020). There was formal approval of all procedures from the Local Institution of Research Review Board.

Statistical Analysis

Firstly, model-based cluster analysis within Gaussian finite mixture models (GMM) was carried out to classify native and foreign players into different groups according to selected variables (Lutz, 2012; Samuel Kalman, 2020). GMM clustering results in a soft assignment, indicating the probability that each player belongs to a cluster (Fraley and Raftery, 1998). The algorithm of GMM clustering calculates the maximum-likelihood estimate (MLE) of Equation 1 to find the optimal distribution underlying the unlabeled data. The above procedure used the "mclust" package in R (Scrucca et al., 2016).

Secondly, we used obtained player clusters to build a lineup of each team. Since the CBA official bans trading native players during the season, the lineup of native players consisted of all native players belonging to the team, and we counted the number of each cluster (including starters and non-starters). As to foreign players, since teams had a limit on the number of foreign players they could replace during the season and only two or three foreign players were allowed at the same time, the lineup of foreign players consisted of foreign players whose number of games played

[1] http://basketball.realgm.com

was in the top 2 in the whole season. The team lineup was combined using native and foreign lineups as follows:

$$Lineup = N1 + N2 + N3 + N4 + N5 + N6 + N7 + N8 + N9 + F1 + F2 + F3 + F4 + F5$$

Where each cluster variable represented the number of players belonging to this cluster in the team.

According to the team rankings of each season, the teams were classified into "playoffs teams" and "non-playoffs teams." Then, a descriptive and inferential analysis was performed using the crosstabs command. The Pearson's Chi-square test was used to analyze the effects between team abilities and the number of each player clusters in the team lineup. Each player cluster in each team was considered an independent sampling unit, the interaction with teammates was disregarded. Effect sizes (ES) were calculated using the Cramer's V-test and their interpretation was based on the following criteria: 0.10 = small effect, 0.30 = medium effect, and 0.50 = large effect (Volker, 2006). The above procedure was run using the IBM SPSS statistical software for Windows, version 20.0 (Armonk, NY: IBM. Corp.).

Thirdly, a classification tree analysis (CART) was used to simulate the decision-making process of team lineup construction. The CART technique splits the sample into segments that are as homogeneous as possible in relation to the dependent variable (playoffs/non-playoffs). Since the algorithm is non-parametric and non-linear, it is often able to uncover complex interactions between predictors which may be difficult or impossible to uncover using traditional multivariate techniques (Lewis, 2000). This statistical analysis was performed using the "Rpart" package in R (Computing, 1991; Therneau et al., 2015), version 4.0.2.

RESULTS

The model-based clustering analysis allowed us to obtain nine clusters of native players (N1-N9) and six clusters of foreign players (F1–F6).

Defining the Nine Playing Positions of Native Players

Figure 1 presents the native players' performance profiles, and the definitions of the nine native players' clusters are as follows:

N1——"Floor General": the average height and weight of N1 were the lowest among all clusters but with the highest in assists and steal. Most of the point guards who prefer pass-first were grouped, most of this cluster were playmakers.
N2——"Sixth Man" had the second lowest average playing time (15 min per game) among all clusters but with high usage.
N3——"Rotation Big" was one of the tree clusters with average height over 205 cm but the average playing time was the lowest among the clusters.
N4——"Shooting Guard": the average playing time of N4 was the second highest among all nine clusters, with high average 3-pointers made and missed but low PER.
N5——"Three-Point Shooting Forward" had the same average 3-pointers made and miss statistics but with the lowest 2-pointers made and missed among all clusters.

TABLE 1 | Selected game related variables.

Variables (abbreviation)	Description
Height	Player height, in centimeters.
Weight	Player weight, in kilograms.
PER	Player efficiency rating statistic created by John Hollinger.
PTS	Points that a player scored per 40 min.
MIN	Minutes a player played on court per game.
2PM	The number of two-point field goals that a player has successfully made per 40 min.
2Pm	The number of two-point field goals that a player or team has unsuccessfully made per 40 min.
3PM	The number of three-point field goals that a player or team has successfully made per 40 min.
3Pm	The number of three-point field goals that a player or team has unsuccessfully made per 40 min.
FTM	The number of free throws that a player or team has successfully made per 40 min.
FTm	The number of free throws that a player or team has unsuccessfully made per 40 min.
FTRATE	The number of free throws made per field goals attempted per 40 min.
TOV	A turnover occurs when a player on offense loses the ball to the defense per 40 min.
AST	An assist occurs when a player completes a pass to a teammate that directly leads to a field goal per 40 min.
STL	A steal occurs when a defensive player takes the ball from a player on offense per 40 min.
BLK	A block occurs when an offensive player attempts a shot, and a defensive player tips the ball, blocking their chance to score per 40 min.
PF	The total number of fouls that a player has committed per 40 min.
OREB	The number of rebounds that a player has collected while they were on offense per 40 min.
DREB	The number of rebounds that a player has collected while they were on defense per 40 min.
USG	The percentage of plays utilized by a player while he is in the game.

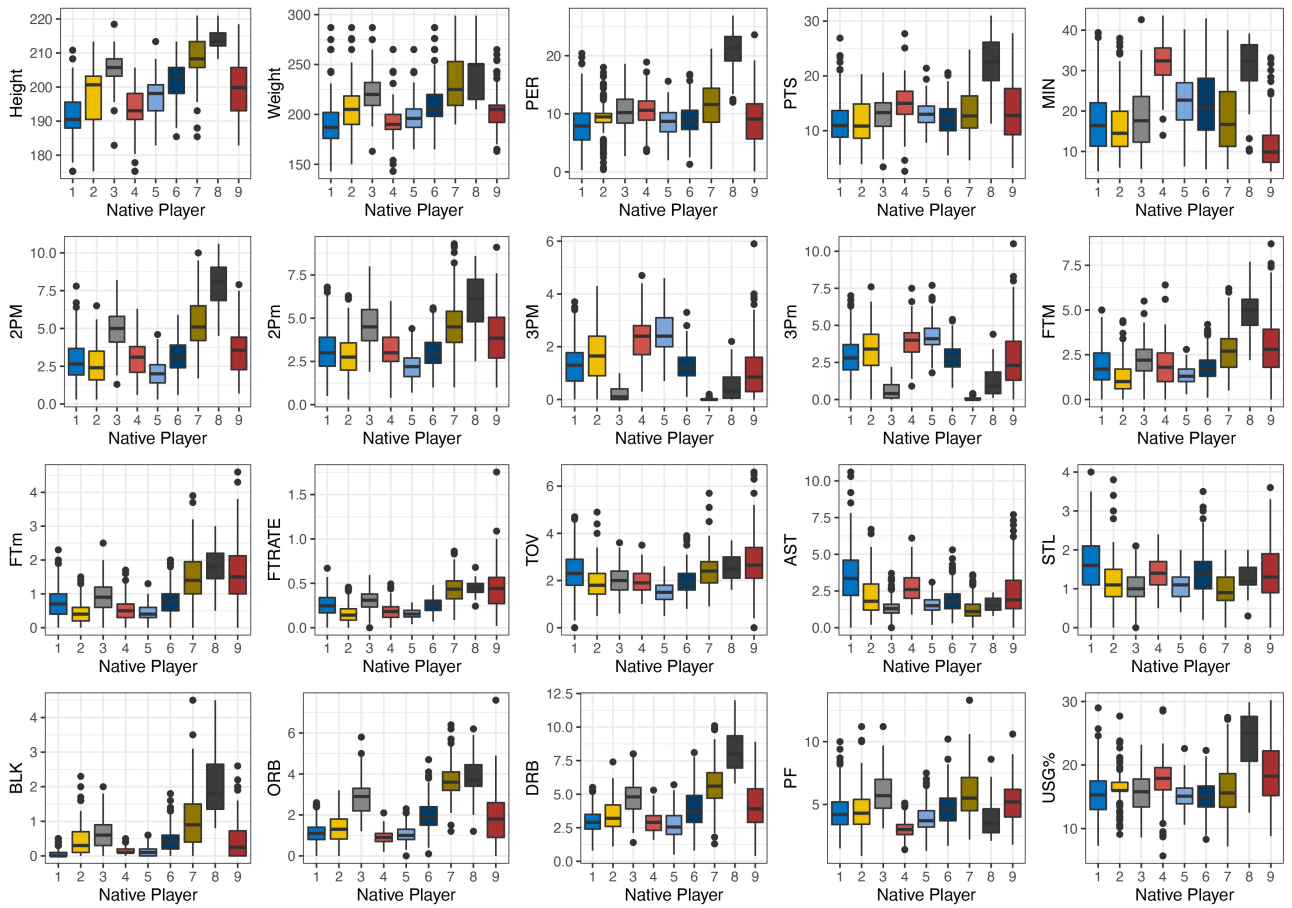


FIGURE 1 | Descriptive statistics of different performance profile clusters in native players.

N6——"Skilled Forward" was slightly higher than average in all game-related statistics but with no outstanding feature.

N7——"Defensive Big" was slight higher than N3, with higher offensive rebound and defensive rebound.

N8——"Dominant Center" was the highest in most statistics (i.e., height, PER, PTS, 2-pointers made, USG%) but low in assists, steal, and 3-pointers made and missed.

N9——"Bench Marginal Players": players from the bench always played below 10 min in garbage time. Most of the clusters were young players.

Defining the Six Playing Positions of Foreign Players

Figure 2 presents the foreign players' performance profiles, and the definitions of six clusters in foreign players are as follows:

F1——"Traditional Centers": players whose average height and weight were the highest among all clusters and excellent in defensive rebound, offensive rebound, and blocking shots.

F2——"Space Stretch Forward": the average height of F2 was more than 200 cm and had good 3-pointer shoot ability, meanwhile they could guarantee some defensive rebounds. These players stood outside the three-point line on offense most of the time, which meant they did not have many opportunities to take offense rebounds than other big players.

F3——"Mid-Range Skilled Forward": players whose role was to get the ball at midrange and low post areas according to its two-point field goals variables were able to create offensive opportunities by isolation and jump shot skill with few assists.

F4——"Three-Point Shooting Guards": the small players who had a high-level 3-pointer shooting ability and infinite shooting privilege was evident on three-point field goals made and missed variables but had the lowest free throws rate. In addition, these players had the second highest usage rate among all foreign player clusters.

F5——"Traditional Point Guard": this cluster includes players with the highest assists and steals but were average at other variables especially in terms of shooting, representing the Traditional Point Guard who prefers to be a team leader by assisting teammates to score than scoring by themselves. It makes them less outstanding on PER compared to other cluster players.

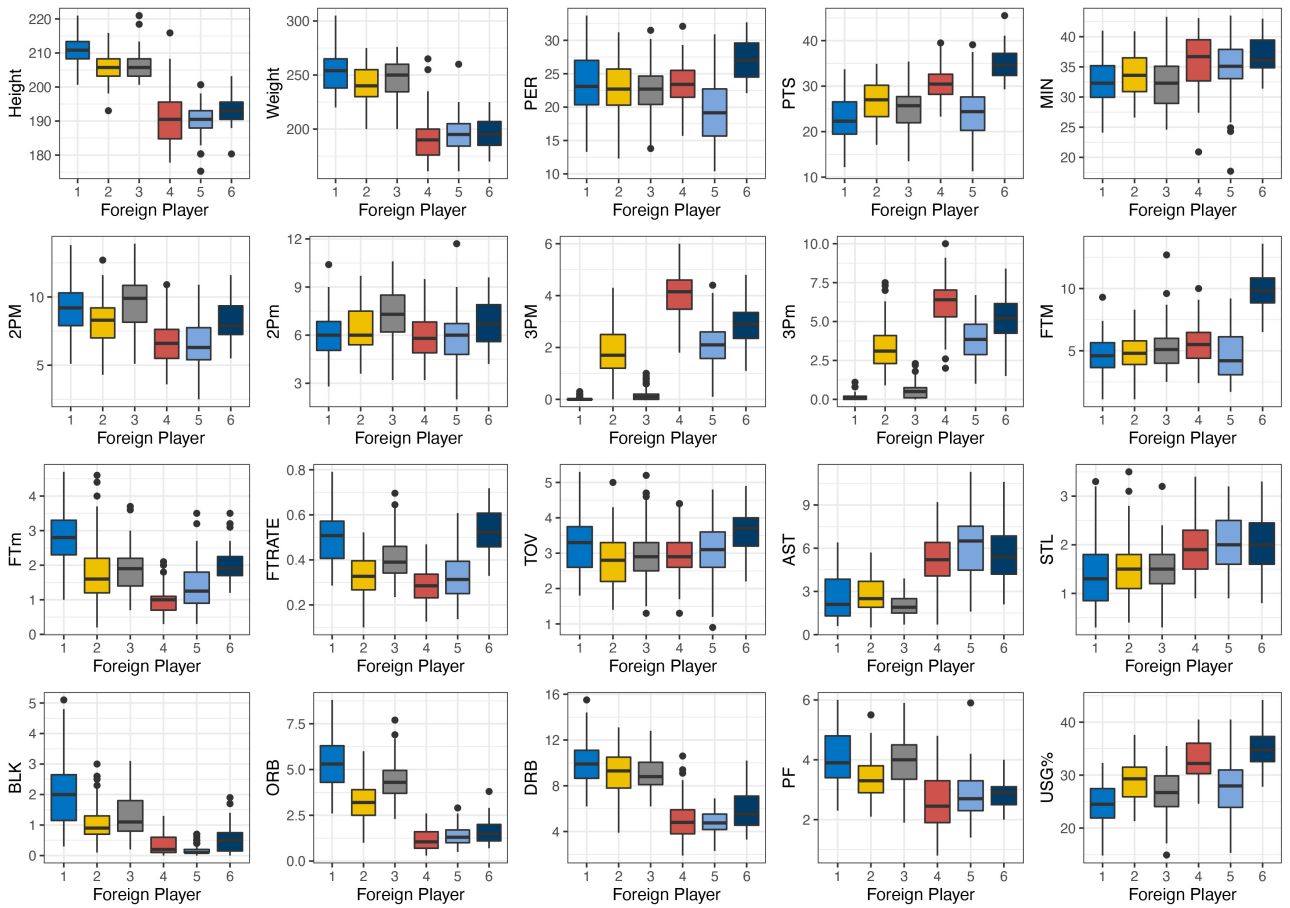


FIGURE 2 | Descriptive statistics of different performance profile clusters in foreign players.

F6——"Dominant Point Guard" includes small players who operate with the ball in their hands and play more aggressively than the Traditional Point Guard. It is worth mentioning that these players are good at scoring by drawing fouls which ensures that they accumulate more free throw field goals than others.

Crosstabs Analysis in Team Composition

The sample distribution of the number of native player clusters for a team is presented in **Table 2**. It shows that in native players only N7 Defensive Big was statistically significant ($P < 0.001$). When a team had more than two Defensive Big players it was easier to reach playoffs (30.6% compared to 25.3% when there were two Defensive Big players in the team; 18.1% compared to 2.2% for two; 1.4% compared to 0.0% for four). Conversely, the team had a low probability of entering the playoffs when no or only one Defensive Big player was in the team lineup (12.5% compared to 28.6% for 0; 37.5% compared to 44.0% for 1).

In addition, N8 Dominant Center had the same positive role as N7 Defensive Big. The result showed that when a team had one Dominant Center player the team had more chances to make playoffs (25.0% compared to 5.5% for one Dominant Center player in the team). But when there was a lack of Dominant Center players in the team, it was more difficult to make the playoffs (75.0% compared to 93.4%).

The result for foreign player clusters (**Table 3**) showed that only F6 Dominant Point Guard was significantly related to team ability. When there was a Dominant Point Guard foreign player in the team, the probability of the team entering the playoffs was lower than not making the playoffs (27.5% compared to 13.9%). Conversely, the team had a high probability of entering the playoffs when no Dominant Point Guard was in the team lineup (86.1% compared to 72.9%).

The classification and regression tree analysis included both native and foreign player cluster variables in the statistical model. **Figure 3** shows that, after pruning by the minimum error algorithm, a total of 21 nodes were defined which included 10 parent nodes and 11 leaf nodes. Each parent node was split by a player cluster variable. The splitting variables for the top 3 parent nodes were the same as the significant variables provided by crosstab analysis (N7 Defensive Big, N8 Dominant Center, F6 Dominant Point Guard). In addition, another four variables (N1 Floor General, N4 Shooting Guard, N9 Bench Marginal Player, N5 Three-Point Shooting Forward) were also considered as splitting variables in the final tree. Each leaf node provided the probability of the team in this cluster of entering the playoffs and

TABLE 2 | Frequency distribution (%) of team ability according to the number of native player clusters (crosstab command: Pearson's Chi-square, degrees of freedom, significance, expected frequency distribution, and effect size).

	Playoffs *n* = 72		Non-playoffs *n* = 91						
Number of players	**%**	***n***	**%**	***n***	χ^2	**df**	***P***	**EFD**	**ES**
N1 Floor general									
0	5.6	4	1.1	1	8.099	6	0.261	1.32†	0.22
1	27.8	20	19.8	18					
2	26.4	19	22.0	20					
3	19.4	14	25.3	23					
4	16.7	12	24.2	22					
5	4.2	3	4.4	4					
6	0.0	0	3.3	3					
N2 Sixth man									
0	62.5	45	52.7	48	8.611	5	0.127	1.32†	0.22
1	23.6	17	30.8	28					
2	5.6	4	9.9	9					
3	1.4	1	5.5	5					
4	2.8	2	1.1	1					
5	4.2	3	0.0	0					
N3 Rotation big									
0	24.7	25	25.3	23	2.183	3	0.57	3.09†	0.11
1	41.7	30	42.6	42					
2	20.8	15	23.1	21					
3	2.8	2	5.5	5					
N4 Shooting guard									
0	33.3	24	37.4	34	0.307	3	0.933	0.88†	0.043
1	55.6	40	52.7	48					
2	9.7	7	8.8	8					
3	1.4	1	1.1	1					
N5 Traditional point guard									
0	18.1	13	18.7	17	5.009	3	0.171	7.06	0.17
1	43.1	31	39.6	36					
2	34.7	25	27.5	25					
3	4.2	3	14.3	13					
N6 Skilled forward									
0	12.5	9	8.8	8	8.535	5	0.109	1.32†	0.22
1	31.9	23	29.7	27					
2	41.7	30	33.0	30					
3	9.7	7	17.6	16					
4	1.4	1	9.9	9					
5	2.8	2	1.1	1					
n7 Defensive big									
0	12.5	9	28.6	26	17.896	4	0.001**	0.44†	0.33
1	37.5	27	44.0	40					
2	30.6	22	25.3	23					
3	18.1	13	2.2	2					
4	1.4	1	0.0	0					
N8 Dominant center									
0	75.0	54	93.4	85	13.226	2	0.001**	0.44†	0.28
1	25.0	18	5.5	5					
2	0.0	0	1.0	1					

(Continued)

TABLE 2 | (Continued)

	Playoffs *n* = 72		Non-playoffs *n* = 91						
Number of players	**%**	***n***	**%**	***n***	χ^2	**df**	***P***	**EFD**	**ES**
N9 Bench marginal player									
0	55.6	40	59.3	54	2.006	4	0.799	0.44†	0.11
1	31.9	23	30.8	28					
2	11.1	8	6.6	6					
3	1.4	1	2.2	2					
4	0.0	0	1.1	1					

**P < 0.05; **P < 0.01; EFD, expected frequency distribution; †When EFD was below 5 or the variable includes values below 1%, the Fisher's exact test was applied; ES, effect size.*

TABLE 3 | Frequency distribution (%) of team ability according to the number of foreign player clusters (crosstab command: Pearson's Chi-square, degrees of freedom, significance, expected frequency distribution, and effect size).

	Playoffs *n* = 72		Non-playoffs *n* = 91						
Number of players	**%**	***n***	**%**	***n***	χ^2	**df**	***P***	**EFD**	**ES**
F1 Traditional center									
0	81.9	59	75.8	69	0.893	1	0.345	15.46	0.07
1	18.1	13	24.2	22					
F2 Space stretch forward									
0	59.7	43	62.6	57	1.919	2	0.513	0.88†	0.11
1	40.3	29	35.2	32					
2	0.0	0	2.2	2					
F3 Mid-range skilled forward									
0	56.9	41	56.0	51	0.152	2	1	3.53†	0.03
1	38.9	28	38.5	35					
2	4.2	3	5.5	5					
F4 Three-point shooting guard									
0	50.0	36	57.1	52	3.745	2	0.198	1.33†	0.15
1	50.0	36	39.6	36					
2	0.0	0	3.3	3					
F5 Traditional point guard									
0	58.3	42	72.5	66	4.201	2	0.0938	1.33†	0.16
1	40.3	29	25.3	23					
2	1.4	1	2.2	2					
F6 Dominant point guard									
0	86.1	62	72.5	66	4.399	1	0.036*	15.46	0.16
1	13.9	10	27.5	25					

**P < 0.05; **P < 0.01; EFD, expected frequency distribution; †When EFD was below 5 or the variable includes values below 1%, the Fisher's exact test was applied; ES, effect size.*

the probabilities of six nodes (8, 36, 148, 150, 302, and 38) were lower than 50%, and five nodes (149, 303, 39, 5, and 3) were more than 50%. Among them, the lowest probability of entering the playoffs was node 148 in which only 7% of teams in this node were likely to enter the playoffs. In contrast, 100% of the teams in node 39 could enter the playoffs, but the sample size was only 4% of the total sample.

The root node (node 1) was split by N7 Defensive Big. High probabilities (88%) to make the playoffs were evident when the obtained values for N7 Defensive Big were higher than 2.5 (node 3) and, conversely, lower chances (39%) to make the playoffs were seen when the obtained values for assists were equal or lower than 2 (node 2). Based on the number of N8 Dominant Centers, node 2 was split into node 4 and leaf node 5. Leaf node 5 showed that in the team that had no more than two N7 Defensive Bigs, if there were more than one N8 Dominant Center, the probability of this team making the playoffs would be 70%. Node 4 was further split into leaf node 8 and node 9 by the number of F6 Dominant Point Guards. When there were less than two N7 Defensive Bigs and no N8 Dominant Centers in the team, and if the team had an F6 Dominant Point Guard in the lineup, the team had an 81% probability of not making the playoffs.

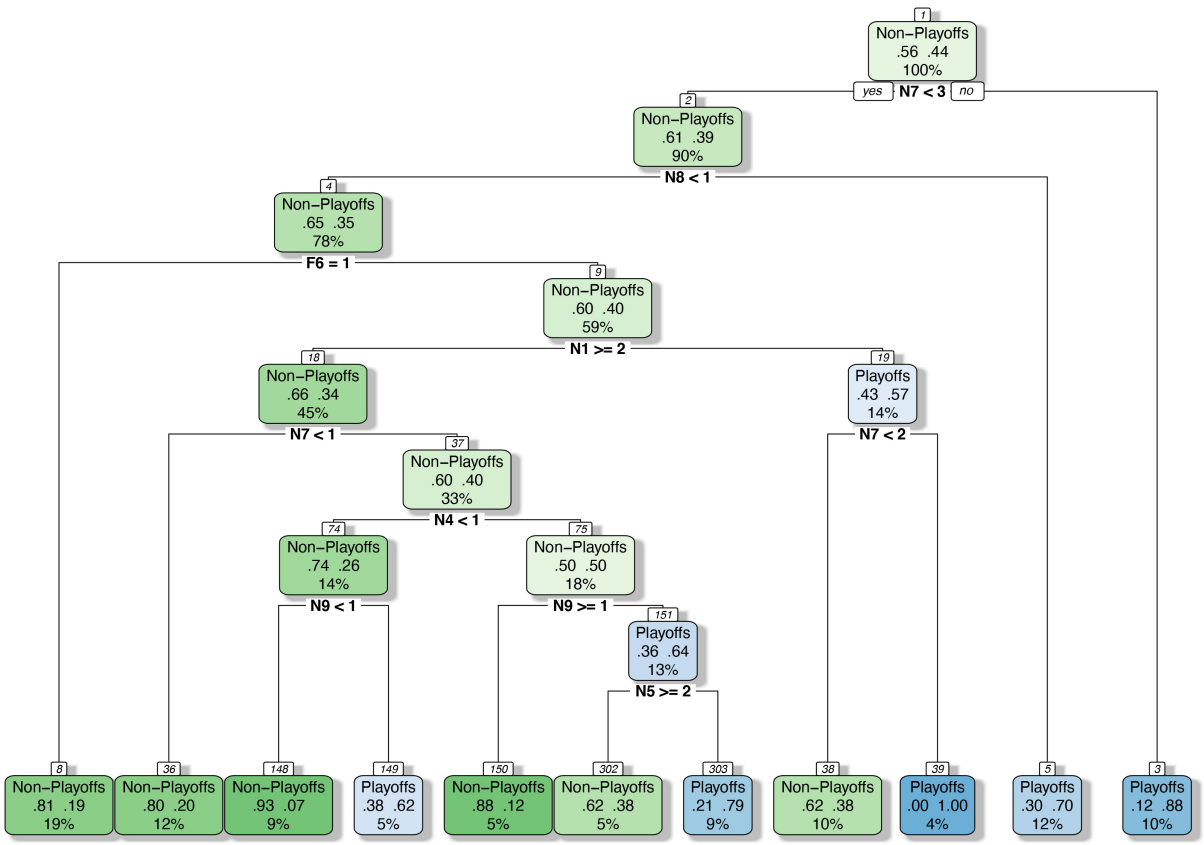


FIGURE 3 | Classification and regression tree analysis of team abilities.

DISCUSSION

The aim of the present study was to (i) use an unsupervised clustering method to classify and identify native and foreign basketball players into similar groups based on game-related statistics; (ii) identify the key clusters that affect whether a team enters the playoffs; and (iii) use classification tree analysis to stimulate the prediction of team ability and the construction of the team roster. It was expected that some players would have a significant impact on the strength of the team (i.e., all-star players, scoring players, or defensive players). Our results revealed a discrepancy of individual performance with nine clusters identified for the native players and six clusters for the foreign players. Furthermore, three clusters of players highlighted significantly different distributions in playoffs and non-playoffs teams. These findings will be of extreme importance for coaches and managers in CBA when recruiting players and building team lineups based on players' strengths and weaknesses, playing position, and nationality.

Difference Between Player Clusters in the Roster

Based on lineups built by new players' clusters, the crosstabs command analysis identified that the number of players from three clusters showed significant differences between playoffs and non-playoffs teams. In native players, the most important playing position was Dominant Center which is linked with previous studies reporting that the most prominent performance characteristics of Dominant Centers are closely related to the team's wins and losses (i.e., two-point field goals made, free throws made, defensive rebounds, and blocked shots) in high-level competition (Çene, 2018). However, talented players can be defined as Dominant Centers and are extremely rare in the league (1.75% of all native players) which means that most teams usually cannot have this cluster of players. Thus, a team that does not have Dominant Centers can only use Defensive Big players as substitutes. The number of this cluster is also significantly different between playoffs and non-playoffs teams. Though Defensive Big is lower than Dominant Center in some offensive variables (i.e., PTS, PER, two-point field goals, free throws made, and USG%), these players have a similar effect to Dominant Center on defensive variables (blocks and defensive rebounds). This finding confirms the conclusions of previous research that centers from winning teams secure more defensive rebounds and make more blocks in contrast to players from the same position in losing teams (Zhang et al., 2019). For foreign players, our study found that Dominant Point Guards are more distributed in non-playoffs teams than playoffs teams. In terms of personal game performances, these players contribute the most

to the teams' wins with higher PTS and USG% than other clusters (Hollinger, 2005; Sampaio et al., 2006a). However, basketball is a competitive team sport that emphasizes teamwork (Melnick, 2001) and according to Oliver's offensive skill curves (Oliver, 2004), the more possessions a team has, the less offense efficiency it has, and then a critical performance occurs. In addition, the high USG% also reflects the imbalance in the overall strength of the team. When teammates cannot score on the court, a player like a Dominant Point Guard has to take over more offensive possessions to win.

Classification and Regression Tree Analysis

Our study used the classification and regression tree model to simulate the prediction of team ability and the construction of the team roster. The results identified that the first two clusters that had the greatest impact on team ability were in native players (N7 Defensive Big and N8 Dominant Centers), and a total of six clusters in native players were selected in the tree but only one cluster in foreign players. This finding is similar to the conclusion that Ozmen got in his research (Ozmen, 2012), that efficiency of foreign players in top teams is no different than that of foreign players in regular (non-top) teams, whereas, native players in top teams are more efficient than native players in other teams. According to the previous clustering result, the main game-related characteristics of Defensive Big and Dominant Center were offensive rebound and defensive rebound, which means if a team could have more such players in their rotation, they can guarantee rebounds at any time during the game. In fact, the defensive rebound is the most important variable for the game outcome (Lorenzo et al., 2010; Gómez et al., 2017). For N1 Floor General, the average PER of this cluster of native players was the lowest. If there are too many Floor Generals in the team roster, inevitably it will pull down the efficiency of the team (Ozmen, 2012). The same explanation can also be used on N4 whose average PER was third among all native player clusters, preceded only by Dominant Center and Defensive Big. However, for the foreign player cluster with particularly high numbers of F6 Dominant Point Guards, the CART tree recommended that the playoff teams try not to recruit these players.

There are limitations in the current research that should be considered in further studies concerning players' and team's performance profiles. Firstly, due to a lack of shooting type and area variables, the new positions obtained by the clustering method cannot fully reflect the ability and style of each player. Secondly, psychological variables and situational variables are also important factors that affect the decision-making of coaches and managers. Finally, after clustering new player positions, the team's lineup just represented the roster of the team in the whole season but not all the players could play in every game. Thus, future studies can be developed based on the data of each game or 5-man lineup on the court and delve into the coach's on-the-spot substitution decision-making.

CONCLUSION

In summary, this study provides a new understanding of playing positions (Floor General, Sixth Man, Rotation Big, Shooting Guard, Three-Point Shooting Forward, Skilled Forward, Defensive Big, Dominant Center, and Bench Marginal Player in native players; Traditional center, Space Stretch Forward, Mid-Range Skilled Forward, Three-Point Shooting Guard, Traditional Point Guard, and Dominant Point Guard in foreign players) and team lineup composition in the CBA. Having a high-level of native big players is the key factor for a team entering the playoffs while the most negative impact is a Dominant Point Guard foreign player. Therefore, basketball coaches and managers will benefit from being aware of these results, particularly to set up teams and optimize preparation for individual player clusters in order to improve game performances of the players and teams.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://basketball.realgm.com/international/league/40/Chinese-CBA/stats.

AUTHOR CONTRIBUTIONS

XW, BH, and SZ contributed to conception of the study. XW organized the database. BH performed the statistical analysis. LZ wrote the first draft of the manuscript. M-ÁG and ALC supervised the design and reviewed the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

REFERENCES

Arnason, A., Sigurdsson, S. B., Gudmundsson, A., Holme, I., Engebretsen, L., and Bahr, R. (2004). Physical fitness, injuries, and team performance in soccer. *Med. Sci. Sports Exerc.* 36, 278–285. doi: 10.1249/01.MSS.0000113478.92945.CA

Blanco Villaseñor, Á, Castellano, J., Hernández Mendo, A., Sánchez López, C. R., and Usabiaga, O. (2014). Aplicación de la TG en el deporte para el estudio de la fiabilidad, validez y estimación de la muestra. *Rev. Psicol. Dep.* 23, 131–137.

Bošnjak, S. (2001). The Declaration of Helsinki – the cornerstone of research ethics. *Arch. Oncol.* 9, 179–184.

Çene, E. (2018). What is the difference between a winning and a losing team: insights from Euroleague basketball. *Int. J. Perform. Anal. Sport* 18, 55–68. doi: 10.1080/24748668.2018.1446234

Computing, S. (1991). *R Foundation for Statistical Computing*. Vienna: The R Foundation.

Fraley, C., and Raftery, A. E. (1998). *How Many Clusters? Which Clustering Method? Answers Via Model-based Cluster Analysis*. USA: University of Washington.

Gasperi, L., Conte, D., Leicht, A., and Gomez-Ruano, M. A. (2020). Game Related Statistics Discriminate National and Foreign Players According to Playing Position and Team Ability in the Women's Basketball EuroLeague. *Int. J. Environ. Res. Public Health* 17:5507. doi: 10.3390/ijerph17155507

Gómez, M. A., Ibáñez, S. J., Parejo, I., and Furley, P. (2017). The use of classification and regression tree when classifying winning and losing basketball teams. *Kinesiology* 49, 47–56. doi: 10.3390/sports5040096

Hernández-Mendo, A., Blanco-Villaseñor, Á, Pastrana, J. L., Morales-Sánchez, V., and Ramos-Pérez, F. J. (2016). SAGT: aplicación informática para análisis de generalizabilidad. *Rev. Iberoamer. Psicol. Ejercicio Deport.* 11, 77–89.

Hollinger, J. (2005). *Pro Basketball Forecast*. Dulles: Potomac Books, Inc.

Ige, C. M., and Kleiner, B. H. (1998). How to coach teams in business: the John Wooden way. *Manag. Res. News* 21, 9–12. doi: 10.1108/01409179810781310

Kubatko, J., Oliver, D., Pelton, K., and Rosenbaum, D. T. (2007). A starting point for analyzing basketball statistics. *J. Q. Anal. Sports* 3, 1–1.

Lewis, R. J. (2000). *An Introduction to Classification and Regression Tree (CART) Analysis*. California: Harbor-UCLA Medical Center.

Lorenzo, A., Gómez, M. Á, Ortega, E., Ibáñez, S. J., and Sampaio, J. (2010). Game related statistics which discriminate between winning and losing under-16 male basketball games. *J. Sports Sci. Med.* 9:664.

Lum, P. Y., Singh, G., Lehman, A., Ishkanov, T., Vejdemo-Johansson, M., Alagappan, M., et al. (2013). Extracting insights from the shape of complex data using topology. *Sci. Rep.* 3, 1–8. doi: 10.1038/srep01236

Lutz, D. (2012). "A cluster analysis of NBA players" in *Proceedings of the MIT Sloan Sports Analytics Conference*, Boston, MA.

Mateus, N., Esteves, P., Goncalves, B., Torres, I., Gomez, M. A., Arede, J., et al. (2020). Clustering performance in the European basketball according to players' characteristics and contextual variables. *Int. J. Sports Sci. Coach.* 15, 405–411. doi: 10.1177/1747954120911308

Melnick, M. J. (2001). Relationship between team assists and win-loss record in the National Basketball Association. *Percept. Mot. Skills* 92, 595–602. doi: 10.2466/pms.2001.92.2.595

Oliver, D. (2004). *Basketball on Paper: rules and Tools for Performance Analysis*. Dulles: Potomac Books, Inc.

Ozmen, M. U. (2012). *Foreign Player Quota, Experience and Efficiency of Basketball Players*. India: UIDAI. doi: 10.1515/1559-0410.1370

Özmen, M. U. (2019). Short-term impact of a foreign player quota liberalisation policy on domestic player performance: evidence from a regression discontinuity design. *Int. J. Sport Pol. Polit.* 11, 39–55. doi: 10.1080/19406940.2018.1488758

Page, G. L., Fellingham, G. W., and Reese, C. S. (2007). *Using Box-Scores to Determine a Position's Contribution to Winning Basketball Games*. United States: Brigham Young University. doi: 10.2202/1559-0410.1033

Pojskic, H., Separovic, V., Uzicanin, E., Muratovic, M., and Mackovic, S. (2015). Positional Role Differences in the Aerobic and Anaerobic Power of Elite Basketball Players. *J. Hum. Kinet.* 49, 219–227. doi: 10.1515/hukin-2015-0124

Reigal, R. E., González-Guirval, F., Pastrana-Brincones, J. L., González-Ruiz, S., Hernández-Mendo, A., and Morales-Sánchez, V. (2020). Analysis of Reliability and Generalizability of One Instrument for Assessing Visual Attention Span: menPas Mondrian Color. *Sustainability* 12:7655. doi: 10.3390/su12187655

Royuela, C. M., Torres, I. E., Pérez, C. F., and Mendo, A. H. (2017). Generalizability theory applied to olympic taekwondo combats. *Eur. J. Hum. Move.* 65–81.

Sampaio, J., Ibáñez, S., Lorenzo, A., and Gómez, M. (2006a). Discriminative game-related statistics between basketball starters and nonstarters when related to team quality and game outcome. *Percept. Mot. Skills* 103, 486–494. doi: 10.2466/pms.103.2.486-494

Sampaio, J., Janeira, M., Ibáñez, S., and Lorenzo, A. (2006b). Discriminant analysis of game-related statistics between basketball guards, forwards and centres in three professional leagues. *Eur. J. Sport Sci.* 6, 173–178. doi: 10.1080/17461390600676200

Sampaio, J., Lago, C., Casais, L., and Leite, N. (2010). Effects of starting score-line, game location, and quality of opposition in basketball quarter score. *Eur. J. Sport Sci.* 10, 391–396. doi: 10.1080/17461391003699104

Samuel Kalman, J. B. (2020). *NBA Lineup Analysis on Clustered Player Tendencies: a New Approach to the Positions of Basketball & Modeling Lineup Efficiency of Soft Lineup Aggregates*. United States: MIT Sloan Sports Analytics Conference.

Scrucca, L., Fop, M., Murphy, T. B., and Raftery, A. E. (2016). mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *R J.* 8:289. doi: 10.32614/rj-2016-021

Tavana, M., Azizi, F., Azizi, F., and Behzadian, M. (2013). A fuzzy inference system with application to player selection and team formation in multi-player sports. *Sport Manag. Rev.* 16, 97–110. doi: 10.1016/j.smr.2012.06.002

Teramoto, M., and Cross, C. L. (2017). Importance of team height to winning games in the National Basketball Association. *Int. J. Sports Sci. Coach.* 13, 559–568. doi: 10.1177/1747954117730953

Therneau, T., Atkinson, B., Ripley, B., and Ripley, M. B. (2015). *Package 'rpart'*. Available online: cran.ma.ic.ac.uk/web/packages/rpart/rpart.pdf (accessed on 20 April 2016).

Tyebkhan, G. (2003). Declaration of Helsinki: the ethical cornerstone of human clinical research. *Indian J. Dermatol. Venereol. Leprol.* 69, 245–247.

Volker, M. A. (2006). Reporting effect size estimates in school psychology research. *Psychol. Sch.* 43, 653–672. doi: 10.1002/pits.20176

WMA (2000). *Press Release: WMA Revises the Declaration of Helsinki. 9 October 2000*. Wayback Machine.

Zhang, S., Lorenzo, A., Gomez, M. A., Mateus, N., Goncalves, B., and Sampaio, J. (2018). Clustering performances in the NBA according to players' anthropometric attributes and playing experience. *J. Sports Sci.* 36, 2511–2520. doi: 10.1080/02640414.2018.1466493

Zhang, S., Lorenzo, A., Zhou, C., Cui, Y., Gonçalves, B., and Angel Gómez, M. (2019). Performance profiles and opposition interaction during game-play in elite basketball: evidences from National Basketball Association. *Int. J. Perform. Anal. Sport* 19, 28–48. doi: 10.1080/24748668.2018.1555738

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Education

ORIGINAL RESEARCH
published: 22 March 2022
doi: 10.3389/feduc.2022.827927

The Representation of Collocational Patterns and Their Differentiating Power in the Speaking Performance of Iranian IELTS Test-Takers

Masoomeh Estaji and Mohammad Reza Montazeri*

Department of English Language and Literature, Allameh Tabataba'i University, Tehran, Iran

Corpus studies have highlighted the role of multiword units in naturally occurring language. Speech theories, too, have underlined the linkage between such formulaic sequences- collocations in particular- and speech production. Few studies, however, have focused their attention on examining collocations in speaking assessment, especially in high-stakes tests. This study investigated the most frequently used collocational patterns and their discriminating power across three groups of participants with band scores 6, 7, and 8 respectively. To collect data, a corpus entailing 60 IELTS speaking samples, 20 samples from each band score, and approximately 110,000 words was gleaned. The results revealed that L1 (adjective + noun) and L7 (verb + noun) were the most frequently used types of lexical collocations, and G8 (verb + preposition) was the most frequently used grammatical collocation. The study also found that L1(adjective + verb), L5 (noun of noun), L8 (phrasal verb and adverb), L9 (noun and phrasal verb), and L10 (phrasal verb and noun) were the five types of lexical collocations with the most discriminating power across the band scores. Given the grammatical collocations, G4 (preposition + noun) and G5 (adjective + preposition) had the power to differentiate across the three band scores.

Keywords: corpus analysis, grammatical collocations, IELTS test-takers, lexical collocations, speaking assessment

OPEN ACCESS

Edited by:
George Waddell,
Royal College of Music,
United Kingdom

Reviewed by:
Yangyu Xiao,
The Chinese University of Hong Kong, China
Musa Nushi,
Shahid Beheshti University, Iran

****Correspondence:***
Masoomeh Estaji
mestaji74@gmail.com

Specialty section:
This article was submitted to Assessment, Testing and Applied Measurement, a section of the journal Frontiers in Education

Received: *02 December 2021*
Accepted: *02 February 2022*
Published: *22 March 2022*

Citation:
Estaji M and Montazeri MR (2022) The Representation of Collocational Patterns and Their Differentiating Power in the Speaking Performance of Iranian IELTS Test-Takers. Front. Educ. 7:827927. doi: 10.3389/feduc.2022.827927

INTRODUCTION

Vocabulary knowledge is an indispensable feature in organizational competence as a subcategory of language competence (Bachman and Palmer, 1996). When it comes to vocabulary learning, not only is the lexical size considered, but importance is also placed on the lexical depth, which is germane to the words relationship (Caro and Mendinueta, 2017). Over the past few years, the topic of formulaic sets has experienced a mushrooming interest, being spotlighted particularly by corpus linguists (e.g., Sinclair, 1991; Biber et al., 1999; Hyland, 2008; Paquot and Granger, 2012; Macis and Schmitt, 2016; Paquot, 2018). Studies pertinent to formulaic sequences have been in vogue for the past four decades. This highlights the importance of such pre-constructed units available to language users of which language is mostly composed (Pawley and Syder, 1983; Sinclair, 1991). Formulaic sequences, i.e., chunks of words with different lengths (Xu, 2018) entail some categories such as phraseological units (Gläser, 1986), collocations, and idioms. As an essential part of lexical cohesion in a language, collocation consists of lexical items which tend to go-together (Halliday and Hasan, 1976).

In L2 production, the speakers' correct use of collocations improves the collocational competence both in writing and speaking (Xu, 2018). This is provided that they use collocations appropriately as there exist certain types of collocations—or as Granger (1998) calls them "significant collocations"—which could differentiate between lower-level and higher-level language learners. That is, lower-level students tend to underuse native-like types of collocations caused by "an underdeveloped sense of salience and of what constitutes a significant collocation" (Granger, 1998, p. 6). Another example of the differentiating power of collocations among learners with different levels is the greater use of highly restricted collocations (i.e., collocations with limited number of substitutions for the headword such as "lunar calendar") on the part of higher-level students (Xu, 2015, 2018). Collocations with phrasal verbs are more challenging for elementary English learners (Xu, 2015). However, most of these studies have mainly scrutinized collocations including one or some types of syntactic patterns, for example, verb + noun collocations (e.g., Bahns and Eldaw, 1993; Laufer and Waldman, 2011) or premodifier-noun collocations (e.g., Durrant and Schmitt, 2009; Granger and Bestgen, 2014), failing to consider all types of collocations altogether so as to better differentiate collocations among language learners or test-takers with different levels of proficiency.

The apropos use of collocations is an emblem of native-like communicative competence (Keshavarz and Salimi, 2007) as it is seen in the native-speakers' propensity to use such chunks in lieu of other words in many situations (Wray, 2000). Thus, achieving this native-like competence and possessing the mastery of a new language would not be feasible without using formulaicity, prefabricated patterns, and collocations in particular (Wray, 2000; Nesselhauf, 2005; Nizonkiza, 2011). Moreover, formulaicity is conducive for faster language processing, leading to the production of language with minimized cognitive load (Bygate, 1987). This, consequently, leads to more improved oral proficiency (e.g., Hsu and Chiu, 2008). Not only have the teaching of collocations and raising collocational competence been accentuated in EFL and ESL classes, but also its benefits accrue to the test-takers. The utilization of collocations goes hand in hand with all of the four skills, i.e., listening, speaking, reading, and writing. Thus, collocation use augments the general level of language proficiency (e.g., Bonk, 2000; Hsu, 2007) which can be assessed through tests. Tests, particularly high-stakes ones, are largely influenced by the use of such native-like sequences.

To illustrate, the International English Language Testing System (IELTS), as one of the most universally recognized high-stakes tests, stresses the importance of lexical resources, in which collocation use is mandatory for obtaining higher scores. Among the four skills tested in IELTS, speaking is one of the most challenging one. Generally, an oft-raised issue in making hindrance for language test-takers has been the speaking portion of the proficiency tests. This is because speaking is certainly an important and perhaps the most puzzling skill among these four skills (Lazaraton, 2014). Due to the dearth of exposure to the target language, language learners have poor speaking ability "especially regarding fluency, control of idiomatic expressions, and understanding of cultural pragmatics" (Shumin, 2002, p. 204). As an efficacious way to ameliorate the test-taker's speaking proficiency, the use of collocations has proved to augment their oral proficiency scores in high-stakes tests (Hsu and Chiu, 2008). In IELTS speaking test, apropos use of collocations eventuates in band scores 7 or above according to the speaking band descriptors (public version).

Using an "evidence-based approach," a corpus-driven study offers invaluable and authentic insights regarding the data the researchers are seeking to analyze (Hyland, 2006). To this end, corpus linguists collect a corpus of representative samples of naturally occurring collocations in spoken or written texts. Previous research has mostly focused on collocation in EFL or ESL contexts or general English corpus, yet few studies have examined collocations in the testing domain, particularly when it comes to the use of grammatical and lexical collocations in important high-stakes tests such as IELTS.

LITERATURE REVIEW

Although many scholars have introduced different definitions of collocations, never have they arrived at a consensus (Mel'cuk, 1998; Wray, 2002; Nesselhauf, 2005; Xu, 2018). However, one important definition has been that of Benson et al. (2010), who referred to collocations as "fixed, identifiable, non-idiomatic" combinations which are used repetitively in a language (p. 19). Despite the multifaceted essence of collocation and formulaic language, researchers agree that collocations are divided into two chief categories, namely grammatical and lexical collocations. Grammatical collocations are collocations between one open class (noun, adjective, and verb) and one close class (a preposition or a grammatical structure) such as "depend on" (Benson et al., 2010). Lexical collocations, on the other hand, contain two content words, both of which are open classes. Lexical collocations include nouns, adjectives, verbs, and adverbs.

Regarding the two types of collocations, grammatical collocations are of eight types while there exist seven types of lexical collocations (Benson et al., 2010). Having added three more collocation categories, Xu (2015) maintained that collocations with phrasal verbs are more complex types of collocations, which call for more language competence. According to Benson et al. (2010), the syntactic patterns of grammatical collocations include: noun + preposition, noun + to + infinitive, noun + that-clause, preposition + noun, adjective + preposition, predicative adjective + to + infinitive, adjective + that-clause, and collocational verb patterns.[1] Based on Xu's (2015) syntactic patterns of lexical collocations, these types of collocations include the following structures: adjective + noun, adverb + adjective, adverb + verb, noun + noun, noun + of + noun, noun + verb, verb + noun, phrasal verb + adverb, noun + phrasal verb, and phrasal verb + noun.

Although there exist different types of grammatical and lexical collocations, previous studies proved that not all types are used

[1]According to Benson et al. (2010), there exist 19 English verb patterns in this category, whose discussion would be irrelevant and time-consuming in this study. However, among these verb patterns, verb + preposition category was chosen in the current study for the analysis of collocations used by IELTS test-takers.

with the same frequency due largely to the difficulty certain types of collocations could pose for language learners (Palmer, 1933; Bahns and Eldaw, 1993; Granger, 1998; Gitsaki, 1999; Nesselhauf, 2005). Additionally, what might impinge on language learners' collocational competence is their underuse, overuse, or misuse of certain types of collocations. This, consequently, would differentiate lower-level L2 learners with more advanced ones or native-speakers. Regarding the underuse of certain types of collocations, Laufer and Waldman (2011), say, found that L2 learners made use of less number of verb + noun collocations (5.9%) compared to native-speakers (10%). Similarly, Granger (1998) concluded that non-native speakers' use of collocations with intensifying adverbs was less than native speakers'. On the other hand, non-native speakers tend to overuse certain types of collocations in situations where more precise meaning is required (e.g., Shih, 2000; Durrant and Schmitt, 2009). Finally, the production of deviant collocations is another problem, differentiating between lower-level and higher-level L2 learners. It has been proven that the chief reason of such misuses is the negative L1 transfer based on the learners' reliance on their first language to produce L2 collocations, resulting in semantically inaccurate collocations (Kormos, 2006; Namvar, 2012; Xu, 2015).

Another proved belief about collocations is their position in the continuum of idiomaticity. The position of collocations is between free combinations such as "under the table" and fixed, figurative combinations such as "under the weather" (Howarth, 1998; Xu, 2018). Therefore, collocations are semifixed structures that contain possibly interchangeable lexical items, but with particular limitations (Xu, 2018). For instance, "make an effort" could only be substituted by "put forth an effort", and not with other elements.

Collocational competence, or the lack thereof, plays an important role in both language production and comprehension. The significance of collocational competence as one of the pivotal components in the four skills and language proficiency, in general, has been highlighted by a plethora of research (e.g., Bonk, 2000; Hsu, 2007; El-Dakhs, 2015). Collocational studies have been on the march since collocation is a widespread phenomenon in languages (El-Dakhs, 2015), helping learners improve efficiency in both comprehension and production by reducing the cognitive load throughout language production (Bygate, 1987).

As authentic language is interwoven with the utilization of collocations, in language testing, particularly spoken evaluations of the test-takers too, the use of collocations has proved to be of considerable use (e.g., Hsu and Chiu, 2008; Attar and Allami, 2013; Keshavarz and Taherian, 2018). The significance of collocations in spoken evaluations of language stems from speech-processing theories in spontaneous speaking tests. These theories substantiate the logical nexus between collocational ability and the construct of L2 oral proficiency, highlighting the role of formulaicity in speech production (Bygate, 1987; Levelt, 1999; Kormos, 2006; Xu, 2015, 2018). To illustrate, Kormos (2006) stated that formulaic expressions are crucially necessary in spoken language, and most of our utterances are replete with such expressions. What advanced L2 speakers mostly resort to is the store of the lexicon rather than L2 declarative rules because their declarative knowledge has changed to procedural knowledge (Kormos, 2006). That is why the memorization of formulaic expressions can pave the way for such a transformation as it creates limited attentional resources.

When it comes to language tests, the role and measurement of collocations has proved to be of high value (Xu, 2015). Drawing on speech production theories, Xu (2015) formulated the new construct of spoken collocational competence. Xu's (2015) study on 60 adults Chinese L2 learners' speech output in an oral English test made it clear that generally the production of accurate, complex, and fluent collocations was the part and parcel of more advanced test-takers' speech. Based on this construct and the results of his study, Xu (2015, 2018) maintained that language assessors need to pay closer attention to the measurement of collocations in language tests.

In recent years, there has been an increasing number of studies on formulaicity and collocations. Collocations have been studied on the four language skills to determine the impact of collocational knowledge on language proficiency. What is common among most of these studies is the fact that proficient L2 language learners have a better performance than lower-level learners in collocational tests (e.g., Bahns and Eldaw, 1993; Zughoul and Abdul-Fattah, 2003; Keshavarz and Salimi, 2007; Namvar, 2012). For instance, having studied 62 Taiwanese EFL students at a university of science and technology to explore the students' utilization of lexical collocations in online writing, Hsu (2007) found positive correlations between the learners' frequency of lexical collocations and their online writing scores. He also found that the variety of lexical collocations was positively correlated with their writing scores.

As to the role of collocations in speaking, Boers et al. (2006) studied the impact of formulaic sequences in general (such as idioms and collocations) on L2 oral proficiency. They studied 32 Belgian college students, 11 of whom were in the experimental group and were made cognizant of formulaic sequences, and the rest 15 were taught in the control group, with the traditional method of instruction. The findings of Boers et al. (2006) demonstrated the better performance of the experimental group in oral proficiency. Similarly, Sarvari et al. (2016) researched 60 EFL learners who were divided into the experimental and control groups. They found that the experimental group, who were taught collocations, were more fluent than the control group, who were taught solely single lexical items, in the IELTS speaking test.

However, there are obvious differences in the research results on the relationship between collocational knowledge and language proficiency. For one thing, the term "collocation" has been defined differently by different researchers, so when a researcher defines collocation in a phraseological sense (e.g., Nesselhauf, 2003), the essence of such a study would differ from another researcher who defines it in a frequency-based sense (e.g., Cowie, 1994). In addition, various researchers have based their collocational studies on a specific collocational pattern. For instance, Bahns and Eldaw (1993) focused on the lexical collocational patterns for investigating the importance of teaching collocations in EFL classes. They studied 58 advanced German learners, 34 of whom were asked to complete a translation task and the rest were asked to complete a cloze task. However, Alsulayyi (2015) attempted to study the grammatical

collocational patterns to ascertain the cause of collocational errors among Arab undergraduate students. He compared the use of grammatical collocations among Arab students majoring in English. The results indicated that noun + preposition and adjective + preposition patterns were the most erroneously used collocations, the crux of which lay in L1 transfer.

None of these studies focused on the test-takers' use of collocations and the frequency of different collocational patterns they represent to analyze their test scores in high-stakes oral proficiency tests such as IELTS. Most of the previous studies, however, researched collocations in non-testing contexts solely based on a limited number of syntactic patterns, disregarding other types of collocations particularly the grammatical ones. To bridge the mentioned gap, this study inspected the most typical grammatical and lexical collocational patterns and their differentiating power in the speaking section of IELTS. To this end, the present study sought to answer the following questions.

1. Which collocational patterns are represented in IELTS test-takers' speaking with various band scores?
2. To what extent do the collocational patterns differentiate among test-takers across various band scores in IELTS speaking?

METHODOLOGY

The Corpus

In the current study, a corpus consisting of 60 recordings of the IELTS speaking mock tests was gleaned. The speech samples were collected from different IELTS centers in Tehran, which administer IELTS mock tests regularly each year. The 60 recordings, approximately 13 h, were selected randomly from among 90 speaking samples, and then through purposive sampling, samples with band scores 6, 7, and 8 were chosen. Purposive sampling was also chosen to select the most reliable IELTS institutes in Tehran through consultation with IELTS experts. As is seen in **Table 1**, each band score involved 20 recordings to be analyzed, and the corpus contained approximately 110,000 words. The mock exams were held in 2020.

Instrumentation

In the current study, different instruments were used to collect data, including (a) IELTS mock tests and (b) IELTS speaking tests.

IELTS Mock Tests

IELTS mock tests held in 2020 were used for the quantitative phase of the study. An IELTS mock test includes four components: Listening, speaking, reading, and writing. The listening and speaking tests are the same for all of the candidates, who take the mock test. However, the reading and writing tests of the academic module are different from the reading and writing tests of the general training module. The reading test consists of three texts with a total of 40 questions. There exist different types of questions, including Multiple choice, identifying information, identifying the writer's views, matching type questions (e.g., matching headings), completion type questions (e.g., flow-chart completion), and short answer questions.

The listening test includes four sections and a total of 40 questions. The first two sections are germane to social contexts, while sections 3 and 4 are concerned with the academic contexts. Half of the sections are set in dialogues between 2 or more people, but the other two sections are set in monologues. The speaking test entails three main parts. The first part questions revolve around general and familiar topics. In the second part, the test-takers are given a cue card with a topic and some suggestions about which they are to talk for two minutes. Finally, in part 3, the examiner asks more detailed questions regarding the topic raised in part 2. The writing test involves two tasks. In task 1 of academic writing, the test-takers are asked to describe a piece of visual information, be it a graph, a table, or a chart. Task 1 of the general training module asks the test-takers to write a letter as a response to a specific situation. Task 2 of both modules requires the test-takers to respond to a problem, provide a solution, compare and contrast different ideas, and evaluate and challenge arguments. The centers administering the mock tests either use the retired or expired versions of the test, whose reliability and validity have already been confirmed, or establish the reliability and validity of the compiled ones through several pilot tests and content analysis of Iranian certified examiners.

TABLE 1 | Details of the Samples.

Speaking band scores	Number of words	Length of recording (h)
6	35,800	4.2
7	36,680	4.3
8	37,820	4.4

IELTS Speaking Tests

IELTS speaking test is an encounter between an examiner and a candidate that is designed to take between 11 to 14 min. It entails three main sections, each of which has a specific function. Part 1 which is called the introduction takes 4 to 5 min, revolving around questions concerning familiar topics such as hobbies, interests, and jobs. The second part, which is also called individual long turn, takes 3–4 min. This part entails a verbal prompt on a card about which the candidates are asked to talk for 2 min. They have one minute to prepare before talking for 2 min. Part 3 is called a two-way discussion in which the examiner asks more abstract concepts related to the topic of part 2. It takes 4–5 min. Scoring of these tasks is reported based on four criteria: Fluency and coherence, lexical resource, grammatical range and accuracy, and pronunciation, each varying from 1 to 9 IELTS bands.

Data Collection and Analysis Procedure

Before the instigation of the study, to ensure the psychometric quality of the study and richness of the data (i.e., the richness of the utilization of collocations in each sample), a speech sample containing 12 IELTS interviews between an examiner and a candidate was scrutinized. Besides, the issue of confidentiality was considered. That is, gaining the consent of the institutes

which provides the researchers with the speech samples was the primary concern of the study. Afterward, a corpus entailing 60 IELTS speaking tests recordings, which were chosen based on the test-takers' speaking section band scores, was gleaned. That is, the IELTS mock test-takers' band scores of the speaking sections were scrutinized, and those samples with the band scores 6, 7, and 8 were selected. Then around 13 h of recordings, which contained around 110,000 words, were transcribed for further analysis utilizing the application Nuance Dragon Professional Individual. Subsequently, the samples were proofread by the researchers. Finally, the examiner's speeches were italicized. A sample of the IELTS speaking test is shown in **Appendix A**.

Having adopted a corpus-driven approach, the frequency of collocations was examined after the manual extraction. That is, the frequency of 10 lexical collocations and four grammatical collocations was analyzed. The lexical and grammatical collocations were coded using Xu's (2015) (**Appendix B**) and Benson et al.'s (2010) (**Appendix C**) coding scheme, respectively. To measure the collocational strength in the English language and judge the acceptability of collocations, the use of a reference native corpus was mandatory (Paquot and Granger, 2012). Such reference corpora are quite large in size, including widely representative samples of the language used in the real and naturally occurring speech (Paquot and Granger, 2012). To do so, the present study made use of the British National Corpus (BNC) and the Corpus of Contemporary American English (COCA). Therefore, the frequency of each collocation was determined with regard to the BNC and COCA reference corpora. Moreover, having analyzed the frequency of each collocation, chi-square analysis was run to determine the differentiating power of the collocations across the three band scores.

RESULTS

Reliability Analyses of Collocation Extraction

To ensure the accuracy of the collocation extraction, the precision and recall equations were taken into account. These equations, which are used to evaluate the NLP applications (Futagi et al., 2008), help the researchers identify and include the most accurate collocations for further analysis. The following equations show the formulae for calculating precision and recall.

$$precision = \frac{\left|Collocation_{extracted} \cap Collocation_{true}\right|}{\left|Collocation_{extracted}\right|}$$

$$Recall = \frac{\left|Collocation_{extracted} \cap Collocation_{true}\right|}{\left|Collocation_{true}\right|}$$

Through analyzing the pilot study sample, the first coder identified 572 collocations. The second coder analyzed this sample and identified 580 collocations. Then the precision and recall equations were utilized to check the accuracy of collocations. The coders also discussed the differences and resolved the discrepancies by referring to the reference corpora (i.e., BNC and COCA). Having computed the precision and recall, the researchers reached the high precision of 0.98 and the recall of 0.97 which substantiate the accuracy of collocation extraction. This means that 98% of the collocations that the first coder had detected were true collocations, and solely 3% of the true collocations embedded in the speaking transcriptions were missed by the first coder.

To ensure that the coded data are reliable, the coding of the different types of collocations was done by two coders. The inter-coder consistency was calculated for 20% of the sample (five cases from each band, $N = 15$) before proceeding to the rest of the coding, using the Kappa formula (**Table 2**).

The result of the Kappa agreement test ($\kappa = 0.89$, SE = 0.017, $p = 0.000 < 0.05$) shows almost perfect agreement (values between 0.81 and 1.00 are considered almost perfect), ensuring that the data obtained from the two coders would have the acceptable consistency.

Results for the Representation of Collocational Patterns in IELTS Test-Takers' Speaking

The first question of the present study was germane to the frequency of different types of lexical and grammatical collocations across the three band scores. Based on the results, it was found that 2,178 collocations were of lexical types, while 1,074 collocations were grammatical ones. Overall, 3,252 collocations were found in the corpus. The sub-corpus of band score 6 included 570 collocations, while more collocations were utilized by the test-takers of band scores 7 and 8 (1,087 and 1,595 collocations, respectively). To analyze the frequency of each type, a frequency analysis of the types of lexical and grammatical collocations found in the analyzed corpus was done. To do so, the number of collocations used by the participants within each band score was counted. **Table 3** below presents the descriptive statistics of the lexical collocations used in each band score.

As reported in **Table 3**, the two most frequently used types of lexical collocations were L1 (adjective + noun) and L7 (verb + noun), followed by L4 (noun + noun), L2 (adverb + adjective), L3 (adverb + verb), L5 (noun of noun), and L10 (phrasal verb + noun). Moreover, L8 (phrasal verb + adverb), L9 (noun + phrasal verb), and L6 (noun + verb) had the overall lowest frequency of uses. To get a better picture of the pattern, a multiple boxplot of the frequencies in each band score was generated (**Figure 1**).

As illustrated in **Figure 1**, in all types of lexical collocations, the frequency of use increased with the increase in the band score. Moreover, the pattern seems still throughout the band sores, with L1 (adjective + noun) and L7 (verb + noun) being the most frequent ones. The following excerpts show the use of these two types of collocations.

Excerpt 1: They should have eh different eh characteristics for start their own business, they should be eh very knowledgeable, and they should have **interpersonal skills** and also manage. **(L1 (adjective + noun)/ band score 7 test-taker)**

Excerpt 2: And eh when eh weather is rainy eh and eh especially in accidents or eh rainy day eh it leads eh to **increase**

TABLE 2 | Kappa agreement: inter-coder consistency.

		Value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance
Measure of Agreement	Kappa	0.89	0.01	54.52	0.00
No. of valid cases		355			

[a]*Not assuming the null hypothesis.*
[b]*Using the asymptotic standard error assuming the null hypothesis.*

eh eh **traffic congestion**. **(L7 (verb + noun)/ band score 6 test-taker)**

After L1 (adjective + noun) and L7 (verb + noun), L4 (noun + noun), L2 (adverb + adjective), L3 (adverb + verb), L5 (noun of noun), and L10 (phrasal verb + noun) were the most frequently used types, respectively. The following excerpts show the use of these types.

TABLE 3 | Descriptive statistics for the frequency of lexical collocations by band scores.

	Band	N	Minimum	Maximum	Mean	SD
L1	6	20	0	12	4.50	3.08
	7	20	6	28	13.75	6.65
	8	20	10	37	21.20	6.37
	Total	60	0	37	13.15	8.82
L2	6	20	0	5	1.20	1.64
	7	20	0	12	2.25	3.52
	8	20	0	10	3.30	2.25
	Total	60	0	12	2.25	2.69
L3	6	20	0	2	0.40	0.68
	7	20	0	5	1.65	1.75
	8	20	0	7	2.10	2.15
	Total	60	0	7	1.38	1.77
L4	6	20	0	7	2.10	1.65
	7	20	0	10	4.95	2.91
	8	20	2	19	7.35	4.20
	Total	60	0	19	4.80	3.74
L5	6	20	0	1	0.25	0.44
	7	20	0	3	1.05	0.82
	8	20	0	6	2.15	1.42
	Total	60	0	6	1.15	1.24
L6	6	20	0	1	0.15	0.36
	7	20	0	2	0.50	0.76
	8	20	0	2	0.55	0.68
	Total	60	0	2	0.40	0.64
L7	6	20	3	12	7.30	3.06
	7	20	4	24	12.60	6.21
	8	20	9	29	16.00	5.28
	Total	60	3	29	11.97	6.12
L8	6	20	0	0	0.00	0.00
	7	20	0	0	0.00	0.00
	8	20	0	2	0.30	0.57
	Total	60	0	2	0.10	0.35
L9	6	20	0	0	0.00	0.00
	7	20	0	1	0.15	0.36
	8	20	0	1	0.40	0.50
	Total	60	0	1	0.18	0.39
L10	6	20	0	2	0.15	0.48
	7	20	0	4	0.85	1.26
	8	20	0	6	2.15	1.84
	Total	60	0	6	1.05	1.54

Excerpt 3: Uh but some of them are really eh motivating, for example, some uh **talent shows** are really motivating for eh young children and eh **(L4 (noun + noun)/band score 6 test-taker)**

Excerpt 4: Well, I had this eh **excruciatingly painful** burden of my parents on me because I was as good at engineering as I was good at mathematics. **(L2 (adverb + adjective)/band score 8 test-taker)**

Excerpt 5: First of all, I have to say that computers have **drastically changed** our lives. **(L3 (adverb + verb)/band score 8 test-taker)**

Excerpt 6: Of course. Uh, I have a lot of **circle of friends** and they I mean they have- I definitely put the- put them on X mm due to their behavior but also my life. **(L5 (noun of noun)/band score 7 test-taker)**

Excerpt 7: You shouldn't uh **bottle up your feelings** and which eh gradually eventuate in some disaster. **(L10 (phrasal verb + noun) band score 7 test-taker)**

Finally, L8 (phrasal verb + adverb), L9 (noun + phrasal verb), and L6 (noun + verb) were the least frequents. The following excerpts show the use of these types of lexical collocations produced by the test-takers.

Excerpt 8: I try to just eh keep eh normal eh sleeping pattern-sleeping pattern and uh not **staying up** much **late**. **(L8 (phrasal verb + adverb)/band score 8 test-taker)**

Excerpt 9: I play the guitar, and I think my whole **life revolves around** the music. **(L9 (noun + phrasal verb)/band score 8 test-taker)**

Excerpt 10: Well, probably those kinda uh cool **titles** that **hook** you. They really can captivate your attention. **(L6 (noun + verb)/band score 8 test-taker)**

Concerning grammatical collocations, only four types were inspected. The rationale for the selection of these four types has been discussed in the delimitations of the study. **Table 4** presents the descriptive statistics for the frequency of use in this type of collocation.

Table 4 shows that the most frequently used type of grammatical collocations was G8 (verb + preposition) and the least G1 (noun + preposition). **Figure 2** illustrates the pattern in a multiple boxplot.

Figure 2 also shows similar patterns throughout the band scores considering the frequency of use, G8 (verb + preposition) having the highest, followed by G4 (preposition + noun) and G5 (adjective + preposition). Moreover, G1 (noun + preposition) had the lowest frequency in all the three band scores. The pattern also exists in the other two types of grammatical collocations. The following examples illustrate the utilization of grammatical collocations by the test-takers.

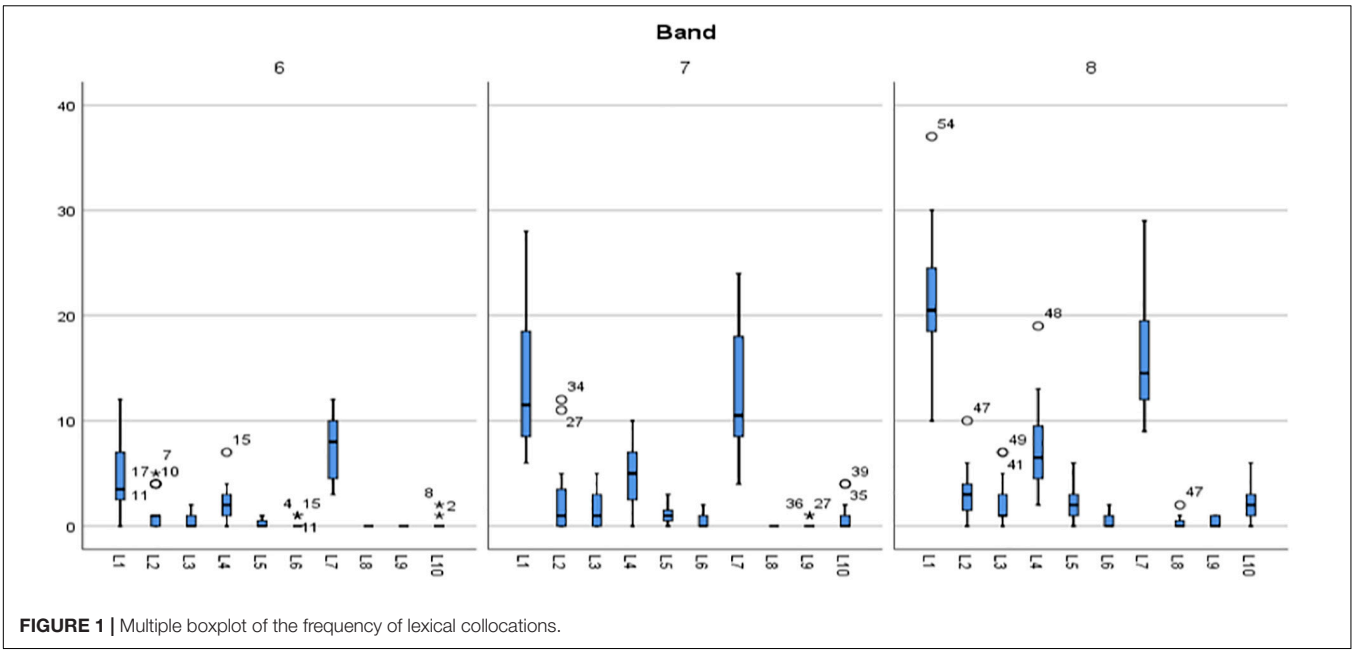


FIGURE 1 | Multiple boxplot of the frequency of lexical collocations.

Excerpt 11: Yes, em . . . to my mind, Iranians are very welcoming and they are hospitable, and I think eh as- our country has a **reputation for** its natural attraction. **(G1 (noun and preposition) band score 7 test-taker)**

Excerpt 12: I guess, um reliable information cannot necessarily be found **on the internet**. **(G4 (preposition + noun) band score 8 test-taker)**

Excerpt 13: Eh mmm as I know, in Iran, eh people are eh... You are so **friendly to** foreigner eh they invite them maybe to their home. **(G5 (adjective + preposition) band score 6 test-taker)**

Excerpt 14: Mmm I think eh, it **depends on** the situation but eh mostly the things that you need eh very strong concentration, for example, doing some... kind of eh drawing. **(G8 (verb + preposition) band score 6 test-taker)**

Results for the Collocational Patterns Differentiating Among IELTS Test-Takers Across Various Band Scores

To capture the differentiating power of the collocations in various band scores, series of chi-square tests were run on each type of collocations. **Table 5** shows the results for lexical collocations and **Table 6** for the grammatical ones.

Based on the results of the chi-square tests, it was found that there did exist certain types which can differentiate test-takers across band scores 6, 7, and 8. As it is evident from **Table 5**, half of the lexical collocations' types can differentiate among the three band scores. They include L1 (adjective + verb) ($\chi^2_{(52)} = 72.8$, $p = 0.03 < 0.05$), L5 (noun of noun) ($\chi^2_{(10)} = 33.4$, $p = 0.000 < 0.05$), L8 (phrasal verb and adverb) ($\chi^2_{(4)} = 10.91$, $p = 0.028 < 0.05$), L9 (noun and phrasal verb) ($\chi^2_{(2)} = 10.91$, $p = 0.004 < 0.05$), and L10 (phrasal verb and noun) ($\chi^2_{(12)} = 32.76$, $p = 0.001 < 0.05$). The following excerpts are indicative of the use of these five collocations among the three band scores. With regard to the use of L1 (adjective + noun), the following excerpts are presented.

Excerpt 15: I actually rarely eh use **public transportations** like bus, subway and uh. **(band score 6 test-taker)**

Excerpt 16: The majority of the time uh we are in faced with eh scarcity and- and water shortage, but I hope in the not- in the **foreseeable future** eh take action to tackle the problem. **(band score 7 test-taker)**

Excerpt 17: Uh, not to mention, I would say it is also famous for uh- for the **heavy traffic** that we face every day here, which makes it actually eh to some extent difficult to < get > around. **(band score 8 test-taker)**

TABLE 4 | Descriptive statistics for the frequency of grammatical collocations by band scores.

	Band	*N*	Minimum	Maximum	Mean	SD
G1	6	20	0	4	.85	1.26
	7	20	0	6	1.45	1.60
	8	20	0	5	1.50	1.53
	Total	60	0	6	1.27	1.48
G4	6	20	1	8	3.55	4.89
	7	20	0	11	4.70	8.32
	8	20	3	12	7.45	4.99
	Total	60	0	12	5.23	2.93
G5	6	20	0	5	1.80	2.37
	7	20	1	7	2.65	3.08
	8	20	1	10	5.00	5.36
	Total	60	0	10	3.15	2.31
G8	6	20	2	14	6.35	9.60
	7	20	2	15	7.65	8.55
	8	20	3	16	10.25	11.67
	Total	60	2	16	8.08	3.50

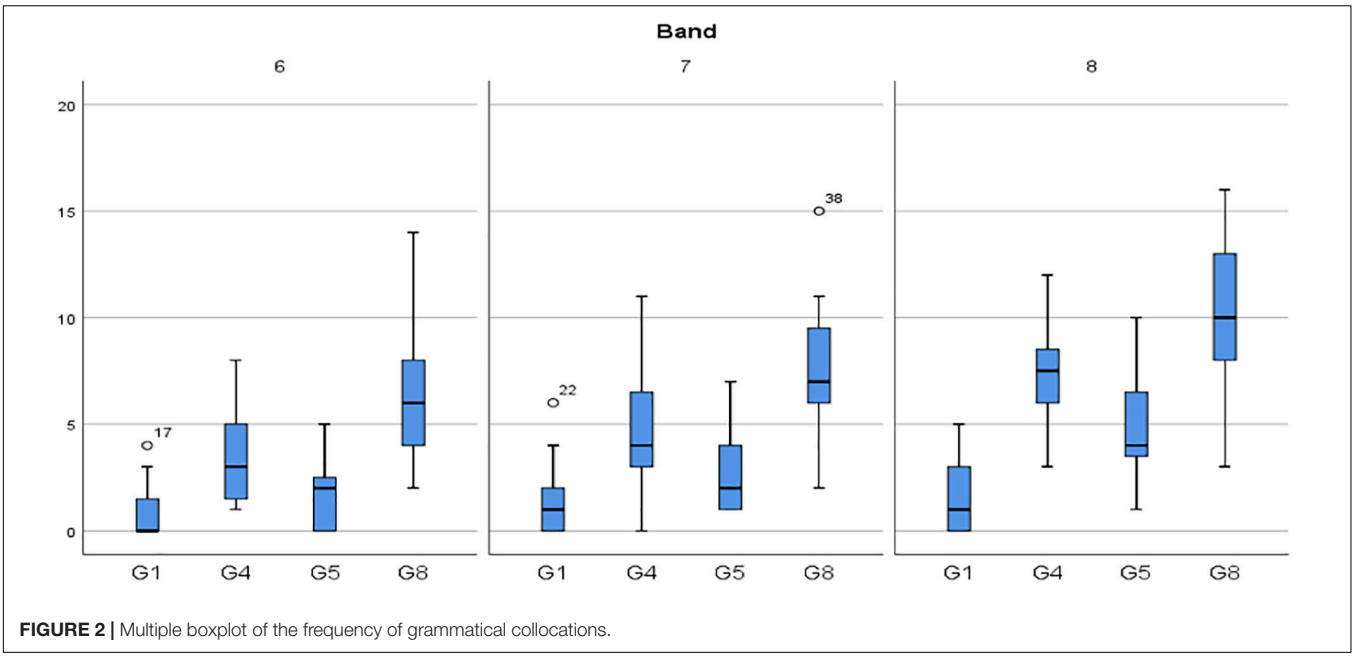


FIGURE 2 | Multiple boxplot of the frequency of grammatical collocations.

The following examples show the use of L5 (noun of noun) across the three groups.

Excerpt 18: It's very eh . . . useful because it makes eh . . . it- it makes eh income for me. And eh . . . it makes income for me and eh it eh eh eh sour- I think it's- it's good eh **source of revenue** for our lives. **(band score 6 test-taker)**

Excerpt 19: Maybe they don't have enough information, or even if they have prepared enough em **piece of information**, it might be nervous to stand in front of so many eyes and they are just staring at you. **(band score 7 test-taker)**

Excerpt 20: And when I exercise, I can feel it in every inch of myself and in every **ounce of blood** that I can think better. **(band score 8 test-taker)**

The use of L8 (phrasal verb and adverb) was solely detected in band score 8, as the excerpt 21 illustrates:

Excerpt 21: Uh, but not everybody. I think uh possesses the eh like ability and capability to **get along well** with English.

The use of L9 (noun and phrasal verb) was only found in band scores 7 and 8. The following examples show the use of this type.

Excerpt 22: He helped us eh about subjects because he is eh very eh resourceful, and whenever we co- come up eh eh to- **a problem comes up**, he helps us a lot. **(band score 7 test-taker)**

Excerpt 23: Their eh potential benefits and eh effect eh might not be eh clear or touchable now for the people but as **time goes by** actually through investing, these can actually uh mm later surprise people. **(band score 8 test-taker)**

Finally, excerpts 24, 25, and 26 show the use of L10 (phrasal verb and noun) across the three band scores.

Excerpt 24: When I was a little girl eh loved eh special **shoes** that eh as a little girl I- I hadn't the uh- the permission to- to **put** them(on). **(band score 6 test-taker)**

Excerpt 25: To this- so this uh you have to be able to **deal with** different kinds of **problems** or **issues** that you will face in the future of your business. **(band score 7 test-taker)**

Excerpt 26: When you cannot make ends meet, you have to eh **look for** a secondary **job** eh that eh- that partially helps you eh- your life to run sm- more smoothly. **(band score 8 test-taker)**

Considering the grammatical collocations, as it is evident from **Table 6**, two out of four types of the collocations could differentiate among the three band scores. They include G4 (preposition + noun) ($\chi^2_{(24)} = 38.49$, $p = 0.031 < 0.05$) and G5 (adjective + preposition) ($\chi^2_{(20)} = 46.6$, $p = 0.001 < 0.05$). The

TABLE 5 | Chi-square test on the type of lexical collocations differentiating the band scores.

	χ^2	df	Asymptotic significance (2-sided)
L1	72.80	52	0.03
L2	26.25	18	0.09
L3	17.48	12	0.13
L4	34.56	24	0.07
L5	33.40	10	0.00
L6	6.02	4	0.19
L7	45.00	40	0.27
L8	10.90	4	0.02
L9	10.90	2	0.00
L10	32.75	12	0.00

TABLE 6 | Chi-square test on the type of grammatical collocations differentiating the band scores.

	χ^2	Df	Asymptotic significance (2-sided)
G1	8.53	12	0.74
G4	38.47	24	0.03
G5	46.60	20	0.00
G8	34.19	28	0.19

following excerpts illustrate the use of G4 (preposition + noun) across the three levels:

Excerpt 27: Uh, this is a difficult question because I usually can't remember my dreams **at night,** but eh eh sometimes eh I usually eh have some dreams about immigration to Australia. **(band score 6 test-taker)**

Excerpt 28: Uh the most popular one I think is the uh subway because it super X uh the other transport systems uh in our town uh and also it's eh so fast and **on time**. **(band score 7 test-taker)**

Excerpt 29: But parents have to educate themselves. Is is- it is not something that you do that just **by nature**. **(band score 8 test-taker)**

Overall, the results indicated that there existed such differences among the test-takers concerning their use of certain types of collocations. Hence, the null hypothesis, which claimed that the collocational patterns do not differentiate among test takers across various band scores in IELTS speaking, was rejected.

DISCUSSION

This study aspired to analyze the representation of collocational patterns and their differentiating power across three band scores in the IELTS speaking test. The first question of the present study was pertinent to the frequency of lexical and grammatical collocations. According to the research results, the use of both types of collocations with all of the syntactic types was observed in each band score. In particular, the production of collocations—both the lexical and grammatical types—increased with the increase of the test-takers' speaking band scores. This highlights that the more advanced the students became, the more collocations they used in their speech.

This finding shows that lower-level test-takers still lack adequate collocational competence to convey their messages through varied types of collocations. This finding is in tune with the literature in that the number of collocation use tends to increase with the increase of the learners' proficiency level (Bahns and Eldaw, 1993; Laufer and Waldman, 2011). This finding must be interpreted with care, though, since the literature has shown that even advanced students tend to make (sometimes more) mistakes. By way of example, some studies (e.g., Obukadeta, 2014; Men, 2018) suggested that "collocation lag" (Men, 2018, p. 2) does exist as the English language learners' proficiency level increases and they tend to encounter more collocational challenges.

The findings indicated that out of the ten patterns of the lexical collocations, the most common patterns were L1 (adjective + noun), L7 (verb + noun), and L4 (noun + noun) among all the three groups. This finding was quite predictable since the literature has substantiated that the majority of collocations produced by English learners include L7 (verb + noun), L1 (adjective + noun), and L4 (noun + noun) (e.g., Xu and Xi, 2010; Xu, 2015). These findings are because such collocations carry the most important information while producing language (Xu, 2015; Men, 2018). In line with the literature (e.g., Mei, 1999; Xu, 2015), the greater frequency of these three syntactic patterns show the role of syntactic transfer in which Iranian IELTS test-takers applied "L1 rules for encoding an L2 phrase" (Kormos, 2006, p. 175). That is, by using such structures, they could lighten their cognitive load, thereby facilitating their L2 oral language production (Kormos, 2006). Although these three types are the commonest types of collocations, it does not mean that error-free production of these combinations is easy. However, research proved that these types carry the most frequent sources of challenge for English L2 learners. To illustrate, regarding L7 (verb + noun), Gitsaki (1999) concluded that verb+noun combinations were the most difficult types of collocations which were acquired at later stages of learning. Moreover, based on the Chinese Learner English Corpus, L1 (adjective + noun) and L4 (noun + noun) were the second and third most deviant collocation types (Gui and Yang, 2003).

The present study also found that L1 (adjective + noun) outnumbered L7 (verb + noun) in general (save for band score 6), although the difference was not significant. This is in contrast with some other studies such as Namvar's (2012) and Xu's (2015). This could be justified by what Johansson and Hofland (1989) stated. They held that L1 (adjective + noun) and L4 (noun + noun) collocations are the most frequently used types by native English speakers. Gitsaki's (1999) study also demonstrated that L1 (adjective + noun) was the easiest combination, being acquired at an early stage of collocational knowledge development.

The paucity of the use of other types of collocations such as L2 (adverb + adjective) and L3 (adverb + verb), especially on the part of band score 6 test takers illustrates their lack of knowledge concerning these collocations. The use of such syntactic patterns increased with the increase of band scores; however, the test-takers still tended not to use these patterns much. This highlights that adverbial collocations are not as common as L1 (adjective + noun), L7 (verb + noun), and L4 (noun + noun), which carry the most important information in oral communication (Xu, 2015). This finding aligns with the literature in that non-native speakers tend to underuse collocations with intensifying adverbs (Granger, 1998). As Xu (2015) maintained, English learners tend to replace such intensifying collocations (such as adverb + adjective) with single words. Collocations with phrasal verbs were also proved to be less utilized in all the three band scores. This is because phrasal verb collocations are more difficult to learn than collocations with single verbs due to their idiomatic nature and syntactic patterns (Xu, 2015; Maeen and Chilukuri, 2019). That is why, in lieu of such advanced syntactic patterns, English learners tend to cling on, as Granger (1998) put it, their "safe bets" or certain types of fixed expressions about which they are more confident (p. 147).

Of the four types of 1,074 grammatical collocations, the most frequent pattern was G8 (verb + preposition), followed by G4 (preposition + noun), G5 (adjective + preposition), and G1 (noun + preposition). This could mean that G8 (verb + preposition) was the easiest for the learners to learn, and they probably had greater exposure to verb + preposition combinations since verb patterns are quite common in English (Benson et al., 2010). On the other hand, the lack of knowledge in the use of G1 (noun + preposition) was pronounced among test-takers, especially among band score 6 test-takers. Lack of exposure to certain types of collocations could result in the test-takers' difficulty of using such structures (Bortfeld and Brennan, 1997; Hsu and Hsu, 2007). Another problem for using

adjective/noun+preposition patterns would be due to the effect of negative transfer as in Persian, the most frequent preposition is "az" (from) (Maeen and Chilukuri, 2019) which makes hindrance for Persian L2 learners to use a wide range of prepositions in using grammatical collocations in English. To illustrate, in English, the adjective "surprised" is followed by "at", the adjective "afraid" by "of", and the adjective "bored" by "with", whereas in Persian all of them are followed by "az" (from). This difficulty for the use of G1 (noun + preposition) was also confirmed by Hatami (2015), who found that this type of collocation was more difficult than the production of G4 (preposition + noun).

The second question of the study examined whether or not the collocational patterns differentiated among IELTS test-takers across various band scores. To answer this question, the study made use of chi-square tests on each type of collocations. It was found that five lexical collocations, i.e., L1 (adjective + verb), L5 (noun of noun), L8 (phrasal verb and adverb), L9 (noun and phrasal verb), and L10 (phrasal verb and noun), were able to differentiate test-takers based on their band scores. The existence of certain types of collocations, which could distinguish higher-level L2 learners from lower-level ones has been supported in literature (e.g., Durrant and Schmitt, 2009; Granger and Bestgen, 2014). In their study, Granger and Bestgen (2014) found that intermediate L2 learners tended to underuse lower-frequency collocations, using a large proportion of high-frequency combinations. They also noticed that adjective+noun types had the differentiating power between intermediate and advanced learners.

The results of the current study were also quite expected since lower-level test-takers tended to avoid complex types of collocations, such as those with phrasal verbs (i.e., L8, L9, and L10), due to their lack of knowledge of these difficult syntactic patterns (Xu, 2015; Maeen and Chilukuri, 2019). This conservative strategy, as is seen in the literature (e.g., Xu and Xi, 2010; Xu, 2015), can eventuate in lower-level test-takers' dearth of utilizing such syntactic patterns and their considerable reliance on easier patterns. Therefore, the apropos use of such complex collocations was proved to be of help to ameliorate the test-takers' speaking band-scores. This differentiating power can help to determine the specific group that the test-takers may belong to as to their use of these collocations. To illustrate, band-score 6 and 7 test-takers did not make use of L8 (phrasal verb and adverb), highlighting the fact that not only is learning phrasal verbs difficult but also more difficulty is created when phrasal verbs collocate with adverbs. This is because collocations with adverbs are not frequent among non-native speakers (Granger, 1998; Xu, 2015).

The two types of grammatical collocations with differentiation power were G4 (preposition + noun) and G5 (adjective + preposition). This finding indicates that learning the right prepositions, which collocate with nouns and adjectives (i.e., before a noun and after an adjective), could considerably change the test-takers' level of proficiency; if applied appropriately. This finding highlights the significance of prepositions in the English language as learning prepositions has proved to be the most challenging part of English learning (Takahaski, 1969). More importantly, the English language contains more diverse prepositions than the Persian language (Ghasemi et al., 2014), meaning that learning them could be an issue particularly for Iranian L2 learners due to the negative transfer of L1. Many adjectives, for instance, are followed by prepositions which cannot be produced by the literal translation of them, such as the most famous preposition in Persian which is "az" (Maeen and Chilukuri, 2019) as used in "عصبانی از" or "کسل از", which are translated into "angry at" and "bored with". In another study, Hatami (2015) found that learning G4 (preposition + noun) was easier for Iranian learners than G1 (noun + preposition). This is in contrast with the current study's results in which G4 (preposition + noun) was proved to have the differentiation power across the test-takers. Therefore, it revealed that with the improvement of their language proficiency and band score, test-takers made more effective use of G4 (preposition + noun) and G5 (adjective + preposition) in their speaking tests.

CONCLUSION AND IMPLICATIONS

In this study, the frequency of collocational use was examined in the IELTS speaking test. It was also an attempt to analyze the differentiating power of collocations across three band scores of 6, 7, and 8. Findings of the current study suggested that collocation use was a quite common phenomenon in the three corpora (i.e., band score 6,7, 8 speaking samples). In particular, the results of the study evinced that both the lexical and grammatical collocations were quite frequent across the three groups. It was also revealed that with the increase in the band scores, the number of collocations uses increased as well. That is, while 570 collocations were used in band score 6 sub-corpus, 1,087 and 1,595 were found in band score 7 and 8 sub-corpora, respectively.

According to the results, L1 (adjective + noun) and L7 (verb + noun) were the most frequently used types of lexical collocations. After these types, L4 (noun+ noun), L2 (adverb + adjective), L3 (adverb + verb), L5 (noun of noun), and L10 (phrasal verb+ noun) were detected to be the most common types. However, L8 (phrasal verb + adverb), L9 (noun + phrasal verb), and L6 (noun + verb) had the lowest frequency of uses across the band scores. In addition, G8 (verb + preposition) was the most frequently used grammatical collocation, followed by the pattern G4 (preposition + noun) and G5 (adjective + preposition), while the least used syntactic pattern was G1 (noun + preposition).

The study also sought to determine the collocational patterns, which differentiated among IELTS test-takers across band scores 6, 7, and 8. Having made use of chi-square tests, the study found that among the lexical collocation types, L1 (adjective + verb), L5 (noun of noun), L8 (phrasal verb and adverb), L9 (noun and phrasal verb), and L10 (phrasal verb and noun) had the most differentiating power. Moreover, G4 (preposition + noun) and G5 (adjective + preposition) possessed the most differentiating power among the grammatical collocation types.

According to the results of the current study, it can be concluded that the more proficient the test-takers became, the more collocations they produced in the speaking test. This means that the apt use of collocations would result in more advancement in language proficiency and communicative

competence (Crossley et al., 2014; Xu, 2015). Moreover, as the differentiating power of the collocations indicated, the more advanced test-takers tended to use more complex syntactic patterns both in terms of lexical and grammatical collocations. As an illustration, more advanced test-takers tended to use more complex patterns of collocational phrasal verbs.

The findings of this study have significant implications for IELTS trainers, IELTS materials developers, and syllabus designers. IELTS trainers should have an encyclopedic knowledge of collocations and be keenly aware of the importance of formulaicity. They, thus, can highlight the merits of formulaic language so that the test-takers would pay heed to them to ameliorate their collocational competence. Moreover, materials developers are required to pay special attention to embedding useful collocations that can contribute to the improvement of test-takers' oral proficiency. They can also use the findings of the study as to the differentiating power of collocations across the three band scores. Stated in another way, they can raise the learners' awareness of the more challenging types of collocations, which would result in the improvement of their band score.

Using lexis-based syllabus as a supplementary syllabus to the core syllabus, syllabus designers can highlight the importance of collocations in IELTS courses. They can also explicitly draw the students' attention to the differences of collocations in L1 and the English language to pre-empt the potential negative-transfer-related collocational errors. Therefore, the importance of formulaic language should be underlined in designing a course, particularly those courses preparing the students for high-stakes tests.

One of the limitations of the study included not having access to a sufficient number of speaking band score 9 since few test-takers may achieve the highest IELTS band score. Thus, the collocations were collected from band scores 6, 7, and 8. In addition, the study's delimitation relates to the choice of those grammatical types of collocations with prepositions for further analysis. According to Takahaski (1969), the most serious problem an L2 learner encounters while learning English is certainly the correct use of specific prepositions. Thus, these specific types would be the most representative of the grammatical competency of an L2 English learner.

As this study did not enjoy an enormous wealth of band scores 9, future researchers can entertain the idea of analyzing the collocations produced by this group of test-takers. Moreover, when it comes to the analysis of different types of collocations, future researchers may wish to merely work on either the lexical or grammatical ones. They can also analyze the other types of grammatical collocations, which were not considered in the present study. IELTS researchers are also recommended to conduct qualitative studies to descry the reasons behind producing miscollocations such as apprehension or anxiety-related problems. Finally, future researchers can carry out greater numbers of corpus-based studies with larger samples to throw more light on the significance of collocations in the speaking parts of the high-stakes tests.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

Both authors have materially participated in the research and manuscript preparation and approved the final manuscript.

REFERENCES

Alsulayyi, M. N. (2015). The use of grammatical collocations by advanced Saudi EFL learners in the UK and KSA. *Int. J. Engl. Linguist.* 5, 32–43. doi: 10.5539/ijel.v5n1p32

Attar, E. M., and Allami, H. (2013). The effects of teaching lexical collocations on speaking ability of Iranian EFL learners. *Theory Pract. Lang. Stud.* 3, 1070–1079. doi: 10.4304/tpls.3.6.1070-1079

Bachman, L. F., and Palmer, A. S. (1996). *Language Testing in Practice: Designing and Developing Useful Language Tests*. Oxford: Oxford University Press.

Bahns, J., and Eldaw, M. (1993). Should we teach EFL students collocations? *System* 2, 101–114. doi: 10.1016/0346-251X(93)90010-E

Benson, M., Benson, E., and Ilson, R. (2010). *The BBI Combinatory Dictionary of English*, 3rd Edn. Amsterdam: John Benjamins. doi: 10.1075/z.bbi

Biber, D., Johansson, J., Leech, G., Conrad, S., and Finegan, E. (1999). *Longman Grammar of Spoken and Written English*. Harlow: Pearson Ltd.

Boers, F., Eyckmans, J., Kappel, J., Stengers, H., and Demecheleer, M. (2006). Formulaic sequences and perceived oral proficiency: putting a lexical approach to the test. *Lang. Teach. Res.* 10, 245–261. doi: 10.1191/1362168806lr195oa

Bonk, W. J. (2000). *Testing ESL Learners' Knowledge of Collocations (ERIC Document Reproduction Service No. ED 442 309)*. Available online at: http://files.eric.ed.gov/fulltext/ED442309.pdf (accessed April 16, 2020).

Bortfeld, H., and Brennan, S. E. (1997). Use and acquisition of idiomatic expressions in referring by native and non-native speakers. *Discourse Process.* 23, 119–148. doi: 10.1080/01638537709544986

Bygate, M. (1987). *Speaking*. Oxford: Oxford University Press.

Caro, K., and Mendinueta, N. R. (2017). Lexis, lexical competence and lexical knowledge: a review. *J. Lang. Teach. Res.* 8, 205–213. doi: 10.17507/jltr.0802.01

Cowie, A. P. (1994). "Phraseology," in *The Encyclopedia of Language and Linguistics*, ed. R. E. Asher (Oxford: Pergamon), 3168–3171.

Crossley, S. A., Salsbury, T., and McNamara, D. S. (2014). Assessing lexical proficiency using analytic ratings: a case for collocation accuracy. *Appl. Linguist.* 36, 1–22. doi: 10.1093/applin/amt056

Durrant, P., and Schmitt, N. (2009). To what extent do native and non-native writers make use of collocations? *IRAL Int. Rev. Appl. Linguist. Lang. Teach.* 47, 157–177. doi: 10.1515/iral.2009.007

El-Dakhs, D. (2015). Collocational competence in English language teaching: an overview. *Arab World Engl. J.* 6, 68–82. doi: 10.24093/awej/vol6no1.5

Futagi, Y., Deane, P., Chodorow, M., and Tetreault, J. (2008). A computational approach to detecting collocation errors in the writing of non-native speakers of English. *Comput. Assist. Lang. Learn.* 21, 353–367. doi: 10.1080/09588220802343561

Ghasemi, F., Janfaza, A., and Soori, A. (2014). A contrastive analysis of the prepositions "of" and "from". *Int. J. Educ. Lit. Stud.* 2, 17–21. doi: 10.7575/aiac.ijels.v.2n.3p.17

Gitsaki, C. (1999). *Second Language Lexical Acquisition: A Study of the Development of Collocational Knowledge*. San Francisco, CA: International Scholars Publications.

Gläser, R. (1986). *Phraseologie der Englischen Sparche.* [Phraseology of English]. Leipzig: VEB Verlag Enzyklopädie. doi: 10.1515/9783111562827

Granger, S. (1998). "Prefabricated patterns in advanced EFL writing: collocations and formulae," in *Phraseology: Theory, Analysis, and Applications*, ed. A. P. Cowie (Oxford: Oxford University Press), 145–160.

Granger, S., and Bestgen, Y. (2014). The use of collocations by intermediate vs. advanced non-native writers: a bigram-based study. *Int. Rev. Appl. Linguist. Lang. Teach.* 52, 229–252. doi: 10.1515/iral-2014-0011

Gui, S. C., and Yang, Z. H. (2003). *Chinese Learner English Corpus*. Shanghai: Shanghai Foreign Language Education Press.

Halliday, M. A. K., and Hasan, R. (1976). *Cohesion in English*. London: Longman.

Hatami, S. (2015). *Collocations in Farsi L2 Learners of English. The Role of Proficiency and L1 Language Transfer* (Unpublished Master's thesis). Tromsø: The Arctic University of Norway.

Howarth, P. A. (1998). Phraseology and second language proficiency. *Appl. Linguist.* 19, 24–44. doi: 10.1093/applin/19.1.24

Hsu, J. Y., and Hsu, L. C. (2007). Teaching lexical collocations to enhance listening comprehension of English majors in a technological university of Taiwan. *Soochow J. Foreign Lang. Cult.* 24, 1–33.

Hsu, J.-y. (2007). Lexical collocations and their relation to the online writing of Taiwanese college English majors and non-English majors. *Electron. J. Foreign Lang. Teach.* 4, 192–209.

Hsu, J.-y., and Chiu, C. y. (2008). Lexical collocations and their relation to speaking proficiency of college EFL learners in Taiwan. *Asian EFL J.* 10, 181–204.

Hyland, K. (2006). *English for Academic Purposes an Advanced Resource Book*. London: Routledge. doi: 10.4324/9780203006603

Hyland, K. (2008). As can be seen: lexical bundles and disciplinary variation. *Engl. Specif. Purp.* 27, 4–21. doi: 10.1016/j.esp.2007.06.001

Johansson, S., and Hofland, K. (1989). *Frequency Analysis of English Vocabulary and Grammar*. Oxford: Oxford University Press.

Keshavarz, M. H., and Salimi, H. (2007). Collocational competence and cloze test performance: a study of Iranian EFL learners. *Int. J. Appl. Linguist.* 17, 81–92. doi: 10.1111/j.1473-4192.2007.00134.x

Keshavarz, M. H., and Taherian, P. (2018). The Effect of explicit instruction of collocations on EFL learners' language proficiency. *Hacettepe Univ. J. Educ.* 33, 987–1001. doi: 10.16986/HUJE.2018038632

Kormos, J. (2006). *Speech Production and Second Language Acquisition*. Mahwah, NJ: Lawrence Erlbaum Associates.

Laufer, B., and Waldman, T. (2011). Verb-noun collocations in second language writing: a corpus analysis of learners' English. *Lang. Learn.* 61, 647–672. doi: 10.1111/j.1467-9922.2010.00621.x

Lazaraton, A. (2014). "Second language speaking," in *Teaching English as a Second or Foreign Language*, eds M. Celce-Murcia, D. M. Brinton, and M. A. Snow (Boston, MA: National Geographic Learning/Heinle Cengage Learning), 106–120.

Levelt, W. J. M. (1999). "Producing spoken language: a blueprint of the speaker," in *The Neurocognition of Language*, eds C. M. Brown and P. Hagoort (Oxford: Oxford University Press), 83–122. doi: 10.1093/acprof:oso/9780198507932.003.0004

Macis, M., and Schmitt, N. (2016). The figurative and polysemous nature of collocations and their place in ELT. *ELT J.* 71, 50–59. doi: 10.1093/elt/ccw044

Maeen, N., and Chilukuri, B. A. (2019). Comparative study of phrasal verbs in English and Persian. *Spec. J. Lang. Stud. Lit.* 3, 16–31.

Mei, J. J. (1999). *Dictionary of Modern Chinese Collocations.* Shanghai: Hanyu Dictionary Press.

Mel'cuk, I. (1998). "Collocations and lexical functions," in *Phraseology: Theory, Analysis, and Application*, ed. A. P. Cowie (New York, NY: Oxford University Press), 23–53.

Men, H. (2018). *Vocabulary Increase and Collocation Learning.* Singapore: Springer. doi: 10.1007/978-981-10-5822-6

Namvar, F. (2012). The relationship between language proficiency and use of collocation by Iranian EFL students. *3L Southeast Asian J. Engl. Lang. Stud.* 18, 41–52.

Nesselhauf, N. (2003). The use of collocations by advanced learners of English and some implications for teaching. *Appl. Linguist.* 24, 223–242. doi: 10.1093/applin/24.2.223

Nesselhauf, N. (2005). *Collocations in a Learner Corpus*. Amsterdam: John Benjamins. doi: 10.1075/scl.14

Nizonkiza, D. (2011). The relationship between lexical competence, collocational competence, and second language proficiency. *Engl. Text Constr.* 4, 113–145. doi: 10.1075/etc.4.1.06niz

Obukadeta, P. (2014). "L2 collocations: a problematic linguistic phenomenon?," in *Proceedings of the Talk Given at the Birmingham English Language Postgraduate Conference*, (Birmingham: University of Birmingham).

Palmer, H. E. (1933). *Second Interim Report on English Collocations*. Tokyo: Kaitakusha.

Paquot, M. (2018). Phraseological competence: a missing component in university entrance language tests? Insights from a study of EFL learners' use of statistical collocations. *Lang. Assess. Q.* 15, 29–43. doi: 10.1080/15434303.2017.1405421

Paquot, M., and Granger, S. (2012). Formulaic language in learner corpora. *Annu. Rev. Appl. Linguist.* 32, 130–149. doi: 10.1017/S0267190512000098

Pawley, A., and Syder, F. H. (1983). "Two puzzles for linguistic theory: nativelike selection and nativelike fluency," in *Language and Communication*, eds J. C. Richards and R. W. Schmidt (London: Longman), 191–226.

Sarvari, S., Gukani, A. H., and Khomami, H. Y. (2016). Teaching collocations: further developments in L2 speaking fluency. *J. Appl. Linguist. Lang. Res.* 3, 278–289.

Shih, R. H.-H. (2000). "Collocation deficiency in a learner corpus of English: from an overuse perspective," in *Paper Presented at the 38th Annual Meeting of the Association for Computational Linguistics*, Hong Kong.

Shumin, K. (2002). "Factors to consider: developing adult EFL student speaking abilities," in *Methodology in Language Teaching: An Anthology of Current Practice*, eds J. C. Richards and W. A. Renandya (Cambridge: Cambridge University Press), 204–211. doi: 10.1017/CBO9780511667190.028

Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Takahaski, G. (1969). Perception of space and the function of certain English prepositions. *Lang. Learn.* 19, 217–234. doi: 10.1111/j.1467-1770.1969.tb00464.x

Wray, A. (2000). Formulaic sequences in second language teaching: principle and practice. *Appl. Linguist.* 21, 463–489. doi: 10.1093/applin/21.4.463

Wray, A. (2002). *Formulaic Language and the Lexicon*. Cambridge: University Press. doi: 10.1017/CBO9780511519772

Xu, J. (2015). *Predicting ESL Learners' Oral Proficiency by Measuring the Collocations in Their Spontaneous Speech* (Unpublished doctoral dissertation). Ames, IA: Iowa State University.

Xu, J. (2018). "Spoken collocational competence" in communicative speaking assessment. *Lang. Assess. Q.* 15, 255–272. doi: 10.1080/15434303.2018.1482900

Xu, J., and Xi, X. (2010). "Comparing human and machine judgments of collocations and relating them to speaking proficiency," in *Paper Presented at the 32nd Language Testing* Research Colloquium (Cambridge: University of Cambridge).

Zughoul, M. R., and Abdul-Fattah, H. (2003). Translational collocational strategies of Arab learners of English: a study in lexical semantics. *Babel* 49, 59–81. doi: 10.1075/babel.49.1.05zug

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

APPENDIX

Appendix A: A Sample Transcription of the IELTS Speaking Test Interview (With a Band-Score 8 Test-Taker):

Examiner: This is The X mock speaking test for the International English Language Testing System conducted on February 6'h, 2020 at the X. The candidate number is X. Good afternoon. Can I have your full name please?
Candidate: My name is XX.

E. In first part, I'm going to ask you some questions about yourself. Let's start by talking about what you do. Do you work or are you a student?
C: Both I should say. Actually, I'm a teacher and I'm about to finish my PhD program.
E. Ohoom. What work do you do exactly?
C: I am an English instructor. This is my main job. I also do some editing and some research if you call it a work.
E: Ok.
C: Yeah.
E. How popular is this job in your country?
C: Well I should say is quite popular eh especially in terms of prestige I think English teachers have a high prestige although the salaries are often very low and most people just look- uh look up to English teachers somehow. Uh well I think many people think of- it consider it as a kind of uh prosperous of I mean job.
E. Ok. Thank you.
C: You're welcome.
E. Let's talk a little bit about television. How much TV do you usually watch?
C: Well uh actually I watched TV every night but actually I don't plan to watch TV it's just on and I just want in front of it while doing my other stuff I just spend time watching those soap operas, too.
E. Do you have a favorite TV program?
C: Not actually I don't watch TV. I mean it's not my favorite hobby to watch TV. I just do it as a kind of every day activity.
E. How much did you watch TV when you were a child?
C: Actually, I watched TV a lot. I used to watch a lot of uh cartoons and animations actually I didn't miss anything on TV and uh well I stayed up late at night watching those soap operas with my family too. And uh I used to watch sports uh like different kinds of matches. It was somehow very interesting.
E. Thank you. Do you think television has changed in the past few decades?
C: Television has changed in the past! Uh I guess yes. The quality of the programs of course the-the quality of everything in-in the society has changed. I should say the quality has raised a little bit. Of course, we are not comparable to other countries but as far as we are talking about Iranian programs compared to the past, I should say yeah, the- the quality has changed, and the types of the programs have also changed because I remember when I was watching TV as a child, uh most of the cartoons we used to watch were just so sad and very
E: Ok.
C: Yeah.
E. Has television changed your life in any way?
C: My life! Uh I don't think so. No. I haven't been influenced by TV.
E. Thank you. Let's talk a little bit about the countryside. Uh would you like to live in the countryside in the future?
C: Uh you mean living permanently? Uh It's an option for me maybe after I get rid of course I should decide later but yeah, I just uh don't uh reject it was an al-alternative because I think we need- we all need all kinds of uh peace in our later lives so yeah maybe.
E. Ok. What are the benefits of living in rural areas?
The benefits in first part- the first one is peace and quiet. The second one is like the fresh air eh you just have to every day I mean it's kind of eh eh it's a very big oppor-opportunity for the people who are living in uh metropolitan cities like Tehran So I guess these two are enough to just convince a person to move to the countryside. The other stuff like are I think privacy If you just have good- good place for yourself
E. Thank you very much.
C: You're welcome.
E. Now I'm going to give you a topic and I'd like you to talk about 1 to 2 minutes. Before you talk you've got one minute to think. You can make some notes if you wish. Do you understand?
C: Sure.
E. So, here's some paper and sometimes
C: Thank you.
E. And here's your topic. I'd like you to describe a small business that you would like to own.
C: The small business! Ok.

[After two minutes]

C: Here you are.

E. You can have it to two minutes.

C: I can have it?

E. Yeah. So, remember you have one to two minutes to talk about this topic so don't worry if I stop you. I'll tell you when the time's up. Can you start please?

C: Sure. Well uh the best small business I would like to uh start uh actually to own by myself is a café. I-I always thought about having a café uh in just uh decorating it the way I wish and running it the way that I would like. And uh before I start this business I think I have to educate myself in uh some related area such as how to run a café first of all and how to manage just uh these kinda business then I have to know about the different kinds of desserts and the food. Uh even I should know about like uh how to deal with the customers I think and I would uh run this business of course run this cafe using some uh qualified maybe eh eh baristas and I don't know waiters and I have to just uh find intricate rules some of those eh best ones may be those who are just uh made for this job and experts so that I can run the away I wish so uh this is my first choice because I've al-I've always enjoyed being in cafés. I-I also uh have good memories of our friends X spending time and I think most-most of the people who just go to café our younger ones and I also like spending time with young generation so I think the energy in a café < > properly somehow very fresh and into me so that's why I just choose this one after that I would just go for other businesses. my own my- husband is just into it so that's another reason I would just uh do it in the future. Yeah

E. Have you ever own your own business?

C: No, I haven't.

E. Would you like to have a family business?

C: Uh if-if my partner is uh considered yes but the other family members I wouldn't say yes because we have different ideas but-but my husband is actually on my side so I think we're just on the same track if we just wanna run a business together so I think he's a good option to be a partner.

E. Ok. thank you very much can I have the card and the notes please? C: Here you go.

E: Thank you very much. We've been talking about a small business you would like to on now I'm going to ask you some more general questions relate to this. Let's consider first of all small businesses. Uh what types of small businesses are most popular in your country?

C: These days I see a lot of start-ups being- working like eh eh being run- run by the young-young people, eh university students and I think mostly in IT and to l-like to start applications to-to pro-program- app- computer programs to design applications I think now this is very popular Iran.

E: Why? Why do you think these businesses are popular?

C: Well it's quite clear you know because internet is just the main uh uh- the main forum- or the let's say the ma- the mainstream uh medium for communication for running different businesses so many people are just- tend to- they just tend to uh get f-familiar with eh how to work with Internet so I think it's the first reason the most important reason is actually the eh of the commonality of Internet-based businesses.

E: Ohoom. What challenges and difficulties do people face when they try to start a small business?

C: A small business in general, you mean?

E: Yeah.

C: Uh first of all you- I mean a person who just starts a business has to have some qualities I mean there are some features that a person has to possess like self-discipline and let's say uh leadership or like the ability to work on his or hor- her own. So, the first thing eh comes to my mind regarding the question uh you just asked is eh having these personal qualities. This-this is a challenge for many people because sometimes they're not used to that- that kind of work and the other stuff is like the support from the government and from the society, so they need just you uh uh work very hard. They have to just uh eh spend a lot of time and energy and sometimes money to attract uh customers and also to attract support in part of their work.

E: Ok thank you. How can small businesses benefit local people?

C: Uh small businesses! uh I think eh there is something eh which has to be considered when we are talking about marketing and stuff that- that is trust. So, for the local people I think it's much easier to trust eh eh to trust eh one's eh whom X uh. Eh I myself I would just trust the person whom I-I just know for a long time. I think is the benefit for the local people around that eh small business. For bigger companies, usually we don't have this kind of uh knowledge and familiarity about them, so it's really hard to just f- trust them.

E: Ok. Let's talk a bit about business owners right now. Why do some people start their own businesses?

C: First of all, it's because of the many of the problems that exists. When you are- when you work for others- when you're employed by someone such- such as lack of freedom or uh sometimes the uh hostile atmosphere some people experience in eh their workplaces. So, they just go for running their own business I mean standing on their own feet somehow. And the other part I guess s- running a business by their own is actually more beneficial financially so people just uh hope to earn more if they run their own businesses, so I guess these are enough reasons to just to uh business of course if they- when this is kind of eh risk-taking job, so they take a risk to uh with the hope of earning some benefit in the future.

E: Ok. Thank you very much. This is the end of the speaking.

Appendix B: Syntactic Patterns of Lexical Collocations (Adapted From Xu, 2015, p. 80)

Type	Example from the corpus
1. Adjective and noun	pensive mood
2. Adverb and adjective	undeniably significant
3. Adverb and verb	sincerely thank
4. Noun and noun	job opportunities
5. Noun of noun	acts of terrorism
6. Noun and verb	the business flourishes
7. Verb and noun	make a decision
8. Phrasal verb and adverb	stay up late
9. Noun and phrasal verb	time goes by
10. Phrasal verb and noun	come across obstacles

Appendix C: Syntactic Patterns of Grammatical Collocations (Adapted From Benson et al., 2010, pp. 19–30)

Type	Example from the corpus
1. Noun and preposition	skill in
4. Preposition and noun	on TV
5. Adjective and preposition	aware of
8. Verb and preposition	depend on

frontiers in Psychology

ORIGINAL RESEARCH
published: 28 March 2022
doi: 10.3389/fpsyg.2022.835340

Creative Togetherness. A Joint-Methods Analysis of Collaborative Artistic Performance

*Vincent Gesbert[1], Denis Hauw[2], Adrian Kempf[3], Alison Blauth[4] and Andrea Schiavio[3]**

[1] Football Club Lorient, Lorient, France, [2] Institute of Sport Sciences, Faculty of Social and Political Sciences, University of Lausanne, Lausanne, Switzerland, [3] Center for Systematic Musicology, University of Graz, Graz, Austria, [4] Artistic Swimming Swiss National Federation, Lausanne, Switzerland

In the present study, we combined first-, second-, and third-person levels of analysis to explore the *feeling of being and acting together* in the context of collaborative artistic performance. Following participation in an international competition held in Czech Republic in 2018, a team of ten artistic swimmers took part in the study. First, a self-assessment instrument was administered to rate the different aspects of togetherness emerging from their collective activity; second, interviews based on video recordings of their performance were conducted individually with all team members; and third, the performance was evaluated by external artistic swimming experts. By combining these levels of analysis in different ways, we explore how changes in togetherness and lived experience in individual behavior may shape, disrupt, and (re-)stabilize joint performance. Our findings suggest that the experience of being and acting together is transient and changing, often alternating phases of decrease and increase in felt togetherness that can be consistently recognized by swimmers and external raters.

Keywords: togetherness, joint performance, individuality, collectivity, sports psychology

OPEN ACCESS

Edited by:
George Waddell, Royal College of Music, United Kingdom

Reviewed by:
Ruud J. R. Den Hartigh, University of Groningen, Netherlands
Juliane J. Honisch, University of Reading, United Kingdom

****Correspondence:***
Andrea Schiavio
andrea.schiavio@gmail.com

Specialty section:
This article was submitted to Performance Science, a section of the journal Frontiers in Psychology

Received: *14 December 2021*
Accepted: *18 February 2022*
Published: *28 March 2022*

Citation:
Gesbert V, Hauw D, Kempf A, Blauth A and Schiavio A (2022) Creative Togetherness. A Joint-Methods Analysis of Collaborative Artistic Performance. Front. Psychol. 13:835340. doi: 10.3389/fpsyg.2022.835340

INTRODUCTION

Individuals displaying high-level expertise in sports and the arts usually operate in a performative niche involving multi-leveled layers of reciprocal interaction (Carron et al., 2002). For example, many skilled musicians perform in ensembles, play for an audience, learn music with and through others, and develop important relationships with the cultural norms and narratives sedimented in their social and historical environment. Similarly, athletes often rely on team effort and develop their skills through training sessions that are in most cases collaborative (e.g., with a coach, other athletes, etc.). Being together with others is therefore increasingly understood as a fundamental resource that can shape skill acquisition and creative performance across a range of individual and collective contexts (see Davids et al., 2007; Hauw, 2018; Schiavio et al., 2019). Although this dimension of being together is perhaps less apparent in instances of solitary musicking and sports performance (but see Høffding and Satne, 2019; Schiavio et al., 2020, in press), it is clearly manifested in a population of musical ensemble and sports team members, whose performances are constantly organized and carried out through a moment-to-moment participation with co-performers, team members, audience, and/or opponents.

There is a vast literature on the psychological dynamics associated with creative teamwork in sports and the performing arts, including studies focused on *group cohesion* (see e.g., Spink, 1990; Heuzé et al., 2006; Lund et al., 2014; Glowinski et al., 2016), *collective creativity* (Santos et al., 2016, 2017; Bishop, 2018), *coordination dynamics* (Keller et al., 2014; Laroche et al., 2014; Miyata et al., 2017; Himberg et al., 2018) as well as *synchrony and self-other overlap* (Lakens, 2010;

Lakens and Stel, 2011; Rabinowitch and Knafo-Noam, 2015; Tunçgenç and Cohen, 2016). Group cohesion is usually defined as "a dynamic process that is reflected in the tendency for a group to stick together and remain united in the pursuit of its instrumental objectives and/or for the satisfaction of member affective needs" (Carron et al., 1998, p. 213). In this process of reciprocal collaboration, novel (e.g., expressive, behavioral) joint configurations can emerge, giving rise to creative outcomes that play out at different layers of awareness (see Walton et al., 2015; Orth et al., 2017; Kimmel et al., 2018).

Examples can be found in studies investigating how action patterns developed in response to unexpected occurrences during performance (e.g., a novel strategy displayed by the opponent team, an unanticipated subtle change in the re-creation of the musical score by a co-performer, etc.) can lead to functional and innovative (or indeed, *creative* – see Runco and Jager, 2012) modifications in behavior. Here contextual adaptations and activities are often negotiated in both local (e.g., what action can be performed individually?) and global (e.g., what collective configuration can emerge from individual behaviors?) terms. Consider how in a soccer game, for instance, a range individual and collective factors, like tiredness or a change in tactics, are highly co-dependent and can shape how teammates respond to particular contextual contingencies – an example being the modification of an existing defensive strategy (see Duarte et al., 2012). Here adaptations are thought to be continuously developed in response to a range of moment-to-moment perturbations that disrupt the stability of the joint activity (see Gesbert and Durny, 2017; van der Schyff et al., 2018; Schiavio et al., 2021).

Within such contexts, these evolving behavioral modifications have been increasingly studied in terms of *coordination dynamics* (see Kelso, 2001, 2003; Tognoli et al., 2020). The main idea is to conceive of a joint performance as a uniquely structured system based on a reciprocal interplay of biological and ecological parameters that recursively change over time (see Chow et al., 2011; Seifert et al., 2013). As such, the set of constraints and open possibilities offered by the physical and social environment in which the performance unfolds is functionally coupled with the shifting behavioral trajectories of the performers, giving rise to a distributed network of co-dependencies that sustains, transforms, and re-orients the joint performance (Chemero, 2009; Hristovski et al., 2012). The constant re-organization of this agent-environment system involves (creative) changes that play out at both macro- and micro-scales (see Demos et al., 2018; Schiavio and Benedek, 2020). Accordingly, not only do kinematics, motor plans, predictions, and outcomes exhibit visible modifications, but the subtle, personal experiences that permeate joint performances are also subject to transformations in the here-and-now. Genuinely subjective descriptions of these dynamics are notoriously difficult to obtain, and they escape the analytic approaches relying on quantitative methods. Yet, gaining a deeper understanding of the individual experiences involved in collective behaviors is of major importance for developing a more integrated view of joint activity – one that places equal emphasis on its local and global components, as well as on both experiential and behavioral dimensions (see Tanaka, 2017; Høffding, 2019).

In this study, we aimed to contribute to this line of research by exploring in greater detail the experience of being and acting together (or "togetherness") emerging from collaborative performance (see Bourbousson and Fortes-Bourbousson, 2017). According to Himberg et al. (2018), the feeling of being and acting with others is indeed an essential part of collective performance as it facilitates the regulation of individual and collective behaviors in light of the direct, immediate experience of others (see Colombetti and Torrance, 2009; Froese and Di Paolo, 2011; He and Ravn, 2018). How do performers describe their experience of being and acting together? What role does it play in regulating and optimizing joint action? And can this sense of togetherness be perceived from the outside, for example, by an audience? To provide some preliminary answers, we report on an original study that focused on collaborative artistic activity (synchronized swimming) and adopted a "joint-methods" approach – one that combines first-, second- and third-person levels of analysis. The first-person data were generated via the administration of a self-assessment instrument, the second-person data were based on interviews (see Petitmengin, 2006), and the third-person data were obtained from independent raters who assessed the swimmers' performance. We suggest that this methodology may provide rich understandings of the creative dynamics of skilled action during participatory activity, offering insights generated on intrapersonal and interpersonal levels. This can mutually "validate and constrain" empirical data generated via more traditional methods (Varela and Shear, 1999, p. 6).

Rationale for the Study

Research exploring the interplay of individual and collective dynamics usually focuses on two levels of description: phenomenological and behavioral (Hauw, 2018; Gesbert and Hauw, 2019). To bring together and complement these two lines of inquiry, Seifert et al. (2016) offered a more unifying approach to study interpersonal coordination in sports based on mixed methods. This approach combines two analytical strategies: the first explores the continuities and discrepancies between behavioral and phenomenological data via side-by-side comparisons, and the second integrates theory-driven categories with tools from ecological dynamics. A good example of the first strategy is the *neurophenomenological* approach, which relies on quantitative analyses of measurable phenomena related to brains, bodies and behaviors, while "embracing the value of first-person reports of experience" (Bockelman et al., 2013; see also Varela, 1996; Lutz, 2002; Lutz et al., 2002; Petitmengin and Lachaux, 2013; Depraz and Desmidt, 2019). Concerning the second line of enquiry proposed by Seifert et al. (2016), one might consider how the recent work by Kimmel and Rogler (2018, 2019) applies theoretical resources from micro-phenomenology, cognitive ethnography, and ecological dynamics to explore qualitatively how agents carry out high-level interacting skills in the context of Aikido. In their examination of the moment-to-moment web of interactivities unfolding between performers, they found that "collective dynamics and individual affordances dialectically engender each other" (Kimmel and Rogler, 2018, p. 251), pointing to the co-specification of individuality and

collectivity in embodied decision-making during emergent collaborative activity.

As the datasets in this type of study are often complex, many researchers have moved from a *mixed-methods* approach (see Anguera et al., 2017) to *joint-methods* (see e.g., Poizat et al., 2012; Sève et al., 2013; R'Kiouak et al., 2016; Hauw et al., 2017; Seifert et al., 2017; Rochat et al., 2020). The former refers to the practice of juxtaposing and/or comparing qualitative and quantitative data to cross-correlate specific aspects of experience with possible behavioral outcomes (see e.g., Vors et al., 2019), whereas the latter instead uses two domains of evidence to enrich an initial analysis based on data with a specific format. In this case, there is *a first choice that determines how additional data can be successfully integrated*. For example, Rochat et al. (2019) recently conducted a study explicitly inspired by such an approach to investigate the experiences of trail-runners interacting with five different water-carrying systems. Nine runners were equipped with a carrying system (e.g., a backpack with two front bottles on the shoulder straps; a waist pack with the bottles on the hips, etc.) and ran a 3-kilometer loop at a regular pace; at the end of each trial, they were instructed to change the water-carrying system, and they then repeated the loop with another carrying system. They repeated the loop five times, each time with a different system. For each trial, the runners were also equipped with inertial sensors to measure both their vertical oscillations and those of the five carrying systems. After the five loops, the runners were individually interviewed and asked about the different "traces" of their past activity, such as pictures and maps of the route and pictures of themselves during the transitions between trials. This confrontation was designed to help the runners access and describe their experience at the moment their activity was unfolding. More specifically, the authors first sought to document the salient aspects associated with the carrying systems during the unfolding activity at the phenomenological level in order to determine the relevant dependent variables to investigate (e.g., when the runners described disturbing system elements like the feeling of the system bouncing in an uncomfortable way). From these qualitative insights, two hypotheses emerged in relation with the behavioral data (i.e., the vertical oscillations of the runners' hip and the backpack) characterizing low- and high-order parameters of behavior, such as the couplings between the accelerations of the runners and the backpacks.

The present study builds on these methodological insights to explore the "feeling of being and acting together" or "sense of togetherness" associated with the ability of team members to successfully coordinate with each other. The study took place during an international competition held in 2018 in Czech Republic, where members of a team of swimmers performed a free combination routine, which was then assessed at the three above-mentioned levels (1st person, 2nd person, and 3rd person). Synchronized swimming is a form of collective performance involving two, eight, or ten swimmers performing a synchronized routine of elaborate moves in the water, accompanied by music. We chose synchronized swimming because this activity demands elaborate individual and collaborative skills – for example, propelling the body through hand movements while performing upside down, achieving stability and height above the water while leaving the hands free to perform arm motions, gaining a sense of how the team is performing, and so on. Here the creative aspects of the process relate to those real-time strategies that swimmers adopt to compensate for possible problems in the exercise, thereby regaining coordination in different ways.

Among others, there are two important questions a joint-methods analysis can help answer: (i) how can changes in individual performance disrupt the unfolding dynamics of interpersonal coordination? and (ii) how can swimmers compensate for destabilizations in team performance and regain individual and collective stability? Moreover, the elision of individual behavior and ecological constraints (i.e., the music guiding performance, the particular environment in which it takes place, etc.) complements the research to date on behavioral co-adaptation and joint action. For example, Froese et al. (2014) investigated how social agents actively co-regulate their interactions in the service of joint action, but no aesthetic or creative dimension was considered. Given its individual, collaborative, and ecological complexity, synchronized swimming is an ideal candidate for investigation.

METHODS

Participants

Ten artistic swimmers participated in this study. They were between 15 and 22 years old (M = 17.8; SD = 2.4) and had been practicing artistic swimming between 8 and 13 years at the time of the study (M = 9.9; SD = 1.7). They were informed of the study purpose and told that their participation was entirely voluntary. Before the study began, they or their families (for those under 18 years) approved, and gave written consent to a protocol agreement that described the study purposes in detail and ensured confidentiality and anonymity (i.e., swimmers were given pseudonyms). Participants were already known to the first author due to an ongoing collaboration. They were not monetarily compensated for their participation. In addition to these swimmers, five other expert swimmers, blind to the aims of the study, were recruited to perform an observational analysis of the swimmers' performance. Their age was 28.4 on average (SD = 5.9). They had all been artistic swimmers at the international level and had trained between 5 and 15 years (M = 8; SD = 4). They were recruited by the fourth author, and they were not monetarily compensated. The study project was not submitted for approval to the Ethics Commission of University of Lausanne as it did not fall within the legal obligations in Switzerland. Indeed, according to Swiss law, only studies dealing with health data must be submitted to an Ethics Commission for authorization. Since this was not the case for our study, we were not required to do so. Nevertheless, the data collection respected the common ethics rules in psychology and was in accordance with the Declaration of Helsinki: the procedures for data collection and analysis were explained in detail to the participants, who gave written informed consent to participate, as did parents/guardians for those under 18 years. The athletes' anonymity was guaranteed by an anonymous login created by

each athlete and only the first researcher knew the link between the athlete and the login.

Data Collection

Data were collected during an international competition that took place in 2018. The main focus of the study was a free combination routine performed by ten swimmers. A *free combination* is a routine that may be a compound set of solos, duets, trios and other team segments. No technical event is "prescribed" for the free combination routine, and swimmers can be quite creative by, for example, presenting higher and bigger lifts. The observed routine lasted about 4 min and 20 s and it was the first time it was performed in competition. Three types of data were collected: (i) first-person, with the swimmers' self-assessments of their feelings of being and acting together during the routine. These were collected just after competition using Likert scales administered to each participant; (ii) second-person, concerning the swimmers' qualitative experiences during the routine. These were collected between 1 and 5 days after the competition by individually confronting the swimmers with the video of their performance; and (iii) third-person, based on the experts' assessments of togetherness. One week after competition, these raters evaluated the team performance and individuated units of behavioral team activity through observational analysis of the video recordings. In what follows, we provide more detailed information concerning the units of analysis and the three types of data reported above.

Before the data were collected, two independent experts (blind to the purposes of the study) observed the choreography in the last training sessions before competition and then were asked to segment the team performance (i.e., routine) into discrete collective units of behavioral activity (see Zacks and Swallow, 2007; Kurby and Zacks, 2008, for a similar approach). The video was segmented into 33 units. This segmentation allowed us to capture the shifting dynamics of being together (at first- and third-person levels) and provided the ground from which the subsequent interviews were structured. The team performance was considered a unitary system, with each component sustaining and regulating the continuous dynamical interplay between its different phases and trajectories (see van der Schyff et al., 2018). On this basis, we conceived of the system transitions as emerging units for analysis. These involved series of episodes or segments with a clearly visible beginning and end. Indications of the end of a unit might be a change in team organization, ending a figure, or moving in a new direction. Put differently, each collective unit corresponded to a behavioral shift within team performance, in a sense similar to what Himberg et al. (2018) described as sudden bifurcations from one state to another. Consider, for example, collective unit of activity 2, which is topographically and contextually represented in, respectively, **Figures 1a,b**.

Here, the position of each swimmer corresponds approximately to her position in the swimming pool during the collective unit of activity. Each swimmer was labeled S1, S2, and so on. The arrows between two swimmers meant bodily contact (e.g., bodily contact was established between S5 and S7, S2 and S6, etc.). For this collective unit, the team was split into eight and two swimmers. Indeed, S1 and S8 were expected to be aligned on both sides of the highlight produced by the eight other swimmers.

Self-Assessments

After competition, all swimmers were invited to assess their feelings of acting together on Likert scales. The purpose was to explore team performance from a *first-person* perspective across the 33 collective units of activity on 7-point scales ranging from: (1) "I had the feeling that we did not act together" to (7) "I had the feeling that we acted together." The scores of all swimmers were then summed up together to obtain a team score for each unit (the maximum score for the team was 70, see **Table 1**).

Interviews

Individual interviews were conducted between 1 and 5 days after competition with all swimmers (on average: 3 days) by the first author, who was present during the competition to collect their self-assessments. He confronted the swimmers with the video of their performance, as well as their post-competition self-assessments. He also encouraged each participant to re-enact her pre-reflective experiences that emerged during the competition, helping them to describe her past lived experience (see Legrand, 2006). According to this *second-person* method (see Froese et al., 2011), it is possible to re-enact an experience when one is guided into an appropriate evocation state by a suitably skilled interviewer (see also Hauw, 2009; Vermersch, 2009; Olivares et al., 2015). In this case, the interviewer had experience in conducting similar interviews (see e.g., Gesbert and Durny, 2017; Gesbert et al., 2017; Hauw et al., 2017; Rochat et al., 2018; Gesbert and Hauw, 2020). Each interview was designed as follows: swimmers were first invited to describe, comment on, and explain their behavior for each collective unit of activity by first helping them to rediscover the spatiotemporal context of their past experience (i.e., when, where, with whom, etc.). The first author asked questions that the swimmers could not reply to without referring to the past competition (e.g., "when you were performing this move, what was your main focus?"). The performance video and self-assessments were therefore used to guide and help each swimmer evoke and describe her own experience during this past competition (see **Figure 2**). To ensure that the swimmers were relating to their own past experience he was attuned to behavioral indicators such as eye shifting or slowing of the word flow, which he considered useful information[1] (see e.g., Petitmengin, 2006). Second, the swimmers were prompted to describe their experience as it occurred in a specific situation, thus without involving retrospective generalizations or comments.

During the interviews, the first author was also sensitive to verbal indicators (e.g., "at this moment," "here," "there") that the swimmers related to their past experience. Once an evocation of the swimmers' past experience was established, the interviewer helped them describe this in detail, using questions concerning

[1]These behavioral indicators were inserted into the interview transcripts. This allowed us to remove elements (such as retrospective generalizations or comments) which we considered had no direct relationship to the swimmer's experience during the competition.

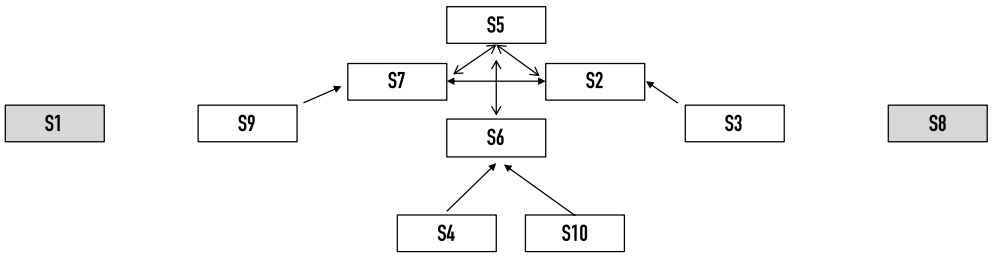


FIGURE 1a | Topographic representation of collective unit of activity 2.

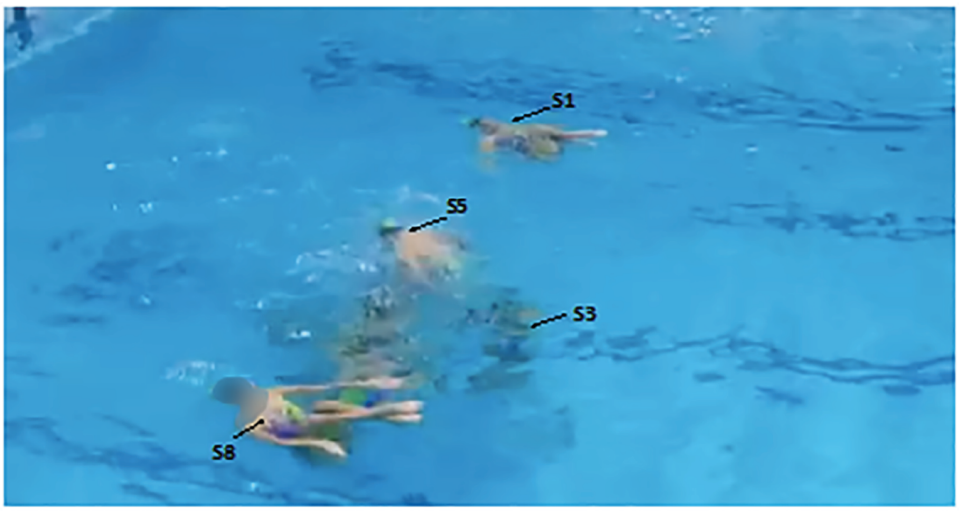


FIGURE 1b | Contextual representation of The collective unit of activity 2.

TABLE 1 | Illustration of first- and third-person data in relation with the swimmers' feeling of being and acting together and the experts' assessment of their togetherness during collective unit 2.

Swimmer	Swimmer's self-assessment	Expert 1 rating	Expert 2 rating	Expert 3 rating	Expert 4 rating	Expert 5 rating
S1	7	5	7	5	6	5
S2	4	3	4	4	2	4
S3	4	3	4	4	2	4
S4	5	3	4	4	2	4
S5	1	3	4	4	2	4
S6	6	3	4	4	2	4
S7	6	3	4	4	2	4
S8	7	5	7	5	6	5
S9	7	3	4	4	2	4
S10	7	3	4	4	2	4
	54	*34*	*46*	*42*	*28*	*42*
Team score	**54/70**			**38.4/70**		

their associated physical or mental activities (e.g., "what were you doing?"; "what were you thinking about?"), bodily sensations (e.g., "can you describe the main body sensations in this situation?"), concerns and volitions (e.g., "what did you want to do at this moment?"; "what were the main worries during this configuration?"), and elements that were drawing their attention (e.g., "what are you focused on at this moment?"). The interviews lasted between 90 and 120 min each; they were video-recorded and transcribed *verbatim* for further analysis.

Experts' Judgment

Five experts in artistic swimming assessed the togetherness displayed by the swimmers from a *third-person* perspective. As they viewed the performance video, the experts were prompted

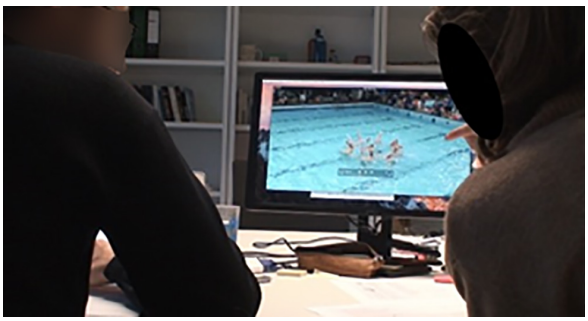


FIGURE 2 | Illustration of the interview situation (the interviewer is on the left and the swimmer is on the right).

to judge the level of togetherness displayed by each swimmer for each collective unit of behavioral activity. To do this, they scored togetherness on a Likert scale from (1) "there is no togetherness" to (7) "there is togetherness." Due to the specificity of the routine investigated (i.e., a routine may be made up of sets of solos, duets, trios and other team segments), the experts were asked to assess this sense of togetherness according to the specific moves performed by the swimmers. As an illustration, if one collective unit of activity was characterized by a group figure (with 8 swimmers) and a duet (see **Figure 1b**), the experts separately assessed togetherness for the eight swimmers and the two remaining swimmers (see **Table 1**). By adding these scores together, *a team score* was obtained for each expert. Then, by averaging these team scores, an expert team score was obtained (for which the maximum was also 70).

Data Analysis

The data analysis consisted of four main phases, each designed to capture a different aspect of the performative experience enacted during the collaborative action. The first three phases corresponded to the analyses of data collected within the three levels described above (1st-, 2nd-, and 3rd-person levels). The fourth phase was a joint analysis of these data to reach a more general level of description, as illustrated in detail in the section "Results." Statistical data analysis was conducted in R (R Core Team, 2020).

Assessing Togetherness From Within

First, using the swimmers' self-assessments about their feeling of being and acting together, each collective unit of behavioral activity was characterized through a team score of being and acting together from the swimmers' point of view. For instance, collective unit 2 was characterized by a team score of 54 (see **Table 1**). Thus, by calculating the score of being and acting together from the swimmers' perspective for each collective unit of activity, the dynamics of this feeling during choreography-performance and the 33 collective units of activity was assessed.

Re-enacting Togetherness

In phase two, the verbal narratives corresponding to the swimmers' experience were processed in three steps (see **Figure 3**) following a technique inspired by the *course-of-action* framework[2] (see e.g., Hauw and Durand, 2008; Hauw, 2009; Poizat et al., 2012; Sève et al., 2013; Mohamed et al., 2015; Theureau, 2015; Rochat et al., 2018). These steps are presented below.

In the first step, the stream of the swimmers' past experience was rebuilt. To do so, the swimmers' experiences were progressively connected by presenting them with the collective units of activity in chronological order. This helped them describe their moment-to-moment experiences with accuracy (see e.g., Gesbert et al., 2017). For example, they were able to check on the spot whether their sense of togetherness was convergent during a specific segment of the performance or not.

In the second step, the participants' transcribed verbal descriptions were categorized as *Units of Meaningful Action* (UMAs) (see e.g., Hauw and Durand, 2008; Sève et al., 2013; Mohamed et al., 2015; Rochat et al., 2018). These UMAs correspond to the smallest units of action that were experienced as meaningful for the swimmer at a given moment. They stemmed from the link between the action and the associated thoughts, or interpretations. UMAs were labeled using a verb followed by a direct object, an adverb, or another complement (e.g., senses that she is unbalanced just before the compression). This coding also helped us simultaneously label the underlying constituents of each UMA, which were identified using a set of more specific questions. Having defined the UMAs, a further categorization differentiated them and described with more precision the range of motor possibilities, perceptive experiences, and proprioceptive feelings that emerged in the choreography. These were labeled as *involvements* (I), *actions* (A), or *perceptions* (P). Involvements were identified by asking the following question: "What were the significant concerns experienced during the choreography?" Actions referred to what the swimmer was actually doing, whereas perceptions included the situations that the participants experienced as significant. As such, P could include the other swimmers' activity (distance, alignment, compression, etc.), the material environment (e.g., the underwater lights, the pool ceiling etc.), the counts of the choreography and/or the key moments in the music, or a sensation (e.g., bodily contact, balance, etc.). Within each UMA, studying the relationships among I, A, and P helped us capture important aspects of what a person felt, thought, and did (see e.g., Hauw and Durand, 2008; Rochat et al., 2018).

Finally, in the third step, the emergent dynamics of togetherness were explored. The starting point was to describe how each swimmer experienced the feeling of being and acting together in each collective unit of behavioral activity (see e.g., Sève et al., 2013; R'Kiouak et al., 2016; Seifert et al., 2017). This description was based on a detailed examination of two components of each UMA (i.e., I and P). We were thus able to identify different ways of experiencing being and acting together from the swimmers' perspective.

[2]As a methodological framework, the *course-of-action* technique (see Theureau, 2015) aims to help athletes rebuild their activity by re-enacting their experience (i.e., what they perceived, felt, and did at a given moment).

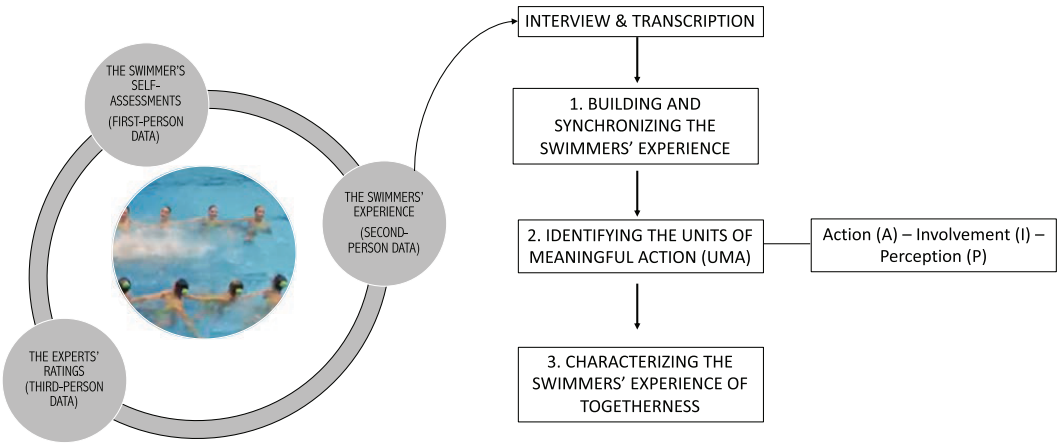


FIGURE 3 | Representation of the three analytic layers, with a special focus on the second-person analytic procedure.

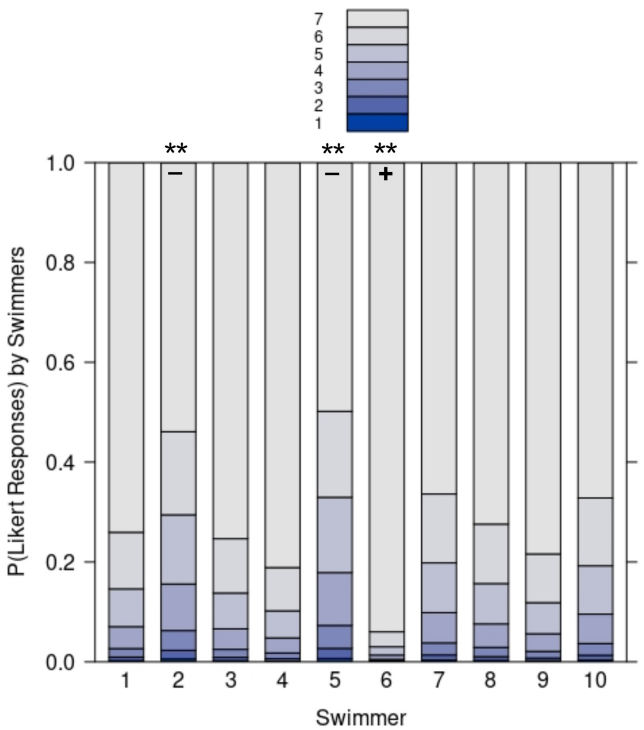


FIGURE 4 | The probability of Likert responses for each swimmer rating their perceived togetherness. Positive (+) and negative (–) deviation of a swimmer from the overall mean are marked if significant (**$p < 0.01$).

Assessing Togetherness From Outside

The third phase of the analytic process focused on the experts' assessments. Here, each collective unit of activity was characterized by a team score of togetherness (see **Table 1**). For instance, collective unit 2 was characterized by a team score of 38.4. This procedure was carried out for all 33 collective units of activity, allowing us to explore the visible dynamics of togetherness from a third-person level as they developed through the performance.

A Joint-Methods Approach

In the fourth and final stage of our analysis, we integrated the first-, second-, and third-person data through different combinations to render the dynamics of creativity and togetherness on the artistic swimming team intelligible, and enrich our understanding of the process. The process can be summarized as follows: *first,* by comparing first- and third-person data, we explored whether the swimmers' feeling of being and acting together corresponded to the togetherness perceived

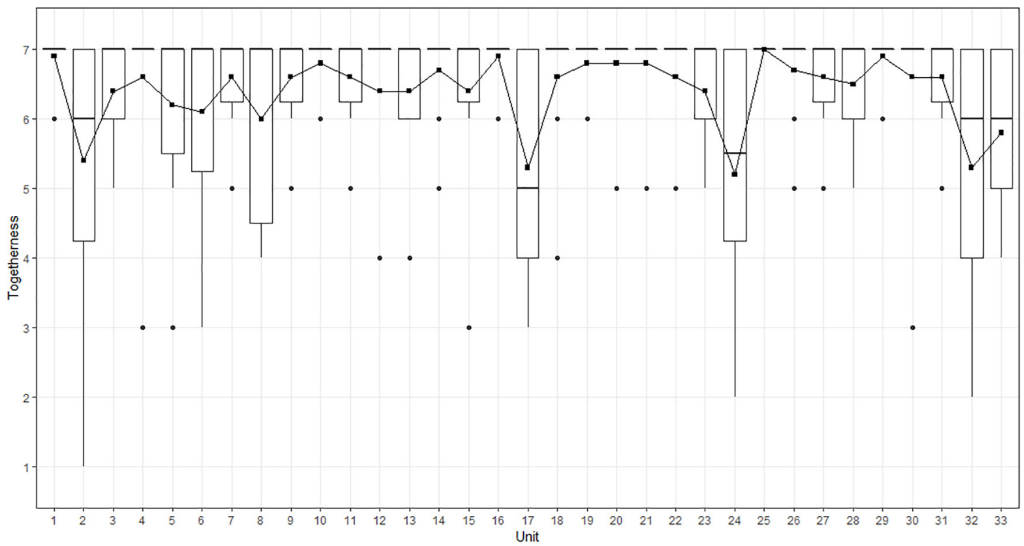


FIGURE 5 | The dynamics of togetherness from the swimmers' point of view during choreography-performance. The *x*-axis corresponds to the collective units of behavioral activity. The *y*-axis corresponds to the rated togetherness by swimmers. Squares indicate the mean rated togetherness by the swimmers for each unit, which corresponds to the scaled team score by factor 1/10.

by the experts during performance. *Second*, by scrutinizing the first-person data, we identified the collective units that were felt as problematic by the swimmers. The second-person data were then used to access and expand on the information offered by the swimmers during these specific units. Finally, the third-person data were used to examine how the experts assessed the swimmers' togetherness in these problematic units. *Third and last,* by examining the second-person data, we first observed how the team (i.e., all the swimmers) experienced a collective feeling of being and acting together during performance. After identifying the collective phenomenological categories for each unit of activity, we focused on specific categories by characterizing them through the first- and third-person data.

RESULTS AND DISCUSSION

In what follows, six main results are presented and contextualized. The first two describe the feeling of being and acting together during the choreography from the swimmers' points of view. These emerged from the self-assessments (1st person), and from the interviews (2nd person). The third result focuses on the experts' judgments (3rd person). The last three results emerged from the joint-analysis of the multiple-leveled data to explore in greater detail how changes in individual performance were able to disrupt the unfolding dynamics of interpersonal coordination and how the swimmers were able to compensate for destabilizations in team performance to regain individual and collective stability.

The Feeling of Being and Acting Together

The median of the swimmers' self-assessments for the 33 collective units during the choreography was 7 (IQR = 1). The divergence of each swimmers' self-assessment from the overall perceived togetherness of the group was quantitatively assessed by calculating an ordinal logistic regression using the polr function from the R package MASS (Venables and Ripley, 2002). To ensure that the parallel regressions assumption was not violated, we ran the brant test (Brant, 1990) using the R package brant (Schlegel and Steenbergen, 2020). In doing so, we determined the statistical model with swimmer as fixed effect factor using sum contrasts (Schad et al., 2020). The analysis showed that self-assessments of swimmer S2 and S5 were significantly lower (S2: $b = -0.86$, $SE = 0.32$, z: −2.69, p: 0.007; S5: $b = -1.02$, $SE = 0.33$, $z = -3.13$, $p = 0.002$) than the overall swimmers' self-assessment. On the contrary, self-assessment of swimmer S6 was significantly higher ($b = 1.73$, $SE = 0.67$, $z = 2.6$, $p = 0.009$) than the overall swimmers' self-assessment. The probabilities for the seven Likert responses according to the model are visualized for each swimmer in **Figure 4**. So, while S2 and S5 tended to rate togetherness during choreography slightly lower than the other team members and that, in contrast, S6 tended to rate togetherness higher than the others. **Figure 5** presents for each collective unit (*x*-axis) the togetherness score at the team level (*y*-axis) corresponding to the addition of the swimmers' self-assessments of each unit. The curve describes the dynamics of togetherness enacted by the ten swimmers during the performance (as a reminder, the maximum score was 70).

The results showed fluctuations in togetherness as the performance unfolded. Some collective units were characterized by a sudden decrease, such as units 2, 17, 24, and 32, which corresponded to collective moves and/or technical figures for which two or more swimmers felt a weakening in togetherness. This is also indicated by the increased dispersion of ratings for these units as depicted in **Figure 5**. Yet, these sudden decreases were systematically followed by an immediate or

progressive regain (see units 3, 4, 18, or 26). Such a regain of togetherness is also connected in a reduction of dispersion of ratings across consecutive units. To better grasp these team scores, we scrutinized the swimmers' individual perspectives (see **Table 1**). For instance, for unit 2, the sudden decrease at the team level may be explained by the perspective of six swimmers (S2, S3, S4, S5, S6, and S7) who were performing a highlight (see **Figure 1**) and felt a weakening in their togetherness during this figure. Then, to better grasp the swimmers' weakening in their togetherness, we explored how they described their lived experience during this specific figure.

Swimmers' Experience Rebuilt

The analysis performed on the swimmers' interview data indicated four ways of experiencing togetherness (see **Table 2**). The 330 UMAs corresponding to the ten swimmers' experience during the 33 collective units were examined.

The first dimension was the *experience of togetherness* (T), corresponding to 71% of the UMAs (236/330). It was characterized by the swimmers' feeling of effectively interacting with the others and producing the choreographic performance that was expected. As an example, consider the following quote:

> *"There, we just did the body boost and then it's the start of the lift. As I'm the swimmer who's the farthest away, I have priority to pass. I do a long breaststroke and I feel that the girls let me through. I thought it was very good, this transition is going well, I don't need to put in extra effort." (S6, Collective Unit 24).*

TABLE 2 | The swimmers' experience of togetherness during choreography-performance.

	Perceptions (P)	Involvements (I)
Togetherness (T)	The habitual bodily contact with another swimmer The right tempo with other swimmers The nice "shape/form" of the formation The right alignment with and right distance from the other swimmers The timing of the movements	Maintain the right distance and stay in contact with the other swimmers Focus on being aligned with the other swimmers
Weakened Togetherness (WT)	Not sufficiently aligned A little too close to or too far from the other swimmers A little too close to or too far from a partner Insufficient compression	Adjust in order to be aligned Slow down or speed up one's movements Push the partner to be in her place Put in a little more effort Reduced time for adjusting self
Absence of Togetherness (AT)	A swimmer has lost the count Body contact is much more condensed than usual The formation has no shape The sensation of being pushed diagonally Chaos while getting into place underwater	Is unable to adjust Has no say in the adjustment process Is prevented from adjusting Has no time to adjust
Meaningless Togetherness (MT)	Each has her own count and her own sensations The coach's instructions	Be as aesthetically pleasing as possible Be tuned into one's own sensations Recall the coach's individual instructions

The second category that emerged from the second-person data was the *experience of weakened togetherness* (WT), corresponding to 14% of the UMAs (46/330). It was characterized by the feeling of not being sufficiently coordinated to produce the expected choreographic performance. The following two quotes exemplify this feeling:

> *"S3 is in front of me, I put my hand on her shoulder to check for the correct distance between us and I have to check the alignment with S10, who's in the other line. I see that we're not too aligned with S10, so I tried to slow down to get into alignment with her." (S9, Collective Unit A7).*

> *"On this mini-lift, I'm with S3 but it doesn't help much. This lift is really hard. I feel it isn't high enough out of the water. . . Yeah, S3 isn't helping me enough." (S10, Collective Unit 28).*

The third category was the *experience of the absence of togetherness* (NT), corresponding to 10% of the UMAs (33/330). This was expressed as bad feelings about collective action, as if something was wrong or was not happening as usual. Consider the following quotes from two swimmers:

> *"Before I jumped, I felt like it wasn't going to work. I'm watching S2 and S7 so I can get on top of them, but I couldn't see their hands but didn't know why! I'm way behind,... here I'm completely behind." (S5, Collective Unit 2).*

> *"I move to the left to be next to S6. The others carry us from behind. I feel the pressure of feet below S7, and the pressure of S5's feet above me. Here we're way too early. When I started to build power, I felt S5 leaving, although for me it was not at all on the count. Usually, I feel that she has time to position her feet and there the moment was very condensed compared to the usual." (S2, Collective Unit 2).*

Finally, the last category that emerged from the interview was the *experience of meaningless togetherness* (MT), for 5% of the UMAs (15/330), especially noted when the swimmers performed technical figures: with the upper body, with the legs (e.g., ballet leg), with spins or solo. This experience could be defined as the feeling of being attuned only to one's own movements, rather than those of the rest of the team. Relevant examples can be found in the following verbal descriptions:

> *"There I have to do a body boost with S2, but I'm only tuned in to my own boost because I can't see S2, who's behind me." (S5, Collective Unit 24).*

> *"Here, I'm paying attention to make the small corrections that the coach gave me and I'm only tuned in to my own figure, looking for my own bodily sensations. The angle is super important, I can't let my legs be too low or too high. . . Once I feel that I have it (the right height), I keep it here." (S8, Collective Unit 30).*

Perceiving Togetherness

The median of the experts' ratings for the swimmers' togetherness for the 33 collective units was 5 (IQR = 2). We assessed the interrater reliability by calculating the intraclass correlation coefficients (ICC) as described by Koo and Li (2016). ICC estimates and the 95% confidence intervals were calculated with the help of the R package "psych" (Revelle, 2021) on base of a mean rating (k = 5), absolute agreement, two-way mixed

effect model. We report a moderate to good ICC for the expert ratings (ICC = 0.73, 95% CI [0.68, 0.78]). As with the analysis of swimmers' self-assessment, we aimed at exploring which swimmers deviated from the group as assessed by the expert ratings. To do so, an ordinal logistic regression was performed using the polr function from the R package Mass (Venables and Ripley, 2002). The model was set up using sum contrasts (Schad et al., 2020) and swimmer as fixed effect factor. Results

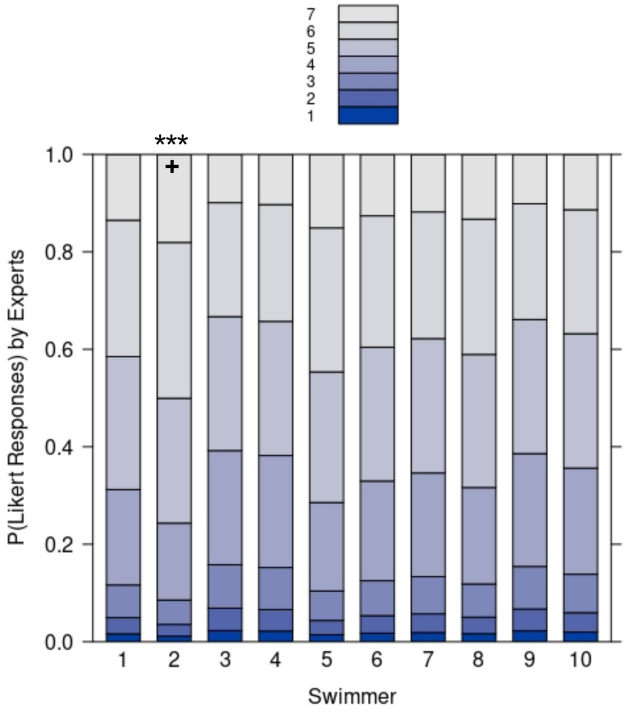


FIGURE 6 | The probability of Likert responses of experts rating the togetherness of each swimmer. Positive (+) deviation of a swimmer from the overall mean are marked if significant (***$p < 0.001$).

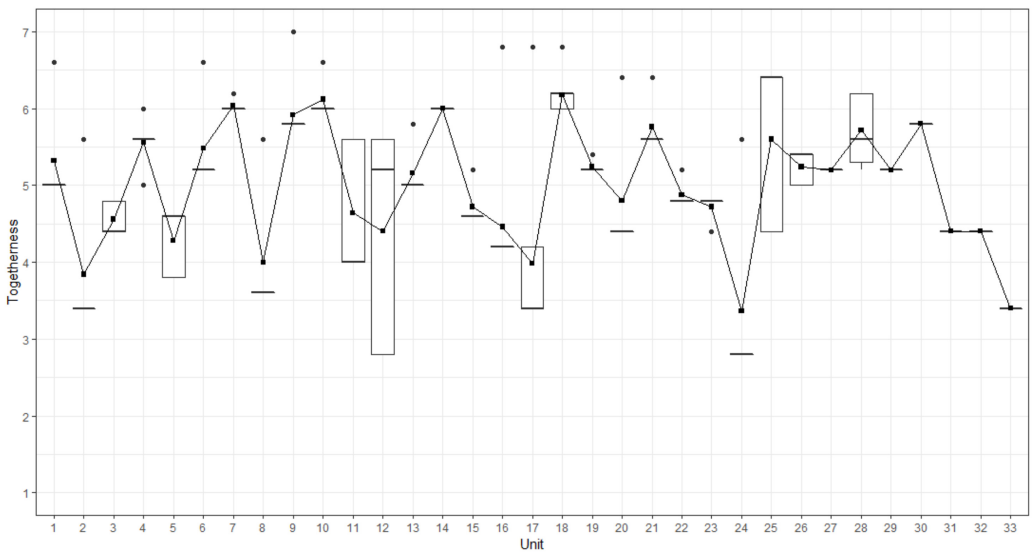


FIGURE 7 | The dynamics of togetherness from the experts' point of view during choreography-performance, rated for each swimmer. The *x*-axis corresponds to the units of behavioral activity. The *y*-axis corresponds to the rated togetherness by experts. Squares indicate the mean rated togetherness by experts for each unit, which corresponds to the scaled team score by factor 1/10.

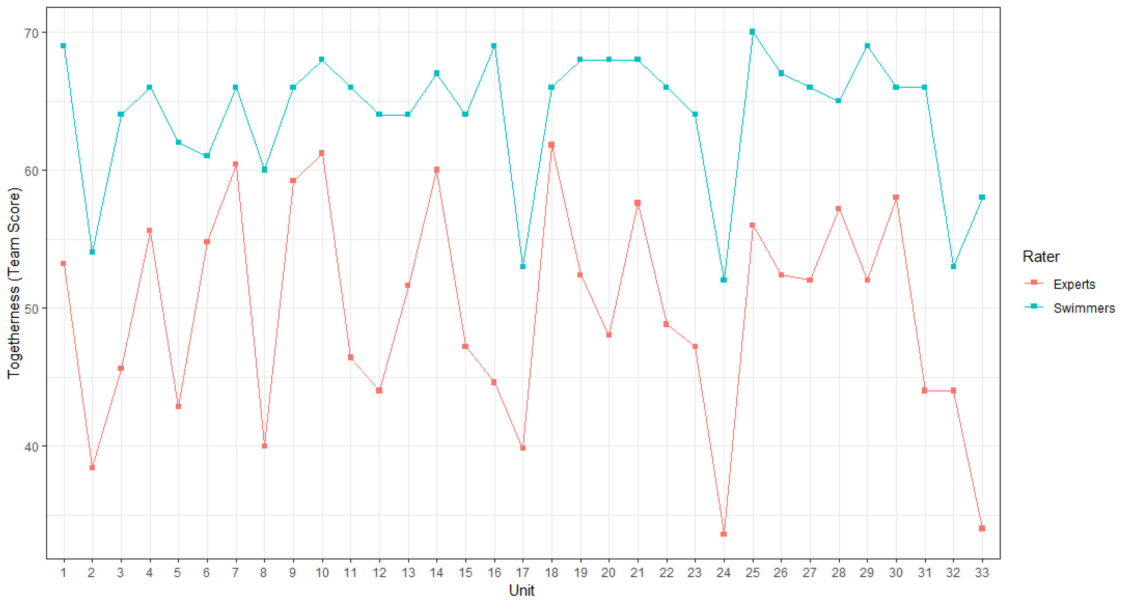


FIGURE 8 | Joint analysis between first- and third-person data. The blue curve above corresponds to the team score of the swimmers' self-assessments (first-person data). The curve below corresponds to the team score of the experts' ratings (third-person data).

showed that experts rated only the togetherness of swimmer S2 as significantly higher ($b = 0.44$, $SE = 0.13$, $z = 3.3$, $p < 0.001$) compared to the overall rated togetherness of the group. The probabilities for the seven Likert responses according to the model are visualized for each swimmer in **Figure 6**. **Figure 7** depicts the mean togetherness score for each collective unit (x-axis) from the experts' point of view (y-axis). The curve represents the dynamics of togetherness at the team level assessed by the experts during the choreography-performance (as a reminder, the maximum score was 70).

Some units were characterized by a sudden decrease, such as units 2, 5, 8, or 24 (corresponding to discrete movements), whereas units 15, 16, and 17 (corresponding to a set of specific and linked movements during the choreography) were characterized by a progressive decrease. To better grasp these fluctuations, we scrutinized the experts' individual perspectives, which revealed how each expert rated togetherness for each swimmer (see **Table 1**). For instance, for unit 2, all the experts assessed the togetherness of the swimmers performing the highlight (i.e., S2, S3, S4, S5, S6, S7, S9, and S10) between 2 and 4 on the 7-point Likert scale, whereas togetherness for the swimmers involved in a duet (S1 and S8) was assessed between 5 and 7.

Joint Analysis Between First-, Third-, and Second-Person Data

For this procedure, we first scrutinized the comparisons between the first- and third-person data in order to delineate the samples of the second-person data to be analyzed. The first objective was to understand how the swimmers' experience of being and acting together fit with the togetherness assessed by the experts (see **Figure 8**). Hence, we ran an ordinal logistic regression using the polr function from the R package Mass (Venables and Ripley, 2002) to compare the ratings of both groups across all units. The model was set up using sum contrasts (Schad et al., 2020) and rater group as fixed effect factor. This comparison revealed that, on average, swimmers rated the perceived togetherness significantly higher ($b = 1.3$, $SE = 0.07$, $z = 18.93$, $p < 0.001$) than experts. This coherent deviation of swimmers and expert ratings can also be found when inspecting the summed ratings for each unit of the choreography in **Figure 8** (i.e., the orange curve was always below the blue curve). The figure shows that the swimmers' feeling of acting together is always exceeded by the experts' perceptions of togetherness for all the units of activity. Three most and least diverging units were identified: units 16, 33, and 31 and units 6, 18, and 7 respectively (in increasing order of divergence).

Overall, the shapes of the curves present the same profile with ascendant and descendant trajectories, except for eight transitions. For instance, between units 18 and 19, the swimmers' feeling of acting together increased, whereas the assessment of togetherness from the experts' perspective decreased. In contrast, between units 29 and 30, the swimmers' feeling of acting together decreased, whereas the assessment of acting together from the experts' point of view increased. When identifying the collective units for which the swimmers' feeling of acting together decreased whereas the assessment of togetherness from the experts' point of view increased (i.e., transitions between units 5–6, 27–28, 29–30, and 31–32), we scrutinized the swimmers' experience from their point of view (see **Table 3**).

For the transition between units 31 and 32, togetherness slightly increased from the experts' point of view (i.e., the team score varied between 40 and 44), whereas the swimmers felt a

TABLE 3 | The swimmers' experience during collective unit 32.

Swimmers	The feeling of being and acting with each other	Unit of meaningful action (UMA)	Involvement (I)	Perception (P)
S1	7	Backs up to be in the circle	To participate in forming a circle with the other swimmers	Aligned with the swimmer opposite her
S2	2	Tells self that they're too spread out just before turning around	To be unable to adjust	Sees swimmers on the side she's on that she shouldn't be able to see
S3	5	Realizes that they're not at all together when she turns around	To be unable to adjust	Swimmers next to her
S4	7	Has the impression that all is correctly positioned behind her	To line up to be in the circle	The positions of S7 and S10
		Realizes that it's not the case when she turns	To turn around to perform the figure	The swimmers' positions
S5	4	Turns around and perceives that they are too far apart to be in formation	To look at the swimmer at the end and be at the right distance from the nearby swimmers	Alignment and distance
S6	7	Backs up to be in the circle	To try to line up and be at the right distance	The alignment in a square
S7	3	Backs up to be in the circle	To be aligned with S2 and the right distance from the nearby swimmers	Alignment and distance with the swimmers
		Realizes that they're not in a circle when she turns around	To be unable to adjust	Poor alignments
S8	7	Gets adjusted with her partner	To adjust with S1	S1
S9	7	Gets a good feeling about the figure being performed	To line up with the opposite swimmer and manage the distance with the other two swimmers	Alignment and distance with the swimmers
S10	4	Backs up to be in formation	To get between S5 and S8 and stay attentive to the alignment with S3	Positions of S3, S5 and S8
		Turns and sees the catastrophe	To be unable to adjust	Poor alignments

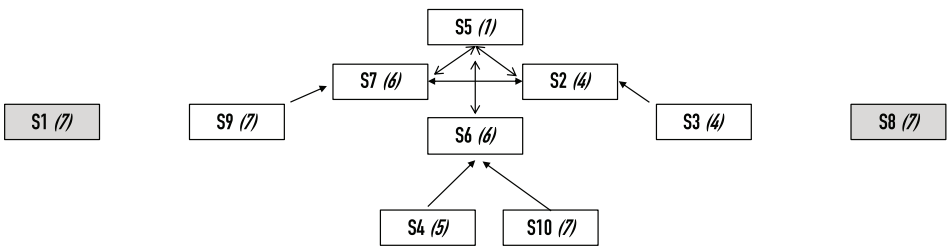


FIGURE 9 | The individual swimmers' perceptions of the feeling of being and acting together for collective unit 2.

substantial weakening in their feeling of being and acting together (i.e., the team score varied between 66 and 54). The second-person data indicated that most of the swimmers (i.e., S3, S4, S5, S7, and S10) perceived that the shape being enacted by the set of swimmers did not correspond to what it was expected only after they had turned around. Only then did they realize that they were in trouble not so much in terms of alignment as in distance and that they were now unable to cope. Only one swimmer (S2) explained that, just before turning around, she had realized that they were not together because she could

TABLE 4 | Contribution from first-, second- and third-person data for unit 2.

Swimmers	Feeling of being and acting together	Perception (P)	Expert 1′s ratings	Expert 2′s ratings	Expert 3′s ratings	Expert 4′s ratings	Expert 5′s ratings
S1	7	Alignment with another swimmer	5	7	5	6	5
S2	4	S5′s push was more condensed than usual	3	4	4	2	4
S3	4	A setback in the boost phase, more difficult for her to push S2	3	4	4	2	4
S4	5	The diagonal position of S6 (rather than the expected vertical)	3	4	4	2	4
S5	1	Feels too far back in the platform (imbalance)	3	4	4	2	4
S6	6	Sensitive to body contact with S2 and S7, then with S3 – some difficulty moving and orienting/localizing herself	3	4	4	2	4
S7	6	The time S2 needed to find her and S6	3	4	4	2	4
S8	7	Alignment with another swimmer	5	7	5	6	5
S9	7	Bodily contact with S7 and perception of S5′s difficulty in finding S2 and S7	3	4	4	2	4
S10	7	The feeling of pushing with S3 and S6	3	4	4	2	4

see the swimmers to the side, which she could not usually do when the figure was performed correctly, though she could see no possibility of adjustment.

Joint Analysis Between First-, Second-, and Third-Person Data

This procedure involved first scrutinizing the first-person data to delineate the samples of second-person data to be analyzed. From the first-person data (see **Figure 7**), we were able to identify all the units that had been experienced as problematic, with the team score of feeling of acting together below 60 (i.e., units 2, 3, 6, 8, 17, 24, 32, and 33). Although this value of 60 was subjective, it indicated that two or more swimmers felt a weakening in their feeling of being and acting together. For instance, unit 2 was characterized by a team score of 53. The individual swimmers' perceptions of feeling of acting together for this unit are indicated in brackets in **Figure 9**.

The number in parentheses indicates the score for feeling of being and acting together expressed by each swimmer from her perspective.

Six swimmers felt a weakening in their feeling of being and acting together (i.e., S2, S3, S4, S5, S6 and S7), and these six swimmers were involved in performing a specific figure (i.e., a highlight[3]). We investigated their experience to determine what was meaningful for them at this instant in the situation (see **Table 3**). The results indicated the rich variety of information that the swimmers drew on to explore their feelings of being and acting together. Included were the alignment with one or more swimmers, the push with one or more swimmers, the expected position of one swimmer, the compression co-built with another swimmer, the balance built with other swimmers, and the time taken to achieve a shape. In the last stage, the five experts' assessments of their togetherness were examined to determine how they fit or did not fit with the swimmers' feelings of being and acting together (see **Table 4**). These third-person data offered the opportunity to observe whether the experts were attuned to the big problems (i.e., characterized through a low team score of togetherness) or the global form enacted by the swimmers. In this last case, low variability was acceptable for the experts.

Joint Analysis of the Second-, First-, and Third-Person Data

In this procedure, we first scrutinized the second-person data to delineate the samples of the first- and third-person data to be examined. As a reminder, the analysis of the swimmers' interview data revealed four ways of experiencing togetherness (see **Table 2**). *Meaningless togetherness* was used to indicate that the swimmers were not paying attention to togetherness at the pre-reflective level of their activity (i.e., labeled MT, in purple). The other types of experience accounted for the UMAs in which the swimmers reported salient experiences of togetherness. The *absence of togetherness* accounted for the UMAs in which the swimmers reported that they were not being and acting together (labeled AT, in red). Weakened togetherness accounted for the UMAs in which the swimmers reported a meaningful experience of a weakening in being and acting together (i.e., labeled WT,

[3]This technical term in artistic swimming refers to an acrobatic movement with a platform of swimmers who unite their maximal power to propel one of them, the flyer (in this case, S5), completely out of the water.

TABLE 5 | The swimmers' experience of togetherness during choreography-performance.

UNIT 1				***PLATFORM LIFT***						***DUET S1 and S8***
T	T	T	T	T	T	T	T	T	T	
UNIT 2				***HIGHLIGHT***						***DUET S1 and S8***
T	WT	WT	WT	AT	AT	AT	AT	T	T	
UNIT 3			***BODY BOOST***						***MINI LIFT***	
T	T	T	T	T	WT	AT	AT	T	WT	
UNIT 4			***HIGHLIGHT***				***FIGURE S7 and S3***		***S2 and S6***	
T	T	T	T	T	T	T	T	T	AT	
UNIT 5			***BODY BOOST***					***S7 and S3 – S2 and S6***		
T	T	WT	WT	WT	MT	T	T	T	T	
UNIT 6				***HIGHLIGHT***					***DUET S9 and S10***	
T	T	T	T	WT	WT	AT	AT	T	T	
UNIT 7				***MOVE IN TWO LINES***					***DUET M + B***	
T	T	T	T	T	T	WT	WT	T	WT	
UNIT 8				***TWO-WAVE LIFT***					***MINI-LIFT***	
T	T	T	T	AT	AT	AT	AT	T	T	
UNIT 9				***FIGURE***					***SOLO***	
T	T	T	T	T	T	WT	WT	WT	T	
UNIT 10				***CIRCLE***					***LIFT S1 and S6***	
T	T	T	T	T	T	WT	WT	T	T	
UNIT 11			***MOVE WITH LEGS***					***MOVE WITH ARMS***		
T	T	T	T	WT	MT	T	T	T	WT	
UNIT 12		***MINI-LIFT***				***MINI-LIFT***			***BODY BOOST***	
T	T	T	T	T	T	T	T	AT	MT	
UNIT 13				***FIGURE***					***DUET***	
T	T	T	T	T	T	WT	WT	T	WT	
UNIT 14				***FIGURE WITH ARMS***					***DUET***	
T	T	T	T	T	T	T	W	T	M	
UNIT 15				***TEAM + SPIN***					***DUET***	
T	T	T	T	WT	AT	MT	MT	T	T	
UNIT 16					***SURFACE FIGURE***				***SOLO***	
T	T	T	T	T	T	T	T	T	MT	
UNIT 17			***BALLET LEGS***					***BODY BOOST***	***SOLO***	
AT	AT	MT	MT	MT	MT	T	T	WT	MT	
UNIT 18		***LIFT S5 (FIRST BASIS)***					***LIFT S5 (SECOND BASIS)***		***SOLO***	
T	T	T	T	WT	T	T	T	T	T	
UNIT 19				***FIGURE WITH ARMS***					***DUET***	
T	T	T	T	T	T	WT	WT	T	T	
UNIT 20				***BOX-BOX FIGURE + 2 LINES***					***DUET***	
T	T	T	T	T	T	T	WT	T	T	
UNIT 21				***ONE-LINE FIGURE***					***DUET***	
T	T	T	T	T	T	T	WT	T	T	
UNIT 22				***BODY BOOST BY 2***					***DUET***	
T	T	T	T	T	T	T	WT	T	WT	
UNIT 23				***TWO LINES***					***DUET***	
T	T	T	T	T	T	T	WT	WT	WT	
UNIT 24				***LIFT S3***					***BODY BOOST***	
T	T	WT	WT	WT	AT	AT	AT	T	MT	
UNIT 25		***BODY BOOST***			***UNDERWATER MOVE***				***BARRACUDA***	
T	T	T	T	T	T	T	T	T	T	
UNIT 26			***MOVE***					***MOVE (KICK)***		
T	T	T	T	T	T	T	T	T	WT	
UNIT 27					***FIGURE***					
T	T	T	T	T	T	T	T	WT	A	
UNIT 28		***BODY BOOST***			***MINI-LIFT***				***BODY BOOST***	
T	T	AT	T	T	T	AT	T	T	T	
UNIT 29				***FIGURE WITH THE UPPER BODY***						
T	T	T	T	T	T	T	T	T	T	
UNIT 30					***BOX-BOX FIGURE***					
T	T	T	T	T	T	T	T	MT	MT	
UNIT 31					***FIGURE WITH TWO LINES***					
T	T	T	T	T	T	T	T	WT	AT	
UNIT 32					***FIGURE IN CIRCLE + BODY BOOST CIRCLE***					
T	T	T	WT	AT	AT	AT	AT	AT	AT	
UNIT 33					***LIFT S5***					
T	T	T	T	WT	WT	WT	AT	AT	AT	

T, Togetherness; WT, Weakening Togetherness; AT, Absence of Togetherness; MT, Meaningless togetherness.

TABLE 6 | Contribution from second-, first-, and Third-person data.

Collective units	CPC	Swimmers' feeling of togetherness	Experts' perceptions of togetherness	Swimmers' feeling of togetherness	Experts' perceptions of togetherness	Swimmers' feeling of togetherness	Experts' perceptions of togetherness
1	CPC1	6.9 *(0.3)*	5.53 *(0.9)*	7 *(0)*	6.6 *(0.5)*		
2	CPC4	5 *(2)*	3.4 *(0.8)*	7 *(0)*	5.6 *(0.8)*		
3	CPC4	6.33 *(1)*	4.4 *(1.4)*	6.5 *(0.6)*	4,8 *(1,4)*		
….	…	….	…	….	…	….	…
12	CPC4	7 *(0)*	5.6 *(1.2)*	7 *(0)*	2.8 *(1.4)*	4 *(0)*	5.2 *(1)*
…	…	….	…	….	…	….	…
32	CPC4	5.3 *(1.9)*	4.4 *(0.5)*				
33	CPC4	5.8 *(1.2)*	3.4 *(0.5)*				

CPC: Collective Phenomenological Categories; CPC1: Simultaneously and Similarly Experienced as Togetherness; CPC2: Simultaneously and Similarly Experienced as Togetherness within a subgroup and simultaneously diverging experience in another subgroup; CPC3: Simultaneously Diverging Experiences in Two subgroups; CPC4: Simultaneously Highly Diverging Experiences.

in green). *Togetherness* (T) accounted for the EUMs in which the swimmers reported a meaningful experience of being and acting together (in yellow). For each collective unit of behavioral activity, the experience of each swimmer was labeled in one of these four phenomenological categories in relation with their position on the team (see **Table 5**).

Four collective phenomenological categories were identified:

- CPC1: Simultaneously and Similarly Experienced as Togetherness at team level.
- CPC2: Simultaneously and Similarly Experienced as Togetherness within a subgroup and simultaneously diverging experience in another subgroup.
- CPC3: Simultaneously Diverging Experiences (i.e., two different ways of experiencing togetherness within two subgroups).
- CPC4: Simultaneously Highly Diverging Experiences (i.e., three or four ways of experiencing togetherness at the team level or within two subgroups).

These categories helped us grasp whether the swimmers similarly or differently perceived their being and acting together. Then, for each collective unit, the average individual scores for the feeling of being and acting together (first-person data) and the perception of togetherness (third-person data) were indicated for each move or figure characterizing this collective unit. For instance, units 1, 2, and 3 were characterized by two distinct moves/figures, whereas unit 12 was characterized by three distinct moves.

For each of these moves, the results indicated the average individual score for the swimmers and the experts, as well as the standard deviation (see **Table 6**).

For collective unit 12, for example, the results showed that CPC4 was characterized from the swimmers' perspective through a single move between two swimmers (i.e., subgroup 3) in which togetherness was rated 4. However, the experts perceived togetherness during this move as higher than the swimmers' perceptions (Msubgroup 3 = 5.2). In contrast, for other moves in which the swimmers' perceived togetherness was equal to 7, the mean of the experts' ratings was lower (Msubgroup 1 = 5.6; Msubgroup 2 = 2.8).

GENERAL DISCUSSION AND CONCLUSION

The aim of this paper was to offer a detailed description of the feeling of being and acting together in the context of collaborative artistic performance. The feeling of being and acting together has often been understood as a crucial dimension for optimal collaborative activity in sports and music (see e.g., Lund et al., 2014; Schiavio and Høffding, 2015; Himberg et al., 2018). We therefore chose to focus on synchronized swimming as it requires skill in both aesthetic and athletic components (e.g., rhythmical entrainment, competitiveness, sportsmanship, etc.). We first developed two assessment instruments so that the swimmers could evaluate their feeling of being and acting together and the expert raters could also evaluate their togetherness, thus providing us with first- and third-person perspectives. We then conducted interviews based on elicitation techniques in order to perform a second-person level of analysis (see e.g., Gesbert et al., 2017; Gesbert and Hauw, 2020). This allowed us to explore in greater detail the moment-to-moment experiences that permeated the swimmers' activities at given moments. By combining these methodological approaches *via* joint analysis, we obtained precise descriptions of how the changes in individual and collective behavior shaped, disrupted, and re-stabilized an artistic performance.

We found that the swimmers who took part in the study were highly attuned to their feeling of being and acting together during the execution of the choreography. Although this result was not fully surprising, the combination of multiple analytical tools helped us provide a detailed description of the interplay between the singular and plural dynamics at the heart of team effort. By integrating the scores of togetherness assigned to each unit of behavioral activity with the verbal descriptions from the interviews, the fluctuating, situated nature of the feeling of being and acting together emerged. In particular, our analysis suggests that despite the planned patterns of behavior defining the choreography, this latter is less static than one might think. Indeed, artistic swimmers often adjust their behaviors in light of an immediate experience of interaction with one or many team members: our first-person data (self-assessments) revealed

how they individually feel togetherness during competition. Since they act and react according to a set of dynamical and evolving constraints (see Davids et al., 2007), their experience of togetherness is transient and constantly oscillating between increases and decreases of felt togetherness.

These first-person data also showed how, after each marked decrease in togetherness at the group level, an immediate active response occurred. The feeling of being and acting together therefore appeared to be crucial to creatively engaging with the contingencies and perturbations of performance, allowing the swimmers to immediately and efficiently adapt to their teammates and their situated activity. Even though the swimmers sought to actively regulate the interaction process as they accomplished a figure or a move together, they were constrained by ecological information: according to their position and their role in the choreography, they experienced different feelings of togetherness. This may explain why, during a highlight, the flyer rated togetherness as a 1 on the 7-point Likert scale, whereas the other swimmers rated it a 7. The flyer, who was situated at the top of the platform lift, felt the set of compressions produced by her teammates, whereas the other swimmers could only feel more attenuated parts of the compressions. It should be noted that the aquatic environment is inauspicious for exchanging information, and the choreography lasts 4 min with a preestablished chaining of figures, highlights, and moves. As such, it is fundamental to establish alternative ways of accessing information related to the behavioral dynamics of others, thereby creating a synergetic "we-experience" that transforms individuality and collectivity on the basis of a subtle sense of togetherness. In artistic swimming, togetherness is bonded in the very tuning of the performers' interactions with others as they strive for accuracy in the lines and positions of the formation, the distance they maintain between themselves, the rate at which the formation shifts, the beauty or aesthetic of the emerging figure, the tempo of their figures, and the musical interpretation expressed *via* their synchronized movements to the music.

Another important outcome of our analysis concerned the four ways in which artistic swimmers experience the sense of togetherness. We labeled these ways as: *togetherness*, *weakened* togetherness, *absence* of togetherness and *meaningless* togetherness. Each of these experiences arguably emerges from the integration of audio-visual and proprioceptive information about the alignment and/or the distance from other swimmers, the position within the formation, and the building of balance, timing and movements in relation with others—or rather their own feelings of staying synchronized with the others (see e.g., Gesbert and Hauw, 2020; Toner and Montero, 2020). In this last case, although a small part of the coded UMAs were characterized as meaningless togetherness, all of them were associated with specific segments of the team performance, thus revealing the situatedness of this experience during the choreography. By engaging with such information, the swimmers were able to adjust their activity and perform together (see I and UMA). Yet, however rich this information may have been, it could not always provide them with all the necessary resources to engage in the complex dynamics of their performance. Again, adaptations need to be made immediately, sometimes with enormous risks for the collective performative outcome. Accordingly, the swimmers also relied on a complementary set of tools centered on a more conative dimension (see Legrand, 2006). This also explains the rich variety of feelings associated with the choreography, which often fluctuated between the concrete immediacy of their activity and the expected outcome that was collectively built through hours of collective practice prior to the performance. Indeed, the swimmers' experience of full togetherness was often hampered by difficulties. These perceived "difficulties" were directly due to the structure of the choreography (i.e., the preestablished chaining between some of the moves was too difficult for some of the swimmers) or the swimmers' activity, such as inadequate positions, insufficient and/or unsatisfactory movements and so on. The interactions affecting their collective and individual activity were analyzed through the interviews, in which the swimmers were prompted to describe, comment on and explain the differences between the ideal and unfavorable conditions of reciprocal interaction and co-regulation.

The analysis of the expert ratings confirmed the key role of togetherness in determining how the various interpersonal synergies unfolded, with a special emphasis on the individual level rather than group level. Unlike other studies that have sought to characterize team performance in artistic and sports contexts (Sève et al., 2013; Vicary et al., 2017; Himberg et al., 2018), the present study dealt with artistic swimming, where a choreography can often be split into two or three interdependent subgroups. Within each subgroup, the performers seek to efficiently adjust their activity to the needs of the collective activity, such as, for instance, sufficiently pushing another swimmer upward before a highlight and adjusting the rate of their leg movements or their position within the formation in order to promote synchronized movements or better alignment among the swimmers. In the present study, only six of the 33 collective units were rated at the team level. For the other collective units, the experts were more attuned to the interaction process between two or more swimmers involved in the same move or figure (or in the achievement of the same task).

One of the main advantages of the joint-methods approach is the mutual enrichment of the domains of evidence, such as is offered by first-, second-, and third-person perspectives. The main idea is that putting together several levels of analysis can generate novel insights that recursively enrich each other, thereby bringing the subtle nuances of intersubjectivity into the daylight of lived experience (see Depraz et al., 2017, p. 192). A good example of this emerged from the integration of first- and third-person data, which permitted us to analyze two main aspects of the swimmers' performance. The first aspect was the coherent discrepancy emerging from the direct comparison of the swimmers' and raters' assessments: across all units, experts rated togetherness 1.3 points lower than swimmers. The second aspect, which the elision of the first and third levels of analysis brought forth, provided a description of the three most and least diverging units of behavioral activity (units 6, 18, and 7 and units 16, 33, and 31 respectively) between swimmers' and experts assessments. This showed that the swimmers' and experts' perspectives on these specific collective units differed from the otherwise coherent divergence between the group ratings.

To account for these differences, the constitutive elements of the swimmers' experience (UMAs, I, P) in these six collective units were given special attention during the interviews. Doing so provided us with new understandings of the swimmers' experience of togetherness as we explored in detail whether they adjusted (and how) their behavior in these specifically problematic moments and thereby enriched the initial data. Combining the first- and third-person data was not enough to yield precise insights into what information was meaningful for the team members to rate togetherness. During the individual interviews, all the performers were thus confronted with their togetherness self-ratings, and invited to comment on, explain, and describe in detail what they had felt during the performance.

This analytic approach builds on and extends the literature by providing an apt counterpoint to studies that focus separately on qualitative and quantitative data. For instance, there is a vast literature in the sports sciences that presents phenomenological data (i.e., second-person data) to identify the relevant dependent variables to be explored quantitatively (see Sève et al., 2013; R'Kiouak et al., 2016; Rochat et al., 2019). In a similar vein, other scholars often focus on behavioral data (i.e., third-person data) and then enrich these data with verbal accounts by the participants (see e.g., Seifert et al., 2017). By adding a further level of analysis, the present contribution provides a more holistic, real-time description of how togetherness evolves and shapes performance. Given the specificity of artistic swimming and the importance of its aesthetic dimension, we used expert perspectives as third-person data. Contrary to other studies in sports that have used biomechanical or behavioral indicators to assess, for instance, the "synchronization of the rowers" (see Sève et al., 2013; R'Kiouak et al., 2016; Seifert et al., 2017), in the present study we asked experts to rate how they perceived the swimmers being and acting together on a 7-point Likert scale. However, it might also be possible to look for behavioral indicators in the swimmers' experience that would account for the togetherness they feel. For instance, future studies could measure the distance and/or alignment between the swimmers at the most difficult moments of the choreography and compare these data with the subjective accounts emerging from the first and second levels of analysis. This can also include a broader examination of the target population's creative potential. Consequently (at least in the contexts of creative togetherness described in this work), not only the joint methods approach has a precious ally in the theoretical resources of works on dynamic systems and ecological dynamics (see e.g., Araújo and Davids, 2004; Araújo et al., 2006; Araújo et al., 2017; Kimmel, 2017, 2019; Schiavio and Kimmel, 2021); it could also be integrated with specific tests and measurements relating to individual and group (motor) creativity (see e.g., Bournelli et al., 2009; Grammatikopoulos et al., 2012; Santos and Monteiro, 2021) in order to increase its explanatory power. This might be beneficial when it allows for broader analytical tools combining interactions at the micro and macro level, the experience and formation of meaningful behavioral patterns, and the creative drive that favors different types of exploration and problem-solving dynamics at multiple scales (see also Hristovski et al., 2012; Orth et al., 2017).

Before concluding, we should note again that the context of artistic swimming may limit generalization, although certain collective activities (such as competitive dance performances[4] for instance) may display similar features. The team performance was segmented into discrete collective units of behavioral activity to both structure the individual interviews and facilitate the comparison between the swimmers' pre-reflective experiences of togetherness during the choreography. Compared with other studies in a sports or artistic context, our study of team performance in artistic swimming relied on the interdependent and independent contributions of the swimmers: interdependent in the sense that the swimmers performing the same task had to act and react together, and independent in the sense that their contributions may also have been focused on performing specific subtasks that involved micro-processes of self-other adaptation. Moreover, unlike other sports contexts in which the use of biomechanical or behavioral indicators has been established in training and performance analysis, this is not the case in artistic swimming, and we therefore focused on other parameters. In conclusion, although the research to date has described being and acting together as an essential aspect of team performance, the present contribution is the first to offer a comprehensive analysis based on joint methods. Our study suggests that the feeling of togetherness experienced by a team of swimmers during a choreography is constantly mutating and is both task-specific and task-general. Interestingly, the continuous increases and decreases of sense of togetherness reported by our swimmers can be consistently recognized by expert raters. And indeed, except for eight transitions, the rated togetherness by experts follows closely the same trajectory of the togetherness emerging from the swimmers' ratings. They both fluctuate in a very limited range of togetherness. We hope future research will engage with similar considerations and extend the analysis to other domains in sports and artistic performance.

DATA AVAILABILITY STATEMENT

The anonymized raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

VG designed the study, collected the data, analyzed the data, and wrote and edited the manuscript. DH designed the study and edited the manuscript. AK analyzed the data and edited the

[4]We thank one of our reviewers for suggesting this link.

manuscript. AB edited the manuscript. AS wrote and edited the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

AS acknowledges the support of the Austrian Science Fund (FWF). This research was funded in part by the Austrian Science Fund (FWF), project number: P 32460. For the purpose of open access, the author has applied a CC BY public copyright license to any author accepted manuscript version arising from this submission.

ACKNOWLEDGMENTS

We wish to thank all participants who took part in the study.

REFERENCES

Anguera, M. T., Camerino, O., Castañer, M., Sánchez-Algarra, P., and Onwuegbuzie, A. J. (2017). The specificity of observational studies in physical activity and sports sciences: moving forward in mixed methods research and proposals for achieving quantitative and qualitative symmetry. *Front. Psychol.* 8:2196. doi: 10.3389/fpsyg.2017.02196

Araújo, D., and Davids, K. (2004). Embodied cognition and emergent decision-making in dynamical movement systems. *Junctures J. Themat. Dial.* 2, 45–57.

Araújo, D., Davids, K., and Hristovski, R. (2006). The ecological dynamics of decision making in sport. *Psychol. Sport Exerc.* 7, 653–676. doi: 10.1016/j.psychsport.2006.07.002

Araújo, D., Hristovski, R., Seifert, L., Carvalho, J., and Davids, K. (2017). Ecological cognition: Expert decision-making behaviour in sport. *Int. Rev. Sport Exerc. Psychol.* 12:1349826. doi: 10.1080/1750984X.2017.1349826

Bishop, L. (2018). Collaborative musical creativity: How ensembles coordinate spontaneity. *Front. Psychol.* 9:1285. doi: 10.3389/fpsyg.2018.01285

Bockelman, P., Reinerman-Jones, L., and Gallagher, S. (2013). Methodological lessons in neurophenomenology: Review of a baseline study and recommendations for research approaches. *Front. Hum. Neurosci.* 7:608. doi: 10.3389/fnhum.2013.00608

Bourbousson, J., and Fortes-Bourbousson, M. (2017). Fluctuations of the experience of togetherness within the team over time: task-cohesion and shared understanding throughout a sporting regular season. *Ergonomics* 60, 810–823. doi: 10.1080/00140139.2016.1229041

Bournelli, P., Makri, A., and Mylonas, K. (2009). Motor creativity and self-concept. *Creativ. Res. J.* 21, 104–110. doi: 10.1080/10400410802633657

Brant, R. (1990). Assessing Proportionality in the Proportional Odds Model for Ordinal Logistic Regression. *Biometrics* 46, 1171–1178. doi: 10.2307/2532457

Carron, A. V., Brawley, L. R., and Widmeyer, W. N. (1998). "The measurement of cohesiveness in sport groups," in *Advances in sport and exercise psychology measurement*, ed. J. L. Duda (Ohio: Fitness Information Technology), 213–226.

Carron, A. V., Colman, M. M., Wheeler, J., and Stevens, D. (2002). Cohesion and performance in sport: A meta-analysis. *J. Sport Exerc. Psychol.* 24, 168–188. doi: 10.1123/jsep.24.2.168

Chemero, A. (2009). *Radical embodied cognitive science*. Cambridge, MA: MIT Press. doi: 10.7551/mitpress/8367.001.0001

Chow, J. Y., Davids, K., Hristovski, R., Araújo, D., and Passos, P. (2011). Nonlinear pedagogy: Learning design for self-organizing neurobiological systems. *New Ideas Psychol.* 29, 189–200. doi: 10.1016/j.newideapsych.2010.10.001

Colombetti, G., and Torrance, S. (2009). Emotion and ethics: An inter-(en) active approach. *Phenomenol. Cognit. Sci.* 8:505. doi: 10.1007/s11097-009-9137-3

Davids, K., Button, C., and Bennett, S. J. (2007). *Dynamics of skill acquisition: A constraints led approach*. Champaign: Human Kinetics.

Demos, A. P., Chaffin, R., and Logan, T. (2018). Musicians body sway embodies musical structure and expression: A recurrence-based approach. *Musicae Sci.* 22, 244–263. doi: 10.1177/1029864916685928

Depraz, N., and Desmidt, T. (2019). Cardiophenomenology: a refinement of neurophenomenology. *Phenomenol. Cognit. Sci.* 18, 493–507. doi: 10.1007/s11097-018-9590-y

Depraz, N., Gyemant, T., and Desmidt, S. (2017). A first-person analysis using third-person data as a generative method: A case study of surprise in depression. *Constructiv. Foundat.* 12, 190–203.

Duarte, R., Araújo, D., Correia, V., and Davids, K. (2012). Sports teams as superorganisms. *Sports Med.* 42, 633–642. doi: 10.1007/BF03262285

Froese, T., and Di Paolo, E. A. (2011). The enactive approach: Theoretical sketches from cell to society. *Pragmat. Cognit.* 19, 1–36. doi: 10.1075/pc.19.1.01fro

Froese, T., Gould, C., and Seth, A. K. (2011). Validating and calibrating first-and second-person methods in the science of consciousness. *J. Conscious. Stud.* 18:38.

Froese, T., Iizuka, H., and Ikegami, T. (2014). Embodied social interaction constitutes social cognition in pairs of humans: a minimalist virtual reality experiment. *Sci. Rep.* 4:3672. doi: 10.1038/srep03672

Gesbert, V., and Durny, A. (2017). A Case Study of Forms of Sharing in a Highly Interdependent Soccer Team During Competitive Interactions. *J. Appl. Sport Psychol.* 29, 466–483. doi: 10.1080/10413200.2017.1287787

Gesbert, V., and Hauw, D. (2019). Commentary: Interpersonal Coordination in Soccer: Interpreting Literature to Enhance the Representativeness of Task Design, From Dyads to Teams. *Front. Psychol.* 10:1093. doi: 10.3389/fpsyg.2019.01093

Gesbert, V., and Hauw, D. (2020). When the 8-count is not enough: An analysis of the interaction modalities enacted by artistic swimmers in a collective choreography. *Int. J. Sport Psychol.* 51, 271–295.

Gesbert, V., Durny, A., and Hauw, D. (2017). How do soccer players adjust their activity in team coordination? An enactive phenomenological analysis. *Front. Psychol.* 8:854. doi: 10.3389/fpsyg.2017.00854

Glowinski, D., Bracco, F., Chiorri, C., and Grandjean, D. (2016). Music ensemble as a resilient system. Managing the unexpected through group interaction. *Front. Psychol.* 7:1548. doi: 10.3389/fpsyg.2016.01548

Grammatikopoulos, V., Gregoriadis, A., and Evridiki, Z. (2012). "Acknowledging the role of motor domain in creativity in early childhood education," in *Contemporary perspectives on research in creativity in early childhood education*, ed. O. Saracho (Charlotte, NC: Information age publishing), 161–178.

Hauw, D. (2009). Reflective practice in the heart of training and competition: the course of experience analysis for enhancing elite acrobatics athletes' performances. *Reflect. Pract.* 10, 341–352. doi: 10.1080/14623940903034671

Hauw, D. (2018). Énaction et intervention en psychologie du sport chez les sportifs élites et en formation. *Canad. J. Behav. Sci.* 50:54. doi: 10.1037/cbs0000094

Hauw, D., and Durand, M. (2008). Temporal dynamics of acrobatic activity: An approach of elite athletes specious present. *J. Sports Sci. Med.* 7:8.

Hauw, D., Rochat, N., Gesbert, V., Astolfi, T., Philippe, R. A., and Mariani, B. (2017). Putting together first-and third-person approaches for sport activity analysis: The case of ultra-trail runners' performance analysis. *Adv. Hum. Factors Sports Outdoor Recreat.* 2017, 49–58. doi: 10.1007/978-3-319-41953-4_5

He, J., and Ravn, S. (2018). Sharing the dance–on the reciprocity of movement in the case of elite sports dancers. *Phenomenol. Cognit. Sci.* 17, 99–116. doi: 10.1007/s11097-016-9496-5

Heuzé, J. P., Raimbault, N., and Fontayne, P. (2006). Relationships between cohesion, collective efficacy, and performance in professional basketball teams: An examination of mediating effects. *J. Sports Sci.* 24, 59–68. doi: 10.1080/02640410500127736

Himberg, T., Laroche, J., Bigé, R., Buchkowski, M., and Bachrach, A. (2018). Coordinated interpersonal behaviour in collective dance improvisation: the aesthetics of kinaesthetic togetherness. *Behav. Sci.* 8:23. doi: 10.3390/bs8020023

Høffding, S. (2019). *A Phenomenology of Musical Absorption*. London: Palgrave Macmillan. doi: 10.1007/978-3-030-00659-4

Høffding, S., and Satne, G. (2019). Interactive expertise in solo and joint musical performance. *Synthese* 2019, 1–19. doi: 10.1007/s11229-019-02339-x

Hristovski, R., Davids, K., Passos, P., and Araújo, D. (2012). Sport performance as a domain of creative problem solving for self-organizing performer-environment systems. *Open Sport. Sci. J.* 5, 26–35. doi: 10.2174/1875399X01205010026

Keller, P. E., Novembre, G., and Hove, M. J. (2014). Rhythm in joint action: psychological and neurophysiological mechanisms for real-time interpersonal coordination. *Philosop. Transact. R. Soc. B Biol. Sci.* 369:0394. doi: 10.1098/rstb.2013.0394

Kelso, J. A. S. (2001). Metastable coordination dynamics of brain and behavior. *Brain Neural Netw.* 8, 125–130. doi: 10.3902/jnns.8.125

Kelso, J. A. S. (2003). "Cognitive coordination dynamics," in *The dynamical systems approach to cognition: Concepts and empirical paradigms based on self-organization, embodiment, and coordination dynamics*, eds W. Tschacher and J.-P. Dauwalder (Singapore: World Scientific Publishing), 45–67. doi: 10.1142/9789812564399_0003

Kimmel, M. (2017). "The complexity of skillscapes: Skill sets, synergies, and meta-regulation in joint embodied improvisation," in *Proceedings of the 13th International Conference on Naturalistic Decision Making*, eds J. Gore and P. Ward (Bath: University of Bath), 102–109.

Kimmel, M. (2019). "A cognitive theory of joint improvisation: The case of tango argentino," in *The Oxford Handbook of Improvisation in Dance*, ed. V. L. Midgelow (Oxford: Oxford University Press), 562–592. doi: 10.1093/oxfordhb/9780199396986.013.32

Kimmel, M., and Rogler, C. R. (2018). Affordances in interaction: the case of aikido. *Ecol. Psychol.* 30, 195–223. doi: 10.1080/10407413.2017.1409589

Kimmel, M., and Rogler, C. R. (2019). The anatomy of antagonistic coregulation: Emergent coordination, path dependency, and the interplay of biomechanic parameters in Aikido. *Hum. Mov. Sci.* 63, 231–253. doi: 10.1016/j.humov.2018.08.008

Kimmel, M., Hristova, D., and Kussmaul, K. (2018). Sources of embodied creativity: interactivity and ideation in contact improvisation. *Behav. Sci.* 8:52. doi: 10.3390/bs8060052

Koo, T. K., and Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *J. Chiropractic Med.* 15, 155–163. doi: 10.1016/j.jcm.2016.02.012

Kurby, C. A., and Zacks, J. M. (2008). Segmentation in the perception and memory of events. *Trends Cognit. Sci.* 12, 72–79. doi: 10.1016/j.tics.2007.11.004

Lakens, D. (2010). Movement synchrony and perceived entitativity. *J. Exp. Soc. Psychol.* 46, 701–708. doi: 10.1016/j.jesp.2010.03.015

Lakens, D., and Stel, M. (2011). If they move in sync, they must feel in sync: movement synchrony leads to attributions of rapport and entitativity. *Soc. Cogn.* 29, 1–14. doi: 10.1521/soco.2011.29.1.1

Laroche, J., Berardi, A. M., and Brangier, E. (2014). Embodiment of intersubjective time: relational dynamics as attractors in the temporal coordination of interpersonal behaviors and experiences. *Front. Psychol.* 5:1180. doi: 10.3389/fpsyg.2014.01180

Legrand, D. (2006). The bodily self: The sensori-motor roots of pre-reflective self-consciousness. *Phenomenol. Cognit. Sci.* 5, 89–118. doi: 10.1007/s11097-005-9015-6

Lund, O., Ravn, S., and Christensen, M. K. (2014). Jumping together: apprenticeship learning among elite trampoline athletes. *Phys. Educ. Sport Pedagogy* 19, 383–397. doi: 10.1080/17408989.2013.769508

Lutz, A. (2002). Toward a neurophenomenology as an account of generative passages: A first empirical case study. *Phenomenol. Cognit. Sci.* 1, 133–167. doi: 10.1023/A:1020320221083

Lutz, A., Lachaux, J. P., Martinerie, J., and Varela, F. J. (2002). Guiding the study of brain dynamics by using first-person data: synchrony patterns correlate with ongoing conscious states during a simple visual task. *Proc. Natl. Acad. Sci.* 99, 1586–1591. doi: 10.1073/pnas.032658199

Miyata, K., Varlet, M., Miura, A., Kudo, K., and Keller, P. E. (2017). Modulation of individual auditory-motor coordination dynamics through interpersonal visual coupling. *Sci. Rep.* 7, 1–11. doi: 10.1038/s41598-017-16151-5

Mohamed, S., Favrod, V., Philippe, R. A., and Hauw, D. (2015). The situated management of safety during risky sport: learning from skydivers' courses of experience. *J. Sports Sci. Med.* 14:340.

Olivares, F. A., Vargas, E., Fuentes, C., Martínez-Pernía, D., and Canales-Johnson, A. (2015). Neurophenomenology revisited: second-person methods for the study of human consciousness. *Front. Psychol.* 6:673. doi: 10.3389/fpsyg.2015.00673

Orth, D., van der Kamp, J., Memmert, D., and Savelsbergh, G. J. (2017). Creative motor actions as emerging from movement variability. *Front. Psychol.* 8:1903. doi: 10.3389/fpsyg.2017.01903

Petitmengin, C. (2006). Describing one's subjective experience in the second person: An interview method for the science of consciousness. *Phenomenol. Cognit. Sci.* 5, 229–269. doi: 10.1007/s11097-006-9022-2

Petitmengin, C., and Lachaux, J. P. (2013). Microcognitive science: bridging experiential and neuronal microdynamics. *Front. Hum. Neurosci.* 7:617. doi: 10.3389/fnhum.2013.00617

Poizat, G., Bourbousson, J., Saury, J., and Sève, C. (2012). Understanding team coordination in doubles table tennis: Joint analysis of first-and third-person data. *Psychol. Sport Exerc.* 13, 630–639. doi: 10.1016/j.psychsport.2012.03.008

R Core Team (2020). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

R'Kiouak, M., Saury, J., Durand, M., and Bourbousson, J. (2016). Joint action of a pair of rowers in a race: shared experiences of effectiveness are shaped by interpersonal mechanical states. *Front. Psychol.* 7:720. doi: 10.3389/fpsyg.2016.00720

Rabinowitch, T.-C., and Knafo-Noam, A. (2015). Synchronous rhythmic interaction enhances children's perceived similarity and closeness towards each other. *PLoS One* 10:e0120878. doi: 10.1371/journal.pone.0120878

Revelle, W. (2021). *psych: Procedures for Personality and Psychological Research. Software*. Vienna: R Core Team.

Rochat, N., Gesbert, V., Seifert, L., and Hauw, D. (2018). Enacting phenomenological gestalts in ultra-trail running: an inductive analysis of trail runners' courses of experience. *Front. Psychol.* 9:2038. doi: 10.3389/fpsyg.2018.02038

Rochat, N., Hacques, G., Ganière, C., Seifert, L., Hauw, D., Iodice, P., et al. (2020). Dynamics of Experience in a Learning Protocol: A Case Study in Climbing. *Front. Psychol.* 11:249. doi: 10.3389/fpsyg.2020.00249

Rochat, N., Seifert, L., Guignard, B., and Hauw, D. (2019). An enactive approach to appropriation in the instrumented activity of trail running. *Cognit. Process.* 20, 459–477. doi: 10.1007/s10339-019-00921-2

Runco, M., and Jager, G. (2012). The standard definition of creativity. *Creativ. Res. J.* 24, 92–96. doi: 10.1080/10400419.2012.650092

Santos, S. D., Memmert, D., Sampaio, J., and Leite, N. (2016). The Spawns of Creative Behavior in Team Sports: A Creativity Developmental Framework. *Front. Psychol.* 7:1282. doi: 10.3389/fpsyg.2016.01282

Santos, S., and Monteiro, D. (2021). Uncovering the Role of Motor Performance and Creative Thinking on Sports Creativity in Primary School-aged Children. *Creativ. Res. J.* 33:1843125. doi: 10.1080/10400419.2020.1843125

Santos, S., Jiménez, S., Sampaio, J., and Leite, N. (2017). Effects of the Skills4Genius sports-based training program in creative behavior. *PLoS One* 12:e0172520. doi: 10.1371/journal.pone.0172520

Schad, D. J., Vasishth, S., Hohenstein, S., and Kliegl, R. (2020). How to capitalize on a priori contrasts in linear (mixed) models: A tutorial. *J. Mem. Lang.* 110:104038. doi: 10.1016/j.jml.2019.104038

Schiavio, A., and Benedek, M. (2020). Dimensions of musical creativity. *Front. Neurosci.* 14:578932. doi: 10.3389/fnins.2020.578932

Schiavio, A., and Høffding, S. (2015). Playing together without communicating? A pre-reflective and enactive account of joint musical performance. *Musica Sci.* 19, 366–388. doi: 10.1177/1029864915593333

Schiavio, A., and Kimmel, M. (2021). "The ecological dynamics of musical creativity and skill acquisition," in *Meaningful Relations: The Enactivist Making of Experiential Worlds*, ed. A. Scarinzi (Sankt Augustin: Academia-Verlag), 123–158. doi: 10.5771/9783896659934-121

Schiavio, A., Gesbert, V., Reybrouck, M., Hauw, D., and Parncutt, R. (2019). Optimizing Performative Skills in Social Interaction: Insights from Embodied Cognition, Music Education, and Sport Psychology. *Front. Psychol.* 10:01542. doi: 10.3389/fpsyg.2019.01542

Schiavio, A., Maes, P.-J., and van der Schyff, D. (2021). The dynamics of musical participation. *Musicae Sci.* [Preprint]. doi: 10.1177/1029864920988319

Schiavio, A., Moran, N., van der Schyff, D., Biasutti, M., and Parncutt, R. (2020). Processes and experiences of creative cognition in seven Western classical composers. *Musicae Sci.* 2020:1029864920943931. doi: 10.1177/1029864920943931

Schiavio, A., Ryan, K., Moran, N., van der Schyff, D., and Gallagher, S. (in press). By myself but not alone. Agency, creativity, and extended musical historicity. *J. R. Musical Associat.*

Schlegel, B., and Steenbergen, M. (2020). *Brant: Test for Parallel Regression Assumption. R package version 0.3-0*. Vienna: R Core Team.

Seifert, L., Adé, D., Saury, J., Bourbousson, J., and Thouvarecq, R. (2016). "Mix of phenomenological and behavioural data to explore interpersonal coordination in outdoor activities: examples in rowing and orienteering," in *Interpersonal Coordination and Performance in Social Systems*, eds P. Passos, K. Davids, and J. Y. Chow (London: Routledge), 109–125.

Seifert, L., Button, C., and Davids, K. (2013). Key properties of expert movement systems in sport. *Sports Med.* 43, 167–178. doi: 10.1007/s40279-012-0011-z

Seifert, L., Lardy, J., Bourbousson, J., Adé, D., Nordez, A., Thouvarecq, R., et al. (2017). Interpersonal coordination and individual organization combined with shared phenomenological experience in rowing performance: two case studies. *Front. Psychol.* 8:75. doi: 10.3389/fpsyg.2017.00075

Sève, C., Nordez, A., Poizat, G., and Saury, J. (2013). Performance analysis in sport: Contributions from a joint analysis of athletes' experience and biomechanical indicators. *Scand. J. Med. Sci. Sports* 23, 576–584. doi: 10.1111/j.1600-0838.2011.01421.x

Spink, K. S. (1990). Group cohesion and collective efficacy of volleyball teams. *J. Sport Exerc. Psychol.* 12, 301–311. doi: 10.1123/jsep.12.3.301

Tanaka, S. (2017). Intercorporeality and aida: developing an interaction theory of social cognition. *Theory Psychol.* 27, 337–353. doi: 10.1177/0959354317702543

Theureau, J. (2015). *Le cours d'action : Enaction & Experience*. Toulouse: Octarès.

Tognoli, E., Zhang, M., Fuchs, A., Beetle, C., and Kelso, J. A. S. (2020). Coordination Dynamics: A Foundation for Understanding Social Behavior. *Front. Hum. Neurosci.* 14:317. doi: 10.3389/fnhum.2020.00317

Toner, J., and Montero, B. (2020). The value of aesthetic judgements in athletic performance. *J. Somaesthet.* 6, 112–126.

Tunçgenç, B., and Cohen, E. (2016). Movement Synchrony Forges Social Bonds across Group Divides. *Front. Psychol.* 7:782. doi: 10.3389/fpsyg.2016.00782

van der Schyff, D., Schiavio, A., Walton, A., Velardo, V., and Chemero, T. (2018). Musical creativity and the embodied mind. Exploring the possibilities of 4E cognition and dynamical systems theory. *Music Sci.* 2018:2059204318792319. doi: 10.1177/2059204318792319

Varela, F. J. (1996). Neurophenomenology: A methodological remedy for the hard problem. *J. Conscious. Stud.* 3, 330–349.

Varela, F. J., and Shear, J. (1999). First-person methodologies: What, why, how. *J. Conscious. Stud.* 6, 1–14.

Venables, W. N., and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Berlin: Springer. doi: 10.1007/978-0-387-21706-2

Vermersch, P. (2009). Describing the practice of introspection. *J. Conscious. Stud.* 16, 20–57.

Vicary, S., Sperling, M., Von Zimmermann, J., Richardson, D. C., and Orgs, G. (2017). Joint action aesthetics. *PLoS One* 12:e0180101. doi: 10.1371/journal.pone.0180101

Vors, O., Cury, F., Marqueste, T., and Mascret, N. (2019). Enactive Phenomenological Approach to the Trier Social Stress Test: a mixed methods point of view. *JoVE* 2019:e58805. doi: 10.3791/58805

Walton, A. E., Richardson, M. J., Langland-Hassan, P., and Chemero, A. (2015). Improvisation and the self-organization of multiple musical bodies. *Front. Psychol.* 6:313. doi: 10.3389/fpsyg.2015.00313

Zacks, J. M., and Swallow, K. M. (2007). Event segmentation. *Curr. Direct. Psychol. Sci.* 16, 80–84. doi: 10.1111/j.1467-8721.2007.00480.x

Conflict of Interest: VG was employed by Football Club Lorient.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers | Frontiers in Psychology

OPINION
published: 14 April 2022
doi: 10.3389/fpsyg.2022.747852

Neuroassessment in Sports: An Integrative Approach for Performance and Potential Evaluation in Athletes

Davide Crivelli[1,2]* *and Michela Balconi*[1,2]

[1] International Research Center for Cognitive Applied Neuroscience (IrcCAN), Catholic University of the Sacred Heart, Milan, Italy, [2] Research Unit in Affective and Social Neuroscience, Department of Psychology, Catholic University of the Sacred Heart, Milan, Italy

Keywords: neuroassessment, sport, neurocognitive efficiency, attention regulation, focusing, self-regulation, neurocognitive fitness, peak performance

OPEN ACCESS

Edited by:
Roberta Antonini Philippe, University of Lausanne, Switzerland

Reviewed by:
Scott Sinnett, University of Hawaii at Manoa, United States
Jonathan Douglas Connor, James Cook University, Australia
Andrew Strong, Umeå University, Sweden

***Correspondence:**
Davide Crivelli
davide.crivelli@unicatt.it

Specialty section:
This article was submitted to Performance Science, a section of the journal Frontiers in Psychology

Received: *26 July 2021*
Accepted: *28 March 2022*
Published: *14 April 2022*

Citation:
Crivelli D and Balconi M (2022) Neuroassessment in Sports: An Integrative Approach for Performance and Potential Evaluation in Athletes. Front. Psychol. 13:747852. doi: 10.3389/fpsyg.2022.747852

PERFORMANCE DETERMINANTS: PHYSICAL, PSYCHOLOGICAL, AND NEUROCOGNITIVE FITNESS

Peak performance—as the traditional Olympic motto "*Citius, Altius, Fortius*" ("*Faster, Higher, Stronger*") points out—is a primary goal for athletes, but it also can be deemed—as posited by Pierre de Coubertin, promoter of the modern Olympic Games—as one of the primary personal development goals which drives peoples every day and professional lives. This makes research in sports a critical window into the advancement of our understanding on how to foster, evaluate, and empower human performance.

The profiling of athletes' psychophysical traits is, in particular, a crucial step in planning, implementing, and monitoring their training schedule and, therefore, receives particular attention by the athletes themselves, their coaches, and sports clubs. Performance data, if generated by appropriately designed assessments, may provide relevant and reliable information to evaluate the progress of athletes over time and/or rank them against their peers (Robertson et al., 2014). Further, such data also help to profile the athlete's strengths/weaknesses, to monitor progress, to adapt training protocols to the athlete's needs, and to identify talent or to predict the athlete's potential (e.g., Morris-Binelli et al., 2021). Yet, the topics of assessment and, specifically, of its implementation in practice are still quite controversial. Indeed, despite the relationship between athletic skill efficiency and success in sports, literature reviewing psychometric properties of assessment methods (e.g., validity, reliability, and predictive value) is still scant (Currell and Jeukendrup, 2008; Robertson et al., 2014; Piggott et al., 2020). Furthermore, as underlined by Robertson et al. (2014), outdated methodology and undefined measurement properties actually limit the usefulness of many performance and skill measures. This is particularly true for their actual implementation in routine professional practice - even though they might be widely used in research - with remarkable implications for assessing the effect of coaching, fatigue, focus of attention, and pre-skill execution routine on participant performance (Mccann et al., 2001; Russell and Kingsley, 2011; Russell et al., 2011; Mckay and Wulf, 2012).

Again, despite the fact that the contribution of psychological factors to individual sport-related experience and that the role of psychological/cognitive load in modulating performance outcomes are widely recognized (Mellalieu et al., 2021), the assessment procedures to evaluate athletes' current performance and their potential often move those factors to the background. The vast majority of empirical work, indeed, focuses primarily on the physical determinants of performance and basic physiological measures. The paucity of suitable tools and standardized procedures for the assessment of psychological and cognitive correlates/determinants of sport performance (e.g., self-regulation skills, set-shifting efficiency, and cognitive control) and of specific training of technical staff in their use has likely contributed to the definition of a theoretical and methodological framework almost uniquely centered on physical fitness.

Namely, *physical fitness* (PhyF) is a well-established construct, whose core components include muscular strength, endurance, and range of motion/joint flexibility (Jeffreys and Moody, 2016). Transdisciplinary as well as sport-specific metrics exist for each of those core elements of PhyF. Similarly, specific training protocols have been developed to enhance strength, endurance and/or flexibility (e.g., conditioning training, high-intensity interval training, and stretching routines; Jeffreys and Moody, 2016). Nonetheless, the interdependence of physical, affective, and cognitive factors in contributing to athletic performance is clearly pointed out by a solid evidence base (Piggott et al., 2020). Across the years, more and more athletes and sport professionals have begun to look at mental factors less as corollary elements for performance assessment (Habay et al., 2021; Mellalieu et al., 2021), and mental training practices are becoming less exclusive.

The construct of *psychological fitness* (PsyF; Heaps, 1978) was then introduced even in sport science and practice to identify various psychological traits—including character strengths, personality traits, personal resources, and resilience against burnout or mental disorders (Heaps, 1978; Cornum et al., 2011; Wesemann et al., 2018)—that promote wellbeing, foster personal development, and sustain performance. Cognitive flexibility and executive functions are, as an example, known as a broad-spectrum protective factor in preventing mental illness (Crivelli and Balconi, 2021) as well as in sustaining personal development (Keyes, 2007; Robinson et al., 2015). Again, empowerment protocols focusing on PsyF have been devised and applied for performance improvement and enhancement of training outcomes even in the army and in sports (Birrer and Morgan, 2010; Boga, 2017). Individual and team ability to manage the psychological load in complex challenging situations is widely deemed as conferring critical performance advantages (Aidman, 2020; Mellalieu et al., 2021).

Yet, we suggest that at least a third level should be considered in defining a comprehensive account for assessment and monitoring of athletes' performance and potential, which could be defined *neurocognitive fitness* (NCF). The construct of NCF builds on the cognitive fitness framework (Aidman, 2020), a working hypothesis intended to integrate available evidences on the contribution of cognitive skills training to performance enhancement, while extending that seminal concept by using a psychophysiological and neurofunctional perspective. Therefore, NCF could be understood as the degree of efficacy and efficiency in exploiting available neural and cognitive resources to complete a task based on the challenge level that the task itself and the environment impose, in order to exert an optimal level of performance. Core components of NCF, as we define it, include: self-awareness, as a primary precursor of intentional goal-directed behavior; self-regulation, as the background for stress management and self-control; and executive control, as the intersection of attention regulation, cognitive flexibility, cognitive workload management, and inhibition skills. As underlined by Nakata et al. (2010), to reach optimal performance and perform skilled movements in real-life sport situations, an athlete needs to be able to flexibly and efficiently adapt movements constituting the athletic gesture based on the perception of environmental information, discrimination of relevant stimuli, rapid decision-making processes, integration of afferent signals, and anticipatory action preparation. Set-shifting, behavioral inhibition, focusing and attention regulation mechanisms are all the same crucial to reach and maintain high-performance levels.

Notably, neurocognitive empowerment protocols based on the combined use of cognitive techniques (e.g., focused attention and self-awareness practices) and neuromodulation devices (e.g., wearable neurofeedback) have already been shown to promote the efficiency of attention regulation and executive control, besides greater stress management skills, even in applied contexts, such as at the workplace and in sports (Balconi et al., 2019a,b; Crivelli et al., 2019a,b). Such protocols, by combining neurotechnologies able to promote greater control on neural resources and cognitive-behavioral practices supporting self-enhancement, represent a valuable example of training programmes implementing NCF in real-life contexts. Such effective integration between psychological, cognitive and neurofunctional levels is likely not that mature in the application field concerning performance profiling and assessment, with the majority of solutions implemented in practice still mainly based on behavioral, psychological, and (limitedly) neuropsychological testing.

FROM ASSESSMENT TO NEUROASSESSMENT

While the investigation of behavioral, psychological and cognitive determinants of personal performance in profiling an athlete's current strengths/weaknesses and potential should be deemed, as above noted, highly relevant, we here intend to point out the potential of an actual integration of those measures with objective neurofunctional and autonomic markers, according to a neuro-behavioral perspective and capitalizing on embodied and perception-action theoretical frameworks (Aglioti et al., 2008; Müller and Abernethy, 2012; Wright et al., 2013). The human brain indeed shows, together with our bodies, the remarkable ability to learn and capitalize from experience to reach mastery and excellence in specialized skills, and that is particularly important when focusing on the investigation of

peak performance and highly specialized activities, such as in sport practice. Namely, by accurately qualifying and quantifying the cascade of physiological events that mark cognitive-motor learning, attention focusing, and self-regulation, as well as planning, programming, and executing athletic gestures, and by linking them to behavior and psychological processes, it is possible to infer how our brain functions during challenging tasks and contexts. Such an evidence base helps identifying factors promoting optimal performance for the examined athlete (Perrey, 2008; Thompson et al., 2008; Piggott et al., 2020).

A vast literature coming from basic neuroscience research already provides a remarkable evidence base for understanding how athletes' brains support them during sport and exercise activity and also when, where and to what extent an athlete brain is different from the one of a common person (Nakata et al., 2010; Li and Smith, 2021). Integrating psychometric, behavioral, and observational assessment tools with neuroscientific devices able to capture covert physiological markers of the efficiency of processes supporting executive control and attention regulation also represents a valuable advantage in the quest for novel, objective, and actually predictive tools for assessment of performance in sport contexts, as is beginning to happen even in other applied contexts (Balconi et al., 2020).

By facilitating insights into sensory, motor, and cognitive processes contributing to the preparation, execution and imagination of the athletic gesture, as well as to the orientation of attention resources on competition, both wearable and lab-based neuroscientific devices (among which, in particular, electroencephalography - EEG, autonomic indices, and—more recently—functional Near Infrared Spectroscopy—fNIRS) may also help in shedding light on specific aspects of sport activity and its neurofunctional characteristics (Crivelli and Balconi, 2017; Balconi and Crivelli, 2019)—such as, for instance, modulations of cognitive effort in pre-competition phases, training-induced enhancement of stimuli detection, resistance to mental/physical fatigue and its relation to performance. Integrating such devices in established assessment procedures could foster the adoption of a multifaceted analytical approach that properly include the investigation of neurocognitive fitness, together with psychological and physical ones. It has to be noted that the topic of feasibility of neurofunctional-psychophysiological assessment in ecological settings is actually still a hot topic in sport and exercise science. Yet methodological debate on such a topic has highlighted interesting developments during recent years and recommendations concerning subject preparation, sensor placement, environmental controls, use of portable recording devices, and signal processing have been proposed to improve quality and informativity of physiological data captured even outside of the lab and during actual exercise/sport activity (for reviews, see Perrey, 2008; Thompson et al., 2008; Park et al., 2015; Cheron et al., 2016; Balconi and Crivelli, 2019).

We therefore propose that the time has come for a more systematic implementation of the perspective change from assessment to *neuroassessment*. Neuroassessment can be defined as a standardized procedure to qualify and quantify the level of physical, psychological and neurocognitive fitness of an athlete via a combination of self-report, behavioral, autonomic, and neurofunctional metrics. Those metrics should be able to capture different and complementary facets of performance during standardized as well as ecologically-valid sport-specific tasks. Notably, the systematic progress in bioengineering and medical technologies allows, to date, to bring neuroscientific tools for biometric measurements even on the field, thus making it possible to collect valuable data during field based sessions of physical and mental training.

Also, such rich sets of data would not only result in a gain in insight for the purposes of athletes profiling and planned development, but will also help building stronger and more complete theoretical frameworks to classify athletes self-awareness, self-regulation, and higher cognition (e.g., attention regulation, inhibitory mechanisms, and information-processing and focusing) skills and their physiological signatures, in keeping with neuroscientific models of such skills.

PROFILING ATTENTION REGULATION SKILLS: AN APPLIED EXAMPLE

By taking attention regulation as an example, we will now briefly introduce a neuroassessment protocol devised to improve evaluation of attention focusing in sports.

The ability to focus attention on target stimuli relevant to the sport context and specific discipline while inhibiting irrelevant information notwithstanding their perceptual or affective salience is considered a key aspect of optimal performance and a valuable transdisciplinary trait in athletes (Memmert, 2009). Yet, notwithstanding the acknowledged relevance of those abilities, profiling athletes on their primary attention regulation and control skills is still commonly based on psychometric testing drawing from the classical theory of attentional styles by Nideffer (1976). Also, psychometric testing, as well as purely behavioral measures, presents a few limitations when used by themselves to try "opening the black box" since, for example, they are susceptible to desirability biases and limited self-reflection abilities. Moreover, they could hardly parse out the contribution of strategic top-down vs. instinctual bottom-up processes to attentional performance. Such fine-grained analysis could be more easily obtained by using neurofunctional techniques.

The proposed neuroassessment protocol, in particular, integrates: self-report evaluation of the ability to regulate attention in sport/competitive contexts (self-awareness component), computerized and neuropsychological testing of the efficiency of attention regulation (behavioral performance component), as well as autonomic (EDA, HR, HRV) and electrophysiological markers of neurocognitive efficiency (task-related alpha-beta modulations as markers of cognitive workload (Janelle and Hatfield, 2008; Thompson et al., 2008), N2 and P3 event-related potentials as markers of attention regulation and stimuli detection (Nakata et al., 2010; Balconi and Crivelli, 2019), Event-Related Negativity as a marker of

monitoring processes (Themanson et al., 2008; Masaki et al., 2017) during challenging ecological tasks implementing salient stimuli evoking sport-specific contexts (psychophysiological and neurofunctional components). Namely, one of the tasks we have developed and tested is an adapted, unpublished, version of a cueing task with multiple target positions and ecological stimuli specific to different sport disciplines (e.g., in fighting sports, fist and kick strikes as target stimuli and guard of an opponent fighter as endogenous cue). Given its characteristics, this exemplifying protocol can be used to profile athletes and inform the design of personalized empowerment programs.

Preliminary validation data from the above-mentioned project highlighted internally consistent profiles across the multi-dimensional metrics of attention regulation and executive control performance, hinting at the potential of the protocol for in-depth assessment of athletes' main characteristics and at the complementarity of chosen performance measures.

CONCLUSION

The conceptualization of a multifactorial construct, as suggested by Atkinson (2002), facilitates the definition of its measurable components. Here we have introduced a multifaceted reference model for the definition of performance in sports by pairing the more established constructs of PhyF and PsyF with the construct of NCF. Such a model might provide the framework for a leaner perspective change from traditional observational or physical assessment procedures to neuroassessment, which we identify with the actual integration of both subjective (self-report, observational) and objective (behavioral, physiological) measures to sketch the profile of athletes' neurocognitive efficiency.

AUTHOR CONTRIBUTIONS

DC and MB contributed to conception of the present work. DC wrote the first draft of the manuscript and MB revised it. All authors have read and approved the submitted version.

REFERENCES

Aglioti, S. M., Cesari, P., Romani, M., and Urgesi, C. (2008). Action anticipation and motor resonance in elite basketball players. *Nat. Neurosci.* 11, 1109–1116. doi: 10.1038/nn.2182

Aidman, E. (2020). Cognitive fitness framework: towards assessing, training and augmenting individual-difference factors underpinning high-performance cognition. *Front. Hum. Neurosci.* 13, 1–9. doi: 10.3389/fnhum.2019.00466

Atkinson, G. (2002). Sport performance: variable or construct? *J. Sports Sci.* 20, 291–292. doi: 10.1080/026404102753576053

Balconi, M., Angioletti, L., and Crivelli, D. (2020). Neuro-empowerment of executive functions in the workplace: the reason why. *Front. Psychol.* 11:1519. doi: 10.3389/fpsyg.2020.01519

Balconi, M., and Crivelli, D. (2019). "Fundamentals of electroencephalography and optical imaging for sport and exercise science. From the laboratory to on-the-playing-field acquired evidence," in *Handbook of Sport Neuroscience and Psychophysiology*, eds R. A. Carlstedt and M. Balconi (New York, NY: Routledge), 40–69. doi: 10.4324/9781315723693-3

Balconi, M., Crivelli, D., and Angioletti, L. (2019a). Efficacy of a neurofeedback training on attention and driving performance: physiological and behavioral measures. *Front. Neurosci.* 13:996. doi: 10.3389/fnins.2019.00996

Balconi, M., Crivelli, D., Fronda, G., and Venturella, I. (2019b). "Neuro-rehabilitation and neuro-empowerment by wearable devices. Applications to well-being and stress management," in *Converging Clinical and Engineering Research on Neurorehabilitation III Biosystems & Biorobotics*, eds L. Masia, S. Micera, M. Akay, and J. L. Pons (Cham: Springer International Publishing), 963–966. doi: 10.1007/978-3-030-01845-0_193

Birrer, D., and Morgan, G. (2010). Psychological skills training as a way to enhance an athlete's performance in high-intensity sports. *Scand. J. Med. Sci. Sports* 20, 78–87. doi: 10.1111/j.1600-0838.2010.01188.x

Boga, D. (2017). "Training mental fitness," in *Human Dimension*, ed Australian Army Headquarters (Puckapunyal: Centre for Army Lessons, Army Knowledge Group), 18–27.

Cheron, G., Petit, G., Cheron, J., Leroy, A., Cebolla, A., Cevallos, C., et al. (2016). Brain oscillations in sport: toward EEG biomarkers of performance. *Front. Psychol.* 7:246. doi: 10.3389/fpsyg.2016.00246

Cornum, R., Matthews, M. D., and Seligman, M. E. P. (2011). Comprehensive soldier fitness: building resilience in a challenging institutional context. *Am. Psychol.* 66, 4–9. doi: 10.1037/a0021420

Crivelli, D., and Balconi, M. (2017). Event-related electromagnetic responses. *Ref. Modul. Neurosci. Biobehav. Psychol.* 4, 1–27. doi: 10.1016/B978-0-12-809324-5.03053-4

Crivelli, D., and Balconi, M. (2021). "Psychopathology of EFs," in *Advances in Substance and Behavioral Addiction - The Role of Executive Functions*, eds M. Balconi and S. Campanella (Cham: Springer), 2. doi: 10.1007/978-3-030-82408-2_2

Crivelli, D., Fronda, G., and Balconi, M. (2019a). Neurocognitive enhancement effects of combined mindfulness–neurofeedback training in sport. *Neuroscience* 412, 83–93. doi: 10.1016/j.neuroscience.2019.05.066

Crivelli, D., Fronda, G., Venturella, I., and Balconi, M. (2019b). Stress and neurocognitive efficiency in managerial contexts: a study on technology-mediatedmindfulness practice. *Int. J. Work. Heal. Manag.* 12, 42–56. doi: 10.1108/IJWHM-07-2018-0095

Currell, K., and Jeukendrup, A. (2008). Validity, reliability and sensitivity of measures of sporting performance. *Sport. Med.* 38, 297–316. doi: 10.2165/00007256-200838040-00003

Habay, J., Van Cutsem, J., Verschueren, J., De Bock, S., Proost, M., De Wachter, J., et al. (2021). Mental fatigue and sport-specific psychomotor performance: a systematic review. *Sport. Med.* 51, 1527–1548. doi: 10.1007/s40279-021-01429-6

Heaps, R. A. (1978). Relating physical and psychological fitness: a psychological point of view. *J. Sports Med. Phys. Fitness* 18, 399–408.

Janelle, C. M., and Hatfield, B. D. (2008). Visual attention and brain processes that underlie expert performance: implications for sport and military psychology. *Mil. Psychol.* 20, 39–69. doi: 10.1080/08995600701804798

Jeffreys, I., and Moody, J. (2016). *Strength and Conditioning for Sports Performance*. London: Routledge. doi: 10.4324/9780203852286

Keyes, C. L. M. (2007). Promoting and protecting mental health as flourishing: a complementary strategy for improving national mental health. *Am. Psychol.* 62, 95–108. doi: 10.1037/0003-066X.62.2.95

Li, L., and Smith, D. M. (2021). Neural efficiency in athletes: a systematic review. *Front. Behav. Neurosci.* 15:698555. doi: 10.3389/fnbeh.2021.698555

Masaki, H., Maruo, Y., Meyer, A., and Hajcak, G. (2017). Neural correlates of choking under pressure: athletes high in sports anxiety monitor errors more when performance is being evaluated. *Dev. Neuropsychol.* 42, 104–112. doi: 10.1080/87565641.2016.1274314

Mccann, P., Lavallee, D., and Lavallee, R. (2001). The effect of pre-shot routines on golf wedge shot performance. *Eur. J. Sport Sci.* 1, 1–10. doi: 10.1080/17461390100071503

Mckay, B., and Wulf, G. (2012). A distal external focus enhances novice dart throwing performance. *Int. J. Sport Exerc. Psychol.* 10, 149–156. doi: 10.1080/1612197X.2012.682356

Mellalieu, S., Jones, C., Wagstaff, C., Kemp, S., and Cross, M. J. (2021). Measuring psychological load in sport. *Int. J. Sports Med.* 42, 782–788. doi: 10.1055/a-1446-9642

Memmert, D. (2009). Pay attention! A review of visual attentional expertise in sport. *Int. Rev. Sport Exerc. Psychol.* 2, 119–138. doi: 10.1080/17509840802641372

Morris-Binelli, K., Müller, S., van Rens, F. E. C. A., Harbaugh, A. G., and Rosalie, S. M. (2021). Individual differences in performance and learning of visual anticipation in expert field hockey goalkeepers. *Psychol. Sport Exerc.* 52:101829. doi: 10.1016/j.psychsport.2020.101829

Müller, S., and Abernethy, B. (2012). Expert anticipatory skill in striking sports. *Res. Q. Exerc. Sport* 83, 175–187. doi: 10.1080/02701367.2012.10599848

Nakata, H., Yoshie, M., Miura, A., and Kudo, K. (2010). Characteristics of the athletes' brain: evidence from neurophysiology and neuroimaging. *Brain Res. Rev.* 62, 197–211. doi: 10.1016/j.brainresrev.2009.11.006

Nideffer, R. M. (1976). Test of attentional and interpersonal style. *J. Pers. Soc. Psychol.* 34, 394–404. doi: 10.1037/0022-3514.34.3.394

Park, J. L., Fairweather, M. M., and Donaldson, D. I. (2015). Making the case for mobile cognition: EEG and sports performance. *Neurosci. Biobehav. Rev.* 52, 117–130. doi: 10.1016/j.neubiorev.2015.02.014

Perrey, S. (2008). Non-invasive NIR spectroscopy of human brain function during exercise. *Methods* 45, 289–299. doi: 10.1016/j.ymeth.2008.04.005

Piggott, B., Müller, S., Chivers, P., Cripps, A., and Hoyne, G. (2020). Interdisciplinary sport research can better predict competition performance, identify individual differences, and quantify task representation. *Front. Sport. Act. Living* 2:14. doi: 10.3389/fspor.2020.00014

Robertson, S. J., Burnett, A. F., and Cochrane, J. (2014). Tests examining skill outcomes in sport: a systematic review of measurement properties and feasibility. *Sport. Med.* 44, 501–518. doi: 10.1007/s40279-013-0131-0

Robinson, P., Oades, L. G., and Caputi, P. (2015). Conceptualising and measuring mental fitness: a Delphi study. *Int. J. Wellbeing* 5, 53–73. doi: 10.5502/ijw.v5i1.4

Russell, M., Benton, D., and Kingsley, M. (2011). The effects of fatigue on soccer skills performed during a soccer match simulation. *Int. J. Sports Physiol. Perform.* 6, 221–233. doi: 10.1123/ijspp.6.2.221

Russell, M., and Kingsley, M. (2011). Influence of exercise on skill proficiency in soccer. *Sport. Med.* 41, 523–539. doi: 10.2165/11589130-000000000-00000

Themanson, J. R., Pontifex, M. B., and Hillman, C. H. (2008). Fitness and action monitoring: evidence for improved cognitive flexibility in young adults. *Neuroscience* 157, 319–328. doi: 10.1016/j.neuroscience.2008.09.014

Thompson, T., Steffert, T., Ros, T., Leach, J., and Gruzelier, J. (2008). EEG applications for sport and performance. *Methods* 45, 279–288. doi: 10.1016/j.ymeth.2008.07.006

Wesemann, U., Willmund, G. D., Ungerer, J., Kreim, G., Zimmermann, P. L., Bühler, A., et al. (2018). Assessing psychological fitness in the military – development of an effective and economic screening instrument. *Mil. Med.* 183, e261–e269. doi: 10.1093/milmed/usy021

Wright, M. J., Bishop, D. T., Jackson, R. C., and Abernethy, B. (2013). Brain regions concerned with the identification of deceptive soccer moves by higher-skilled and lower-skilled players. *Front. Hum. Neurosci.* 7:851. doi: 10.3389/fnhum.2013.00851

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Psychology

HYPOTHESIS AND THEORY
published: 03 May 2022
doi: 10.3389/fpsyg.2022.758123

Performance Monitoring, Subordinate's Felt Trust and Ambidextrous Behavior; Toward a Conceptual Research Framework

Farooque Ahmed[1], Shuaib Ahmed Soomro[1], Fayaz Hussai Tunio[2], Yi Ding[3] and Naveed Akhtar Qureshi[1]*

[1] Sukkur IBA University, Sukkur, Pakistan, [2] Shaheed Zulfiqar Ali Bhutto University of Law, Karachi, Pakistan, [3] Central University of Finance and Economics, Beijing, China

OPEN ACCESS

Edited by:
Kittisak Jermsittiparsert,
Dhurakij Pundit University, Thailand

Reviewed by:
Supat Chupradit,
Chiang Mai University, Thailand
Parinya Siriattakul,
Mahidol University, Thailand
Nattavud Pimpa,
Mahidol University, Thailand
Priyanut Wutti Chupradit,
Chiang Mai University, Thailand

***Correspondence:**
Yi Ding
dingruohan@163.com

Specialty section:
This article was submitted to Performance Science, a section of the journal Frontiers in Psychology

Received: 13 August 2021
Accepted: 28 January 2022
Published: 03 May 2022

Citation:
Ahmed F, Soomro SA, Tunio FH, Ding Y and Qureshi NA (2022) Performance Monitoring, Subordinate's Felt Trust and Ambidextrous Behavior; Toward a Conceptual Research Framework. Front. Psychol. 13:758123. doi: 10.3389/fpsyg.2022.758123

The present research proposes an electronic performance monitoring framework based on ambidextrous leadership and social exchange theories in a dynamic environment. It reviews and integrates essential literature on electronic performance management (EPM), trust, and ambidextrous behavior. For this, authors have reviewed relevant literature on various themes and underpinned them for managing EPM. The study emphasizes individuals' psychological foundations that demonstrate trust behavior and relationship with their leaders. Eventually, through an ambidextrous approach, managers gain steady performance and relationships with their subordinates through EPM. The study shows that ambidexterity benefits organizations; it enhances employees' resources, resulting in enhanced performance that leads to the performance of an organization. The authors discuss the theoretical as well as practical implications of this study.

Keywords: performance, monitoring, trust, leader-member exchange (LMX), ambidextrous behavior, ambidextrous leadership

INTRODUCTION

Leader's role has always remained in the limelight in business and academia (Cortellazzo et al., 2019). In this era of quick change, evolution, and technological improvements, organizational leaders expect their subordinates to be experts in dealing with current and upcoming challenges (Hunter and Perreault, 2007). The literature on leadership describes that employees and associates want humble, insightful, empowering leaders (Owens et al., 2013). Contrary to the literature, business leaders in practice exercise swashbuckling in all-seeing and all-doing ways in an organization to get work done. Hence, these leaders are not humble and quiet (Johnson et al., 2012), they are the exact opposite of what we have in literature and what practicing leaders do in business

(Matos et al., 2018). These swashbuckling practices create a dilemma for leaders. They think about what kind of leadership style they can apply to succeed in the competitive market that exercises a humble way to get their work done. At the same time, the correct type of leadership depends on the business situation. For instance, competition among service providers requires leaders to focus on service employees' quality service to meet increasing customer demands (Agnihotri et al., 2017). Leaders who combine service quality with sales enable competitive advantage which helps them to motivate workers to perform simultaneously (Gabler et al., 2017).

In today's competitive environment, sales leaders have doubted their subordinates' performance working in the field. As a practice, they cannot be with them working in the area, so they pay surprise visits to see how they are workings in the field. The sales leaders' objective is to meet the service quality and sales results through their subordinates. Typically, sales employees work remotely and away from their leaders (Cascio, 2000). Physical space separates them and may affect their relationship, leading to a decline in their performance (Wieseke et al., 2008). In contrast, technological gadgets provided leaders with the convenience of accessing their subordinates' performance in the field. In this regard, they can assess and manage their performance efficiently to achieve subordinates' success. With the arrival of electronic performance monitoring (EPM), leaders are further benefiting from a myriad of valuable services like performance measurement and improvement, productivity reports and communication services.

Prior research has shown employees' significant concerns with monitoring; consequently, it creates a working environment by reducing trust and unpleasant working relationships (Greengard, 1996; Lewis and Sobhan, 1999). It has adverse outcomes like work stress (Kolb and Aiello, 1996) and perceived distrust (Frey, 1993; Ariss, 2002; Smith and Tabak, 2009). Accordingly, organizations demonstrate ambidextrous behavior among leaders to handle subordinate behavior (Kao and Chen, 2016). Such behavior labels employees simultaneously to exploring new skills and exploiting present skills in their job responsibilities and obligations (Mom et al., 2009; Kauppila and Tempelaar, 2016). Few firms, for example, train sales staff to simultaneously engage in cross-selling and up-selling (Jasmand et al., 2012; Johnson and Friend, 2015). Previous studies described how employees and organizational performance are certainly affected by leaders' ambidextrous behavior (Auh and Menguc, 2005; Cao et al., 2009; Mom et al., 2009; Kauppila, 2010; Patel et al., 2013). More precisely, individual-level ambidexterity has been found to increase sales performance (Jasmand et al., 2012).

Previous research shows that ambidexterity benefits many companies, it may also enhance employees' resources, resulting in greater performance (Gabler et al., 2017). There is progress in exploring individual ambidexterity and its influence on workers (Kao and Chen, 2016; Kauppila and Tempelaar, 2016; Gabler et al., 2017). More is needed to explore how sales leaders support subordinates' ambidextrous behavior. Besides this, subordinates expect trust and respect from their leaders (Ullrich et al., 2009); they want independence and dignity in their employment (Deci and Ryan, 2002). Research has shown if EPM employees think that they are being viewed with integrity and dignity by their leader, they can respond by believing the leader more (McNall and Roch, 2009). The positive effect led to leader-member exchange (LMX) due to a rise in the trust level (Newcombe and Ashkanasy, 2002). Further, individuals who experience high-quality LMX openly address challenges in achieving their job goals amid the monitoring process (Audenaert et al., 2019). The ambidextrous individuals refine and update their expertise, knowledge, and skills (Schnellbächer and Heidenreich, 2020), specifically in a sales job to build and retain clients. In this track, EPM purpose can enhance an employee's desire to progress in knowledge, skills, and abilities (Ravid et al., 2020).

This paper contributes to the existing literature in many ways.

- First, in light of ambidextrous behavior (Gabler et al., 2017), and a future research call from Ravid et al. (2020), there is a need to study EPM to achieve employees' engagement in exploitation and exploration behaviors.
- Second, past research on the developmental purpose of EPM was restricted to attitudinal outcomes (Wells et al., 2007). We add to the literature through its behavioral effects on LMX and ambidextrous behavior. Hence, understanding of the perceived purpose of EPM will be enhanced.
- Third, to better record reciprocal reactions of subordinates after their perception of EPM as developmental, the research proposed serial mediation of felt trust and perceived LMX quality. The serial mediation can help in gaining a solid knowledge of underlying mechanism between perceived developmental EPM and ambidextrous behaviors.
- Fourth, the study is proposed in sales field where subordinates normally work away from their leader (Cascio, 2000). Under fixed EPM, sales people can only achieve sales goals consistently by demonstrating in ambidextrous behavior. Keeping monitoring as developmental, this model offers trust, respect and LMX to subordinates with purpose to reciprocate in the form of behavioral ambidexterity.
- Finally, the research framework contributes to social exchange theory (Blau, 1964) that posits that how individual engage in exchange relationship that explicitly brings a win– win situation for all the stakeholders.

LITERATURE REVIEW

Electronic Performance Monitoring and Its Developmental Purpose

Electronic performance monitoring (EPM) is an integral part of new information mechanisms and working environments. Monitoring employees' performance helps companies determine whether to pay or not (Alder, 2001). Motivations for EPM implementation are to assess both constructive (productivity, task performance) and detrimental behaviors of employees, like counterproductive work behaviors (CWBs) (Tomczak et al., 2018). EPM can enhance employees' performance (Bhave, 2014), and it has shown its positives and negatives like data protection,

health monitoring, and safety protection (Alge and Hansen, 2014), stress (Kolb and Aiello, 1996), and distrust (Frey, 1993; Ariss, 2002; Smith and Tabak, 2009).

Further, organizing constructs relevant to monitoring indicate that perceivable monitoring features influence the employees' feelings, opinions, and assessments about monitoring and then the effect on thoughts such as fairness, trust, and happiness (Stanton and Weiss, 2000). Purpose, probably more than any other EPM feature, can most clearly express what a company values and anticipates from employees (Wells et al., 2007; Jeske and Kapasi, 2018). When observed as developmental, monitoring is regarded as fairer compared to when it is alleged as a warning to future conduct (Wells et al., 2007). Expanding on the social exchange model of McNall and Roch (2009) on EPM reactions, we believe that the perceived developmental purpose of EPM can be a base for felt trust and perceived LMX quality and ultimately ambidextrous behavior of sales workers.

Felt Trust

The felt trust refers to a judgment about the degree to which others trust you (Gill et al., 2019). To feel others' trust, the individual has to recognize that the trustor has the impression that the trustee will complete specific actions worthy of the trustor (Lau et al., 2007). Subordinates who sense trust realize that another party expects intelligent behavior from them without monitoring (Lau et al., 2014). The subordinate's felt trust has a significant positive impact on subordinate's psychological empowerment (Karunarathne, 2019). In the context of leader and subordinate relationships, trust can lead to subordinates' positive behavior toward the leader (Dirks and Ferrin, 2002) and brings exchange relationship quality and teamwork (Chiu and Chiang, 2019). If trusted partnerships are not formed and sustained, salespeople and sales managers alike can waste precious time on efforts intended to defend themselves from each other (Strutton et al., 1993). Since workers cannot know instantly to what degree their leader trusts them, the sense of trust is likely to evolve based on behavioral and situational signals perceived as demonstrations of trust or absence thereof. Therefore, in our suggested framework based on Social Exchange theory (Blau, 1964), we place the felt trust after the perceived developmental purpose of EPM and before the perceived LMX. Accordingly, subordinates can feel the trust of their supervisors if they perceive the EPM's purpose as developmental. Further, the felt trust can increase their perception of LMX quality and obligation to pay back as social exchange.

Perceived Leader-Member Exchange

In the exchange relationship, when an individual feels the delight of receiving more support than allocated, he/she considers it as a high-quality relationship (Byun et al., 2017). Consequently, a high-quality LMX relationship evolves with a high degree of loyalty and mutual trust between a leader and his members (Sparrowe and Liden, 2005). The higher the perceived quality of the LMX, the more inspired members are to participate in the social exchange with the leader to keep gaining tangible benefits, e.g., information, and intangible benefits, e.g., the leader's trust (Erdogan and Enders, 2007). The high-quality LMX gives rise to employees' psychological empowerment (Rafique et al., 2022). Researchers have agreed that to respond to high-quality LMX, members will go beyond the necessary in-role performance and participate in organizational citizenship behavior to maintain a stable social exchange (Ilies et al., 2007). Employees receiving preferential treatment from superiors should promote positive actions, e.g., ambidextrous behavior (Rhoades and Eisenberger, 2002).

Ambidextrous Behavior

Ambidextrous behavior is the tool of employees to effectively adapt to complex scenarios by effectively controlling their exploitation and exploration responses. In the organizational setting, individual exploitation is the ability to maintain concentration on the relevant content and the task at stake, whereas exploration includes the quest for innovation and creativity (Good and Michel, 2013). Findings reveal that such an ambidextrous technique contributes to more consistency in their efficiency by adding to their competitive advantage. When individuals are skilled in two qualities, they can become adaptable, happy to extend their perceptions to possibilities, and function well according to situation (Zhang et al., 2019). Considering the worth of ambidextrous behavior, research constantly redefines it as different conflicting demands like adaptability versus alignment (Gibson and Birkinshaw, 2004), flexibility versus efficiency (Adler et al., 1999; Yu et al., 2020) creativity versus attention to detail (Sok and O'Cass, 2015), sales and service quality (Agnihotri et al., 2017). Specifically for sales workers, it is central to behave ambidextrously (Van der Borgh et al., 2017). For example, they can achieve sales growth by selling higher quantities to current customers through exploitation and prospecting new clients to achieve sales growth through exploration. Hence, we have kept ambidextrous behavior as an outcome in a social exchange process to create a win–win situation.

THEORETICAL FRAMEWORK AND PROPOSITIONS DEVELOPMENT

Developmental Purpose of Electronic Performance Monitoring and Subordinate's Felt Trust

Subordinates cannot know immediately to what degree their leader trusts them. Feelings of being trusted or not trusted are expected to occur after perception of behavioral and situational signals as demonstrations of trust or absence thereof (Lau et al., 2014). Normally, monitoring is a signal of no confidence, and it is expected to be perceived by subordinates as a symbol of distrust (Frey, 1993; Ariss, 2002; Smith and Tabak, 2009). However, the employee's perception that the intent of EPM is developmental would give the impression that the individual is capable of the time and investment needed for development efforts. The perception of EPM as developmental could convey the signal to the individual that you are trusted and respected (Wells et al., 2007). Therefore, it is expected that the perceived

developmental purpose of EPM can induce feelings of being trusted in subordinates. Thus, we propose that;

> P:1 There will be a positive relationship between perceived developmental EPM and felt trust.

Developmental Electronic Performance Monitoring and Leader-Member Exchange

Earlier research has criticized EPM as it invades privacy, increases stress, decreases job satisfaction, and creates a work environment characterized by weakened trust and undesirable work relationships (Piturro, 1989; Greengard, 1996; Lewis, 1999). The belief that the function of EPM system is to limit employees from practicing unnecessary and undesirable behaviors may indicate that the company has neither confidence nor trust. And employees can't work satisfactorily in the absence of monitoring. Such a perception would not probably convey an acknowledgment as a respected member and instead may consider the employee for investigation (Wells et al., 2007). So, in the exchange relationship, it is evident that the deterrent perception of electronic performance monitoring can harm the relationship between a subordinate and his/her leader. On the other hand, this relationship can be improved when subordinates perceive the purpose of monitoring as developmental. When EPM employees think that they are being viewed with integrity and dignity by their leader, they can respond by believing the leader more (McNall and Roch, 2009). The developmental motive for EPM (Tomczak et al., 2018) may positively impact the perception of relationship quality among subordinates. Hence, we expect that the perceived developmental purpose of EPM can enhance the perception of subordinate's LMX quality. Thus, it is suggested that

> P:2 There will be a positive relation between developmental EPM perception and subordinate's perceived LMX.

Developmental Electronic Performance Monitoring and Ambidextrous Behavior

It is well established in organizational literature that employees' attitudes and behaviors are linked to their electronic performance monitoring (Tomczak et al., 2018). So, the way EPM is introduced and conveyed to workers becomes critical. The broad difference in views on EPM indicates that subordinates don't respond similarly to monitoring in all situations (Alder, 2001). The findings has shown that when EPM is seen as developmental, it is acknowledged as fair and brings a commitment to the organization and felt obligation (Wells et al., 2007). In the organizational setting, individual exploitation is the ability to maintain concentration on relevant content and the task at stake, whereas exploration includes' quest for innovation and creativity. However, employees can succeed in ambidexterity by situational alignment (Gibson and Birkinshaw, 2004).

In contrast, organizational structures are occasionally needed to facilitate behavioral ambidexterity individually (Volery et al., 2015). They proposed developing ambidexterity through a suitable organizational framework, comprising attributes of support, discipline, stretch, and trust (Gibson and Birkinshaw, 2004). They demonstrate that mutual respect, transparency, and trust among personnel lead to promoting an environment of information sharing that has a meaningful impact on individual ambidexterity (Ajayi et al., 2017). Recently, Ravid et al. (2020) expected that the developmental purpose of EPM could enhance an employee's desire to progress in existing skill or develop a new one. Therefore, we have supposed our third proposition.

> P:3 There will be a positive relation between developmental EPM perception and Ambidextrous behavior.

Felt Trust and Perceived Leader-Member Exchange Quality

Subordinates significantly like trust and respect from their leader and organization (Ullrich et al., 2009). Felt trust relates to subordinates' opinions about how strongly their superiors trust them (Lester and Brower, 2003). From a social exchange point of view, this indicates their leader's willingness to spend additional effort to strengthen and enhance their relationships, which excites subordinates to contribute to the social exchange (Dulebohn et al., 2012). While few scholars have examined the effect of felt trust on quality LMX, results were either positive (Lau et al., 2014; Kim et al., 2018) or negative (Baer et al., 2015). However, without trust, it is unlikely to have high-quality LMX (McKnight et al., 1998), as trust enables a more efficient exchange partnership between two parties (Colquitt et al., 2007). Therefore, it would be beneficial to add this relationship in the context of the sales force. Hence, we propose,

> P:4 There will be a positive relationship between subordinate's felt trust and perceived LMX quality.

Felt Trust and Ambidextrous Behavior

Being under the umbrella of trust can lead to a sense of responsibility or obligation in trusted individuals to perform tasks or roles required by trustors (Lau et al., 2014). The social exchange is one mechanism through which trust can bring positive work results (Gill et al., 2019). Earlier work on felt trust has shown its importance for multiple positive organizational and employees' outcomes like job satisfaction, less intention to leave, organizational citizenship, job performance, psychological empowerment, and trust in the supervisor (Lester and Brower, 2003; Brower et al., 2009; Gill et al., 2019). If subordinates recognize that their leader trusts them, their organizational self-esteem is improved, encouraging them to perform much better in the field (Lau et al., 2014). Research has also shown a positive link between trust and an individual's behavior to provide novel ideas (Rodrigues and Veloso, 2013), an employee's intrinsic motivation and experience of mastery (Bernstrøm and Svare, 2017). Trust as an organizational factor also encourages ambidexterity at an individual level (Zhang et al., 2019). Therefore, we expect that a subordinate will demonstrate ambidextrous behavior if he feels his leader's trust. Thus, we propose

> P:5 There will be a positive relationship between felt trust and ambidextrous behavior.

Felt Trust as a Mediator Between Perceived Developmental Electronic Performance Monitoring and Leader-Member Exchange Quality

The positive perceptions toward leaders are important in developing high-quality LMX since these offer pleasant feelings to subordinates and direct subordinates to have faith in continuous advantage from the exchange relationship (Liao and Chun, 2016). Developing subordinates' competencies through EPM can enhance LMX quality as it leads to the positive perception of subordinates toward their leader. This perception could convey the message to the subordinates that they are trusted (Wells et al., 2007). The degree to which a manager trusts a subordinate has implications for the nature of the relationship between the subordinate and the supervisor and for autonomy at work (Seppälä et al., 2011). Trust is also recognized as a significant mediator between different organizational activities and worker outcomes (Bernstrøm and Svare, 2017). Felt trust has also gained considerable support in the existing literature as a mediating variable. Earlier, Falk and Kosfeld (2006) tested the mediating effect of felt trust between monitoring and intrinsic motivation, contributing to a decrease in trust. In other study, the relationship between monitoring and intrinsic motivation and monitoring and mastery was fully mediated by the felt trust (Bernstrøm and Svare, 2017). Therefore, this research proposes,

P:6 Felt trust will mediate the relationship between perceived developmental EPM and perceived LMX quality.

Perceived Leader-Member Exchange and Ambidextrous Behavior

Subordinates with a high LMX partnership believe that they are operating in an inspiring psychological atmosphere. Obligatory, they participate in discretionary processes and innovative work by responding positively to their leader's favors (Atwater and Carmeli, 2009; Volmer et al., 2012). Subordinates in high-quality LMX partnerships are considered to be knowledgeable and credible, and acquire additional tools relevant to tasks and relational help to execute assignments (Gu et al., 2015) efficiently. Earlier research revealed that high-quality relationships between a leader and members create a psychological atmosphere that promotes salespeople's empowerment by raising subordinates' feelings of autonomy and care (Martin and Bush, 2006). Research has confirmed the subordinates' empowerment through leadership style brought service-sales ambidexterity (Yu et al., 2013). In the near past, cross-functional teamwork, association, and confidence with the supervisor were the features that visibly led to behavioral ambidexterity of subordinates (Yu et al., 2013; Patterson et al., 2014; Van der Borgh et al., 2017). Therefore, we argue that the quality of LMX can influence the ambidextrous behavior of sales employees. Thus, we propose that,

P:7 There will be a positive relationship between perceived LMX and ambidextrous behavior.

Perceived Leader-Member Exchange as a Mediator Between Felt Trust and Ambidextrous Behavior

Felt trust increases followers' beliefs in their functioning capacity, leading to success in tasks and different behaviors (Zheng et al., 2019). From the social exchange view, members who are trusted feel obligated to keep that trust and reciprocate by working hard to enhance their task performance (Lau and Liden, 2008). The output of LMX represents a particular type of social exchange within the company (Cropanzano et al., 2002) that could be a possible mediator because of psychological impact that a leader exerts. According to the social exchange theory, leaders' positive actions may build liabilities among subordinates by creating a favor exchange. The favors exchange leads subordinates to feel advantages at many levels, including organizational resource control, competence, consideration, and trust (Li et al., 2012). In a recent study, perceived LMX was found as a mediator between the leader's trust and subordinate's task performance (Byun et al., 2017). So, it is expected that perceived LMX can mediate the relationship between the perceived developmental purpose of EPM and ambidextrous behavior. Drawing on the social exchange theory, we propose

P:8 Perceived LMX by subordinate will mediate the relation between felt trust and ambidextrous behavior.

Felt Trust and Leader-Member Exchange as Mediators Between Perceived Developmental Electronic Performance Monitoring and Ambidextrous Behavior

In this research model as revealed in **Figure 1**, developmental EPM offers a signal of being trusted to subordinates, which leads to the perception of high LMX and consequently enables them to engage in exploitation and exploration in the form of ambidextrous behavior. Reciprocity is the social exchange law whereby the two sides satisfy their gains and accomplish exchange (Cropanzano and Mitchell, 2005). Social exchange relationships also have implicit, instead of explicit, obligations regarding reciprocity. Trust between two people is imperative to continue the relationship (Blau, 1964). Subordinates are more willing to recognize a leader's authority if they believe they are trusted by the leader (Zheng et al., 2019). Earlier research has recommended that employees who are assumed competent are inclined to construct and sustain a higher LMX level with their superiors, but those who are considered ineffective are expected to maintain a lower LMX quality (Graen and Uhl-Bien, 1995; Gerstner and Day, 1997; Liden et al., 1997). Subordinates show more desirable habits, like higher performance, when the standard of LMX is higher rather than poor, since they want to give back the advantages of their supervisor's high-quality relationship (Gerstner and Day, 1997). Against these characteristics of Felt trust and LMX, we expect that felt trust and LMX will mediate the relationship between perceived developmental EPM and ambidextrous behavior. Drawing on social exchange theory (Blau, 1964), we propose,

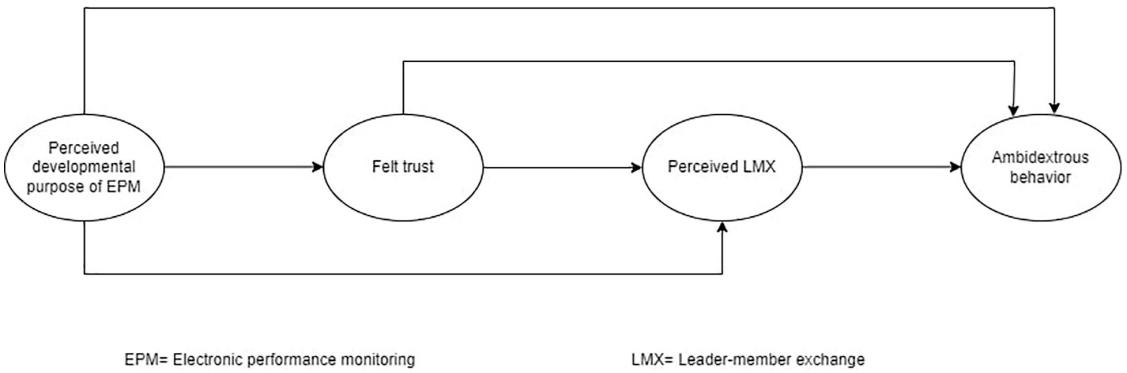


FIGURE 1 | Research model.

P:9 Felt trust and perceived LMX will mediate the relationship between Perceived developmental EPM and ambidextrous behavior.

CONCLUSION

In this competitive world, organizational leaders want their workers to be capable of dealing with current and future problems on the job (Hunter and Perreault, 2007). In contrast to popular belief, most business leaders in practice are swashbuckling, strong, in all-doing and in all-seeing. These are not rulers who are polite and calm (Johnson et al., 2012). This is completely opposite to what we read in the literature and what they do (Matos et al., 2018).

The arrival of EPM has added more power to leaders as they can monitor their subordinates continuously. However, a major monitoring issue is that it produces working environments marked by reduced trust and unpleasant working relationships (Greengard, 1996; Lewis, 1999). We believe that reduced trust and unfriendly relationships can distract the performance of subordinates working in the field. Hence, here is a solid case for monitoring as developmental instead of a deterrent. It is the starting point of our social exchange framework. Thus, we propose that if subordinates perceive EPM as developmental, they can feel their sales manager's trust and the perceived quality of LMX. According to social exchange theory (Blau, 1964), leaders' positive actions may build liabilities among subordinates by creating a favor exchange.

Consequently, we have reasoned that subordinates' ambidextrous behavior can be an outcome of the perceived developmental purpose of monitoring. Our framework is consistent with the social exchange model of McNall and Roch (2009) on reactions of employees to the developmental purpose of EPM.

In the context of the leader-subordinate relationship, trust can lead to subordinates' positive behavior toward the leader (Dirks and Ferrin, 2002) and brings exchange relationship quality and teamwork (Chiu and Chiang, 2019). Being under the trust can lead to the sense of responsibility or obligation in trusted individuals to perform tasks or roles as required by trustors (Lau et al., 2014). Accordingly, we have suggested a positive relationship among felt trust, perceived LMX, and ambidextrous behavior. The entire relationship between variables is backed by social exchange theory (Blau, 1964). Hence, this inclusive framework can bring mutual gain for a sales leader and sales subordinates working in the field. Under the umbrella of trust and LMX, sales workers are expected to demonstrate ambidextrous behavior if they perceive monitoring as developmental. Sales Managers will also gain steady performance and professional partnership with their subordinates by creating a trustful environment. In this framework, we have well-adjusted the power between leader and associate and termed it as a win-win situation for all.

We contribute to social exchange theory by providing a novel and useful framework in the context of sales. Based on reciprocity, this analytical framework demonstrates a positive relationship between the perceived developmental purpose of EPM, Felt trust, perceived LMX quality, and ambidextrous behavior of sales workers. Earlier studies on the developmental purpose of EPM were limited to attitudinal outcomes (Wells et al., 2007). We add to the literature through its behavioral effects on LMX and ambidextrous behavior. Hence, understanding of the perceived purpose of EPM will be enhanced. In earlier studies, the relationship between felt trust and leader- member exchange was conflicting. Few studies exposed it as positive (Lau et al., 2014; Kim et al., 2018) and others revealed it as negative (Baer et al., 2015). But, we believe that in the context of sales, the relationship will be positive. Previous research has shown creativity, and efficient execution of assignments as positive outcomes of LMX; (Gu et al., 2015). Thus, ambidextrous behavior is expected as the positive outcome of perceived high-quality LMX, which is another contribution to literature. We expect this entire framework as a win-win situation for sales leaders and subordinates and will find the right place in literature.

Practical Implications

Sales managers in electronic performance monitoring need to value subordinates' esteem. There may be some organizational factors that potentially encourage managers to monitor employees. However, the purpose should be developmental instead of deferral. There are many logics for this. First, it is the ethically right thing to do. Second, evidence from many case studies shows that subordinates monitored through EPM

feel additional stress than associates observed through other methods (Kolb and Aiello, 1996). Third, the expected benefits of monitoring may be reduced or even removed if workers have an adverse reaction to the EPM system (Jeske and Santuzzi, 2015). It sheds light on the value of trust in the relationship. Since workers cannot know instantly to what degree their leader trusts them, the sense of trust is likely to evolve based on behavioral and situational signals perceived as demonstrations of trust or absence thereof (Kim et al., 2018). Hence, through perceived developmental EPM, subordinates will feel the trust of their leader. Felt trust is specifically essential for sales workers as they work remotely and physically away from their leader. This physical distance may decrease belonging between sales leaders and salespeople (Cascio, 2000). However, trust is recognized as the single most critical feature of any successful professional partnership (Kramer, 1999). Therefore, leaders will be able to enjoy a required professional collaboration with their subordinates. Subordinates will also perceive high LMX relationship quality on the perception of developmental EPM and felt trust.

Further, felt trust has importance for multiple positive organizational and employees' outcomes like job satisfaction, less intention to leave, organizational citizenship, job performance, psychological empowerment, and trust in the supervisor (Lester and Brower, 2003; Brower et al., 2009; Gill et al., 2019). Salespeople serve a pivotal role in successfully implementing the organizational strategy of selling new and existing products (Van der Borgh et al., 2017). In this framework, the most important implications for sales managers can be for ambidextrous behavior of salesforce. If they want subordinates to sell new and existing products, they can achieve it by promoting developmental EPM and exchange relationship quality.

Limitations and Suggestions for Future Research

Based on the social exchange theory, this model offers new theoretical relationships that are needed to be tested empirically. The research opens a call for future studies to test this framework in the field of sales. We limited ourselves to the developmental perception of EPM and focused only on a positive feature. However, the purpose of EPM can also be a deterrent. Earlier research has revealed a negative relationships between deterrent perception of EPM and employees' outcomes like organizational commitment, felt obligation, job satisfaction, and perceived fairness (Wells et al., 2007). Future research can study the deterrent purpose of EPM's impact on variables in this study and confirm how it will outline the relationship among variables. The theory of social exchange predicts that if subordinates consider their company is less eager to contribute to social exchange relation, they are also less inclined to engage in social exchange (Blau, 1964). Future studies can examine the implicit employment contract between perceived EPM and ambidextrous behavior of employees. In this study, we could only integrate LMX quality perceived by subordinates. Future researchers can also fill this limitation by including LMX perceived by subordinates and LMX perceived by leaders. Finally, we have proposed this model specifically in the context of the salesforce. However, future research can adapt similar framework in other industries, specifically in banking, where monitoring is the most common and importance of trust and LMX relationship is higher.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

FA and SS initially worked together to finalize the all of parts, including the idea of the work, the conceptual framework, systematic literature review, development of propositions, discussion, and conclusion of the manuscript, and produced the initial draft of the manuscript. FT, YD, and NQ helped in addressing the reviewer's comments in the manuscript and contributed to re-drafting the manuscript and also re-evaluating the review of the past literature. All authors made substantial contributions to the revision of the manuscript.

REFERENCES

Adler, P. S., Goldoftas, B., and Levine, D. I. (1999). Flexibility versus efficiency? A case study of model changeovers in the toyota production system. *Org. Sci.* 10, 43–68. doi: 10.1287/orsc.10.1.43

Agnihotri, R., Gabler, C. B., Itani, O. S., Jaramillo, F., and Krush, M. T. (2017). Salesperson ambidexterity and customer satisfaction: examining the role of customer demandingness, adaptive selling, and role conflict. *J. Pers. Sell. Sales Manage.* 37, 27–41. doi: 10.1080/08853134.2016.1272053

Ajayi, O. M., Odusanya, K., and Morton, S. (2017). Stimulating employee ambidexterity and employee engagement in SMEs. *Manage. Decis.* 55, 662–680. doi: 10.1108/md-02-2016-0107

Alder, G. S. (2001). Employee reactions to electronic performance monitoring: a consequence of organizational culture. *J. High Technol. Manage. Res.* 12, 323–342. doi: 10.1016/s1047-8310(01)00042-6

Alge, B. J., and Hansen, S. D. (2014). "Workplace monitoring and surveillance research since 1984: a review and agenda," in *The Psychology of Workplace Technology*, eds M. D. Coovert and L. F. Thompson (Oxfordshire: Routledge), 209–237.

Ariss, S. S. (2002). Computer monitoring: benefits and pitfalls facing management. *Inform. Manage.* 39, 553–558. doi: 10.1016/s0378-7206(01)00121-5

Atwater, L., and Carmeli, A. (2009). Leader–member exchange, feelings of energy, and involvement in creative work. *Leadersh. Q.* 20, 264–275. doi: 10.1016/j.leaqua.2007.07.009

Audenaert, M., Decramer, A., George, B., Verschuere, B., and Van Waeyenberg, T. (2019). When employee performance management affects individual innovation in public organizations: the role of consistency and LMX. *Int. J. Hum. Resour. Manage.* 30, 815–834. doi: 10.1080/09585192.2016.1239220

Auh, S., and Menguc, B. (2005). Balancing exploration and exploitation: the moderating role of competitive intensity. *J. Bus. Res.* 58, 1652–1661. doi: 10.1016/j.jbusres.2004.11.007

Baer, M. D., Dhensa-Kahlon, R. K., Colquitt, J. A., Rodell, J. B., Outlaw, R., and Long, D. M. (2015). Uneasy lies the head that bears the trust: the effects of

feeling trusted on emotional exhaustion. *Acad. Manage. J.* 58, 1637–1657. doi: 10.5465/amj.2014.0246

Bernstrøm, V. H., and Svare, H. (2017). Significance of monitoring and control for employees' felt trust, motivation, and mastery. *Nordic J. Work. Life Stud.* 7, 29–49.

Bhave, D. P. (2014). The invisible eye? Electronic performance monitoring and employee job performance. *Pers. Psychol.* 67, 605–635. doi: 10.1111/peps.12046

Blau, P. M. (1964). Justice in social exchange. *Soc. Inq.* 34, 193–206.

Brower, H. H., Lester, S. W., Korsgaard, M. A., and Dineen, B. R. (2009). A closer look at trust between managers and subordinates: understanding the effects of both trusting and being trusted on subordinate outcomes. *J. Manage.* 35, 327–347. doi: 10.1177/0149206307312511

Byun, G., Dai, Y., Lee, S., and Kang, S. (2017). Leader trust, competence, LMX, and member performance: a moderated mediation framework. *Psychol. Rep.* 120, 1137–1159. doi: 10.1177/0033294117716465

Cao, Q., Gedajlovic, E., and Zhang, H. (2009). Unpacking organizational ambidexterity: dimensions, contingencies, and synergistic effects. *Org. Sci.* 20, 781–796. doi: 10.1287/orsc.1090.0426

Cascio, W. F. (2000). Managing a virtual workplace. *Acad. Manage. Perspect.* 14, 81–90. doi: 10.5465/ame.2000.4468068

Chiu, H.-C., and Chiang, P.-H. (2019). A trickle-down effect of subordinates' felt trust. *Pers. Rev.* 48, 957–976. doi: 10.1108/pr-01-2018-0036

Colquitt, J. A., Scott, B. A., and LePine, J. A. (2007). Trust, trustworthiness, and trust propensity: a meta-analytic test of their unique relationships with risk taking and job performance. *J. Appl. Psychol.* 92, 909–927. doi: 10.1037/0021-9010.92.4.909

Cortellazzo, L., Bruni, E., and Zampieri, R. (2019). The role of leadership in a digitalized world: a review. *Front. Psychol.* 10:1938. doi: 10.3389/fpsyg.2019.01938

Cropanzano, R., and Mitchell, M. S. (2005). Social exchange theory: an interdisciplinary review. *J. Manage.* 31, 874–900. doi: 10.1177/0149206305279602

Cropanzano, R., Prehar, C. A., and Chen, P. Y. (2002). Using social exchange theory to distinguish procedural from interactional justice. *Group Org. Manage.* 27, 324–351. doi: 10.1177/1059601102027003002

Deci, E. L., and Ryan, R. M. (2002). "The paradox of achievement: the harder you push, the worse it gets," in *Improving Academic Achievement*, ed. J. Aronson (Amsterdam: Elsevier), 61–87.

Dirks, K. T., and Ferrin, D. L. (2002). Trust in leadership: meta-analytic findings and implications for research and practice. *J. Appl. Psychol.* 87, 611–628. doi: 10.1037/0021-9010.87.4.611

Dulebohn, J. H., Bommer, W. H., Liden, R. C., Brouer, R. L., and Ferris, G. R. (2012). A meta- analysis of antecedents and consequences of leader-member exchange: integrating the past with an eye toward the future. *J. Manage.* 38, 1715–1759. doi: 10.1177/0149206311415280

Erdogan, B., and Enders, J. (2007). Support from the top: supervisors' perceived organizational support as a moderator of leader-member exchange to satisfaction and performance relationships. *J. Appl. Psychol.* 92, 321–330. doi: 10.1037/0021-9010.92.2.321

Falk, A., and Kosfeld, M. (2006). The hidden costs of control. *Am. Econ. Rev.* 96, 1611–1630. doi: 10.1257/aer.96.5.1611

Frey, B. S. (1993). Does monitoring increase work effort? The rivalry with trust and loyalty. *Econ. Inq.* 31, 663–670. doi: 10.1111/j.1465-7295.1993.tb00897.x

Gabler, C. B., Ogilvie, J. L., Rapp, A., and Bachrach, D. G. (2017). Is there a dark side of ambidexterity? Implications of dueling sales and service orientations. *J. Serv. Res.* 20, 379–392. doi: 10.1177/1094670517712019

Gerstner, C. R., and Day, D. V. (1997). Meta-Analytic review of leader–member exchange theory: correlates and construct issues. *J. Appl. Psychol.* 82, 827–844. doi: 10.1037/0021-9010.82.6.827

Gibson, C. B., and Birkinshaw, J. (2004). The antecedents, consequences, and mediating role of organizational ambidexterity. *Acad. Manage. J.* 47, 209–226. doi: 10.2147/PRBM.S332222

Gill, H., Cassidy, S. A., Cragg, C., Algate, P., Weijs, C. A., and Finegan, J. E. (2019). Beyond reciprocity: the role of empowerment in understanding felt trust. *Eur. J. Work Org. Psychol.* 28, 845–858. doi: 10.1080/1359432x.2019.1678586

Good, D., and Michel, E. J. (2013). Individual ambidexterity: exploring and exploiting in dynamic contexts. *J. Psychol.* 147, 435–453. doi: 10.1080/00223980.2012.710663

Graen, G. B., and Uhl-Bien, M. (1995). Relationship-based approach to leadership: development of leader-member exchange (LMX) theory of leadership over 25 years: applying a multi- level multi-domain perspective. *Leadersh. Q.* 6, 219–247. doi: 10.1016/1048-9843(95)90036-5

Greengard, S. (1996). Privacy: entitlement or illusion? *Pers. J.* 75, 74–88.

Gu, Q., Tang, T. L.-P., and Jiang, W. (2015). Does moral leadership enhance employee creativity? Employee identification with leader and leader–member exchange (LMX) in the Chinese context. *J. Bus. Ethics* 126, 513–529. doi: 10.1007/s10551-013-1967-9

Hunter, G. K., and Perreault, W. D. Jr. (2007). Making sales technology effective. *J. Mark.* 71, 16–34. doi: 10.1509/jmkg.71.1.16

Ilies, R., Nahrgang, J. D., and Morgeson, F. P. (2007). Leader-member exchange and citizenship behaviors: a meta-analysis. *J. Appl. Psychol.* 92, 269–277. doi: 10.1037/0021-9010.92.1.269

Jasmand, C., Blazevic, V., and De Ruyter, K. (2012). Generating sales while providing service: a study of customer service representatives' ambidextrous behavior. *J. Mark.* 76, 20–37. doi: 10.1509/jm.10.0448

Jeske, D., and Kapasi, I. (2018). "Electronic performance monitoring: lessons from the past and future challenges. Organizing for digital economy: societies, communities and individuals," in *Proceedings of the 14th Annual Conference of the Italian Chapter of the AIS*, Rome, 119–132.

Jeske, D., and Santuzzi, A. M. (2015). Monitoring what and how: psychological implications of electronic performance monitoring. *New Technol. Work Employ.* 30, 62–78. doi: 10.1111/ntwe.12039

Johnson, J. S., and Friend, S. B. (2015). Contingent cross-selling and up-selling relationships with performance and job satisfaction: an MOA-theoretic examination. *J. Pers. Sell. Sales Manage.* 35, 51–71. doi: 10.1080/08853134.2014.940962

Johnson, R. E., Venus, M., Lanaj, K., Mao, C., and Chang, C.-H. (2012). Leader identity as an antecedent of the frequency and consistency of transformational, consideration, and abusive leadership behaviors. *J. Appl. Psychol.* 97, 1262–1272. doi: 10.1037/a0029043

Kao, Y.-L., and Chen, C.-F. (2016). Antecedents, consequences and moderators of ambidextrous behaviours among frontline employees. *Manage. Decis.* 54, 1846–1860. doi: 10.1108/md-05-2015-0187

Karunarathne, R. A. I. C. (2019). The impact of subordinate's trust in supervisor and felt trust on subordinate psychological empowerment. *Glob. J. Manage. Bus. Res.* 19:33.

Kauppila, O.-P. (2010). Creating ambidexterity by integrating and balancing structurally separate interorganizational partnerships. *Strateg. Org.* 8, 283–312. doi: 10.1177/1476127010387409

Kauppila, O.-P., and Tempelaar, M. P. (2016). The social-cognitive underpinnings of employees' ambidextrous behaviour and the supportive role of group managers' leadership. *J. Manage. Stud.* 53, 1019–1044. doi: 10.1111/joms.12192

Kim, T.-Y., Wang, J., and Chen, J. (2018). Mutual trust between leader and subordinate and employee outcomes. *J. Bus. Ethics* 149, 945–958. doi: 10.1007/s10551-016-3093-y

Kolb, K. J., and Aiello, J. R. (1996). The effects of electronic performance monitoring on stress: locus of control as a moderator variable. *Comput. Hum. Behav.* 12, 407–423. doi: 10.1016/0747-5632(96)00016-7

Kramer, R. M. (1999). Trust and distrust in organizations: emerging perspectives, enduring questions. *Ann. Rev. Psychol.* 50, 569–598. doi: 10.1146/annurev.psych.50.1.569

Lau, D. C., Lam, L. W., and Wen, S. S. (2014). Examining the effects of feeling trusted by supervisors in the workplace: a self-evaluative perspective. *J. Org. Behav.* 35, 112–127. doi: 10.1002/job.1861

Lau, D. C., and Liden, R. C. (2008). "Antecedents of coworker trust: leaders' blessings. *J. Appl. Psychol.* 93, 1130–1138. doi: 10.1037/0021-9010.93.5.1130

Lau, D. C., Liu, J., and Fu, P. P. (2007). Feeling trusted by business leaders in China: antecedents and the mediating role of value congruence. *Asia Pac. J. Manage.* 24, 321–340. doi: 10.1007/s10490-006-9026-z

Lester, S. W., and Brower, H. H. (2003). In the eyes of the beholder: the relationship between subordinates' felt trustworthiness and their work attitudes

and behaviors. *J. Leadersh. Org. Stud.* 10, 17–33. doi: 10.1177/107179190301000203

Lewis, C. (1999). American workers beware: big brother is watching. *USA Today Mag.* 127, 20–23.

Lewis, D., and Sobhan, B. (1999). Routes of funding, roots of trust? Northern NGOs, Southern NGOs, donors, and the rise of direct funding. *Dev. Pract.* 9, 117–129. doi: 10.1080/09614529953269

Li, X., Sanders, K., and Frenkel, S. (2012). How leader–member exchange, work engagement and HRM consistency explain Chinese luxury hotel employees' job performance. *Int. J. Hosp. Manage.* 31, 1059–1066. doi: 10.1016/j.ijhm.2012.01.002

Liao, E. Y., and Chun, H. (2016). Supervisor monitoring and subordinate innovation. *J. Organ. Behav.* 37, 168–192. doi: 10.1002/job.2035

Liden, R. C., Sparrowe, R. T., and Wayne, S. J. (1997). Leader-member exchange theory: the past and potential for the future. *Res. Pers. Hum. Resour. Manage.* 15, 47–120.

Martin, C. A., and Bush, A. J. (2006). Psychological climate, empowerment, leadership style, and customer-oriented selling: an analysis of the sales manager–salesperson dyad. *J. Acad. Mark. Sci.* 34, 419–438. doi: 10.1177/0092070306286205

Matos, K., O'Neill, O., and Lei, X. (2018). Toxic leadership and the masculinity contest culture: How "win or die" cultures breed abusive leadership. *J. Soc. Issues* 74, 500–528. doi: 10.1111/josi.12284

McKnight, D. H., Cummings, L. L., and Chervany, N. L. (1998). Initial trust formation in new organizational relationships. *Acad. Manage. Rev.* 23, 473–490. doi: 10.5465/amr.1998.926622

McNall, L. A., and Roch, S. G. (2009). A social exchange model of employee reactions to electronic performance monitoring. *Hum. Perform.* 22, 204–224. doi: 10.1080/08959280902970385

Mom, T. J., Van Den Bosch, F. A., and Volberda, H. W. (2009). Understanding variation in managers' ambidexterity: investigating direct and interaction effects of formal structural and personal coordination mechanisms. *Org. Sci.* 20, 812–828. doi: 10.1287/orsc.1090.0427

Newcombe, M. J., and Ashkanasy, N. M. (2002). The role of affect and affective congruence in perceptions of leaders: an experimental study. *Leadersh. Q.* 13, 601–614. doi: 10.1016/s1048-9843(02)00146-7

Owens, B. P., Johnson, M. D., and Mitchell, T. R. (2013). Expressed humility in organizations: implications for performance, teams, and leadership. *Org. Sci.* 24, 1517–1538. doi: 10.1287/orsc.1120.0795

Patel, P. C., Messersmith, J. G., and Lepak, D. P. (2013). Walking the tightrope: an assessment of the relationship between high-performance work systems and organizational ambidexterity. *Acad. Manage. J.* 56, 1420–1442. doi: 10.5465/amj.2011.0255

Patterson, P., Yu, T., and Kimpakorn, N. (2014). Killing two birds with one stone: cross-selling during service delivery. *J. Bus. Res.* 67, 1944–1952. doi: 10.1016/j.jbusres.2013.11.013

Piturro, M. C. (1989). Employee performance monitoring. Or meddling? *Manage. Rev.* 78:31.

Rafique, S., Khan, N. R., Soomro, S. A., and Masood, F. (2022). Linking LMX and schedule flexibility with employee innovative work behaviors: mediating role of employee empowerment and response to change. *J. Econ. Adm. Sci.* doi: 10.1108/JEAS-11-2021-0238 [Epub ahead of print].

Ravid, D. M., Tomczak, D. L., White, J. C., and Behrend, T. S. (2020). EPM 20/20: a review, framework, and research agenda for electronic performance monitoring. *J. Manage.* 46, 100–126. doi: 10.1177/0149206319869435

Rhoades, L., and Eisenberger, R. (2002). Perceived organizational support: a review of the literature. *J. Appl. Psychol.* 87, 698–714. doi: 10.1037/0021-9010.87.4.698

Rodrigues, A. F. C., and Veloso, A. L. (2013). Organizational trust, risk and creativity. *Rev. Bras. Gestão Negócios* 15, 545–561. doi: 10.7819/rbgn.v15i49.1334

Schnellbächer, B., and Heidenreich, S. (2020). The role of individual ambidexterity for organizational performance: examining effects of ambidextrous knowledge seeking and offering. *J. Technol. Transfer* 45, 1535–1561. doi: 10.1007/s10961-020-09781-x

Seppälä, T., Lipponen, J., Pirttila-Backman, A.-M., and Lipsanen, J. (2011). Reciprocity of trust in the supervisor–subordinate relationship: the mediating role of autonomy and the sense of Power. *Eur. J. Work Org. Psychol.* 20, 755–778. doi: 10.1080/1359432X.2010.507353

Smith, W. P., and Tabak, F. (2009). Monitoring employee e-mails: Is there any room for privacy? *Acad. Manage. Perspect.* 23, 33–48. doi: 10.5465/amp.2009.45590139

Sok, P., and O'Cass, A. (2015). Examining the new product innovation - performance relationship: optimizing the role of individual-level creativity and attention-to-detail. *Indus. Mark. Manage.* 47, 156–165. doi: 10.1016/j.indmarman.2015.02.040

Sparrowe, R. T., and Liden, R. C. (2005). Two routes to influence: integrating leader-member exchange and social network perspectives. *Adm. Sci. Q.* 50, 505–535. doi: 10.2189/asqu.50.4.505

Stanton, J. M., and Weiss, E. M. (2000). Electronic monitoring in their own words: an exploratory study of employees' experiences with new types of surveillance. *Comput. Hum. Behav.* 16, 423–440. doi: 10.1016/s0747-5632(00)00018-2

Strutton, D., Pelton, L. E., and Lumpkin, J. R. (1993). The relationship between psychological climate and salesperson-sales manager trust in sales organizations. *J. Pers. Sell. Sales Manage.* 13, 1–14.

Tomczak, D. L., Lanzo, L. A., and Aguinis, H. (2018). Evidence-based recommendations for employee performance monitoring. *Bus. Horiz.* 61, 251–259. doi: 10.1016/j.bushor.2017.11.006

Ullrich, J., Christ, O., and van Dick, R. (2009). Substitutes for procedural fairness: prototypical leaders are endorsed whether they are fair or not. *J. Appl. Psychol.* 94, 235–244. doi: 10.1037/a0012936

Van der Borgh, M., de Jong, A., and Nijssen, E. J. (2017). Alternative mechanisms guiding salespersons' ambidextrous product selling. *Br. J. Manage.* 28, 331–353. doi: 10.1111/1467-8551.12148

Volery, T., Mueller, S., and von Siemens, B. (2015). Entrepreneur ambidexterity: a study of entrepreneur behaviours and competencies in growth-oriented small and medium-sized enterprises. *Int. Small Bus. J.* 33, 109–129. doi: 10.1177/0266242613484777

Volmer, J., Spurk, D., and Niessen, C. (2012). Leader–member exchange (LMX), job autonomy, and creative work involvement. *Leadersh. Q.* 23, 456–465. doi: 10.1016/j.leaqua.2011.10.005

Wells, D. L., Moorman, R. H., and Werner, J. M. (2007). The impact of the perceived purpose of electronic performance monitoring on an array of attitudinal variables. *Hum. Resour. Dev. Q.* 18, 121–138. doi: 10.1002/hrdq.1194

Wieseke, J., Homburg, C., and Lee, N. (2008). Understanding the adoption of new brands through salespeople: a multilevel framework. *J. Acad. Mark. Sci.* 36, 278–291. doi: 10.1007/s11747-007-0055-z

Yu, T., Gudergan, S., and Chen, C. F. (2020). Achieving employee efficiency–flexibility ambidexterity. *Int. J. Hum. Resour. Manage.* 31, 2459–2494. doi: 10.1080/09585192.2018.1449762

Yu, T., Patterson, P. G., and de Ruyter, K. (2013). Achieving service-sales ambidexterity. *J. Serv. Res.* 16, 52–66. doi: 10.1177/1094670512453878

Zhang, Y., Wei, F., and Van Horne, C. (2019). Individual ambidexterity and antecedents in a changing context. *Int. J. Innovat. Manage.* 23:1950021. doi: 10.1142/s136391961950021x

Zheng, X., Hall, R. J., and Schyns, B. (2019). Investigating follower felt trust from a social cognitive perspective. *Eur. J. Work Org. Psychol.* 28, 873–885. doi: 10.1080/1359432x.2019.1678588

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers | Frontiers in Psychology

ORIGINAL RESEARCH
published: 09 May 2022
doi: 10.3389/fpsyg.2022.859466

Construction and Validation of the Research Misconduct Scale for Social Science University Students

Saba Ghayas[1], Zaineb Hassan[1], Sumaira Kayani[2] and Michele Biasutti[3]**

[1] Department of Psychology, University of Sargodha, Sargodha, Pakistan, [2] Department of Psychology, Zhejiang Normal University, Jinhua, China, [3] Department of Philosophy, Sociology, Education, and Applied Psychology, University of Padova, Padova, Italy

The current study aims to construct and validate a measure of research misconduct for social science university students. The research is comprised of three studies; Study I presents the scale construction in three phases. In Phase I, the initial pool of items was generated by reviewing the literature and considering the results of semi-structured interviews. Phase II involved a psychometric cleaning of items, after which 38 items were retained. In Phase III, those 38 items were proposed to 652 university students, and data were exposed to exploratory factor analysis, which extracted a one-factor structure with 15 items and 55.73% variance. Study II confirmed the factorial structure of the scale using an independent sample (N = 200) of university students. Confirmatory factor analysis of the scale demonstrates a good model fit to the data with the one-factor structure established through the exploratory factor analysis. The scale exhibits good internal consistency, with a Cronbach's alpha of 0.95. Study III involves validation of the scale, with evidence for convergent validity collected from a sample of university students (N = 200). The results reveal that the research misconduct scale has significant positive correlations with academic stress and procrastination and a significant negative correlation with academic achievement. The obtained convergent validity testifies that the scale can be considered a psychometrically sound instrument to measure research misconduct among social science university students.

Keywords: research misconduct, exploratory factor analysis, validation, confirmatory factor analysis, academic dishonesty, psychometric properties, scale development

OPEN ACCESS

Edited by:
Silvia Riva,
St Mary's University, Twickenham, United Kingdom

Reviewed by:
Yovav Eshet,
Zefat Academic College, Israel
Emanuela Brusadelli,
University of Wollongong, Australia

****Correspondence:***
Sumaira Kayani
sumaira@zjnu.edu.cn
Michele Biasutti
michele.biasutti@unipd.it

Specialty section:
This article was submitted to Performance Science, a section of the journal Frontiers in Psychology

Received: *21 January 2022*
Accepted: *07 April 2022*
Published: *09 May 2022*

Citation:
Ghayas S, Hassan Z, Kayani S and Biasutti M (2022) Construction and Validation of the Research Misconduct Scale for Social Science University Students. Front. Psychol. 13:859466. doi: 10.3389/fpsyg.2022.859466

INTRODUCTION

Misconduct in research and academic dishonesty are important, persistent issues for universities, as most students have engaged in academic misconduct at some point of their careers (Peled et al., 2019). Almost all (92%) surveyed students reported having cheated at least once or knowing someone who had (Eshet et al., 2021). Unethical research practices and research misconduct can also be found among scholars. A study conducted in Africa showed that about 68.9% of a group of researchers admitted being involved in one of the following forms of research misconduct: plagiarism, falsifying data, intentional protocol violations, selective dropping of data, falsification of biosketches, disagreements about authorship, and pressure from study weight (Okonta and Rossouw, 2013). Another report regarding research misconduct in Nigeria revealed that about 54.6% of researchers acknowledged engaging in at least one practice listed under the criteria of research misconduct (Adeleye and Adebamowo, 2012). There are many indices of research

misconduct in other countries, including South Korea, Japan, Taiwan, and China (Bak, 2018; Tsai, 2018); thus, considering the evidence of high misconduct rates, research misconduct needs to be comprehensively examined (Steneck, 2006; Steen, 2011).

A crucial point for developing studies to investigate the effects of research misconduct behavior is having the necessary tools for assessing the phenomenon; to our knowledge, such tools are scarce. The present study seeks to answer the call for tools by constructing and validating a measure of research misconduct among social science students and was conducted with the following objectives:

1. To develop a self-report research misconduct scale for university students.
2. To examine the psychometric properties of the research misconduct scale.

The research is comprised of three studies. Study I presents the scale construction in three phases. Study II illustrates the verification of the factorial structure of the scale. Study III demonstrates the convergent validity of the scale. The background section deals with the definition of misconduct in research, terms that can indicate different types of research misconduct, and the factors connected with misconduct in research.

BACKGROUND

Ethical standards and morals play a central role in maintaining research integrity in the scientific community. Abiding norms and ethics promote research based on truthful information and knowledge, discouraging scholars from making errors. Despite detailed and comprehensive guidelines on the standards, ethics, and rules to be followed in research, some scholars still become involved in research misconduct. That behavior is also found among university students who have to conduct research in their final years to earn their degrees.

Research misconduct can be defined as a transgression that occurs when a researcher is involved in fabricating data, falsifying data, or plagiarizing ideas and information in a research project, article, or report. The definition of research misconduct can also be extended to involve wrongdoing related to publication, authorship, and standards of confidentiality (American Psychological Association [APA], 2019). According to Okonta and Rossouw (2013), research misconduct involves various malpractices and actions, such as plagiarizing data, falsification, fabrication, and intentionally violating protocols relevant to research procedures and the enrolment of participants. Other issues involve selectively dropping or skipping outlier cases, conflicts regarding authorship, and pressure from those sponsoring the research study (e.g., an organization or pharmaceutical company) to indulge in research wrongdoing. For the past 20 years in particular, research misconduct and research integrity have been widely discussed and sometimes hotly debated on a variety of platforms. According to the Office of Research Integrity (ORI) in the United States, adhering to commonly accepted rules, standards, principles, norms, and morals is called research integrity (Office of Research Integrity, 2001).

To understand research misconduct, terms like fabrication, falsification, and plagiarism have to be defined, as each is a form of research misconduct. Making up results is called fabrication (Yamamoto and Lennon, 2018), which includes creating and reporting false data or information in a study (Office of Research Integrity, 2019). Fabrication can consist of constructing or adding information, observations, or data that were never actually found during data collection or any other research process. Claims and comments made about incomplete or falsified data sets are also considered a form of fabrication (The Pennsylvania State University, 2018). Fabrication of data is an important research misconduct behavior that can be found in scientific research across disciplines. According to Stretton et al. (2012), almost 52.1% of articles from the medical sciences were found to contain incidences of misconduct, including the fabrication of data.

Falsification can also entail the manipulation of materials, instruments, or processes involved in research or excluding information or results so that representation of the actual research work is compromised (De Vries et al., 2006; Steneck, 2006; Krimsky, 2007; Office of Research Integrity, 2019).

Plagiarism refers to the outright theft or the surreptitious, uncredited use of another person's ideas, work results, or research. It also includes confidential reviews of other research proposals, reports, synopses, and manuscripts. Another core aspect is that research misconduct is performed intentionally and does not involve genuine differences of opinion or honest errors that can occur in the normal course of research. The World Association of Medical Editors (2019) defines plagiarism as the use of others' distributed and unpublished thoughts, words, or other licensed innovation without authorization and presenting them as novel. Several motivations could induce authors to resort to plagiarism, including the pressure to publish and having substandard research skills (Jawad, 2013).

Clear and explicit examples of research wrongdoing include plagiarism, falsification, and fabrication. However, many other practices can fall into the category of research misconduct because they deviate from ethical standards in research (Eshet et al., 2021). These include misrepresentation of data in publications, selectively reporting results, characterizing results with low power as unfavorable, improper use of funds, violations of safety protocols, gift authorship, conflicts of interest, and duplicate publications (Federman et al., 2003; Maggio et al., 2019; Haven and van Woudenberg, 2021). Although the aspects that characterize misconduct research are well known, the factors that influence misconduct in research have been less intensively scrutinized.

Factors of Misconduct in Research

There is a growing interest in research regarding the factors that enable misconduct in research, and we argue that academic stress, procrastination, and academic achievement could all be relevant aspects in the construct validity of the research misconduct scale.

Academic stress can lead a student to commit misconduct in research, according to a qualitative study by Devlin and Gray (2007) on why university students plagiarize. Academic

and external pressures were cited as reasons for committing this type of misconduct. In this era of heated academic competition, students feel pressure and complain that the workload placed on them by teachers is difficult to manage and argue that stress due to academic workload could be the cause of research misconduct (Khadem-Rezaiyan and Dadgarmoghaddam, 2017). This happens mainly with university students; because teachers want to make the most of their time and do their best for their students, they often assign maximum tasks, research projects, and assignments. This can make it quite challenging for students to complete all tasks with maximum proficiency and details; sometimes, they are unable to meet all their deadlines. Recruiting participants for research is difficult, and a scarcity of participants can induce students to indulge in unethical—or even illegal—means to complete their assignments. The extreme stress of anticipated failure compels them to conduct fake interviews, complete falsified forms, collect fake data, and finally fabricate and plagiarize data. It is evident that academic stress can play a significant role in influencing academic procrastination among students.

Academic procrastination refers to staying away from academic duties for as long as possible, which can cause students to fail to meet their academic requirements (Ferrari et al., 1995). Several studies have shown that students who demonstrate a careless academic attitude face a variety of negative effects of procrastination (Kandemir, 2010). This kind of educational carelessness inevitably leads to adverse outcomes, such as failing exams (Ferrari et al., 1995; Knaus, 1998), falling behind the rest of the class (Rothblum et al., 1986), and skipping classes and dropping out of school (Knaus, 1998). Ferrari (2004) found that postponing starting or completing an academic task is one of the primary characteristics of academic procrastination, regardless of the student's intention of ultimately doing the work. Because students who procrastinate begin to work later than those who do not procrastinate, they run out of time to complete work, even something as important as a thesis, before the deadline (Schouwenburg and Groenewoud, 2001). Procrastinating behavior leaves students in a situation where they find themselves out of time and resources. We can assume that students might find it easier to plagiarize, fabricate, or falsify data than do actual work and thus commit research misconduct. Procrastination can have a significant impact on students' lives, as it can result in low grades that impact various aspects of life and have objectively negative outcomes like poor academic performance (Hussain and Sultan, 2010). In terms of research, procrastination can be classified as a negative attitude. Research needs to be performed through proper planning, design, and investigation rather than being rushed. It demands the investment of adequate time and effort to achieve the best and most accurate outcomes. With continued procrastination, students are left only with the option of faking or plagiarizing by simply fabricating or copying and pasting others' research results. Plagiarism, which is one of the most common forms of research misconduct, has been highly correlated to procrastination among university students (Siaputra, 2013).

Regarding academic achievement and research misconduct, Finn and Frone (2004) found that students with poor academic performance were likely to commit academic misconduct, including plagiarism, fabrication, and falsification of data. Other research has shown that students with high CGPAs or good academic achievements were less likely to commit plagiarism (Guo, 2011).

Research Misconduct Assessment

As to assessing research misconduct, only a few scales measuring research misconduct among students or researchers are available, such as the academic dishonesty scale (Bolin, 2004) and a scale examining the perceptions of research coordinators who manage clinical trials regarding different perspectives on misconduct (Broome et al., 2005).

The academic dishonesty scale (Bolin, 2004) is composed of nine items describing behaviors like "copied material and turned it in as your own work," "used unfair methods to learn what was on a test before it was given," "copied a few sentences of material from a published source without giving the author credit," and "cheated on a test in any way." The academic dishonesty scale is not focused solely on research but also considers other behaviors that could involve students' academic tasks.

The different perspectives of the misconduct questionnaire (Broome et al., 2005) were used as part of a broader study aimed at analyzing scientific misconduct from a research supervisor's or coordinator's point of view. However, it does not measure the tendency for deliberately committing or slowly becoming involved in research misconduct. Previously, information was collected by directly asking scientists if they were involved in any kind of research misconduct (i.e., fabrication, falsification, and plagiarism) in surveys or interviews. Questions regarding primary forms of misconduct were asked, ignoring minor details like authorship credit details, data screening, and violation of protocols (Greenberg and Goldberg, 1994; Martinson et al., 2005). To the best of our knowledge, there are currently no instruments for measuring the level of research misconduct, and the present study intends to fill this gap by developing a tool for that purpose.

MATERIALS AND METHODS

This study aims to construct and validate a measure of research misconduct for students, including both quantitative and qualitative research misconduct, although they are different in nature. This paper reports on three studies.

(1) Study I focuses on the construction of a scale in three phases (generation of item pool, item cleaning, and exploration of factor structure). In Phase I, the initial pool of items was generated by reviewing the literature and considering the results of semi-structured interviews. Phase II involved psychometric cleaning of items, after which 38 items were retained. In Phase III, the items were proposed to 652 university students, and an exploratory factor analysis (EFA) was calculated.
(2) In Study II, the factorial structure was tested through a confirmatory factor analysis (CFA) on data collected from an independent sample of 200 university students.

(3) Study III includes validation (convergent validity) of the scale through the administration of questionnaires to a sample of 200 university students. The correlation between research misconduct and the following tools were used to provide evidence for convergent validity: academic procrastination scale (short form), academic stress scale, and academic achievements as measured by cumulative grade point average (CGPA).

The institutional review board of the University of Sargodha gave ethical approval for the research. After that approval was granted, a letter was submitted to the head of the Department of Psychology asking for permission to gather data from students. Once that permission was obtained, questionnaires were presented to the participants, who were assured that the data collected would be used solely for research purposes and that their identities and personal information would be kept confidential. Participants were approached personally and given detailed instructions regarding the purpose of the study and how to complete the questionnaires. Informed consent was obtained from the participants before data collection. Participants were thanked for their support in the research and were provided contact details if they wanted to obtain any further information about the research. Data collection began in September 2021 and finished in November 2021.

Study I: Development of Research Misconduct Scale

Study I consists of the construction of the scale and has three phases. The first phase identified the pool of items for the research misconduct scale, the second involved a psychometric cleaning of items, and the third involved EFA and psychometric properties.

Phase I: Initial Item Pool for Research Misconduct Scale

An initial pool of items was generated using empirical and deductive approaches. With the literature review in mind, items for the research misconduct scale were generated in English. All available literature related to research misconduct was reviewed, giving access to a wide array of concepts and ideas of research misconduct. Qualitative, unstructured individual interviews were also carried out by the researchers to expand their knowledge and obtain subjective viewpoints about research misconduct. Item pool generation was completed by using the following steps:

1. Literature-based: different domains of research misconduct were analyzed to identify aspects that could be used in a single psychometric measure of research misconduct and might provide a quantitative score of research misconduct as a whole.
2. To obtain an understanding of research misconduct and generate additional items for the scale, detailed unstructured interviews were carried out with professors (n = 20) and students (n = 50) from the University of Sargodha in Punjab. Interview participants were assured of the confidentiality of any information they provided. After they provided signed informed consent, they were asked to report the type of student misconduct incidents they regularly face, different examples of research misconduct, and unique cases of research misconduct. Students enrolled in MPhil and BS (Hons) programs were interviewed regarding the types of misconduct they had committed or observed in their peers. In addition, students were asked to report any factors that they thought might compel them to indulge in such acts.

Phase II: Psychometric Cleaning of Items

1. The initial item pool produced from the literature review and interviews yielded 40 items. After the initial item pool was generated, experts (n = 10) in psychology and psychometry offered reviews and opinions of the suitability of each item for the research misconduct scale. These expert opinions ensured the relevance and applicability of items to the target population. Information about the purpose of the scale was provided to the experts, who individually analyzed whether the items were culturally and contextually relevant and suggested any additions to or elimination of items in the scale. In response to the experts' views, 38 items were retained as best fitting the literature, cultural context, and target population.
2. That final scale of 38 items used a five-point Likert-type approach (1 = strongly disagree to 5 = strongly agree). The five-point response format was chosen based on its property of maintaining a balance between both poles while providing respondents with more freedom to choose from the response range that best depicted their views (Gregory, 2004).

Phase III: Exploratory Factor Analysis and Psychometric Properties

Participants

A sample of 652 university students (300 male, 352 female) belonging to social sciences departments was recruited for the study through a convenience sampling technique. Only full-time students with at least one research experience were included. Participant age ranged from 20 to 24 (M = 21.5, SD = 5.12).

Results

To determine the final structure of the scale, an EFA was carried out on the sample of 652 participants through principal axis factoring and the direct oblimin method. This rotation method was used based on the assumption that if more factors were yielded, they would share some sort of covariance (Field, 2013). A single factor was clearly obtained, accounting for a substantial amount of variance (55.73%) as you can see in **Table 1**.

The Kaiser-Meier-Olkin (KMO) test and Bartlett's Test of Sphericity were applied to assess whether the sample was adequate. The KMO value was 0.92, showing perfect sample sufficiency and adequacy (Kaiser, 1974). Bartlett's Test of Sphericity was also significant, indicating that the items are significantly correlated and that the sample is appropriate for further analysis (Field, 2013). According to Coakes and Steed (2003), factor analysis is quite sensitive to assumptions of normality. Therefore, skewness and kurtosis were calculated to

TABLE 1 | Factor loadings through principal axis factoring for the research misconduct scale (N = 652).

	Standardized factor loadings
Item no.	**F1**
1	0.71
2	0.73
3	0.72
4	0.75
5	0.71
6	0.72
7	0.81
8	0.73
9	0.73
10	0.78
11	0.72
12	0.76
13	0.70
14	0.72
15	0.78
Eigenvalue: 9.91	
% of variance: 55.73	

assess the normality of the data; good normality was obtained. All the commonalities were considerably high, suggesting that factor analysis could proceed; thus, all variables were selected for further analysis.

The EFA yielded a one-factor solution with a direct oblimin rotation method and eigenvalues > 1.0. The obtained one-factor structure was well defined and interpretable with theoretical reliability and construct relevance. Of 38 items, 15 were retained for their substantial loadings (≥0.70) on a single factor. A single-factor structure was interpreted as satisfactory factor loading and theoretical relevance of all items to the factor.

The following 15 items showed exclusive loading on a single factor: "faking," "cheating," "misconducting," "manipulating," "plagiarism," "fabricating," and "favored authorship" were the hallmarks of the obtained factor. The scale's Cronbach's alpha was computed; the value of 0.95 indicated very good reliability and excellent internal consistency.

Study II: Confirmatory Factor Analysis

Participants

A sample of 200 university students (101 male, 99 female) was recruited through a convenience sampling technique, using the same affiliation and criteria as applied in Study I.

Results

Based on the initial criteria (i.e., item loading > 0.70), the model obtained through EFA was analyzed *via* CFA; the factor structure obtained showed an excellent fit with the data. The goodness of fit (GFI) value and the comparative fit index (CFI) are fairly close to one, and the value of root mean square error of approximation (RMSEA) is significantly close to zero, indicating a good model fit. The value of chi-square/df is 2.91; as that is less than three, it is considered good (Hatcher and Stepanski, 1994). The final model obtained through CFA consisted of 15 items and presented a good model fit.

Table 2 reports the final model obtained through CFA; factor loadings ranged from 0.66 to 0.79. **Figure 1** shows the standardized factor loadings in the CFA.

Study III: Convergent and Discriminant Validation of Research Misconduct Scale for University Students

Study III was designed to verify the validity for the research misconduct scale for university students. To provide evidence for this validity, the study tested the following hypotheses:

- H1: A positive relationship between research misconduct and academic procrastination would provide evidence of convergent validity.
- H2: A positive relationship between research misconduct and academic stress would provide evidence of convergent validity.
- H3: A negative relationship between research misconduct and CGPA would provide evidence of convergent validity.

Participants

A sample of 200 university students (100 male, 100 female) was recruited through a convenience sampling technique, with the same affiliation and criteria as applied in Study I.

Instruments

Research Misconduct Scale

The research misconduct scale is a 15-item self-report measure (see **Appendix 1**) developed to examine research misconduct in university students. The response format of the scale is a five-point Likert-type format (1 = strongly disagree, 2 = disagree, 3 = neutral, 4 = agree, and 5 = strongly agree). A high score on the scale represents a high level of research misconduct, while low scores represent low levels of research misconduct. There are no reverse scored items on the scale, which is comprised of only one factor. The Cronbach's alpha of the reliability index of the scale is 0.95.

Academic Procrastination Scale (Short Form)

A short form of the academic procrastination scale (Yockey, 2016), which consists of five items, was used in the present study. The response format of the scale is a five-point Likert-type format (1 = disagree, 5 = agree). There are no reverse scored items on the scale. A high score on this scale represents a high level of procrastination, while low scores represent low levels of procrastination. The scale shows good internal consistency, with a Cronbach's alpha of 0.87.

Academic Stress Scale

Developed by Lin and Chen (2009), the academic stress scale consists of 34 items with responses using a five-point Likert-type

TABLE 2 | Model fit indices of CFA for research misconduct scale for university students (N = 200).

Indexes	Chi-square	*df*	Chi-square/df	CFI	RMSEA	GFI	TLI
Model	245.26	84	2.91	0.92	0.05	0.90	0.91

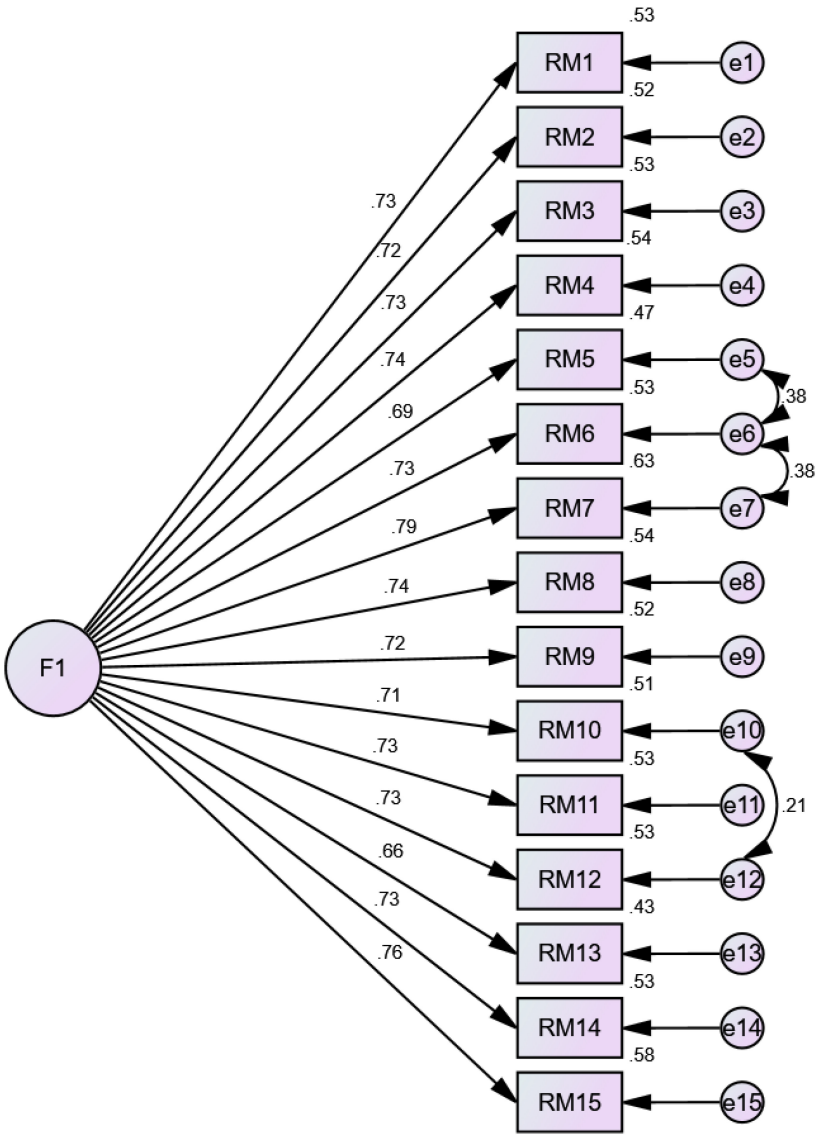


FIGURE 1 | Standardized factor loadings in the CFA of the research misconduct scale.

format (1 = strongly disagree, 2 = disagree, 3 = do not know, 4 = agree, 5 = strongly agree). High scores on this scale represent high levels of academic stress, and low scores represent low levels of academic stress. There are no reverse scored items on this scale; its Cronbach's alpha is 0.90, showing good internal consistency.

Academic Achievement as Measured by Cumulative Grade Point Average

Student CGPAs were taken as evidence for the convergent validity of the study; CGPAs range between 0.00 and 4.00. Students were asked to provide information about their CGPAs in the previous semester.

Results

Table 3 shows correlations between the research misconduct scale, academic stress scale, academic procrastination scale, and CGPA. The research misconduct scale has a significant positive correlation with the academic stress scale ($r = 0.74$, $p < 0.01$) and the academic procrastination scale ($r = 0.58$, $p < 0.01$) but a significant negative correlation with CGPA ($r = -0.38$, $p < 0.01$). The academic stress scale has a significant positive relation with the academic procrastination scale ($r = 0.67$, $p < 0.01$) and a significant negative correlation with CGPA ($r = -0.25$, $p < 0.05$). Finally, the academic procrastination scale has a significant negative correlation with CGPA ($r = -0.22$, $p < 0.05$).

DISCUSSION

The present study has illustrated the steps for developing a 15-item self-report measure for research misconduct among students. The items on the scale are general and related to

TABLE 3 | Correlation of research misconduct scale with academic stress scale, academic procrastination scale, and CGPA (N = 200).

	RMS	ASS	APS	CGPA
RMS	–	0.74**	0.58**	−0.38**
ASS		–	0.67**	−0.25*
APS			–	−0.22*
CGPA				–

RMS, Research misconduct scale; ASS, Academic stress scale; APS, Academic procrastination scale; CGPA, Cumulative grade point average.
$^{**}p < 0.01$; $^{*}p < 0.05$.

various processes and ethical issues involved in conducting research. The EFA highlighted that the scale is unidimensional and that one factor explained 55.73% of the total variance, while CFA confirmed the one-factor structure obtained through EFA and showed that the model fits for the data and alpha reliability were 0.95, indicating excellent internal consistency (Biasutti and Frate, 2017, 2018).

The research misconduct scale has been validated by testing correlations with the academic stress scale, the academic procrastination scale, and CGPA. Research misconduct had a significant positive correlation with academic stress, and academic procrastination, and a negative correlation with CGPA.

Regarding hypothesis one ("A positive relationship between research misconduct and academic procrastination would provide evidence of convergent validity"), it was assumed that a positive correlation of academic procrastination with research misconduct would provide evidence of convergent validity. The present study's findings support hypothesis one because research misconduct had a significant positive correlation with academic procrastination. Most of the time, students procrastinate on their academic tasks while prioritizing non-academic activities. Procrastinating behaviors place students in situations where they find themselves out of time and resources; for some, the only solution appears to be inappropriate behaviors like faking and plagiarizing. Plagiarism is one of the most common forms of research misconduct, with several studies (e.g., Siaputra, 2013) finding a high correlation between research misconduct and procrastination in university students. The findings of the present study are in line with another study that suggested a significant positive correlation between plagiarism and academic procrastination (Roig and DeTommaso, 1995), as well as with a panel study of German university students, which revealed that higher levels of academic procrastination results were connected to higher levels of plagiarism, falsification, and data fabrication (Patrzek et al., 2014).

Concerning hypothesis two ("A positive relationship between research misconduct and academic stress would provide evidence of convergent validity"), the fact that research misconduct had a significant positive correlation with academic stress is unsurprising in today's academic environment (Devlin and Gray, 2007). Keeping in mind the role of academic stress in research misconduct, it was argued that a positive correlation of academic stress with research misconduct would provide evidence of convergent validity. The analysis supports this second hypothesis of the study, which is in line with previous research in which university students noted that academic stress and pressure might be reasons for research misconduct (Khadem-Rezaiyan and Dadgarmoghaddam, 2017). In addition, the findings here are supported by a qualitative study on why university students plagiarize that found academic and external pressures to be reasons behind that type of misconduct (Devlin and Gray, 2007). Students engage in research misconduct when they encounter a task that is more demanding than their capabilities and skills, which places them under stress. Plagiarism is correlated with the difficulty and nature of students' tasks, which can also be considered connected to academic stress (Tindall and Curtis, 2020).

With regard to hypothesis three ("A negative relationship between research misconduct and CGPA would provide evidence of convergent validity"), it was conjectured that a negative correlation of CGPA with research misconduct would provide evidence of convergent validity. The study's findings supported the third hypothesis and revealed that research misconduct has a significant negative correlation with CGPA. Conducting research demands a mix of convergent and divergent abilities such as critical thinking, analyzing and comparing situations, planning, scheduling, comprehension, and creativity. Not everyone can think outside the box, which is crucial for designing research. These skills are also related to academic scores. High IQ is a determinant of high scores on certain kinds of tests, but university courses now focus on skills that go beyond merely cramming for exams. Knowing the practical applications of knowledge is more demanding than simply digesting an entire syllabus. Hence, the focus of exams is now more on applied principles of knowledge. Students who are not able to apply knowledge in practical terms do not score well. Low scorers also fail to conduct good research, as they do not have the prerequisite knowledge and skills for conducting good research. The findings of the present study are aligned with those reported by Comas-Forgas and Sureda-Negre (2010), who stated that students with low academic success are more likely to plagiarize. As plagiarism is one of the significant components of research misconduct, this evidence can be taken to support the hypothesis. Similarly, research has shown that students with higher CGPAs or good academic performance were less likely to commit plagiarism (Guo, 2011).

Limitations

The present study has certain limitations as to participants. The study sample was a convenience sample and consisted solely of students from one university in Punjab. The sample was not representative of the total student population, so the results cannot be generalized to all students. The students available for this study did not represent the actual percentages of students from different religions and races. In addition, only students aged 20 to 24 were included.

The techniques for scale refinement—EFA and CFA—that were used in the present study were specific to sample size, and it is advisable to confirm or refine the findings in further research using a larger sample. All the constructs of the present study were measured through self-report measures, which might have resulted in inflated correlation among the study's

variables. However, an inspection of the correlation matrix among the variables of the present study revealed that none of the correlations was too high, which reduces the likelihood of common method bias.

Educational Implications

Several educational implications of this study could be discussed. Regarding the positive correlation between research misconduct and academic stress, it could be suggested to university professors do not put students too much under pressure because this could generate wrongdoing behaviors. Prevent to generate stress could be a strategy to avoid misconduct behaviors in students. Other actions could be taken to discourage academic procrastination, which is in correlation with research misconduct. The variables of academic procrastination could be examined to identify factors to be controlled in the educational process of university students.

Future Research

Future studies should explore other potential correlations of research misconduct to expand its nomological network. Research misconduct could be studied in association with aspects such as personality traits and academic ethics. The research design used here was cross-sectional and does not provide any causal evidence. Therefore, investigating research misconduct using an experimental research design is suggested. Future research should verify the test-retest reliability of the research misconduct scale. For this purpose, a longitudinal research design should be adopted to assess the temporal stability of research misconduct as operationalized through the research misconduct scale. Future studies should investigate the criterion-related validity of the research misconduct scale by examining the concurrent validity in cross-sectional designs and the predictive validity in longitudinal designs. In addition, different populations of students could be involved, including doctoral students.

Applications of the Research Misconduct Scale

The research misconduct scale developed and validated in the present study opens new avenues of research. Often, research misconduct stays in the dark and is not reported to save an institution's integrity. This scale should be used in future studies to strengthen research integrity and the scientific community's ability to assess research misconduct and its various correlations. Furthermore, cross-cultural research on misconduct might help students evaluate the actual and perceived seriousness and consequences of research misconduct.

The research misconduct scale could be used in a triangulation approach for assessing the causes and consequences of research misconduct among various institutes. Both instructors' and students' perceptions of research misconduct could be analyzed. More specifically, the administration of the research misconduct scale in various universities might help supervisors, policymakers, curriculum developers, and administrators identify specific factors related to research misconduct and take appropriate actions to stop or at least reduce the practice.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

Ethical review and approval were not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

SG: conceptualization, writing original draft, and supervision. SK: methodology and formal analysis. ZH: formal analysis. MB: critical rewriting and supervision. All authors have read and agreed to the published version of the manuscript.

REFERENCES

Adeleye, O. A., and Adebamowo, C. A. (2012). Factors associated with research wrongdoing in Nigeria. *J. Empir. Res. Hum. Res. Ethics* 7, 15–24. doi: 10.1525/jer.2012.7.5.15

American Psychological Association [APA] (2019). *Research Misconduct*. Washington, DC: American Psychological Association.

Bak, H. J. (2018). Research misconduct in east asia's research environments. *East Asian Sci. Technol. Soc.* 12, 117–122. doi: 10.1215/18752160-6577620

Biasutti, M., and Frate, S. (2017). A validity and reliability study of the attitudes toward sustainable development scale. *Environ. Educ. Res.* 23, 214–230. doi: 10.1186/s12913-016-1423-5

Biasutti, M., and Frate, S. (2018). Group metacognition in online collaborative learning: validity and reliability of the group metacognition scale (GMS). *Educ. Technol. Res. Dev.* 66, 1321–1338. doi: 10.1007/s11423-018-9583-0

Bolin, A. U. (2004). Self-control, perceived opportunity, and attitudes as predictors of academic dishonesty. *J. Psychol.* 138, 101–114. doi: 10.3200/JRLP.138.2.101-114

Broome, M. E., Pryor, E., Habermann, B., Pulley, L., and Kincaid, H. (2005). The scientific misconduct questionnaire—revised (SMQ-R): validation and psychometric testing. *Account. Res.* 12, 263–280. doi: 10.1080/08989620500440253

Coakes, S. J., and Steed, L. G. (2003). Multiple Response And Multiple Dichotomy Analysis. SPSS: Analysis Without Anguish: Version 11.0 For Windows. Singapore: Wiley, 215–224.

Comas-Forgas, R., and Sureda-Negre, J. (2010). Academic plagiarism: explanatory factors from students' perspective. *J. Acad. Ethics* 8, 217–232. doi: 10.1007/s10805-010-9121-0

De Vries, R., Anderson, M. S., and Martinson, B. C. (2006). Normal misbehavior: scientists talk about the ethics of research. *J. Empir. Res. Hum. Res. Ethics* 1, 43–50. doi: 10.1525/jer.2006.1.1.43

Devlin, M., and Gray, K. (2007). In their own words: a qualitative study of the reasons Australian university students plagiarize. *High Educ. Res. Dev.* 26, 181–198. doi: 10.1080/07294360701310805

Eshet, Y., Steinberger, P., and Grinautsky, K. (2021). Relationship between statistics anxiety and academic dishonesty: a comparison between learning environments in social sciences. *Sustainability* 13:1564. doi: 10.3390/su13031564

Federman, D., Hanna, K., and Rodriguez, L. (2003). *Responsible Research*. Washington, DC: Institute of Medicine National Academies Press.

Ferrari, J. R. (2004). "Trait procrastination in academic settings: an overview of students who engage in task delays," in *Counseling The Procrastinator In Academic Settings*, eds H. C. Schowuenburg, C. Lay, T. A. Pychyl, and J. R. Ferrari (Washington, DC: American Psychological Association), 19–28. doi: 10.1037/10808-002

Ferrari, J. R., Johnson, J. L., and McCown, W. G. (1995). *Procrastination And Task Avoidance: Theory, Research, And Treatment*. New York, NY: Plenum Press.

Field, A. (2013). *Discovering Statistics Using IBM SPSS Statistics*. Thousand Oaks, CA: Sage.

Finn, K. V., and Frone, M. R. (2004). Academic performance and cheating: moderating role of school identification and self-efficacy. *J. Educ. Res.* 97, 115–121. doi: 10.3200/JOER.97.3.115-121

Greenberg, M., and Goldberg, L. (1994). Ethical challenges to risk scientists: an exploratory analysis of survey data. *Sci. Technol. Hum. Values* 19, 223–241. doi: 10.1177/016224399401900206

Gregory, R. J. (2004). *Psychological Testing: History, Principles, And Applications*. Boston, MA: Allyn and Bacon.

Guo, X. (2011). Understanding student plagiarism: an empirical study in accounting education. *Account. Educ.* 20, 17–37. doi: 10.1080/09639284.2010.534577

Hatcher, L., and Stepanski, E. J. (1994). *A Step-By-Step Approach To Using The SAS System For Univariate And Multivariate Statistics*. Cary, NC: SAS Institute.

Haven, T., and van Woudenberg, R. (2021). Explanations of research misconduct, and how they hang together. *J. Gen. Philos. Sci.* 52, 543–561. doi: 10.1007/s10838-021-09555-5

Hussain, I., and Sultan, S. (2010). Analysis of procrastination among university students. *Proc. Soc. Behav. Sci.* 5, 1897–1904. doi: 10.1016/j.sbspro.2010.07.385

Jawad, F. (2013). Plagiarism and integrity in research. *JPMA J. Pakistan Med. Assoc.* 63, 1446–1447.

Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika* 39, 31–36. doi: 10.1007/BF02291575

Kandemir, M. (2010). Akademik erteleme davranıçını açıklayıcı bir model. *Pegem Eğitim Öğretim Dergisi* 4, 51–72.

Khadem-Rezaiyan, M., and Dadgarmoghaddam, M. (2017). Research misconduct: a report from a developing country. *Iran. J. Public Health* 46:1374.

Knaus, W. J. (1998). *Do It Now! Break The Procrastination Habit*. New York, NY: John Wiley and Sons.

Krimsky, S. (2007). When conflict-of-interest is a factor in scientific misconduct. *Med. Law* 26:447.

Lin, Y. M., and Chen, F. S. (2009). Academic stress inventory of students at universities and colleges of technology. *World Trans. Eng. Technol. Educ.* 7, 157–162.

Maggio, L., Dong, T., Driessen, E., and Artino, A. Jr. (2019). Factors associated with scientific misconduct and questionable research practices in health professions education. *Perspect. Med. Educ.* 8, 74–82. doi: 10.1007/s40037-019-0501-x

Martinson, B. C., Anderson, M. S., and De Vries, R. (2005). Scientists behaving badly. *Nature* 435:737. doi: 10.1038/435737a

Office of Research Integrity (2001). *Research Misconduct*. Washington, DC: Office of Research Integrity.

Office of Research Integrity (2019). *Definition of Research Misconduct*. Washington, DC: Office of Research Integrity.

Okonta, P., Rossouw, T. (2013). Prevalence of scientific misconduct among a group of researchers in Nigeria. *Dev. World Bioethics* 13, 149–157. doi: 10.1111/j.1471-8847.2012.00339.x

Patrzek, J., Sattler, S., van Veen, F., Grunschel, C., and Fries, S. (2014). Investigating the effect of academic procrastination on the frequency and variety of academic misconduct: a panel study. *Stud. High. Educ.* 40, 1014–1102. doi: 10.1080/03075079.2013.854765

Peled, Y., Eshet, Y., Barczyk, C., and Grinautski, K. (2019). Predictors of academic dishonesty among undergraduate students in online and face-to-face courses. *Comput. Educ.* 131, 49–59. doi: 10.1016/j.compedu.2018.05.012

Roig, M., and DeTommaso, L. (1995). Are college cheating and plagiarism related to academic procrastination? *Psychol. Rep.* 77, 691–698. doi: 10.2466/pr0.1995.77.2.691

Rothblum, E. D., Solomon, L. J., and Murakami, J. (1986). Affective, cognitive, and behavioral differences between high and low procrastinators. *J. Couns. Psychol.* 33, 387–394. doi: 10.1037/0022-0167.33.4.387

Schouwenburg, H. C., and Groenewoud, J. T. (2001). Study motivation under social temptation: effects of trait procrastination. *Pers. Individ. Diff.* 30, 229–240. doi: 10.1016/S0191-8869(00)00034-9

Siaputra, I. B. (2013). The 4PA of plagiarism: a psycho-academic profile of plagiarists. *Int. J. Educ. Integrity* 9, 50–55. doi: 10.21913/IJEI.v9i2.892

Steen, R. G. (2011). Retractions in the scientific literature: is the incidence of research fraud increasing? *J. Med. Ethics* 37, 249–253. doi: 10.1136/jme.2010.040923

Steneck, N. H. (2006). Fostering integrity in research: definitions, current knowledge, and future directions. *Sci. Eng. Ethics* 12, 53–74. doi: 10.1007/PL00022268

Stretton, S., Bramich, N. J., Keys, J. R., Monk, J. A., Ely, J. A., Haley, C., et al. (2012). Publication misconduct and plagiarism retractions: a systematic, retrospective study. *Curr. Med. Res. Opin.* 28, 1575–1583. doi: 10.1185/03007995.2012.728131

The Pennsylvania State University (2018). *Falsification, Fabrication, Plagiarism*. State College, PA: The Pennsylvania State University.

Tindall, I. K., and Curtis, G. (2020). Negative emotionality predicts attitudes toward plagiarism. *J. Acad. Ethics* 18, 89–102. doi: 10.1007/s10805-019-09343-3

Tsai, D. (2018). Reflections on the research misconduct cases in east asia. *East Asian Sci. Technol. Soc.* 12, 181–184. doi: 10.1215/18752160-6577762

World Association of Medical Editors (2019). *Publication Ethics Policies For Medical Journals*. Available online at: https://wame.org/recommendations-on-publication-ethics-policies-for-medical-journals (accessed April 10, 2022).

Yamamoto, K., and Lennon, M. L. (2018). Understanding and detecting data fabrication in large-scale assessments. *Qual. Assurance Educ.* 26, 196–212. doi: 10.1108/qae-07-2017-0038

Yockey, R. D. (2016). Validation of the short form of the academic procrastination scale. *Psychol. Rep.* 118, 171–179. doi: 10.1177/0033294115626825

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

APPENDIX

Appendix 1 | The research misconduct scale for students.
Please indicate the extent of your agreement or disagreement with the statements by using the following scale:

	1	2	3	4	5
	Strongly disagree	**Disagree**	**Neutral**	**Agree**	**Strongly agree**
(1) Rather than putting in effort, I would prefer to pay someone to get my research projects or some of their parts done	1	2	3	4	5
(2) I have reported fake research studies in my research project	1	2	3	4	5
(3) To please someone, I might add their name as an author without a significant contribution to my research	1	2	3	4	5
(4) If collecting data from a sample is difficult (challenging population, poor response rate, etc.), I might complete data collection tools (questionnaires, interviews, checklists, etc.) myself	1	2	3	4	5
(5) If an author is taking too much time to respond, I might be compelled to use the research tool without his or her permission	1	2	3	4	5
(6) Performing the main analysis without checking its assumptions is not misconduct	1	2	3	4	5
(7) If the results do not turn out as expected, I might manipulate data to send it in my desired direction	1	2	3	4	5
(8) I am not particularly concerned about keeping research information (demographics, responses, etc.) confidential	1	2	3	4	5
(9) If there was no risk of being caught, I would not mind claiming someone else's work as my own	1	2	3	4	5
(10) It is fine to report high reliability even if it is actually lower than is required in my research	1	2	3	4	5
(11) I have reported non-significant findings as significant ones	1	2	3	4	5
(12) I have manipulated demographics to balance the ratio between groups in my research	1	2	3	4	5
(13) In my opinion, adding fake references in research is not misconduct	1	2	3	4	5
(14) If there is no fear of being caught, I might easily report false results in my research	1	2	3	4	5
(15) I have mixed original and fake data (questionnaires, interviews, documented records, etc.) during data collection in my research	1	2	3	4	5

frontiers | Frontiers in Education

BRIEF RESEARCH REPORT
published: 09 June 2022
doi: 10.3389/feduc.2022.757012

Building Student Entrepreneurship Activities Through the Synergy of the University Entrepreneurship Ecosystem

Eriana Astuty, Okky Rizkia Yustian† and Chyntia Ika Ratnapuri†*

Entrepreneurship Department, BINUS Business School Undergraduate Program, Bina Nusantara University, Jakarta, Indonesia

Student entrepreneurship activities can be a driving force for the emergence of young entrepreneurs. Therefore, universities are making efforts to equip their students with the requisite entrepreneurial knowledge and skills for a conducive university entrepreneurial ecosystem. The present study employs a quantitative approach and survey-type research. The research method uses the explanatory method with research objects, including the internal environment of the institution, external environmental support, student entrepreneurial orientation, student entrepreneurial intentions, and student entrepreneurial activities. Data were collected through online questionnaires, which were randomly distributed to 456 students of 7 state universities and 11 private universities across Java and Sumatra, Indonesia. Descriptive and multivariate data analyses with a structural equation model was carried out using the IBM SPSS Amos 20.0 software. The study has propounded a research novelty called Entrepreneurship Eclectic Education, which combines several techniques, designs, and methods that have been proven valid, reliable, and feasible for adoption in universities. Such novelty is likely to trigger student performance in their entrepreneurial activities in the university's entrepreneurial ecosystem. This is realized through a synergy between the internal and external environment of the institution that can foster an entrepreneurial orientation and then trigger students' entrepreneurial intentions, which leads to the creation of student entrepreneurial activities. This study offers valuable recommendations for higher education decision-makers to re-orient the entrepreneurship curriculum and create a conducive university entrepreneurship ecosystem.

Keywords: entrepreneurship ecosystem, institution environment, entrepreneurship activity, entrepreneurship intention, entrepreneurship orientation

OPEN ACCESS

Edited by:
George Waddell, Royal College of Music, United Kingdom

Reviewed by:
Karin Širec, University of Maribor, Slovenia
Guido Baltes, Hochschule Konstanz University of Applied Sciences, Germany

****Correspondence:***
Eriana Astuty
eriana.astuty@binus.ac.id

†These authors have contributed equally to this work and share last authorship

Specialty section:
This article was submitted to Assessment, Testing and Applied Measurement, a section of the journal Frontiers in Education

Received: *11 August 2021*
Accepted: *20 April 2022*
Published: *09 June 2022*

Citation:
Astuty E, Yustian OR and Ratnapuri CI (2022) Building Student Entrepreneurship Activities Through the Synergy of the University Entrepreneurship Ecosystem. Front. Educ. 7:757012. doi: 10.3389/feduc.2022.757012

INTRODUCTION

The World Economic Forum states that higher education is the fifth pillar of the 12 pillars supporting a country's global competitiveness index through its role as an efficiency enhancer that increases its productivity and long-term prosperity (Schwab, 2018). As one of the factors driving the nation's economic growth, universities must have policies to create a higher education environment that thoughtfully and comprehensively supports entrepreneurial activities. Previous studies confirmed that entrepreneurial activity is positively correlated with economic growth (Galindo-Martín et al., 2019). For this reason, universities need to develop and encourage potential

in young entrepreneurs while they are still in college through student entrepreneurial activities. It is believed that the entrepreneurial activities that students undertake during college will encourage them to be future entrepreneurs (Kourilsky and Walstad, 1998). It is also believed that they are more aware of the latest technology, market trends, and latest product ideas, are more friendly and sociable, and have more energy and high enthusiasm to begin and engage in new ventures at their age (Bosma et al., 2020). Entrepreneurial activities carried out by students are confirmed to result from simultaneous interaction between social values and individual attributes (Bosma et al., 2020), where the university environment and support from communities, societies, or organizations are considered social values. Individual attributes include orientation, intention, and motivation of students to become entrepreneurs, to create job opportunities for others and add values to their business (Bosma et al., 2020). In addition, previous research has confirmed that many things can improve the entrepreneurial skills of students, including entrepreneurship education (Odewale et al., 2019), support from external practitioners such as existing entrepreneurs (Bazan et al., 2019), and a university ecosystem that can have a positive influence on student entrepreneurial activities, although not significantly (Bazan et al., 2019).

The trend in Indonesia shows that only 16% of university graduates become entrepreneurs (BPS Indonesia, 2016). As of early 2020, the ratio of Indonesian national entrepreneurs was only 3.47% of the total population, which was still below the entrepreneurial ratio of neighboring countries, such as Malaysia, 4.74%, Thailand, 4.26%, and Singapore, 8.76% (Indonesian Ministry of Cooperatives and MSMEs, 2020). Therefore, the micro condition of these universities has a significant impact on Indonesia's macro needs, especially the ratio of Indonesian national entrepreneurs. Analysis of previous research regarding student entrepreneurial activities linking it to the conditions in Indonesian universities led the authors to formulate the main problem in this study, namely the low ability of universities to produce quality graduates capable of entrepreneurship after completing their studies.

Based on the research focus, this study aimed to obtain the best-fit model that could serve as an empirical basis to answer the research questions. In addition, it was expected that the results of this study could: (a) identify the most dominant indicator that contributes to each research variable, (b) analyze the relationship and the magnitude of influence between each variable, and (c) analyze the role of the mediator variable that is entered into the model. It was imperative to obtain empirical evidence from a higher education ecosystem in developing countries, such as Indonesia, which encourage student entrepreneurial activities.

LITERATURE REVIEW

Social Values Shaping Student Entrepreneurial Activities

Institution Internal Environment, External Environment Support

The internal environment of the institution that interacts with external parties in conducting entrepreneurship education, entrepreneurship research, and the development of the university's entrepreneurship programs can motivate students' intention to become entrepreneurs (Walter et al., 2013). An effective interaction synergizing internal strategies with the surrounding environment is purportedly able to strengthen entrepreneurship, particularly the performance of the entrepreneurs (Lumpkin and Dess, 1996). This interaction is also able to generate strong support for the education system within the institution so that the perception of potential obstacles in entrepreneurial activities is reduced (Mehtap et al., 2017), which in the end can provide learning to improve students' work abilities in the future (Knight and Yorke, 2003).

Disruptive innovations that are changing the organizational structure and current market conditions increasingly require universities to prepare an entrepreneurial ecosystem that can equip students with entrepreneurial competencies and skills to face the current demands (Hulme et al., 2014; Kuratko and Morris, 2018). Effective education and training could improve entrepreneurial skills to spark students' entrepreneurial intentions (Gieure et al., 2019). Entrepreneurship is based on complex learning, so universities need to formulate a comprehensive learning strategy to stimulate students' interests (Knight and Yorke, 2003). Entrepreneurship education is not a topic restricted just to the business but also includes a complex set of attitudes, beliefs, skills, and qualities. It is likely to have a positive impact if other relevant disciplines are linked to the entrepreneurship sub-topic as part of the curriculum. This would provide the program with a scientific context, enhance career relevance (Bridgstock, 2013), and facilitate finding solutions to complex problems (McDonald et al., 2018). Entrepreneurship is also a dynamic study with a history of being related to many variables such as knowledge and experience (Gupta et al., 2016), being very contextual (Thomassen et al., 2019), and comprehensive in nature (Hägg and Kurczewska, 2019). Entrepreneurship's proven ability to improve students' critical thinking skills (Ratten and Usmanij, 2020) has been presented in various forms of learning (Thomassen et al., 2019).

Since it has been proven to positively impact the experience of staff and students (Brown, 2018), universities could accommodate this need to re-orient their entrepreneurship curriculum by facilitating mentorship by practitioners (Baluku et al., 2019; Williams Middleton et al., 2019). Other research findings confirm that facilitating business incubators could teach students the need to recognize, adapt, and appreciate the tensions/dynamics of the business environment (Ollila and Williams-Middleton, 2011). The work-based entrepreneurship learning programs improve the student's business learning experience in all relevant disciplines (Gibson and Tavlaridis, 2018), which, in turn, could increase the student's business setup (Galvão et al., 2020). In addition, other researchers confirm that entrepreneurship education in line with local wisdom is a must. Therefore, entrepreneurship educators should always keep "regional attachments in mind" when developing and implementing entrepreneurship learning programs at their universities (Franco et al., 2010). Viewed from the external environment of the institution, support for providing training services, guidance, and financial support for business startups in an incubator is a significant issue that

needs to be resolved urgently (Fong, 2020). Support from outside the institution in the form of investment activities for student entrepreneurial activities is proven to directly relate to the student's entrepreneurship (Zabelina et al., 2019).

Based on the literature review that was used as a reference in this study, the authors compiled six hypotheses as follows:

Hypothesis 1a: Institutional Internal Environment directly affects Student Entrepreneurial Activity.
Hypothesis 1b: Institutional Internal Environment directly affects Student Entrepreneurial Orientation.
Hypothesis 1c: Institutional Internal Environment directly affects Student Entrepreneurial Intention.
Hypothesis 2a: External Environment Support directly affects Student Entrepreneurial Activity.
Hypothesis 2b: External Environment Support directly affects Student Entrepreneurial Orientation.
Hypothesis 2c: External Environment Support directly affects Student Entrepreneurial Intention.

Individual Attributes Forming Student Entrepreneurship Activities

Student Entrepreneurship Orientation, Student Entrepreneurship Intention

Entrepreneurial performance has increased significantly with the strengthening of the entrepreneurial orientation (Lumpkin and Dess, 2001; Walter et al., 2006; Keh et al., 2007; Li et al., 2009; Bayarçelik and Özşahin, 2014; Emoke-Szidónia, 2015; Gupta et al., 2016; Zhang et al., 2016; Chavez et al., 2017); however, entrepreneurial orientation cannot directly improve performance. Therefore, we needed a model to capture and illustrate the relationship between entrepreneurial orientation and performance (Lumpkin and Dess, 1996; Lyon et al., 2000). Entrepreneurial orientation is one of the essential individual attributes observed in this study to monitor its impact on entrepreneurial performance.

Entrepreneurial intention is an individual attribute that is key to building the foundations of student entrepreneurial activity (SEA), but not every entrepreneurial intention can ultimately be realized intostarting and operating a new business (Shirokova et al., 2016). Many things affect the gap between entrepreneurial intentions and the outcome: a individual factors such as family entrepreneurial background, age, gender, and entrepreneurship education at the previous level of education, and the characteristics of the university environment, among others, where policymakers work in harmony with academicians in designing academic curricula to incorporate relevant theoretical elements with entrepreneurial practice (Iwu et al., 2019). Intentions are important outcomes of learning that are widely adopted across educational contexts. An increase in entrepreneurial intention is a desired outcome of entrepreneurship education (Nabi et al., 2017; Lavelle, 2019) because it is the first step in setting up a business (Sancho et al., 2020). Entrepreneurial intention is also shaped by various personality types and entrepreneurial self-efficacy (Sahin et al., 2019). Studies confirm that the greater the students' intention to become entrepreneurs, the higher their tendency to become nascent entrepreneurs and complete the tasks related to it (Souitaris et al., 2007). Against this background, the authors proposed the following hypotheses:

Hypothesis 3a. Student Entrepreneurial Orientation directly affects Student Entrepreneurial Activity.
Hypothesis 3b. Student Entrepreneurial Orientation directly affects Student Entrepreneurial Intention.
Hypothesis 4. Student Entrepreneurial Intention directly affects Student Entrepreneurial Activity.

As a result of reviewing previous studies, nine hypotheses were successfully constructed and formulated in the following conceptual model presented in **Figure 1**.

MATERIALS AND METHODS

Research Design

This study used a deductive approach with a survey research strategy on five research objects: internal institutional environment, external environment support, student entrepreneurial orientation, student entrepreneurial intention, and student entrepreneurial activity. The research subjects were universities in Indonesia with students as the unit of analysis. Data were collected using a cross-sectional time horizon with a web-based questionnaire as the research instrument. The data analysis technique used quantitative methods through descriptive analysis and explanatory analysis.

Population and Sample

There are 122 public universities in Indonesia spread across five major islands, namely Java, Sumatra, Kalimantan, Sulawesi, and Papua. The distribution of universities in Java and Sumatra is 63%, and the rest is spread over the three other islands (45). There are 3,171 private universities in Indonesia, with the largest distribution (75%) on the islands of Java and Sumatra and the rest spread over the islands of Kalimantan, Sulawesi, and Papua. Based on data published by the Indonesian Ministry of Research, Technology, and Higher Education in 2018, the largest number of public and private universities are located on the islands of Java and Sumatra. This was the determining factor for choosing the public and private universities located in Java and Sumatra for this study.

Seven public universities and 11 private universities in Java and Sumatra were chosen as samples, and the survey was conducted at these universities. The questionnaires were distributed within ±2 months with the help of several fellow lecturers who were permanent lecturers at the 18 sample universities. These colleagues then distributed the online questionnaires through social media accounts connected to several social media groups on campus, including social media groups of lecturers distributing the questionnaires and students on the campus. The online questionnaires were also distributed with the help of several colleagues who were members of lecturer associations, official lecturer forums, and several business incubators managed by lecturers at the sampled universities.

A total of 477 students agreed to participate as respondents in this study by filling out the research questionnaire. After

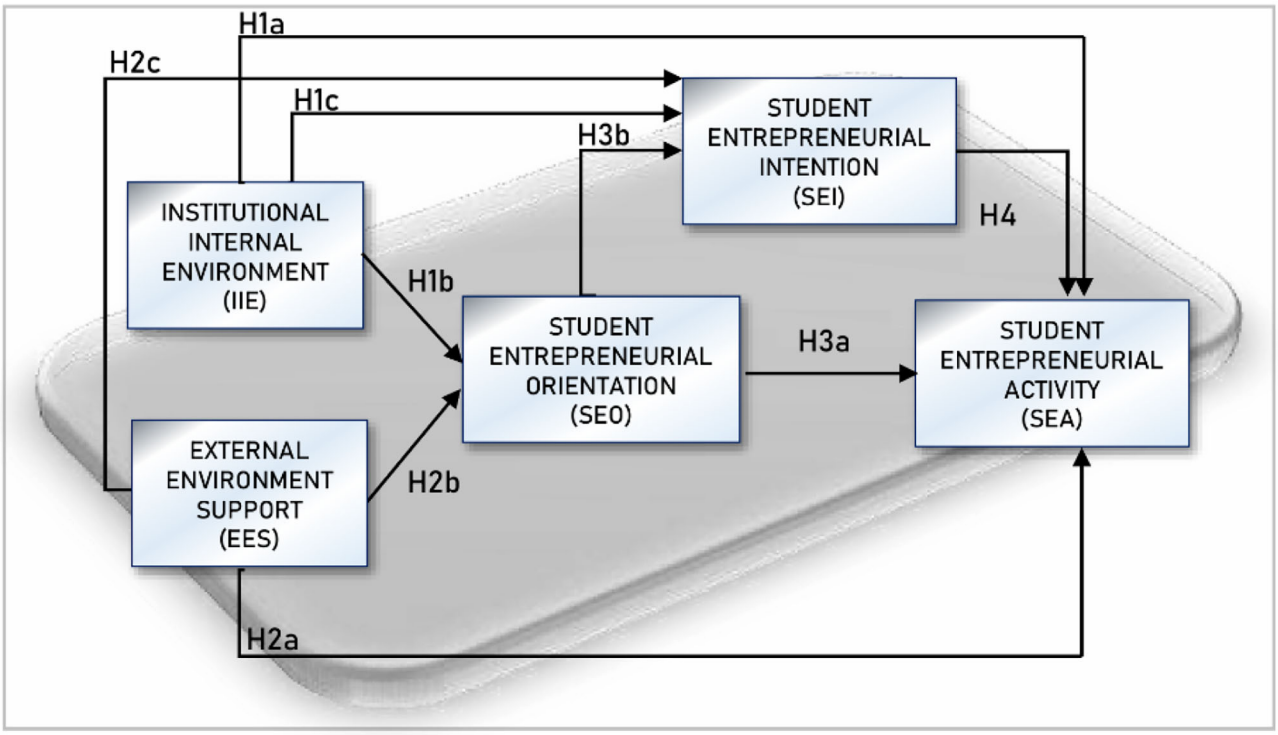


FIGURE 1 | Conceptual model.

checking for missing data, straight-line responses, and outliers, 456 respondents were deemed to be eligible for this study. The mean age of the respondents was 20 years, with a standard deviation of 1.47 years. A total of 41.2% of the respondents were men, 58.8% were women. Respondents included students enrolled in business study programs (51%) and non-business study programs (49%).

Demographically, of the 456 respondents, 39.9% did not have entrepreneurship education experience before entering college, 42.5% had received prior entrepreneurship education for 1–2 years, 10.5% for 3–4 years, and 7% for 5–6 years. Furthermore, 25.2% of the respondents had no entrepreneurial experience before entering college, 50% had entrepreneurial experience, 21.3% had been entrepreneurs to meet personal needs, and 3.5% had been entrepreneurs to meet family needs.

Variable Operationalization

Research Instrument

The research instrument used is a web-based questionnaire with question items arranged based on the operationalization of the variables in **Table 1**. The validity of this research instrument was measured using SPSS 20 by measuring the consistency of the correlation between item scores and the overall score on each research variable. In contrast, the correlation coefficient used was the Pearson correlation coefficient considering that the research data was interval scale and ratio scale. The instrument is "valid" if the significance level measured by the $p < 0.05$ (Hair et al., 2014) and "very valid" if the resulting p-value is much smaller than 0.05 (Hair et al., 2014). **Table 2** presents the results of the validity of the research instrument.

It can be seen that the research instrument was valid and mostly very valid because the resulting p-value was much smaller than 0.05 (Hair et al., 2014). Instrument reliability can be measured through three perspectives—stability, equivalence, and consistency (Cooper and Schindler, 2014). The instrument is (1) stable, if repeated measurements are made on the same person with the same instrument and still produces the same answer; (2) equivalent, if the level of variation in answers obtained from several different respondents is relatively low; and (3) Consistent, if the response given by the respondent shows a homogeneous answer. Reliability is determined when the value of the resulting reliability coefficient is >0.7 (Hair et al., 2014). In this study, the reliability coefficient value was measured using the Cronbach's Alpha value on SPSS20. **Table 3** shows the results of the reliability test of this research instrument.

Since the Cronbach's Alpha value of all variables was higher than 0.7, the instrument was reliable and could be used as a measuring tool for this research.

RESULTS

Descriptive Analysis

The respondents' replies to each question were distributed between the lowest value of one point and the highest value of five points. Based on their responses, the 456 respondents were grouped into three categories, namely "low" when the total

TABLE 1 | Variable operationalization.

Variable	Dimension	Indicator	Question items	Scale
The university's internal environment that can influence students' intentions to become entrepreneurs includes entrepreneurship education, entrepreneurship support programs, industrial ties, and research orientation (Walter et al., 2013)	Eclectic Entrepreneurship's Education (Pittaway and Edwards, 2012)	The "About" dimension (EEEA) emphasizes the practice of a pedagogic approach and usually didactic	EEEA1: The availability of courses on basic business knowledge, basic knowledge of entrepreneurship, or similar	Interval (5-point numeric scale)
			EEEA2: The availability of subject courses related to the concept of creativity, creative and innovative thinking, or equivalent	
			EEEA3: The availability of courses pertaining to entry in a business, startup, or similar courses	
		The "For" dimension (EEEF) concerns the involvement of students in assignments, activities, and projects that enable them to acquire the competencies needed for entrepreneurship	EEEF1: the availability of courses about entrepreneurial behavior, ethics, attitudes, and skills	
			EEEF2: The availability of learning opportunities aimed at increasing general entrepreneurial competencies such as business planning and other course packages in the form of case studies or projects	
			EEEF3: The availability of learning opportunities related to entrepreneurial competence through student participation in entrepreneurship competitions/business competitions, or similar courses	
		The "Through" dimension (EEET) concerns learning by doing, but in a "safe" mode, like in business incubators in universities, and internship programs for specific business units or companies	EEET1: The existence of a business incubator	
			EEET2: The presence of experiential learning through internship programs for various business/industry	
			EEET3: Experiential learning through collaboration with entrepreneurs	
			EEET4: Facilitation of students in business funding activities from outside parties	
			EEET5: Availability of mentoring programs from business actors to students	
			EEET6: Facilitation of student interaction activities with the market	
		The "Embedded" dimension (EEEE) is directly applied in a particular discipline so that it is relevant to the field of specialization and can generate intellectual property	EEEE1: The availability of embedding entrepreneurship courses in the curriculum of non-business study programs	
			EEEE2: The inclusion of business study students into non-business study programs to enroll in several classes according to their specialization and to further improve the diversity of skills needed by students	
	Research lecturers in the field of entrepreneurship (Walter et al., 2013)	Availability of student involvement in research activities of lecturers/researchers in entrepreneurship theme	ER1: Student involvement in lecturer research activities	
	Community service in entrepreneurial activities (Walter et al., 2013)	Availability of student involvement in community service activities related to entrepreneurship activities	ECS1: student involvement in community service activities	

(Continued)

TABLE 1 | Continued

Variable	Dimension	Indicator	Question items	Scale
Environmental factors that influence entrepreneurial intentions include economic/financial conditions, politics, social relations, technology, and cultural characteristics (Kristiansen, 2001, 2002)	Economics/ Financial	Capital support from outside the institution	EES1: Availability of capital support from the environment outside the institution, such as from entrepreneurs, financial institutions, investors, and others for students who will/are starting a business	Interval (5-point numeric scale)
	Information	Information Support from outside the institution	EES2: Availability of information from outside the institution regarding entrepreneurial activities, such as training, mentoring, and competitions related to entrepreneurship activities held by the government or certain organizations.	
	Technology	Technology support from outside the institution	EES3: There is support for training using certain technologies in the form of workshops, and others offered from outside the institution	
			EES4: There is technology procurement support offered from outside the institution	
	Networking	Business network support from outside the institution	EES5: Availability of easy access to community/network for business actors, for example, startup community, MSME community, and others	
	Social/Culture	Social/cultural life support from the local government which is oriented toward business development	EES6: Availability of events provided by the local government or non-institutional organizations to support entrepreneurial activities periodically, such as city bazaar activities, product exhibition events, and others	
Student Entrepreneurial Orientation (SEO): Characteristics related to views, tendencies, styles, methods, which reflect entrepreneurial behavior (Lumpkin and Dess, 1996, 2001)	Autonomy	Independent in expressing ideas/visions, able to turn ideas into concrete actions, making decisions, taking real actions even though they are constrained by resources, directing oneself in pursuing opportunities	SEOA1: I have independence in expressing ideas	Interval (5-point numeric scale)
			SEOA2: I have independence in decision making	
			SEOA3: I have the independence to take concrete action even though it is limited by resources	
			SEOA4: I have the independence in directing myself to pursue opportunities	
	Risk taking	Willingness to do things that are high risk for profit or opportunity	SEOR1: I am willing to do things that are high risk to get an opportunity	
			SEOR2: I am willing to do things that are high risk for profit	
	Proactiveness	Responsive in anticipating problems, responsive to change, responsive in seizing new opportunities	SEOP1: I am responsive in anticipating problems	
			SEOP2: I am responsive to change	
			SEOP3: I am responsive in seizing new opportunities	
	Competitive aggressiveness	Aggressive behavior to compete in improving the business position	SEOC1: I have aggressive behavior to compete to improve my business position	
	Innovativeness	The tendency of creative behavior in product/service development	SEOI1: I tend to be creative in developing products/services	

(Continued)

TABLE 1 | Continued

Variable	Dimension	Indicator	Question items	Scale
Student Entrepreneurial Intention (SEI): the basic foundation in the entrepreneurial process (Shirokova et al., 2016)		Preferred Priority	SEIPP: Becoming an entrepreneur is the main priority of my career choice after graduating from college	Interval (5-point numeric scale)
		Tend to Like	SEITL: I'd rather be an entrepreneur than an employee of a company	
		Think Seriously	SEITS: I'm seriously thinking about the things that need to be done to start a business	
		Determined	SEIDT: I have made up my mind to become an entrepreneur	
		Ready to Do (Bazan et al., 2019)	SEIRD: I am ready to do what it takes to be an entrepreneur	
Student Entrepreneurial Activity (SEA): activities that generate added value in an effort to create prosperity and economic equity (Bosma et al., 2020)	Student Entrepreneurial Activity By Phase (Bosma et al., 2020)	NPE: Non-Potential Entrepreneurs.	NPE: I have not started a business at all, and I am not familiar with the business opportunity, have no knowledge, and do not have the skills required to do business.	Ratio
		PE: Potential Entrepreneur.	PE: I have not started a business at all, but I am familiar with business opportunities, understand knowledge, and have the skills needed in doing business.	
		NE: Nascent Entrepreneur	NE: I have started a business but have not been able to pay salary for 3 months or more, including the founder	
		NBE: Owner/Manager of A New Business Entrepreneurs	NBE: I have started a business and have been able to pay salaries including to the founder for 3 months or more, but <42 months	
		EB: Owner-Manager of An Established Business:	EB: I am already running a business and have been able to pay wages for 42 months or more	
	Student Entrepreneurial Activity By Impact (Bosma et al., 2020)	Market Scope	The scope of the product/service sales area of the business being carried out: (1) No products/services have been sold yet; (2) Internal Campus; (3) City; (4) Province; (5) National	
		Innovation	SEAII: Product/service innovation growth that has been carried out during the business: (1) 0%; (2) 1–25%; (3) 26–50%; (4) 51–75%; (5) 76–100%; (6) >100%	

perception score on each question was between 456 and 1,064, "medium" between 1,065 and 1,672, and "high" between 1,673 and 2,280. The results were then averaged for each dimension and each variable. All the results of this descriptive statistical analysis are presented in **Tables 4–8**.

Measurement of the SEA by phase dimension showed that:

1) 11.6% of the respondents had not started a business at all and had no business knowledge/skills
2) 45.8% had not started a business at all but had business knowledge/skills
3) 32.5% had started a business but had not been able to pay salaries for 3 months or more, including for the founders
4) 8.3% had started a business and were able to pay salaries including the founders for 3 months, but <42 months
5) 1.8% had run a business and had been able to pay wages for 42 months.

Furthermore, reviewing the performance of SEA by impact dimension with indicators of sales area coverage showed that:

1) 47.1% of the respondents stated that there was no product/service sales or active business
2) 12.9% of them stated that there still were sales around the campus
3) 28.9% of them stated that they had reach within the city
4) 4.2% of them stated that they had reach up to the provincial level
5) 6.8% of them stated that they had reach up to the national level.

For the product innovation growth indicator:

1) 40.8% stated that there was no growth in product innovation

TABLE 2 | Validity test result.

	EEEA1	**EEEA2**	**EEEA3**	**EEEF1**	**EEEF2**	**EEEF3**	**EEET1**	**EEET2**	**EEET3**
Pearson Cor.	0.581	0.510	0.747	0.511	0.384	0.687	0.709	0.631	0.736
Sig. (2-tailed)	0.001	0.004	0.000	0.004	0.036	0.000	0.000	0.000	0.000
	EEET4	**EEET5**	**EEET6**	**EEEE1**	**EEEE2**	**ER1**	**ECS1**	**IIE_TOT**	
Pearson Cor.	0.794	0.735	0.636	0.712	0.678	0.385	0.545	1	
Sig. (2-tailed)	0.000	0.000	0.000	0.000	0.000	0.036	0.002		
	EES1	**EES2**	**EES3**	**EES4**	**EES5**	**EES6**	**EES_TOT**		
Pearson Cor.	0.756	0.622	0.672	0.779	0.676	0.690	1		
Sig. (2-tailed)	0.000	0.000	0.000	0.000	0.000	0.000			
	SEOA1	**SEOA2**	**SEOA3**	**SEOA4**	**SEOR1**	**SEOR2**			
Pearson Cor.	0.812	0.582	0.486	0.791	0.771	0.735			
Sig. (2-tailed)	0.000	0.001	0.006	0.000	0.000	0.000			
	SEOP1	SEOP2	SEOP3	SEOC1	SEOI1	**SEO_TOT**			
Pearson Cor.	0.626	0.778	0.745	0.714	0.549	1			
Sig. (2-tailed)	0.000	0.000	0.000	0.000	0.002				
	SEIPP	**SEITL**	**SEITS**	**SEIDT**	**SEIRD**	**SEI_TOT**			
Pearson Cor.	0.845	0.856	0.767	0.883	0.846	1			
Sig. (2-tailed)	0.000	0.000	0.000	0.000	0.000				
	SEAP	**SEAIM**	**SEAII**	**SEA_TOT**					
Pearson Cor.	0.699	0.831	0.901	1					
Sig. (2-tailed)	0.000	0.000	0.000						

TABLE 3 | Reliability test result.

	IIE	**EES**	**SEO**	**SEI**	**SEA**
Cronbach's Alpha	0.898	0.873	0.886	0.909	0.735
No. of items	16	6	11	5	3

2) 26.5% indicated that they experienced product innovation growth of 1–25%
3) 10.6% stated that they had product innovation growth of 26–50%
4) 10.0% stated that they had product innovation growth of 51–75%
5) 12.0% stated that the development of their product innovation was 76–100%.

Demographic data showed that 74% of male students and 76% of female students had entrepreneurial experience before entering college. Their entrepreneurial experience was motivated by various factors, including:

1) just trying (44% for male students, 54% for female students);
2) fulfilling personal needs (25% for male students, 19% for female students);
3) to fulfill family needs (5% for male students, 3% for female students).

We conducted the contingency test on respondents' demographic data between the entrepreneurial experience of students before entering college and the entrepreneurial activities they do while in college. The results showed a significant relationship between the two periods with a contingency coefficient of 37.5% for female students and 40.6% for male students (Astuty et al., 2021). This validates the choice of the research model for this study, which shows that, for the formation of a student's entrepreneurial activities, the student needs an entrepreneurship education that is more than just knowledge-based. It also strengthens the results of the contingency test, which proved that the student's entrepreneurial activities while in college were not related to their knowledge-based

TABLE 4 | Descriptive analysis of IIE perception.

	Average of IIE variable → 1,826 (High)																
	EEE—About			EEE—For			EEE—Through						EEE-Embeddeed		Research	Com-Serv	
	EEEA1	EEEA2	EEEA3	EEEF1	EEEF2	EEEF3	EEET1	EEET2	EEET3	EEET4	EEET5	EEET6	EEEE1	EEEE2	ER1	ECS1	
Mean	4.12	4.20	4.07	4.22	4.13	3.93	3.99	4.07	4.02	3.99	4.23	3.97	4.05	4.06	3.90	3.80	
Std. Dev	0.77	0.78	0.78	0.71	0.81	0.96	0.96	0.97	0.97	1.01	0.84	0.90	0.89	0.83	0.96	0.95	
Sum	1,877	1,915	1,858	1,924	1,885	1,790	1,819	1,857	1,835	1,819	1,928	1,809	1,845	1,853	1,778	1,735	
	Avg of EEEA = 1,883 (H)			Avg of EEEF = 1,866 (H)			Avg of EEET = 1,845 (H)						EEEE = 1,849 (H)		1,778 (H)	1,735 (H)	

H, high; M, moderate; L, low, com-serv, community services.

entrepreneurship education when they attended high school (Astuty et al., 2021).

Bivariate testing of demographic data in this study proves that the emergence of students' entrepreneurial intentions is not related to the study program taken by students. Even non-business study programs can trigger the emergence of students' intentions to become entrepreneurs, even though these non-business study programs do not directly present an entrepreneurial learning curriculum as much as business study programs (Astuty and Yustian, 2021).

Confirmatory Factor Analysis and Measurement Model Analysis

The results from the measurement model show that the main characteristics of each variable observed in this study were validity, reliability, and having a good model fit. The results of calculations using the related AMOS software are presented in **Table 9**.

Not all proposed indicators in the conceptual model were valid, and only the indicators in **Table 9** have a loading factor value of >0.5, so they were declared valid (Hair et al., 2014). The high loading factor indicated that the latent constructs converge on a common point (have good convergent validity). Variance Extract (VE) of the IIE variable is 55.94%, which means that all indicators representing the IIE construct have a convergence rate of 55.94%. This implies that the data variance described in the IIE variable can be considered a communality (Hair et al., 2014). This also applies to the VE values of other constructs in the model, including VE of EES = 62.37%, VE of SEO = 50.00%, VE of SEI = 71.38%, and VE of SEA = 67.41%.

Good reliability means that all indicators can consistently represent the measured latent variables. Good reliability is achieved if the Construct Reliability (CR) value is higher than 0.7 (Hair et al., 2014). As seen in **Table 9**, all the CR values in the five latent variables were >0.7, so it can be confirmed that the IIE, EES, SEO, SEI, and SEA variables have high internal consistency. All the values of CR are >0.7 and VE >0.5; therefore, all existing variables are declared reliable (Hair et al., 2014).

Based on the confirmatory factor analysis, it can be firmly established that:

(1) The institutions that facilitate business funding activities from outside parties for its students are the most dominant internal variable and an indicator of the social values of a university's entrepreneurial ecosystem.
(2) The offer of technology procurement support from outside is the main characteristic of the external support variable.
(3) Student behavior that is proactive toward change is the most dominant indicator reflecting student entrepreneurial orientation.
(4) Determination to become an entrepreneur is the main characteristic of entrepreneurial intentions.
(5) The growth of product/service innovation is the most dominant indicator that reflects the impact of student entrepreneurial activities at universities in Indonesia.

TABLE 5 | Descriptive analysis of EES perception.

Average of EES variable → 1,806 (High)						
	EXTERNAL ENVIRONMENT SUPPORT					
	EES1	**EES2**	**EES3**	**EES4**	**EES5**	**EES6**
Mean	3.75	4.11	4.01	3.86	4.03	4.02
Std. Dev	1.05	0.88	0.97	1.00	0.91	0.97
Sum	1,710 (H)	1,872 (H)	1,827 (H)	1,762 (H)	1,836 (H)	1,831 (H)

H, high; M, moderate; L, low.

TABLE 6 | Descriptive analysis of SEO perception.

Average of SEO variable → 1,743 (High)											
	Autonomy (SEOA)				**Risk taking (SEOR)**		**Proactiveness (SEOP)**			**Competitive aggressiveness (SEOC)**	**Innovativeness (SEOI)**
	SEOA1	**SEOA2**	**SEOA3**	**SEOA4**	**SEOR1**	**SEOR2**	**SEOP1**	**SEOP2**	**SEOP3**	**SEOC1**	**SEOI1**
Mean	3.81	3.92	3.82	3.93	3.79	3.75	3.97	4.01	3.95	3.69	3.80
Std. Dev	0.77	0.76	0.81	0.84	0.92	0.95	0.76	0.77	0.80	0.92	0.90
Sum	1,739	1,787	1,741	1,791	1,726	1,712	1,812	1,828	1,800	1,742	1,699
	Avg of SEOA = 1,765 (H)				Avg of SEOR = 1,719 (H)		Avg of SEOP = 1,813 (H)			1,682 (H)	1,734 (H)

H, high; M, moderate; L, low.

TABLE 7 | Descriptive analysis of SEI perception.

Average of SEI variable → 1,825 (High)					
	Preferred priority (SEIPP)	**Tend to like (SEITL)**	**Think seriously (SEITS)**	**Determined (SEIDT)**	**Ready to do (SEIRD)**
Mean	3.92	4.04	4.19	3.83	4.04
Std. Dev	1.02	1.03	0.83	1.03	0.96
Sum	1,788 (H)	1,840 (H)	1,909 (H)	1,748 (H)	1,841 (H)

H, high; M, moderate; L, low.

TABLE 8 | Descriptive analysis of SEA perception.

Average of SEA variable → 1,009 (Low)			
	SEA By phase	**SEA By impact**	
	(SEAP)	**Market Scope (SEAIM)**	**Innovation (SEAII)**
Mean	2.43	2.11	1.90
Std. Dev	0.87	1.24	1.13
Sum	1,107	960	865
	Avg = 1,107 (M)	Avg of by impact = 912 (L)	

The relationship magnitude between the latent variables is shown in the correlation matrix in **Table 10**.

The highest correlation occurs in the interaction between the internal environment of the institution and external parties in the form of support for entrepreneurial learning. A value of 0.685 indicates a moderate to strong relationship (Hair et al., 2014). This result is in line with the findings that formal learning in institutions and informal learning obtained through external support, such as training support, mentoring, and network access support, which strengthens education and entrepreneurial competence through knowledge and access to the entrepreneurial resources available outside the institution (Williams Middleton et al., 2019).

Figure 2 below shows the complete measurement model analysis where the valid and reliable indicators of the CFA results have been constructed.

According to the estimation results of Goodness of Fit (GOF) test results at the **Table 11**, there are two GOF measures met the fit criteria in the goodness of fit (Hair et al., 2014).

Root mean square error of approximation (RMSEA) is one of the most widely used GOF measurements to correct the χ^2 significance test that rejects models with a large sample or a large number of observed variables so that the RMSEA better represents how well the model fits the population, not just the

TABLE 9 | Loading factor, construct reliability, and variance extract of IIE, EES, SEO, SEI, and SEA.

Indicators	Variables					
	IIE	**EES**	**SEO**	**SEI**	**SEA**	
EEET1	0.698					Valid
EEET2	0.722					Valid
EEET3	0.816					Valid
EEET4	**0.848**					Valid
EEET5	0.734					Valid
EEET6	0.792					Valid
EEEE1	0.637					Valid
EEEE2	0.714					Valid
EES1		0.706				Valid
EES2		0.776				Valid
EES3		0.827				Valid
EES4		**0.831**				Valid
EES5		0.803				Valid
EES6		0.789				Valid
SEOA1			0.666			Valid
SEOA2			0.617			Valid
SEOA3			0.685			Valid
SEOA4			0.729			Valid
SEOR1			0.653			Valid
SEOP1			0.695			Valid
SEOP2			**0.754**			Valid
SEOP3			0.744			Valid
SEOC1			0.652			Valid
SEOI1			0.678			Valid
SEIPP				0.847		Valid
SEITL				0.842		Valid
SEITS				0.747		Valid
SEIDT				**0.900**		Valid
SEIRD				0.880		Valid
SEAP					0.814	Valid
SEAIM					0.824	Valid
SEAII					**0.825**	Valid
Construct reliability (CR)	0.9098	0.9084	0.8998	0.9255	0.8612	
Variance extract (VE)	55.94%	62.37%	51.00%	71.38%	67.41%	
	Reliable	Reliable	Reliable	Reliable	Reliable	

The bold values indicates the largest loading factor value for each of the IIE, EES, SEO, SEI, and SEA variables.

TABLE 10 | Correlation between variables.

	SEA	**SEI**	**SEO**	**EES**
SEI	0.341			
SEO	0.361	0.526		
EES	0.111	0.192	0.336	
IIE	0.189	0.198	0.309	0.685

sample used for estimation. A low RMSEA value indicates a good model fit. Previous studies used a limit of <0.05 as a good RMSEA value, while some others used <0.08 as a good RMSEA limit. Recent research does not recommend a certain limit in expressing the RMSEA value (Hair et al., 2014). Whether the limit is <0.05 or <0.08, the results of this study, with an RMSEA value of 0.047, fulfill both limits. This also confirms that the RMSEA value in this study can correct the value of the χ^2 significance test to meet the model fit criteria.

The Comparative Fit Index (CFI) is an incremental progression toward the model fit criteria. The range of CFI values is between 0 and 1. A higher value indicates a better fit, so a CFI value >0.90 is usually associated with a model fit (50). In this study, the CFI value was 0.9470, thus meeting the model fit criteria. Considering that the two goodness of fit measures (RMSEA and CFI) met the model fit criteria, it can be stated that the model built based on the empirical evidence meets the criteria of the best-fit model. The model can estimate the

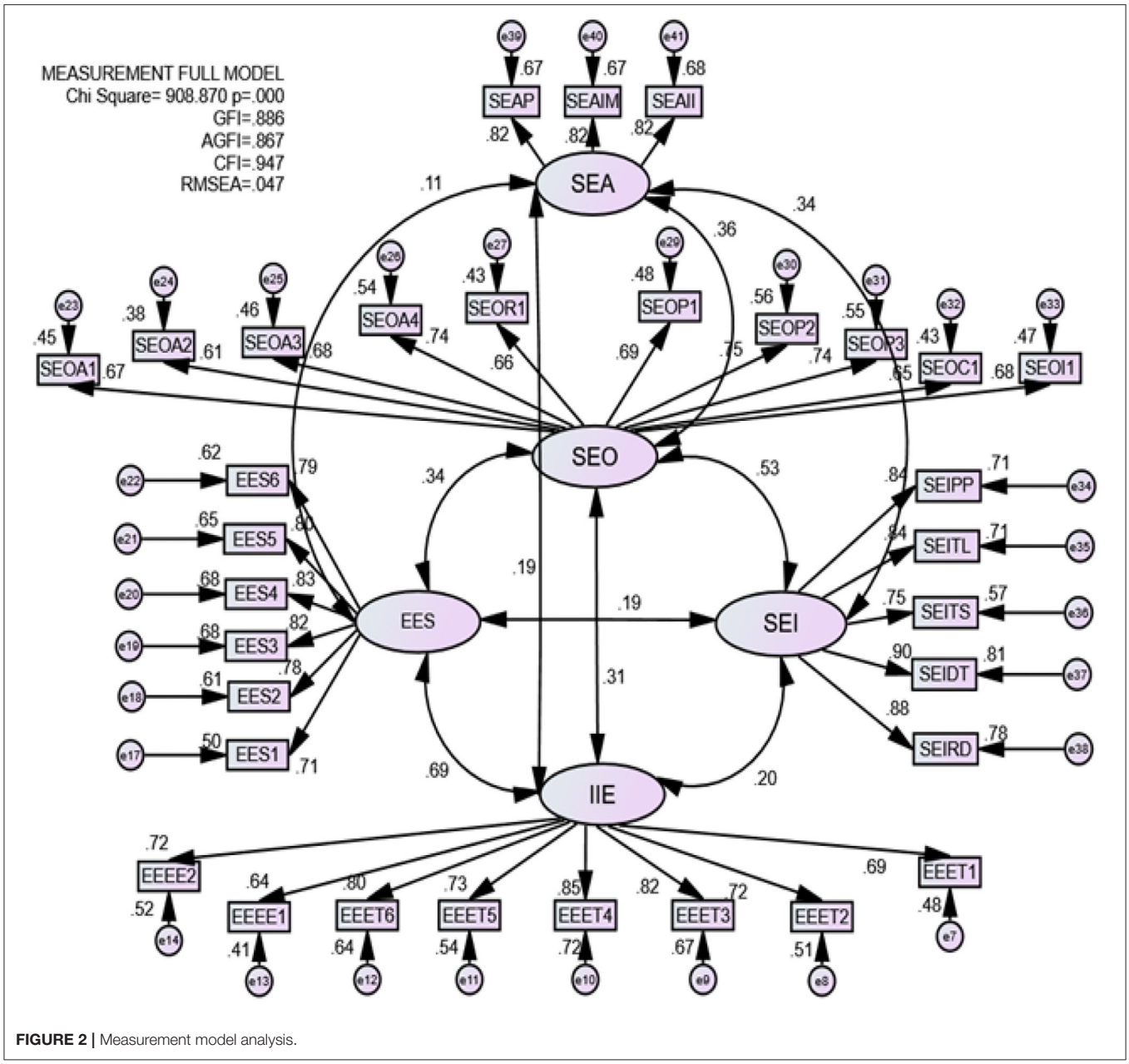


FIGURE 2 | Measurement model analysis.

population covariance matrix, which tends not to differ from the sample data covariance matrix.

Structural Model Analysis

Nine hypotheses proposed in the conceptual model were tested through a structural model using the AMOS20 software, as shown in **Figure 3**.

The statistical parameters of the structural model test results in **Figure 3** are presented in the summary of estimation results for the structural model parameter in **Table 12**.

TABLE 11 | Goodness of fit test results.

GOF	Fit criteria (Hair et al., 2014)	Results	Conclusion
Chi-square significance test (χ^2 Test)	*p*-value ≥ 0.000	0.0000	Not fit
Goodness of fit index (GFI)	≥0.90	0.8860	Not fit
Adjusted goodness of fit index (AGFI)	≥0.90	0.8670	Not fit
Comparative fit index (CFI)	≥0.90	0.9470	Fit
Root mean square error of approximation (RMSEA)	≤0.08	0.0470	Fit

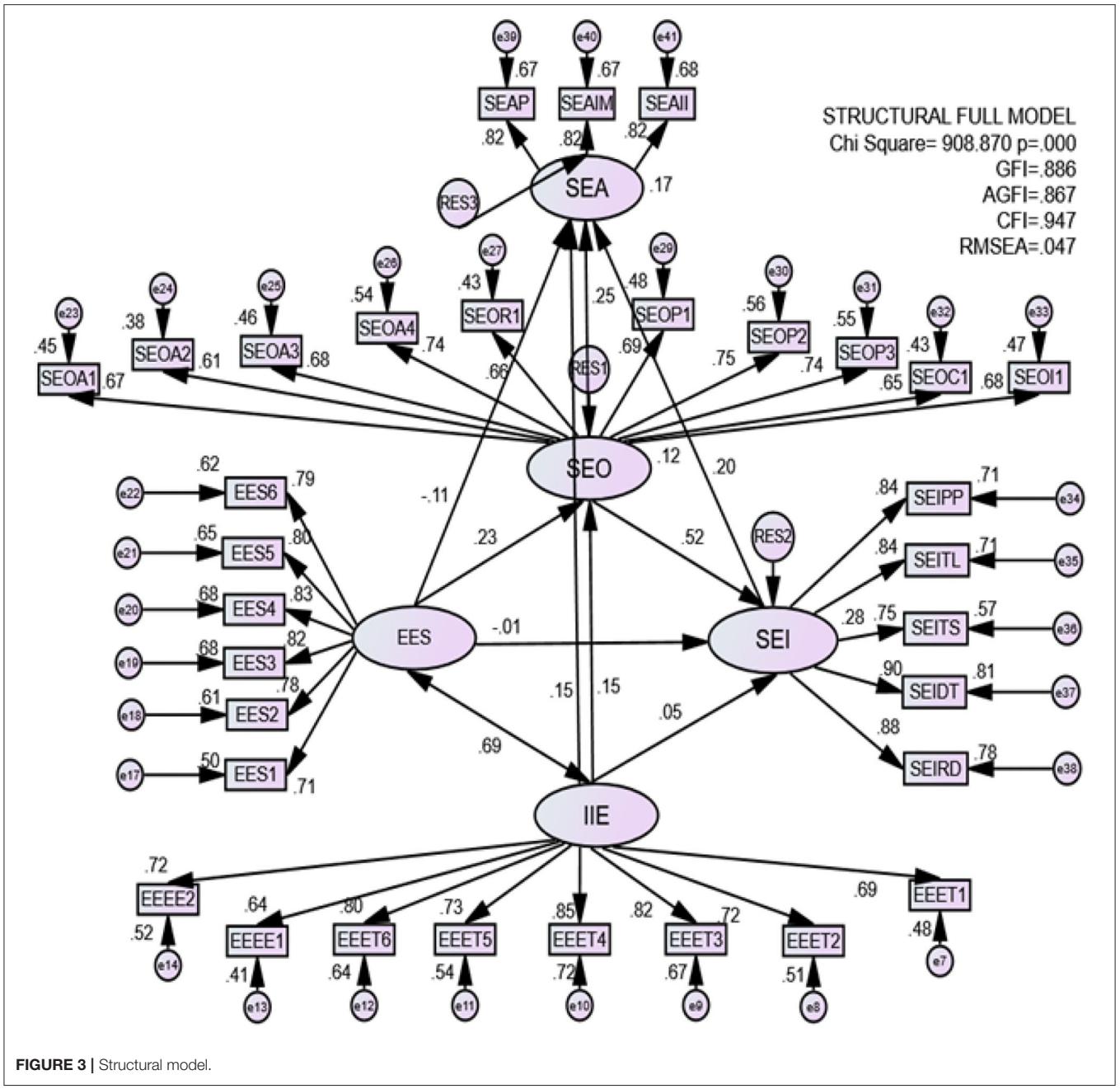


FIGURE 3 | Structural model.

The significance test results for each estimated path coefficient in **Table 12** show that the nine hypotheses proposed were not wholly accepted. Three hypotheses, namely H1c, H2c, and H2a, were rejected.

Based on the structural model summary of the estimation in **Table 12**, this research model consists of three main substructures.

Substructure 1:

SEO = 0.15*IIE + 0.24*EES.

($p = 0.041$; IIE→SEO) and ($p = 0.002$; EES→SEO).

This means that the IIE and EES can significantly increase the SEO by 15 and 24%, respectively (H1b and H2b accepted), with the strength in explaining the variation of sample data to predict the population being classified as weak ($R^2 = 12\%$; Chin, 1998). This finding aligns with the previous finding and supports the statement that various factors are conducive to the formation of SEO, including personality traits, internal and external motivation, family environment, personality features of organizational leaders, and dynamics of the internal and external environment of an organization (Pittino et al.,

TABLE 12 | Summary of estimation results for structural model parameters.

H	MODEL			Estimate		S.E.	C.R	*p*		R^2
				RW	SRW					
	SEO				**(β1)**					
H1b	SEO	←	IIE	0.115	0.148	0.056	2.039	0.041	Sig	0.12
H2b	SEO	←	EES	0.161	0.235	0.051	3.175	0.002	Sig	
	SEI				**(β2)**					
H1c	SEI	←	IIE	0.062	0.048	0.085	0.734	0.463	Not sig	0.28
H2c	SEI	←	EES	−0.016	−0.014	0.077	−0.207	0.836	Not sig	
H3b	SEI	←	SEO	0.862	0.515	0.096	8.965	***	Sig	
	SEA				**(γ)**					
H1a	SEA	←	IIE	0.162	0.151	0.078	2.07	0.038	Sig	0.17
H2a	SEA	←	EES	−0.109	−0.114	0.07	−1.548	0.122	Not sig	
H3a	SEA	←	SEO	0.339	0.246	0.089	3.791	***	Sig	
H4	SEA	←	SEI	0.169	0.204	0.049	3.431	***	Sig	

*RW, regression weights; SRW, standardized regression weights. *** = 0.000; Sig, significant at the level 0.05; not sig, not significant at level 0.05.*

2017). Thus, it is clear that $R^2 = 12\%$, which indicates that the dynamics of Indonesian higher education's internal and external environment are quite realistic because these two variables are part of several antecedents that form the SEO.

Substructure 2:
SEI = 0.05*IIE – 0.01*EES + 0.51*SEO.
($p = 0.463$; IIE→SEI), ($p = 0.836$; EES→SEI), and ($p = 0.000$; SEO→SEI).

This means that student entrepreneurial intention can directly be influenced by the entrepreneurial orientation of the students themselves (H3b accepted). At the same time, the internal and external environments do not directly foster SEI. Substructure 2 has the power to predict that the population tends to be moderate ($R^2 = 0.28$; Chin, 1998). This finding confirms that a strong student entrepreneurial orientation can increase the entrepreneurial intentions of the students (Martins and Perez, 2020).

Substructure 3:
SEA = 0.15*IIE – 0.11*EES + 0.25*SEO + 0.20*SEI.
($p = 0.038$; IIE→SEA), ($p = 0.122$; EES→SE A), ($p = 0.000$; SEO→SEA), and ($p = 0.000$; SEI→SEA).

Substructure 3 indicates that EES does not have the power to influence students to engage in SEA directly but that IIE can foster SEA (H1a accepted). Individual internal factors related to entrepreneurship orientation and entrepreneurial intentions have been proven to accelerate student entrepreneurship activities (H3a and H4 accepted). The strength of substructure 3 ($R^2 = 0.17$) in predicting the population is still relatively weak (Chin, 1998).

Referring to **Table 12**,

- EES has a direct negative impact on SEI and SEA. Even after analyzing the impact, it is proven that EES is not significant in influencing students' intentions to become entrepreneurs. In addition, it has no impact on the emergence of student entrepreneurial activities (H2a and H2c, rejected).

  The external environmental support cannot arouse students' intentions to become entrepreneurs, or even to carry out entrepreneurial activities. This finding is in line with previous research in Indonesia and Japan, which stated that the readiness of instruments (access to capital, social networks, and information) had no significant effect on the entrepreneurial intentions of Indonesian and Japanese students (Indarti, 2015).
  For Indonesian students, capital assistance from government and private financial institutions is quite burdensome as they are obliged to repay the capital and the interest charged while still studying. In addition, the procurement of technology and the forms of cooperation offered by entrepreneurs from outside the institution require them to provide feedback on the collaboration results, and they are not sure that they can fulfill such an offer.

- IIE is not significant in influencing students' intentions to become entrepreneurs (H1c rejected).

  Based on lexical meaning, the intention is the will (desire in the heart) to do something (Ministry of Education and Culture, 2016) that requires several factors to ignite it. This study confirms that the university's internal and external environment can directly increase student entrepreneurial orientation. Still, it cannot directly influence students' intentions to become entrepreneurs, especially encouraging them to engage in entrepreneurial activities.
  The results of this study are in line with previous research, which succeeded in constructing various factors conducive to the formation of entrepreneurial orientation, including personality traits, internal and external motivation, family environment, personality traits of organizational leaders, and dynamics of an organization's internal and external

TABLE 13 | Direct effect, indirect effect, and the total effect of each variable.

		SEO	SEI	SEA
Standardized direct effects	IIE →	0.148**	0.048$^{(ns)}$	0.151**
	EES →	0.235**	−0.014$^{(ns)}$	−0.114$^{(ns)}$
	SEO →	0	0.515**	0.246**
	SEI →	0	0	0.204**
Standardized indirect effects	IIE →	0	0.076	0.062
	EES →	0	0.121	0.079
	SEO →	0	0	0.105
	SEI →	0	0	0
Standardized total effects	IIE →	0.148**	0.125$^{(ns)}$	0.213**
	EES →	0.235**	0.107$^{(ns)}$	−0.035$^{(ns)}$
	SEO →	0	0.515**	0.351**
	SEI →	0	0	0.204**

***Significant at the level 0.05. ^{ns}, not significant at level 0.05.*

environment (Pittino et al., 2017). Therefore, it is clear that IIE can influence entrepreneurial orientation first and then create entrepreneurial intentions.

The ability of the variable mediators to mediate their antecedents to their consequences in a particular relationship is presented in the decomposition analysis in **Table 13**.

Institutional Internal Environment as an entrepreneurial ecosystem entity created through eclectic entrepreneurship education with the "Through" and "Embedded" types has proven to significantly influence the performance of SEA directly by 15.1%. This influence was found to increase when mediated by the growth of SEO. The students' eclectic entrepreneurship education increased their entrepreneurial intentions and ultimately impacted SEA performance up to 21.3%. This proves that SEO and SEI are significant mediating variables for IIE in influencing the performance of SEA. Furthermore, increasing SEO can directly and significantly influence the SEI by 51.5%. This is an interesting finding because, when viewed with respect to the direct influence of SEO on SEA, which is only 25%, increasing SEO further increases SEI to 35% because of the SEI that develops in students due to their increased entrepreneurial orientation.

DISCUSSION

"Eclectic Entrepreneurship Education" Is a Novelty in Indonesian Higher Education

Through contingency testing of demographic data in this study, it has been empirically proven that the pattern of knowledge-based entrepreneurship education alone is not significantly correlated with entrepreneurial activities carried out by students while in the college (Astuty et al., 2021). Meanwhile, the contingency relationship between students' entrepreneurial experiences before entering college and entrepreneurial activities in college shows a significant relationship. It proves that a practical education pattern is significantly correlated with the entrepreneurial activities of young entrepreneurs.

The next question that arises is, what kind of education and learning pattern can trigger an increase in entrepreneurial activities for college students and the growth of students' new businesses?

"Through" and "Embedded" types of entrepreneurship education were found to be valid, reliable, and suitable for Indonesian entrepreneurship education. "Through" type accelerates the development of students' intentions to start their own business (Ollila and Williams-Middleton, 2011; Baluku et al., 2019), improve entrepreneurial competence (Gibson and Tavlaridis, 2018; Hägg and Kurczewska, 2019; Williams Middleton et al., 2019), and enhance the effectiveness of entrepreneurial learning (Ratten and Usmanij, 2020). Meanwhile "Embedded" type refers to embedding entrepreneurship education in a particular discipline to make it relevant for that specific discipline so that it can create entrepreneurial activities that have career relevance and the potential to generate intellectual property (Blake Hylton et al., 2020), facilitate the embedding of business study program students to take specific courses according to their specialization in non-business study programs so that they are expected to provide the necessary context (Thomassen et al., 2019), and to improve the ability to construct knowledge in the work-life of entrepreneurs (Bridgstock, 2013). "Through" and "Embedded" types simultaneously provide direct knowledge and practical learning to equip students with comprehensive entrepreneurial experience (Gieure et al., 2019).

Considering these findings, the authors formulate this entrepreneurship learning type as ***"Eclectic Entrepreneurship Education."*** This nomenclature aligns with the Indonesian dictionary, which states that training or education carried out using various techniques, approaches, and methods simultaneously is called eclectic education (Ministry of Education and Culture, 2016). "Eclectic Entrepreneurship Education" is a novelty term from this research that higher education decision-makers in Indonesia can use to re-orient the entrepreneurship learning curriculum to create a conducive university entrepreneurial ecosystem.

"***Through***" dimension of Eclectic Entrepreneurship Education (EEET) emphasizes learning by doing but in a "safe" mode, such as in university business incubators and internship programs for specific business units or companies. Examples from the "through" dimension include:

- the existence of business incubators at universities
- the presence of experiential learning through internship programs for various businesses/industry
- the availability of experiential learning through collaboration with entrepreneurs
- facilitating students to obtain business funding from outside parties

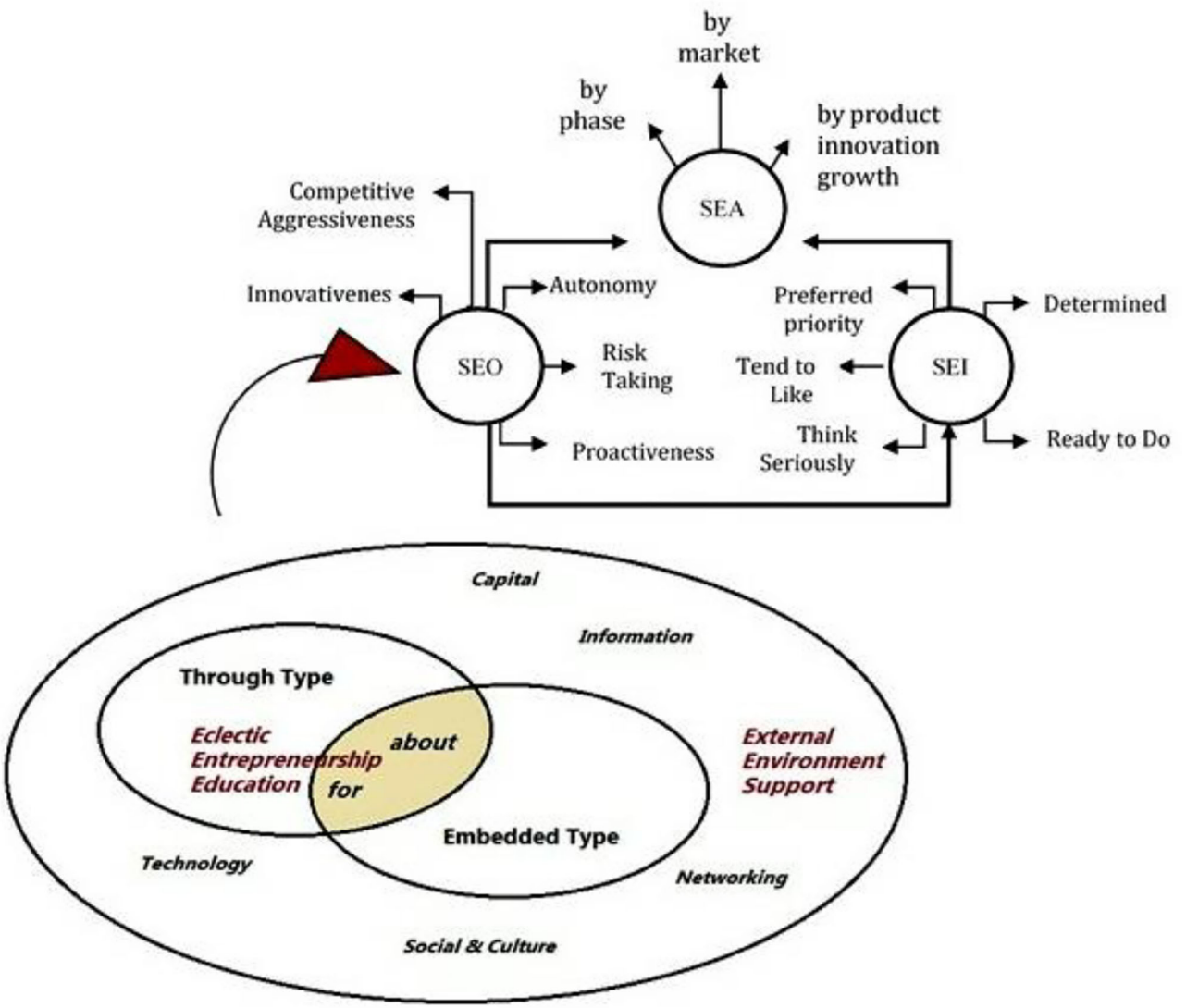


FIGURE 4 | Empirical model of Indonesian University entrepreneurship ecosystem.

- providing mentoring programs from entrepreneurs to students
- facilitating students to interact with the market

"Embedded" dimension of the Eclectic Entrepreneurship Education (EEEE) emphasizes direct entrepreneurship learning in a particular discipline so that it is relevant to the field of specialization and can generate intellectual property. Examples from the "Embedded" dimension include:

- the availability of embedding entrepreneurship courses in the curriculum of non-business study programs
- the inclusion of business study students in non-business study programs
- the opportunity for students to enroll in several classes according to their specialization
- to further improve the diversity of skills needed by students

Entrepreneurship Eclectic Education strengthens the university's entrepreneurial ecosystem because of the intersection of interactions between the university's internal and external environment. This learning model then strengthens the entrepreneurial orientation of students, which in turn can generate entrepreneurial intentions that result in entrepreneurial activities carried out by students. The empirical model of the university ecosystem in forming student entrepreneurship activities in several universities in Indonesia is presented in **Figure 4**.

The educators and institutions need to understand the importance of a pragmatic and comprehensive approach for their students to generate their interest, express increased willingness to become entrepreneurs, and provide the necessary motivation to engage in entrepreneurial activities to become potential entrepreneurs. The finding from this study confirms

that SEI is proven to be a positive and significant mediator for SEO in increasing SEA. Hence, it can be concluded that the entrepreneurial ecosystem in which there is a conducive internal environment accompanied by interactions with external parties can contribute positively to individual attributes such as SEO and SEI for the realization of the SEA (Knight and Yorke, 2003; Hulme et al., 2014; Mehtap et al., 2017; Kuratko and Morris, 2018).

RESEARCH LIMITATIONS AND SUGGESTIONS FOR FUTURE RESEARCH

This study has several research limitations, including:

a) Sampling was only in 18 universities in Java and Sumatra. Yet, it was done because 75% of universities in Indonesia are spread over two of the five major islands in Indonesia, namely Java and Sumatra. Meanwhile, another 25% of universities spread across the islands of Kalimantan, Sulawesi, and Papua were not included in the sampling. Researchers had limited resources in distributing samples to universities spread across three other major islands in Indonesia. However, the goodness of fit test included the RMSEA and CFI values in the model fit criteria. It can be emphasized that the model built based on the empirical evidence meets the requirements of the best-fit model, so it means that the model can estimate the population covariance matrix, which tends not to differ from the sample data covariance matrix. It is suggested that future research on this study's results adopts the empirical model to study universities that are spread over three other major islands in Indonesia.
b) The empirical evidence revealed something quite surprising regarding SEO, which significantly increases SEI by 51% directly though this intention results in the realization of SEA by 20%. Therefore, future research can study the effect of intervening variables between the SEI and SEA and investigate and verify the previous findings that all entrepreneurial intentions do not lead to concrete actions in the form of business creation (Shirokova et al., 2016). This, therefore, could be the "next task" to find the factors that can facilitate students' entrepreneurial intentions into actual entrepreneurial activities.
c) This study included just one direct antecedent of SEI in its research model, namely SEO. It is suggested that further research be undertaken to examine the competence of lecturers in frontline universities who can directly deliver "entrepreneurship eclectic education" to students, as recommended in this study. This is also based on other research findings that institutions that provide entrepreneurship education must have competent lecturers who can ignite the fire of entrepreneurial intentions in students (Iwu et al., 2019).

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Prof. Dr. Tirta Nugraha Mursitama, S.Sos., M.M., Ph.D. (BINUS University), Prof. Dr. Mts. Arief, M.M., MBA., CPM. (BINUS University), Prof. Dr. Sasmoko, M.Pd. (BINUS University), and Nugroho Juli Setiadi, S.E., M.M., Ph.D. (BINUS University). Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

EA contributed to the literature review, model development, data processing, and analysis. OY and CR contributed to data collection and report preparation. All authors contributed to the article and approved the submitted version.

FUNDING

Research and publications are funded by BINUS University through Basic Research Internal Grants with implementation agreement No: 079/VR.RTT/VIII/2020.

ACKNOWLEDGMENTS

The authors are thankful to all parties who approved undertaking this research and BINUS University for funding this research. The authors also thank all colleagues who helped disseminate the research instrument and all respondents who participated in this study. We would like to thank Editage (www.editage.com) for English language editing.

REFERENCES

Astuty, E., and Yustian, O. R. (2021). Analysis of the entrepreneurial intention's emergence at business and non-business students in Indonesia. *J. Pendidikan Prog.* 11, 27–38. doi: 10.23960/jpp.v11.i1.202103

Astuty, E., Yustian, O. R., and Ratnapuri, C. I. (2021). "Gender, student background, and the implications for the emergence of student entrepreneurial activities," in *Proceedings of the International Conference on Industrial Engineering and Operations Management* (Singapore), 1729–1737.

Baluku, M. M., Matagi, L., Musanje, K., Kikooma, J. F., and Otto, K. (2019). Entrepreneurial socialization and psychological capital: cross-cultural and multigroup analyses of impact of mentoring, optimism, and self-efficacy on entrepreneurial intentions. *Entrep. Educ. Pedagogy* 2, 5–42. doi: 10.1177/2515127418818054

Bayarçelik, E. B., and Özşahin, M. (2014). How entrepreneurial climate effects firm performance? *Proc. Soc. Behav. Sci.* 150, 823–833. doi: 10.1016/j.sbspro.2014.09.091

Bazan, C., Shaikh, A., Frederick, S., Amjad, A., Yap, S., Finn, C., et al. (2019). Effect of memorial university's environment and support system in shaping entrepreneurial intention of students. *J. Entrep. Educ.* 22, 1–35.

Blake Hylton, J., Mikesell, D., Yoder, J.D., and LeBlanc, H. (2020). Working to instill the entrepreneurial mindset across the curriculum. *Entrep. Educ. Pedagogy* 3, 86–106. doi: 10.1177/2515127419870266

Bosma, N., Hill, S., Ionescu-Somers, A., Kelley, D., Levie, J., Tarnawa, A., et al. (2020). *The Global Entrepreneurship Monitor (GEM): 2019/2020 Global Report*. London.

BPS Indonesia (2016). *Profil UsahaPerusahaan Ekonomi Kreatif*. Jakarta: BPS Indonesia.

Bridgstock, R. (2013). Not a dirty word: arts entrepreneurship and higher education. *Arts Human. High. Educ.* 12, 122–137. doi: 10.1177/1474022212465725

Brown, A. (2018). Embedding research and enterprise into the curriculum adopting Student as Producer as a theoretical framework. *High. Educ. Skills Work Based Learn.* 8, 29–40. doi: 10.1108/HESWBL-09-2017-0064

Chavez, R., Yu, W., Jacobs, M. A., and Feng, M. (2017). Manufacturing capability and organizational performance: the role of entrepreneurial orientation. *Int. J. Prod. Econ.* 184, 33–46. doi: 10.1016/j.ijpe.2016.10.028

Chin, W. (1998). *Structural Equation With Latent Variables*. New York, NY: John Willey and Sons.

Cooper, D. R., and Schindler, P. S. (2014). *Business Research Methods. 12th Edn.* New York, NY: McGraw-Hill/Irwin.

Emoke-Szidónia, F. (2015). International entrepreneurial orientation and performance of romanian small and medium-sized firms: empirical assessment of direct and environment moderated relations. *Proc. Econ. Fin.* 32, 186–193. doi: 10.1016/S2212-5671(15)01381-7

Fong, T. W. M. (2020). Design incubatees' perspectives and experiences in Hong Kong. *High. Educ. Skills Work Based Learn.* 10, 481–496. doi: 10.1108/HESWBL-10-2019-0130

Franco, M., Haase, H., and Lautenschläger, A. (2010). Students' entrepreneurial intentions: an inter-regional comparison. *Educ. Train.* 52, 260–275. doi: 10.1108/00400911011050945

Galindo-Martín, M. A., Méndez-Picazo, M. T., and Castaño-Martínez, M. S. (2019). The role of innovation and institutions in entrepreneurship and economic growth in two groups of countries. *Int. J. Entrep. Behav. Res.* 26, 485–502. doi: 10.1108/IJEBR-06-2019-0336

Galvão, A. R., Marques, C. S. E., Ferreira, J. J., and Braga, V. (2020). Stakeholders' role in entrepreneurship education and training programmes with impacts on regional development. *J. Rural Stud.* 74, 169–179. doi: 10.1016/j.jrurstud.2020.01.013

Gibson, D., and Tavlaridis, V. (2018). Work based learning for enterprise education? The case of Liverpool John Moores university "live" civic engagement projects for students. *High. Educ. Skills Work-Based Learn.* 8, 5–14. doi: 10.1108/HESWBL-12-2017-0100

Gieure, C., Benavides-Espinosa, M., and Roig-Dobón, S. (2019). Entrepreneurial intentions in an international university environment. *Int. J. Entrep. Behav. Res.* 25, 1605–1620. doi: 10.1108/IJEBR-12-2018-0810

Gupta, V. K., Dutta, D. K., Guo, C. G., and Javadian, G. (2016). Classics in entrepreneurship research: enduring insights, future promises. *N. Engl. J. Entrep.* 19, 7–23. doi: 10.1108/NEJE-19-01-2016-B001

Hägg, G., and Kurczewska, A. (2019). Toward a learning philosophy based on experience in entrepreneurship education. *Entrep. Educ. Pedagogy* 2019:251512741984060. doi: 10.1177/2515127419840607

Hair, J. F., et al. (2014). *Multivariate Data Analysis, 7th Edn*. England: Pearson.

Hulme, E., Thomas, B., and DeLaRosby, H. (2014). Developing creativity ecosystems: preparing college students for tomorrow's innovation challenge. *About Campus* 19, 14–23. doi: 10.1002/abc.21146

Indarti, N. (2015). Factors affecting entrepreneurial intentions among indonesian students. *Fact. Affect. Entrep. Intent. Among Indones. Stud.* 19, 57–70. doi: 10.22146/jieb.6585

Indonesian Ministry of Cooperatives and MSMEs (2020). *Through Training, Ministry of Cooperation and SMEs Increase the Competitiveness of MSMEs in Super Priority Tourism Destinations*. Jakarta. Available online at: https://kemenkopukm.go.id/read/melalui-pelatihan-kemenkop-dan-ukm-tingkatkan-daya-saing-umkm-di-destinasi-wisata-super-prioritas (accessed June 6, 2020).

Iwu, C. G., et al. (2019). Entrepreneurship education, curriculum and lecturer-competency as antecedents of student entrepreneurial intention. *Int. J. Manage. Educ.* 2019:100295. doi: 10.1016/j.ijme.2019.03.007

Keh, H. T., Nguyen, T. T. M., and Ng, H. P. (2007). The effects of entrepreneurial orientation and marketing information on the performance of SMEs. *J. Bus. Ventur.* 22, 592–611. doi: 10.1016/j.jbusvent.2006.05.003

Knight, P. T., and Yorke, M. (2003). Employability and good learning in higher education. *Teach. High. Educ.* 8, 3–16. doi: 10.1080/1356251032000052294

Kourilsky, M. L., and Walstad, W. B. (1998). Entrepreneurship and female youth: knowledge, attitudes, gender differences, and educational practices. *J. Bus. Ventur.* 13, 77–88. doi: 10.1016/S0883-9026(97)00032-3

Kristiansen, S. (2001). Promoting African pioneers in business: what makes a context conducive to small-scale entrepreneurship? *J. Entrep.* 10, 43–69. doi: 10.1177/097135570101000103

Kristiansen, S. (2002). Individual perception of business contexts: the case of smallscale entrepreneurs in Tanzania. *J. Dev. Entrep.* 7.

Kuratko, D. F., and Morris, M. H. (2018). Corporate entrepreneurship: a critical challenge for educators and researchers. *Entrep. Educ. Pedagogy* 1, 42–60. doi: 10.1177/2515127417737291

Lavelle, B. A. (2019). Entrepreneurship education's impact on entrepreneurial intention using the theory of planned behavior: evidence from chinese vocational college students. *Entrep. Educ. Pedagogy* 2019:251512741986030. doi: 10.1177/2515127419860307

Li, Y. H., Huang, J. W., and Tsai, M. T. (2009). Entrepreneurial orientation and firm performance: the role of knowledge creation process. *Indus. Mark. Manage.* 38, 440–449. doi: 10.1016/j.indmarman.2008.02.004

Lumpkin, G. T., and Dess, G. G. (1996). Clarifyang the entrepreneurial orientation construct and linkingit to performance. *Acad. Manage. Rev.* 21, 135–172. doi: 10.2307/258632

Lumpkin, G. T., and Dess, G. G. (2001). Linking two dimensions of entrepreneurial orientation to firm performance: the moderating role of environment and industry life cycle. *J. Bus. Ventur.* 16, 429–451. doi: 10.1016/S0883-9026(00)00048-3

Lyon, D. W., Lumpkin, G. T., and Dess, G. G. (2000). Enhancing entrepreneurial orientation research: operationalizing and measuring a key strategic decision making process. *J. Manage.* 26, 1055–1085. doi: 10.1177/014920630002600503

Martins, I., and Perez, J. P. (2020). Testing mediating effects of individual entrepreneurial orientation on the relation between close environmental factors and entrepreneurial intention. *Int. J. Entrep. Behav. Res.* 26, 771–791. doi: 10.1108/IJEBR-08-2019-0505

McDonald, S., Gertsen, F., Rosenstand, C. A. F., and Tollestrup, C. (2018). Promoting interdisciplinarity through an intensive entrepreneurship education post-graduate workshop. *High. Educ. Skills Work Based Learn.* 8, 41–55. doi: 10.1108/HESWBL-10-2017-0076

Mehtap, S., Pellegrini, M. M., Caputo, A., and Welsh, D. H. B. (2017). Entrepreneurial intentions of young women in the Arab world: socio-cultural and educational barriers. *Int. J. Entrep. Behav. Res.* 23, 880–902. doi: 10.1108/IJEBR-07-2017-0214

Ministry of Education and Culture (2016). *KBBI Online*. Available online at: https://kbbi.kemdikbud.go.id/ (accessed July 2, 2021).

Nabi, G., Liñán, F., Fayolle, A., Krueger, N., and Walmsley, A. (2017). The impact of entrepreneurship education in higher education: a systematic review and research agenda. *Acad. Manage. Learn. Educ.* 16, 277–299. doi: 10.5465/amle.2015.0026

Odewale, G. T., Hani, S. H., Migro, S. O., and Adeyeye, P. O. (2019). Entrepreneurship education and students' views on self-employment among international postgraduate students in universiti utara malaysia. *J. Entrep. Educ.* 22, 1–15.

Ollila, S., and Williams-Middleton, K. (2011). The venture creation approach: integrating entrepreneurial education and incubation at the university. *Int. J. Entrep. Innov. Manage.* 13, 161–178. doi: 10.1504/IJEIM.2011.038857

Pittaway, L., and Edwards, C. (2012). Assessment: examining practice in entrepreneurship education. *Educ. Train.* 54, 778–800. doi: 10.1108/00400911211274882

Pittino, D., Visintin, F., and Lauto, G. (2017). A configurational analysis of the antecedents of entrepreneurial orientation. *Eur. Manage. J.* 35, 224–237. doi: 10.1016/j.emj.2016.07.003

Ratten, V., and Usmanij, P. (2020). Entrepreneurship education: time for a change in research direction? *Int. J. Manage. Educ.* 2019:100367. doi: 10.1016/j.ijme.2020.100367

Sahin, F., Karadag, H., and Tuncer, B. (2019). Big five personality traits, entrepreneurial self-efficacy and entrepreneurial intention: a configurational approach. *Int. J. Entrep. Behav. Res.* 25, 1188–1211. doi: 10.1108/IJEBR-07-2018-0466

Sancho, M. P. L., Martín-Navarro, A., and Ramos-Rodríguez, A. R. (2020). Will they end up doing what they like? The moderating role of the attitude towards entrepreneurship in the formation of entrepreneurial intentions. *Stud. High. Educ.* 45, 416–433. doi: 10.1080/03075079.2018.1539959

Schwab, K. (2018). *The Global Competitiveness Index Report 2017-2018*. World Economic Forum. Available online at: http://ci.nii.ac.jp/naid/110008131965/

Shirokova, G., Osiyevskyy, O., and Bogatyreva, K. (2016). Exploring the intention-behavior link in student entrepreneurship: moderating effects of individual and environmental characteristics. *Eur. Manage. J.* 34, 386–399. doi: 10.1016/j.emj.2015.12.007

Souitaris, V., Zerbinati, S., and Al-Laham, A. (2007). Do entrepreneurship programmes raise entrepreneurial intention of science and engineering students? The effect of learning, inspiration and resources. *J. Bus. Ventur.* 22, 566–591. doi: 10.1016/j.jbusvent.2006.05.002

Thomassen, M. L., Middleton, K. W., Ramsgaard, M. N., and Neergaard, H., and Warren, L. (2019). Conceptualizing context in entrepreneurship education: a literature review. *Int. J. Entrep. Behav. Res.* 26, 863–886. doi: 10.1108/IJEBR-04-2018-0258

Walter, A., Auer, M., and Ritter, T. (2006). The impact of network capabilities and entrepreneurial orientation on university spin-off performance. *J. Bus. Ventur.* 21, 541–567. doi: 10.1016/j.jbusvent.2005.02.005

Walter, S. G., Parboteeah, K. P., and Walter, A. (2013). University departments and self-employment intentions of business students: a cross-level analysis. *Entrep. Theory Pract.* 37, 175–200. doi: 10.1111/j.1540-6520.2011.00460.x

Williams Middleton, K., Padilla-Meléndez, A., Lockett, N., Quesada-Pallarès, C., and Jack, S. (2019). The university as an entrepreneurial learning space: the role of socialized learning in developing entrepreneurial competence. *Int. J. Entrep. Behav. Res.* 26, 887–909. doi: 10.1108/IJEBR-04-2018-0263

Zabelina, E., Deyneka, O., and Tsiring, D. (2019). Entrepreneurial attitudes in the structure of students' economic minds. *Int. J. Entrep. Behav. Res.* 25, 1621–1633. doi: 10.1108/IJEBR-04-2018-0224

Zhang, J. A., Edgar, F., Geare, A., and O'kane, C. (2016). The interactive effects of entrepreneurial orientation and capability-based HRM on firm performance: the mediating role of innovation ambidexterity. *Industrial Marketing Management* 59, pp. 131–143. doi: 10.1016/j.indmarman.2016.02.018

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Education

ORIGINAL RESEARCH
published: 10 June 2022
doi: 10.3389/feduc.2022.885853

Performing Meaningful Movement Analysis From Publicly Available Videos Using Free Software – A Case of Acrobatic Sports

Pui Wah Kong, Alexiaa Sim* and Melody J. Chiam*

Physical Education and Sports Science Academic Group, National Institute of Education, Nanyang Technological University, Singapore, Singapore

This paper illustrates how movement analysis could be performed using publicly available videos and freeware to generate meaningful information for sports practitioners and researchers. Using acrobatic sports as a case, we performed kinematic analysis on 206 YouTube videos of high-level competitions in diving and gymnastics using Kinovea. Results revealed good to excellent inter-rater reliability of variables analyzed. Significant differences in angular speed ($p < 0.001$, $\eta^2_p = 0.213$) and flight time ($p < 0.001$, $\eta^2_p = 0.928$) were found among eight different events. Divers had longer flight time ($p < 0.001$, $\eta^2_p = 0.569$) and were somersaulting faster than gymnasts ($p = 0.021$, $\eta^2_p = 0.026$). Angular speed was higher in tuck than pike somersaults ($p < 0.001$, $\eta^2_p = 0.214$). Shorter the flight time was significantly correlated with faster angular speed ($rho = -0.533$, $p < 0.001$) in gymnastics events. Coaches and scientists can consider applying the proposed method to monitor the athletes' performance and to identify errors (e.g., insufficient flight time). The kinematics measurements can also be used to guide the transition plan across different apparatus and categories (e.g., 10-m platform to 3-m springboard). In conclusion, the present study highlights the potential of using readily available information and open-source freeware to generate scientific data for sports applications. Such data analysis approach can accommodate a wide range of video qualities, is easily accessible, and not restricted by situations such as social distancing, quarantine, lockdown or other restrictive measures.

Keywords: YouTube, Kinovea, somersaults, sport, diving, gymnastics

OPEN ACCESS

Edited by:
Aaron Williamon, Royal College of Music, United Kingdom

Reviewed by:
Germina Cosma, University of Craiova, Romania
José Eugenio Rodríguez-Fernández, University of Santiago de Compostela, Spain

****Correspondence:***
Pui Wah Kong
puiwah.kong@nie.edu.sg
Alexiaa Sim
nie184704@e.ntu.edu.sg

Specialty section:
This article was submitted to Assessment, Testing and Applied Measurement, a section of the journal Frontiers in Education

Received: *28 February 2022*
Accepted: *18 May 2022*
Published: *10 June 2022*

Citation:
Kong PW, Sim A and Chiam MJ (2022) Performing Meaningful Movement Analysis From Publicly Available Videos Using Free Software – A Case of Acrobatic Sports. Front. Educ. 7:885853. doi: 10.3389/feduc.2022.885853

INTRODUCTION

In the field of sports science, biomechanical analysis of athletes' movements can provide useful information for coaches, scientists and athletes to evaluate performance, identify errors, and develop training and competition plans (Ae, 2020; Barbosa et al., 2021). To obtain detailed and accurate movement data, athletes are often tested comprehensively in a biomechanical laboratory using expensive and advanced equipment such as multi-camera 3D motion capture systems and force platforms (Ferdinands et al., 2009; Augustus et al., 2021). COVID-19 and the associated restrictions impose great challenge to sports (Bobo-Arce et al., 2021). Many laboratories around

the world were closed for a sustained period of time. Others required users to adhere to strict safe management measures such as limiting the number of people per group, maintaining social distancing, and avoiding physical contact. These restrictions make it difficult if not impossible to perform standard biomechanical movement analysis in the laboratory. Fortunately, there are alternative solutions to perform meaningful movement analysis in sports (Morgulev et al., 2018; Pan et al., 2021). This paper presents a method that makes use of publicly available videos and free software to conduct biomechanical movement analysis. When using publicly available videos, the researchers have no control over the video recording process such as camera position, sampling rate, shutter speed and field of view. The wide range of video qualities imposes challenges on movement analysis, for example, blurry images due to slow shutter speeds and temporal errors associated with low frame rate. Despite these technical difficulties, it is yet possible to perform remote video analysis by carefully selecting variables that can be determined with good confidence even when the videos were taken from different sources with sub-optimal qualities. The present study will illustrate a practical data analysis approach that can accommodate a wide range of videos of varied qualities.

To demonstrate the idea of video-based movement analysis, a case of acrobatic sports was selected. Acrobatic sports athletes such as divers and gymnasts performed highly skilled moves across different events (e.g., platform, springboard, floor, and gymnastic apparatus) in specialized competition venues (e.g., gymnasium and diving pool) (King and Yeadon, 2004; Vladimir et al., 2015; Sayyah et al., 2018, 2020; King et al., 2019, 2022). While great efforts have been made to develop inertial measurement units for on-water data collection (Walker et al., 2017, 2019), kinematic information of diving movements are mostly obtained from videos recorded at diving pools (Sayyah et al., 2018, 2020; King et al., 2019, 2022). Similarly in gymnastics, video-based analysis has been a useful means to assess movement quality, provide feedback and inform training plans (Mizutori et al., 2021; Fujihara, 2022). Analyzing publicly available competition videos is a promising approach to offer insights into these very difficult moves that are approaching human limit. At high-level international competitions, it is interesting to observe that acrobatic sports athletes perform similar somersaulting moves despite competing in different sport events. For example, triple back somersaults are seen in floor routines and dismounts from apparatus in artistic gymnastics as well as the armstand dive group from the 10-m platform. Although the numbers of rotation appear similar, one cannot assume that the biomechanical characteristics of these moves are the same for divers and gymnasts. Factors that can influence the somersaulting speed may include the approach speed, compliance and height of the take-off and landing surfaces, flight time available to complete the required number of somersaults, body shape, and moment of inertia of the athlete while somersaulting (Hamill et al., 1986; Gronbech, 1993; Miller and Sprigings, 2001; Cheng and Hubbard, 2008; Kong, 2010; Mkaouer et al., 2013; Mikl, 2018). Previous biomechanical studies on acrobatic sports typically reported detailed kinematics of one move (Vladimir et al., 2015; Park and Yoon, 2017; Walker et al., 2019) or comparing across a few somersaulting moves (Sanders and Gibson, 2003; Walker et al., 2017; King et al., 2022) within a single sport. To the authors' best knowledge, no studies have compared the somersaulting moves performed by athletes across different acrobatic sports events.

The purpose of this study was to use acrobatic sports as a case to illustrate a video-based movement analysis approach using open-source freeware to perform meaningful analysis from publicly available footages of varied qualities. In specific, key biomechanical variables of divers and gymnasts when performing similar acrobatic moves in real competition settings would be compared. Findings from this study can help us better understand the mechanical demand of the highly difficult skills and the transferability of skills in acrobatic sports. Such information may also be useful for coaches and sports scientists to develop training plan and select movements for competitions. The proposed data analysis approach has economic and practical benefits for scientists, coaches and athletes because the measurements can potentially be performed by anyone in any location, without the need to access specialized and expensive facilities, equipment or software.

MATERIALS AND METHODS

This study was approved by the Nanyang Technological University Institutional Review Board (Protocol Number: IRB-2021-02-029). Publicly available videos of from high-level diving and gymnastics competitions in the past 6 years were retrieved for kinematics analysis.

Selection of Video

To systematically retrieve videos for analysis, a set of pre-set criteria should be developed. These criteria should be objective and clearly defined such that other analysts can repeat the search if they follow the described methods. In the present study, the acrobatic sports videos were included for analyses if they fulfilled all the following criteria:

(1) Publicly available on the YouTube site, uploaded by the competition organizers (e.g., Olympics YouTube Channel) and/or the international federation of diving or gymnastics [e.g., Fédération Internationale de Natation (FINA), Fédération Internationale de Gymnastique (FIG)].
(2) Athletes performed either a 3.0 or 3.5 back somersault move in a tuck or pike position without any twists in the diving or gymnastics competitions. These moves were considered very difficult moves that were near the athletes' limits. We did not limit the move to only the triple somersault because divers do not usually land feet first at high level competition. In the backward dive group, divers typically perform 3.5 somersaults and entered the water with their hands first to minimize splash.
(3) The move had to be executed in a high-level competition setting (e.g., Olympic Games, world championships, European championships, British Gymnastics Championships, United States Gymnastics

Championships, China Diving Star, and China National Games).
(4) The competitions took place in the past 6 years (2016–2021).
(5) The videos must be in normal time and not slow motion.

For diving videos, the search terms on the YouTube site included dive number "207C" (backward 3.5 somersaults in tuck), "207B" (backward 3.5 somersaults in pike), "626C" (arm stand 3.0 somersault in tuck), "626B" (arm stand 3.0 somersault in pike), followed by the term "diving". For gymnastics videos, the terms used were "triple somersault," "triple back tuck," "triple back pike," followed by the term "gymnastics." In addition, the search also included terms related to official high-level competitions such as "Olympics Gymnastics," "Olympics Diving," "World Championship Gymnastics," "FINA Diving," and "Diving World Cup." The selected videos consisted of two categories in diving – 3-m springboard and 10-m platform (backward group and armstand group), and five categories in gymnastics – floor exercise, rings, horizontal (high) bar, double-mini trampoline, and tumbling. Both male and female athletes were included in the video search process.

In the present study, a tuck position was considered when there is hip and knee flexion whereby the knees should be bent to a maximum range and pulled toward the chest (Mitchell et al., 2002). A pike position was considered when there is hip flexion (less than 90°) and straight legs, and this is also known as a closed pike position (Mitchell et al., 2002). If the athlete attempted to somersault in a pike position but bent the knees midway, the video would be excluded from the analysis because no distinct body position could be identified. Every eligible video was extracted from YouTube through screen recording at 30 Hz through these two software – OBS Studio 27.0.1, Xbox Game Bar 5.721.6282.0. This allows slow motion viewing, frame-to-frame navigation, and kinematic analysis.

Biomechanical Analysis

Biomechanical analysis of the extracted videos was performed using the Kinovea software (version 0.8.15, available for download at: http://www.kinovea.org). Kinovea is an open source software under GPLv2 license for 2D motion analysis. It can be downloaded for free, and is widely used in the field of kinematic analyses in sports (Nor Adnan et al., 2018). This software has been demonstrated to show good accuracy in measuring angles and distances (Puig-Diví et al., 2019). Specific to acrobatic sports, excellent intra-rater and inter-rater reliability was found in angle measurements and identification of key events in dynamic gymnastics movements (Khong and Kong, 2016). Moreover, the authors reported good agreement among three independent raters with varying levels of gymnastics experience suggesting that 2D video analysis can be readily adopted by users regardless of their personal experience in acrobatic sport.

When selecting biomechanical variables to be analyzed, we need to consider the recording rate, camera position and scaling factors. Temporal or timing variables can be easily obtained as long as the key events (e.g., takeoff and landing) can be visually seen from the videos. The accuracy of timing data will depend on the video recording frame rates (Tay and Kong, 2018). Angular kinematic measurements (e.g., angle and angular speed) require the movement plane to lie perpendicular to the optical axis of the camera (Tay and Kong, 2018). Any out-of-plane angle measurements would result in perspective errors. Linear kinematic variables (e.g., displacement and velocity) are the most difficult ones since calibration procedures must be performed to obtain appropriate scaling factors. In some sports, the known dimensions of equipment (e.g., billiards table) can be used for calibration purposes (Pan et al., 2021).

In the present study on acrobatic sports, there are technical challenges associated with the online videos which were captured under different camera settings with varied qualities. First, the cameras were moving to follow the athletes and that no consistent calibration objects were available. Second, the cameras were not always positioned perpendicular to the movement plane and therefore angle measurements would be subjected to perspective errors. Third, most videos were recorded at low sampling rate and slow shutter speeds. This temporal limitation restricted us from analyzing frame-by-frame instantaneous kinematics such as joint angle-time history or peak angular speed. Having consider these technical challenges, the present study proposed to measure flight time and average angular speed of the somersaulting moves. Flight time is a temporal variable which does not require calibration. Compared with instantaneous kinematic measurements (e.g., maximum hip flexion angle, peak angular speed), taking the average angular speed over a longer period of time would be more accurate because this method was less affected by the temporal errors associated with low frame rate. We believe that flight time and average angular speed information can be determined with good confidence across videos of varied qualities to facilitate remote analysis of acrobatic sports. Details of each variable are described below.

Flight Time

Flight time was defined as the duration of the athlete in the air (**Figure 1**). It started from last video frame that any part of the athlete's body is in contact with the take-off surface (springboard, platform, floor, and tumbling track) or apparatus (rings, high bar, and mini trampoline) before somersaulting in the air, to the first video frame that the athletes' hands breaking the water (diving) or feet touching the landing surfaces (gymnastics).

Average Angular Speed

Previous studies illustrated that when divers performed multiple somersaulting dives, the angular speed increased sharply in the beginning of the flight phase, remained fairly unchanged in the somersault phase when a tight body position was formed, and decreased toward the end when the body extended from the tight somersault position (Walker et al., 2017, 2019). In the present study, the average angular speed was quantified while the athlete was somersaulting in a tight and rigid body position with minimal joint flexion or extension movements. While there are variations in the way how divers and gymnasts hold a tuck or pike position (Mikl, 2018), we considered a tight and rigid position was assumed when the athlete pulled the legs closest to the chest during the rotation. Once the athlete loosened from the tight tuck

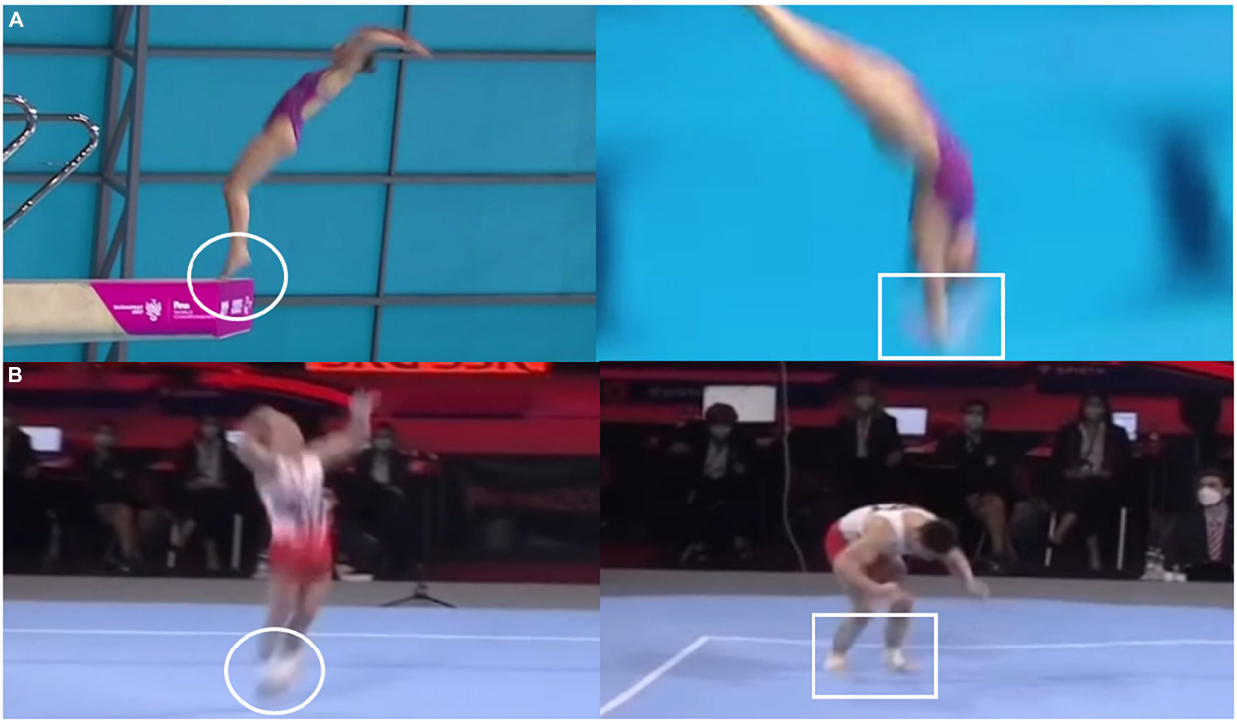


FIGURE 1 | An example of measurement of flight time in **(A)** diving and **(B)** gymnastics. The start of flight time is denoted by the circle, while the end of flight time is denoted by the rectangle.

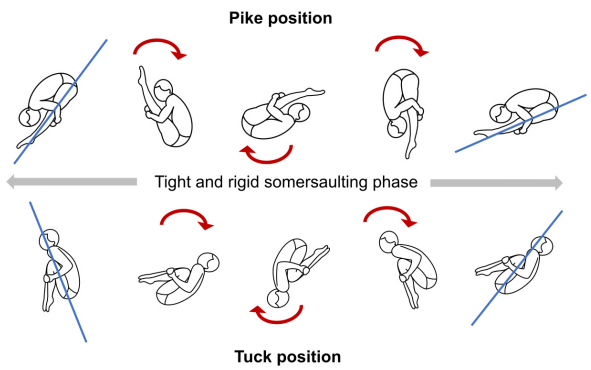


FIGURE 2 | A straight line passing through the hip and knee joints of the athletes was drawn in indicate the orientation of the athlete in the beginning and end of the tight and rigid somersaulting phase.

or pike position and started to extend the body, the tight and rigid body was no longer assumed. Among all back somersaulting videos retrieved for analysis, the tight and rigid body position was commonly seen during or slightly after the completion of the first 360° rotation. A straight line passing through the athlete's hip and knee joints were drawn to represent the orientation of the athlete in first and last frame of the tight and rigid shape duration (**Figure 2**). Based on these two lines, the total degree of rotation that the athlete performed were measured using the angle tool in the Kinovea software.

The average angular speed was calculated as the total degrees of rotation divided by the time in which the athlete held a tight and rigid body position. In some tumbling videos, the cameras were not placed perpendicular to the movement plane, and hence measuring angles from these camera views would lead to prospective errors. To overcome this limitation, we determined the time required for the athlete to complete either 0.5, 1.0, or 1.5 somersaults in a tight and rigid body position. We then use the known degrees of rotation of 180°, 360°, or 540° in the computation of angular speed to avoid the perspective errors associated with camera position.

Video analyses were performed by two raters under the guidance of the senior author who has a Ph.D. in sports biomechanics and competitive background in springboard diving. All gymnastics videos were analyzed by rater 1 who has a bachelor's degree in sports science and 10 years of competitive background in artistic gymnastics. All diving videos were processed by rater 2 who has a bachelor's degree in sports science, 4 years of competitive background in artistic gymnastics, and 1 year of competitive background in cheerleading.

Statistical Analysis

A total of 206 videos (105 diving and 101 gymnastics) were retrieved, including 150 in tuck and 56 in pike positions (**Figure 3**). The athletes took off from their feet in most moves (75 diving and 95 gymnastics), with 30 armstand dives and 6 dismounts from apparatus. All divers entered the water

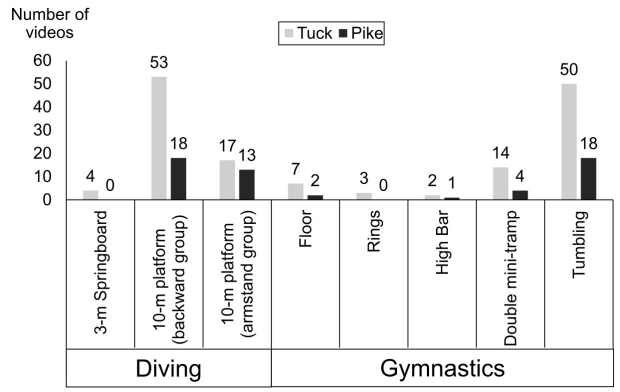


FIGURE 3 | Number of tuck and pike multiple (3.0 or 3.5) back somersault moves in competitive diving (n = 105) and gymnastics (n = 101) videos retrieved in years 2016–2021 (total = 206 video).

with their hands first while all gymnasts landed the triple somersaults on their feet.

All gymnastics triple somersaults were performed by male gymnasts and hence no female data were available. In diving, there were only four female divers performing backward 3.5 somersaults from the 10-m platform. Since the flight time (1.40–1.74 s) and angular speed (831°/s to 1169°/s) of these four female divers fell within the range of those in their male counterparts (flight time: 1.39–1.80 s; angle speed: 818°/s to 1,200°/s), all divers' data were analyzed as a group regardless of sex.

To check the inter-rater reliability, rater 1 and rater 2 independently analyzed a sub-sample of 34 videos (17 diving and 17 gymnastics). The intraclass correlation coefficients (ICC) of average angular speed and flight time was calculated using IBM SPSS Statistics for Windows (Version 27.0. Armonk, NY, United States: IBM Corp.). Subsequently, the standard error of measurements (SEM) of each variable were computed.

For the main dataset, JASP (version 0.14.1) was used for statistical analysis with significance level set at $p < 0.05$. The two outcome variables were angular speed and flight time. As the data were not normally distributed, descriptive data are expressed as median [interquartile range (IQR)]. Data were log transformed prior to running Analysis of Variance (ANOVA). First, a One-way Analysis of Variance (ANOVA) was performed to detect differences across all diving and gymnastics events. In addition, two-way ANOVA was used to compare between sports (diving/gymnastics) and body position (tuck/pike). Effect size was calculated as partial eta-squared (η^2_p). *Post hoc* tests with Bonferroni adjustments were performed where appropriate. To examine the relationship between angular speed and flight time, Spearman rho's correlation analysis was performed.

RESULTS

The inter-rater reliability between rater 1 and rater 2 were *excellent* for flight time (ICC = 0.994, SEM = 0.02 s) and *good* for angular speed (ICC = 0.788, SEM = 54°/s). Typical tight and rigid body positions during the somersaulting phase of divers and gymnasts are illustrated in **Figure 4**. In general, divers hold a tighter shape than gymnasts especially in the pike position. There were significant differences in angular speed ($p < 0.001$, $\eta^2_p = 0.213$) and flight time ($p < 0.001$, $\eta^2_p = 0.928$) among the different events in diving and gymnastics (**Table 1**). The 3.5 back somersaults in 3-m springboard diving were the fastest, while the triple back somersaults in double mini-tramp was the slowest. Flight time was the longest in 3.5 backward somersaults from the 10-m platform and the shortest when dismounting from rings.

When comparing between sports, ANOVA results revealed that divers had longer flight time ($p < 0.001$, $\eta^2_p = 0.569$) and were somersaulting faster than gymnasts ($p = 0.021$, $\eta^2_p = 0.026$, **Table 2**). While the flight time did not differ between body positions ($p = 0.627$), the angular speed was higher in tuck than pike somersaults ($p < 0.001$, $\eta^2_p = 0.214$, **Table 2**). No significant interaction effect between sport and body shape was found in angular speed or flight time variables.

Regarding the relationship between angular speed and flight time, there was a significant negative correlation in gymnastics ($rho = -0.533$, $p < 0.001$) but not in diving ($rho = -0.016$, $p = 0.873$). The shorter the flight time available, the faster the gymnasts' angular speed in order to complete required degrees of rotation (**Figure 5**).

DISCUSSION

Using acrobatic sports as an example, this study illustrated how movement analysis can be performed using publicly available videos and free software to offer insights into high performance sports. The findings showed that the biomechanical characteristics of multiple back somersaults differed among various diving and gymnastics events despite the numbers of rotation were similar. The source videos were real performance of elite athletes at high-level competitions, providing greater ecological validity over analyzing training videos (Sayyah et al., 2018, 2020; King et al., 2022) or laboratory biomechanical tests (Mkaouer et al., 2013; Mizutori et al., 2021; Fujihara, 2022). While the videos were recorded with different camera settings and of varied qualities, the proposed data analysis approach yields good to excellent inter-rater reliability when measuring average angular velocity and flight time of acrobatic moves. We demonstrate that it is possible to make good use of readily available information such YouTube videos to enhance our understanding of sport movements at low cost. Although we used acrobatic sports as a case to illustrate the method, the potential applications of video-based analysis are certainly not limited to diving and gymnastics. In surfing, for example, time-motion analysis from videos have also been shown to reveal the timing of various activity that are deem useful for the development of tailored condition training programs (Minghelli et al., 2019).

Comparison Across Diving and Gymnastics Events

The present study found significant differences in angular speed and flight time across eight events in competitive diving and

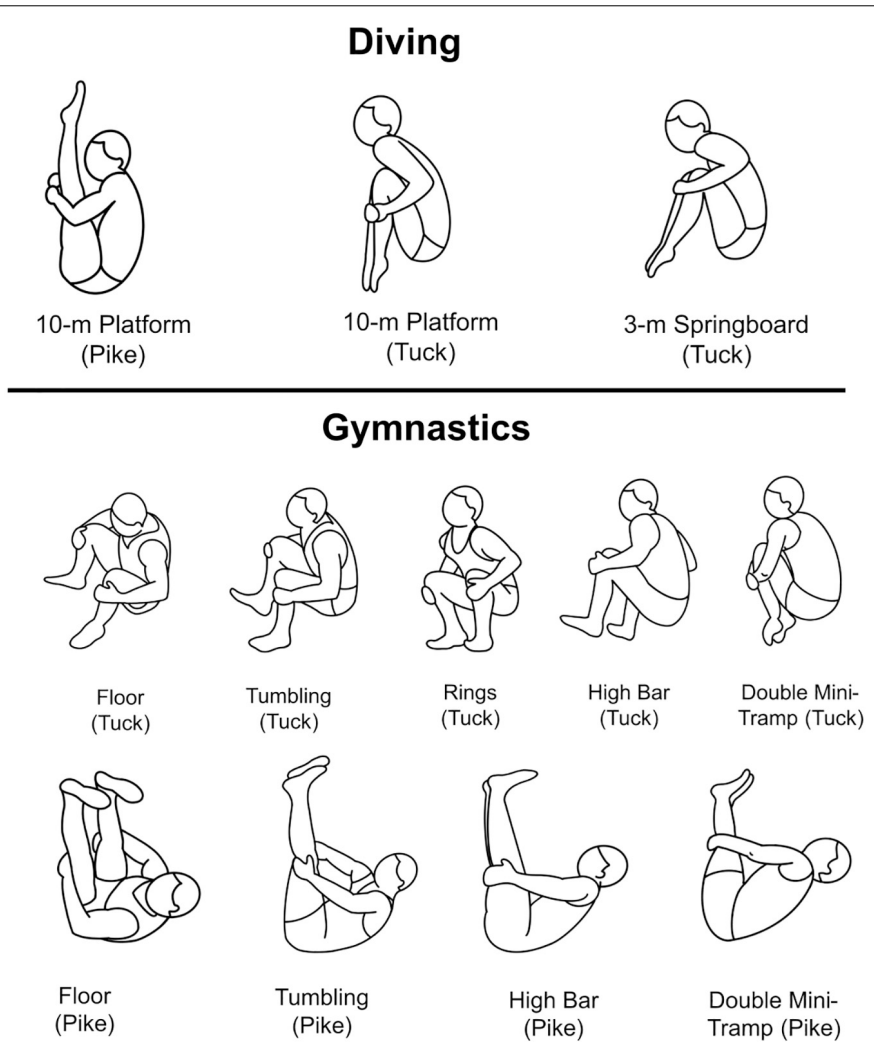


FIGURE 4 | Illustration of typical tuck and pike positions of divers and gymnastics during the tight and rigid somersaulting phase of 3.0 or 3.5 back somersaults at high-level competitions.

gymnastics. The range of angular speeds (median 900°/s to 1,122°/s) observed from our video analysis were similar to those quantified using inertial measurement units for forward 3.5 piked somersaults (855°/s to 904°/s) and forward 4.5 tuck somersaults (1,090°/s) in 3-m springboard diving (Walker et al., 2019). The observed differences in biomechanical characteristics can be explained by the varied approach speeds, compliance and height of the take-off and landing surfaces. For example, tumblers and artistic gymnasts have long run-up to build up momentum for their somersaults (Mkaouer et al., 2013) where divers perform backward and armstand dives from a stationary start (Hamill et al., 1986; Cheng and Hubbard, 2008). Platform divers have 10-m distance to travel before entering the water while 3-m springboard divers can take advantage of the springboard to increase flight time (Sanders and Wilson, 1988; Miller et al., 1989).

It was observed that divers rotated faster than gymnasts when performing multiple back somersaults of similar numbers of rotation. With the 3.5 back somersaults in 3-m springboard diving being the fastest move (1,122 [85]°/s), it is not surprising that only four male divers performed this move at high-level competitions. While four female divers chose to perform 3.5 back somersault from the 10-m platform, females are yet to master this dive on the 3-m springboard. Previous studies on springboard diving observed that males achieved greater heights than females (Miller et al., 1989; Sanders and Gibson, 2003). When performing 2.5 backward somersaulting dives, females were suggested to increase the amount of work done on the springboard to generate sufficient dive height and angular momentum (Sanders and Gibson, 2003).

The 3.5 back somersaults back somersaults from the 10-m platform was the second fastest move (987 [987, 106]°/s) and had the longest flight time (1.70 [0.06]°/s) among all events. This was the only event in which female athletes were able to perform 3.5 backward somersaults at high-level competitions. Starting from a stationary position with no run-up, platform divers must take off

TABLE 1 | Angular speed and flight time of multiple back somersaults in diving and gymnastics events.

Variables		Events	Number of Somersaults	Median [IQR]	ANOVA		*Post hoc*
					p	η^2_p	
Angular speed (°/s)	D_1	Diving: 3-m Springboard	3.5	1,122 [85]	<0.001	0.213	$D_1 > D_3, G_3, G_4, G_5$
	D_2	Diving: 10-m Platform (Backward Group)	3.5	987 [106]			$D_2 > G_5$
	G_1	Gymnastics: Rings	3.0	983 [45]			
	D_3	Diving: 10-m Platform (Armstand Group)	3.0	977 [132]			$D_3 > G_5$
	G_2	Gymnastics: Floor	3.0	973 [300]			$G_2 > G_5$
	G_3	Gymnastics: Tumbling	3.0	973 [62]			$G_3 > G_5$
	G_4	Gymnastics: High Bar	3.0	908 [23]			
	G_5	Gymnastics: Double mini-Trampoline	3.0	900 [63]			
Flight time (s)	D_1	Diving: 3-m Springboard	3.5	1.47 [0.02]	<0.001	0.928	$D_2 > D_1 > G_1, G_2, G_3$
	D_2	Diving: 10-m Platform (Backward Group)	3.5	1.70 [0.06]			D_2 > all other 7 events
	G_1	Gymnastics: Rings	3.0	1.10 [0.04]			G_1 < all other 7 events
	D_3	Diving: 10-m Platform (Armstand Group)	3.0	1.40 [0.07]			$D_3 > G_1, G_2, G_3$
	G_2	Gymnastics: Floor	3.0	1.26 [0.04]			$D_1, D_2, D_3, G_3, G_4, G_5 > G_2 > G_1$
	G_3	Gymnastics: Tumbling	3.0	1.30 [0.04]			$D_1, D_2, D_3, G_5 > G_3 > G_1, G_2$
	G_4	Gymnastics: High Bar	3.0	1.34 [0.06]			$D_2 > G_4 > G_1, G_2$
	G_5	Gymnastics: Double Mini-Trampoline	3.0	1.47 [0.09]			$D_2 > G_5 > G_1, G_2, G_3$

Event orders are arranged from the fastest to the slowest based on median angular speeds for diving ($D_1 > D_2 > D_3$) and gymnastics ($G_1 > G_2 = G_3 > G_4 > G_5$). Data are reported as median [interquartile range]. Raw data were log transformed prior to running ANOVA.

TABLE 2 | Comparison of angular speed and flight time by sports (diving/gymnastics) and body positions (tuck/pike).

Variables	Body position	Diving	Gymnastics	ANOVA					
				Sports		Body positions		Interaction	
				p	η^2_p	*p*	η^2_p	*p*	η^2_p
Angular speed (°/s)	Tuck	1019 [107]	973 [100]	0.021	0.026	<0.001	0.214	0.080	0.015
	Pike	916 [92]	900 [87]						
Flight time (s)	Tuck	1.67 [0.26]	1.30 [0.09]	<0.001	0.569	0.627	0.001	0.159	0.010
	Pike	1.66 [0.26]	1.34 [0.13]						

Data are reported as median [interquartile range]. Raw data were log transformed prior to running ANOVA.

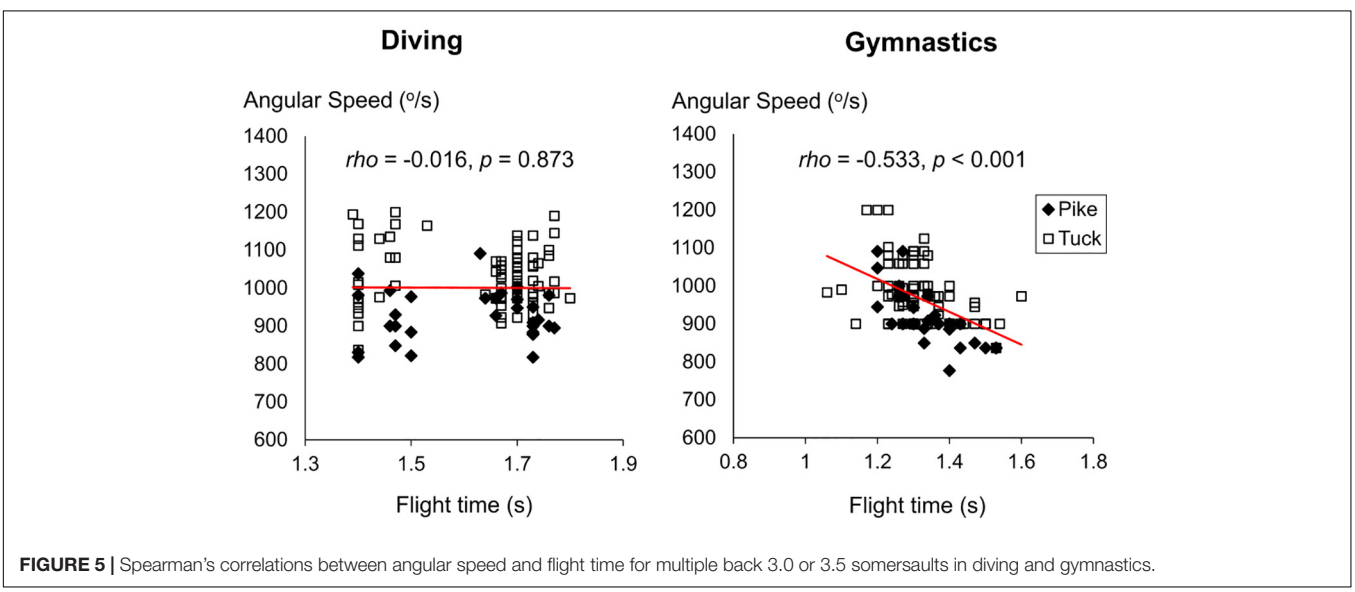


FIGURE 5 | Spearman's correlations between angular speed and flight time for multiple back 3.0 or 3.5 somersaults in diving and gymnastics.

powerfully to generate sufficient height and angular momentum for rotation (Sanders and Wilson, 1988). The 10 m free-falling distance from the platform to the water provides the advantage of long flight time even if the athletes do not generate a high vertical velocity at take-off. Interestingly, even without a powerful take-off from the legs, the flight time of the armstand 3.0 somersaults (1.44 [1.42, 1.46]°/s) was comparable to other events such as 3-m springboard (1.48 [1.43, 1.53]°/s) and double-mini trampoline (1.47 [1.44, 1.50]°/s). The advantage of having a long flight time resulted in many divers being able to perform highly difficult moves from the 10-m platform.

In gymnastics, the triple back somersaults dismount from rings was the fastest move (983 [45]°/s) and only three male gymnasts managed to perform this difficult move. In the literature, there are very few detailed biomechanical studies on triple back somersaults in gymnastics. One case study on the double back somersaults in women's artistic gymnastics floor routine reported a range of flight heights of the center of mass from 1.85 to 2.06 m across a few competitions (Vladimir et al., 2015). The author did not examine the flight time or angular speed during the somersaulting phase and hence no direct comparison could be made. In the current study, the flight time for the triple back somersault on floor was rather short (1.26 [0.04]s, the second shortest among all events). As a result, the gymnasts must rotate at a high angular speed in order to complete the required number of rotations before extending the body to land safely on their feet. Gymnasts build momentum from the run-up, round-off and back handspring(s) before taking off to perform the triple back somersaults. It is postulated that the full tumbling sequence in floor routines contributed to generating sufficient angular momentum required to complete the difficult move of triple back somersaults despite the short flight time.

Why Do Divers Rotate Faster Than Gymnasts?

With the exception of rings dismount, the multiple back somersaults in diving events required faster angular speeds than all other gymnastics events (**Table 1**). As illustrated in **Figure 3**, divers tend to hold a tighter shape during the somersaulting phase. Holding a tight body position while somersaulting require muscular efforts from the arms, torso, and legs (Gronbech, 1993; Miller and Sprigings, 2001; Kong, 2010; Mikl, 2018). Mikl (2018) used an 11-segment model to calculate the joint forces and moments required to maintain a tuck and pike shape while somersaulting. For both body positions, it was found that the greatest joint moments are at the hip, followed by the pelvis-abdomen segment. The author discussed that isometric strength at various joints may play a role in limiting the diver or gymnast's ability to hold a rigid body shape while somersaulting. In addition to strength, the flexibility of the athletes can also determine whether one can form a tight shape, especially for the pike position. Gymnastics is a sport that is generally characterized by high levels of strength and power relative to body weight, as well as high flexibility (Donti et al., 2014), especially for male artistic gymnasts who are required to perform strength elements in various events (Jemni et al., 2006; Mkaouer et al., 2018). It is unlikely that gymnasts are limited by their isometric strength or poor flexibility to form tighter tuck and pike shapes during the somersaults.

One possible reason for why gymnasts do not hold a tighter shape to rotate faster may be related to the margin of error for landing. In diving, if a diver makes a mistake in the timing of come-out from the tight somersaulting shape, he/she cannot enter the water in a desirable near-vertical position. While a poor dive entry may lead to a low performance score and sometimes physical pain or bruising, the chance of severe injuries is low when entering the water. In gymnastics, the margin of error for landing is very narrow (Hiley and Yeadon, 2003, 2005). A poor landing resulting from mis-judging the timing of extending the body from a tight somersault position may lead to severe injuries or even fatal consequences. Thus, gymnasts may be more cautious toward pushing for the limit in their somersaulting speed. Vision may also play a role (Natrup et al., 2020, 2021; Barreto et al., 2021) as it is difficult to spot for landing when the athlete is rotating in a deeply closed pike position. During a back tuck somersault on a trampoline, the gymnasts' gaze behavior depended on their position in the bed (Natrup et al., 2020). Kicking-out from a somersault early allows gymnasts to orient themselves to prepare for landing. A final point to note is that there is no advantage in the performance score for holding a very tight shape. According to gymnastics code of points for gymnastics, there will be no deduction as long the hip angle is less than 90° for the pike position. It is therefore not surprising to note that the typical pike position across all gymnastics events were much more open than that in 10-m platform diving (**Figure 3**). Within the requirements in code of points, gymnasts can regulate the tightness of the shape and not necessarily tucking or piking to the maximum extent. Holding a not-too-tight shape would mean that the gymnasts rotate at a slower angular speed, allowing a larger window of error to open their body and land safely on their feet.

Tuck Versus Pike Positions

To maintain a tight body position while somersaulting at fast speeds, muscular efforts from the arms, torso and legs are required to provide the necessary centripetal force (Gronbech, 1993; Miller and Sprigings, 2001; Kong, 2010; Mikl, 2018). Differences are noted in the way how divers and gymnast hold a tuck or pike shape while somersaulting and the joint moments required to hold the shape are influenced by the location and direction of how the arms pulling on the leg (Mikl, 2018). Since the angular momentum of the athlete in the flight phase is constant due to no external torque, the tightness of the body shape can substantially influence the rotational speed. The present study showed that somersaulting in a tuck shape was faster than that in a pike shape in which the athlete's legs are extended and further away from the center of mass. This increases the moment of inertia about the transverse axis, making it more difficult to somersault (Mikl, 2018). Interesting, flight time was similar regardless of the body position during somersaulting. This reflects that athletes who were able to perform the same number of somersaults in a pike shape did not compromise the

height of move, compared with those who chose to somersault in the easier tuck shape. These results align with previous data on female divers which showed similar dive height (mean value, pike: 0.82 m, tuck: 0.81 m) and flight time (pike: 1.29 s, tuck: 1.29 s) regardless of body positions when performing backward 2.5 somersaults from a 3-m springboard (Sanders and Gibson, 2003). When divers and gymnasts progress from tuck to pike somersaulting moves, they should aim to maintain the same jump height and expect a slower rotating speed.

Relationship Between Angular Speed and Flight Time

Among all gymnastics events, shorter the flight time was significantly correlated with faster angular speed. The shorter the flight time available, the faster the gymnasts must rotate to complete required degree of rotation. When taking off from a compliant surface – double mini-trampoline, the gymnasts gained the longest flight time and therefore could rotate more slowly when compared with the other gymnastics events. In diving, however, there was no significant relationship between angular speed and flight time. The lack of relationship may be due to the narrow range of flight time data in diving when compared with gymnastics (**Figure 5**). Only four dives were performed from the 3-m springboard and the other 101 dives were from 10-m platform in which the flight time was rather similar due to the long free-falling distance. From the same 10-m platform, divers performed 0.5 less somersault in the armstand group than the backward group. Despite significant less flight time in the armstand 3.0 somersaults (1.40 [0.07]s) than backward 3.5 somersaults (1.70 [0.06]s), the somersaulting speed were similar in these two dive groups (armstand: 977 [132]°/s; backward: 987 [106]°/s). Thus, when combing all diving moves as a group, the relationship between angular speed and flight time was unclear.

Transferability of Acrobatic Sports Skills

Diving coaches and athletes should be aware the high somersaulting speeds required to perform multiple back somersaults from the 3-m springboard. If a diver can perform 3.5 back somersaults from 10-m platform, he/she may not be able to execute the same move on 3-m springboard. Similarly, gymnasts can compete in different events and should be aware of the different rotational requirements. A gymnast who can dismount with a triple back somersault from the high bars may not be able to do so from the rings which require a faster angular speed. As different disciplines of acrobatic sports share common techniques and training approaches, some athletes may crossover from one sport to another. There are examples of retired gymnasts who pick up diving and continue to compete after. While we can reasonably expect high transferability of similar acrobatic skills, one previous study has showed that the kinematics of divers performing somersaulting moves differed between dry-land training and actual on-water diving (Barris et al., 2013). Thus, it is important to acknowledge the subtle biomechanical differences between similar acrobatic moves which can be critical at high performance level.

Limitations

There are a few limitations to the present study. First, the kinematic analysis was done on publicly available YouTube videos that were screen recorded at 30 Hz. Some videos were not filmed with fast shutter speeds and hence the extracted images were blurry at times. While higher quality videos recorded at higher sampling rate will allow more accurate kinematic analysis, access to data collection at high level competitions are often restricted. The kinematic video analysis in the present study was performed in a consistent manner and the inter-rater reliability was good to excellent. It is believed that the data obtained could provide sufficiently accurate and reliable information and allow a fair comparison across different acrobatic sports events. Second, the present study observed that divers in general hold a tighter shape than gymnasts and the typical body positions were illustrated in **Figure 4**. However, we were unable quantify the moments of inertia of the athletes owing to the limitation of video analysis.

CONCLUSION

This study demonstrated how movement analysis of elite sport performance can be done via analysis of publicly available videos using free software without the need of specialized laboratory, equipment or software. In the case of acrobatic sports, findings from this study showed that the biomechanical variable multiple back somersaults differed among various diving and gymnastics moves that appeared similar in the numbers of rotation. In general, divers rotated faster than gymnasts and that somersaulting in a tuck position has higher angular speed that the pike position. Gymnasts tended to hold a less tight shape than divers, possibly because a slower somersaulting speed allows larger margin of error in the timing of opening up for safe landing. Coaches can consider applying the proposed method to monitor the athletes' performance and to identify errors. For example, if a gymnast struggles to complete a move, is it due to insufficient flight time or slow angular speed? The kinematics measurements can also be used to guide the transition plan across different apparatus (e.g., dismount from high bar and rings) and categories (e.g., 10-m platform to 3-m springboard). In conclusion, the present study highlights the potential of using readily available information and open-source freeware to generate scientific data for sports applications. Such data analysis approach can accommodate a wide range of video qualities, is easily accessible, and not restricted by situations such as social distancing, quarantine, lockdown or other restrictive measures.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: NIE Data Repository (https://doi.org/10.25340/R4/HRQ4PS).

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Nanyang Technological University Institutional Review Board (Protocol Number: IRB-2021-02-029). Written informed consent for participation was not required for this study in accordance with the National Legislation and the Institutional Requirements.

AUTHOR CONTRIBUTIONS

AS and MC performed the material preparation and data collection and analysis. PK wrote the first draft of the manuscript. All authors commented on previous versions of the manuscript, contributed to the study conception and design, and read and approved the final manuscript.

FUNDING

AS was supported by the Nanyang President's Graduate Scholarship, Nanyang Technological University.

ACKNOWLEDGMENTS

We would like to thank Stefanie C. Lee and Jing W. Pan for their assistance with the preparation of figures.

REFERENCES

Ae, M. (2020). The next steps for expanding and developing sport biomechanics: winner of the 2019 ISBS Geoffrey Dyson Award. *Sports Biomech*. 19, 701–722. doi: 10.1080/14763141.2020.1743745

Augustus, S., Hudson, P. E., and Smith, N. (2021). The effect of approach velocity on pelvis and kick leg angular momentum conversion strategies during football instep kicking. *J. Sports Sci*. 39, 2279–2288. doi: 10.1080/02640414.2021.1929008

Barbosa, T. M., Barbosa, A. C., Simbaña Escobar, D., Mullen, G. J., Cossor, J. M., Hodierne, R., et al. (2021). The role of the biomechanics analyst in swimming training and competition analysis. *Sports Biomech.* [Epub ahead of print]. doi: 10.1080/14763141.2021.1960417

Barreto, J., Casanova, F., Peixoto, C., Fawver, B., and Williams, A. M. (2021). How task constraints influence the gaze and motor behaviours of elite-level gymnasts. *Int. J. Environ. Res. Public Health* 18:6941. doi: 10.3390/ijerph18136941

Barris, S., Davids, K., and Farrow, D. (2013). Representative learning design in springboard diving: is dry-land training representative of a pool dive? *Eur. J. Sport Sci.* 13, 638–645. doi: 10.1080/17461391.2013.770923

Bobo-Arce, M., Sierra-Palmeiro, E., Fernández-Villarino, M. A., and Fink, H. (2021). Training in rhythmic gymnastics during the pandemic. *Front. Psychol*. 12:658872. doi: 10.3389/fpsyg.2021.658872

Cheng, K. B., and Hubbard, M. (2008). Role of arms in somersaulting from compliant surfaces: a simulation study of springboard standing dives. *Hum. Mov. Sci*. 27, 80–95. doi: 10.1016/j.humov.2007.05.004

Donti, O., Tsolakis, C., and Bogdanis, G. C. (2014). Effects of baseline levels of flexibility and vertical jump ability on performance following different volumes of static stretching and potentiating exercises in elite gymnasts. *J. Sports Sci. Med.* 13, 105–113.

Ferdinands, R. E., Kersting, U., and Marshall, R. N. (2009). Three-dimensional lumbar segment kinetics of fast bowling in cricket. *J. Biomech.* 42, 1616–1621. doi: 10.1016/j.jbiomech.2009.04.035

Fujihara, T. (2022). Real-time video and force analysis feedback system for learning strength skills on rings in men's artistic gymnastics. *Sports Biomech.* [Epub ahead of print]. doi: 10.1080/14763141.2021.2024873

Gronbech, E. (1993). Sport-specific strength training for the rotation about the transverse axis in diving and gymnastics. *Natl. Strength Cond. Assoc. J.* 15, 62–66.

Hamill, J., Ricard, M. D., and Golden, D. M. (1986). Angular momentum in multiple rotation nontwisting platform dives. *Int. J. Sport Biomech.* 2, 78–87. doi: 10.1123/ijsb.2.2.78

Hiley, M. J., and Yeadon, M. R. (2003). The margin for error when releasing the high bar for dismounts. *J. Biomech.* 36, 313–319. doi: 10.1016/S0021-9290(02)00431-1

Hiley, M. J., and Yeadon, M. R. (2005). The margin for error when releasing the asymmetric bars for dismounts. *J. Appl. Biomech.* 21, 223–235. doi: 10.1123/jab.21.3.223

Jemni, M., Sands, W. A., Friemel, F., Stone, M. H., and Cooke, C. B. (2006). Any effect of gymnastics training on upper-body and lower-body aerobic and power components in national and international male gymnasts? *J. Strength Cond. Res.* 20, 899–907. doi: 10.1519/R-18525.1

Khong, S., and Kong, P. A. (2016). Simple and objective method for analyzing a gymnastics skill. *Eur. J. Physic. Educ. Sport* 12, 46–57. doi: 10.13187/ejpe.2016.12.46

King, M. A., Kong, P. W., and Yeadon, M. R. (2019). Maximising forward somersault rotation in springboard diving. *J. Biomech.* 85, 157–163. doi: 10.1016/j.jbiomech.2019.01.033

King, M. A., Kong, P. W., and Yeadon, M. R. (2022). Differences in the mechanics of takeoff in reverse and forward springboard somersaulting dives. *Sports Biomech.* [Epub ahead of print]. doi: 10.1080/14763141.2022.2034929

King, M. A., and Yeadon, M. R. (2004). Maximising somersault rotation in tumbling. *J. Biomech.* 37, 471–477. doi: 10.1016/j.jbiomech.2003.09.008

Kong, P. W. (2010). Hip extension during the come-out of multiple forward and inward pike somersaulting dives is controlled by eccentric contraction of the hip flexors. *J. Sports Sci.* 28, 537–543. doi: 10.1080/02640411003628030

Mikl, J. (2018). Joint moments required to hold a posture while somersaulting. *Hum. Mov. Sci.* 57, 158–170. doi: 10.1016/j.humov.2017.12.001

Miller, D. I., Hennig, E., Pizzimenti, M. A., Jones, I. C., and Nelson, R. C. (1989). Kinetic and kinematic characteristics of 10-m platform performances of elite divers: I. Back takeoffs. *Int. J. Sport Biomech.* 5, 60–88. doi: 10.1123/ijsb.5.1.60

Miller, D. I., and Sprigings, E. J. (2001). Factors influencing the performance of springboard dives of increasing difficulty. *J. Appl. Biomech.* 17, 217–231. doi: 10.1123/jab.17.3.217

Minghelli, B., Paulino, S., Graça, S., Sousa, I., and Minghelli, P. (2019). Time-motion analysis of competitive surfers: Portuguese championship. *Rev. Assoc. Méd. Bras*. 65, 810–817. doi: 10.1590/1806-9282.65.6.810

Mitchell, D., Davis, B., and Raim, L. (2002). *Teaching Fundamental Gymnastics Skills*, 1st Edn. Champaign, IL: Human Kinetics.

Mizutori, H., Kashiwagi, Y., Hakamada, N., Tachibana, Y., and Funato, K. (2021). Kinematics and joints moments profile during straight arm press to handstand in male gymnasts. *PLoS One* 16:e0253951. doi: 10.1371/journal.pone.0253951

Mkaouer, B., Hammoudi-Nassib, S., Amara, S., and Chaabène, H. (2018). Evaluating the physical and basic gymnastics skills assessment for talent identification in men's artistic gymnastics proposed by the International Gymnastics. *Biol. Sport* 35, 383–392. doi: 10.5114/biolsport.2018.78059

Mkaouer, B., Jemni, M., Amara, S., Chaabène, H., and Tabka, Z. (2013). Kinematic and kinetic analysis of two gymnastics acrobatic series to performing the backward stretched somersault. *J. Hum. Kinet.* 5, 17–26. doi: 10.2478/hukin-2013-0021

Morgulev, E., Azar, O. H., and Lidor, R. (2018). Sports analytics and the big-data era. *Int. J. Data Sci. Anal.* 5, 213–222. doi: 10.1007/s41060-017-0093-7

Natrup, J., Bramme, J., de Lussanet, M. H. E., Boström, K. J., Lappe, M., and Wagner, H. (2020). Gaze behavior of trampoline gymnasts during a back tuck somersault. *Hum. Mov. Sci.* 70:102589. doi: 10.1016/j.humov.2020.102589

Natrup, J., de Lussanet, M. H. E., Boström, K. J., Lappe, M., and Wagner, H. (2021). Gaze, head and eye movements during somersaults with full twists. *Hum. Mov. Sci.* 75:102740. doi: 10.1016/j.humov.2020.102740

Nor Adnan, N. M., Ab Patar, M. N. A., Lee, H., Yamamoto, S. I., Jong-Young, L., Mahmud, J., et al. (2018). Biomechanical analysis using Kinovea for sports application. *IOP Conf. Ser. Mater Sci. Eng.* 342:012097. doi: 10.1088/1757-899X/342/1/012097

Pan, J. W., Komar, J., Sng, S. B. K., and Kong, P. W. (2021). Can a good break shot determine the game outcome in 9-ball? *Front. Psychol.* 12:691043. doi: 10.3389/fpsyg.2021.691043

Park, J., and Yoon, S. (2017). Kinematic analysis of back somersault pike according to skill level in platform diving. *Korean J. Sport Biomech.* 27, 157–164. doi: 10.5103/KJSB.2017.27.3.157

Puig-Diví, A., Escalona-Marfil, C., Padullés-Riu, J. M., Busquets, A., Padullés-Chando, X., Marcos-Ruiz, D., et al. (2019). Validity and reliability of the Kinovea program in obtaining angles and distances using coordinates in 4 perspectives. *PLoS One* 14:e0216448. doi: 10.1371/journal.pone.0216448

Sanders, R., and Gibson, B. (2003). Diving: technique and timing in women's backward two and one half somersault tuck (205C) and the men's backward two and one half somersault pike (205B) 3m springboard dives. *Sports Biomech.* 2, 73–84. doi: 10.1080/14763140308522809

Sanders, R. H., and Wilson, B. D. (1988). Factors contributing to maximum height of dives after takeoff from the 3M springboard. *Int. J. Sport Biomech.* 4, 231–259. doi: 10.1123/ijsb.4.3.231

Sayyah, M., Hiley, M. J., King, M. A., and Yeadon, M. R. (2018). Functional variability in the flight phase of one metre springboard forward dives. *Hum. Mov. Sci.* 59, 234–243. doi: 10.1016/j.humov.2018.04.014

Sayyah, M., King, M., Hiley, M. J., and Yeadon, M. R. (2020). Functional variability in the takeoff phase of one metre springboard forward dives. *Hum. Mov. Sci.* 72:102634. doi: 10.1016/j.humov.2020.102634

Tay, C. S., and Kong, P. W. (2018). A video-based method to quantify stroke synchronisation in crew boat sprint kayaking. *J. Hum. Kinet.* 65, 45–56. doi: 10.2478/hukin-2018-0038

Vladimir, P., Carmen, M., and Daniel, N. (2015). Andreyeva No. Didactic technologies of learning the double back somersault on floor based on the biomechanical analysis of sports technique in women's artistic gymnastics. *J. Phys. Educ. Sport* 15, 120–127. doi: 10.7752/jpes.2015.01020

Walker, C., Sinclair, P., Graham, K., and Cobley, S. (2017). The validation and application of inertial measurement units to springboard diving. *Sports Biomech.* 16, 485–500. doi: 10.1080/14763141.2016.1246596

Walker, C., Warmenhoven, J., Sinclair, P. J., and Cobley, S. (2019). The application of inertial measurement units and functional principal component analysis to evaluate movement in the forward $3^1/_2$ pike somersault springboard dive. *Sports Biomech.* 18, 146–162. doi: 10.1080/14763141.2019.1574887

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Education

ORIGINAL RESEARCH
published: 15 June 2022
doi: 10.3389/feduc.2022.877129

Psychometric Properties of the Competencies Compound Inventory for the Twenty-First Century

Macarena-Paz Celume[1,2*†] *and Haïfat Maoulida*[1,2†]

[1] Laboratoire de Psychologie et d'Ergonomie Appliquée, Institut de Psychologie, Université Paris Cité and Univ Gustave Eiffel, LaPEA, F-92100 Boulogne-Billancourt, France, [2] Beyond Education, Paris, France

The world is evolving rapidly, implying that the jobs of tomorrow, the socio-economic problems and the technologies we will have to interact with, will no longer exist. For this, a new set of skills and competencies will be necessary and these will allow us to face the twenty-first century. The "Four-Dimensional Education" model from the Center of Curriculum Redesign (CCR), which is developed by Fadel and his collaborators in 2015, stands out by proposing a framework that organizes twelve competencies for the twenty-first century, defines them in a clear and usable way, and provides levels for action for all education stakeholders. Based on this model, a self-reported scale was built to assess these competencies. The purpose of this study is to present the psychometric properties of this scale with the objective of measuring this specific set of competencies. The results showed good psychometric properties, presenting a sensitive, reliable, and valid scale to measure twenty-first century competencies.

Keywords: twenty-first century competencies, assessment, tools, psychometrics, education

OPEN ACCESS

Edited by:
George Waddell,
Royal College of Music,
United Kingdom

Reviewed by:
Ángel Freddy Rodríguez Torres,
Central University of Ecuador, Ecuador
Wei Shin Leong,
Ministry of Education, Singapore

***Correspondence:**
Macarena-Paz Celume
mp.celume@cri-paris.org

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Assessment, Testing and Applied Measurement, a section of the journal Frontiers in Education

Received: *16 February 2022*
Accepted: *10 May 2022*
Published: *15 June 2022*

Citation:
Celume M-P and Maoulida H (2022) Psychometric Properties of the Competencies Compound Inventory for the Twenty-First Century. Front. Educ. 7:877129. doi: 10.3389/feduc.2022.877129

INTRODUCTION

The world is evolving rapidly, implying that the jobs of tomorrow, the socio-economic problems and the technologies that people will have to interact with, will no longer exist (WEF, 2020). However, in the context of all these problems and issues, humanity must be able to face them successfully to continue to evolve in this world. If education, as it is thought, consists of a transmission of knowledge, which must be applied, the consequences of this rapid change imply this knowledge must be used to create new adapted ones (Trilling and Fadel, 2009; Fadel et al., 2015).

For this, a new set of skills and competencies will be necessary, and this will allow us to face the twenty-first century. It is therefore necessary to rethink education to transmit, on the one hand, this traditional knowledge, but above all, to give the necessary skills to individuals, not only to apply this knowledge but to exploit it (Trilling and Fadel, 2009; Fadel et al., 2015), on the other hand. The competencies for the twenty-first century are therefore defined as all the skills and competencies needed by individuals to face, with adaptability and consciousness, by themselves and together, the technological, societal, and economic challenges that cannot be anticipated or thought through in the present because of their fast evolution and uncertain nature. Many models propose not only defining these competencies but also listing them [e.g., P21 Framework for Twenty-first Century Learning (Lai et al., 2017); Framework for Twenty-first Century Learning from the Partnership for Twenty-first Century Skills (P21, 2009); Twenty-first Century Skills System from ATC21S (Griffin and Care, 2014);

twenty-first century skills and competences for new millennium learners from Organisation for Economic Co-operation and Development, OECD (Ananiadou and Claro, 2009), etc.]. Among these models, the "Four-Dimensional Education" model from the Center of Curriculum Redesign (CCR), which is developed by Fadel et al. (2015), stands out by proposing a framework that organizes competencies of the twenty-first century, defines them in a clear and usable way, and provides levers for action for all education stakeholders.

The Four-Dimensional Education Framework

The Center of Curriculum Redesign framework proposes to divide education, as it has to be from now, for the twenty-first century, into four main dimensions: knowledge, skills, character, and meta-learning (Fadel et al., 2015).

In the CCR model that is used here, 12 competencies are grouped to form 3 dimensions: skills (4 of the 12 competencies), character (6 of the 12 competencies), and meta-learning (2 of the 12 competencies). A fourth dimension is also present in the model, the "Knowledge" dimension. It includes all the "classical" knowledge, which is generally transmitted through the school curriculum (literature, science, mathematics, etc.). If this knowledge dimension has been taught for decades or even centuries, it is also widely evaluated, especially in a numerical way (i.e., through grades), which makes it possible, among other things, to situate the student among his or her age peers and to assess his or her level at a given moment and to propose lessons adapted to his or her understanding and evolution. The idea of this research is to propose a unique tool measuring the 3 of these 4 dimensions that are not taught in traditional education and, therefore, do not benefit from a systematic common/unique evaluation.

Skills Dimension

Skills refer to the way a person uses what he or she has learned (Bialik et al., 2015b). In general, a skill is defined as "an ability or proficiency acquired through training and practice" (VandenBos, 2015).

In the CCR frameworks, these skills refer to what used to be identified as social skills (i.e., "a set of learned abilities that enable an individual to interact competently and appropriately in a given social context," VandenBos, 2015) because they encompassed these 4 key competencies: critical thinking, communication, collaboration, and creativity. They are also known as the 4C.

Creativity refers to the interaction between aptitude, process, and the environment by which an individual or group produces a perceptible product that is both novel and useful as defined within a social context (Bialik et al., 2015b; Fadel et al., 2015). The person will be able to come up with ideas and implement actions that are both new and useful. When faced with an obstacle or an uncertain situation, a creative person can propose multiple ideas, adopt a variety of perspectives, and adjust previous actions or ideas.

Critical thinking designs the mental processes, strategies, and representations people use to solve problems, make decisions, and learn new concepts (Bialik et al., 2015b; Fadel et al., 2015). It also refers to the ability to critically evaluate information and claims one is confronted with (Bialik et al., 2015b; Fadel et al., 2015). The individual has to be able to solve problems, make decisions, and learn new things using logic and reasoning. A person who thinks critically considers alternative and opposing perspectives and is able to identify the strengths and weaknesses of each solution or draw conclusions. This involves organizing information, knowing what questions to ask, and making sense of confusing ideas.

Communication is the ability to listen to and understand information and ideas presented through spoken words and sentences (and other media,; Bialik et al., 2015b; Fadel et al., 2015). It also relates to the aim to possess adequate ability to pass along or give information through public speaking, design, presentations, and use of media (Bialik et al., 2015b; Fadel et al., 2015). A skilled communicator is able to listen actively, understand others, and express clearly and precisely knowledge and ideas. This person is able to adapt their communication style according to the audience and deliver the message using various methods, such as verbal, non-verbal, written, or digital.

Collaboration could be defined as the coordinated and synchronous activity that is the result of a continued attempt to construct and maintain a shared conception of a problem (Bialik et al., 2015b; Fadel et al., 2015). A skilled collaborator will be a solid part of a group activity or/and project with the willingness to create and maintain a shared understanding of a problem. A collaborative person can share and take responsibility, give and receive feedback, and face a conflict, if needed. This kind of person works with empathy and others can rely on them.

Characters Dimension

Characters design how people behave and engage in the world (Bialik et al., 2015a). The notion of a character or character trait is often confused, intermingled, and used instead of terms like personality, temperament, or mood. It will be more generally apprehended as the set of personality traits and attributes which include, among others, the set of social, moral, belief, and ethical characteristics of individuals (Allport, 1921; VandenBos, 2015). It relates to half of the 12 competencies of the CCR framework: mindfulness, curiosity, courage, resilience, ethics, and leadership.

Mindfulness describes a present-oriented state of conscious awareness, in which the individual is aware of multiple perspectives (Bialik et al., 2015a; Fadel et al., 2015). There is an element of openness to novelty in which the individual actively constructs categories and distinctions. The person is able to be in the present moment and aware of their own state and the state of the world. A mindful person can have multiple perspectives, be aware of and express emotions appropriately, and understand the world in its complexity. Novelty is welcomed with calmness, happiness, and openness to it.

Curiosity refers to an interest in ideas and a love for learning, understanding, and intellectual exploration; an inquisitive, playful mindset; being drawn to thinking and playing with ideas (Bialik et al., 2015a; Fadel et al., 2015). Doing reflective activities or investigations are among the favorite activities that attract a

curious person. Drawn to inquiry and new ideas, this person has an open and playful mindset.

Courage is the ability to voice opinions, needs, and feelings, aiming to exert social influence; the capacity to assert one's own will to accomplish goals in the face of opposition or consequences, such as speaking out, taking a stand, and confronting others if needed (Bialik et al., 2015a; Fadel et al., 2015). One will be able to accomplish aims and goals no matter what potential constraints or obstacles there are. Courageous people do not hesitate to express their opinions, needs, and feelings regardless of the potential consequences. A courageous person can speak up, take a stand, mobilize, and confront others when necessary.

Resilience refers to the ability to deal appropriately with the ambiguity, changes, and challenges that different perspectives and experiences can present and to maintain one's identity and/or develop personally, or as a result (Bialik et al., 2015a; Fadel et al., 2015). When facing obstacles, a resilient person will see things through with patience, flexibility, and a positive mindset.

Ethics refers to a system of moral principles that affect how people make decisions and lead their lives with concern for what is good for individuals and society (i.e., it will then include moral dilemmas and decisions, rights, and responsibilities, "good or bad," and "right or wrong" dualities; Bialik et al., 2015a; Fadel et al., 2015). An individual can make decisions based on a strong system of moral principles, such as respect, equality, honesty, and justice. An ethical person is concerned with what is good for him or herself, other individuals, and society. Ethical people consider the consequences of their own actions and try to make decisions that will leave a positive mark on the world.

Leadership is defined as a relational and ethical process of people aiming to accomplish positive change together (Bialik et al., 2015a; Fadel et al., 2015). This type of person can set goals for themselves and inspire others to follow them to accomplish a positive change. A leader collaborates and manages ethically all the personal, financial, and material resources. Leaders have a strategic mindset and a clear vision, which are needed to accomplish goals and face challenges. A leader serves as an inspiration for their community.

Meta-Learning Dimension

Meta-learning concerns reflection on oneself and the fact that a person constantly adapts, continues to grow, and learn, reaching their goals, and purposes (Bialik and Fadel, 2015). It is then divided into two main competencies: metacognition and growth-mindset.

Metacognition refers to both the ability to recognize one's knowledge, skills, attitudes/values, and way of learning and to set goals and adapt learning strategies based on outcomes (Bialik and Fadel, 2015; Fadel et al., 2015). A person would then be capable of recognizing their own knowledge, skills, behaviors, and ways of learning. A metacognitive person can reflect on and adapt their own learning strategies and adjust goals accordingly. In general, a highly meta-cognitive person demonstrates flexibility in choices, decisions, and actions, thanks to in-depth self-knowledge and the ability to self-regulate.

A growth-mindset is believing that one can change and can learn, grow, and improve one's personal future as much as seeing progress as contingent on effort and obstacles (Bialik and Fadel, 2015; Fadel et al., 2015). With this mindset, a person will be able to internalize that they have the power to effect change in themselves, others, and the world. A person with a growth-mindset suggests that one can always learn, grow, and improve and is able to see progress from the efforts they put in. These people are always capable of seeing failure, drawbacks, and feedback as a chance to grow and improve themselves.

From the CCR to CCI: A Three-Dimensional Model-Based Assessment

The first dimension of the CCR-model, knowledge, is all about what a person knows and understands (Bialik and Fadel, 2018). Even if current knowledge areas covered in curricular subjects might still need to be carefully redesigned, to include modern educational subjects, interdisciplinary courses, and some of the traditional school subjects may need to be reshaped, this dimension remains the one that has been and still is, the most worked on and developed by educators around the world. In the same way, skills, characters, and meta-learning are private study objects and have been the focus of research in the social sciences and epistemology, in which this paper is rooted. The objective of this research will be to propose, through the analysis of psychometric qualities, a sensitive, faithful, and valid measurement tool, which can be used in the evaluation of each of the previously 12 competencies defined. The objective of this research will be to propose, through the analysis of psychometric qualities, a sensitive, faithful, and valid measurement tool, which can be used in the evaluation of each of the 12 competencies previously defined and cited.

A Proposal to Measure Twenty-First-Century Competencies

The importance of developing twenty-first century competencies has been underlined by actors from different backgrounds, many of whom emphasize the need to develop them through their evaluation.

Also, in addition to proposing a model that allows us to better understand and define them, the CCR has advanced the need to measure them (Bialik et al., 2016). As said before, the knowledge dimension is already widely measured, especially in the individual's school career, but the other dimensions (skills, characters, and meta-learning) are much less so, if at all (Bialik et al., 2016). Therefore, the CCR will propose measuring the latter by using an evaluation from several angles, including the use of measurement scales to measure the level of individuals for each of these skills (Bialik et al., 2016). In fact, they present tools to assess each of the 12 competencies wishing to be measured here. Yet, no tool of an individual or an actual knowledge can measure all of these competencies in a single, quick, robust, and self-reported way.

Thus, a specific tool measuring each of those 12 competencies, with sensitive, reliable, and strongly valid items, will help caregivers, educational professionals (particularly those who develop programs to develop each of them), kids, and themselves,

to evaluate first, then monitor, and finally progress on these indispensable competencies.

Item Construction and Psychometric Procedure

Based on the definitions of the 12 competencies proposed by Fadel et al. (2015), researchers specializing in the field constructed an initial pool of items in English to assess these competencies. Items were constructed to fit with children's comprehension and memory or use words that were not beyond their vocabulary level. They were designed to make them as short and explicit as possible, written in simple language, and contextualized with examples taken in their familiar environment. Then, these items were compared to the four scales: the Values in Action Inventory of Strength (and for youth version, VIA96-youth, Park and Peterson, 2006); the Global Assessment of Character Strengths (GACS-24; McGrath, 2017), which is presented as a reduced version of the VIA-96; the Motivated Strategies for Learning Questionnaire (MSLQ; Pintrich et al., 1991); and the Character Strengths Inventory for Children (CSI-C; Shoshani and Shwartz, 2018). Indeed, Bialik et al. (2016) proposed a set of measures for each competency to assess them. This research focused only on tools using self-reported measurements that were associated specifically with one of the CCR-model competencies. Also, there were many dimensions measured by the VIA-96 that overlapped with those measured by the scale that this study aimed to develop, especially for skills and characters. The MSLQ covered mostly the meta-learning dimensions. By comparing their items to the others, a set of items were removed. Finally, the remaining items were scored independently on a 0–3 scale. Items with a score of 3 were retained, and those with an average score above 2.5 were discussed and either modified or retained. The objective was to retain 4 items for each dimension. Then, the 48 items were submitted to native English speakers for a check of the language and their adaptability to a population of young people between 13 and 18 years old. After the first pre-test of these items on a sample of fewer than 100 people, the first results in terms of the global sensitivity of the scale and fidelity of each skill were encouraging; the validation procedure for the scale was then launched. The analysis started with the sensitivity of each item and each competency. It continued by checking the three dimensions for sensitivity, and finally the scale as a whole. Second, the internal consistency of the three dimensions and the whole scale was assessed. Moreover, the validity of this scale in terms of structure was tested, analyzing its hierarchical structure; a global factor of twenty-first century competencies, divided into 3 dimensions: skills, character, and meta-learning.

Study Purposes

The purpose of this study is to assess the psychometric qualities of a twenty-first century competencies tool, in particular, the CCI21 presented in this paper. To this end, two concatenated studies were conducted: a first study measuring the sensitivity and factorial validity of the CCI21 scale to iterate the scale and find a better fit, and a second study measuring the sensitivity, reliability, and validity of the revised version of the CCI21 scale after study 1.

TABLE 1 | Crosswalk between competencies compound inventory for the twenty-first century (CCI21) competencies and values in action inventory of strength (VIA-96)/global assessment of character strengths (GACS-24).

CCI21 competencies	VIA-96 and GACS-24-character strengths
Creativity	Creativity
Critical thinking	Judgement
Collaboration	Teamwork
Communication	Perspective
Mindfulness	Self-Regulation
Curiosity	Curiosity
Courage	Bravery
Ethics	Fairness
Leadership	Leadership
Resilience	Perseverance
Metacog	Love of learning
Growth mindset	Love of learning

In particular, it can be hypothesized, on the one hand, that the scale will measure twenty-first century competencies and that it does not measure social desirability (Crowne and Marlowe, 1960) or it does not measure all the character strengths of the VIA (or GACS for the short version) that were not associated with one of the twelve competencies. On the hand, it can be hypothesized that there will be a link between CCI competencies and the one from the VIA/GACS that previous authors (Bialik et al., 2016) or ourselves have made in correspondence. Also, these matched character strengths are used as convergent validity criteria while both SDS and other character strengths are used as divergent validity criteria. We will observe a positive moderate correlation between VIA/GACS correspondent strengths and CCI-21 competencies, no or weak correlation between CCI-21competencies, and both VIA/GACS non-correspondence competencies and SDS scores.

Based on this, it is expected that after both studies, the following assumptions have been made:

1. The CCI21 psychometric properties will reach moderate to high scores in sensitivity, reliability, and validity indices:
   a. Item sensitivity, as well as dimension and whole scale sensitivity of the CCI21, will reach normal levels in both skewness and kurtosis indices.
   b. Dimension and whole scale reliability of the CCI21 will reach good to excellent Cronbach's alpha scores.
   c. The CCI21 competencies' scores will present a moderate to strong positive correlation with eleven of the character strengths of the VIA-96 (Peterson and Seligman, 2004) and the GACS-24 (McGrath, 2017) scales, identified as related to the CCI21 (**Table 1**), regarding the observed correlations between these two previously validated scales.
   d. Competencies' scores will show null or weak correlations with the VIA-96 competencies that are unrelated, specifically, religion and spirituality, and with the social desirability scale (SDS).

2. The CCI21 global score will present no effect on age or gender.

The following table presents the crosswalk between the competencies of the present tool (CCI21) and the VIA-96 and GACS-24 character strengths (see **Table 1**).

STUDY 1

Materials and Method

Participants

A total of 349 English-speaking school-age students (M = 15.33; SD = 2.57) from different countries (88% from South Africa, 12% from other countries such as England and France) were recruited to complete the protocol. The participants in this study were primarily South African school-aged students who took this study as part of their school curriculum. The study was presented to them as a study of their psychosocial competencies, the competencies that will enable them to cope with the current and future challenges of the twenty-first century. This scale was also proposed to a lesser extent to English-speaking education students in France, as part of a university course. Only participants who completed the entire study (i.e., filled out all the questionnaires) were included in this study. Of those recruited, 269 started the protocol and fully completed the scale for the twenty-first century competencies. The final sample ($N = 269$) consisted of 115 boys (M = 15.5; SD = 2.14), 134 girls (M = 15.1; SD = 2.91), and 7 who did not specify gender (M = 15; SD = 2.31). Twenty students did not specify their age.

Material

Socio-Demographic Questionnaire

Students were asked about their age, gender, country of study, nationality, school level, name of the school, type of school, primary language, level of English, level of studies for both parents, home conditions (own room, computer, and internet access), and their interests. These questions had to be answered by choosing from a predefined range or by inputting a text when needed.

Social Desirability Scale

To measure social desirability and prevent response bias, the Barger (2002) version of Marlow's Social Desirability Scale (SDS, Crowne and Marlowe, 1960) was used. It reported good criterion validity (Barger, 2002). To match students' comprehension, six items were adapted (e.g., I sometimes try to take revenge rather than forgive and forget), arriving at a final dichotomous scale consisting of 13 items (true vs. false).

The Competencies Compound Inventory for the Twenty-First Century

This scale proposes measuring twenty-first century competencies using a 5-point Likert scale and 48 items. Each set of 12 competencies belongs to one of these 3 dimensions: skills (CRE for creativity, CRI for critical thinking, COL for collaboration, COM for communication), characters (MIN for mindfulness, CUR for curiosity, COU for courage, RES for resilience, ETH for ethics, and LEA for leadership); and Meta-learning (MET for metacognition and GRO for growth-mindset). Every competency is measured with 4 items presented randomly.

Procedure

Participants needed to register through an online platform, the DreamShaper platform, before starting the testing procedure. After registering, participants were invited to read a detailed information notice where all types of data that would be collected and stored were clarified and a presentation of the study and its objectives was explained, and then they gave their consent to data collection through a consent form. The detail of this information notice and consent form can be sent on request. A parent or guardian had to provide explicit permission for students under 16 years old, in accordance with the legal and ethical procedures. As the study was conducted in France, European rules apply, which means that over 16 years old, participants can give their own consent for the collection of their data. Moreover, ethical advice, previous to the submission of an official ethical form, was asked to an ethical university committee, which responded that according to national laws, there was no need for ethical approval as no medical intervention was conducted on participants, and information collected could not provoke prejudice in the participants. The advice was mostly focused on two points: (1) providing an information notice and a consent form to the participants before starting the data collection, which was done through the platform; and (2) detailing data storage information and procedure of non-identification of participants. Data collected from participants were stored in a secured European server to comply with European orientations. A General Data Protection Regulation document was also created to detail safety measures in case of national control, as required in France by the French National Commission for Information Technology and Civil Liberties (CNIL). Data were anonymized by automatically assigning a number to every participant within the platform, which was non-traceable, as data exports did not include any personal identification, and analyses were only conducted with anonymized data.

Then, the study was conducted through the online platform. After having filled in the socio-demographic information, they completed the social desirability scale, and finally, the CCI21 scale.

Results

Sensitivity

Sensitivity was analyzed at three levels: items, competencies, and global.

Item Sensitivity

With means ranging from 2.59 (MC1) to 3.97 (CU3), participants responded predominantly to the right on the 5-point Likert scale. Kurtosis and skewness indices showed an abnormal rightward skew for the following items, with dimensions in parenthesis: CR1, CR3, CR4, CT1, CM1 (Skills); MF2, MF3 (Characters); and MC1, MC2 (Meta-learning) (see **Table 2**).

TABLE 2 | Skewness and Kurtosis scores for abnormal items of the CCI21.

	M	*MODE*	*MED*	*SD*	*Range*	*Skewness*	*Kurtosis*
Item CR1	3.08	2.00	3.00	1.23	1–5	0.17	−1.03
Item CR3	3.20	3.00	3.00	1.28	1–5	−0.00	−1.05
Item CR4	3.37	5.00	3.50	1.31	1–5	−0.33	−1.02
Item CT1	3.40	5.00	4.00	1.34	1–5	−0.35	−1.06
Item CM1	3.25	3.00	3.00	1.26	1–5	−0.14	−1.05
Item MF2	3.51	5.00	4.00	1.38	1–5	−0.44	−1.06
Item MF3	3.40	5.00	3.00	1.32	1–5	−0.33	−1.03
Item MC1	2.59	1.00	2.00	1.34	1–5	0.35	−1.06
Item MC2	3.17	4.00	3.00	1.26	1–5	−0.15	−1.02

CR, Creativity item; CT, Critical thinking item; CM, Communication item; CL, Collaboration item; MF, Mindfulness item; CU, Curiosity item; CO, Courage item; RS, Resilience item; ET, Ethics item; LD, Leadership item; MC, Metacognition; GM, Growth mindset item.

TABLE 3 | Global score sensitivity checklist.

Steps	Results	Criteria
Normal distribution?	Tends to normal curve, slightly to the right, although not extreme Normality indexes: Skew = −0.10; Kurt = −0.33	Yes
Obs. centrality similar among them?	Mean 3.6 ≃ median 3.5 ≃ mode 4	Yes
Th. centrality indexes similar to observed?	Th Mean = 3 VS. Obs. = 3.6 Th Median = 3 VS. Obs. = 3.5 Th Mode = 3 VS. Obs. = 4	Yes, except for mode
Theoretical range, similar to observed?	Th. range 4 ≃ Obs. range 3.06	Doubtful
Observed SD higher than theoretical SD	Obs.SD.61 > Th SD.66	No

Competencies Sensitivity

With means ranging from 3.27 (MET) to 3.87 (GRO), participants responded centered on the 5-point Likert scale, following a normal distribution. Kurtosis and skewness indices confirmed normal distribution with no abnormal scores for any of the competencies.

Global Sensitivity

This analysis is presented in the **Table 3**. It was checked for normality indexes (skewness and kurtosis) and the shape of the scale distribution (normal curve). Both demonstrate a slight trend to the right of the scale. The observed centrality indices are close to each other and close to the expected theoretical values. It is almost the same for the observed range, which is close to the theoretical value. Finally, the observed SD is slightly below the theoretical SD (0.67) contrary to what was expected.

Reliability

Internal Consistency

The reliability of the CCI21 and its dimensions were measured using Cronbach's alpha, showing excellent reliability for the whole scale ($\alpha_{CCI21} = 0.94$), and good for each dimension ($\alpha_{skills} = 0.82$; $\alpha_{character} = 0.89$; $\alpha_{meta-learning} = 0.82$). Competencies' reliability presented scores ranging from poor to acceptable [α (0.51, 0.77)].

Split-Half Reliability

A correlational approach was used to determine if there is a consistency between the odds items and even items of the scale (each part was composed of 24 items). A strong positive correlation between these two parts ($r = 0.986$) was found.

Validity

Factor Structure of the CCI21

Principal component analysis (PCA) was conducted on the whole 48-item inventory and on each dimension. An oblimin rotation was used because it was supposed that there might be correlations between the factors. For each analysis, the factor loadings above 1 as a criterion were used to determine the number of factors to be retained. This analysis was conducted to see if any items needed to be removed from the scale.

First, PCA was performed on the first 16 items to see if the four-dimensional structure of the Skills dimension was found. The PCA [KMO = 0.83, $\chi^2_{(120)} = 952$ $p < 0.001$] leads to a four-factor solution ($\lambda1 = 2.71$, $\lambda2 = 2.25$, $\lambda3 = 1.89$, $\lambda4 = 1.43$) explaining 51.8% of the variance (Hayton et al., 2004). Analysis showed that not all the items match their factors; thus, after iteration, the item that contributed the least to the factor loadings and the competency reliability was deleted for each competency. The first factor was composed of a mix of items belonging to creativity (CR) and communication (CM). Yet, when items among the least sensitive and least saturating on this factor are removed (CR4 = 0.54 and CM4 = 0.41), two factors are observed. One gathers three items related to critical thinking, and the other gathers three items related to Communication. Factor 3 gathered the four Collaboration (CL) items, although CL2 (= 0.36) was removed based on the same criteria (lowest weight in the competency within the dimension matrix). Critical thinking (CT) items were divided between factors 3 and 4, but when deleting CT1 (= 0.84), even if it presented an increased weight, CT2, CT3, and CT4 formed one single factor. The final PCA [KMO = 0.82, $\chi^2_{(66)} = 660$ $p < 0.001$] led us to a four-factor

TABLE 4 | Principal component analysis (PCA) on Skills dimension items (12) for the CCI-21 36 items version.

	Component				MSA
	1	2	3	4	
CR1		0.42			0.859
CR2		0.866			0.773
CR3		0.705			0.833
CT2			0.565	0.473	0.891
CT3				0.871	0.704
CT4		0.431		0.384	0.872
CM1			0.801		0.833
CM2			0.456		0.89
CM3			0.713		0.831
CL1	0.527				0.878
CL3	0.823				0.727
CL4	0.872				0.730

solution ($\lambda 1 = 1.97$, $\lambda 2 = 1.92$, $\lambda 3 = 1.87$, $\lambda 4 = 1.41$), explaining 59.7% of the variance (Hayton et al., 2004, see **Table 4**).

For the next 24 items, the same analysis was conducted in order to find the six-factor model that would correspond to the Characters dimension. The PCA [KMO = 0.90, $\chi^2_{(276)} = 1{,}795$ $p < 0.001$] led to a six-factor solution ($\lambda 1 = 3.22$, $\lambda 2 = 2.56$, $\lambda 3 = 2.20$, $\lambda 4 = 2.20$, $\lambda 5 = 2.09$, $\lambda 6 = 1.21$), explaining 56.2% of the variance (Hayton et al., 2004), although some items did not completely match each factor as expected. Thus, the same iteration process was followed for the skill analysis and the following items that presented either the lowest weights, or abnormal scores in item sensitivity analyses were removed: MF4 (Mindfulness item), CU3 (Curiosity item), CO4 (Courage item), RS1 (Resilience item), ET3 (Ethics item), and LD2 (Leadership item). The final PCA [KMO = 0.88, $\chi^2_{(153)} = 1{,}362$ $p < 0.001$] led to a six-factor solution ($\lambda 1 = 2.47$, $\lambda 2 = 2.18$, $\lambda 3 = 1.88$, $\lambda 4 = 1.91$, $\lambda 5 = 1.77$, $\lambda 6 = 1.41$), explaining 56.2% of the variance (Hayton et al., 2004, see **Table 5**).

Finally, the same analysis was conducted for the last 8 items to find the two-factor solution for the Meta-learning dimension. The PCA [KMO = 0.82, $\chi^2_{(28)} = 594$ $p < 0.001$] led to a two-factor solution ($\lambda 1 = 3.09$, $\lambda 2 = 1.62$), explaining 58.8% of the variance (Hayton et al., 2004, see **Table 5**). The same path as before was followed and indicated the need to delete items MC4 (Metacognition item) and GM2 (Growth Mindset item). The final PCA [KMO = 0.74, $\chi^2_{(15)} = 411$ $p < 0.001$] led to a two-factor solution ($\lambda 1 = 2.38$, $\lambda 2 = 1.57$), explaining 65.8% of the variance (Hayton et al., 2004, see **Table 6**).

Discussion of Study 1

Competency sensitivity presented normal scores, although item sensitivity and global sensitivity presented some issues. The sensitivity of the items suggests a need to review or delete certain items, which was confirmed by the results for global sensitivity. Minor changes would be needed to improve the mode and dispersion indexes. In the same way, reliability analyses presented excellent and good scores for scale and dimension,

TABLE 5 | PCA on Characters dimension items (6) for the CCI-21 36 items version.

	Component						MSA
	1	2	3	4	5	6	
MF1						0.900	0.778
MF2					0.531	0.449	0.788
MF3	0.602				0.454		0.907
CU1			0.865				0.752
CU2			0.611				0.884
CU4			0.557				0.919
CO1	0.661						0.939
CO2	0.701						0.891
CO3	0.700						0.900
RS2					0.354		0.927
RS3					0.769		0.870
RS4					0.479		0.888
ET1				0.738			0.879
ET2				0.710			0.903
ET4				0.378			0.913
LD1		0.836					0.838
LD3		0.807					0.838
LD4		0.619					0.928

nevertheless, competencies' reliability presented some poor scores, suggesting that items might need minor changes. Considering this information, the PCA analyses were satisfactory once iterations on items were made for the dimensions "skills and characters". Nevertheless, the dimension "meta-learning" presented one item that was more linked to growth-mindset competency than to its original competency, the metacognition. In this sense, to arrive at a fully fitted model, a proposal to modify this item can be formulated, and also a new study with a revised version of the scale could be undertaken. All the other dimensions, after iteration, presented a good fit, which led us to continue the criterion validity study with this renewed 36-item CCI scale. This study will be presented in the following section.

STUDY 2

Materials and Method

Participants

A total of 162 English-speaking students (M = 15.6; SD = 2.43; 66 boys, 88 girls, and 6 who did not specify gender) fully completed the protocol. One student did not provide their age.

Material

Besides the already described scales, such as the socio-demographic questionnaire (Demo) and the Social Desirability Scale (SDS, Crowne and Marlowe, 1960; Barger, 2002), the following scales were used:

The Competencies Compound Inventory for the Twenty-First-Century—36-Item Version

This scale is the 36-item version that was created after factor analyses of the 48-item scale. The CCI21-36 is a self-reporting

scale that measures 12 of the twenty-first century competencies through 3 items per competency on a 5-point Likert scale, ranging from *not like me at all* to *very much like me*.

The Values in Action Inventory of Strength 96

The scale is a youth adapted version of 24-character strengths measured with 4 items on a 5-point Likert scale, presenting a pole of 96 items (VIA 96-Youth, Park and Peterson, 2006). Traits evaluated are as follows: appreciation of beauty and excellence, bravery, creativity, curiosity, fairness, forgiveness, gratitude, honesty, hope, humility, humor, judgment, kindness, leadership, love, love of learning, perseverance, perspective, prudence, self-regulation, social intelligence, spirituality, teamwork, and zest. This scale presented good validity across different countries, including European and American countries (McGrath, 2017). Reliability scores of global scores for our sample achieved excellent scores for Cronbach's alpha ($\alpha = 0.96$).

Global Assessment of Character Strengths-24

Students' character strength was also measured by asking them to read each character definition and then establish, using a Likert 7-point scale, how accurately this trait describes them (GACS, McGrath, 2017). This scale has been validated for adult use, although the author was contacted to ask if the scale could be used on a youth population, establishing it as having vocabulary adapted for a youth student population.

Procedure

Within the online platform Dreamshaper, after completing the questionnaires Demo, SDS, and CCI21-36, in order to establish convergent and divergent validity, students were asked to complete the VIA-96 and the GACS-24 scales.

Data Analysis

Kurtosis and skewness indices were calculated to measure sensitivity. Cronbach's alpha was calculated to measure internal consistency. A split-half method was used to measure test reliability through correlation analysis. Then, confirmatory factor analysis were performed to verify the factor structure of each dimension, that is to say for Skills dimension a four-factor solution, for Characters a six-factor solution and for Meta-learning dimension a two-factor solution. Finally, correlational analyses were performed to test for convergent and divergent validity.

Results

Sensitivity

Global scores and competencies scores followed a normal distribution.

Reliability

Internal Consistency

The global scores for the CCI21-36 presented excellent internal consistency ($\alpha_{CCI21-36} = 0.93$). Similar results were found for each dimension ($\alpha_{skills} = 0.79$, $\alpha_{character} = 0.87$, $\alpha_{meta-learning} = 0.78$).

TABLE 6 | PCA on Meta-learning dimension items (6) for the CCI-21 36 items version.

	Component		MSA
	1	2	
MC1		0.943	0.603
MC2		0.778	0.674
MC3	0.739		0.763
GM1	0.682		0.807
GM3	0.786		0.791
GM4	0.813		0.777

Split-Half Reliability

A correlational analysis was undertaken to determine if there was consistency between the odd items and even items of the scale (each part is composed of 18 items). It found a strong positive correlation between those two equal parts ($r = 0.94$).

Validity

Structural Validity

For each dimension, confirmatory factor analysis was conducted to test whether the multiple-factor structure fitted our data better than a single-factor structure.

Confirmatory factor analysis was conducted to test whether a four-factor structure fitted our data better than a single-factor structure (see **Table 7**). The analysis confirmed our hypothesis and led us to adopt this four-factor solution with 3 items per factor.

Again, for the character dimension, the analysis continued with confirmatory factor analysis to confirm how the six-factor model matched the data better than the one-factor model (see **Table 7** below). As expected, this multiple-factor solution was the one solution that fitted the best, with 3 items per factor.

Confirmatory factor analyses were then conducted to see if the two-factor model matched the data better than the one-factor model, but this was not confirmed (see **Table 7** below). Yet, this result is consistent with the PCA performed in this study, as the item MC3 matches more with "growth-mindset" than with the factor identified as "metacognition."

Convergent/Concurrent Validity

As the table above shows, correlations between corresponding/concurrent dimensions (the characters of the VIA-96 that are aligned with the competencies of the CCI21-36) showed significant moderate correlations as expected in comparison with the correlations between the VIA-96 and the GACS-24 (see **Table 8**). The only correlations that appear to be weaker are as follows: (a) the competency "critical thinking," aligned with the character "judgment," but that presents better correlations with the GACS-24 character "judgment;" and (b) the competency "collaboration," aligned with the character "teamwork," which presents weaker correlations between the CCI21-36 and both concurrent scales, than the correlations between VIA-96 and GACS-24. Almost all other correlations between CCI21-36 and the concurrent scales appear

TABLE 7 | Confirmatory factor analysis on CCI21-36.

Model	X^2	ddl	*p*	CFI	TLI	SRMR	RMSEA	AIC	BIC
Skills									
One-Factor	180	54	0.000	0.797	0.752	0.068	0.093	9,821	9,951
Four-Factors	224	98	0.000	0.856	0.824	0.059	0.128	13,084	13,279
Characters									
One-Factor	359	135	0.000	0.831	0.808	0.062	0.079	13,713	13,907
Six-Factors	337	237	0.000	0.940	0.930	0.069	0.039	18,198	18,511
Meta-Learning									
One-Factor	97.6	20	0.000	0.867	0.814	0.060	0.125	5,805	5,889
Two-Factors	73.2	8	0.000	0.840	0.700	0.078	0.181	4,276	4,443

TABLE 8 | Comparative table of final correlation scores between VIA-96 and GACS-24 characters, VIA-96 characters and CCI21-36 competencies, and GACS-24 characters and CCI21-36 competencies.

CCI21-36 competencies	VIA-96 and GACS-24 characters	r (VIA-96 × GACS-24)	r (VIA-96 × CCI21-36)	r (GACS-24 × CCI21-36)
Creativity	Creativity	0.41	0.36	0.34
Critical thinking	Judgement	0.40	0.32	0.38
Collaboration	Teamwork	0.48	0.39	0.38
Communication	Perspective	0.33	0.46	0.34
Mindfulness	Self-Regulation	0.26	0.27	0.33
Curiosity	Curiosity	0.46	0.43	0.35
Courage	Bravery	0.41	0.55	0.43
Ethics	Fairness	0.39	0.37	0.36
Leadership	Leadership	0.48	0.38	0.42
Resilience	Perseverance	0.34	0.31	0.38
Metacognition	Love of learning	0.35	0.45	0.31
Growth mindset	Love of learning	0.35	0.49	0.35

All correlations between VIA-96 and CCI21-36 are significant at $p < 0.001$.

to be stronger than the correlations between the VIA-96 and the GACS-24 (see **Table 8**).

Divergent Validity

As expected, global scores for the CCI21-36 and each competency score presented weak to moderately weak correlations with the social desirability scale [M $r = 0.26$; (0.10–0.37)], as well as the "spirituality" character from the VIA-96 [M $r = 0.25$; (0.12–0.35)] (see **Table 9**).

It is also observed that the global score strongly correlates with each of the competencies scores from the CCI21-36 scale (M $r =$ 0.69; (0.56–0.79)] (see **Table 9**).

Gender and Age-Related Differences

One-way ANOVA analysis showed no effect for gender or for age.

Discussion of Study 2

The results of sensitivity analyses showed us that the scale presents normal levels of sensitivity in all the evaluated forms, leading us to conclude that the scale is sensitive enough in terms of items, dimensions, and as a whole scale. Regarding reliability, the scale presented excellent internal consistency, confirmed by PCA analyses. Finally, validity analyses confirmed the structure of each dimension and also showed consistent moderate correlations with concurrent dimensions of scales that measured similar constructs. In the same way, when competencies from the scale were contrasted with scales or dimensions measuring different constructs, the scale presented weak to moderate correlations. Moreover, the global score correlated strongly with each of its items, allowing us to conclude that the CCI21-36 presents correct validity indexes.

GENERAL DISCUSSION

The CCI-21 scale is a self-report questionnaire that measures twenty-first century competencies based on the Four-Dimensional Education framework from the Center for Curriculum Redesign (CCR). They propose dividing Education and twenty-first century competencies into four main dimensions: knowledge, skills, character, and meta-learning. The scale presented in this study aimed to measure the three last dimensions of this model consisting of the 12 competencies that it covers. This study focused on two main objectives: (1) to present CCI21 psychometric properties that will give moderate to high scores in terms of sensitivity, reliability, and validity indices; and (2) to not find an age or gender effect on the CCI21

global and competency-based score. As a whole, these objectives have been achieved, allowing us to propose a scale, not in 48 but in 36 items, measuring the skills of the twenty-first century.

Thereby, to reach these goals, for each of the previous psychometrics qualities, analyses were conducted at several levels. The first version of the scale presented 48 items and was administered to a sample of international English-speaking middle- to high-school students. Results from psychometric analyses confirmed that there were some issues regarding the scale, particularly sensitivity and factor analyses, that suggested that some of the items needed further revision or to be deleted to correspond to a good model fit. After iterations of the scale, CCI21 was reduced from 48 items to arrive at a 36-item scale that was tested on the sample of students who completed the whole protocol (a protocol that would permit convergent and divergent validity analyses). This sample produced satisfying results in terms of sensitivity, reliability, and validity. These results permitted us to confirm that the 36-item version presents good psychometric qualities, which was confirmed by correlational analyses.

In this context, the hypotheses laid out under hypothesis 1 were confirmed as all psychometric properties presented good scores in the evaluated indices: for 1a, both skewness and kurtosis indices presented normal levels for each of the competencies of the scale; for 1b, Cronbach's alpha scores were excellent; for 1c, even if correlations were mainly moderate or moderately strong, they corresponded to the correlations expected in this type of analysis, sometimes presenting even stronger correlations than expected; for 1d, correlations with the "spirituality" character and the social desirability scale were weak and moderately weak, as expected, being always weaker than the correlations presented by the global score of the CCI21-36 and its competencies. Finally, for hypothesis 2, the global score for the CCI21-36 showed no effect from age or gender.

This study encountered some difficulties inherent in the validation of the tool. First of all, the recruitment of a large sample of English-speaking school-age children, a population that is difficult to interview, and the fact that the research team was not based in an English-speaking country caused difficulties. Access to the sample was coupled with the absence of a test-retest, which was very difficult to conduct without a high percentage of lost subjects. Moreover, the length of the study (i.e., too long) did not favor this research. Finally, it was not possible to retest the sample, one of the reasons being the health crisis context related to the coronavirus disease 2019 (COVID-19). Indeed, the global health context in the year 2021 may have slowed down the conduct of this study, the priority being the maintenance of educational continuity and programs, which was not easy for the schools in this pandemic context. In addition to the difficulties encountered, the study has limitations.

First of all, even if the sample was representative of a population of English-speaking students, future studies might require a larger sample than would confirm these claims. In this sense, colleagues have to consider this study cautiously regarding variables that could diminish the replicability of the results. In line with this, differences in the sample of the study, such as cultural differences, level of English, or beliefs regarding

TABLE 9 | Correlations between CCI21-36 & competencies, SDS, and VIA-96 spirituality character.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1. CCI36	–														
2. SDS	0.370***	–													
3. CRE	0.673***	0.27***	–												
4. CRI	0.615***	0.168**	0.412***	–											
5. COM	0.676***	0.307***	0.46***	0.396***	–										
6. COL	0.634***	0.195**	0.291***	0.328***	0.407***	–									
7. MIN	0.564***	0.296***	0.219***	0.227***	0.306***	0.350***	–								
8. CUR	0.667***	0.104	0.447***	0.408***	0.379***	0.384***	0.295***	–							
9. COU	0.777***	0.307***	0.479***	0.493***	0.475***	0.434***	0.457***	0.422***	–						
10. RES	0.729***	0.285***	0.437***	0.341***	0.364***	0.390***	0.508***	0.530***	0.555***	–					
11. ETH	0.734***	0.332***	0.486***	0.446***	0.468***	0.423***	0.334***	0.387***	0.537***	0.464***	–				
12. LEA	0.706***	0.227***	0.352***	0.31***	0.466***	0.463***	0.322***	0.421***	0.504***	0.463***	0.448***	–			
13. MET	0.676***	0.28***	0.426***	0.338***	0.420***	0.338***	0.303***	0.348***	0.416***	0.396***	0.549***	0.477***	–		
14. GRO	0.785***	0.217***	0.501***	0.456***	0.480***	0.435***	0.370***	0.497***	0.601***	0.526***	0.527***	0.539***	0.544***	–	
15. VIA SPIRIT	0.352***		0.247**	0.118	0.184*	0.176*	0.193*	0.253***	0.263***	0.232**	0.266***	0.299***	0.342***	0.271***	–

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

social and emotional learning should be considered as potential variables that might impact the results.

Self-report scales always represent a limit in studies, since it is almost impossible to ensure that the results of the scale are incontestable, as social desirability might play an important role in the type of response given by participants, particularly when working with adolescents. Here it can be observed that there is a link, yet weak, between the SDS and CCI-21 scores. Then, the results of those with very high scores (i.e., more than 2.5 standard deviations) on the desirability scale must be interpreted with caution or could even be uninterpretable.

Moreover, the perspectives of parents, educators, or peers were not considered in the construction of the tool, but only the student perspective, which some educational actors might consider to be an issue in terms of the evaluation of these competencies. A way to improve accuracy in the results could be to complement this tool with other measurements (e.g., educators' evaluation, parents' evaluation, and peers' evaluation) to have cross regards and perspectives and be able to provide a full profile of the students regarding their twenty-first century competencies. Nevertheless, the present tool presents reliable properties that permit us to claim that self-reported competencies can be measured accurately with it, which might have positive implications for the educational sector, particularly in terms of social and emotional learning.

Future research could be conducted on the CCI-21 36 version. For this CCI-21, shorter versions (with 36 items), it should be found the same (non-)correspondence made between the character of strengths of the VIA/GACS (or global score of SDS) and competencies of the CCI-21 48-version. Then, correlations of the same order, with similar values will be expected. Yet, to make the evaluation better and more user friendly, we can do without VIA-96 (only keeping the GACS), if, again, the convergent validity procedure is considered necessary (it is recommended but not mandatory in our opinion). Indeed, redoing and rechecking the fidelity, using an analysis of the internal consistency (Cronbach's alpha), of stability over time (using a test-retest) and confirmatory factorial analysis, seems to be the minimum and the best to consider. Going further, other variables can also be related to this measure, such as emotional intelligence (Mayer and Salovey, 1997), or we could even take an interest in the predictive validity of the scale in terms of academic success or educational and professional orientation.

Finally, some recommendations can be made on the use of this assessment tool. It could serve to evaluate students' perception of their level of competencies for the twenty-first century during their development throughout their schooling, to check their evolution in time, or even to check which competencies might need improvement or more work. It may be useful to analyze both the overall score and the competency score of school-age youths who would be assessed on this test. The global level will allow professionals to apprehend the general level of the person in terms of twenty-first century competencies, but considering the imbalance of the different dimensions in the global score, a competency-based analysis will refine the analysis and the possibilities for intervention. Indeed, this will give a more precise mapping of these competencies, allowing educators to identify if the student presents a homogeneous profile on the whole for these competencies or is rather heterogeneous with competency assets (or "strengths," high scores on particular competencies), or competencies to develop (or "weaknesses," low scores on particular competencies). Such an assessment will not only allow the participant to be situated among his or her peers (with the help of calibration). This will also allow individual or collective intervention or remediation work to help them optimize each of their twenty-first century competencies. For the professional, this will also allow the construction of educational programs in accordance with the level of the student (or students) and individualized or personalized to the class group. It could also be of great use to guide annual reports on students, interviews with parents, and even to help identify upcoming issues in the classroom. The scale could also be useful to test the effectiveness of learning programs concerning twenty-first century competencies or social and emotional learning programs that are to be used in the classroom, and even as a measure to define improvements in actions that have already been carried out within the school. In this sense, the scale could be a good asset for educational actors, which is why the team is currently working on the English adaptation (actually to extend it) of the scale for children aged 10–12, and a study is being conducted in order to validate the French version of the scale.

CONCLUSION

The CCI21-36 is a valid tool to measure the twenty-first century competencies of 13-year-olds and older youths. The scale not only presents good psychometric properties but also has promising implications in the field of the social and emotional education of youths.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

Both authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

REFERENCES

Allport, G. W. (1921). Personality and character. *Psychol. Bull.* 18, 441.

Ananiadou, K., and Claro, M. (2009). *21st Century Skills and Competences for New Millennium Learners in OECD Countries.* Paris, France: The Organisation for Economic Co-operation and Development, OECD.

Barger, S. D. (2002). The Marlowe-Crowne affair: Short forms, psychometric structure, and social desirability. *J Pers Assess* 79, 286–305. doi: 10.1207/S15327752JPA7902_11

Bialik, M., Bogan, M., Fadel, C., and Horvathova, M. (2015a). *Character Education for the 21st Century: What Should Students Learn?* Boston, MA: Center for Curriculum Redesign.

Bialik, M., and Fadel, C. (2015). *Meta-Learning for the 21st Century: What Should Students Learn?* Boston, MA: Center for Curriculum.

Bialik, M., and Fadel, C. (2018). *Knowledge for the Age of Artificial Intelligence: What Should Students Learn?* Boston, MA: Center for Curriculum.

Bialik, M., Fadel, C., Trilling, B., and Groff, J. S. (2015b). *Skills for the 21st Century: What should Students Learn?* Boston, MA: Center for Curriculum.

Bialik, M., Martin, J., Mayo, M., and Trilling, B. (2016). *Evolving Assessments for a 21st Century Education The Ellen Koshland Family Fund* . Center for Curriculum Redesign. Available online at: http://curriculumredesign.org/wp-content/uploads/Evolving-Assessments-for-the-21st-Century-Report-Feb-15-Final-by-CCR-ARC.pdf

Crowne, D. P., and Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *J. Consult. Psychol.* 24, 349–354. doi: 10.1037/h0047358

Fadel, C., Bialik, M., and Trilling, B. (2015). *Four-Dimensional Education: The Competencies Learners Need To Succeed.* In C. Fadel (Ed.), Boston, MA: Center for Curriculum Redesign.

Griffin, P., and Care, E. (2014). "The ATC21S method," In Griffin, P., and Care, E. (Eds.), *Assessment and Teaching of 21st Century Skills: Methods and Approach* (Dordrecht: Springer), 3–33. doi: 10.1007/978-94-017-9395-7_1

Hayton, J. C., Allen, D. G., and Scarpello, V. (2004). Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis. *Organ. Res. Methods* 7, 191–205.

Lai, E., DiCerbo, K., and Foltz, P. (2017). *Skills for Today: What We Know About Teaching and Assessing Collaboration.* Pearson.

Mayer, J. D., and Salovey, P. (1997). What is emotional intelligence. *Emot. Dev. Emot. Intell. Educ. Implic.* 3, 31.

McGrath, R. E. (2017). *Technical Report: The VIA Assessment Suite for Adults: Development and Evaluation*. Cincinnati, OH: VIA Institute on Character.

P21, Partnership for 21st Century Skills (2009). *A Framework for Twenty-First Century Learning.* Available online at: https://static.battelleforkids.org/documents/p21/P21_Framework_DefinitionsBFK.pdf

Park, N., and Peterson, C. (2006). Moral competence and character strengths among adolescents: the development and validation of the values in action inventory of strengths for youth. *J. Adolesc.* 29, 891–909. doi: 10.1016/j.adolescence.2006.04.011

Peterson, C., and Seligman, M. E. (2004). *Character Strengths and Virtues: A Handbook and Classification (Vol. 1)*. Oxford, LO: Oxford University Press.

Pintrich, P. R., Smith, D. A. F., Garcia, T., and McKeachie, W. J. (1991). *A Manual for the Use of the Motivated Strategies for Learning Questionnaire (MSLQ). (Vol. 1)*. Washington, DC: ERIC, Educational Research and Improvement.

Shoshani, A., and Shwartz, L. (2018). From character strengths to children's well-being: development and validation of the character strengths inventory for elementary school children. *Front. Psychol.* 9, 2123. doi: 10.3389/fpsyg.2018.02123

Trilling, B., and Fadel, C. (2009). *21st Century Skills, Enhanced Edition: Learning for Life in Our Times.* San Francisco, CA: Jossey-Bass.

VandenBos, G. R (Ed.). (2015). *APA Dictionary of Psychology.* American Psychological Association. Available online at: https://dictionary.apa.org/

WEF: World Economic Forum. (2020). *The Global Risks Report 2020 Insight Report 15th Edition.* https://www.weforum.org/reports/the-global-risks-report-2020/

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers | Frontiers in Education

TYPE Original Research
PUBLISHED 28 July 2022
DOI 10.3389/feduc.2022.854342

OPEN ACCESS

EDITED BY
George Waddell,
Royal College of Music,
United Kingdom

REVIEWED BY
Lilly Augustine,
Jönköping University, Sweden
Y. Selvamani,
International Institute for Population Sciences (IIPS), India

*CORRESPONDENCE
Björn Boman
contact@bjornboman.com

SPECIALTY SECTION
This article was submitted to Assessment, Testing and Applied Measurement, a section of the journal Frontiers in Education

RECEIVED 13 January 2022
ACCEPTED 01 July 2022
PUBLISHED 28 July 2022

CITATION
Boman B (2022) Regional differences in educational achievement: A replication study of municipality data.
Front. Educ. 7:854342.
doi: 10.3389/feduc.2022.854342

Regional differences in educational achievement: A replication study of municipality data

Björn Boman*

Department of Education, Stockholm University, Stockholm, Sweden

The current study analyzed the relationships between explanatory variables such as socioeconomic status (SES), migration background (MB), and formal teacher competence, and aggregated grades in the Swedish lower-secondary school context by using aggregated municipality data from 2013, 2018, and 2019. SES indicators had larger effect sizes when data from different years were merged and when the outcome variable was changed to an alternative measure of educational achievement. In one model, the MB variable even became statistically insignificant. These results indicate that SES is an important variable which explains a substantial amount of variance in regard to school achievement indicators such as grade point average. Nonetheless, aggregated data may still suffer from omitted variable bias and biased effect size estimates.

KEYWORDS

Sweden, educational achievement, grades, socioeconomic status, migration

Introduction

It is well established that socioeconomic status (SES) is intimately linked to academic achievement such as grades and scholastic assessment tests in a variety of national and international contexts (e.g., Sirin, 2005; Tan, 2015; Kim et al., 2019; Boman, 2022a,c), including Sweden (Gustafsson and Yang Hansen, 2018). Although measurements of SES may differ between countries and regions (e.g., Kim et al., 2019), parental education and/or average (parental) income is typically used to indicate SES. These two types of measures tend to be highly correlated and estimates do not differ substantially depending on the use of only one of these two forms of indicators (e.g., Falk et al., 2021). In the Programme for International Student Assessment Survey (PISA) context, the estimated number of books at home is typically used as an indicator of SES (e.g., Tan, 2015; Reimer et al., 2018). More books at home indicate a higher level of SES (Reimer et al., 2018, p. 35). Aside from PISA and similar international scholastic assessments, in Sweden, average income and/or parental education is typically used to indicate the level of SES. The relationship between SES and grades is moderate but has partially been affected by migration during recent decades (Gustafsson and Yang Hansen, 2018). Thus, the relative impact of migration background (MB), and its relation to SES, on school results has

increasingly become an important area of research within the Swedish context, due to changes in demographics related to an increased influx of migrants (e.g., Ruist, 2015; Gustafsson and Yang Hansen, 2018; Boman, 2022b).

For at least a decade, public debates on school results in Sweden have been pervasive (e.g., Wiklund, 2018). Poor results in the PISA 2012 survey and in relation to grades and completion rates at lower-secondary level (i.e., the transition from Grade 9 to Grade 1 at upper-secondary level) appear to be associated with either a neoliberal decentralization of the school system that leads to increased socioeconomic and sociodemographic inequalities (e.g., Bunar, 2010), poor teachers (e.g., SVT, 2018), and increased immigration from low-skilled countries (e.g., Holmlund et al., 2019; Sanandaji, 2020).

While all of these factors might contribute to lower results but to different extents, there is little support of poor teacher quality in Sweden. First, teachers have, on average, high cognitive skills in literacy and numeracy (Hanushek et al., 2019) and although the intake requirements are low, teacher students who graduate at Swedish tertiary institutions like the University of Gothenburg have good grades and cognitive and non-cognitive skills (Hasselgren, 2018). The grade point averages of recently graduated teacher (social science teachers in particular) have decreased in several subjects (Alatalo et al., 2021) but the effects of teacher grades on student grades or test scores have not yet been investigated. Thus, such relationships remain unknown. Furthermore, it may be difficult to draw causal inferences from correlational and descriptive data on the relationship between teacher competence and student's academic performance because, among other reasons, students are typically taught by several teachers and therefore current teachers may or may not have an impact on the recent performance (Björklund et al., 2010; Holmlund et al., 2019).

Student segregation might be a more plausible explanation for a decline of academic performance in Sweden but is difficult to disentangle from migration factors. Moreover, school segregation effects linked to decentralization do typically occur at upper-secondary level in the Swedish context (Lindbom, 2010; Yang Hansen and Gustafsson, 2016), why differences in achievement at lower levels might be due to other factors such as the SES of parents and thus students. Instead, it seems more reasonable to assert that substantial immigration leads to larger shares of low-performing individuals, schools, and municipalities (Holmlund et al., 2019; Boman, 2022b) and a relative shortage of certified teachers among schools and municipalities because increasing populations, driven by migration to Sweden (Sanandaji, 2020), lead to mismatches between the rate of certified teachers and demand for such staff within schools and municipalities in today's Sweden. Nonetheless, indicators of SES such as the share of highly educated at the municipality level are unhesitatingly correlated with school results in Sweden (e.g., Boman, 2022b). However, the impact of SES and migration background, as well as their interaction, must be further examined. For instance, aggregated school and municipality data requires replication, as well as the addition of individual level data of students in order to discern inferences at the individual level and control for more factors than can be provided at the aggregated level.

Hence, the current article adds to the corpus of research on educational achievement at lower-secondary level in Sweden, specifically aggregated data on Grade 9 in 2013 and 2018, as well as additional data from 2019. It used multiple regression analyses at the national-regional level, covering 289 out of 290 municipalities in Sweden for which data exist at the municipality database Kolada. In regard to the municipality level data, explanatory variables related to SES and migration background were analyzed in relation to grades. In addition, teacher certification rates (TCR) in each municipality, as well as geographical distance was added. Then, the independent and dependent variables of these two cross-sectional data sets were compared in order to capture fluctuations and performance trends over time and introduced some additional control variables. This article underlines the importance of replication in educational research (Makel and Plucker, 2014) as it contributes with a replication of the relationships found in Boman (2022b) by utilizing similar data sources, analytical strategies, dependent, and independent variables. It does also contribute to scholarly discussions on using aggregated data in relation to educational research.

However, it does not merely replicate the data used in Boman (2022b), as this study was recently published, but includes more data sources from in total 3 years (2013, 2018, 2019), additional predictors (e.g., teacher competence, classroom climate) and alternative dependent variables (e.g., national test results, grade point average), as well as robustness checks (e.g., a moderator analysis). As such it provides more robust and reliable results in this regard.

The following research question was addressed:

RQ1: What are the relationships between socioeconomic status, migration background, and grades at the municipality level?

The article proceeds with a theoretical background section which briefly sketches the contours of the contemporary Swedish educational context at the lower-secondary level, a literature review, and a discussion on using aggregated data, a method and data section, results, and a conclusive discussion.

Theoretical background

The contemporary Swedish education context

As Gustafsson and Yang Hansen (2018) underscore, Sweden has since 1998 moved from a normed-referenced grading system to a criterion-based grading system. The Swedish national

curriculum has been revised several times, and the last major revision for the entire school system was introduced in 2011 (Lgr 11), with a partial revision in 2018. Sweden offers 9 years of mandatory elementary school education, in which "Årskurs 7–9" can be translated into lower-secondary education (Swedish National Agency for Education, 2018). In Grade 8 ("Årskurs 8") students are typically 14–15 years old and 15–16 years old in Grade 9.

Apart from the knowledge-centered national curriculum of compulsory education, the current education system is affected by the free school choice on a voucher based quasi-market, which leads to increased school segregation (Larsson, 2019; Hennerdal et al., 2020), and high rates of low-skilled migrants the 1990s onward (Ekberg, 1999). Furthermore, many smaller municipalities in Sweden have very few schools at the secondary levels (Fjellman et al., 2019).

Fundamentally, the Swedish educational context is signified by a hybrid system of on the one hand neoliberal-oriented decentralization and school accountability and centralized control and bureaucracy on the other hand (Bunar, 2010; Hennerdal et al., 2020). These characteristics might affect patterns at the national level, such as TCR because low-performing municipalities are often compensated with more, if not better school resources such as a more beneficial student–teacher ratio (Holmlund et al., 2019). However, OECD (2022) underscores that these compensatory efforts, in reality, are quite small.

Related literature and theoretical concepts

It is a general pattern that high-SES students outperform their lower-SES peers (e.g., Sirin, 2005). That might occur because high-SES parents are more involved in their children's lives and have higher aspirations as regards their children's future trajectories in education and the labor market. In the home environment, high-SES parents spend more time reading with and to their children which likely stimulates cognitive growth and learning (e.g., Myrberg and Rosén, 2009). Moreover, children whose parents have higher SES do generally develop higher cognitive ability levels which are then transmitted to their children through both genes and the environment (Turkheimer et al., 2003; Engelhardt et al., 2018; Falk et al., 2021). Conversely, the lower degree of nurture and support among lower-SES children can affect the cognitive ability levels, and hence outputs in for example school tests, negatively (Turkheimer et al., 2003; Sackett et al., 2009; Flynn, 2012; Boman (2022d)). However, it is possible that earlier academic achievement among larger groups of students might decrease over time because of various social (e.g., immigration, fewer adolescents read in their leisure time), socioeconomic (e.g., increased relative poverty), and school-related factors (e.g., Flynn, 2012; Dutton et al., 2016). That is known as the anti-Flynn effect (Dutton et al., 2016).

Larsson (2019) underlines a socio-educational division in the current Swedish society, especially the capital Stockholm. Whereas some rural areas are associated with high achievement and high-SES family characteristics (e.g., Lundsberg in Värmland), the division is now between the renowned inner-city schools and the suburban schools with low completion rates and intake ratings (Gynatagningen, 2019). In other countries such as the US, inner-city schools are associated with poor school results, low discipline, and high rates of violence (e.g., Devine, 1996). The contention is not that all Swedish inner-city schools are high-performers, but that generally there is a positive relationship between urban locality and school results, although with many exceptions (Boman, 2022b).

Sanandaji (2020) stresses that the three most ethnically diverse municipalities in Sweden have increased their municipal tax equalization revenues since 1996 due to poor performance in the local and national labor market. A report published by Swedish National Agency for Education, 2016 indicates that approximately 6–7% of the decreases in PISA 2000–2012 are attributable to increased migrant participation. Differences between natives and non-natives in regard to educational achievement are also accentuated in a report conducted by Holmlund et al. (2019; see also Swedish National Agency for Education, 2009). In different European countries, natives do typically outperform immigrant students. Azzaloni et al. (2012) found that in Italy and Spain achievement differences in the PISA 2009 survey were associated with family background and socioeconomic status. In Italy, the effect of academic tracking was also substantial while immigrants from Latin America in Spain had an advantage in PISA reading due to their ability in Spanish. However, socioeconomic, institutional, and linguistic factors cannot fully explain the variance in both models (Azzaloni et al., 2012). Similar results were found among first- and second-generation immigrants in Switzerland when analyzing data from the PISA 2000 wave. For example, larger family sizes for first-generation immigrants, typically from Albania, had a negative statistical relationship with PISA scores (Meunier, 2011). Moreover, as Flores-Mendoza et al. (2021; see also Deary et al., 2007) underline, much of the impact of SES disappears when cognitive ability indicators are included in regression analyses. Furthermore, non-cognitive abilities such as conscientiousness or grit might have a partial impact or mediate the relationship between SES and school results (Thorsen et al., 2021).

Granvik Saminathen et al. (2018) found that teachers' ratings of school leadership, teacher education and school ethos differ across school segregation profiles. Gustafsson and Yang Hansen (2018) examined changes in the impact

of family education on student educational achievement in Sweden, 1988–2014, and found that correlations increased by 0.04 units between the early 1990s and 2014, in part because of increased immigration. However, despite efforts to adjust the dependent variable (grades) from different grading systems, this study suffers in part from the insufficient comparability over time.

In relation to teacher effects, there is a rich research literature from the United States which indicates that teachers have at least a moderate relationship with academic performance (e.g., Hanushek et al., 1996). In Sweden, there is a dearth of studies but Holmlund et al. (2019) stress that the totality of national and international evidence suggests that teachers do have a substantial effect on academic performance at lower-secondary level. However, findings from correlational and regression analyses might be spurious due to that certified teachers may be more frequent in low-performing schools and municipalities as a consequence of compensatory efforts (Holmlund et al., 2019). Edmark and Persson (2021) accentuate that there seem to be indications of grade inflation, especially in the independent schools. This might also be shown by a comparison between Sweden's PISA results from the PISA 2012 survey (e.g., 478 in mathematics compared to 510 in the 2000 survey) and the increasing grade point averages during approximately the same period (Holmlund et al., 2019).

As said, SES is intimately linked to academic achievement (e.g., Sirin, 2005; Tan, 2015; Kim et al., 2019), including Sweden (Gustafsson and Yang Hansen, 2018). However, educational results in Sweden are typically affected by both migration, geographical location, and school factors such as teacher competence or certification (Swedish National Agency for Education, 2009; Björklund et al., 2010; Gustafsson and Yang Hansen, 2018; Holmlund et al., 2019). Thus, all these factors must be considered in this context. However, these should typically not be considered as completely separate but connected factors because highly educated parents and certified teachers are generally located in the vicinities of urban areas where students might perform best (Yang Hansen and Gustafsson, 2016; Holmlund et al., 2019). Because of the impact of migration and the Swedish hybrid model in education these socioeconomic and sociodemographic patterns might be fuzzy or unexpected. For instance, a substantial portion of non-natives might be well-integrated while larger portions are not (Boman, 2021, 2022b). The geographical aspects of cultural and social capital, which are intricately linked with SES, have been underlined by Bourdieu-inspired research which is relevant for the Swedish context (e.g., Larsson, 2019). For instance, major university cities in Sweden attract students with higher cultural capital and SES (Carlhed, 2017) and some inner-city schools in Stockholm are considered as prestigious locations to attain upper-secondary education (Larsson, 2019).

The current study adds to the earlier literature such as Gustafsson and Yang Hansen (2018) and Boman (2022b). Overall, this is regarded as a replication study of Boman (2022b). More broadly, it does also add to the literature on SES–academic achievement relationships (e.g., Sirin, 2005; Tan, 2015) and migration–academic achievement relationships (e.g., Meunier, 2011), as well as research on grades in the Swedish context (e.g., Molin and Fjellborg, 2021). Specifically, it aims to replicate the results of Boman (2022b) but also contribute with a more reliable research design and as such it stands in its own feet. Akin to Molin and Fjellborg (2021), it uses school level data to discern relationships at the meso and macro levels of Swedish school students at the end of the lower-secondary level, specifically Grade 9.

Aggregated data and the ecological fallacy

Piandosi et al. (1988; see also Bronfenbrenner, 1977) have described *the ecological fallacy* as a kind inferential fallacy. The authors describe it as follows:

Variables that describe groups of individuals, rather than the individuals themselves, are termed "ecological" and are often used when the analysis of individuals' data is not possible (1). Ecological analyses may be preferred when (1) variables are more conveniently defined or measured on groups because the analysis on individuals would require excessive time or extensive data gathering; (2) ecological analyses permit study of a wider range of values for the independent variable, as in international studies of diet; (3) the precision of aggregate measures like alcohol consumption is likely to be higher for groups than for individuals; and (4) population responses such as smoking quit rates may be of primary interest.

With regard to educational research there are both pros and cons with individual level data compared to aggregated data. For instance, because it is not the same students who take the PISA tests in the triennial surveys only aggregated average scores and aggregated inputs at the school level can be used to compare changes and relationships over time (Gustafsson, 2008; Boman, 2022a). On the other hand, in relation to many other types of research such as longitudinal studies the same individuals should be followed up (e.g., Guglielmi and Brekke, 2017). Moreover, the effect sizes might be biased when for example schools and not students constitute the unit of analysis (Hennerdal et al., 2020; see also Singer, 1961).

Obviously, there are many reasons to focus on the individual level of analysis in the Swedish education context and elsewhere. However, there are good reasons for the use of public data sources where school level data are available.

First of all, these data are open to the public and it might be better that researchers, rather than organizations, politicians, pundits and journalists, analyze these data (Gustafsson, 2008). In the next step, other researchers and stakeholders may scrutinize the data and research results in accordance with open data policies. Moreover, these data cover both the meso and macro contexts of educational performance (e.g., Bennett et al., 2012; Aurini et al., 2020; see also Dopfer et al., 2004). In other words, they say something broad but substantial about school achievement and their underlying explanations.

Methods, data, and variables

Dependent variable

When examining educational achievement among Swedish students at lower-secondary level, there are typically four standardized procedures to evaluate them: grades, national tests, international tests (e.g., PISA), and specific tests constructed by researchers (Björklund et al., 2010; Holmlund et al., 2019). While all these assessments have their set of strengths and weaknesses (Lundahl, 2014), grades—non-self-reported such in particular (Kuncel et al., 2005)—are pertinent because they are the results of both cognitive and non-cognitive abilities and are predictive for the future life course among individuals (Borghans et al., 2016).

This study used aggregated data from the municipality database Kolada for information on lower-secondary level indicators in 289 of Sweden's 290 municipalities in relation to *Grades in All Subjects* (GAS) in the first step. Percentage per municipality on the degree to which students receive the documented, non-self-reported grades E–A in all subjects from Kolada is understood as the dependent variable. This measure captures both "equity and excellence" as it shows that in high-performing municipalities a large share of students performs at least "good enough" (E) in all subjects or better. Hence, it was preferred over average national test scores. Moreover, as underlined in Boman (2022b) national test results and grades are highly intercorrelated (r = 0.899) why a single measure might be quite sufficient. As Gustafsson and Yang Hansen (2018) emphasize, the inclusion of more grades increases reliability why GAS is reliable in that respect. However, might grade point averages (GPA) be a more suitable measure of grades than GAS? (e.g., Molin and Fjellborg, 2021). That aspect was considered by a consecutive bivariate analysis of the data which shows that the correlation between municipality GAS and municipality GPA for the year 2019 was r = 0.802, which indicates that a researcher may use GAS, GPA, NTR, or a combination of these three indicators in regard to educational achievement in the Swedish context. Whereas GPA is more usual than GAS it was also important to use the exact same outcome variable in order to have the replicate the earlier findings. Nonetheless, some robustness checks included GPA.

Migration background

The current study has some differences compared to Boman (2022b). Instead of using three different but moderately intercorrelated socioeconomic and sociodemographic variables (i.e., percentage of highly educated, welfare recipients, and non-natives in each municipality) it used only non-natives and only data on public schools instead of average scores for both municipal and independent schools from the regional Kolada database.[1] An advantage with Kolada in this regard, compared to Sweden's Statistics (2022) open data, is that percentages of natives and non-natives are found at the elementary and lower-secondary school level (i.e., Grade 1–9) and not just at the municipality level where only data on broad age groups (16–64) are available.

Socioeconomic status

With regard to SES, it included average income (in SEK, Swedish kronor) at the municipality level from Sweden Statistics (2022)[2] as a broad indicator of SES. The parsimony of SES in this context is because of the large intercorrelations (above r = 0.50, see Cohen, 1988) between average income and percentage of highly educated (r = 0.669), and average income and welfare recipients (r = −0.691) in Sweden. Hence, a single SES indicator was deemed sufficient. Boman (2022b), on the other hand, included the percentage of highly educated and percentage of welfare recipients as SES indicators.

It is also worth to pay attention to that Kolada does not offer comprehensive statistics from 2013 but only full data on the years 2016–2018. Therefore, only comparable independent variables were sound to use. Moreover, a comparison between for instance 2016 and 2018 would be too narrow to identify meaningful trends over time in relation to both the dependent variable and the independent variables. The selection of specific years is also related to the curriculum (Lgr 11) and its new grading system (F–A) that was implemented around 2012. Hence, 2013 constitutes the first year for students who graduated from Grade 9 at the lower-secondary level. Thus, 2013 and 2018 comprise 2 years of complete comparability in terms of the dependent variables and independent variables.

1 https://www.kolada.se/

2 https://kommunsiffror.scb.se/

TABLE 1 Regression output for GAS 2018.

	B	β	Standard error
(Constant)	32,597*		9,533
Income	0.001*	0.328	0.000
Non-natives	−0.278**	−0.173	0.059
NNI	−0.339*	−0.210	0.096
TCR	0.195*	0.155	0.089
TC	0.097**	0.090	0.051
DMC	0.004***	0.044	0.005
Classroom	0.006***	0.026	0.051

Adjusted R^2: 429. *p-value: 0.001, **p-value: 0.005, ***p-value: 0.050, ****p-value: 0.50 NNI stands for non-native increase (%). TCR, stands for teacher certification ratio. TC, stands for teacher competence. DMC, stands for distance to major city.

Control variables

As in Boman (2022b), it also included geographical distance to major city (DMC) and teacher certification rates (TCR). The DMC variables constitute the distance from any municipality to cities with at least 100,000 inhabitants, as well as the important university town Kalmar which has slightly less than 70,000 inhabitants (see **Supplementary Material**). This variable captures important nuances regarding residence. For instance, urban clusters in Sweden can only partially predict academic achievement due to their heterogenous populations (Manhica et al., 2018; Vogiazides and Mondani, 2019) and many certified teachers work in towns or cities at a commuting distance from various urban clusters or vice versa. Concretely, this was measured with https://sv.distance.to/in kilometers by matching each municipality with the closest major city. A slight negative relationship was expected, because the further from a major city, the less likely such municipalities being inhabited by high-SES families and certified teachers, on average. While imperfect, this variable captures geographical dynamics and relationships within the urban–rural continuum in a more nuanced way than binary delineations (i.e., urban/rural).

In addition, it added two more school level variables for 2018 (these were not included in the 2013, as the School inspection's survey is only available for the years 2016–2018): classroom climate (percentage of students who claim to have a peaceful classroom climate, "studiero") and teacher competence (percentage of students who claim that their teachers explain what the students should do, shortened as TC in **Supplementary Material**). These are indicators of school quality, especially teachers' basic didactic competence as it is perceived by students. For a few missing values on these two measures in some municipalities, 60% were used a standard replacement value for teacher competence and 50% for classroom discipline, which are slightly below the national averages.

TABLE 2 Regression output for GPA 2019.

	B	β	Standard error
(Constant)	230,845*		4,843
Highly educated	0.082*	0.417	0.010
Welfare recipients	−0.176*	−0.402	0.023

Adjusted R^2: 545. *p-value: 0.001.

Analytical strategy

The initial estimation model at the municipality level can be described as follows:

$$Grade^m = X^{Income}{}_m\beta + X^{Migrant}{}_m\beta + X^{Controls}{}_m\beta + e^m,$$

where *Grade* is the percentage of the degree to which pupils receive the grades E–A in all subjects in municipality ^{m}, $X^{Income}{}_m\beta$ are the coefficients of average income levels in municipality ^{m}, $X^{Migrant}{}_m\beta$ are the coefficients of average share of population with a migrant (i.e., non-native) background, and $X^{Controls}{}_m\beta$ are the coefficients of the included control variables (i.e., TCR, DMC, TC, classroom climate), and e^m is an error term.

Results

Descriptive statistics

In 2013, average aggregated GAS was 76.06% with a standard deviation of 7.96. In 2018, average GAS were 72.4% and with standard deviation of 8.86 (see **Supplementary Material**), which implies an average drop at the macro-national level by roughly 3.6%.

Regression output

Several OLS models (**Tables 1**, **2**) had problems with multicollinearity because of the medium to high intercorrelations between some of the independent variables, but average income had the strongest relationship with GAS 2013 and GAS 2018, followed by the share of non-natives. However, the differences between these effect sizes are rather small, implying that both SES and migration/socio-demographics are crucial factors in relation to grades at this level. Moreover, the increase of non-natives (NNI) from 2013 to 2018 had a rather strong bivariate relationship with GAS 2018 (r = −0.513). This is roughly equal to the positive relationship between GAS 2018 and average income (r = 0.511). Hence, some of the fluctuations in relation to the aggregated grades might be partly attributable to the increase of non-natives because of migration.

Only a few of the school related control variables (teacher competence) had a significant relationship with GAS.

TABLE 3 Regression output for GAS 2013.

	B	β	Standard error
(Constant)	41,536*		9,533
Income	0.001*	0.367	0.000
Non-natives	−0.278*	−0.271	0.059
TCR	0.178*	0.110	0.089
DMC	0.006***	0.069	0.005

Adjusted R^2: 562. *p-value: 0.001, **p-value: 0.005, ***p-value: 0.020 TCR, stands for teacher certification ratio. DMC, stands for distance to major city.

However, classroom climate and teacher competence were not strongly related to GAS.

Robustness checks

To test the robustness of the models, an alternative model was introduced which had GAS (2019) as the dependent variable in tandem with the independent variables from 2018 (see **Supplementary Material**). This is in part because the 2019 dataset (see Boman, 2022b) consists of both municipal and independent schools. The relationships were, however, robust (e.g., average income for 2018, β = 0.406).

In additional robustness models, sum scores for the 2013, 2018, and 2019 variables were created. GASsum was the dependent variable, whereas SESsum, NNsum, and TCRsum were the independent variables. The relationships were similar, although the composite SES variable had a slightly stronger relationship (β = 0.444). Moreover, GPA instead of GAS was used as the dependent variable in two additional models with 2019 data. In the first model, the highly educated SES variable was stronger (β = 0.651) when welfare recipients and geographical distance were omitted. In the second model, both geographical distance, TCR, and non-native shares became statistically significant, and the highly educated indicator effect size became smaller (β = 0.446, see **Supplementary Material**). **Table 2** shows an OLS model which had no problems with multicollinearity and included these two SES indicators. This lends some support to the explanation that it is SES rather than MB "*per se*" that explains a substantial portion of grade point averages (Holmlund et al., 2019; Boman, 2021). Moreover, this also indicates that GAS is more sensitive to migration background than GPA. This might be due to the fact that especially first-generation migration students experience difficulties in school because of language abilities (Holmlund et al., 2019; Boman, 2022b) and therefore "fail" to attain E (pass) in all 17 school subjects at end of the compulsory school.

Lastly, the author created an interaction term between SES and migration background (i.e., percentages of non-natives) with grades all subjects (2018) as the outcome variable. However, the interaction term was not statistically significant (see **Supplementary Material**), even though the interaction term was aggregated and mean centered.

Discussion

Summary of the findings

One of the aims of this study was to replicate the findings from Boman (2021) by using very similar types of data (school level data aggregated to the municipality level), methods (OLS regression models), and variables. Indeed, the effect sizes (i.e., the standardized beta coefficients) are very similar to the ones which are reported in Boman (2022b). Boman (2022b) found an effect size of β = 0.301 for SES (there measured as the percentage of highly educated individuals) for the 2019, whereas the current study found β = 0.367 for 2013 β = 0.328 for 2018. The slight differences might be partly attributable to that the SES variable was not completely identical, and that Boman (2022b) also included another SES predictor (i.e., the percentage of welfare recipients). The effect size order for the independent variables (SES, migration background, teacher certification) were, however, the same. In two robustness models, the SES beta coefficient had larger effect sizes (β = 0.406 and 0.444). This might be related to the fact Boman (2022b) used a combination of public municipal schools and free schools whereas this study used only municipal school data. Hence, when the data and variables from the two studies were merged the relationships became somewhat affected. Nonetheless, the broader patterns are the same. However, it is difficult to make reliable inferences on the relationship between TCR and GAS 2013 and GAS 2018 because two cross-sectional data sets are compared. It might also be the case that earlier teachers affect students' achievement to a larger degree than the current ones (e.g., Hanushek et al., 2019).

The overall findings of the study suggest that grades at the aggregated macro-national level have decreased by roughly 3.6% since 2013 and the overall share of non-natives has increased (in 2018), whereas TCR decreased from the first point of measure to the second. That does also indicate a partial relationship between non-native increases and decreases of certified teachers. Theoretically and empirically, this might be related to the so-called anti-Flynn effect (Dutton et al., 2016), which means that the earlier secular IQ gains (i.e., the Flynn effect, see e.g., Flynn, 2012) have turned negative because of, for example, social and educational factors (Dutton et al., 2016; Hernaes et al., 2019; Vainikainen and Hautamäki, 2022).

While distance to major cities (DMC) remains constant, the relationship between DMC and GAS has fluctuated somewhat but not significantly. Overall, based on this particular measurement there is no substantial relationship between urbanicity and grades at the national-regional level, although other studies have found such associations (e.g., Yang Hansen and Gustafsson, 2016). However, as emphasized in

Boman (2022b) it is notable that many of the high-performing municipalities in Sweden are located in the Stockholm region (e.g., Danderyd, Lidingö, see **Supplementary Material**).

The overall picture suggests the share of non-natives affects school results, and that optimal migration, integration and school policies therefore are important for the facilitation of heterogenous groups of students and to increase achievement among those that are, at least initially, low-performers. This relationship is of a similar magnitude as SES, here measured by average income levels. This is also stressed in earlier research (e.g., Gustafsson and Yang Hansen, 2018; Boman, 2022b). Specifically, here a SES variable was associated with academic achievement but because of the rather strong interrelationship with socio-demographic variables such as the share of non-natives (an indicator of migration background) it is difficult to interpret these findings. However, it appears likely that current and previous migration patterns increase the share of low-SES households and individuals (e.g., Ruist, 2015).

Some municipalities have increased GAS in 2018 compared to 2013, in spite of increased shares of non-natives, which might inspire qualitative research such as interviews and case studies with minority students that could counter negative stereotypes (e.g., Jutengren and Medin, 2019). The negative relationship between non-natives on GAS is probabilistic, not causal, or entirely pessimistic (Boman, 2021). This leaves room for further school improvement. As is notable in **Supplementary Material** (2018 data), both classroom climate, teacher competence and TCR may be improved because the values appear rather low. All such are, ceteris paribus, comparative advantages for schools and schoolchildren in general (Granvik Saminathen et al., 2018), although the effects may be small. It seems that the most crucial variables are SES, migration background, and teacher certification, which is one aspect of teacher competence (Alatalo et al., 2021). However, because of the unexplained variance (see **Tables 2**, **3**) there might be omitted variables that stem from the limitations of using aggregated data. Two candidates are cognitive and non-cognitive abilities (e.g., Flores-Mendoza et al., 2021; Thorsen et al., 2021; Boman, 2022a,c). However, it appears unlikely that such data will be collected for all or even the majority of Sweden's schools and municipalities. Attendance and classroom climate, which may be aggregated to for instance the school level, are loosely tangential to non-cognitive abilities such as conscientiousness but also related to SES (e.g., Ning et al., 2015). Thus, these are not appropriate indicators of conscientiousness. Hence, the aggregated data will not just provide somewhat biased effect sizes but also omitted variables in such regards. Nonetheless, SES (e.g., average income) and migration background (e.g., the percentage of students with a non-native background) are two important cues for school results in Sweden, and perhaps in other contexts, that seem to explain much of the variability in school results such as aggregated grade point averages or similar outcome variables.

The study has several limitations as aggregated data are comprehensive but crude and may provide misleading information on the relationships between for instance teacher quality indicators and school achievement (e.g., Hanushek et al., 1996). Furthermore, TCR and perceived basic teacher competence were weakly correlated in the present study and therefore it seems implausible that they capture a similar underlying factor. Hence, it is difficult to draw reliable conclusions in that respect. Lastly, other factors such as lower quantities of school-related reading might have influenced grades negatively over time (Vinterek et al., 2020) and this was not controlled for in the present study. Furthermore, as height is associated with educational attainment and achievement (e.g., Cinnirella et al., 2011) it would be useful to include additional biosocial covariates. However, such are not found within the frames of current data sources.

Nonetheless, the study shows that GAS were lower in 2018 than in 2013. A reason for this decline seems to be the increase of non-natives in Swedish school at lower-secondary level (and overall in society), although average income levels increase over time and the positive relationship with grades was somewhat stronger in 2018 (this was also the case with TCR). Hence, it is possible that this statistical relationship is masked by grade inflation, or conversely, that a grade inflation might take place even though the aggregated grades point to a decline of 3.6% because of the increase of non-natives due to migration. It is also plausible that the increase of non-natives has led to a decrease of certified teachers in a large share of Swedish municipalities. All in all, this shows that research on academic performance, as measured by for example grades, in the Swedish context must account for especially SES and secondarily migration background and teacher competence.

Data availability statement

The original contributions presented in this study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

Author contributions

The author confirms being the sole contributor of this work and has approved it for publication.

Funding

Stockholm University was a part of the Bibsam agreement.

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/feduc.2022.854342/full#supplementary-material

References

Alatalo, T., Hansson, A., and Johansson, S. (2021). Teacher's academic achievement: evidence from Swedish longitudinal register data. *Eur. J. Teach. Educ.* 1–21.

Aurini, J., Missaghian, R., and Milian, R. (2020). Educational status hierarchies, after-school activities, and parenting logic: lessons from Canada. *Sociol. Educ.* 93, 173–189.

Azzaloni, D., Schnell, P., and Palmer, J. R. (2012). Educational achievement gaps between immigrant and native students in two "new" immigration countries: Italy and Spain in comparison. *ANN. Am. Acad. Polit. Soc. Sci.* 643, 46–77. doi: 10.1177/0002716212441590

Bennett, P. R., Lutz, A. C., and Jayaram, L. (2012). Beyond the schoolyard: the role of parenting logics, financial resources, and social institutions in the social class gap in structured activity participation. *Sociol. Educ.* 85, 131–157. doi: 10.1177/0038040711431585

Björklund, A., Fredriksson, P., Gustafsson, J.-E., and Öckert, B. (2010). *Den svenska utbildningspolitikens arbetsmarknadseffekter. Vad säger forskningen?* Available online at: https://www.ifau.se/Forskning/Publikationer/Rapporter/2010/Den-svenska-utbildningspolitikens-arbetsmarknadseffekter-vad-sager-forskningen/ (accessed August 10, 2022).

Boman, B. (2021). Parallelization: the fourth leg of cultural globalization theory. *Integ. Psychol. Behav. Sci.* 55, 354–370. doi: 10.1007/s12124-021-09600-4

Boman, B. (2022a). PISA achievement in Sweden from the perspective of both individual data and aggregated cross-country data. *Front. Educ.* 6:753347. doi: 10.3389/feduc.2021.753347

Boman, B. (2022c). Educational achievement among East Asian school children 1967-2022: A thematic review of the literature. *International Journal of Educational Research Open*, 3.

Boman, B. (2022b). Regional differences in educational achievement among Swedish Grade 9 students. *Scand. J. Educ. Res.* 66, 610–625.

Boman, B. (2022d). The influence of SES on grades: crossectional and longitudinal evidence from two Swedish cohorts. *Eur. J. Psychol. Educ.*

Borghans, L., Golsteyn, B., Heckman, J., and Humphries, J. (2016). What achievement tests measure. *Proc. Natl. Acad. Sci. U.S.A.* 51, 13355–13359. doi: 10.1073/pnas.1601135113

Bronfenbrenner, U. (1977). Toward an experimental ecology of human development. *Am. Psychol.* 32, 513–531.

Bunar, N. (2010). Choosing for quality or inequality: current perspectives on the implementation of school choice policy in Sweden. *J. Educ. Policy* 25, 1–18.

Carlhed, C. (2017). The social space of educational strategies: exploring patterns of enrolment, efficiency and completion among Swedish Students in undergraduate programmes with professional qualifications. *Scand. J. Educ. Res.* 61, 503–525.

Cinnirella, F., Piopiunik, M., and Winter, J. (2011). Why does height matter for educational attainment? Evidence from German children. *Econ. Hum. Biol.* 9, 407–418. doi: 10.1016/j.ehb.2011.04.006

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. London: Routledge.

Deary, I., Strand, S., Smith, P., and Fernandes, C. (2007). Intelligence and educational achievement. *Intelligence* 35, 13–21.

Devine, J. (1996). *Maximum Security: The Culture of Violence in Innercity Schools*. Chicago, IL: The University of Chicago Press.

Dopfer, K., Foster, J., and Potts, J. (2004). Micro–meso–macro. *J. Evol. Econ.* 14, 263–279.

Dutton, E., van der Linden, D., and Lynn, R. (2016). The negative Flynn effect: a systematic literature review. *Intelligence* 59, 163–169.

Edmark, K., and Persson, L. (2021). The impact of attending an upper secondary school: evidence from Sweden using a school ranking data. *Econ. Educ. Rev.* 84:102148.

Ekberg, J. (1999). Immigration and the public sector: income effects for the native population in Sweden. *J. Popul. Econ.* 12, 411–430.

Engelhardt, L., Church, J., Harden, P., and Tucker-Drob, E. (2018). Accounting for the shared environment in cognitive abilities and academic achievement with measured socioecological contexts. *Dev. Sci.* 22:e12699. doi: 10.1111/desc.12699

Falk, A., Kosse, F., Pinger, P., Schildberg-Hörisch, H., and Deckers, T. (2021). Socioeconomic status and inequalites in Children's IQ and economic preferences. *J. Polit. Econ.* 129, 2504–2545.

Fjellman, A. M., Yang Hansen, K., and Beach, D. (2019). School choice and implications for equity: the new political geography of the Swedish upper secondary school market. *Educ. Rev.* 71, 518–539.

Flores-Mendoza, C., Ardila, R., Gallegos, M., and Reategui-Colareta, N. (2021). General intelligence and socioeconomic status as strong predictors of student performance in Latin American schools: evidence from PISA items. *Front. Educ.* 6:632289. doi: 10.3389/feduc.2021.632289

Flynn, J. (2012). *Are We Getting Smarter? Rising IQ in the Twenty-First Century*. Cambridge: Cambridge University Press.

Granvik Saminathen, M. G., Brolin Låftman, S., Almquist, Y. B., and Modin, B. (2018). Effective schools, school segregation, and the link with school achievement. *Sch. Effect. Sch. Improve.* 29, 464–484.

Guglielmi, S., and Brekke, N. (2017). A framework for understanding cross-national and cross-ethnic gaps in math and science achievement: the case of the United States. *Comp. Educ. Rev.* 61, 176–213.

Gustafsson, J.-E. (2008). Effects of international comparative studies on educational quality on the quality of educational research. *Eur. Educ. Res. J.* 7, 1–17.

Gustafsson, J.-E., and Yang Hansen, K. (2018). Changes in the impact of family education on student educational achievement in Sweden 1988–2014. *Scand. J. Educ. Res.* 62, 719–736.

Gynatagningen (2019). *Preliminär Antagning 2019 (Preliminary admission 2019)*. Available online at: https://www.gyantagningen.se/antagningspoang-och-rapporter/preliminar-antagning/prel-antagning-2019.html (accessed August 10, 2022).

Hanushek, E., Piopiunik, M., and Wiederhold, S. (2019). The value of smart teachers: international evidence on teacher cognitive skills and student performance. *J. Hum. Resour.* 54, 857–899.

Hanushek, E. A., Rivkin, S. G., and Taylor, L. L. (1996). Aggregation and the estimated effects of school resources. *Rev. Econ. Stat.* 78, 611–627.

Hasselgren, E. (2018). *Varför Lärarstudenter Hoppar av. En Studie om Orsakerna Bakom från Lärarutbildning vid Göteborgs Universitet. Göteborg, report 2018:2*. Göteborg: Göteborgs universitet.

Hennerdal, P., Malmberg, B., and Andersson, E. K. (2020). Competition and school performance: swedish school leavers from 1991–2012. *Scand. J. Educ. Res.* 64, 70–86.

Hernaes, Ø, Markussen, S., and Røed, K. (2019). Television, cognitive ability and high school completion. *J. Hum. Resour.* 54, 371–400.

Holmlund, H., Sjögren, A., and Öckert, B. (2019). *SOU, Jämlikhet i möjligheter och utfall i den svenska skolan. Bilaga 7 till Långtidsutredningen 2019*. Available online at: https://www.regeringen.se/4adad2/contentassets/23c13d7ae0ef48e4bed43b68917573d3/jamlikhet-i-mojligheter-och-utfall-i-den-svenska-skolan-sou-201940.pdf (accessed August 10, 2022).

Jutengren, G., and Medin, E. (2019). Cross-ethnic friendship and prosocial behavior's potential significance to elementary children's academic competence. *J. Educ. Res.* 112, 38–45.

Kim, S. W., Cho, H., and Kim, L. Y. (2019). Socioeconomic status and academic outcomes in developing countries: a meta-analysis. *Rev. Educ. Res.* 89, 875–916.

Kuncel, N. R., Credé, M., and Thomas, L. L. (2005). The validity of self-reported grade point averages, class ranks, and test scores: a meta-analysis and review of the literature. *Rev. Educ. Res.* 75, 63–82.

Larsson, E. (2019). *Innerstadsgymnasierna. En studie av tre Elitpräglade Gymnasieskolor i Stockholm och deras Positionering på Utbildningsmarknaden*. Ph.D. thesis. Stockholm: Stockholm University.

Lindbom, A. (2010). School choice in Sweden: effects on student performance, school costs, and segregation. *Scand. J. Educ. Res.* 54, 615–630.

Lundahl, C. (2014). *Bedömning för Lärare*. Lund: Studentlitteratur.

Makel, M., and Plucker, J. (2014). Facts are more important than novelty: replication in the education sciences. *Educ. Res.* 43, 304–316.

Manhica, H., Berg, L., Almquist, Y. B., Rostila, M., and Hjern, A. (2018). Labour market participation among young refugees in Sweden and the potential of education: a national cohort study. *J. Youth Stud.* 22, 533–550.

Meunier, M. (2011). Immigration and student achievement: evidence from Switzerland. *Econ. Educ. Rev.* 30, 16–38.

Molin, L., and Fjellborg, A. A. (2021). Geographical variations in the relation between final course grades and results on the national tests in social sciences, 2015–2017. *Educ. Rev.* 73, 451–469.

Myrberg, E., and Rosén, M. (2009). Direct and indirect effects of parents' education on reading achievement among third graders in Sweden. *Br. J. Educ. Psychol.* 79, 695–711.

Ning, B., Van Damme, J., Liu, H., Vanlaar, G., and Gielen, S. (2015). The influence of school disciplinary climate on reading achievement: a cross-country comparative study. *Sch. Effect. Sch. Improve.* 26, 586–611.

OECD (2022). *Policy Dialogues in Focus for Sweden. International Insights for School Funding Reform*. Paris: OECD.

Piandosi, S., Byar, D., and Green, S. (1988). The Ecological fallacy. *Am. J. Epidemiol.* 127, 893–904.

Reimer, D., Jensen, S. S., and Kjeldsen, C. (2018). "Social inequality in student performance in the Nordic countries," in *Northern Lights on TIMSS and PISA 2018* (Copenhagen: TeamNord), 31–60.

Ruist, J. (2015). The Fiscal cost of refugee immigration: the example of Sweden. *Popul. Dev. Rev.* 41, 567–581.

Sackett, P. R., Kuncel, N. R., Arneson, J. J., Cooper, S. R., and Waters, S. D. (2009). Does socioeconomic status explain the relationship between admissions tests and post-secondary academic performance? *Psychol. Bull.* 135, 1–22. doi: 10.1037/a0013978

Sanandaji, T. (2020). *Mass Challenge: The Socioeconomic Impact of Migration to a Scandinavian Welfare State*. New York, NY: Palgrave Macmillan.

Singer, J. D. (1961). *World Politics. The International System: Theoretical Essays*. Cambridge: Cambridge University Press.

Sirin, S. (2005). Socioeconomic status and academic achievement: a meta-review of research. *Rev. Educ. Res.* 75, 417–453.

SVT (2018). *0,05 på Högskoleprovet – Tillräckligt för att bli Antagen till Lärarutbilningen*. Available online at: https://www.svt.se/nyheter/inrikes/0-05-poangpa-hogskoleprovet-tillrackligt-for-att-bli-larare (accessed August 10, 2022).

Sweden Statistics (2022). *Statistics of Sweden*. Available online at: https://www.scb.se/ (accessed August 10, 2022).

Swedish National Agency for Education (2009). *Vad påverkar resultaten i svensk grundskola? Kunskapsöversikt om betydelsen av olika faktorer [What affects the results in Swedish elementary education?]*. Stockholm: Sweden's Ministry of Education.

Swedish National Agency for Education (2016). *Invandringens betydelse för skolresultaten*. Stockholm: Sweden's Ministry of Education.

Swedish National Agency for Education (2018). *Curriculum for the Compulsory School, Preschool Class and School-age Educare*. Stockholm: Sweden's Ministry of Education.

Tan, C. Y. (2015). The contribution of cultural capital to students' mathematics achievement in medium and high socioeconomic gradient economies. *Br. Educ. Res. J.* 41, 1050–1067.

Thorsen, C., Yang Hansen, K., and Johansson, S. (2021). The mechanisms of interest and perseverance in predicting achievement among academically resilient and non-resilient students: evidence from Swedish longitudinal data. *Br. J. Educ. Psychol.* 91, 1481–1497. doi: 10.1111/bjep.12431

Turkheimer, E., Haley, A., Waldron, M., D'Onofrio, B., and Gottesman, I. (2003). Socioeconomic status modifies heritability of IQ in young children. *Psychol. Sci.* 14, 623–628.

Vainikainen, M.-P., and Hautamäki, J. (2022). Three studies on learning to learn in Finland: anti-flynn effects 2001–2017. *Scand. J. Educ. Res.* 66, 43–58.

Vinterek, M., Winberg, M., Tegmark, M., Alatalo, T., and Liberg, C. (2020). The decrease of school related reading in Swedish Compulsory School – Trends Between 2017 and 2017". *Scand. J. Educ. Res.* 66, 119–133.

Vogiazides, L., and Mondani, H. (2019). A geographical path to integration? Exploring the interplay between regional context and labour market integration among refugees in Sweden. *J. Ethnic Migrat. Stud.* 46, 23–45.

Wiklund, M. (2018). The Media apparatus in the becoming of education policy: education media discourse during two electoral periods. *J. Crit. Educ. Policy Stud.* 16, 99–134.

Yang Hansen, K., and Gustafsson, J.-E. (2016). Causes of educational segregation in Sweden – school choice or residential segregation. *Educ. Res. Eval.* 22, 23–44.

frontiers | Frontiers in Education

TYPE Correction
PUBLISHED 05 January 2023
DOI 10.3389/feduc.2022.1125527

OPEN ACCESS

APPROVED BY
Frontiers Editorial Office,
Frontiers Media SA, Switzerland

*CORRESPONDENCE
Björn Boman
✉ contact@bjornboman.com

SPECIALTY SECTION
This article was submitted to Assessment, Testing and Applied Measurement, a section of the journal Frontiers in Education

RECEIVED 16 December 2022
ACCEPTED 20 December 2022
PUBLISHED 05 January 2023

CITATION
Boman B (2023) Corrigendum: Regional differences in educational achievement: A replication study of municipality data.
Front. Educ. 7:1125527.
doi: 10.3389/feduc.2022.1125527

Corrigendum: Regional differences in educational achievement: A replication study of municipality data

Björn Boman*

Department of Education, Stockholm University, Stockholm, Sweden

KEYWORDS

Sweden, educational achievement, grades, socioeconomic status, migration

A corrigendum on
Regional differences in educational achievement: A replication study of municipality data

by Boman, B. (2022). *Front. Educ.* 7:854342. doi: 10.3389/feduc.2022.854342

In the published article, there was an error in Table 2 as published. The standard error value for the intercept bad been placed in the column for the standardized beta coefficient. Now it has been moved to its proper place. The corrected Table 2 and its caption appear below.

The authors apologize for this error and state that this does not change the scientific conclusions of the article in any way. The original article has been updated.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

TABLE 2 Regression output for GPA 2019.

	B	β	Standard error
(Constant)	230,845*		4,843
Highly educated	0.082*	0.417	0.010
Welfare recipients	−0.176*	−0.402	0.023

Adjusted R^2: 545. *p-value: 0.001.

frontiers | Frontiers in Education

TYPE Systematic Review
PUBLISHED 28 July 2022
DOI 10.3389/feduc.2022.902899

OPEN ACCESS

EDITED BY
George Waddell,
Royal College of Music, United Kingdom

REVIEWED BY
Justin Dimmel,
University of Maine, United States
Dave Hewitt,
Loughborough University, United Kingdom

*CORRESPONDENCE
Ian Benson
ian.benson@roehampton.ac.uk

SPECIALTY SECTION
This article was submitted to Assessment, Testing and Applied Measurement, a section of the journal Frontiers in Education

RECEIVED 23 March 2022
ACCEPTED 30 June 2022
PUBLISHED 28 July 2022

CITATION
Benson I, Marriott N and McCandliss BD (2022) Equational reasoning: A systematic review of the Cuisenaire–Gattegno approach. *Front. Educ.* 7:902899. doi: 10.3389/feduc.2022.902899

Equational reasoning: A systematic review of the Cuisenaire–Gattegno approach

Ian Benson[1,2]*, Nigel Marriott[3] and Bruce D. McCandliss[4]

[1]School of Education, University of Roehampton, London, United Kingdom, [2]Ian Benson and Partners Ltd, London, United Kingdom, [3]Marriott Statistical Consulting Ltd., Bath, United Kingdom, [4]Educational Neuroscience, Graduate School of Education and Department of Psychology, Stanford University, Stanford, CA, United States

The Cuisenaire–Gattegno (Cui) approach to early mathematics uses color coded rods of unit increment lengths embedded in a systematic curriculum designed to guide learners as young as age five from exploration of integers and ratio through to formal algebraic writing. The effectiveness of this approach has been the subject of hundreds of investigations supporting positive results, yet with substantial variability in the nature of results across studies. Based on an historical analysis of one of the highest-fidelity studies (Brownell), which estimated a treatment effect on equation reasoning with an effect size of 1.66, we propose that such variability may be related to different emphases on the use of the manipulatives or on the curriculum from which they came. We conducted a systematic review and meta-analysis of Cui that sought to trace back to the earliest investigations of its efficacy. Results revealed the physical manipulatives component of the original approach (Cuisenaire Rods) have had greater adoption than efforts to retain or adopt curriculum elements from the Cuisenaire–Gattegno approach. To examine the impact of this, we extended the meta-analysis to index the degree to which each study of Cuisenaire Rods included efforts to align or incorporate curricular elements, practices, or goals with the original curriculum. Curriculum design fidelity captured a significant portion of the variability of efficacy results in the meta-analysis.

KEYWORDS

aptitude-treatment interactions, arithmetic fluency, NCTM pre-algebra, Cuisenaire-Gattegno, Cuisenaire rods

1. Introduction

Educational policy changes have shifted the focus of early mathematics education research from arithmetic (computation with numbers) towards algebra (computation with *types*) (NCTM, 2000; Greenes and Rubenstein, 2008). In this paper we offer some *technical vocabulary* to elucidate this transition. We recover some of the intellectual history of early algebra research: the use of the Cuisenaire–Gattegno (Cui) curriculum. Section 2 introduces the historical context of research on introducing equational reasoning into the early years of school mathematics and the mixed results of employing manipulatives within contemporary curricula. In Section 3, we describe the distinctive characteristics of the Cui programme, an integrated approach to manipulatives and

curriculum. We review William Brownell's post-test experiment with Cui on which we based our study design. We structure our meta-analysis of the literature on Cui effectiveness to test his hypothesis that it's not just using Cuisenaire rods that leads to the significant effects, but fidelity to Gattegno's curriculum and pedagogy. Section 4 reports the results of that analysis.

Section 5 discusses the contribution of this work and next steps. Two online appendices provide supplementary material. Appendix A documents the 37 studies from which the meta-analysis is drawn. Cui was developed by Gattegno in collaboration with the developmental psychologist Jean Piaget and with Jean Dieudonné, an author of the Bourbaki reforms to mathematics education. A similar initiative taken by Davydov and his colleagues in the Soviet Union is receiving renewed attention in the contemporary literature (Coles, 2021). Like Cui Davydov's curriculum "develops algebraic structure from the relationships between quantities such as length, area, volume, and weight. The arithmetic of the real numbers follows as a concrete application of these algebraic generalizations... In a study in which the entire 3-year elementary curriculum of Davydov was implemented in a US school setting, children using the curriculum developed the ability to solve algebraic problems normally not encountered until the secondary level in the US" (Schmittau and Morris, 2004, p 60). Gattegno goes further than Davydov. He advocates from the outset the study all four arithmetic operations and unit fractions as operators for small numbers. Appendix B discusses the relationship of Cui to Piaget's theories and Davydov's experiments.

2. Early algebra research, manipulatives, and the reform of school mathematics

Algebra encompasses the relationships between quantities, the use of notation, the modeling of phenomena, and the mathematical study of change. While the word algebra is not often heard in elementary school classrooms, the mathematical experiences and conversations of students in early grades frequently include elements of pattern recognition and related algebraic reasoning.

Much of the debate about the nature of algebra in secondary school mathematics ignores this pre-algebraic experience. It focusses instead on the problems students face with techniques of symbol manipulation when algebra is introduced after arithmetic. For example, in discussing seventh grade student difficulties (Herscovics and Linchevski, 1994, p. 76) notes that "the detachment of a number from the preceding minus sign had a high incidence and this indicates that evaluating strings of operations is not a trivial problem. These difficulties indicate that some of the problems in early algebra find their origin in the students' arithmetic background and warrant further investigation."

Hewitt (2011) in his study of secondary mathematics with the virtual manipulative *Grid Algebra* notes that to achieve proficiency in algebraic reasoning students need to be able to switch between several levels of abstraction:

- Algebra as appearance of letters.
- Algebra as working with or on the unknown.
- Algebra as an expression of generality using actions, words and gestures.
- Algebra as seeing the general in the particular and the particular in the general, and after Gattegno.
- Algebra as an attribute of the mind. Here he argues that "students were working with operations in order to carry out these tasks and the awareness of equivalence of different sets of operations was certainly operating upon operations" (Hewitt, 2011, p 9).

In this paper we will be concerned with how and how well early algebra might serve as an enabler of *arithmetic proficiency* (accuracy) and *arithmetic fluency* (accuracy and response time) and as preparation for future learning of equational reasoning. We review the role that physical and virtual manipulatives play in supporting both conventional school mathematics, and the conceptually enriched curriculum of Cuisenaire–Gattegno (Cui). Equational reasoning is a particularly important activity in elementary algebra and in reasoning about the behavior of computer programs (O'Donnell et al., 2006; Sangwin, 2015). *Equational reasoning*, operating on equations, includes substituting equivalent expressions within part of an equation as well as other forms of reasoning such as operating on both sides of an equation or splitting a single equation into cases. Asserting that two expressions A and B are equivalent means that in certain circumstances A may be replaced by B and vice versa. Asserting that two equations A = 0 and B = 0 are equivalent is subtly but crucially different. It means that the solutions of A = 0 are precisely the solutions of B = 0 i.e., those particular values of the variables coincide.

Equational reasoning is important for several reasons. For example students might be asked to give an example of a quadratic equation whose roots are $x = 3$ and $x = 5$. (Sangwin, 2005, p. 441) reports that most of his first year undergraduate students tackled this without the slightest hesitation. Nevertheless some of his weaker students "(enough to notice a pattern) did not realize that the factored form of a quadratic would provide an almost immediate answer and instead wrote the quadratic as $p(x) = ax^2 + bx + c$" and attempted to solve the simultaneous equations that resulted from substituting in the two roots.

Reasoning by equivalence is a refinement of equational reasoning: a repetitive formal symbolic procedure where algebraic expressions, or terms within an expression, are replaced by an equivalent until a "solved" form is reached. The goal is to replace an expression or a sub-expression in a problem

by an equivalent expression to provide a new problem having the same solutions.

In high school graduation examinations a third of examinable content is reasoning by equivalence (Rasila and Sangwin, 2016). Students typically do not pay attention to domains of definition or explicitly indicate which steps guarantee equivalence of adjacent lines and which do not. For example when undergraduate students are asked to solve equations such as $(x + 5)/(x - 7) - 5 = (4x - 40)/(13 - x)$, they typically reason by equivalence working line by line. Most students need many lines of working, for this example typically about a dozen. This is problematic because "elementary algebra contains a number of subtle 'traps', including division by zero, or gaining/loosing solutions by squaring/square rooting both sides of an equation" (Rasila and Sangwin, 2016, p. 4).

Sangwin (2016) notes that equational reasoning is as important in undergraduate mathematics as it is in computer science since:

1. It exists at every level from solving linear equations onwards.
2. It is the start of proof & rigor (deductive geometry).
3. It contains logic and extended calculation.
4. It is a part of many methods, e.g., solving ordinary differential equations.
5. It is a key part of many pure mathematics proofs: the induction step, epsilon-delta proofs.
6. It enables reasoning about and verification of software.

The Massachusetts Comprehensive Assessment System (MCAS) is a high stakes standardized test that has been used as an efficient opportunity to gather data on early algebra interventions over time. Narrative reports of small scale quasi-experiments with early algebra suggest that even a limited exposure to equational reasoning can help children to out perform their peers when they take part in MCAS (Kaput and Blanton, 2000; Schliemann et al., 2007). A longitudinal intervention study in Boston has shown that introducing algebra as part of the early mathematics curriculum is highly feasible. Specific representational tools—manipulatives, tables, graphs, numerical and algebraic notation, and certain natural language structures—can be employed to help students express functional relations among numbers and quantities and solve algebra problems (Carraher et al., 2008).

The evidence that given an appropriate "mathematising situation" young learners are capable of sophisticated reasoning continues to mount. It accumulates in the developing market for customized apps and in the literature recounting small scale experiments with pattern making with physical manipulatives and structured drawings (Radford, 2014, 2018; Borthwick et al., 2021). It has led to a renewed attention to equational reasoning. Some of this activity builds explicitly on the pioneering work of Caleb Gattegno and his collaborators working with Cuisenaire rods in the 1950's (Mason, 2008; Benson, 2011; Goutard, 2017; Adom and Adu, 2020). Other researchers, working from first principles, have independently discovered many of Gattegno's findings especially those relating to the central importance of early algebra, pattern making, and mathematical equivalence (Davydov, 1962; Kaput, 1995a,b; Healy et al., 2002; Schmittau and Morris, 2004; Carraher et al., 2005; Schliemann et al., 2007; Baez, 2009; Mulligan and Mitchelmore, 2009; Blanton and Kaput, 2011; Cai and Knuth, 2011; Empson et al., 2011; McNeil et al., 2011; Rittle-Johnson et al., 2011; Kieran et al., 2016; Gadanidis et al., 2018; Kieran, 2018; Matthews and Fuchs, 2018; Simsek et al., 2021).

Gattegno was a working mathematician and educator, and an early collaborator on mathematics teaching reform with the influential developmental psychologist Jean Piaget (Piaget and Szeminska, 1952; Sfard, 1995). Piaget had a substantial influence on the school mathematics curriculum in the West. He identified human thought itself with logico-mathematical structures and held a rigorous view on how children would grow their understandings. Both he and Gattegno paid attention to integrating conceptual mathematics into their theories of mathematical cognition (Choquet, 1963; Piaget et al., 1992). Piagetian commentators "have almost universally accepted that his 'mathematisation' is at worst 'ideosyncratic' and left it alone, concentrating on his claim to have demonstrated the process of acquiring knowledge through the clinical method" (Seltman and Seltman, 1985, p. viii).

By contrast Gattegno brought together a Commission of mathematicians that included Evert Beth, inventor of the semantic tableau used in formal reasoning, Jean Dieudonné, a prime mover in the Nicolas Bourbaki group that reformed university mathematics after WWII and Gustave Choquet whose work on capacities and integral representations found many applications in analysis and probability. Choquet was founding Commission President. He studied the Cui experiments teaching young children with Cuisenaire rods and became both an adept at the approach and a skilled user of the rods. Choquet's "What is Modern Mathematics," became the Commission's manifesto. In it he drew attention to some of the key tools of Bourbaki's axiomatic method: *sets, functions, morphisms, categories*, and *functors*.

We have adopted this conceptual mathematical definition of algebraic structure, in particular the notion of a *type* as found in mathematics and computer science, where amongst other things it names a property common to the elements of a set. Expressed in these terms Gattegno's definition of *algebraic awareness* may be regarded as an appreciation that the composition of two elements of the same type can result in a third element with the same property.

Choquet wrote "Since Bourbaki has such clear-cut concepts and is so intimately associated with the development of mathematics in our time, we can hope that a study of 'his' philosophical and mathematical work may lead us to the essence of modern trends in analysis. Such as study may serve to develop for all levels of education a teaching of mathematics better

adapted to the needs of our time and the level of awareness of our generation" (Choquet, 1963, p. 3).

Manipulatives like the Cuisenaire, Stern, and Montessori materials have found a place in Western mathematics classrooms from the time of diagnostic testing with counters (a la Piaget), to contemporary bead strings, Numicon tiles and the Rekenrek abacus. Today they are often augmented by toys such as the Rubik cube, animations such as BBC Numberblocks and "virtual" manipulatives, delivered through the web, on a tablet or on a standalone computer.

For the most part physical manipulatives such as Dienes blocks and animations such as Numberblocks represent decimal numbers. The Cui approach is an exception in that the rods are not given prescribed number names, rather names are first encoded as letters and then resolved to values by measuring one rod with another (the unit). This emphasis on measurement as a basis for number is shared with Davydov's approach. He writes, "such introduction of whole numbers greatly facilitates the subsequent mastering of fractions—both simple and decimal—since the child understands from the very outset, first that abstract number as a relationship, and, second, the value being measured as a homogeneous object that may be measured with any degree of precision" (Davydov, 1962, p. 35).

In their definitive meta-analysis of physical manipulatives, Carbonneau et al. (2013) found that "simply incorporating manipulatives into mathematics instruction may not be enough to increase student achievement in mathematics." They identified several factors that determined the size of effect: "instructional variables such as the perceptual richness of an object, level of guidance offered to students during the learning process, and the development status of the learner moderate the efficacy of manipulatives." Jones et al. (2019) note that a major drawback in such quantitative research studies is that while many studies seek to measure conceptual understanding most observations assess only procedural or surface understanding. They have shown how to create more sophisticated metrics in their work of the efficacy of computer applications for learning algebra.

Gilmore et al. (2017) explored the procedural skill, conceptual understanding and working memory capacity of 75 children aged 5–6 years as well as their overall mathematical achievement. They found that, not only were all three capabilities independently associated with mathematics achievement, but there was also a significant interaction between them. In fact levels of conceptual understanding moderate the relationship between procedural skill and mathematics achievement. Fuchs et al. (2014) conducted a controlled experiment with fourth grade at risk students with interventions in fraction learning, emphasizing fluency and conceptual knowledge. Results revealed a significant aptitude-treatment interaction, in which students with very weak working memory learned better with conceptual activities but children with more adequate (but still low) working memory learned better with fluency activities.

Virtual manipulatives enable an even more customized interaction although "something may be being lost in the translation" (Nemirovsky and Sinclair, 2020, p. 107). Especially for young children, technology manipulatives may be more manageable and extensible. In one study, third graders working with technology manipulatives made statistically significant gains learning fraction concepts (Reimer and Moyer, 2005). Although most apps for young learners concentrate on handwriting training and drill and practice, some create direct manipulation situations in which the underlying mathematical structure can be accessed (Bakos and Pimm, 2020). For example for 3–5 year olds, Little Digits is an iOS app that uses fingers to work out all permutations and combinations (number bonds) for small numbers, one author's *notHiding* is a one or two player pelmanism game to develop strategies to map between colors and their letter codes and between upper and lower case letter forms and Dragon Box introduces linear equations (Benson, 2012; CowleyOwl, 2012; DragonBox, 2012). Thai et al. (2021) reports on a cluster randomized study of a digital game-based learning environment that provides personalized content and adaptive embedded assessments which shows that it can improve mathematics knowledge of transitional kindergarten and kindergarten students.

3. Methods

Our goal was to review through a systematic analysis the historical development of Gattegno's pioneering work and its reception, with the intention of subsequently abstracting, replicating and extending the most promising statistical findings. In Section 3.1, we describe the distinctive aspects of the Cuisenaire–Gattegno approach. One of the highest-fidelity studies was due to William Brownell who designed an unusual longitudinal experiment to investigate the efficacy of Cui. In Section 3.2, we explain how Brownell created a balanced quasi-experiment. We do this to highlight some important effects and to motivate both the meta-analysis and a subsequent study (forthcoming) which examines the long term transfer effects.

3.1. Cuisenaire–Gattegno: An integrated approach to manipulative and curriculum

Cuisenaire rods are cuboids, the length of each a multiple of the length of the smallest—a 1 cm white cube. Rods of the same size have the same color. Each student has a box containing sufficient rods of different sizes to construct all the partitions of the smaller rods (Figure 1). In Cui physical and diagrammatic set combination and mathematical writing interacts with domain general reasoning aptitude

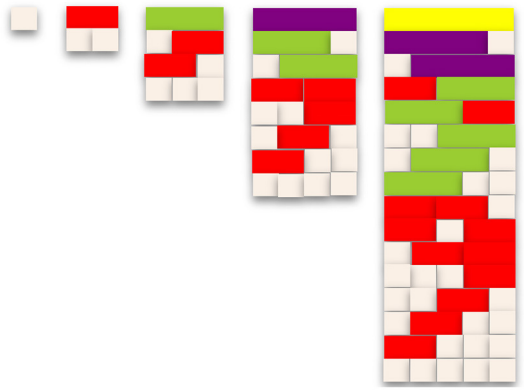


FIGURE 1
"Complete patterns" for small integers [1..5].

as a preparation for arithmetic proficiency. Figure 2 shows how this educates learners' sensitivity to common patterns of mathematical relations by coordinating "vision, audition, haptic, sensorimotor and introspective modalities through constructions with color-coded rods of unit increments" (ATM, 1977, p. 185). Gattegno introduces the integers to teachers as the "numeral names for a sequence of diagrams constructed by partitioning" (Gattegno, 2010a, p. 80).

This experience of number is enhanced by the use of mathematical vocabulary, symbols and notation. From the outset Gattegno introduces the concept of *"equivalence"* as a generalization of "equivalent color" and "equivalent length." Each "complete pattern" in the sequence of diagrams corresponds to an equivalence class of partitions of an integer (Figure 1). Other examples of equivalence are "equivalent expressions" (such as *"w+r"*, *"r+w"*) and equivalent equations. Figure 2 shows the conceptual coverage in the first 2 years of schooling. Concepts such a powers, roots, and logarithms go beyond the entitlements of the statutory UK National Curriculum. They prepare the way for the study of number systems of different bases: multi-digit numerals being formed by juxtaposing polynomial coefficients. This brings out the structure of the number system directly, in contrast with the conventional emphasis on the *"place-value"* reading of written numerals which takes up so much time in the early grades.

Color codes and expressions are at the same time named integer values, computed by measuring the length of one rod by another, and recipes for colored rod constructions: "+" for example being the action of placing two rods end to end to form a "train". Gattegno generalizes the concepts of school algebra to encompass sensitivity to the dynamic that combines two objects of the same type (*"w"*,*"r"*) to form a third of that type (a named rod construction). He intended to make teachers and pupils aware of this dynamic which transforms rod constructions, diagrams, written expressions and equations into equivalent forms. He contrasted this *"algebraic awareness"* of the nature of number systems with traditional symbol manipulation in school algebra and with drill-based factual fluency (Gattegno, 1983). He summarized his philosophy in these terms, "the most important lesson that teachers can learn is that rather than teach mathematics we should strive to make people into mathematicians" (Gattegno, 2010a, p 82).

Gattegno uses operations with the rods—placing them end to end, side by side or stacked as towers—to model sets with structure such as the integer and rational number systems. In the Cui approach "all the operations with integers and fractions can be studied simultaneously (with colored rods); whole numbers being recognized as the equivalence class of their partitions and fractions as ordered pairs, one serving to measure the other, or as operators belonging to classes of equivalence which are the rational numbers involved in the operations" (Fedon, 1966, p. 201). He demonstrated that "Children of six or seven are thoroughly familiar with their tables, children of five conceive and compare fractions easily and accurately, children of eight solve simultaneous equations and at 10 they understand permutations and combinations which they themselves form and analyse" (Gattegno, 1956, p. 88).

The Cui programme has four distinctive characteristics.

Firstly, it consists of a suite of textbooks and teachers' guides with exercises with permutations of rods. These encourage the learner to pay attention to the relationship between quantities. They give rise to a substantial experience with integers and rational numbers (Cuisenaire and Gattegno, 1953, 1962; Gattegno, 1959, 2010a, 2011a; Benson, 2011; Goutard, 2017; Adom and Adu, 2020).

Secondly, the exercises are organized in a concept graph with 55 key mathematical concepts and their inter-dependencies. Gattegno calls this a map of elementary mathematics derived from tables of partitions. The map is drawn as a directed graph—a data structure studied in computer science. Nodes representing concepts are linked by a network of arrows. The graph introduces learners from the outset to concepts such as *equivalence, set, function and domain*. The arrows illustrate the dependencies between the concepts. The graph has four *root nodes* based on a study of subsets of the complete patterns of partitions. The hierarchy of conceptual dependencies is in places eight levels deep (Gattegno, 2010a; ATM, 2017; Cane, 2017). The technical vocabulary in the concept graph covers two sets of ideas: concepts that appear both in the graph and the textbooks are intended for learners, concepts that appear only in the graph are for teacher education. Coverage of the concepts means that teachers understand the graph in its totality. The idea that teachers need to know more than the statutory school curriculum in order to teach mathematics well is sometimes called "subject matter knowledge at the mathematical horizon" (Zazkis and Mamolo, 2011).

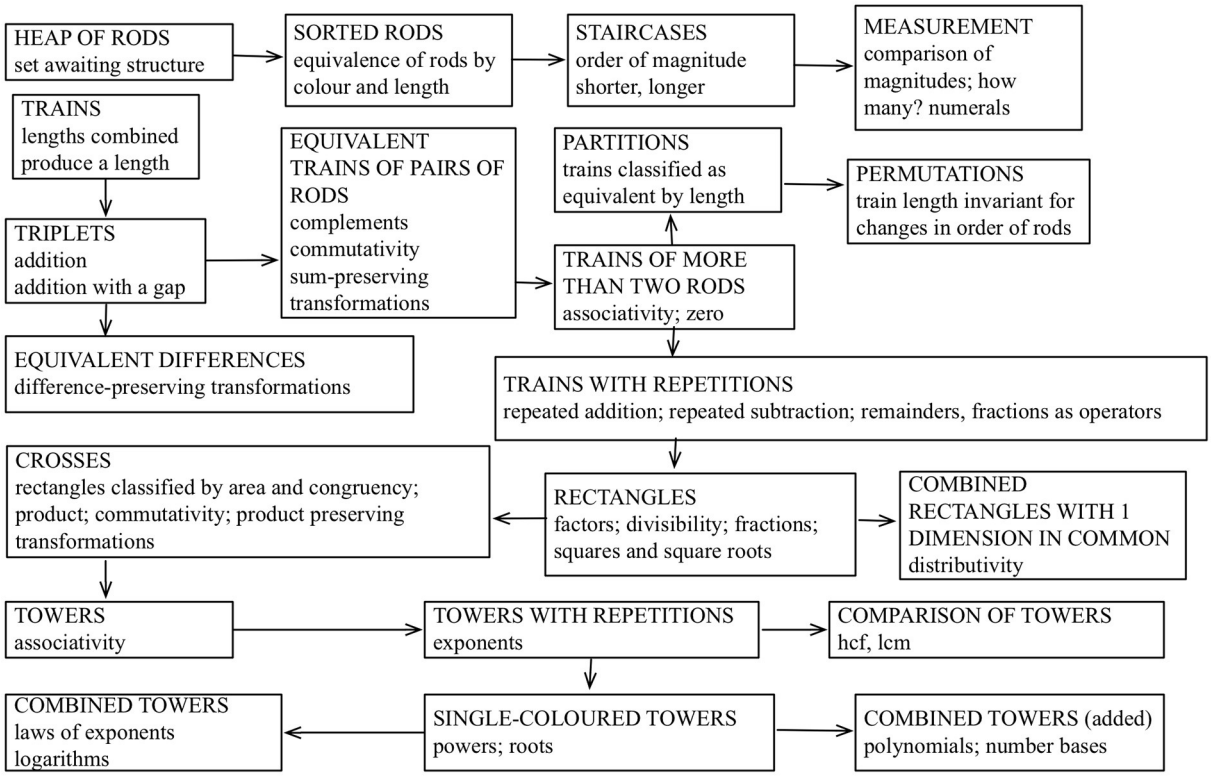


FIGURE 2
A flow chart for the Cui approach to coordinating vision, audition, haptic, sensorimotor and introspective modalities.

Thirdly, young children write expressions and equations in all four arithmetic operations and unit fractions as operators—initially for computation with types and subsequently for computation with small numbers. Gattegno called this sequence *"algebra first"* in contrast with conventional *"counting first"* school mathematics.

Fourthly, the *"subordination of teaching to learning"*: a theory of learning based on conscious (or unconscious) *"awareness"* as the unit of study (Gattegno, 1970, 1987, 2010c; ATM, 2018). Young and Messum (2011) have reviewed this model of human learning and shown how it can be applied both inside and outside the classroom. Griffin (2018) has described the questions teachers ask themselves when designing mathematical tasks in this approach:

- What might students (or teachers) be noticing (inside themselves) when engaged in the activity—what awarenesses might arise?
- How can I maximize the possibility that these awarenesses are available to the students (or teachers), that there is an awareness of these awarenesses so that it enables action—i.e., that the awareness can be educated.
- What is my role as the teacher in all this? When do I "step in" and when do I "step away" in order that the student is genuinely working with their own awareness but I am supporting that process and maybe helping it to be more efficient—how can my teaching be subordinated to the learning? and, when working with teachers:
- What activities and approaches enable teachers to be aware of this phenomenon themselves (that it is profitable for students to be aware of their own awareness) and consider how they might support this in their students—awareness of awareness of awareness.

3.2. Study design

Observational studies of early adopters of Cui were generally positive and in British Columbia a Royal Commission recommended a large-scale study with a view to integrating the method into elementary teacher training programmes (Howard, 1957; Ellis, 1964). Such findings encouraged researchers to compare the Cui vs. Conventional approach. Robinson cites 50

qualitative comparisons employing 15,000 students over several grade levels. He writes, "One could say that research reported to date has compared the effects of some 20,000 student years of Cuisenaire exposure to the effects of the equivalent amount of 'traditional' instruction" (Robinson, 1964).

Gattegno's work caught the attention of William Brownell, a pioneer of educational research and sometime president of the American Educational Research Association (Kilpatrick and Weaver, 1977). Brownell was open to Gattegno's intellectual ambition since he believed that "Children differ markedly in the ways in which they think of numbers and in the ways in which they learn number facts. No adequate measurement of degrees of development can be made, therefore, unless the measures of speed and accuracy are supplemented by a measure of the maturity of the processes employed in dealing with numbers" (Brownell, 1928, p. 201). As Dean Emeritus of the Berkeley School of Education Brownell undertook several large scale quantitative and qualitative studies of Cui (Brownell, 1967a,b).

Our study design drew on Brownell (1967b), an unusual design for this kind of evaluative research and one of the larger longitudinal studies. We will describe the study in some detail as it was the most comprehensive study to date. It was conducted in Scotland and California. Brownell administered pen and pencil tests to (n = 1,109) learners who remained in the program after 3 years of schooling—at the end of Scottish Primary III. It was a post-test-only control quasi-experiment classified as design type 6 by Campbell and Stanley (1963). Brownell recruited classrooms from 24 schools. Half of the classes had followed a pure Cuisenaire–Gattegno course of study, and half the traditional "counting first" curriculum. Teaching intensity averaged between 33 and 67 min per day. Accordingly Brownell divided his data into longer and shorter durations of study. Brownell assessed children's domain general cognitive skills that fall outside mathematics *via* a standardized verbal reasoning test, although he conceptualized this scholastic aptitude as "IQ" (sic) at the time. This test was administered at the end of the 3 years (Brownell, 1967b). Learners were selected at random from each group, matched by age, gender and verbal reasoning skills. High and Low scholastic aptitude subjects were determined by removing the middle 20% from the verbal reasoning distribution. This resulted in a smaller sample of 405 X and 453 C. The data was then divided into eight cells based on treatment (X, C), scholastic aptitude (Hi, Lo) and teaching intensity (high, low). Teaching in the range 31–34 min per day was classified as low intensity, and the range 47–64 min per day was taken as high intensity (Brownell, 1967b). From these eight cells, one cell would have been identified as having the smallest sample which in this case was 38. For statistical inference testing, it is desirable to have equal sample sizes in each cell. The reason why Brownell does this is to eliminate unwanted correlations between the additional variables e.g., scholastic aptitude and intensity of teaching. By doing this, he ended up mimicking a balanced experimental design which in an ideal world would have been achieved before the tests were administered. Obviously in this case it was not practical since children are allocated to schools by their parents and local authorities and not by Brownell. To achieve a balanced design Brownell removed samples from the other seven cells at random until he had 38 pupils in each cell. His final sample size was 304. This meant 1,003 of the original 1,337 population were excluded. By setting aside data in this way Brownell introduced a potential risk that the excluded pupils might have given different results.

He tested material covered in both courses of study (the Common test), and content covered in only one of them (the CUI and TRA tests). Brownell used an ANOVA test to confirm that the differences and interactions between effects were significant. High teaching intensity studies showed evidence of a treatment effect in all three tests. The interactions between treatment and scholastic aptitude in all three tests were statistically significant. Referring to the aptitude-treatment interaction Brownell wrote that "it is reasonable to suggest that children identified as low in intelligence and exposed to a relatively long period of instruction in arithmetic will gain more through involvement in the Cui program" (Brownell, 1967b). In the case of the CUI test it is children who scored highest on his scholastic aptitude task who gained the most.

3.3. Systematic review protocol

The goal of the meta-analysis was to evaluate the effectiveness of the Cuisenaire–Gattegno interventions on measures of mathematical performance. To find all studies that met our criteria, we conducted a literature search using the search terms Cuisenaire, Cuisenaire Gattegno, and Cuisenaire Gattegno quasi-experiment in the full text databases of ProQuest dissertations, theses and scholarly journals, ERIC, Google Scholar, JSTOR and Association of Teachers of Mathematics. We included the journals Educational Studies in Mathematics, Arithmetic Teacher, Mathematics Teacher, Review of Educational Research, For the Learning of Mathematics, Mathematical Gazette, Journal for Research in Mathematics Education and Zentralblatt für Didaktik der Mathematik. In the case of Masters and Doctoral dissertations we followed up bibliographic references. Where possible we consulted or obtained copies of the primary sources and repeated our enquiries on subsequent bibliographic references.

An initial search was conducted in Stanford libraries in 2005. It was last updated in March 2022. In total, the Cuisenaire searches returned 1,189 Proquest items and 5,490 Google Scholar items. Cuisenaire Gattegno returned 151 Proquest and 1,310 Google Scholar items. These abstracts were investigated for relevance to the topic. Relevant abstracts included general reviews of the use of manipulatives and references to experiments and quasi-experiments in elementary schools. This produced a long list of 37 quantitative studies

for which abstracts were available (with full-text examination if necessary to determine inclusion). These are summarized in a table in Appendix A.

These 37 studies examined the impact of Cuisenaire rods on arithmetic development in children including those which reported a metric for arithmetic understanding. These tests quantify performance with arithmetic operations. They range from evaluating simple addition and subtraction expressions to missing number sentences to working with fractions. We looked for tests that could inform our research with the Woodcock-Johnson Mathematics Fluency subscale, a metric widely used in cognitive, educational and neuro-imaging studies (Woodcock et al., 2007). We excluded four foreign language dissertations that did not have an English translation, observational studies and studies where the control did not follow a traditional curriculum. Our analysis required reported means and standard deviation or sufficient statistical detail to allow us to impute these values. One dissertation was excluded as it did not report means.

These experiments can be distinguished by the experience of teachers with the Cui approach, type of intervention and control, the number of final sample subjects (n), grade level, duration, design [Experiment (EX), Quasi-experiment (QEX), Observational (OB)], availability of pre-test and post-test means and standard deviations, within and between subjects analysis and fidelity to the Cui approach. Unless otherwise reported, as in Brownell (1967a), a school year is taken as 180 days of teaching at five mathematics lessons of 50 min per week.The direction of the reported effect is shown as Cui = Control, Cui > Control, or Cui < Control. Peer reviewed findings were equally balanced between Cui and conventional teaching. Other studies were more favorable to Cui.

In preparation for the meta-analysis we excluded foreign language studies, du Bon Pasteur (1966), Bellemare (1967), Lin (2013) and Huang (2019) and all of which reported a direction for the effect of Cui > Control. We also excluded Brownell (1967a, 1968) which was a three way study in the relative conceptual development achieved by Cui, Tra (Traditional), and Dienes programs assessed using the techniques of observation and interview. We excluded observational studies in which there was no explicit control (Beard, 1964; Steencken, 2001; Bulgar, 2002; Marchese, 2009; Yankelewitz, 2009) or where the control didn't follow a conventional curriculum (Gell, 1963; Fedon, 1966; Sweeney, 1968; Lamon and Scott, 1970; Fennema, 1972; Keagle and Brummett, 1993). Rich (1972) was excluded as his experiment was not restricted to Cuisenaire. Rodman (1964), Rawlinson (1965) and Allen (1978) were excluded as they did not report means.

Whilst the remaining papers and dissertations recorded means and sample sizes, many were poor at recording the standard deviations. We included studies where a standard error of difference in the means, p-value, T or F statistic was included under the assumption that the coefficient of variation would be the same for experimental and control samples. This allowed us to impute the standard deviations for Nasca (1966) and Dairy (1969) although Dairy (1969) only reported means for her Kindergarten sample. Hollis (1964, 1965) reported means for three different pre-post tests. We excluded her evidence as we found no basis to estimate the relative coefficients of variation for the 3 different types of tests.

Haynes (1963) described two experimental ($E1, E2$) samples with a single control sample ($C3$). It was possible to explicitly derive 3 sample standard deviations using simultaneous equations and compare these with our imputation method. When we did this the largest error was 7%. Since the standard error of the difference between mean experimental and mean control was known for each pair, this allowed us to compute the pooled variance, PV, as follows (nX and nC being the size of the experimental and control samples):

$$PV = \frac{Standard\_Error\_Diff^2}{\frac{1}{nX} + \frac{1}{nC}}$$

Similarly pooled variance may be calculated as a weighted average of the sample variances where the weights are the sample degrees of freedom. Since the experimental and control sample sizes were identical we were able to derive each pooled variance as a straightforward average of the sample variances. Thus we ended up with 3 simultaneous equations

$$\begin{aligned} V_{E1} + V_{E2} &= 2 * PV_{E1E2} \\ V_{E1} + V_{C3} &= 2 * PV_{E1C3} \\ V_{E2} + V_{C3} &= 2 * PV_{E2E2} \end{aligned}$$

Which were solved to derive the sample variances (V_S).

Robinson (1978) like Haynes (1963) reported two experimental classes matched with a single control. In both cases we amalgamated the two experiments by taking a weighted average of the means and calculating the combined standard deviation. Egan (1990) uses different measures for pre and post tests and is included only in the post-test analysis.

It was not possible to recover the standard deviations for Passy (1963a,b) as we could not discover the true sample sizes. The sample sizes given in the peer-reviewed article are much higher than implied by the degrees of freedom in an ANOVA table in his dissertation. This suggests some data has been removed but no explanation is given as to how and why the data was removed. Ellis (1964) doesn't mention p-values, T or F statistics or standard error of difference so we were not able to recover the standard deviation. Adom and Adu (2020) reported an effect size of 5 with a T2X standard deviation more or less the same as the T1X data. Since the standard deviation is normally proportional to the mean, and the mean doubled we

would expect a doubling of the standard deviation. We therefore excluded it from the meta-analysis.

3.4. Meta-analysis

Meta-analysis was performed using the open-source statistical software package R, and employing the `metafor` package. Analyses were carried out using the standardized mean difference (effect size) as the outcome measure. A random-effects model was fitted to the data. The amount of heterogeneity (i.e., τ^2), was estimated using the restricted maximum-likelihood estimator (Viechtbauer, 2005). In addition to the estimate of τ^2, the Q-test for heterogeneity (Cochran, 1954) and the I^2 statistic are reported (Higgins and Thompson, 2002). In case some amount of heterogeneity is detected (i.e., $\tau^2 > 0$, regardless of the results of the Q-test), a prediction interval for the true outcomes is also provided and shown at the bottom of the forest plot. It is centered at the summary estimate, and its width accounts for the uncertainty of the summary estimate, the estimate of between study standard deviation in the true treatment effects (τ), and the uncertainty in the between study standard deviation estimate itself. It indicates the possible treatment effect in an individual setting (Riley et al., 2011). Studentized residuals and Cook's distances are used to examine whether studies may be outliers and/or influential in the context of the model (Viechtbauer and Cheung, 2010). Studies with a studentized residual larger than the $100 \times (1 - 0.05/(2 \times k))$th percentile of a standard normal distribution are considered potential outliers (i.e., using a Bonferroni correction with two-sided $\alpha = 0.05$ for k studies included in the meta-analysis). Studies with a Cook's distance larger than the median plus six times the interquartile range of the Cook's distances are considered to be influential. The rank correlation test (Begg and Mazumdar, 1994) and the regression test (Sterne and Eggar, 2005), using the standard error of the observed outcomes as predictor, are used to check for funnel plot asymmetry.

4. Results

After systematic application of these inclusion principles, 13 studies were deemed to pass all the above criteria. The process of selection of studies is summarized in Figure 3. These remaining studies gave rise to a collections of post test reports and pre-post test reports. To investigate the effect of fidelity to Cui we created a weighted ranking of the 13 studies, according to dimensions of fidelity suggested by Brownell. Several of these studies contained more than one comparison between control and treatment conditions appropriate for inclusion in the meta-analysis, such as when results were reported separately for males and females and by grade. In all this gave rise to k = 23 post-test contrasts at grade and gender level and k = 8 pre-post contrasts, each contrast representing an independent and distinct population of students. Where studies presented results from two or more independent samples (each with a control group) that received the same intervention they were coded as distinct assessments in our analysis. This gave a final assessment count of 23 (n = 1,968, nX = 1,096, nC = 928) for the post-test meta-analysis and 8 (n = 465, nX = 244, nC = 221) for the pre-post meta-analysis.

In each study we selected an outcome measure that best captured the construct of arithmetic fluency and best approximated the Woodcock-Johnson Mathematics Fluency subscale. Five studies reported the Metropolitan Readiness or Achievement Test, two studies the Science Research Associates Arithmetic test and other studies measured proficiency with fractions and missing number sentences (see Table 1). Brownell reported his raw data results at a test item level. We used the items below to construct a measure of arithmetic proficiency from his Common test missing number sentences that we could compare with the studies in our meta-analysis and we could use in our replication and extension study (forthcoming) (Brownell, 1967b, and Appendix):

$$
\begin{aligned}
2 \times \square &= 12 \\
12 - \square &= 7 \\
6 + \square &= 14 \\
9 - \square &= 0 \\
\square + 7 &= 10 \\
\square - 5 &= 7 \\
\square \div 2 &= 7 \\
\square + 8 &= 8
\end{aligned}
$$

Studies can be distinguished by the experience of teachers with the Cui approach, the number of final sample subjects (n), grade level, gender, frequency and duration of mathematics lessons, experiment design, control design, statistical tools and fidelity to the Cui approach.

Effect sizes were computed directly from the means and standard deviation values obtained from the manuscripts without regard for statistical significance reported in the source materials. For example, in one case (Haynes, 1963), a contrast originally reported as a null result appears in Table 1 as a small effect.

4.1. Quantifying fidelity to central Cui scholarship, curriculum, and pedagogy

The Cui approach was transmitted to the world through specific artifacts: an original curriculum and text books intended for children, scholarly books and papers, secondary literature that related Cui to main currents of mathematics education research and accounts of adoption. We explored an hypothesis that transmission became less effective the further a study

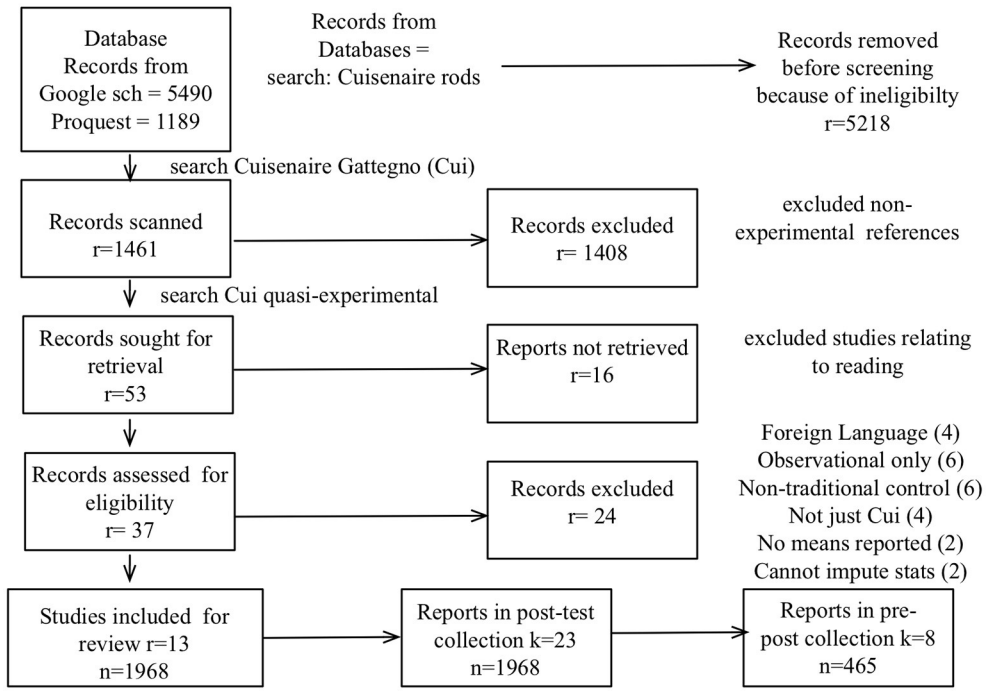


FIGURE 3
A Prisma flow chart for the selection of studies.

TABLE 1 Experiments included in the post-test meta-analysis ranked in order of fidelity (Peer reviewed findings are marked *).

Study	*n*	Grade	Days	Effect(d) C.I.	Metric
*Brownell (1967b)	304	3	540	1.66 (1.40, 1.92)	Missing number sentences
Wallace (1974)	154	4–6	15	0.99 (0.66, 1.33)	Area model for fractions (Wallace, 1974, p. 85-9)
Steiner (1964)	102	4	180	0.53 (0.12, 0.93)	Metropolitan Achievement Test, Arithmetic Computation
Aurich (1963)	90	1	180	1.38 (0.92, 1.84)	Science Research Associates Arithmetic
Robinson (1978)	119	3, 4	5	0.10 (-0.29, 0.48)	Decimal fractions (Robinson, 1978, p. 95-114)
Haynes (1963)	63	3	30	0.37 (-0.16, 0.90)	Metropolitan Achievement Test, Arithmetic Computation
Crowder (1965)	425	1	143	0.25 (0.06, 0.45)	Science Research Associates Arithmetic
Egan (1990)	81	2	180	−0.30 (-0.74, 0.14)	Missouri Mastery Achievement Test (Mathematics)
Dairy (1969)	53	K	540	0.85 (0.29, 1.42)	Metropolitan Readiness Test
*Nasca (1966)	45	2	180	−0.09 (-0.68, 0.49)	Metropolitan Achievement Test, Mathematics
Romero (1977)	240	1–6	160	0.44 (0.19, 0.70)	Metropolitan Achievement Test, Mathematics
Keagle and Brummett (1993)	38	4	4	−0.56 (-1.12, 0.09)	Custom Fraction Test
*Lucow (1962)	254	3	30	0.65 (0.40, 0.90)	Growth in × and ÷

drifted away from these benchmarks and that this might account for a significant element of the heterogeneity in the true effects/outcomes in the meta-analysis.

We quantified these aspects of the studies in four dimensions: the curriculum experienced by the learner ($rank_{learn}$), the teacher's experience with Cui ($rank_{teach}$), the teachers' Cui training ($rank_{train}$) and the preparation of the research team ($rank_{research}$). The 13 studies were compared by an independent adjudicator against one another in each dimension and ranked in order from most (1) to least (13) faithful. The adjudicator holds a PhD in applied mathematics. She was familiar with the overall literature, Cui classrooms and the criteria for ranking. The studies themselves were anonymized. In the event that all 13 studies were distinctive she ranked them from 1 to 13. In other dimensions were there were fewer distinctions some rankings were duplicated or not assigned.

The relative weights for these dimensions were chosen to reflect Brownell's account of his studies. He wrote "Dr. Gattegno stressed algebra more, and arithmetic less, than had M. Cuisenaire; and he formulated a system of instruction to which British teachers who follow the "Cui. program" adhere more or less scrupulously: Cuisenaire and Gattegno (1953), Gattegno (1957), Gattegno (2010b), Gattegno (2011b)" (Brownell, 1967a, p.14). We gave the highest weighting (4) to this curriculum and pedagogy as this is what the learners experience moment by moment. Then we weigh teacher experience (3) and preparation to deliver the curriculum with fidelity (2) and finally we weigh the evidence of researcher awareness of the debate on "number first" vs. "algebra first" progression (1). The overall metric for fidelity for a study was computed with the formula

$$fidelity = 4*rank_{learn} + 3*rank_{teach} + 2*rank_{train} + rank_{research}$$

In the learn dimension the highest ranking was given to reports that exhibited evidence that they used Gattegno's curriculum in the classroom. Credit was given if the study reproduced a précis of the Cuisenaire–Gattegno approach and cited the seminal text-books for pupils (Gattegno, 1957, 1963). Brownell (1967a), for example, devoted seven pages to a description of "computation in the Cuisenaire program" written by the teacher who coordinated teacher training for his study. The lowest ranking studies have only a rudimentary account of Cui. They do not cite the seminal books.

In the teaching experience dimension the highest rankings were given to studies that reported more than 1 year's prior teaching experience with the approach.

In the teacher training dimension we looked for citations of Gattegno's seminal teacher training books and his writing on educational research. These influential works are listed in the bibliography below. This was taken to be evidence of the quality of teacher training.

In the research dimension we assessed the preparation of the research team by examining the extent to which the study's bibliography and Sections 5 covered the contemporary literature on early algebra and manipulatives.

Once the set of fidelities for the 13 studies had been computed it was mapped into an ordinal variable with values 1–8. This was calculated by dividing the difference between highest and lowest value into eight equal intervals, and assigning the resulting "fidelity rank" to each of the 13 studies. We did this because we wanted to design a moderator with a granularity that took account of the subjective nature of the classification. We didn't think that it was warranted to use the precision that the raw fidelity statistic implied.

These measures can only be informed by what the authors choose to report in their papers or dissertations. It could be that the authors did not mention something that was very significant within one or more of these dimensions. Nevertheless the literature as a whole conforms to Mason's observation that educational research tends to privilege novelty over coherence. He writes, "In the early 1980s I had the chance to attend a number of seminars led by Caleb Gattegno when he tried to re-vivify his science of education in the mathematics education community in England. ...I began to get a taste of what it is like when an experienced "gray-beard" assembles their to-them-coherent-and-comprehensive framework or theory. Whereas when the fragments were being worked on and described there is often considerable interest amongst colleagues, once the whole is assembled, people don't really want to know" (Mason, 2010, p. 5). Brownell's early attention to fidelity in study design, which was echoed by du Bon Pasteur (1966) and Bellemare (1967), is exceptional in the literature by the care taken to reflect the original Cui framework.

4.2. Results of the meta-analysis

The analysis was carried out using R (version 4.0.4) (R Core Team, 2020) and the metafor package (version 2.5.82) (Viechtbauer, 2010). Analysis was carried out using two different approaches: a random effects model for three analysis of arithmetic proficiency (k = 8, 13, 23), and a mixed effects model for the analysis of the fidelity rank as a moderator (k = 13). Several of the 13 studies in Table 1 presented results from two or more independent samples (each with a control group) that received the same intervention. They were coded as distinct assessments in our analysis, giving an assessment count of k = 23 (n = 1,968) for the post-test meta-analysis and k = 8 (n = 425) for the pre-post meta-analysis.

Metafor takes pooled standard deviation from the samples at T1 and T2. This assumes that the subjects are different at the two time points—which they are not in general. As a result the pooled standard deviation is an overestimate and the effect size is an underestimate.

In the first r = 13 analysis we used a single measure per study (i.e., k, the number of contrasts, was 13) as shown in Table 1. The weighted average effect size was d = 0.5 (95% C.I. 0.16, 0.84)

TABLE 2 Pre-post-test effect size (d), Confidence Intervals (C.I.) for the influence of Cui on arithmetic proficiency outcomes.

Study	Grade	Effect (d)	Effect C.I.
Wallace (1974)	4	1.29	(0.68,1.89)
Wallace (1974)	5	0.59	(0.03,1.15)
Wallace (1974)	6	0.24	(−0.30,0.79)
Steiner (1964)	4	0.43	(0.03,0.83)
Aurich (1963)	1 Boys	1.12	(0.53,1.72)
Aurich (1963)	1 Girls	1.04	(0.36,1.70)
Robinson (1978)	3	0.59	(0.01,1.17)
Robinson (1978)	4	−0.24	(−0.82,0.34)

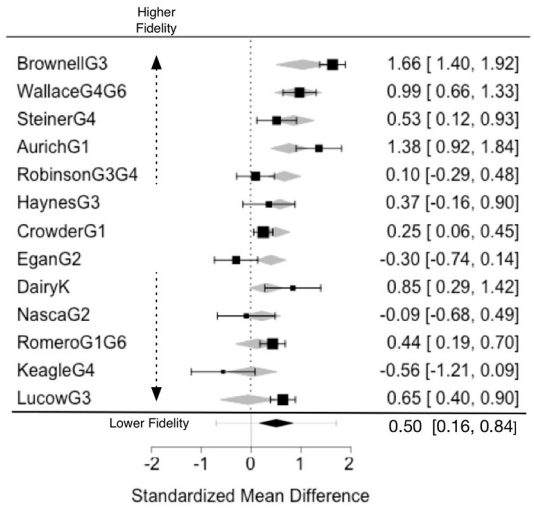


FIGURE 4
Post-test effect size (d) showing predicted (diamond) and observed (bar) proficiency outcome effect sizes by experiment in order of fidelity. Prediction interval and summary "diamond" for C.I. for estimate.

with the majority of estimates being positive (77%). Therefore, the average outcome differed significantly from zero ($z = 2.8969$, $p = 0.0038$). Cohen suggested that $d = 0.2$ be considered a "small" effect size, 0.5 represents a "medium" effect size and 0.8 a "large" effect size (Cohen, 1988). That is, if two groups' means do not differ by 0.2 standard deviations or more, the difference is trivial, even if it is statistically significant. We analyzed subgroups of studies according to the measure chosen. For the nine independent studies using the Metropolitan Achievement Test ($n = 450$) there was a small effect size of 0.34 (95% C.I. 0.10, 0.59) and for the 3 Science Research Associates arithmetic tests ($n = 515$) there was a large effect size 0.94 (95% C.I. 0.16, 1.72).

We calculated the prediction interval for the $k = 13$ analysis (−0.70, 1.71) with the metafor `predict` function. This indicates that the average effect does not tell us much about what happens in any particular study as there is a great deal of heterogeneity, that is between study variance. In Section 4.3, we explore how we might account for this variation. The $r = 13$ studies gave rise to $k = 23$ post-test reports, and $k =$ 8pre-post reports.

The weighted effect size for the $k = 23$ post-test experiments was $d = 0.55$ (experimental sample size nX = 1,040, control sample nC = 928). The Confidence interval was (0.3, 0.8) and prediction interval (−0.56, 1.66).

The pre-post meta-analysis is shown in Table 2. These assessments used the same metrics as those in Table 1. The prediction interval was (−0.24, 1.47) with a weighted effect size of $d = 0.61$ (nX = 244, nC = 221).

Figure 4 shows the observed outcome effects for the $r = 13$ studies in Table 1. The three random effects models confirm that our findings are broadly robust to treating each study as one observation rather than treating independent samples within each study as separate assessments.

4.3. Assessing the effect of fidelity

We built a mixed effects model to study the extent to which arithmetic proficiency was influenced by fidelity to the Cui approach. The 13 experiments were ordered within each dimension by an external adjudicator. A weighted average ranking from 1 to 8 was calculated for each experiment and the results entered as a moderator in the meta-analysis.

Figure 4 shows the observed proficiency outcomes and a prediction based on the mixed effects model by experiment in order of fidelity. The gray diamonds show the predicted effects and their CI limits. The model shows that when fidelity changes by 1 on the 1 to 8 scale we used, the estimated effect size decreases by 0.19. The effect size for fidelity 1 was 1.2 which reduced to effect size −0.06 for fidelity 8. We checked to see if the effect of fidelity was non-linear but the model showed no sign of that and so our final model assumes the effect of fidelity is linear.

According to the Q-test, the true outcomes appear to be heterogeneous [$Q_{(12)} = 135.7691$, $p < 0.0001$, $\tau^2 = 0.3461$, $I^2 = 91.8758\%$]. A 95% prediction interval for the true outcomes is given by −0.6990 to 1.7054. Hence, although the average outcome is estimated to be positive, in some studies the true outcome may in fact be negative.

An examination of the studentized residuals revealed that none of the studies had a value larger than ±2.8905 and hence there was no indication of outliers in the context of this model. According to the Cook's distances, none of the studies could be considered to be overly influential. Neither the rank correlation nor the regression test indicated any funnel plot asymmetry ($p = 0.6754$ and $p = 0.1617$, respectively).

A statistically significant relationship between treatment effect size and the rank order of fidelity to Gattegno's curriculum/pedagogy was revealed by a QM test of moderators [$Q_M(\text{df} = 1) = 5.8416$, $p = 0.0157$] (Viechtbauer, 2021). As evident in Figure 4 studies with the highest fidelity rankings produced effect sizes >1, while effects fell off systematically as evidence of fidelity to the original work waned. In fact, rank order of fidelity to the seminal work accounted for 32% of the heterogeneity of outcomes (R^2).

5. Discussion

5.1. Findings

In this paper we have brought together two pieces of scholarship that interact and combine to form a new view of Cuisenaire–Gattegno. We have reappraised (Brownell, 1967b) one of the most rigorous previous studies and conducted

a meta-analysis guided by Brownell's observations on the need for fidelity. In a forthcoming paper we report on a replication-extension of Brownell's experiment to investigate his hypothesis that the algebraic understanding gained by following the Cui approach will underpin later arithmetic and algebraic proficiency.

Brownell held that "one cannot "play around" with the Cui program.... expertness of the teachers is a prime requisite to success. Otherwise, classroom activities with the Cuisenaire rods may amount to no more than the haphazard manipulation of colored sticks" (Brownell, 1967a, p. 195). Our meta-analysis concurred that fidelity of transmission of the Cui equational reasoning approach is a moderator in arithmetic proficiency.

Attribute-treatment interactions such as the one reported by Brownell are increasingly studied in mathematics education research. This is because individual differences in children's cognitive resources are associated with mathematics learning, even when individual differences in elementary mathematics knowledge are statistically controlled. This indicates that mathematics intervention should be designed to help students with poor foundational mathematics skills compensate for limitations in the cognitive resources associated with poor learning.

5.2. Conclusions

Gattegno's work promoting Cuisenaire's invention and developing the Cui curriculum was seen by Brownell and his colleagues as a promising direction for mathematics education research. Their appraisal was endorsed by teachers' associations across the francophone and anglophone worlds. Our meta-analysis has highlighted that Cuisenaire rods can have a large effect on arithmetic proficiency and algebraic understanding if rigorous attention is given to the appropriate curriculum and pedagogy.

The meta-analysis showed that the average outcome is estimated to be of medium effect size, yet the efficacy of this approach is remarkably heterogeneous. Rather than attributable to noise, efficacy results appear to follow a pattern of diffusion, in which strong effects associated with the seminal curriculum materials and pedagogical practices dissipated as the teaching aides were adapted and the curriculum materials that inspired them were left behind. A high fidelity to the Cui approach was associated with a large effect size (1.2). This impact was reduced by 16% for each of eight levels of divergence from a benchmark we based on Brownell.

The policy implications are significant. As with all pedagogical interventions we have asked the key questions, who does it benefit? and, in what contexts? Our findings endorse Brownell's conclusions that learners falling below expected levels of academic performance may benefit most from gains in arithmetic fluency while leaners of all aptitudes will gain in algebraic reasoning. While his study can be readily adapted by researchers and teachers as a successful intervention in early years algebra through equational reasoning these results suggest that adoption of the Cuisenaire rods alone may be insufficient, and that careful consideration of how to effectively adopt the original curriculum and pedagogy is advisable.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

IB, NM, and BM contributed to conception and design of the meta-analysis and performed the statistical analysis. IB organized the dataset and wrote the first draft of the manuscript. NM and BM wrote sections of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

Funding

The work has been part funded by a network of schools, the Ogden, Sutton and Shuttleworth Foundations, the Greg and Rosie Lock Charitable Foundation and Sociality Mathematics CIC. The UK government provided funding through the Maths Hubs of the Department of Education National Centre for Excellence in the Teaching of Mathematics and the Department's Primary Strategy Learning Networks. Funding also came from the UK Department of Trade and Industry Global Watch programme. The authors declare that this study received funding from Apple Inc. The funder was not involved in the study design, collection, analysis, interpretation of data, the writing of this article or the decision to submit it for publication.

Acknowledgments

Our thanks go to the senior leadership, teachers and learners in participating schools and to friends and colleagues for their hospitality at Stanford. Piers Messum, Anne Haworth, Greg Gomberg, and Steve Everhard helped devise and deliver professional development support. We are grateful for the assistance of library staff at the University of Laval, Quebec, Bibliotheque Nationale de France, the National Library of Australia, the Moore and University Libraries, Cambridge University and Cubberley Library, Stanford University. The authors acknowledge the valuable contributions of Jan Atkinson, Oliver Braddick, Colin Foster, Martin Hyland, and Anna Vignoles and the reviewers who commented on earlier drafts of this paper. We are grateful for an equipment grant and

advice from John Couch and Janet Wozniak at Apple Inc and guidance on outreach project design from Bob Moses' Algebra Project and John Chowcat of Prospect, the UK union for school improvement consultants. The project forms part of the Tizard outreach initiative of Churchill College, Cambridge, 1967 mathematicians.

Conflict of interest

Author IB is the Director of a non-profit entity Sociality Mathematics CIC and Director of Ian Benson and Partners Ltd. He provides professional development services to a network of schools in the UK and US related to topics and findings reported in this manuscript. Author NM was employed by Marriott Statistical Consulting Ltd.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/feduc.2022.902899/full#supplementary-material

References

Adom, G., and Adu, E. O. (2020). The use of Cuisenaire rods on learners' performance in fractions in grade 9 in Public High Schools in Chris Hani West District, South Africa. *Int. J. Sci. Res. Publ.* 10, 2250–3153. doi: 10.29322/IJSRP.10.06.2020.p10215

Allen, H. R. (1978). *The Use of Cuisenaire Rods to Improve Basic Skills (Addition-Subtraction) in Seventh Grade* (Ph.D. thesis). New Brunswick, NJ: Rutgers, The State University of New Jersey.

ATM (1977). *Notes on Mathematics for Children*. Association of Teachers of Mathematics.

ATM (2017). *Working with the Rods and Why*. Association of Teachers of Mathematics.

ATM (2018). *On Teaching and Learning Mathematics with Awareness*. Association of Teachers of Mathematics.

Aurich, S. M. R. (1963). *A comparative study to determine the effectiveness of the Cuisenaire method of arithmetic instruction with children at first grade level* (master's thesis). Catholic University of America, Washington, DC, United States.

Baez, J. (2009). *Can Five-Year-Olds Compute Coproducts? n-Category Cafe*. Available online at: http://golem.ph.utexas.edu/category/2009/12/can_fiveyearolds_compute_copro.html

Bakos, S., and Pimm, D. (2020). Beginning to multiply (with) dynamic digits: fingers as physical-digital hybrids. *Digit. Exp. Math. Educ.* 6, 145–165. doi: 10.1007/s40751-020-00066-4

Beard, D. K. (1964). *An intensive study of the development of mathematical concepts through the Cuisenaire method in three year olds* (master's thesis). Southern Connecticut State University, New Haven, CT, United States.

Begg, C., and Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics* 50, 1088–1101. doi: 10.2307/2533446

Bellemare, T. (1967). *La Methode Cuisenaire-Gattegno et le development operatoire de la pensee* (Ph.D. thesis). University Laval, Quebec, QC, Canada.

Benson, I. (2012). *notHiding iOS app*. Available online at: https://apps.apple.com/us/app/nothiding/id521900115

Benson, I. (2011). *The Primary Mathematics: Lessons from the Gattegno School*. Saarbrücken: Lambert Academic.

Blanton, M. L., and Kaput, J. J. (2011). "Functional thinking as a route into algebra in the elementary grades," in *Early Algebraization: A Global Dialogue From Multiple Perspectives*, eds J. Cai and E. Knuth (Springer). doi: 10.1007/978-3-642-17735-4_2

Borthwick, A., Gifford, S., and Thouless, H. (2021). *The Power of Pattern: Pattening in the Early Years*. Association of Teachers of Mathematics.

Brownell, W. (1968). Conceptual maturity in arithmetic under differing systems of instruction. *Element. Schl. J.* 69, 151–163. doi: 10.1086/460493

Brownell, W. A. (1928). *The Development of Children's Number Ideas in the Primary Grades*. University of Chicago.

Brownell, W. A. (1967a). *Arithmetical Abstractions: The Movement Towards Conceptual Maturity Under Differing Systems of Instruction*. University of California, Berkeley, CA.

Brownell, W. A. (1967b). *Arithmetical Computation: Competence After Three Years of Learning Under Differering Instructional Programmes*. Available onelin at: https://eric.ed.gov/?id=ED022703

Bulgar, S. (2002). *Through a teacher's lens: Children's constructions of division of fractions* (Ph.D. thesis). New Brunswick, NJ: Rutgers.

Cai, J., and Knuth, E. J. (2011). *Early Algebraization*. Heidelberg: Springer. doi: 10.1007/978-3-642-17735-4

Campbell, D. T., and Stanley, J. (1963). "Experimental and quasi-experimental designs for research on teaching," in *Handbook of Research on Teaching*, ed N. L. Gage (London: Rand McNally) 25–27.

Cane, J. (2017). Mathematical journeys: our journey in colour with Cuisenaire rods. *Math. Teach.* 257, 7–11.

Carbonneau, K. J., Marley, S. C., and Selig, J. P. (2013). A meta-analysis of the efficacy of teaching mathematics with concrete manipulatives. *J. Educ. Psychol.* 105, 380–400. doi: 10.1037/a0031084

Carraher, D. W., Martinez, M. V., and Schliemann, A. D. (2008). Early algebra and mathematical generalization. *ZDM Int. J. Math. Educ.* 40, 3–22. doi: 10.1007/s11858-007-0067-7

Carraher, D. W., Schliemann, A. D., and Brizuela, B. (2005). "Treating the operations of arithmetic as functions," in *Journal for Research in Mathematics Education, volume 13 of Monograph Medium and Meaning: Video Papers in Mathematics Education Research* (Reston, VA: NCTM) 1–17. Available online at: https://www.jstor.org/stable/30037

Choquet, G. (1963). *What Is Modern Mathematics*? Available online at: https://issuu.com/eswi/docs/1162_what-is-modern-mathematics

Cochran, W. G. (1954). The combination of estimates from different experiments. *Biometrics* 10, 101–129. doi: 10.2307/3001666

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Abingdon: Routledge.

Coles, A. (2021). Commentary on a special issue: Davydov's approach in the XXI century: views from multiple perspectives. *Educ. Stud. Math.* 106, 471–478. doi: 10.1007/s10649-020-10018-9

CowleyOwl (2012). *Little Digits app*. Available online at: https://apps.apple.com/gb/app/little-digits/id511606843

Crowder, A. B. (1965). *A Comparative study of two methods of teaching arithmetic in the first grade* (Ph.D. thesis). North Texas State University, Denton, TX, United States.

Cuisenaire, G., and Gattegno, C. (1953). *Numbers in Colour: A New Method of Teaching the Processes of Arithmetic to All Levels of the Primary School, 3rd Edn*. London: Heinemann.

Cuisenaire, G., and Gattegno, C. (1962). *Initiation a la méthode, Les nombres en couleurs*. Denges: Delachaux et Niestlé.

Dairy, L. (1969). *Does the Use of Cuisenaire Rods in Kindergarten, First and Second Grades Upgrade Arithmetic Achievement*? Available online at: https://eric.ed.gov/?id=ED032128

Davydov, V. V. (1962). An experiment in introducing elements of algebra in elementary school. *Soviet Educ.* 5, 27–37. doi: 10.2753/RES1060-9393050127

DragonBox (2012). *Dragon Box iOS App*. Available online at: https://itunes.apple.com/gb/app/dragonbox-algebra-5/id522069155

du Bon Pasteur, T. (1966). *La méthode Cuisenaire et le développement opératoire de la pensée: recherche psychopédagogique sur l'efficacité de la méthode Cuisenaire* (Ph.D. thesis). University Laval, Quebec, QC, Canada.

Egan, D. L. (1990). *The effects of using Cuisenaire rods on the math achievement of second grade students* (master's thesis). Warrensburg, MI: Central Missouri State University.

Ellis, E. N. (1964). *The Use of Coloured Rods in Teaching Primary Number Work in Vancouver Public Schools*. Available online at: http://www.eric.ed.gov/PDFS/ED028823.pdf

Empson, S. B., Levi, L., and Carpenter, T. P. (2011). "The algebraic nature of fractions: developing relational thinking in elementary school," in *Early Algebraization: A Global Dialogue from Multiple Perspectives*, eds J. Cai and E. Knuth (Berlin: Springer) 409–428. doi: 10.1007/978-3-642-17735-4_22

Fedon, J. P. (1966). *A study of the Cuisenaire-Gattegno method as opposed to an eclectic approach for promoting growth in operational technique and concept maturity with first grade children* (master's thesis). Temple University, Philadelphia, PA, United States.

Fennema, E. (1972). The relative effectiveness of a symbolic and a concrete model in learning a selected mathematical principle. *J. Res. Math. Educ.* 3, 233. doi: 10.2307/748490

Fuchs, L. S., Schumacher, R. F., Sterba, S. K., Long, J., Namkung, J., Malone, A., et al. (2014). Does working memory moderate the effects of fraction intervention? An aptitude-treatment interaction. *J. Educ. Psychol.* 106, 499–514. doi: 10.1037/a0034341

Gadanidis, G., Clements, E., and Yiu, C. (2018). Group theory, computational thinking, and young mathematicians. *Math. Think. Learn. Int. J.* 20, 1403542. doi: 10.1080/10986065.2018.1403542

Gattegno, C. (1956). New developments in arithmetic teaching in Britain: introducing the concept of 'Set'. *Arithmetic Teach.* 3, 85–89. doi: 10.5951/AT.3.3.0085

Gattegno, C. (1957). *Arithmetic with Numbers in Colour*. London: William Heinemann.

Gattegno, C. (1959). Thinking afresh about arithmetic. *Arithmetic Teach.* 6, 30–32. doi: 10.5951/AT.6.1.0030

Gattegno, C. (1963). *Mathematics with Numbers in Colour: Numbers from 1 to 20*, Vol. 1. Fishguard: Educational Explorers.

Gattegno, C. (1970). *What We Owe Children: The Subordination of Teaching to Learning*. New York, NY: Outerbridge and Dienstfrey.

Gattegno, C. (1983). On algebra. *Math. Teach.* 105, 34–36.

Gattegno, C. (1987). *Science of Education: Part I Theoretical Considerations*. New York, NY: Educational Solutions.

Gattegno, C. (2010a). *Common Sense of Teaching Mathematics*. New York, NY: Educational Solutions.

Gattegno, C. (2010b). *Now Johnny Can Do Arithmetic: A Handbook on the Use of Coloured Rods*. Fishguard: Educational Explorers.

Gattegno, C. (2010c). *Science of Education: Part 2B Awareness of Mathematisation*. New York, NY: Educational Solutions.

Gattegno, C. (2011a). *Modern Mathematics with Numbers in Colour*. Fishguard: Cuisenaire Company Ltd.

Gattegno, C. (2011b). *A Teacher's Introduction to the Cuisenaire-Gattegno Method of Teaching Arithmetic*. New York, NY: Educational Solutions.

Gell, J. A. (1963). *An evaluation of the Cuisenaire method of teaching arithmetic* (Master's thesis). Southern Connecticut State University, New Haven, CT, United States.

Gilmore, C., Keeblea, S., Richardson, S., and Cragg, L. (2017). The interaction of procedural skill, conceptual understanding and working memory in early mathematics achievement. *J. Num. Cogn.* 3, 400–416. doi: 10.5964/jnc.v3i2.51

Goutard, M. (2017). *Mathematics and Children*. Fishguard: Educational Explorers.

Greenes, C. E., and Rubenstein, R. (eds.). (2008). *Algebra and Algebraic Thinking in School Mathematics*. Reston, VA: NCTM.

Griffin, P. (2018). "A diary of a working group," in *On Teaching and Learning Mathematics With Awareness*, eds D. Brown, A. Coles, and J. Ingram (Derby: Association of Teachers of Mathematics) 4–19.

Haynes, J. O. (1963). *Cuisenaire rods and the teaching of multiplication to third-grade children* (Ph.D. thesis). Tallahassee, FL: Florida State University.

Healy, L., Pozzi, S., and Sutherland, R. (2002). "Reflections on the role of the computer in the development of algebraic thinking," in *Perspectives on School Algebra*, eds R. Sutherland, T. Rojano, A. Bell, and R. Lins (Dordrecht: Springer), 231–247. doi: 10.1007/0-306-47223-6_13

Herscovics, N., and Linchevski, L. (1994). A cognitive gap between arithmetic and algebra. *Educ. Stud. Math.* 27, 59–78. doi: 10.1007/BF01284528

Hewitt, D. (2011). "What is algebraic activity?" in *Proceedings of the 7th Congress of the European Society for Research in Mathematics (CERME)*, eds M. Pytlak, T. Rowland, and E. Swoboda, (Rzeszñw).

Higgins, J. P. T., and Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine* 21:1539–1558. doi: 10.1002/sim.1186

Hollis, L. Y. (1964). *A study to compare the effects of teaching first and second grade mathematics by the Cuisenaire-Gattegno method with a traditional method* (Ph.D. thesis). Texas Technological College, Lubbock, TX, United States. doi: 10.1111/j.1949-8594.1965.tb13550.x

Hollis, L. Y. (1965). A study to compare the effects of teaching first and second grade mathematics by the Cuisenaire-Gattegno method with a traditional method. *Schl. Sci. Math.* 65, 683–687.

Howard, C. F. (1957). British teachers' reactions to the Cuisenaire-Gattegno materials. *Arithmetic Teach.* 4, 191–195. doi: 10.5951/AT.4.5.0191

Huang, Y. (2019). *The effects of Cuisenaire rods on lower grade students' mathematical learning interests and learning achievements* (master's thesis). Huafan University, New Taipei City, Taiwan.

Jones, I., Bisson, M.-J., Gilmorea, C., and Inglisa, M. (2019). Measuring conceptual understanding in randomised controlled trials: Can comparative judgement help? *Brit. Educ. Res. J.* 45, 662–680. doi: 10.1002/berj.3519

Kaput, J. J. (1995a). "Overcoming physicality and the eternal present: cybernetic manipulatives," in *Exploiting Mental Imagery with Computers in Mathematics Education*, eds R. Sutherland and J. Mason (Berlin: Springer) 161–177. doi: 10.1007/978-3-642-57771-0_11

Kaput, J. J., and Blanton, M. L. (2000). *Algebraic Reasoning in the Context of Elementary Mathematics: Making It Implementable on a Massive Scale*. Available online at: https://eric.ed.gov/?id=ED441663

Kaput, J. J. (1995b). *Transforming Algebra from an Engine of Inequity to an Engine of Mathematical Power by "Algebrafying" the K-12 Curriculum*. NCTM.

Keagle, M. A., and Brummett, A. J. (1993). *Manipulative versus traditional teaching for mathematics concepts: Instruction-testing match* (Master's thesis). Ball State University, Muncie, IN, United States.

Kieran, C., Pang, J. S., Schifter, D., and Ng, S. F. (2016). *Early Algebra. Research into Its Nature, Its Learning, Its Teaching*. Cham: Springer. doi: 10.1007/978-3-319-32258-2

Kieran, C. (2018). "Conclusions and looking ahead," in *Teaching and Learning Algebraic Thinking with 5-to 12- Year-Olds, ICME-13 Monographs*, eds C. Kieran (Cham: Springer), 427–438. doi: 10.1007/978-3-319-68351-5

Kilpatrick, J., and Weaver, J. F. (1977). Place of William A. Brownell in mathematics education. *J. Res. Math. Educ.* 8, 382–384. doi: 10.5951/jresematheduc.8.5.0382

Lamon, W. E., and Scott, L. F. (1970). An investigation of structure in elementary school mathematics: isomorphism. *Educ. Stud. Math.* 3, 95–110.

Lin, H.-C. (2013). *The study of relationship between concepts of place value and academic achievement of the first and second graders in elementary school in Taoyuan county* (master's thesis). Chung Yuan University, Taoyuan City, Taiwan.

Lucow, W. H. (1962). Cuisenaire method compared with the current methods of teaching multiplication and division. Winnepeg, MB: Manitoba Teachers Society.

Marchese, C. (2009). *Representation and generalization in algebra learning of 8th grade students* (Ph.D. thesis). New Brunswick, NJ: Rutgers.

Mason, J. (2008). "Making use of children's powers to produce algebraic thinking," in *Algebra in the Early Grades*, eds J. J. Kaput, D. W. Carraher, and M. L. Blanton (Reston, VA: NCTM) 57–94.

Mason, J. (2010). Mathematics education: theory, practice and memories over 50 years. *Learn. Math.* 30, 3–9.

Matthews, P. G., and Fuchs, L. S. (2018). Keys to the gate? Equal sign knowledge at second grade predicts fourth-grade algebra competence. *Child Dev.* 91, e14–e28. doi: 10.1111/cdev.13144

McNeil, N. M., Fyfe, E. R., Petersen, L. A., Dunwiddie, A. E., and Brletic-Shipley, H. (2011). Benefits of practicing 4 = 2 + 2: nontraditional problem formats facilitate children's understanding of mathematical equivalence. *Child Dev.* 82, 620–1633. doi: 10.1111/j.1467-8624.2011.01622.x

Mulligan, J., and Mitchelmore, M. (2009). Awareness of pattern and structure in early mathematical development. *Math. Educ. Res. J.* 21, 33–49. doi: 10.1007/BF03217544

Nasca, D. (1966). Comparative merits of a manipulative approach to second grade arithmetic. *Arithmetic Teach.* 13, 221–226. doi: 10.5951/AT.13.3.0221

NCTM (2000). *Principles and Standards for School Mathematics*. National Council of Teachers of Mathematics.

Nemirovsky, R., and Sinclair, N. (2020). On the intertwined contributions of physical and digital tools for the teaching and learning of mathematics. *Digit. Exp. Math. Educ.* 6, 107–108. doi: 10.1007/s40751-020-00075-3

O'Donnell, J., Hall, C., and Page, R. (2006). *Discrete Mathematics Using a Computer*. London: Springer.

Passy, R. A. (1963a). The effect of the Cuisenaire materials on reasoning and computation. *Arithmetic Teach.* 10, 439–440. doi: 10.5951/AT.10.7.0439

Passy, R. A. (1963b). *How do Cuisenaire materials in a modified elementary mathematics program affect the mathematical reasoning and computational skill of third-grade children*? (Ph.D. thesis). New York University, New York, NY, United States.

Piaget, J., Henriques, G., and Ascher, E. (1992). *Morphisms and Categories: Comparing and Transforming*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Piaget, J., and Szeminska, A. (1952). *Genése du nombre chez l'enfant (The Child's Conception of Number)*. Transl. by C. Gattegno and F. M. Hodgson. Delachaux et Niestle (Abingdon: Routledge and Kegan Paul).

R Core Team (2020). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Radford, L. (2014). Towards an embodied, cultural, and material conception of mathematics cognition. *ZDM Math. Educ.* 46, 349–361. doi: 10.1007/s11858-014-0591-1

Radford, L. (2018). "The emergence of symbolic algebraic thinking in primary school," in *Teaching and Learning Algebraic Thinking with 5- to 12- Year-Olds*, ed C. Kieran (Cham: Springer), 3–25. doi: 10.1007/978-3-319-68351-5_1

Rasila, A., and Sangwin, C. (2016). "Development of stack assessments to underpin mastery learning," in *Proceedings of 13th International Congress on Mathematical Education* (Hamburg).

Rawlinson, R. W. (1965). *An Assessment of the Cuisenaire-Gattegno Approach to the Teaching of Number in the First Year at School*. Sydney, NSW: Australian Council for Educational Research.

Reimer, K., and Moyer, P. S. (2005). Third-graders learn about fractions using virtual manipulatives: a classroom study. *J. Comput. Math. Sci. Teach.* 24, 5–25.

Rich, L. W. (1972). *The effects of a manipulative instructional mode in teaching mathematics to selected 7th grade inner city students* (Ph.D. thesis). Temple University, Philadelphia, PA, United States.

Riley, R. D., Higgins, J. P. T., and Deeks, J. J. (2011). Interpretation of random effects meta-analyses. *Brit. Med. J.* 342, 964–967. doi: 10.1136/bmj.d549

Rittle-Johnson, B., Matthews, P. G., Taylor, R. S., and McEldoon, K. L. (2011). Assessing knowledge of mathematical equivalence: a construct-modeling approach. *J. Educ. Psychol.* 103, 85–104. doi: 10.1037/a0021334

Robinson, E. B. (1978). *The effects of a concrete manipulative on attitude toward mathematics and levels of achievement and retention of a mathematical concept among elementary students* (Ph.D. thesis). East Texas State University, Commerce, TX, United States.

Robinson, F. G. (1964). "A note on the quantity and quality of Canadian research on the Cuisenaire method," in *Canadian Experience with the Cuisenaire Method,* eds F. G. Robinson (Ottawa, ON: Canadian Council for Research in Education), 181–2.

Rodman, J. T. (1964). Equal time. *Arithmetic Teach.* 11, 342–343. doi: 10.5951/AT.11.5.0342

Romero, R. C. (1977). *Student achievement in a pilot Cureton reading, Cuisenaire mathematics program, and a bilingual program of an elementary school* (Ph.D. thesis). Northern Arizona University, Flagstaff, AZ, United States.

Sangwin, C. (2016). "How does CAS change mathematics?" in *International Congress on Mathematics Education* (Hamburg).

Sangwin, C. J. (2005). On building polynomials. *Math. Gazette* 89, 441–450. doi: 10.1017/S0025557200178295

Sangwin, C. J. (2015). An audited elementary algebra. *Math. Gazette* 99, 298–316. doi: 10.1017/mag.2015.38

Schliemann, A., Carraher, D., and Brizuela, B. (2007). *Bringing out the Algebraic Character of Arithmetic*. New Jersey, NJ: Lawrence Erlbaum Associates. doi: 10.4324/9780203827192

Schmittau, J., and Morris, A. (2004). The development of algebra in the elementary mathematics curriculum of V. V. Davydov. *Math. Educ.* 8, 60–87.

Seltman, M., and Seltman, P. (1985). *Piaget's Logic: A Critique of Genetic Epistemology*. London: George Allen and Unwin.

Sfard, A. (1995). The development of algebra: confronting historical and psychological perspectives. *J. Math. Behav.* 14, 15–39. doi: 10.1016/0732-3123(95)90022-5

Simsek, E., Jones, I., Hunter, J., and Xenidou-Dervou, I. (2021). Mathematical equivalence assessment: measurement invariance across six countries. *Stud. Educ. Eval.* 70, 101046. doi: 10.1016/j.stueduc.2021.101046

Steencken, E. P. (2001). *Tracing the growth in understanding of fraction ideas: a fourth grade case study* (Ph.D. thesis). New Brunswick, NJ: Rutgers.

Steiner, K. E. (1964). *A comparison of the Cuisenaire method of teaching arithmetic with a conventional method* (Master's thesis). North Texas State University, Denton, TX, United States.

Sterne, J. A. C., and Eggar, M. (2005). "Regression methods to detect publication and other bias in meta-analysis," in *Publication Bias in Meta-analysis: Prevention, Assessment and Adjustment, Chapter 6,* eds Editors H. R. Rothstein, A. J. Sutton and M. Borenstein (Hoboken, NJ: Wiley), 99–110. doi: 10.1002/0470870168.ch6

Sweeney, J. (1968). *A comparative study of the use of the Cuisenaire method and materials and a non-Cuisenaire approach and materials in a grade one mathematics program* (master's thesis). University of Toronto, Toronto, ON, Canada.

Thai, K.-P., Bang, H. J., and Li, L. (2021). Accelerating early math learning with research-based personalized learning games: a cluster randomized controlled trial. *J. Res. Educ. Effect.* 15, 1–24. doi: 10.1080/19345747.2021.1969710

Viechtbauer W. (2010). Conducting meta-analyses in R with the metafor package, *J. Stat. Softw.* 36, 1–48.

Viechtbauer, W. (2005). Bias and efficiency of meta-analytic variance estimators in the random-effects model. *J. Educ. Behav. Stat.* 30, 261–293. doi: 10.3102/10769986030003261

Viechtbauer, W., and Cheung, M. W.-L. (2010). Outlier and influence diagnostics for meta-analysis. *Res. Synthesis Methods* 1, 112–125. doi: 10.1002/jrsm.11

Viechtbauer, W. (2021). "Model checking in meta-analysis," in *Handbook of Meta-Analysis,* eds C. H. Schmid, T. Stijnen, and I. White (CRC Press) 219–254. doi: 10.1201/9781315119403-11

Wallace, P. (1974). *An investigation of the relative effects of teaching a mathematical concept via multisensory models in elementary school mathematics* (Ph.D. thesis). East Lancing, MI: Michigan State.

Woodcock, R. W., McGrew, K. S., and Mather, N. (2007). *Woodcock Johnson III Tests of Achievement*. Itasca, IL: Riverside Publishing.

Yankelewitz, D. (2009). *The development of mathematical reasoning in elementary school students' exploration of fraction ideas* (Ph.D. thesis). New Brunswick, NJ: Rutgers.

Young, R., and Messum, P. (2011). *How We Learn and How We Should Be Taught: An Introduction to the Work of Caleb Gattegno*. London: Duo Flumina.

Zazkis, R., and Mamolo, A. (2011). Reconceptualizing knowledge at the Mathematical Horizon. *Learn. Math.* 31, 8–13.

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 29 July 2022
DOI 10.3389/fpsyg.2022.795759

OPEN ACCESS

EDITED BY
George Waddell,
Royal College of Music, United Kingdom

REVIEWED BY
Laura Angioletti,
Catholic University of the Sacred Heart, Italy
Christos I. Ioannou,
CYENS Centre of Excellence, Cyprus

*CORRESPONDENCE
Niclas Wisén
niclas.wisen@ki.se

SPECIALTY SECTION
This article was submitted to Performance Science, a section of the journal Frontiers in Psychology

RECEIVED 15 October 2021
ACCEPTED 11 July 2022
PUBLISHED 29 July 2022

CITATION
Wisén N, Larsson G, Risling M and Arborelius U (2022) Is conduct after capture training sufficiently stressful? *Front. Psychol.* 13:795759. doi: 10.3389/fpsyg.2022.795759

Is conduct after capture training sufficiently stressful?

Niclas Wisén[1]*, Gerry Larsson[2], Mårten Risling[1] and Ulf Arborelius[1]

[1]Department of Experimental Traumatology, Institution of Neuroscience at Karolinska Institute, Stockholm, Sweden, [2]Department of Leadership and Command and Control, Swedish Defence University, Karlstad, Sweden

Conduct after capture (CAC) training is for personnel at risk of being captured. To be effective, it needs to be stressful. But how do we know if it is stressful enough? This study uses biomarkers and cognitive measures to evaluate CAC. Soldiers undergoing CAC were measured by the stress hormone cortisol from saliva samples at baseline and during training. The training consisted of being taken capture and put through a number of realistic and threatening scenarios, targeting survival strategies taught in the preceding week. Between scenarios, the trainees were held in a holding cell where they were monitored by a guard. The saliva samples were taken in conjunction with the scenarios. The whole training took place over a period of ~24h. Cognitive performance was measured at baseline and after training. Three groups took part Group A (n=20) was taken after 48h of intense tasks leaving them in a poor resting state. Group B (n=23) was well rested at CAC onset. Group C (n=10) was part of a survival, evasion, resistance, and escape (SERE) instructor course. The CAC training was the same for all groups. Group A exhibited a high increase in cortisol during CAC, compared to baseline levels were multiple times as high as "expected" values. Group B exhibited elevated levels slightly lower than those of group A, they also "dropped" to "normal" levels during the latter part of the exercise. Group C displayed the least increase with only slightly elevated levels. CAC training is stressful and cortisol levels were elevated enough to satisfy the prerequisite for effective stress inoculation. No cognitive performance drop could be identified; however, several participants "froze" during the exercise due to intensive stress.

KEYWORDS

SERE, conduct after capture, stress inoculation, military training, cognition, salivary cortisol

Introduction

Stress inoculation training (SIT) is a method developed in the 1980s as a form of cognitive behavioral therapy (CBT) by Donald Meichenbaum and colleagues (Meichenbaum and Novaco, 1985). The fundamental idea of SIT is to expose a person to a stressor while providing coping strategies and methods to handle the subsequent stress successfully. The experience of successfully managing a stressful situation should "inoculate" the trainee not only against specific stressors but also other similar stressors and

thus be generalizable. The idea of SIT has been transferred from the therapy environment to the military environment, where it has been used in several settings. For a comprehensive in-depth description of SIT's modern military use, see Robson and Manacapilli (2014). SIT is also used in areas in which performance under stress is key to survival, such as Conduct After Capture (CAC). The intense stress from a life-threatening situation will activate survival responses, often labeled as fight, flight, or freeze. The survival response will mainly benefit physical performance compared to cognitive performance concerning functions such as, problem-solving, decision-making, and making use of theoretical tactical models during the response. Providing the experience of overcoming stressors and successfully making use of cognitive strategies, will according to the SIT paradigm, increase self-efficacy when faced with similar challenges.

CAC training is a form of SIT that aims to replicate the intense stress of being captured, kidnapped, or exposed to a threatening interrogation. CAC training is provided not only to military personnel but also to journalists active in unstable parts of the world as well as to ship crews traveling in waters frequented by pirates. Real-life experience, e.g., the experience of individuals who have undergone CAC training and then been taken capture, has shown at least anecdotally that the training lessens stress and increases a sense of control. As stated by a journalist taken captive in Syria in 2013, "in the middle of the stress and fear, the training provided comfort, I recognized the situations and knew what to expect" (Helmertz, 2019).

Military CAC training with the Swedish Armed Forces is a part of Survival, Evasion, Resistance, and Escape training (SERE), SERE C training. SERE A and B are basic survival training with a focus on finding food, keeping the right body temperature, and signal for help, etc., it is a part of basic soldier training. SERE C represents the highest course level, aimed at high-risk personnel. There have been several studies on CAC, which have examined stress hormones, cognitive performance, and mood among a variety of physical measures (Lieberman et al., 2016; Suurd Ralph et al., 2017). The findings include significant effects on those measures as a result of the intense stress experienced during the exercises, concluding that SERE or CAC training is stressful. However, SERE or CAC training might share the same label between nations, but the components and setting might differ in ways that calls for evaluation of the intended effect. One cannot evaluate the effect in real-life settings, e.g., being captured, other than in the unfortunate events where an individual has been taken capture. The intended effect from the training, however, is to satisfy the requisites for SIT. That is, in order to gain the self-efficacy of managing a high-stress capture situation one must pass a similar challenge with a successful outcome. Thus, our research question was as follows: Is the CAC training sufficiently stressful? If the experienced stress is too low, the requisite for SIT is not met (Suurd Ralph et al., 2017).

SERE C at the Swedish Armed Forces Survival School is a 2-week course, with the CAC event occurring within those weeks. During the exercise, trainees are exposed to different scenarios and situations that place them in stressful scenarios (interrogations) that they face alone, termed "ploys." Between ploys, the trainees are placed in a "holding cell" (a large concrete room with no furniture), under the surveillance of a hostile guard.

CAC training has undergone continuous change over the years from a focus on physical stress (e.g., rough treatment, stress positions, and exposure to cold) to a more controlled, safer approach with more focus on psychological stress. This change warrants a structured evaluation to validate the effectiveness of the training. In this evaluation initiated by the SERE School, three consecutive groups (A, B, and C) undergoing CAC training, were examined with a focus on stress measured by cortisol and the effects of the exercise on cognition. The groups differed due to the natural selection of participants, in several ways. Two of the groups (A and B) were selected military staff that serve in high-risk position often in the air force, the third group (C) were a SERE instructor course including a full CAC training. In this study, we looked at the longitudinal data for each group firsthand. We did however compare the groups acknowledging the fact that they differed both in their demographics and the state they were in when entering the exercise. Taking that into account, we argue that the groups are similar enough to warrant a comparison on some of the identified confounders. Age is an example; Group A were younger than B and C, and most research on cortisol and ageing focus on the ageing adult (around 70 years). Research has shown that levels of cortisol are relatively stable in adulthood even though a small decline is observed during the early 20s to the 40s where it increases again (Moffat et al., 2020), making a comparison of cortisol between the groups that differ slightly in age valid. The groups also had different resting states when entering the training, sleep deprivation has been shown to affect cortisol increasing reactivity the following day (Hirotsu et al., 2015).

Since CAC is a resource-intensive training, it is important to evaluate whether the training creates the desired effects.

What is stressful enough during CAC training? Stress responses are psychological and physiological, the interaction is to some extent individual and based on previous experience and exposure to stressors as well as the appraisal of the situation (McEwen, 1998). A strong psychological stress evokes a significant physiological response including a release of stress-related hormones such as cortisol. If the training is perceived as stressful, we expect to see elevated levels that are so high that they cannot be result as over the day fluctuations to normal stress. Since a strong stress response can cause a temporary decline in cognitive performance (Lupien et al., 2007; Juster et al., 2010), instructors will adapt the stress during a ploy to a level that allows the participant to perform with sufficient function and utilize methods and strategies ending the ploy with a success. We therefore are not concerned with providing too much stress. As pointed out, changes have been made towards the exercise paradigm, that has lowered the amount of physical stressors and rough handling (due to the risk of injuries), leaving the instructor with nonphysical stressors such as shouting, isolation, false information, moral

dilemmas, etc. Since psychological stressors in a training environment can be mitigated by keeping a focus on that, it is just a training exercise, it could mitigate the stress to such an extent that it loses its function as a core component in stress mitigation training.

We choose cortisol as our objective stress measure; cortisol is frequently used as a biomarker of stress because it is easy to collect from saliva and since modern technology facilitates on-site analysis. Cortisol has a relatively stable diurnal curve over time. Its peak occurs in the morning, a measurement referred to as the cortisol awakening response (CAR; Wust et al., 2000; Hellhammer et al., 2009), and then slowly declines in the course of the day (Wust et al., 2000; Matsuda et al., 2012). However, acute stress can significantly increase cortisol as a reaction to the stressor (Kirschbaum and Hellhammer, 1989; Hellhammer et al., 2009; Bozovic et al., 2013). There is, however, a rather substantial Intra Individual Variation (IIV; Segerstrom et al., 2017), meaning that the smooth slope we infer from morning to evening is not so smooth after all. Over the course of the day, there are natural fluctuations due to everyday activity and the magnitude of the fluctuation is related to the perceived magnitude of the stressor (Schlotz et al., 2011). Nevertheless, there is a clear downward slope from the morning peak to the evening–night nadir. In addition to psychological events, variations in rest, food intake, nicotine, coffee, and other substances can affect daytime cortisol levels (Kudielka et al., 2009). There are limitations to using cortisol to measure minor effects of stress due to IIV. Previous studies (Morgan et al., 2000; Suurd Ralph et al., 2017) show that the effects of CAC go well beyond what could be expected from IIV.

SIT builds on successfully using learned strategies in a stressful situation where access to and execution of the strategies is perceived as hard or demanding due to loss of ability to use one's full cognitive potential. Therefore, it is relevant to assess how cognition is affected by the CAC training. Here, cognition is defined as a mental action of processing information in the brain with the goal of producing a favorable response. It is impacted by stress in several ways. It affects recall (memory) and problem solving as well as perception. Stress impacts cognition following the Yerkes–Dodson law (Yerkes and Dodson, 1908), that is that an optimal level of stress will increase performance, while too little or too much stress will result in less than optimal performance. Since it follows an inverted u-shape, it implies that we need some stress to recruit resources in order to perform at peak level. However if stress is too intense, we pass the peak and our performance declines with increased stress or load. Measuring cognitive performance can be done using Reaction Time measures which had been previously utilized in similar studies (Harris et al., 2005).

Question/Hypothesis

Based on the literature and the course setup, we designed an evaluation study with the following hypotheses:

> A/ CAC training will increase psychological stress, as measured by salivary cortisol, during the exercise compared to baseline measures.
>
> B/ CAC training will have a negative impact on cognitive performance, assessed directly after the exercise, compared to a baseline assessment.
>
> C/ Explorative: Will there be a difference between the groups.

Materials and methods

Design

Data were collected as an evaluation of the CAC part of the SERE C course. Using the design described below, the collected data were subsequently included in this study for scientific evaluation. There are several ploys during the CAC event. We chose to sample saliva from four evenly spaced ploys over the entire period of captivity which provided us a spread of data over time that were comparable to the baseline measures and normal fluctuations of cortisol levels during a 24 h cycle. Salivary cortisol was collected right after the ploys. The ploys had the same setup for all the trainees. However, the ploys unfold partly due to the interaction of the trainee. Therefore, they differ in intensity and length. The nature of the exercise, which is supposed to represent a relatively novel and unknown situation for the participants, prevents us from describing the setup in detail. It is a 2-week course; in the first week, they are taught methods and strategies to increase the likelihood of survival. How to create value and how to use information strategically to keep you an asset over time. During the CAC exposure exercise, its put to a test how effectively the participants can apply the methods and strategies taught the previous week. The exposure training covers 24h starting with participants taken capture, they are then put through the different plojs and supervised confinement both together with peers and in solitary confinement. The aim of the CAC part of the course is to put participants through a challenging setup where the theoretical foundation taught at week one is put into a practical test. As described unless a participant fails completely, they will be guided thorough the ploys in a way that lets them experience that they can perform under pressure.

Participants

The number of participants was as follows: Group A ($n = 20$), group B ($n = 23$), and group C ($n = 10$). In total, there were 53 participants in the study, (all participants of the course) and there was one female participant. The age distribution (years) for the groups was as follows: A (mean = 24.4, SD = 7.1), B (mean = 28.8, SD = 3.91), and C (mean = 28.9, SD = 4.2).

TABLE 1 Group specifics.

Group	*n*	Age mean	Branch	Specifics
A	20	24.4, SD = 7.1	Two Army 18 Air Force	Pre exhausted (sleep and food deprived) the 24 h before preceding capture.
B	23	28.8, SD = 4.0	23 Air Force	Well rested before capture
C	10	28.9, SD = 4.2	Five Army One Navy Four Air Force	Well rested before capture. Voluntary participation in course, and previous similar experience.
tot	53	27.1, SD = 5.8	One Navy Seven Army 45 Air Force	

The participants were from all branches of the armed forces: Air Force ($n = 45$), Army ($n = 7$), and Navy ($n = 1$). The participants in groups A and B underwent the training as part of their ordinary training (mandatory), while group C participants had applied for the longer SERE instructor course (voluntary). This study was initiated as a training evaluation study and participation was expected still they were verbally informed on that they were tested. Subsequent use of these data for a scientific approach has been subject to ethical evaluation by the Swedish Ethics Review Authority, (review nr: 2019-05361). The committee concluded the data do not meet the requirements to be subject for individual consent to be used, since no identification from data is possible.

Grouping

Trainees from three SERE C courses (groups A, B, and C) were all included in the CAC evaluation. Group A entered the CAC after a period of 48 h containing an "urban evasion" training where they are to avoid capture by instructors in a small Swedish town. The night before the CAC exercise they also were put through a heat chamber exercise covering several hours. They were taken capture when going back to have the next day off resulting in limited sleep and food intake. Group B was given the night before exercise off, ending at around 6 p.m. They were well rested and fed when entering the CAC exercise the following morning. Group C differed from the others in that its members were trainees undergoing an instructor course for the SERE A and B levels. The instructor course is longer than the SERE C course. Thus, the participants have an opportunity to get to know one another. Compared to the other groups, group C consisted of a wider array of individuals from the different branches of the armed forces (five Army, one Navy, four Air Force) than group A (two Army, 18 Air Force) and group B (23 Air Force). The previous experience with SERE, however, did not include CAC training. Group C also entered the CAC exercise well rested and fed. Table 1. Group specifics.

Procedure

The three groups A, B, and C were all subjected to baseline testing the week before the CAC exposure. The baseline tests were given on a "lecture" day starting at wake up collecting three saliva samples (at awakening and after 15 and 30 min) to obtain the CAR, and then at ~06:00 p.m. to measure the evening level. Drinking, food intake, nicotine use and teethbrushing were prohibited during the 30 min prior to saliva collection to avoid their affecting the saliva content. During CAC, saliva was collected after each of the 4 ploys selected for sampling. Since all participants were subjected to the same ploys in a consecutive manner, the sampling time varied within the range of each ploy (i.e., 1–2 h) and depending on participant performance. The saliva collected from the ploys are referred to as sample events 3–6. The Ploys were evenly distributed during the day staring at around 9 a.m. (after the capture and incarceration procedure). Last ploy were sampled around 8–10 p.m.

Two cognitive tests were given, one at the same lecture day as the cortisol sampling. Follow-up testing was performed right after the end of the exercise.

Variables

Stress

The main stressor that is applied through the training is the psychological stress provided by the CAC exercise. Since physical stressors are excluded (no stress positions, rough handling, cold or heat exposure, etc.,) all remaining stressors are psychological in nature. The training environment is designed to be realistic, that is in isolation, filled with uncomfortable smell and sounds, participants are also put into captive clothing's and from time to time they wear a hood covering their eyes. The handling from instructors is mainly shouting, threatening, degrading, or trying to play participants against one and other and to take away the feeling of being in control of the situation. There is a component of pass or fail that can be a source of stress; the participants are not aware that the instructors will adapt and guide them through the ploys since the idea is not to test the individuals but to have them experience that they made it even though it took some effort.

Initial "rest" status

The second impact factor was the initial resting state of the group when entering the CAC exercise. Group A were in a status of sleep and food deprivation. While Group B and C were well rested

TABLE 2 Between-group comparison one-way analysis of variance.

Variable	Group A ($n = 20$)		Group B ($n = 23$)		Group C ($n = 10$)		Variable Shapiro–Wilk	
	M	SD	*M*	SD	*M*	SD	Statistic	*p*
TIME 1								
Simple reaction time[a]	335.25	31.02	359.29	46.46	349.59	25.41	0.904	0.001
Choice reaction time[a]	707.30	87.96	773.49	129.61	724.31	95.21	0.919	0.002
Go or no go[a]	461.40	54.00	514.21	69.46	487.22	46.38	0.941	0.014
TIME 2								
Simple reaction time[a]	350.40	46.95	360.36	38.55	349.55	35.36	0.959	0.079
Choice reaction time[a]	753.95	143.44	759.80	89.48	740.15	65.92	0.963	0.115
Go or no go[a]	472.85	53.90	517.33	44.40	498.78	74.54	0.961	0.096

[a]Scores show ms.

and feed. The factor was not measured due to the training schedule. Group A were doing overnight training the preceding 48 h, while group B and C finished at noon the day before the CAC training.

Data collection procedures

Cognitive measurements

The participants were given a digital cognitive test based on reaction time (RT) measurements. They were given the test at twice once at baseline and once directly after the exercise. The cognitive test battery consisted of three RT-based cognitive tests as follows. Simple reaction time (*SRT*) is defined as the time required to elicit a simple defined reaction to a stimulus, often using a visual stimulus with a motoric response. SRT is assessed by touching a dot (stimuli) as fast as possible on a screen when the stimulus appears, the test covers 20 stimuli events (dots) with randomly varied intervals between. Choice reaction time (*CRT*) adds stimuli identification and response selection to the SRT paradigm, compared to the SRT, the test has four independent symbols. When they appear on the screen, the respondent touches the "button" with the corresponding symbol on it. The response buttons have two symbols each and are situated below the area where the stimuli appear. As with SRT, the test has 20 events with varied time intervals between. The go or no go (*GNG*) test paradigms present two different stimuli appearing in a grid with six possible places for the stimuli to appear. When a blue dot appears in one position in the grid the correct response is to refrain (inhibit) a response, while a red dot is correctly responded to by touching the "shoot" button, the test covers 10 response stimuli and 10 inhibit stimuli randomly distributed over the test (Littman and Takács, 2017).

RT-test paradigms are a well-established way to measure information-processing performance (Woods et al., 2015; Burke et al., 2017), and SRT has been shown to be a valuable test paradigm in measuring stress-induced deterioration (Harris et al., 2005). Our version was given on an Android-based tablet, using a program developed inhouse based on well-established test protocols and paradigms, for bringing test availability to the field.

Biological measurements

Cortisol was used as a biomarker for stress. The saliva was sampled using Salivette™ collectors. Cortisol analysis was performed using mobile salivary cortisol assays (I-calQ, LLC; Scottsdale, Arizona, United States). The I-calQ is developed for field use (medical), which makes it possible to test the collected saliva onsite with no storage or delays still they were kept frigerated during the exercise before analyzed.

It uses the immunoassay test strips and image analysis algorithm to analyze the saliva. The cortisol assay utilizes affinity chromatography. That is Antibodies developed with a high affinity for particles of cortisol, These antibodies adhered to cortisol produce a visible signal. The intensity of this signal correlates with the amount of cortisol present in the saliva sample, which is also correlated to the blood concentration of cortisol.

Statistics

Cognitive measures were analyzed using a MANOVA repeated-measure design covering between-group and within-group baseline-post-measurements, comparisons. The use of a MANOVA was motivated by the assumption that all cognitive subtests measure an underlying function that could indicate an overall effect.

Cortisol measures were compared based on group means and complemented with analysis of AUC area under the curve.

Statistics were analyzed using, SPSS version 26 and R.

Results

Cognitive measures

Table 2 presents means and standard deviations of the three groups on both measurement occasions. The Shapiro–Wilk statistic shows that the response distributions of all three reaction time tests were statistically non-normal on the first assessment. On the second measurement occasion, the three tests did not deviate significantly

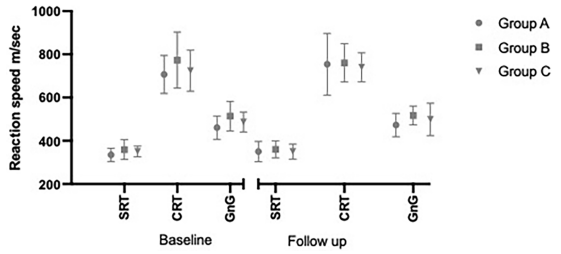


FIGURE 1
Cognitive test for all groups A–C at baseline and follow-up; the vertical bars represent one standard deviation. SRT, simple reaction time; CRT, choice reaction time; GnG, go or no go.

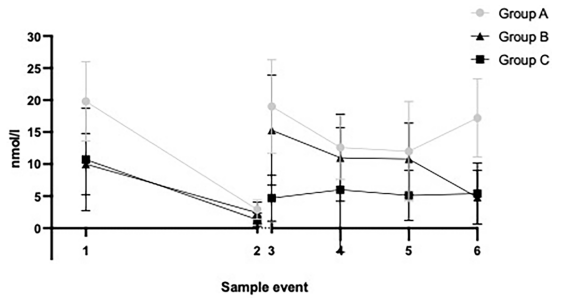


FIGURE 2
All groups over all events: 1–2 baseline, 3–6 during CAC (vertical bars one SD).

from normal. This was confirmed by the Kurtosis values. On the first measurement occasion, they ranged from 1.034 to 1.404. On the second assessment, they ranged between 0.511 and 0.594.

A MANOVA repeated-measures design was used to test within-and between-group differences on the cognitive tests. Since there were only two levels (time and group), the assumption of sphericity was met and the Mauchly's test was not applicable (Field, 2000). The Box's *M*-test score was 62.20, $F = 1.156$, (42, 2839,38), $p = 0.228$, indicating that the observed covariance matrices of the three reaction time tests were equal across the three groups. Mahalanobis' distance showed one extreme value which caused the critical value for six dependent variables (the three cognitive tests pre-and post) to slightly exceed the maximum limit (22.72 where the limit is 22.46). Beginning with within-subjects effects across time, the multivariate tests (Pillai's trace) did not show any significant differences. Turning to the between-subjects effects, also here no significant differences emerged the graphical distrubution is shown in Figure 1.

Biological measures

The cortisol values are presented in Figure 2 using the mean nmol/l. The sample events refer to the time the samples were collected. Sample event 1 is Baseline CAC (the mean of the three awakening measures), and Sample event 2 is the baseline pm measure (covering ~12 h during the day). Sample events 3–6 are the measures taken after each ploy during CAC. The diurnal fluctuation of cortisol with its peak in the morning and the nadir in the evening gives an estimate of the slope over the day. The graph in the figure shows that the baseline levels follow that natural decrease during the day. Sample events 3–6 show a different path, with levels being elevated throughout the exercise. Cortisol levels were also compared using the area under the curve (AUC), it also shows a significant elevation during the CAC training. AUC was calculated for individuals with a full dataset (all measures), 8 individuals had non-complete tests due to not sufficient saliva when sampling. Therefore, AUC results are based on 45 individuals. The mean AUC for each group at baseline, training, and the absolute difference is presented in Table 3. An overall repeated-measures ANOVA showed significant differences between groups, time, and group:time. Pairwise comparisons over time (within each group) show that groups A and B have a significant difference in AUC, while group C does not Tables 4, 5.

TABLE 3 Area under the curve (AUC) for all groups A, B and C at baseline and during Conduct after training (CAC).

Group	Variable	*n*	Mean	SD
A	AUC at baseline	19	34.04	10.43
B	AUC at baseline	21	18.59	7.42
C	AUC at baseline	10	17.94	12.96
A	AUC during CAC	19	42.73	13.03
B	AUC during CAC	23	30.43	13.32
C	AUC during CAC	10	15.67	9.70
A	Absolute difference in AUC	19	8.69	13.00
B	Absolute difference in AUC	20	13.36	13.78
C	Absolute difference in AUC	10	−2.27	13.06

The last three lines show the difference between the two times.

TABLE 4 Overall repeated measures ANOVA.

Effect	*dfn*	*dfd*	*F*	*p*
group	2	46	22.616	0.000
time	1	46	10.848	0.002
group:time	2	46	4.587	0.015

Discussion

The results support the main hypothesis that cortisol levels increase during the exercise with the exception for group C. The cortisol taken at baseline affords an estimate of a normal decline of levels during a day without intense stress (Cochrane et al., 2014). Based on the baseline measures, we can compare the assumed "slope" with the values from the exercise, as presented in the graphs. The soldiers in all groups exhibited increased levels over the entire period of captivity. This phenomenon might not be as obvious in the morning when the

TABLE 5 Pairwise comparisons between time points (within each group).

omg	Time 1	Time 2	n1	n2	Statistic	*df*	*p*
A	Baseline	Post	19	19	−2.91	18	0.0090
B	Baseline	Post	20	20	−4.34	19	0.0004
C	Baseline	Post	10	10	0.55	9	0.5960

system is naturally saturated. However, comparing the evening (baseline late sample) and the third test event (taken during the afternoon/evening ploy), we found elevated levels during CAC 4–5 to be times higher than baseline levels. The results compared between groups both using means and AUC indicate that there were effects on cortisol levels from preceding stress, such as sleep deprivation and low food intake. Group C however did not show the same profile, they were more prepared, had a more established team feeling, and were voluntary participants, all factors that could possibly mitigate the stress response during CAC. However, the first measure (CAR) was so low that one can suspect an possible artifact in sampling due to some error in reading, testing equipment, or other factors that influenced all the samples taken at that time.

The cognitive tests performed before and after the exercise indicated no change in cognitive processing speed. Although the response distributions on the three reaction time tests deviated from normality on the first test occasion, the deviance was limited and they met the normality requirement on the second assessment. Thus, we regard the use of the MANOVA repeated-measures design as legitimate. RT tests are commonly used to measure cognitive performance. However, the ability measured requires a significant stress to show any decline in performance. This is a challange for military perfromance research in general. Its hard to find test that has ecological validity and that test clos to what would be an actual responce in a real-life situation. During the exercise, the participants were observed to have difficulties accessing the methods and strategies they were taught the preceding week. At an extreme, one soldier became "stuck" for over 45 min in a ploy that usually required <10 min to pass. He struggled to handle the stress and find the correct responses, even with instructors providing "hints" in their role-play. One plausible explanation is that the observed decrease in cognitive performance only occurs while the participant is exposed to the stressor, and the recovery time for cognitive function is immediate. This outcome calls for improved understanding and future testing. Cognitive function is a broad area, and when it appears to fail in a situation that makes use of taught models (i.e., recall and activation), we must determine which properties are responsible for the "freeze" or lack of access to cognitive resources. The ambition to bring participants through CAC training with a feeling of success is the reason for instructors "hinting" or leading them through the ploys, one could argue that it might affect the stress response and following cortisol sample. The "hinting" or "support" is however not obvious and the instructors are trained at keeping their hostile approach even when offering a way out by presenting obvious use of the tools they have been taught.

Since we did not have performance measures for each ploy, we cannot compare ploy success with cortisol levels. Could individuals with higher cortisol response be more prone to perform worse than those who had a lower response, an indication of less perceived stress? This question can be addressed in future research. Further, there are other limitations to this study, since it was performed on already planned groups and curriculums the only thing we could affect was the resting state at onset. Therefore, there is no random assignment between groups, and Group C differs in many aspects of the group composition.

Conclusion

We hypothesized that the CAC exercise would increase stress to such an extent that it could be measured in salivary cortisol, there would be a cognitive performance drop directly after the exercise, and the magnitude of these effects would be affected by the rest and food status of the participants at the time of exercise onset. The results supported the cortisol hypothesis but not the cognitive performance hypothesis. Saliva was easy to collect with little impact on the exercise. However, cognitive ability in the form of RT tests cannot be used during the exercise without interrupting and exerting a negative effect on exercise realism. What we observed was that the soldiers who were sleep- and food-deprived had the highest levels of cortisol reaction, indicating a higher stress response. Therefore, pre-exhaustion of participants might be a way to amplify the intended stress effect on participants with less intense stress stimuli. There is, however, a risk of less learning when sleep deprived (Pierard et al., 2004). As noted, the fundamental idea of SIT and CAC is to create a stress exposure in response to which learned skillsets can be successfully applied. Instructor reports and ploy observations revealed that there are temporary cognitive limitations due to stress. Ploys are fairly standardized, and it should be possible to find ways to assess cognitive performance during each ploy. Such a design could possibly identify which cognitive components are most affected by stress. And if there are individuals who are more resilient or susceptible to CAC stressors. This question warrants further research and could be helpful in the further development of CAC training. Studies such as this one are relevant in that we must evaluate and validate training paradigms to develop them further. Operational demands and training regulations might have an impact on their intended effects (regulations and limitations). Therefore, because such training comes at a great cost for the organization providing it and for the participant, constant training evaluation is required.

Data availability statement

The data that support the findings of this study are openly available in Zendo.org at: https://doi.org/10.5281/zenodo.4543557.

Ethics statement

Ethical review and approval were not required for this study in accordance with the national legislation and the institutional requirements. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

Author contributions

NW conducted the study, analyzed the results, and wrote the paper. MR and UA provided the resources and facilitated the study. GL supervised and helped with analysis aside from providing continuous feedback thought the writing process. All authors contributed to the article and approved the submitted version.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Bozovic, D., Racic, M., and Ivkovic, N. (2013). Salivary cortisol levels as a biological marker of stress reaction. *Med. Arch.* 67, 374–377. doi: 10.5455/medarh.2013.67.374-377

Burke, D., Linder, S., Hirsch, J., Dey, T., Kana, D., Ringenbach, S., et al. (2017). Characterizing information processing with a mobile device: measurement of simple and choice reaction time. *Assessment* 24, 885–895. doi: 10.1177/1073191116633752

Cochrane, K. C., Coburn, J. W., Brown, L. E., and Judelson, D. A. (2014). Effects of diverting activity on strength, electromyographic, and mechanomyographic signals. *J. Strength Cond. Res.* 28, 1203–1211. doi: 10.1519/jsc.0000000000000378

Field, A. (2000). *Discovering statistics using IBM SPSS statistics*. SAGE.

Harris, W. C., Hancock, P. A., and Harris, S. C. (2005). Information processing changes following extended stress. *Mil. Psychol.* 17, 115–128. doi: 10.1207/s15327876mp1702_4

Hellhammer, D. H., Wust, S., and Kudielka, B. M. (2009). Salivary cortisol as a biomarker in stress research. *Psychoneuroendocrinology* 34, 163–171. doi: 10.1016/j.psyneuen.2008.10.026

Helmertz, F. (2019). Kunskap som räddar liv. *Försvarsmaktens Forum* 1:3.

Hirotsu, C., Tufik, S., and Andersen, M. L. (2015). Interactions between sleep, stress, and metabolism: From physiological to pathological conditions. *Sleep Sci.* 8, 143–152. doi: 10.1016/j.slsci.2015.09.002

Juster, R. P., McEwen, B. S., and Lupien, S. J. (2010). Allostatic load biomarkers of chronic stress and impact on health and cognition. *Neurosci. Biobehav. Rev.* 35, 2–16. doi: 10.1016/j.neubiorev.2009.10.002

Kirschbaum, C., and Hellhammer, D. H. (1989). Salivary cortisol in psychobiological research: An overview. *Neuropsychobiology* 22, 150–169. doi: 10.1159/000118611

Kudielka, B. M., Hellhammer, D. H., and Wust, S. (2009). Why do we respond so differently? Reviewing determinants of human salivary cortisol responses to challenge. *Psychoneuroendocrinology* 34, 2–18. doi: 10.1016/j.psyneuen.2008.10.004

Lieberman, H. R., Farina, E. K., Caldwell, J., Williams, K. W., Thompson, L. A., Niro, P. J., et al. (2016). Cognitive function, stress hormones, heart rate and nutritional status during simulated captivity in military survival training. *Physiol. Behav.* 165, 86–97. doi: 10.1016/j.physbeh.2016.06.037

Littman, R., and Takács, Á. (2017). Do all inhibitions act alike? A study of go/no-go and stop-signal paradigms. *PLoS One* 12:e0186774. doi: 10.1371/journal.pone.0186774

Lupien, S. J., Maheu, F., Tu, M., Fiocco, A., and Schramek, T. E. (2007). The effects of stress and stress hormones on human cognition: implications for the field of brain and cognition. *Brain Cogn.* 65, 209–237. doi: 10.1016/j.bandc.2007.02.007

Matsuda, S., Yamaguchi, T., Okada, K., Gotouda, A., and Mikami, S. (2012). Day-to-day variations in salivary cortisol measurements. *J. Prosthodont. Res.* 56, 37–41. doi: 10.1016/j.jpor.2011.04.004

McEwen, B. S. (1998). Protective and damaging effects of stress mediators. *N. Engl. J. Med.* 338, 171–179. doi: 10.1056/nejm199801153380307

Meichenbaum, D., and Novaco, R. (1985). Stress inoculation: a preventative approach. *Issues Ment. Health Nurs.* 7, 419–435. doi: 10.3109/01612848509009464

Moffat, S. D., An, Y., Resnick, S. M., Diamond, M. P., and Ferrucci, L. (2020). Longitudinal change in cortisol levels across the adult life span. *J. Gerontol. A Biol. Sci. Med. Sci.* 75, 394–400. doi: 10.1093/gerona/gly279

Morgan, C. A. 3rd, Wang, S., Mason, J., Southwick, S. M., Fox, P., Hazlett, G., et al. (2000). Hormone profiles in humans experiencing military survival training. *Biol. Psychiatry* 47, 891–901. doi: 10.1016/S0006-3223(99)00307-8

Pierard, C., Beracochea, D., Peres, M., Jouanin, J. C., Liscia, P., Satabin, P., et al. (2004). Declarative memory impairments following a military combat course: parallel neuropsychological and biochemical investigations. *Neuropsychobiology* 49, 210–217. doi: 10.1159/000077369

Robson, S., and Manacapilli, T. (2014). *Enhancing Performance Under Stress: Stress Inoculation Training for Battlefield Airmen*. Santa Monica: RAND Corporation, Air Force.

Schlotz, W., Hammerfald, K., Ehlert, U., and Gaab, J. (2011). Individual differences in the cortisol response to stress in young healthy men: testing the roles of perceived stress reactivity and threat appraisal using multiphase latent growth curve modeling. *Biol. Psychol.* 87, 257–264. doi: 10.1016/j.biopsycho.2011.03.005

Segerstrom, S. C., Sephton, S. E., and Westgate, P. M. (2017). Intraindividual variability in cortisol: approaches, illustrations, and recommendations. *Psychoneuroendocrinology* 78, 114–124. doi: 10.1016/j.psyneuen.2017.01.026

Suurd Ralph, C., Vartanian, O., Lieberman, H. R., Morgan, C. A. 3rd, and Cheung, B. (2017). The effects of captivity survival training on mood, dissociation, PTSD symptoms, cognitive performance and stress hormones. *Int. J. Psychophysiol.* 117, 37–47. doi: 10.1016/j.ijpsycho.2017.04.002

Woods, D. L., Wyma, J. M., Yund, E. W., Herron, T. J., and Reed, B. (2015). Factors influencing the latency of simple reaction time. *Front. Hum. Neurosci.* 9:131. doi: 10.3389/fnhum.2015.00131

Wust, S., Wolf, J., Hellhammer, D. H., Federenko, I., Schommer, N., and Kirschbaum, C. (2000). The cortisol awakening response – normal values and confounds. *Noise Health* 2, 79–88.

Yerkes, R. M., and Dodson, J. D. (1908). The relation of strength of stimulus to rapidity of habit-formation. *J. Compa. Neurol. Psychol.* 18, 459–482. doi: 10.1002/cne.920180503

frontiers | Frontiers in Education

TYPE Original Research
PUBLISHED 05 August 2022
DOI 10.3389/feduc.2022.910039

OPEN ACCESS

EDITED BY
George Waddell,
Royal College of Music,
United Kingdom

REVIEWED BY
John Hattie,
The University of Melbourne, Australia
Chia-Lin Tsai,
University of Northern Colorado,
United States

*CORRESPONDENCE
Belén Gutiérrez-de-Rozas
bgutierrezderozas@edu.uned.es

SPECIALTY SECTION
This article was submitted to Assessment, Testing and Applied Measurement, a section of the journal Frontiers in Education

RECEIVED 31 March 2022
ACCEPTED 11 July 2022
PUBLISHED 05 August 2022

CITATION
Gutiérrez-de-Rozas B, López-Martín E and Carpintero Molina E (2022) Defining the profile of students with low academic achievement: A cross-country analysis through PISA 2018 data.
Front. Educ. 7:910039.
doi: 10.3389/feduc.2022.910039

Defining the profile of students with low academic achievement: A cross-country analysis through PISA 2018 data

Belén Gutiérrez-de-Rozas[1]*, Esther López-Martín[1] and Elvira Carpintero Molina[2]

[1]Department of Methods of Research and Diagnosis in Education II, Universidad Nacional de Educación a Distancia (UNED), Madrid, Spain, [2]Department of Research and Psychology in Education, Universidad Complutense de Madrid, Madrid, Spain

The explanation of underachievement and the search for its associated factors have been of constant interest in educational research. In this regard, the number of variables that have been involved in its description and explanation has increased over the years, as has the number of studies at an international level on this topic. Although much research has focused on identifying the personal, family, and school aspects that exert the greatest influence on students' low academic performance, the literature shows the need to study the differential effects of said variables according to the countries in which the studies are conducted. The objective of this article is therefore to analyse cross-national differences in the effect of personal, family, and school characteristics on students' academic underachievement based on data derived from the Programme for International Student Assessment (PISA) 2018. Furthermore, it aims to identify the profile that characterises students with the lowest academic performance and to estimate the importance of the selected variables in explaining low achievement across countries. To reach these goals, the multivariate technique of decision trees through the binary CART (Classification and Regression Trees) algorithm was used, allowing the estimation of both a global model and nine specific models for each of the selected countries. The results show that, despite slight differences between the countries analysed, the variables that define the general profile of students with the lowest achievement and which have shown the strongest predictive capacity for low performance are mainly linked to the students themselves. These variables are followed in importance by family aspects, which present great differences between the territories that compose the sample. Finally, teacher and school variables have shown to have a low explanatory capacity in this study. It can therefore be concluded that, although personal characteristics continue to be those that best explain academic performance, a series of contextual

variables, especially related to families, appear to influence academic achievement differentially and may even hide or cancel out certain personal characteristics.

KEYWORDS

academic achievement, low achievement, PISA, cross-country analysis, decision trees, CART

Introduction

A reduction in low academic achievement, related to the non-attainment of learning goals for a student's level, age, or ability (Lamas, 2015), is one of the main objectives of current education systems. However, there is a clear lack of agreement when it comes to establishing the most appropriate standards or criteria for its definition (Gorard and Smith, 2003).

These standards may refer to the students' performance, to the performance of the group they belong to, or even to previously established external criteria (Gutiérrez-de-Rozas and López-Martín, 2020). For this reason, in a particular situation of low performance, a student can present insufficient attainment—by not achieving the educational objectives established for all the students—or an unsatisfactory performance—by performing below what could be expected based on his or her abilities (Jiménez Fernández, 2010). Hence, a situation of low academic performance may or may not exist depending on the standard used.

The International Association for the Evaluation of Educational Achievement (IEA) and the Organisation for Economic Co-operation and Development (OECD) apply standards based on external criteria in their assessments—that is, Trends in International Mathematics and Science Study (TIMSS) and Progress in International Reading Literacy Study (PIRLS) by IEA, or the Programme for International Student Assessment (PISA) by OECD. In these assessments, students are considered to be low-performing students if they are placed at the Low International Benchmark (Mullis et al., 2020) in TIMSS and PIRLS, or below Level 2 in PISA (Organisation for Economic Co-operation and Development [OECD], 2019a). Despite the utility of these standards for making international comparisons in the level of academic achievement between different countries, they do not capture the variability existing within countries. In this regard, results from the PISA 2018 assessment showed that around 71.8% of students in the Philippines were low-performing students in the three areas considered, while only 1.1% of the students were low-performing students in Beijing, Shanghai, Jiangsu, and Zhejiang (Organisation for Economic Co-operation and Development [OECD], 2019b). These unequal results may hide, among other things, different socio-cultural realities. Thus, being placed at one level or another has different implications in each territory.

Regardless of the contextualisation of student underachievement and the way of evidencing it, there is no doubt that knowing the aspects that facilitate or hinder academic performance is key to providing an adequate response to the educational needs of students. For this reason, numerous empirical studies have focused on identifying and analysing the predictive capacity of the conditioning factors of academic performance (Kornilova et al., 2009).

Within these factors, and despite the interrelationship among the variables that influence learning (Bhowmik, 2019; Akbas-Yesilyurt et al., 2020), since the past century, academic literature has been highlighting the strong influence of students' personal characteristics, together with other contextual aspects, on their educational outcomes. As proof of this, in the review conducted by Sipe and Curlette (1997), student characteristics had the largest effect sizes on academic achievement, followed by school variables and, finally, family aspects. Subsequently, Hattie (2003) showed that, when the interactions between variables were ignored, student characteristics predicted 50% of performance, while teacher characteristics explained 30%. The author attributed much smaller influences to school, peers, school leaders, and family characteristics (between 5 and 10%). In line with these results, the most recent meta-analytic evidence shows the effects of some specific personal aspects, such as the use of self-regulated learning strategies (Ergen and Kanadli, 2017), intelligence (Zaboski et al., 2018), or some personality types (Poropat, 2009), on academic performance. Therefore, there is ample scientific evidence, generated since the last century, for the differential influence of personal, family, and school variables on students' academic performance.

In this sense, Hattie's (2009, 2017) work should be highlighted as one of the most important international review studies in the field, since this author identified the influence of personal, family, and school variables on student academic performance by compiling the existing meta-analytical evidence. His research is of particular interest due to the vast amount of evidence that it summarises and also for its systematicity, as the author classifies these conditioning factors of academic performance into 22 categories and 66 subcategories. The results of his research show that previous

high academic performance and self-efficacy are the personal variables that most positively influence academic achievement. On the contrary, some personal factors, such as boredom, depression, minority language use, superficial motivation, sleep problems, attention deficit hyperactivity disorder, or hearing difficulties showed the strongest negative effects. Among the family variables, the author demonstrated the positive influence of a favourable home environment and high socio-economic status and highlighted the negative effects of corporal punishment, excessive television viewing, or being a beneficiary of welfare policies. Finally, among school and teacher variables, this author found that teacher efficacy had the strongest positive influence, while student suspension, excessively long summer holidays, or changes of school by students were the aspects with the most negative influence on academic performance.

However, despite these general findings, the literature warns the differential effects of conditioning factors on academic performance depending on the countries in which the studies are conducted. For example, the study by Ghasemi and Burley (2019) revealed the existence of differences in the predictive capacity of gender in mathematics in the countries analysed. Also, Ning et al. (2015), using PISA 2009 data, showed that the influence of school disciplinary climate on students' academic performance presented cross-national differences. This is to be expected given that each territory has its own socio-economic, cultural, political, and educational characteristics; that the aspects that condition academic achievement are interrelated (Bhowmik, 2019; Akbas-Yesilyurt et al., 2020); and that, in accordance with the ecological systems theory, inhabitants are influenced by the countries they live in (Hampden-Thompson et al., 2013). It can, therefore, be deduced that any macro-level differential aspects between countries may affect the characteristics of the students, the education given by families, and the education provided in schools.

Programme for International Student Assessment: Assessment of competences and associated factors

The Programme for International Student Assessment (PISA), aims to evaluate the extent to which students in the participating countries have acquired the knowledge and skills that are required to fully participate in today's societies by the end of compulsory education (Organisation for Economic Co-operation and Development [OECD], 2019c).

This assessment analyses students' proficiency in science, mathematics, and reading—the 2018 edition also includes global competence—through a series of tests that provide an updated and comparative overview of students' academic performance at the age of 15 years. Said performance does not only refer to the level of knowledge acquired in the areas assessed but also to the degree of skills and competence development in these domains. In each PISA edition, the OECD focuses its analysis and conclusions on one of the skills assessed, thus establishing it as the main domain. In the 2018 edition, as was the case in 2000 and 2009, the focus was placed on reading literacy—understood as "students' ability to understand, use, evaluate, reflect on and engage with text to achieve their purposes" (Organisation for Economic Co-operation and Development [OECD], 2019c, p. 15).

Through the PISA assessments, the OECD aims not only to provide countries with information on the performance of adolescents in their education systems but also to enable them to understand the results obtained by students in other participating countries and to analyse and compare educational policies (Organisation for Economic Co-operation and Development [OECD], 2019a). Therefore, context questionnaires are applied as a supplement to achievement tests to identify the characteristics of education systems, interpret the results obtained, and understand the factors that are linked to success or failure from both a national and a comparative perspective (López-Martín et al., 2018).

These questionnaires collect contextual data of students—including personal, family, and school aspects—but only a small part of the contextual information is provided by teachers and families. This aspect deserves special consideration, as students' perceptions often explain variation in learning outcomes beyond what could be attributed to background characteristics themselves (Van Petegem et al., 2007).

Another issue that also deserves consideration is that the OECD has not only added new items to the contextual questionnaires over the successive editions of PISA but has also developed and implemented new full questionnaires, such as the ICT familiarity questionnaire, the educational career questionnaire, the financial literacy questionnaire, or the well-being questionnaire.

Regarding the topic of student well-being, defined as "the psychological, cognitive, social and physical functioning and capabilities that students need to live a happy and fulfilling life" (Organisation for Economic Co-operation and Development [OECD], 2017, p. 35), it should be mentioned that, while information on this construct was collected through certain items of the student questionnaire in previous PISA editions, the specific well-being questionnaire was applied for the first time in 2018. Thus, the importance given to this construct by the OECD is in line with the findings of current empirical evidence, which is highlighting the prominent role of adolescent well-being in both positive adolescent development and success in learning processes (Holzer et al., 2021).

Present study

For all the above, this study aims to analyse cross-national differences in the effect of personal, family, and school characteristics on students' academic underachievement based on the data derived from the PISA 2018 assessment. Also, this article aims to identify the profile that characterises students with the lowest academic performance and to estimate the importance of the selected variables in the explanation of low performance across countries.

Therefore, this research seeks to help eliminate existing knowledge gaps relating to the differential influence of personal, family, and school variables on students' low academic performance across countries. Thus, the present study goes beyond the conventional research approach into the conditioning factors of academic performance and, more specifically, academic underachievement, in which the particularities of each territory are usually not considered.

For this purpose, the multivariate decision tree technique, through the binary CART (Classification and Regression Trees) algorithm (Breiman et al., 1984), is used. This technique is considered to be particularly suitable for yielding insights into the research question posed because, as Razi and Athappilly (2005) state, being a non-parametric procedure that allows the prediction of a continuous dependent variable from categorical independent variables, it fits the data perfectly. Also, as the cited authors affirm, CART models provide better predictions than regression models when the predictors are binary or categorical and the dependent variable is continuous. In addition, this model allows the creation of subsets of homogeneous data for the dependent variable, and calculation of the relative importance of each of the independent variables in explaining said dependent variable. Moreover, this technique allows a more thorough study of the variables that influence low performance not only globally but also comparatively across countries. Finally, it is noteworthy that several studies have already used this technique satisfactorily to analyse PISA data (Asensio Muñoz et al., 2018; López-Martín et al., 2018; Arroyo Resino et al., 2019; She et al., 2019). For all these aspects, the multivariate decision tree technique, through the binary CART algorithm, is used here to achieve the objectives proposed in this article.

After the above introduction, the rest of this article is structured as follows: first, the method is described. Then, the profiles of students with the lowest academic performance are presented together with the standardised importance of the analysed variables in explaining low performance in the selected countries. The article concludes with a discussion of the main results.

Materials and methods

Population and sample

The study population was composed of 15-year-old students from the countries that completed all the student context questionnaires in PISA 2018. After excluding from the selection process all the territories that did not apply all context questionnaires to their students, the final selection was based on nine countries. The final sample consisted of 97,878 students from Bulgaria (5.4%), Georgia (5.7%), Hong Kong SAR (China) (6.2%), Ireland (5.7%), Mexico (7.5%), Panama (6.4%), Serbia (6.8%), Spain (36.7%), and the United Arab Emirates (19.7%). Therefore, information was available from Europe, Asia, and Latin America. The main socio-demographic, political, and economic characteristics of the selected countries are described in **Appendix A**.

As reflected in **Table 1**, the sample was weighted using the normalised weight variable SENWT—when analysing the overall information from the set of countries considered—or the student sampling weight W_FSTUWT—when analysing the data individually for each of the countries (Organisation for Economic Co-operation and Development [OECD], n.d.). In this regard, it should be clarified that the SENWT variable assigns the same value to the samples of each country to ensure an equal contribution to the analysis, while the W_FSTUWT variable adjusts the samples to the population size of each country so that each contribution depends on population size.

Materials

Information derived from achievement tests and context questionnaires administered to students in the PISA 2018 assessment was selected. Reliability and validity evidence of the

TABLE 1 Sample description.

Country	Number of students		
	Unweighted	Weighted (W_FSTUWT)	Weighted (SENWT)
Bulgaria	5,294	47,851	5,000
Georgia	5,572	38,489	5,000
Hong Kong SAR (China)	6,037	51,101	5,000
Ireland	5,577	59,639	5,000
Mexico	7,299	1,480,904	5,000
Panama	6,270	3,854	5,000
Serbia	6,609	61,895	5,000
Spain	35,943	416,703	5,000
United Arab Emirates	19,277	54,403	5,000
Total	97,878	2,249,526	45,000

scales can be found in the PISA 2018 Technical Report,[1] in which information about the sampling procedure, the questions included in each questionnaire, the sample items, and the response scale are provided.

As the dependent variable of this study, low academic performance in reading, which was the core subject in the PISA 2018 assessment, was considered. Taking into account that one of the main features of the PISA assessment design is that it reports students' academic performance through 10 plausible values, the following procedure was followed to estimate the variable "low academic performance":

- Calculation of average reading literacy performance for each country.
- Classification of students as having low academic achievement: YES—in cases where their performance was below the estimated average performance for their country—and NO—if their reading literacy score was equal to or above the average performance for their country. This classification was made for each of the 10 plausible values provided by PISA.

In this regard, it should be noted that establishing the average performance of each country as a reference point for calculating this variable intends to overcome the limitation of PISA performance levels by considering the internal variability of each territory.

Contextual information was obtained by selecting some of the indices estimated by the OECD from students' responses to the student and well-being questionnaires. For this selection, Hattie's (2009) work was taken as a reference, as it proposes a classification of personal, family, and school aspects whose influence on performance has been analysed and demonstrated in the meta-analytical literature (**Table 2**).

The response rate for all variables was above 70%—except for *Learning time (in total)-minutes per week*, which had a response rate of 55%.

Procedure

The multivariate technique of decision trees through the binary CART algorithm (Breiman et al., 1984) was used. As mentioned previously, this technique allows the creation of subsets of data that are as homogeneous as possible with respect to the dependent variable, as well as the calculation of the relative importance of each independent variable in explaining the dependent variable.

To identify the personal, family, and school factors that explain low academic performance, first, 10 global models were estimated for the set of countries in the study sample. Second, the average importance of the predictors obtained in the 10 models was calculated. Finally, the relative importance of each independent variable with respect to the predictor that emerged as most relevant in explaining academic performance was calculated. In other words, the standardised importance reflects the impact of each of the independent variables in the model, so that it is possible to observe which are the most important (de Oña et al., 2012). Thus, the relative importance of the most relevant variable would be 100%, while the rest of the variables would be attributed importance proportional to that of 100%. This same procedure was performed with each of the subsamples corresponding to the different countries.

As shown in **Table 1**, the data were weighted using the SENWT variable in the estimation of the global model to ensure that the contributions of each of the countries were equal, regardless of their sample size. In the models estimated for each of the countries, the values were weighted by the final student weight (W_FSTUWT).

Together with the standardised importance of each of the independent variables, we present the variables that compose the profiles of students with the lowest reading performance in the global model and in each of the nine selected countries.

SPSS Statistics version 27 was used to conduct the analyses.

Results

The results of the global model—relative to the set of countries that comprise the sample—and of the specific models—for each of the nine countries—are presented below. As can be seen in **Table 3**, the overall average classification rate of the estimated models, which reveals the models' ability to correctly classify the variables through a percentage, is situated between 68.70% (average for the 10 general models) and 77.26% (average for the 10 Irish models).

Based on these decision trees—calculated for the whole sample and for each of the countries—the general most extreme profile of students with the lowest academic achievement in reading literacy is presented, along with the specific models that reflect cross-national differences in the effect of personal, family, and school characteristics on low performance.

Profile of students with the lowest reading achievement

This section shows the general most extreme profile of students with low academic achievement in reading using the decision tree estimated for the whole sample. In this estimation, a depth of six levels was established. However, since this article seeks to analyse the variables that best describe the profile

1 https://www.oecd.org/pisa/data/pisa2018technicalreport/

TABLE 2 Independent variables according to Hattie's (2009) categories and subcategories.

Type of variable	Category	Subcategory	Index/independent variable	PISA code	Questionnaire
Student	Attitudes and dispositions	Motivation	Mastery goal orientation	MASTGOAL	Student
			Work mastery	WORKMAST	Student
			Expected occupational status	BSMJ	Student
			Eudaimonia: meaning in life	EUDMO	Student
			Subjective well-being: positive affect	SWBP	Student
		Attitude to school subjects	Enjoyment of reading	JOYREAD	Student
			Attitude towards school: learning activities	ATTLNACT	Student
			Subjective well-being: sense of belonging to school	BELONG	Student
		Concentration/persistence/ engagement	Learning time (in total)-minutes per week	LMINS	Student
		Personality	Competitiveness	COMPETE	Student
			General fear of failure	GFOFAIL	Student
			Resilience	RESILIENCE	Student
		Self-concept	Self-concept of reading: perception of competence	SCREADCOMP	Student
			Self-concept of reading: perception of difficulty	SCREADDIFF	Student
			Body image	BODYIMA	Well-being
	Background	Background	Understanding and remembering	UNDREM	Student
			Summarising	METASUM	Student
			Assessing credibility	METASPAM	Student
	Physical influences	Illness	Student's body mass index	BMI	Well-being
	Preschool experiences	Early interventions	Early childhood education and care	DURECEC	Student
Family	Home environment	Parental involvement in learning	Parents' emotional support perceived by student	EMOSUPS	Student
			Social connection to parents	SOCONPA	Well-being
	Socioeconomic and cultural status	Socioeconomic and cultural status	Educational level of parents	PAREDINT	Student
			Highest occupational status of parents	HISEI	Student
			Immigration background	IMMIG	Student
			Household possessions	HOMEPOS	Student
Teacher	Quality of teaching	Quality of teaching	Perceived feedback	PERFEED	Student
			Teacher's stimulation of reading engagement	STIMREAD	Student
			Perceived teacher's interest	TEACHINT	Student
	Teacher–student relationships	Teacher–student relationships	Adaptation of instruction	ADAPTIVITY	Student
			Teacher-directed instruction	DIRINS	Student
			Teacher support in test language lessons	TEACHSUP	Student
School	Classroom influences	Classroom behaviour	Disciplinary climate in test language classes	DISCLIMA	Student
		Group cohesion	Perception of competitiveness at school	PERCOMP	Student
			Perception of co-operation at school	PERCOOP	Student
		Peer influences	Student's experience of being bullied	BEINGBULLIED	Student

TABLE 3 Overall average classification rate of the models estimated from the 10 plausible values.

	Global	Bulgaria	Georgia	Hong Kong SAR (China)	Ireland	Mexico	Panama	Serbia	Spain	United Arab Emirates
Overall classification rate	68.70%	74.82%	73.13%	74.76%	77.26%	71.94%	69.23%	72.40%	72.21%	74.53%

of students with the lowest achievement in reading, only the variables that appear in the first three positions of the branch in which said profile is represented are displayed below (Table 4).

The student profile with the lowest achievement in reading is best defined by the three variables that emerged ordered by their discriminatory capacity, which decreases on descending

TABLE 4 Variables that best describe the profile of students with the lowest academic achievement in reading literacy.

Variable	Global	Bulgaria	Georgia	Hong Kong SAR (China)	Ireland	Mexico	Panama	Serbia	Spain	United Arab Emirates
Work mastery						2^{nd} 3 $3rd$ 5	2^{nd} 4 3^{nd} 1			$3rd$ 6
Student's expected occupational status (SEI)								2^{nd} 10		
Joy/like reading	2^{nd} 10	1^{st} 2 2^{nd} 2	1^{st} 6 2^{nd} 2	2^{nd} 8 3^{rd} 2		3^{rd} 1	3^{rd} 1	3^{rd} 2		2^{nd} 5
Attitude towards school: learning activities					3^{rd} 1					
Subjective well-being: sense of belonging to school		2^{nd} 1 3^{rd} 7								
Learning time (minutes per week) – in total							3^{rd} 1			
Self-concept of reading: perception of competence	3^{rd} 2				1^{st} 10	3^{rd} 1	2^{nd} 2		3^{rd} 2	
Self-concept of reading: perception of difficulty	3^{rd} 8		1^{st} 1 2^{nd} 4 3^{rd} 4						3^{rd} 8	1^{st} 4 3^{rd} 3
Meta-cognition: understanding and remembering			1^{st} 3 2^{nd} 3	3^{rd} 4						
Meta-cognition: summarising	1^{st} 10	1^{st} 8 2^{nd} 1		2^{nd} 2 3^{rd} 3	2^{nd} 3 3^{rd} 1	1^{st} 10		1^{st} 10	2^{nd} 10	1^{st} 1 3^{rd} 1
Meta-cognition: assess credibility				1^{st} 10	2^{nd} 7 3^{rd} 8			3^{rd} 1	1^{st} 10	
Social connection to parents			3^{rd} 1					3^{rd} 4		
Highest occupational status of parents						2^{nd} 1				
Immigration background										1^{st} 5 2^{nd} 5
Household possessions		2^{nd} 6 3^{rd} 3	2^{nd} 1 3^{rd} 5			2^{nd} 6 3^{rd} 3	1^{st} 10 2^{nd} 4 3^{rd} 7			
Teacher-directed instruction								3^{rd} 1		
Perception of co-operation at school				3^{rd} 1						
Student's experience of being bullied								3^{rd} 2		

The subscript represents the number of times each predictor emerges in the 1^{st}, 2^{nd}, and 3^{rd} position in each of the 10 models estimated for each country and for the whole sample (1–10).

TABLE 5 Mean performance in reading for each country and number of students with low achievement.

Country	Bulgaria	Georgia	Hong Kong SAR (China)	Ireland	Mexico	Panama	Serbia	Spain	United Arab Emirates
Mean achievement	419.84	379.75	524.28	518.08	420.47	376.97	439.47	476.54	431.78
Number of students with low achievement (weighted)	2,341	2,630	2,318	2,460	2,541	2,558	2,494	2,221	2,557
Percentage of students with low achievement	46.8%	52.59%	46.36%	49.20%	50.82%	51.15%	49.87%	44.42%	51.13%

the nodes of the tree. Thus, the variable *Meta-cognition: summarising* appears in first place with regard to segmentation of the sample, as it has the greatest discriminatory capacity when characterising students with the lowest performance in the 10 global models. The second variable that allows charactirisation of students with the lowest achievement is *Joy/like reading*, meaning that a lack of enjoyment of this activity may be another key aspect in low performance. After that, *Self-concept of reading: perception of difficulty* and *Self-concept of reading: perception of competence* emerge (eight times and twice, respectively) in the third position, indicating that students with very low reading achievement would also consider reading as a difficult task and would regard themselves as not sufficiently prepared to complete it satisfactorily.

Table 4 also shows the aspects that characterise students with the lowest achievement across countries. As can be seen, personal variables linked to *meta-cognition, joy/like reading,* and *cognitive self-concept* present a notable discriminatory capacity in the profile of students with the lowest academic achievement in most of the countries considered, which is in line with the results of the general model. In this sense, it is worth noting that Panama is the only country in which no metacognitive variable appears in the profile.

Work mastery is also a prevalent variable in relation to the personal dimension, being found in the models of three of the countries (Mexico, Panama, and the United Arab Emirates). In addition, *Subjective well-being: sense of belonging to school* appears as a defining variable in Bulgaria and *Student's expected occupational status* does in Serbia. Finally, *Attitude towards school: learning activities* is found in Ireland and *Learning time (minutes per week)-in total* in Panama; nevertheless, each of these only appears in one of the 10 estimated models.

Home possessions emerge as a family variable with a high discriminatory capacity in Bulgaria, Georgia, Mexico, and Panama. However, it is not present in Hong Kong SAR (China), Ireland, Spain, or in the United Arab Emirates. Also, *Immigration background* is found in the profile of students with very low academic performance in the United Arab Emirates. Finally, *Social connection to parents* and *Highest occupational status of parents* show some presence in Georgia and Serbia and in Mexico, respectively.

School and teacher variables are poorly represented among those aspects characterising students with the lowest academic achievement, except for *Teacher-directed instruction* and *Student's experience of being bullied* in Serbia and *Perception of cooperation* in Georgia, all of which show a minor presence among the first three positions of the models.

Beyond these results, it is necessary to consider that there are differences in the average level in the achievement of the selected countries. **Table 5** shows the mean achievement for each country and the data on student underachievement. These data were used to estimate the models that describe the profile of students with very low academic performance in each country. Hence, the percentage of students with a low performance can be observed to be below 50% in Hong Kong SAR (China), Ireland, Serbia, and Spain, but above 50% in the remaining countries.

Variables with the highest explanatory capacity for low reading achievement

To analyse the cross-national differences in the personal, family, and school characteristics of students that most contribute to explaining low academic achievement in reading, this section presents the standardised importance that these predictors have in both the overall model and each of the models estimated for the nine countries.

In this vein, it should be clarified that the models have been estimated by considering all the indices together. However, the results are presented in three sections based on each of the three ambits considered—personal, family, or school—for easier reading. It is also noteworthy that the standardised importance corresponding to each of the independent variables is established according to the variable that best contributes to explaining the dependent variable, to which a value of 100% is attributed.

Student variables

The results presented in **Table 6** show that the variable *Joy/like reading* has the greatest explanatory capacity for low performance in the global model. Although the importance of this ability differs among the countries considered, it can be found among the top five positions in all the territories except for the United Arab Emirates, Mexico, and Panama. The rest of the variables that constitute the subcategory *Attitude to school subjects* present low standardised importance in the overall model.

The three metacognitive indicators can be considered to show a high explanatory capacity both in the overall model and in each of the selected countries. However, this explanatory capacity differs among territories. Hence, *Meta-cognition: summarising* is the variable that most contributes to explaining low performance in Spain. It can also be found among the top five positions in all the remaining countries. Moreover, *Meta-cognition: understanding and remembering* leads the chart in Serbia and occupies the second position in Georgia, and *Meta-cognition: assess credibility* occupies the first position in Hong Kong SAR (China), and Ireland.

Cognitive self-concept is another construct that plays an important role in explaining low reading achievement across countries. At least one of the two cognitive self-concept variables appears at the top of the table in all the countries considered, except for Serbia. By contrast, *Body image*, which is related to physical self-concept, has an unremarkable explanatory capacity in all the selected countries, ranking below the middle of the table in most of the models.

Regarding motivational variables, only *Work mastery* and *Student's expected occupational status* appear to be remarkable in the overall model. However, differences between the nine countries in relation to both aspects are noteworthy. On one hand, *Work mastery* is in the top five positions in Panama, while it is situated at the bottom of the ranking in Hong Kong SAR (China). On the other hand, although *Student's expected occupational status* appears in the first half of the table in all the countries analysed, it only ranks among the top five positions in Serbia. Finally, the variables *Eudaimonia: meaning in life*, *Subjective well-being: positive affect,* and *Mastery goal orientation* are situated outside the top positions in all the models.

Regarding personality indices, *Resilience* is the variable in this category with the greatest explanatory capacity in the overall model and that is best positioned in the said category in most of the selected countries. However, this set of variables plays a minor role in all the countries analysed.

The variable *Learning time (minutes per week)-in total* shows a standardised importance rate of 19.2% in the overall model. However, the cross-country differences in relation to this aspect are notable: while this variable is located near the centre of the table in most countries, it can be found leading the table in Panama.

TABLE 6 Standardised importance of the independent student variables.

Independent variable	Global	Bulgaria	Georgia	Hong Kong SAR (China)	Ireland	Mexico	Panama	Serbia	Spain	United Arab Emirates
Motivation										
Mastery goal orientation	11.9%	8.7%	20.2%	15.2%	15.1%	19.2%	6.5%	14.9%	15.8%	19.4%
Work mastery	30.2%	58.4%	61.3%	7.3%	7.6%	60.9%	41.6% (4th)	41.8%	12.4%	64.2%
Student's expected occupational status (SEI)	42.3%	49.2%	22.5%	7.5%	48.0%	31.4%	35.9%	90.4% (2nd)	55.1%	23.1%
Eudaimonia: meaning in life	–	6.0%	11.1%	5.7%	16.7%	10.1%	5.9%	17.3%	5.1%	7.9%
Subjective well-being: positive affect	6.7%	10.8%	15.6%	–	–	12.0%	9.2%	6.6%	–	23.5%
Attitude to school subjects										
Joy/like reading	100% (1st)	83.2% (2nd)	100% (1st)	47.9% (4th)	72.8% (3rd)	54.5%	12.1%	90.4% (2nd)	66.1% (5th)	48.3%
Attitude towards school: learning activities	5.9%	18.1%	7.5%	15.2%	8.0%	25.8%	7.0%	11.2%	5.5%	19.8%
Subjective well-being: sense of belonging to school	8.6%	55.5%	24.3%	7.5%	–	22.2%	15.0%	22.7%	–	10.8%
Concentration/persistence/engagement										
Learning time (minutes per week) – in total	19.2%	22.7%	30.6%	12.0%	18.4%	73.8%	100% (1st)	60.0%	11.3%	20.3%
Personality										
Competitiveness	6.3%	15.7%	16.8%	8.1%	–	20.2%	11.9%	5.9%	–	22.1%
General fear of failure	–	7.4%	9.1%	9.3%	6.1%	7.5%	6.3%	–	–	5.5%
Resilience	17.3%	40.6%	34.8%	6.8%	–	20.8%	28.0%	24.7%	9.4%	30.3%
Self-concept										
Self-concept of reading: perception of competence	79.9% (2nd)	80.6% (3rd)	52.8%	35.5% (5th)	76.6% (2nd)	83.9% (4th)	26.0%	71.6%	78.6% (3rd)	67.9% (5th)
Self-concept of reading: perception of difficulty	64.2% (5th)	63.5% (5th)	75.9% (3rd)	26.7%	56.0% (4th)	66.8%	47.4% (3rd)	69.1%	76.7% (4th)	100% (1st)
Body image	–	8.8%	9.1%	8.6%	8.1%	7.7%	5.1%	7.4%	6.1%	8.2%
Background										
Meta-cognition: understanding and remembering	56.7%	55.5%	98.2% (2nd)	70.9% (3rd)	30.8%	50.8%	25.6%	100% (1st)	47.2%	84.6% (2nd)
Meta-cognition: summarising	73.9% (3rd)	76.1% (4th)	67.6% (4th)	83.1% (2nd)	54.8% (5th)	86.9% (3rd)	40.5% (5th)	82.7% (4th)	100% (1st)	69.0% (4th)
Meta-cognition: assess credibility	65.3% (4th)	28.4%	12.1%	100.0% (1st)	100% (1st)	64.7%	30.7%	80.1% (5th)	81.8% (2nd)	64.2%
Illness										
Body mass index of student	–	15.3%	11.3%	13.0%	10.0%	11.9%	20.3%	8.5%	–	8.2%
Early interventions										
Duration in early childhood education and care	–	–	12.6%	7.1%	–	5.6%	15.7%	–	5.8%	19.6%

Only standardised importance values above 5% are displayed.

Finally, the variables *Body mass index of student* and *Duration in early childhood education and care* are located in the second half of the table in most of the countries with the exception of the former in Hong Kong SAR (China), Ireland, and Panama and the latter in Panama.

Family variables

The results reflecting the standardised importance and position of the family variables (Table 7) show that socio-economic and cultural factors are the aspects related to the family environment that most contribute to explaining low performance in the global model—with the exception of *Educational level of parents*. However, some remarkable differences between the territories must be considered.

First, the variable *Household possessions* is among the top positions in four of the countries (Bulgaria, Georgia, Mexico, and Panama)—being the first variable of the model in Mexico and Bulgaria—and is situated in the first half of the ranking in all the countries analysed. Also, although *Highest occupational status* greatly contributes to explaining academic performance in almost all the countries (especially in Mexico), it only occupies the 18th position in Georgia. In addition, *Immigration background* presents wide diversity across the countries, being the third most important variable in the United Arab Emirates

TABLE 7 Standardised importance of the independent family variables.

Independent variable	Global	Bulgaria	Georgia	Hong Kong SAR (China)	Ireland	Mexico	Panama	Serbia	Spain	United Arab Emirates
Parental involvement in learning										
Parents' emotional support perceived by student	21.2%	44.9%	39.9%	5.4%	–	21.1%	9.4%	51.9%	7.5%	10.1%
Social connection to parents	18.2%	49.6%	44.3%	8.1%	–	10.6%	5.7%	41.1%	5.4%	23.5%
Socioeconomic and cultural status										
Educational level of parents	18.5%	58.0%	14.2%	7.5%	11.5%	80.1% (5^{th})	38.7%	31.7%	22.3%	45.5%
Highest occupational status of parents	41.9%	60.3%	15.8%	33.3%	20.8%	87.3% (2^{nd})	28.2%	64.8%	35.7%	55.3%
Immigration background	20.7%	14.7%	9.7%	–	–	15.4%	27.4%	10.5%	8.5%	81.2% (3^{rd})
Household possessions	35.1%	100% (1^{st})	66.4% (5^{th})	31.5%	21.2%	100% (1^{st})	55.1% (2^{nd})	74.6%	39.5%	46.3%

Only standardised importance values above 5% are displayed.

and the second last in Hong Kong SAR (China), and Ireland. Finally, the disparity is also evident for the variable *Educational level of parents*, as this appears in the 24th place in Hong Kong SAR (China) but occupies the 5th position in Mexico.

In relation to parental involvement, the standardised importance values for the two variables considered (*Parents' emotional support perceived by student* and *Social connection to parents*) are around 20% in the overall model. Again, the results show cross-country differences, with both variables ranking close to the 10th position in Georgia but appearing situated at the bottom of the table in Ireland.

School and teacher variables

The results in **Table 8** show the low explanatory capacity of school and teacher variables in the overall model. Moreover, none of the variables within this domain are among the main predictors of low achievement in any of the countries considered. However, some differences among the territories require further analysis.

In terms of the aspects more directly related to school, having an *Experience of being bullied* plays an important role in explaining underachievement in some countries. However, some diversity can be observed, as this variable is in the lower half of the table in Ireland and the United Arab Emirates but reaches the upper half in all the remaining countries. Another school factor, *Disciplinary climate in test language classes,* ranks Considerably higher in Serbia and Hong Kong SAR (China) than in the other countries. Finally, the variables related to group cohesion have a low explanatory capacity both in the overall model and in all the countries considered.

With regards to teacher-related variables, only *Teacher-directed instruction* shows a standardised importance of above 6%. However, these teacher-related variables are still below the middle of the ranking in almost all the countries considered.

Discussion

The explanation of underachievement and the search for its associated factors have been of constant interest in educational research. In this regard, the number of variables involved in the description and explanation of achievement—and, more specifically, of underachievement—has increased over the years, as so the number of studies. The very selection of the explanatory variables itself poses a bias in the analysis.

Therefore, to get as broad a picture as possible of this phenomenon, we have analysed the factors affecting low academic achievement using the data that—at least up to now—offer the most complete overview of performance: that is, taking the data from the international PISA assessment. For this purpose, and to adjust the explanation to the reality of each country, all available information on possible associated variables has been incorporated into the analyses. For this reason, only countries that had applied all context questionnaires were included in the sample.

The results of this study show the effects of personal, family, and school characteristics on low academic achievement. Hence, despite slight differences between the countries analysed, the variables that most influence low academic performance are mainly linked to the students themselves (low metacognition, lack of enjoyment of reading, poor self-concept, and low expectations about their future occupational status). In addition, at the family level, socio-economic aspects also play a significant role in explaining low academic achievement. These results are in line with those obtained by Sipe and Curlette (1997) and Hattie (2003), who reported personal variables, followed by family variables, to have the highest predictive capacity for low academic performance. However, in contrast to the results obtained in the aforementioned review papers, in our research, teacher variables were shown to have a low explanatory capacity.

Focusing attention on students' characteristics, this research shows the major role of a lack of enjoyment of reading in

TABLE 8 Standardised importance of the independent teacher and school variables.

Independent variable	Global	Bulgaria	Georgia	Hong Kong SAR (China)	Ireland	Mexico	Panama	Serbia	Spain	United Arab Emirates
Quality of teaching										
Perceived feedback	–	5.9%	5.1%	–	–	14.8%	5.4%	–	–	–
Teacher's stimulation of reading engagement	5.8%	11.6%	9.3%	10.6%	6.6%	9.1%	–	27.1%	–	12.4%
Perceived teacher's interest	5.4%	7.9%	8.1%	8.4%	5.2%	9.2%	5.9%	25.8%	–	7.9%
Teacher–student relationships										
Adaptation of instruction	5.6%	11.5%	6.7%	7.8%	–	6.5%	–	13.8%	–	23.9%
Teacher-directed instruction	11.3%	9.2%	10.7%	5.2%	6.5%	22.0%	10.4%	42.5%	6.8%	24.2%
Teacher support in test language lessons	–	7.9%	6.7%	–	–	6.6%	5.0%	12.2%	–	6.9%
Classroom behaviour										
Disciplinary climate in test language classes	13.5%	21.4%	19.8%	15.4%	–	15.0%	7.0%	50.3%	–	16.9%
Group cohesion										
Perception of competitiveness at school	–	5.4%	9.7%	9.9%	–	7.7%	10.2%	8.8%	–	7.5%
Perception of co-operation at school	5.2%	11.0%	6.9%	11.7%	–	15.0%	9.8%	9.8%	6.2%	–
Peer influences										
Student's experience of being bullied	11.7%	34.8%	23.7%	12.9%	–	32.8%	30.9%	46.2%	10.6%	15.4%

Only standardised importance values above 5% are displayed.

students' low performance. In this vein, this variable not only presents the greatest predictive capacity in the global model but also plays a relevant role in almost all the countries considered. These results coincide with the conclusions of the meta-analysis conducted by Tze et al. (2016), in which high levels of boredom were significantly related to low levels of academic performance, as well as to low levels of motivation and poor study strategies. In this sense, it is necessary to consider the role of emotional self-regulation when managing boredom. As this is one of the main components of emotional intelligence, it has been shown to be a good predictor of academic results (Checa et al., 2008; Calero et al., 2014). Similarly, in the study by Chang et al. (2016), cognitive engagement is positively correlated with academic performance. As Calero et al. (2014) describe, this could be explained by the fact that engagement is an essential element of motivation when it comes to predicting academic performance, as it is linked to the subjective value that students give to the task they are performing and thus influences their desire to carry it out and the results obtained.

Along with reading enjoyment, metacognition and self-concept have also played a major role in explaining low student achievement in all the countries analysed in our study; these results are in line with the findings of Hattie's (2009, 2017). Furthermore, Ohtani and Hisasaka (2018) analysed 118 papers and reported that once intelligence is controlled, metacognition appears to be a good predictor of academic performance. These results seem to be reasonable, as metacognition refers to a person's knowledge of his or her own information processing abilities, cognitive processes, and strategies for developing said processes, and includes the executive skills responsible for monitoring and self-regulating them (Schneider, 2010). Therefore, if students can recognise and understand their mental processes properly, they are likely to apply them optimally while learning.

On the contrary, the personal characteristics that consistently show a low explanatory capacity for academic underachievement across the selected countries correspond to duration in early childhood education and care, *body image*, and *body mass index of student*. Regarding the first of these variables, it is often claimed that attending early childhood education improves academic outcomes in the long term. However, although findings about this aspect seem to be contradictory in meta-analytic literature, they tend to show that the effects depend more on factors such as the quality of education than on whether this level of education is attended (Van Huizen and Plantenga, 2018). In this vein, it is also worth mentioning that the results obtained by the aforementioned authors revealed that the positive effects of attending early childhood education are greater for disadvantaged children, thus demonstrating the interrelation of this predictor with the socio-economic status of the families.

Finally, results for the remaining personal variables differ depending on the countries analysed, although not substantially. A greater variability is only observed for learning time, especially in the two, especially in the two Latin American countries

considered. This difference is particularly noticeable in the case of Panama, where this variable ranks first in the model. Among other aspects, this could be related to the fact that in this region, some children spend part of their time at work rather than at school or studying, with the negative consequences this has on academic performance (Murillo and Román, 2014).

The importance of family factors in the explanation of academic achievement has also been analysed in this study, as the relationship between parents and children can be one of the most significant throughout a person's life (Vasquez et al., 2016). Regarding said factors, the overall results are, again, in line with the findings of existing reviews, which reveal a medium-low predictive capacity for this dimension (Sirin, 2005; Castro et al., 2015; Pinquart, 2016; Vasquez et al., 2016; Tan, 2017). However, in contrast to the similar patterns which personal variables follow in each of the countries analysed, there are notable differences between the explanatory capacity of family variables across the territories.

First, the explanatory capacity of parental involvement varies greatly across countries: while the highest explanatory capacity is found in Georgia, the influence of this dimension is very small in Ireland. In this regard, Hampden-Thompson et al. (2013) state that there is a growing body of research that, in line with ecological systems theory, suggests that countries exert social, cultural, political, public, and institutional influences on their inhabitants. Due to this, cross-country differences in the association between family involvement and educational outcomes could be the consequence of national variations in relation to very diverse economic, cultural, social, or political aspects.

Meanwhile, cross-country differences in relation to families' cultural and socio-economic status can be explained by the differences that exist between the territories analysed (World Bank, 2022a). In this vein, the United Arab Emirates, as a territory with very high rates of temporary labour immigration (Möller, 2022), is the country where immigrant status and occupational status have shown the greatest explanatory capacity for low academic performance. In this regard, the results also probably reflect education inequalities derived from differences between locals and immigrants in the school system. Also noteworthy is the prominent role played by the educational and occupational status of families and the material household resources in Bulgaria—being a country where inequality has been rising during the last decade (Peshev, 2015; Hallert, 2020). The same phenomenon is also observed in Mexico which, despite being the country with the lowest inequality in Latin America, is still affected by this problem (Amarante and Colacce, 2018). Finally, the great importance that these household material resources and the occupational status of families have in Serbia—as well as the students' expectations about their employment—should be highlighted. These aspects may be related in part to the unemployment rates that this country still faces despite their gradual reduction (World Bank, 2022b), and also to the recent economic and social stabilisation that this territory has faced after the war suffered between 1991 and 2001. Therefore, the results suggest that these variables are more important in countries where the discrimination between family socio-economic resources is greater.

Variables relating to the characteristics of schools and teachers have shown little influence both in defining the general profile of students with low achievement and in explaining underachievement in each of the countries analysed. These results are partly in line with the findings of Hattie (2003), who, while attributing a minor role to school characteristics in explaining academic performance, found that teachers had an explanatory capacity of about 30%. However, the author did not consider the interrelationship between variables, which could explain the differential results with respect to our work. In any case, some slight differences between the analysed territories are observed in relation to school variables.

Although the explanatory capacity of the selected variables is, in general, very similar in all the countries analysed, the high standardised importance of school variables in Serbia stands out. This may be linked to the great variance found in reading achievement across the schools in this territory (Organisation for Economic Co-operation and Development [OECD], 2019a). Also, the fact of having suffered bullying plays a remarkable role in seven of the nine countries considered (Bulgaria, Hong Kong SAR (China), Spain, Mexico, Panama, Georgia, and Serbia). Of these territories, only Spain has lower rates of exposure to bullying than the OECD average (Organisation for Economic Co-operation and Development [OECD], 2019d). Finally, the variability found in the influence of classroom disciplinary climate on academic underachievement must be highlighted, which may be related to differences in cultural and behavioural standards across countries (Ning et al., 2015).

In conclusion, although much research has been focused on identifying the personal, family, and school aspects that exert the greatest influence on students' low academic performance, our findings suggest the need to examine cross-national differences in greater depth and to consider the specificities of each territory. Despite the interest in these results, the main limitations of this study should be noted. On one hand, the selection of variables was based on the indicators of the PISA context questionnaires. Hence, there could be other explanatory variables for low academic performance—such as intelligence (Zaboski et al., 2018), self-regulation (Kyriakides et al., 2013), or perfectionism (Madigan, 2019) at the personal level, or the type of leadership of the school leaders (Chin, 2007) at the school level—which have not been considered in this article. On the other hand, cross-country comparisons have only been made between territories where all the context questionnaires were applied, which has reduced the number of international comparisons.

It would therefore be desirable to explore further the realities of the countries analysed in this study, as well as to

explore comparisons between the explanatory capacity of the variables in a larger number of territories. In this vein, this study lays the theoretical foundations for future research, as it demonstrates the need to go a step further in research into the conditioning factors of low academic performance, which have traditionally been addressed from an international non-comparative perspective that establishes universal conclusions. Thus, this work has demonstrated the existence of differences in the personal, family, and school variables depending on the territories analysed. The results of this research also contribute to laying the foundations to develop specific policies addressed for preventing and improving underachievement, in which the whole educational community should be involved (Vera Sagredo et al., 2021). In this sense, this article demonstrates that, in addition to the development of international common policies aimed at reducing educational problems—as is the case, for example, of those conducted in the European Union—every country should develop its own robust policies aimed at improving students' academic performance, which should be based on the specific influence that each of the variables exerts on said performance.

In conclusion, despite some common trends in the countries analysed, the variables that explain underachievement are different across them, given that socio-demographic and contextual conditions also differ. Therefore, although personal characteristics continue to be the ones that best explain academic performance, a series of contextual variables, especially related to families, exert a greater or lesser influence on performance depending on the level of development and characterisation of each country, and may even hide or annul certain personal characteristics.

In this vein, although personal variables have shown the greatest impact on students' underachievement, there is an interrelation between all the factors that influence students' academic performance (Bhowmik, 2019; Akbas-Yesilyurt et al., 2020). For this reason, consideration should be given to the possibility that personal factors may, in turn, be influenced by family or school variables. For example, we should inquire whether students' expected occupational status or attitude toward school may be conditioned by family or teachers' expectations, which also depend on their socio-economic status. We may also ask to what extent teachers might be influencing students' level of metacognition, self-concept, or boredom.

Thus, policies and interventions should not only target students but should also consider the context in which they live, paying special attention to their families. Also, adequate pre-service and continuous training should be guaranteed for all teachers to ensure that students receive an adequate educational response, paying special attention to those who work in disadvantaged socio-educational contexts (Fernández Batanero, 2011).

For all these reasons, there is a clear need to continue working toward equity as a starting point, so that once equal opportunities are achieved, other personal variables can flourish.

Data availability statement

Publicly available datasets were analysed in this study. These data can be found here: https://www.oecd.org/pisa/data/2018database/.

Ethics statement

Ethical review and approval were not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

Author contributions

BG: conceptualization and design of the study, validation, data analysis, interpretation of results, discussion and conclusions, and reviewing and editing. EL-M: conceptualization and design of the study, validation, data analysis, interpretation of results, and reviewing and editing. ECM: conceptualization and design of the study, validation, and reviewing and editing. All authors listed have made a substantial, direct, and intellectual contribution to the work, and approved the submitted version for publication.

Funding

This research has been conducted under the support of the Ayudas para la Formación de Profesorado Universitario (FPU).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Akbas-Yesilyurt, F., Kocak, H., and Yesilyurt, M. E. (2020). Spatial models for identifying factors in student academic achievement. *Int. J. Assess. Tools Educ.* 7, 735–752. doi: 10.21449/ijate.722460

Amarante, V., and Colacce, M. (2018). ¿Más o menos desiguales? Una revisión sobre la desigualdad de los ingresos a nivel global, regional y nacional. *Rev. Cepal* 2018, 7–34. doi: 10.18356/1d244513-es

Arroyo Resino, D., Constante Amores, I. A., and Asensio Muñoz, I. (2019). La repetición de curso a debate: Un estudio empírico a partir de PISA 2015. *Educación* 22, 69–92. doi: 10.5944/educxx1.22479

Asensio Muñoz, I., Carpintero Molina, E., Exposito Casas, E., and Lopez Martin, E. (2018). ¿Cuánto oro hay entre la arena? Minería de datos con los resultados de España en PISA 2015/How much gold is in the sand? Data mining with Spain's PISA 2015 results. *Rev. Española de Pedagog.* 76, 225–246. doi: 10.22550/REP76-2-2018-02

Bhowmik, M. K. (2019). "Ethnic minority young people's education in Hong Kong: factors influencing school failure," in *Education, Ethnicity and Equity in the Multilingual Asian Context*, eds J. Gube and F. Gao (New York, NY: Springer), 179–195. doi: 10.1007/978-981-13-3125-1_11

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees*. Pacific Grove: Wadsworth and Brooks-Cole.

Calero, M. D., Carles, R., Mata, S., and Navarro, E. (2014). Diferencias en habilidades y conducta entre grupos de preescolares de alto y bajo rendimiento escolar. *RELIEVE-Rev. Electrón. de Investig. y Eval. Educ.* 16, 1–17. doi: 10.7203/relieve.16.2.4137

Castro, M., Expósito-Casas, E., López-Martín, E., Lizasoain, L., Navarro-Asencio, E., and Gaviria, J. L. (2015). Parental involvement on student academic achievement: A meta-analysis. *Educ. Res. Rev.* 14, 33–46.

Chang, D. F., Chien, W. C., and Chou, W. (2016). Meta-analysis approach to detect the effect of student engagement on academic achievement. *ICIC Express Lett.* 10, 2241–2246. doi: 10.1186/s13054-016-1208-6

Checa, P., Rodríguez—Bailón, R., and Rueda, M. R. (2008). Neurocognitive and temperamental systems of self—regulation and early adolescents' social and academic outcomes. *Mind Brain Educ.* 2, 177–187. doi: 10.1111/j.1751-228X.2008.00052.x

Chin, J. M. C. (2007). Meta-analysis of transformational school leadership effects on school outcomes in Taiwan and the USA. *Asia Pac. Educ. Rev.* 8, 166–177. doi: 10.1007/BF03029253

de Oña, J., de Oña, R., and Calvo, F. J. (2012). A classification tree approach to identify key factors of transit service quality. *Expert Syst. Appl.* 39, 11164–11171. doi: 10.1016/j.eswa.2012.03.037

Ergen, B., and Kanadli, S. (2017). The effect of self-regulated learning strategies on academic achievement: A meta-analysis study. *Eurasian J. Educ. Res.* 17, 55–74. doi: 10.14689/ejer.2017.69.4

Fernández Batanero, J. M. (2011). Abandono escolar y prácticas educativas inclusivas. *Rev. Latinoam. Educ. Incl.* 5, 43–58.

Ghasemi, E., and Burley, H. (2019). Gender, affect, and math: A cross-national meta-analysis of Trends in International Mathematics and Science Study 2015 outcomes. *Large-Scale Assess. Educ.* 7, 1–25. doi: 10.1186/s40536-019-0078-1

Gorard, S., and Smith, E. (2003). "What Is" Underachievement at School"?," in *Working Paper Series Paper*, (Cardiff: Cardiff University School of Social Sciences).

Gutiérrez-de-Rozas, B., and López-Martín, E. (2020). *Contextualización y evaluación del fracaso escolar*. Madrid: Sanz y Torres.

Hallert, J. J. (2020). *Inequality, poverty, and social protection in bulgaria*. SSRN [Working Paper]. doi: 10.2139/ssrn.3688532

Hampden-Thompson, G., Guzman, L., and Lippman, L. (2013). A cross-national analysis of parental involvement and student literacy. *Int. J. Comp. Sociol.* 54, 246–266. doi: 10.1177/0020715213501183

Hattie, J. (2003). *Teachers Make a Difference. What is the research evidence?*. Camberwell: Australian Council for Educational Research.

Hattie's, J. (2009). *Visible Learning: A Synthesis of 800+ meta-Analyses on Achievement*. Milton Park: Routledge.

Hattie's, J. (2017). *Visible Learning plus. 250+ Influences on Student Achievement*. Available Online at: https://visible-learning.org/wp-content/uploads/2018/03/250-Influences-Final-Effect-Size-List-2017_VLPLUS.pdf (accessed March 3, 2022).

Holzer, J., Bürger, S., Samek-Krenkel, S., Spiel, C., and Schober, B. (2021). Conceptualisation of students' school-related wellbeing: Students' and teachers' perspectives. *Educ. Res.* 63, 474–496. doi: 10.1080/00131881.2021.1987152

Jiménez Fernández, C. (2010). *Diagnóstico y evaluación de los más capaces*. Hoboken: Prentice Hall.

Kornilova, T. V., Kornilov, S. A., and Chumakova, M. A. (2009). Subjective evaluations of intelligence and academic self-concept predict academic achievement: Evidence from a selective student population. *Learn. Individ. Diff.* 19, 596–608. doi: 10.1016/j.lindif.2009.08.001

Kyriakides, L., Christoforou, C., and Charalambous, C. Y. (2013). What matters for student learning outcomes: A meta-analysis of studies exploring factors of effective teaching. *Teach. Teach. Educ.* 36, 143–152. doi: 10.1016/j.tate.2013.07.010

Lamas, H. A. (2015). Sobre el rendimiento escolar. *Propósitos y Representaciones* 3, 313–386. doi: 10.20511/pyr2015.v3n1.74

López-Martín, E., Expósito-Casas, E., Carpintero Molina, E., and Asensio Muñoz, I. (2018). ¿Qué nos dice PISA sobre la enseñanza y el aprendizaje de las ciencias? una aproximación a través de árboles de decisión. *Rev. de Educ.* 382, 133–162.

Madigan, D. J. (2019). A meta-analysis of perfectionism and academic achievement. *Educ. Psychol. Rev.* 31, 967–989. doi: 10.1007/s10648-019-09484-2

Möller, L. M. (2022). "United Arab Emirates: temporary multiculturalism, but permanent legal pluralism?," in *Normativity and Diversity in Family Law*, eds N. Yassari and M.-C. Foblets (New York, NY: Springer), 101–117. doi: 10.1007/978-3-030-83106-6_5

Mullis, I. V. S., Martin, M. O., Foy, P., Kelly, D. L., and Fishbein, B. (2020). *TIMSS 2019 International Results in Mathematics and Science*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center.

Murillo, J., and Román, M. (2014). Consecuencias del trabajo infantil en el desempeño escolar. Estudiantes latinoamericanos de educación primaria. *Lat. Am. Res. Rev.* 49, 84–106. doi: 10.1353/lar.2014.0031

Ning, B., Van Damme, J., Van Den Noortgate, W., Yang, X., and Gielen, S. (2015). The influence of classroom disciplinary climate of schools on reading achievement: A cross-country comparative study. *Sch. Effect. Sch. Improv.* 26, 586–611. doi: 10.1080/09243453.2015.1025796

Ohtani, K., and Hisasaka, T. (2018). Beyond intelligence: A meta-analytic review of the relationship among metacognition, intelligence, and academic performance. *Metacogn. Learn.* 13, 179–212. doi: 10.1007/s11409-018-9183-8

Organisation for Economic Co-operation and Development [OECD] (2017). *PISA 2015 Results (volume III): Students' well-being*. Paris: OECD Publishing.

Organisation for Economic Co-operation and Development [OECD] (2019). *PISA 2018 Results. Combined Executive Summaries. Volume I, II & III*. Paris: OECD Publishing.

Organisation for Economic Co-operation and Development [OECD] (2019a). *PISA 2018 Insights and Interpretations*. Paris: OECD Publishing.

Organisation for Economic Co-operation and Development [OECD] (2019b). *PISA 2018 results. combined executive summaries*, Vol. I, II, III. Paris: OECD Publishing.

Organisation for Economic Co-operation and Development [OECD] (2019c). *PISA 2018 Assessment and Analytical Framework*. Paris: OECD Publishing, doi: 10.1787/b25efab8-en

Organisation for Economic Co-operation and Development [OECD] (2019d). *PISA 2018 Results (Volume III): What School Life Means for Students' Lives*. Paris: OECD Publishing, doi: 10.1787/acd78851-en

Organisation for Economic Co-operation and Development [OECD] (n.d.). "Chap. 19 - International Data Products," in *PISA 2018 technical report* (Paris: OECD Publishing).

Peshev, P. (2015). Analysis of the wealth inequality dynamics in Bulgaria: Different approach. *Econ. Altern. J.* 4, 29–33.

Pinquart, M. (2016). Associations of parenting styles and dimensions with academic achievement in children and adolescents: A meta-analysis. *Educ. Psychol. Rev.* 28, 475–493. doi: 10.1007/s10648-015-9338-y

Poropat, A. E. (2009). A meta-analysis of the five-factor model of personality and academic performance. *Psychol. Bull.* 135:322. doi: 10.1037/a0014996

Razi, M. A., and Athappilly, K. (2005). A comparative predictive analysis of neural networks (NNs), nonlinear regression and classification and regression tree (CART) models. *Expert Syst. Appl.* 29, 65–74. doi: 10.1016/j.eswa.2005.01.006

Schneider, W. (2010). "The development of metacognitive competences," in *Metacognition, Strategy Use, and Instruction*, eds H. Salatas Waters and W. Schneider (New York, NY: Guilford Press).

She, H. C., Lin, H. S., and Huang, L. Y. (2019). Reflections on and implications of the Programme for International Student Assessment 2015 (PISA 2015) performance of students in Taiwan: The role of epistemic beliefs about science in scientific literacy. *J. Res. Sci. Teach.* 56, 1309–1340. doi: 10.1002/tea.21553

Sipe, T. A., and Curlette, W. L. (1997). A meta-synthesis of factors related to educational achievement: A methodological approach to summarizing and synthesizing meta-analyses. *Int. J. Educ. Res.* 25, 583-698. doi: 10.1016/S0883-0355(96)00021-3

Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Rev. Educ. Res.* 75, 417–453. doi: 10.3102/00346543075003417

Tan, C. Y. (2017). Examining cultural capital and student achievement: Results of a meta-analytic review. *Alberta J. Educ. Res.* 63, 139–159.

Tze, V. M. C., Daniels, L. M., and Klassen, R. M. (2016). Evaluating the relationship between boredom and academic outcomes: A meta-analysis. *Educ. Psychol. Rev.* 28, 119–144. doi: 10.1007/s10648-015-9301-y

Van Huizen, T., and Plantenga, J. (2018). Do children benefit from universal early childhood education and care? A meta-analysis of evidence from natural experiments. *Econ. Educ. Rev.* 66, 206–222. doi: 10.1016/j.econedurev.2018.08.001

Van Petegem, K., Aelterman, A., Rosseel, Y., and Creemers, B. (2007). Student perception as moderator for student wellbeing. *Soc. Indic. Res.* 83, 447–463. doi: 10.1007/s11205-006-9055-5

Vasquez, A. C., Patall, E. A., Fong, C. J., Corrigan, A. S., and Pine, L. (2016). Parent autonomy support, academic achievement, and psychosocial functioning: A meta-analysis of research. *Educ. Psychol. Rev.* 28, 605–644. doi: 10.1007/s10648-015-9329-z

Vera Sagredo, A., Cerda Etchepare, G., Aragón Mendizábal, E., and Pérez Wilson, C. (2021). Rendimiento académico y su relación con variables socioemocionales en estudiantes chilenos de contextos vulnerables. *Educación* 24, 375–398. doi: 10.5944/educXX1.28269

World Bank (2022a). *Datos de desempleo*. Washington, DC: World Bank.

World Bank (2022b). *World Development Indicators*. Washington, DC: World Bank.

Zaboski, B. A., Kranzler, J. H., and Gage, N. A. (2018). Meta-analysis of the relationship between academic achievement and broad abilities of the Cattell-Horn-Carroll theory. *J. Sch. Psychol.* 71, 42–56. doi: 10.1016/j.jsp.2018.10.001

10.3389/feduc.2022.910039

Appendix

APPENDIX A General description of the nine selected countries (2018 data).

Descriptor	Bulgaria	Georgia	Hong Kong SAR (China)	Ireland	Mexico	Panama	Serbia	Spain	United Arab Emirates
Population	7,025,037	3,726,549	7,451,000	4,867,316	126,190,782	4,176,868	6,982,604	46,797,754	9,630,966
Net migration (2017 data)	−24,001	−50,000	146,542	118,020	−300,000	56,000	20,000	200,000	200,000
GDP per capita (US$ at constant 2010 prices)	7,860	4,539	45,285	72,608	9,946	14,881	6,262	27,726	39,671
Unemployment rate (% of the labour force)	5.2	12.7	2.8	5.7	3.3	3.8	12.7	15.3	2.4
Education expenditure (% of GNI)	4.0	1.8	2.8	4.1	4.5	2.8	3.7	4.0	N/A

Source: World Bank (2022a).

frontiers | Frontiers in Education

TYPE Original Research
PUBLISHED 12 August 2022
DOI 10.3389/feduc.2022.914445

OPEN ACCESS

EDITED BY
George Waddell,
Royal College of Music,
United Kingdom

REVIEWED BY
Stevo Popovic,
University of Montenegro, Montenegro
Matthieu E. M. Lenoir,
Ghent University, Belgium
Annike Bekius,
University of Amsterdam, Netherlands

*CORRESPONDENCE
Valentina Biino
valentina.biino@univr.it

SPECIALTY SECTION
This article was submitted to Assessment, Testing and Applied Measurement, a section of the journal Frontiers in Education

RECEIVED 06 April 2022
ACCEPTED 11 July 2022
PUBLISHED 12 August 2022

CITATION
Biino V, Giustino V, Guidetti L, Lanza M, Gallotta MC, Baldari C, Battaglia G, Palma A, Bellafiore M, Giuriato M and Schena F (2022) Körperkoordinations test für Kinder: A short form is not fully satisfactory. *Front. Educ.* 7:914445. doi: 10.3389/feduc.2022.914445

Körperkoordinations test für Kinder: A short form is not fully satisfactory

Valentina Biino[1]*, Valerio Giustino[2], Laura Guidetti[3], Massimo Lanza[1], Maria Chiara Gallotta[4], Carlo Baldari[5], Giuseppe Battaglia[2], Antonio Palma[2], Marianna Bellafiore[2], Matteo Giuriato[1] and Federico Schena[1]

[1]Department of Neurosciences, Biomedicine and Movement Sciences, University of Verona, Verona, Italy, [2]Sport and Exercise Sciences Research Unit, Department of Psychology, Educational Science and Human Movement, University of Palermo, Palermo, Italy, [3]Department of Unicusano, University Niccolò Cusano, Rome, Italy, [4]Department of Physiology and Pharmacology "Vittorio Erspamer", Sapienza University of Rome, Rome, Italy, [5]Department of Theoretical and Applied Sciences, eCampus University, Novedrate, Italy

Assessment of motor competence (MC) is crucial to finding deficiencies in children's motor development. Because of the need to ensure validity, reliability, and feasibility, the selection of contemporary testing batteries is a difficult task. Many papers report the validity of the KTK test in describing MC in school aged children. KTK consists of 4 four separate items: walking back, jumping sideways, moving sideways, and hopping for height. Some authors suggested the use of a short version of KTK that includes 3 items excluding one subtest: hopping for height. This study aimed to evaluate the effectiveness of short versions of Körperkoordinationstest für Kinder (KTK). A sample of 2,231 participants (boys: n=1,188; girls: n = 1,043; age range: 6–12 years) were recruited from Italian schools between January 2019 and February 2020 and they performed the complete version of KTK. Stepwise linear regression was performed on the dataset to evaluate the ideal number of variables to describe the KTK short form version. Data for both the sexes and all ages indicated that considering the item combinations of each short version, the highest R squares were obtained in those that included exactly the deleted subtest (ranging from 0.881–0.979). The adoption of a short form does not seem to provide a fully satisfactory condition for measuring MC in children 6–12 years.

KEYWORDS

KTK, KTK3, motor competence, motor coordination, motor assessment, children

Introduction

Identifying motor competence (MC) during childhood is crucial, not only to finding excellence in sports or future talents (O'Brien-Smith et al., 2019) but also to assessing an appropriate coordination level (Giuriato et al., 2021). Indeed, an adequate MC allows for the functioning of daily motor skills (Barnett et al., 2022) and the achievement of physical fitness (Lopes et al., 2018; Stodden et al., 2019). MC is generally evaluated to define children at risk of developing poor-motor coordination, which may result in an inability to perform daily activities and participate in health-related physical activity or organized

sport. MC is based on the components of physical fitness and motor coordination such as locomotion, manipulation, and stability (Gallahue et al., 2012; Barnett et al., 2016), which represent the primary focus for the development of motor skills.

Measuring pupils' level of MC performance provides an assessment of motor development. There are several test batteries to assess MC in children of all ages. A systematic review by Griffiths et al. (2018) outlines the majority of tools used to assess motor skills, including: the Bayley Scale of Infant and Toddler Development III (Bayley-III) (Urlich, 2019; Duncan et al., 2021), age range from 1 month to 3 years of age (Viezel et al., 2014); Test of Motor Proficiency (BOT-2), age range 4–21 years (Beitel and Mead, 1980); Movement Assessment Battery for Children (MABC-2), age range 3–16 years (Henderson and Sugden, 1992; Henderson et al., 2007); McCarron Assessment of Neuromuscular Development (MAND), 3–25 years (McCarron, 1997); Neurological Sensory Motor Developmental Assessment (NSMDA), from 1 month to 6 years (Burns et al., 1989); and Peabody Developmental Motor Scales second edition (PDMS-2), birth to 5 years (Folio and Fewell, 2000). All these test batteries provide an evaluation of fine and gross motor development and include balance, locomotion, object control, and an estimate of muscle strength (Barnett et al., 2016, 2022; Nascimento et al., 2019). Brian et al. (2016) investigated MC using test batteries that were converted to a standardized Motor Quotient (MQ) based on normative data: the Motor-Proficiency-Test for children (MOT 4-6) was validated with 548 children aged 4–6 years (Zimmer and Volkamer, 1987) and the Körperkoordination test für Kinder (KTK) was normalized on data of 1,128 German children aged 5–14 years (Kiphard and Schilling, 1974, 2007), both incorporated evaluation of fine and/or gross motor coordination skills.

Batteries such as the KTK have become widely used to measure general MC in young athletes (12 articles 2010–2014, 21 post-2015 in PubMed) (O'Brien-Smith et al., 2019), in addition, more than 50 articles about the assessment of gross motor coordination with the KTK test have been written from 2015 to date. The KTK assesses GMC through four subtests, i.e., walking backward (WB), jumping sideways (JS), moving sideways (MS), and hopping for height (HH) (Kiphard and Schilling, 1974, 2007). Its reliability and validity are well established ($r = 0.97$; WB = 0.80, JS = 0.95, MS = 0.85, and HH = 0.96).

Many studies (Fransen et al., 2014; Brian et al., 2016; Rudd et al., 2016; Ré et al., 2018) have focused on the validity and reliability of some test batteries such as BOT- 2 short forms vs. KTK, MOT 4–6 y vs. KTK, or TGMD-2 vs. KTK, showing their consistency. Although such consistency between these test batteries was moderate to adequate, these researchers recommended the scientific community not to evaluate children's MC based on the result of a single subtest, or a single test tool, but to consider a wider range of tests. This advice highlights that motor assessment is a complex issue, although these MC test batteries have excellent values of validity and reliability (Vandorpe et al., 2011; Griffiths et al., 2018). In addition, many researchers have raised the question of the feasibility of the test batteries, particularly, whether the performance measure is straightforward and easy to set up and administer. In terms of feasibility, the duration of the test and the potential risk of carrying out the tests are of significant importance. To reduce the duration, which often depends on the experience of the administrator and the age of the children, many authors have made short forms of their tests such as the BOT-2, or the TGMD-3 (Urlich, 2019).

Another reason for the proposal of short forms of some test batteries is the injury risk while performing the test (Novak et al., 2017). Therefore, some authors have shifted their attention to the short version of KTK (KTK 3) (Novak et al., 2017; O'Brien-Smith et al., 2019). The reason that led to the reduction of KTK to KTK 3, excluding the HH subtest is related to different causes, that is, the execution time which for KTK 3 is about 10 min instead of 20 min of the original KTK 4 version, and the reduced possibility of injury risk. KTK 3 is a reliable testing battery crosstab that revealed substantial agreement between the classification using KTK 3 and KTK 4 over all the age groups (Novak et al., 2017) and has therefore been used to assess MC across numerous individual and team sports such as soccer (Vandendriessche et al., 2012; Deprez et al., 2015), volleyball (Pion et al., 2015), and figure skating (Mostaert et al., 2016).

Moreover, the subtest of the monopodalic jump requires a large component of explosive force to gradually overcome greater heights, given by the addition of a 5 cm high cushion after each jump is performed correctly. A task strongly correlated with age and maturation is overcoming a single leg jump performed with both the right and the left leg of ∼60 cm total height. In this task, by the age of 8–9 years, men tend to perform better than women in all the subtests of the KTK except for the monopodalic jumping test (Luz et al., 2016). The difference in *the* performance of this task may be due to the peak of growth, which is reached on average about 2 years earlier by girls compared to boys. Therefore, HH does not exclusively measure gross motor coordination, such as walking backward, moving sideways, or jumping sideways (which is a simple coordination task becoming difficult under time pressure, the 15 s of the KTK execution). This is also one reason why it has been removed by many researchers. Therefore, HH has been removed in the short version because it relies more on strength or explosivity, while the focus in the short version is more on pure coordination, i.e., how well a child can organize their body movements for a given task, regardless of how tall/strong they are.

Based on this evidence, some researchers have used only a subset of the test batteries found in the literature (Fransen et al., 2014; van der Fels et al., 2015; Valentini et al., 2018; Coppens et al., 2021). Coppens et al. (2021) to assesses MC in children and adolescents and added to KTK 3 an alternating one-handed throwing and catching task of a tennis ball, validating the KTK 3 "plus." This version replaces the single leg jump with

TABLE 1 Anthropometrics characteristics, divided by age and sex.

		6-y		7-y		8-y		9-y		10-y		11-y		12-y	
		Boys	Girls	Boys	Girls	Boys	Girls	Boys	Girls	Boys	Girls	Boys	Girls	Boys	Girls
	n	136	124	163	161	223	210	203	166	202	178	124	99	137	104
Height (m)	Mean	1.21	1.20	1.28	1.26	1.34	1.32	1.39	1.38	1.45	1.44	1.49	1.50	1.57	1.55
	S.D.	0.06	0.05	0.06	0.07	0.06	0.07	0.07	0.06	0.07	0.09	0.08	0.08	0.08	0.06
Weight (Kg)	Mean	23.43	23.71	28.17	27.18	32.40	31.33	35.50	35.05	40.50	40.45	42.77	42.69	48.34	47.01
	S.D.	4.35	4.29	5.43	5.38	7.38	7.09	8.78	9.52	9.78	10.69	10.17	8.79	11.06	9.60

a task of throwing and catching a tennis ball against the wall (Platvoet et al., 2018). This recent review of the KTK test is also aimed at selecting the subtests of the KTK for tasks that are not conditioned by differences in body shape. BMI has been negatively associated with gross motor coordination (Battaglia et al., 2021) and does not seem to negatively affect the control abilities of the object (Barnett et al., 2016). Numerous studies (Fransen et al., 2014; Deprez et al., 2015; Pion et al., 2015) described significant differences in MS scores between high and low levels in athletes of both genders. It appears that this novel task can adequately represent gross motor coordination in children (O'Brien-Smith et al., 2019). The exercise consists of two valid tests of moving the tablets sideways over a period of 20 s as many times as possible.

Concerning the issue of the usefulness of test batteries, scientific literature agrees that MC is a complex and articulated question and that MC assessment should be approached carefully. Therefore, opting for a quick form of evaluation might not be consistent. It is uncertain whether shorter forms have a real advantage or if it is practical, in either case, this reduction removes information on the children's motor coordination.

This study aimed to identify whether there may be a short form or an extra-short form of the KTK that could be used to evaluate gross motor coordination in a sample of Italian children aged 6–12 years old, exploring whether shortcuts can assess MC. Analysis of the literature led us to hypothesize that the short form version of the KTK did not reduce information on MC levels.

Materials and methods

Participants

The present study is a part of research conducted between January 2019 and February 2020, in which 2,231 participants (boys: $n = 1{,}188$; girls: $n = 1{,}043$; age range: 6–12 years) were enrolled in primary and middle schools of northern, central, and southern Italy to assess motor coordination among a representative sample of the Italian children and early adolescents. Children with certification of movement disorders were excluded from the study. Table 1 shows the characteristics of the sample.

The authors received written informed consent from all the parents to participate in the study. To ensure participants' confidentiality, all the data were used anonymously. The study, in compliance with the Declaration of Helsinki, was approved by the Ethical Board of Verona University (N. 2019-UNVRCLE-0298910).

Anthropometric measurements

Anthropometric measurements (Table 1) were carried out using an electronic scale and a standard stadiometer to the nearest 0.5 kg and 0.1 cm to assess the participants' body mass (kg) and height (m) (Cole et al., 2000).

Motor coordination assessment

Motor competence was assessed with the Körperkoordinations test für Kinder (KTK) (Kiphard and Schilling, 1974, 2007), a test standardized for children aged 5 to 14 in which MC is expressed using the MQ index, a norm-referenced measure for age and gender. The KTK data were collected by the examiners trained in administering the test protocol and scored according to the manual guidelines.

The test protocol included the following four subtests: (1) walking backward (WB) on a balance beam of 3 m in the length of decreasing widths (6 cm, 4.5 cm, 3 cm); (2) two-legged jumping sideways (JS) from side-to-side over a beam (60 cm × 4 cm × 2 cm) as fast as possible for 15 s; (3) one-legged hopping for height (HH) over a foam obstacle of increasing height (consecutive increments of 5 cm); (4) moving sideways (MS) on the floor in 20 s by stepping from one plate (25 cm × 25 cm × 2 cm, supported on four legs 3.7 cm high) to the next, moving onto the first plate, stepping on it, and so on (Figure 1).

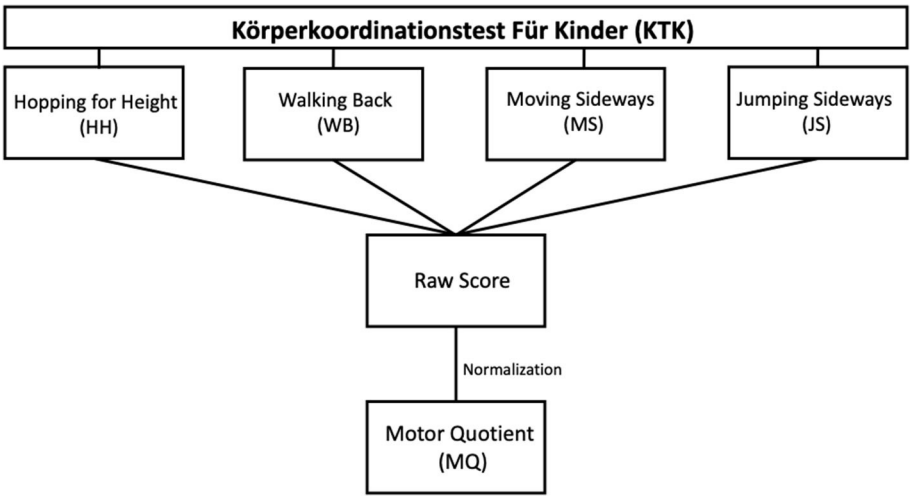


FIGURE 1
KTK test setup of 4 items.

TABLE 2 Data from stepwise linear regression declared among 6-year aged subject of both sexes.

Model	*R*	*R* square	Adjusted *R* square	Std. error of the estimate	*R* square change
Girls					
WB	0.797	0.635	0.632	18.555	0.635
HH	0.810	0.657	0.654	17.991	0.657
HH. JS	0.913	0.833	0.830	12.595	0.176
HH. WB	0.926	0.857	0.854	11.680	0.200
WB. JS	0.925	0.856	0.854	11.686	0.221
HH. JS. MS	0.934	0.872	0.868	11.094	0.038
WB. JS. MS	0.959	0.920	0.918	8.757	0.064
HH. WB. JS	0.987	0.974	0.973	4.991	0.117
Boys					
JS	0.806	0.649	0.646	19.547	0.649
JS. WB	0.888	0.788	0.784	15.266	0.139
JS. HH	0.916	0.839	0.836	13.295	0.190
JS. WB. MS	0.923	0.853	0.849	12.768	0.065
JS. HH. MS	0.940	0.884	0.881	11.348	0.045

The models are ordered by adjusted R square value from smallest to largest. All model reported have a significance $p < 0.001$.
HH, Hopping for Height; WB, Walking Back; JS, Jumping Sideways; MS, Moving Sideways.

TABLE 3 Data from stepwise linear regression declared among 7-year aged subject of both sexes.

Model	*R*	*R* square	Adjusted *R* square	Std. error of the estimate	*R* square change
Girls					
JS	0.850	0.722	0.720	20.081	0.722
JS. WB	0.912	0.831	0.829	15.695	0.109
JS. HH	0.937	0.877	0.875	13.401	0.155
JS. WB. MS	0.949	0.901	0.899	12.039	0.070
JS. HH. MS	0.949	0.901	0.899	12.051	0.024
JS. HH. WB	0.984	0.968	0.967	6.871	0.091
Boys					
JS	0.834	0.696	0.694	20.684	0.696
JS. WB	0.906	0.820	0.818	15.969	0.124
JS. HH	0.922	0.850	0.848	14.556	0.154
JS. HH. MS	0.935	0.875	0.872	13.374	0.024
JS. WB. MS	0.938	0.879	0.877	13.139	0.059
JS. HH. WB	0.986	0.973	0.973	6.194	0.123

The models are ordered by adjusted R square value from smallest to largest. All model reported have a significance $p < 0.001$.
HH, Hopping for Height; WB, Walking Back; JS, Jumping Sideways; MS, Moving Sideways.

According to the manual, the MQ was computed by adding the four raw scores of each item and then standardized for age and gender.

Kiphard and Schilling (2007) reported that the KTK showed acceptable construct validity, indeed, the test–retest for the raw score on the test protocol detected a reliability coefficient of 0.97, and for each item, reliability coefficients ranged from 0.80 to 0.96 (Kiphard and Schilling, 1974, 2007).

TABLE 4 Data from stepwise linear regression declared among 8-year aged subject of both sexes.

Model	*R*	*R* square	Adjusted *R* square	Std. error of the estimate	*R* square change
Girls					
WB	0.719	0.518	0.515	27.491	0.518
HH	0.793	0.628	0.627	24.126	0.628
WB. JS	0.877	0.769	0.766	19.087	0.251
HH. JS	0.911	0.830	0.828	16.373	0.201
HH. WB	0.917	0.840	0.838	15.868	0.212
WB. JS. MS	0.920	0.847	0.845	15.554	0.079
HH. JS. MS	0.934	0.873	0.871	14.161	0.044
HH. WB. JS	0.989	0.978	0.977	5.935	0.138
Boys					
JS	0.701	0.492	0.489	29.075	0.492
HH	0.814	0.663	0.662	23.664	0.663
JS. WB	0.850	0.723	0.720	21.514	0.231
JS. WB. MS	0.899	0.808	0.805	17.965	0.085
HH. JS	0.926	0.858	0.856	15.412	0.195
HH. WB	0.931	0.867	0.866	14.913	0.204
HH. JS. MS	0.945	0.893	0.891	13.425	0.035
HH. WB. JS	0.990	0.979	0.979	5.904	0.112

The models are ordered by adjusted R square value from smallest to largest. All model reported have a significance $p < 0.001$.
HH, Hopping for Height; WB, Walking Back; JS, Jumping Sideways; MS, Moving Sideways.

TABLE 5 Data from stepwise linear regression declared among 9-year aged subject of both sexes.

Model	*R*	*R* square	Adjusted *R* square	Std. error of the estimate	*R* square change
Girls					
WB	0.738	0.545	0.542	26.731	0.545
HH	0.795	0.632	0.630	24.033	0.632
WB. JS	0.853	0.727	0.723	20.785	0.182
WB. JS. MS	0.894	0.800	0.796	17.842	0.073
HH. JS	0.914	0.835	0.833	16.158	0.203
HH. WB	0.916	0.840	0.838	15.914	0.207
HH. JS. MS	0.929	0.862	0.860	14.806	0.027
HH. WB. JS	0.987	0.974	0.973	6.436	0.134
Boys					
JS	0.684	0.467	0.464	31.002	0.467
HH	0.807	0.651	0.649	25.087	0.651
JS. WB	0.860	0.740	0.737	21.707	0.273
JS. WB. MS	0.887	0.786	0.782	19.764	0.046
HH. JS	0.919	0.844	0.842	16.823	0.193
HH. JS. MS	0.948	0.899	0.897	13.585	0.055
HH. JS. WB	0.988	0.975	0.975	6.698	0.131

The models are ordered by adjusted R square value from smallest to largest. All model reported have a significance $p < 0.001$.
HH, Hopping for Height; WB, Walking Back; JS, Jumping Sideways; MS, Moving Sideways.

Statistical analysis

All data were manually entered into a spreadsheet and checked for transcription errors, with corrections made where appropriate analyzing data with a linear model to exclude them. The dataset was also suspended for outliers. Stepwise linear regression was performed on the dataset to evaluate the ideal number of variables to describe the KTK test, short-form version: MQ as a dependent variable, and different items (HH, MS, JS, WB) as independent variables. All analyses were performed in SPSS Statistics (v21, IBM, Chicago, IL, USA). *P*-value of <0.05 was fixed. The reliability of the regression models was expressed with adjusted R Square (R^2) and standard error of estimate (SEE).

Results

Stepwise linear regression was conducted to evaluate which subtest significantly predicted the model that described the KTK test. Adjusted R^2 and SEE for both the sexes were summarized in Tables 2–8, respectively, for 6–12 years, and in Figure 2 only for three subtest versions. The model that better describes KTK with the highest R^2 (0.979) and the lowest SEE (7.176) was composed of three items (WB, HH, and JS), accounting whole sample (all the ages) for both sex ($p < 0.001$). Accounting, only for girls HH, JS, and WB were the best significant model ($p < 0.001$) with high R^2 (0.980) and the lowest SEE (6.966); further for boys with it was composed of HH, JS, and WB ($R^2 = 0.978$; SEE = 7.365; $p < 0.001$).

Furthermore, reducing two items the model with the highest R^2 (0.900) and the lowest SEE (15.619) was composed of HH, and JS for all ages and both sexes ($p < 0.001$). Moreover, only for girls the best model was composed of HH and JS ($R^2 = 0.900$; SEE = 15.450; $p < 0.001$) and for boys composed of HH and JS ($R^2 = 0.902$; SEE = 15.527; $p < 0.001$).

Reducing the description of KTK at one subtest, the best predictor with the highest R^2 (0.694) and the lowest SEE (27.303) results HH ($p < 0.001$) for all ages and both sexes. Only for girls, it was HH ($R^2 = 0.716$; SEE = 26.082; $p < 0.001$) and for boys JS ($R^2 = 0.673$; SEE = 28.407; $p < 0.001$).

Dividing per age, the models that describe KTK (three, two, one item) vary slightly according to sex. The three-item version that best predicts KTK was composed of HH, JS, and WB (Tables 2–8). Particularly for girls, it was confirmed that HH, JS, and WB it was the best model with three items (Tables 2–8). From 7-y to 12-y is the best model (HH, WB, JS) for boys (Tables 3–8). Indeed, for 6-y, the best model was composed of JS, HH, and MS (Table 2).

TABLE 6 Data from stepwise linear regression declared among 10-year aged subject of both sexes.

Model	*R*	*R* square	Adjusted *R* square	Std. error of the estimate	*R* square change
Girls					
WB	0.761	0.579	0.576	25.282	0.579
HH	0.776	0.602	0.600	24.562	0.602
WB. JS	0.869	0.755	0.752	19.349	0.176
WB. JS. MS	0.892	0.795	0.792	17.731	0.041
HH. JS	0.905	0.819	0.816	16.646	0.216
HH. WB	0.913	0.833	0.831	15.971	0.230
HH. JS. MS	0.929	0.862	0.860	14.538	0.044
HH. WB. JS	0.985	0.971	0.971	6.653	0.138
Boys					
WB	0.715	0.512	0.509	27.849	0.512
HH	0.732	0.536	0.534	27.137	0.536
WB. JS	0.839	0.704	0.700	21.753	0.192
WB. JS. MS	0.877	0.769	0.765	19.281	0.065
HH. JS	0.922	0.850	0.849	15.456	0.314
HH. JS. MS	0.939	0.882	0.880	13.780	0.031
HH. JS. WB	0.984	0.969	0.969	7.032	0.119

The models are ordered by adjusted R square value from smallest to largest. All model reported have a significance $p < 0.001$.
HH, Hopping for Height; WB, Walking Back; JS, Jumping Sideways; MS, Moving Sideways.

TABLE 7 Data from stepwise linear regression declared among 11-year aged subject of both sexes.

Model	*R*	*R* square	Adjusted *R* square	Std. error of the estimate	*R* square change
Girls					
WB	0.533	0.284	0.275	23.698	0.284
HH	0.677	0.458	0.451	20.630	0.458
WB. MS	0.744	0.554	0.543	18.830	0.269
HH. JS	0.840	0.705	0.698	15.305	0.248
WB. MS. JS	0.851	0.724	0.713	14.906	0.170
HH. WB	0.856	0.733	0.727	14.558	0.276
HH. JS. MS	0.865	0.748	0.738	14.252	0.042
HH. WB. JS	0.960	0.922	0.919	7.919	0.189
Boys					
WB	0.533	0.284	0.275	23.698	0.284
WB. MS	0.744	0.554	0.543	18.830	0.269
HH	0.808	0.653	0.650	25.885	0.653
HH. JS	0.942	0.887	0.885	14.831	0.234
JS. MS. WB	0.945	0.893	0.890	14.521	0.111
HH. JS. MS	0.954	0.910	0.908	13.271	0.023
HH. JS. WB	0.984	0.967	0.967	8.001	0.080

The models are ordered by adjusted R square value from smallest to largest. All model reported have a significance $p < 0.001$.
HH, Hopping for Height; WB, Walking Back; JS, Jumping Sideways; MS, Moving Sideways.

KTK model described with two items it was found different between age. Girls' best model is described with HH, WB from 8 to 12 years (Tables 4–8); at 6-y, the best model was described with WB, JS (Table 2), and 7-y with JS, HH (Table 3). The model composed of items for boys is described at 6-, 7-, 10-, and 11-y with HH, JS, respectively in Tables 2, 3, 6, 7; indeed, at 8-, 9-, and 12-y with the items HH, WB (Tables 4, 5, 8).

Reducing at one predictor of KTK, the best item HH at 6, 8–12 years (Tables 2, 4–8), at 7-y the best predictor was JS (Table 3). Boys' best model was explained at 6-, 7-y with JS (Tables 2, 3), and 8- to 12-y with HH (Tables 4–8).

Discussion

Based on the KTK protocol developed by Kiphard and Schilling (Kiphard and Schilling, 1974, 2007), the purpose of the present study was to explore the validity of KTK short forms including 3 subtests by investigating their accuracy.

Our results suggest that the most accurate KTK short form is composed of HH, WB, and JS for all ages both girls and boys, though the impact of removing a subtest in the other short forms (i.e., WB, JS, and MS vs. HH, JS, and MS) showed good values of fit of these forms to the classical KTK protocol. In particular, the best short form we found was composed of HH, WB, and JS and showed an adjusted coefficient of determination ranging from 0.919 to 0.977 in girls of 11 and 8 years, respectively; and from 0.947 to 0.979 in boys of 12 and 8 years, respectively.

To the best of our knowledge, this is the first study that aimed to assess the most accurate KTK form including 3 subtests. Indeed, although previous studies have used a KTK version with 3 items (Vandendriessche et al., 2012; Deprez et al., 2015; Opstoel et al., 2015; Pion et al., 2015; Mostaert et al., 2016; de Niet et al., 2021), all of them have removed the HH item, probably because of the duration of the subtest. Novak et al. (2017) investigated the validity of a short form of this test battery, referred to as KTK3, following the removal of a specific subtest (i.e., the HH) by comparing it to the standard KTK protocol. The choice of the authors to consider a short version excluding the HH subtest was made based on the time necessary for the administration of this item which, as reported by the authors, takes about 10 min, that is, the same duration as the other three subtests put together (Novak et al., 2017).

In addition to considering the duration required to administer the KTK test, the potential risks in performing tasks with strong performance, and the variables due to factors related to maturation, height, and body weight, it should be taken into account that the removal of the subtest HH changes the evaluation results of gross motor coordination developmental age obtained with the Körperkoordinationtest für Kinder (Kiphard and Schilling, 1974).

TABLE 8 Data from stepwise linear regression declared among 12-year aged subject of both sexes.

Model	*R*	*R* square	Adjusted *R* square	Std. error of the estimate	*R* square change
Girls					
WB	0.700	0.490	0.483	29.493	0.490
HH	0.748	0.559	0.553	27.435	0.559
WB. MS	0.877	0.768	0.762	20.013	0.278
HH. JS	0.930	0.865	0.861	15.292	0.306
HH. WB	0.936	0.876	0.872	14.656	0.317
HH. JS. MS	0.942	0.887	0.883	14.049	0.023
WB. MS. JS	0.944	0.891	0.886	13.823	0.123
HH. WB. JS	0.982	0.965	0.964	7.819	0.089
Boys					
MS	0.648	0.420	0.414	27.593	0.420
HH	0.703	0.494	0.489	25.781	0.494
MS. WB	0.871	0.760	0.755	17.850	0.340
HH. JS	0.897	0.804	0.801	16.096	0.311
HH. WB	0.908	0.825	0.821	15.240	0.331
HH. JS. MS	0.925	0.856	0.852	13.855	0.052
MS. WB. JS	0.943	0.890	0.886	12.154	0.130
HH. WB. JS	0.974	0.948	0.947	8.322	0.124

The models are ordered by adjusted R square value from smallest to largest. All model reported have a significance $p < 0.001$.
HH, Hopping for Height; WB, Walking Back; JS, Jumping Sideways; MS, Moving Sideways.

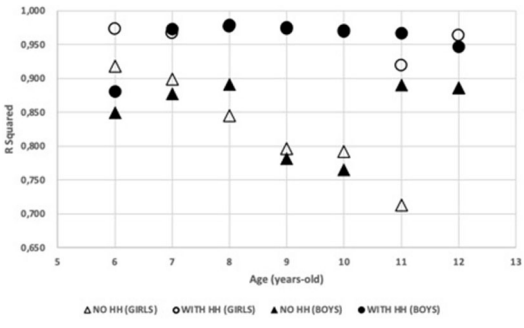


FIGURE 2
Comparison of short form's R squared. All the short forms that appear are composed of three items with HH (circle) and without HH (triangle), per sex (boys, full; girls, empty). NO HH, it is considering three items KTK with the highest R squared without HH; WITH HH, it is considering three items KTK with the highest R squared with HH. HH, Hopping Height item.

Although motor competence typically improved with age, maturity, and experience, children of the same chronological age may demonstrate significant variance in motor competence when they have a different skeletal age. The state of skeletal maturation interacts with the size of the body influencing the tests of motor coordination in children (Freitas et al., 2015); children who turn out to be precocious in the maturational state have higher levels of gross motor coordination. Age is directly correlated with motor competence (Barnett et al., 2016) but there are differences between subjects who develop at different rates (Lloyd and Oliver, 2012). Children who turn out to be precocious in the maturational state have higher levels of gross motor coordination. However, the benefits of accelerated development are not the same for boys and girls. Furthermore, Freitas et al. (2015) show that the differences in skeletal age explained only 9% of motor coordination. We believe that accuracy is a fundamental aspect of estimating the real level of motor coordination in children and adolescents (Yoon et al., 2006) and we agree with Kiphard and Schilling (1974) in believing that the data should not be read only based on age but also in the context of the whole person. Therefore, these authors have considered the advantage of letting the child reach the limit of their performance slowly (especially in HH), thus, reducing the renunciation of performance due to fear, inhibition, and disregard. Consistent with the previous research where the validity of short forms of other gross motor assessment tools has been demonstrated, the construct, and validity of this short form seem to indicate that a 3 -member KTK can be adopted to assess motor coordination, but it cannot replace the result in gross motor coordination obtained with KTK 4.

Although for the administration of a test battery researchers and practitioners should consider several factors (Goodway et al., 2019) such as the duration of the administration (for the latter, as properly examined by Novak et al., 2017) and other aspects such as setting assessment (e.g., educational or sports settings), characteristics of the population, cost of the test, and sample size, it is essential also to take into account the validity of the test battery (Cools et al., 2009), as suitably done by the authors who developed short versions of other gross motor assessment tools. In fact, over the past years, the possibility of using short forms of different gross motor tests battery for children and adolescents has been examined (Hassan, 2001; Cairney et al., 2009; Cools et al., 2009; Valentini et al., 2018). For instance, the short form of the second edition of the Bruininks-Oseretsky Test of Motor Proficiency (BOT-2 SF), including 14 items, has been validated based on the high correlation found with the complete form (BOT-2 CF), a tool composed of 53 items to measure fine and gross motor skills (Cairney et al., 2009; Cools et al., 2009). Recently, based on the Test of Gross Motor Development-2 (TGMD-2), an instrument for measuring 12 fundamental motor skills in children including "locomotor" and "object control" skills, Valentini et al. (2018) developed a valid and reliable short form of the TGMD-2 (TGMD-2 SF) with six skills, three for each of the two subtests.

The KTK has been specifically developed to adequately assess MC in children and adolescents both in educational and sports settings representing a valid and reliable instrument (Giuriato et al., 2019). The validity and reliability of the KTK (*r*

= 0.97) obviously depends on that of its subtests that ranging from 0.80 to 0.96 in which the highest coefficient of stability has been found precisely in the HH subtest (i.e., WB = 0.80, JS = 0.95, MS = 0.85, HH = 0.96, respectively) (Kiphard and Schilling, 2007). Hence, besides the duration required to administer the KTK, the removal of an item should be also taking into account the highest accuracy level in the measurement of motor coordination. Indeed, accuracy, which refers to the degree of correspondence of the measured value with the real value, represents a key aspect to estimate the real level of motor coordination in children and adolescents (Yoon et al., 2006).

For the creation of the KTK test, the main problem was identifying a difficulty in the tasks that embraced both the weakest 5 years-olds and the best 14-years-olds. The solution was found either in the time spent or in the task, through the attribution of scores; instead, for the HH subtest the performance limit was given by the increasing levels of difficulty (Kiphard and Schilling, 1974). In this light, the assumption that the development of the ability to move can only occur through the growing and self-determined comparison with environmental circumstances, puts in the foreground the role of the HH subtest.

Conclusions

Based on our results, the item removed in the most accurate short version (i.e., the MS) is a test that requires a high level of intersegmental coordination of both the upper and lower limb joints and also the trunk (Assaiante et al., 2005; Bekius et al., 2021). Hence, considering that this item requires very short administration times and that its exclusion would eliminate specific information on MC, we suggest using the classic version of the KTK (i.e., with 4 items) for a detailed assessment of motor coordination in children and adolescents. Compared with other gross motor test batteries the classic version of the KTK requires a shorter administration time, thanks to the fact that only four subtests are included. In addition, despite there being four items, it is possible to obtain a global index of motor coordination. Furthermore, two of the four KTK subtests are administered "in levels" and enable assessment of the maximum level of MC that each child can reach. Indeed, in the HH, foam obstacles of increasing height are added as long as the child can overcome them; while in the WB, the child can walk backward on balance beams of decreasing widths according to abilities. However, it should be noted that, although the KTK measures motor coordination in children and adolescents, the major limitation of this test battery is that it does not provide items measuring the upper limb, meaning it lacks specific information (e.g., throwing, catching skills).

Strengths and limitations of the study

The numerousness within the age groups is a limitation of this study because the age ranges are the main component of the KTK test (Kiphard and Schilling, 1974). In particular, this may have played a role in determining age-related differences in motor performance that required a strength component (HH and JS) (Vandorpe et al., 2011). In addition, this sample reduces the generalizability of our results. However, our data may provide useful information for computing a coefficient for future well-powered studies. A further limitation is we did not consider the time spent to perform the subtests for each age group of children. This would have made it possible to identify the presence of an effective difference in the timing of the tests in favor of the older children.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

Ethics statement

The studies involving human participants were reviewed and approved by the Ethical Board of Verona University (N. 2019-UNVRCLE-0298910). Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

Author contributions

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Assaiante, C., Mallau, S., Viel, S., Jover, M., and Schmitz, C. (2005). Development of postural control in healthy children: a functional approach. *Neural Plast.* 12, 109–272. doi: 10.1155/NP.2005.109

Barnett, L. M., Lai, S. K., Veldman, S., Hardy, L. L., Cliff, D. P., Morgan, P. J., et al. (2016). Correlates of gross motor competence in children and adolescents: a systematic review and meta-analysis. *Sports Med.* 46, 1663–1688. doi: 10.1007/s40279-016-0495-z

Barnett, L. M., Webster, E. K., Hulteen, R. M., De Meester, A., Valentini, N. C., Lenoir, M., et al. (2022). Through the looking glass: a systematic review of longitudinal evidence, providing new insight for motor competence and health. *Sports Med.* 52, 875–920. doi: 10.1007/s40279-021-01516-8

Battaglia, G., Giustino, V., Tabacchi, G., Lanza, M., Schena, F., Biino, V., et al. (2021). Interrelationship between age, gender, and weight status on motor coordination in italian children and early adolescents aged 6-13 years old. *Front. Pediat.* 9, 738294. doi: 10.3389/fped.2021.738294

Beitel, P. A., and Mead, B. J. (1980). Bruininks-oseretsky test of motor proficiency: a viable measure for 3- to 5-year-old children. *Percept. Motor Skills* 51, 919–923. doi: 10.2466/pms.1980.51.3.919

Bekius, A., Bach, M. M., van de Pol, L. A., Harlaar, J., Daffertshofer, A., Dominici, N., et al. (2021). Early development of locomotor patterns and motor control in very young children at high risk of cerebral palsy, a longitudinal case series. *Front. Hum. Neurosci.* 15, 659415. doi: 10.3389/fnhum.2021.659415

Brian, A., Bardid, F., Barnett, L. M., Deconinck, F. J. A., Lenoir, M., and Goodway, J. D. (2016). Actual and perceived motor competence levels of belgian and united states preschool children. J. Motor Learn. Develop. 6, S320–S336. doi: 10.1123/jmld.2016-0071

Burns, Y. R., Ensbey, R. M., and Norrie, M. A. (1989). The neuro-sensory motor developmental assessment. Part 1: development and administration of the test. *Aust. J. Physioth.* 35, 141–149.

Cairney, J., Hay, J., Veldhuizen, S., Missiuna, C., and Faught, B. E. (2009). Comparing probable case identification of developmental coordination disorder using the short form of the bruininks-oseretsky test of motor proficiency and the movement ABC. *Child Care Health Dev.* 35, 402–408. doi: 10.1111/j.1365-2214.2009.00957.x

Cole, T. J., and Bellizzi, M. C., Flegal, K. M., and Dietz, W. H. (2000). Establishing a standard definition for child overweight and obesity worldwide: international survey. *BMJ (Clin. Res. Ed.).* 320, 1240–1243. doi: 10.1136/bmj.320.7244.1240

Cools, W., De Martelaer, K., Samaey, C., and Andries, C. (2009). Movement skill assessment of typically developing preschool children: a review of seven movement skill assessment tools. *J. Sports Sci. Med.* 8, 154–168.

Coppens, E., Laureys, F., Mostaert, M., D'Hondt, E., Deconinck, F., and Lenoir, M. (2021). Validation of a motor competence assessment tool for children and adolescents (KTK3+) with normative values for 6- to 19-year-olds. *Front. Physiol.* 12, 652952. doi: 10.3389/fphys.2021.652952

de Niet, M., Platvoet, S. W. J., Hoeboer, J. J. A. A. M., de Witte, A. M. H., de Vries, S. I., and Pion, J. (2021). Agreement between the KTK3+ test and the athletic skills track for classifying the fundamental movement skills proficiency of 6- to 12-year-old children. *Front. Educ.* 6, 37. doi: 10.3389/feduc.2021.571018

Deprez, D. N., Fransen, J., Lenoir, M., Philippaerts, R. M., and Vaeyens, R. (2015). A retrospective study on anthropometrical, physical fitness, and motor coordination characteristics that influence dropout, contract status, and first-team playing time in a high-level soccer players aged eight to eighteen years. *J. Strength Condit. Res.* 29, 1692–1674. doi: 10.1519/JSC.0000000000000806

Duncan, M. J., Martins, C., Ribeiro Bandeira, P. F., Issartel, J., Peers, C., Belton, S., et al. (2021). TGMD-3 short version: evidence of validity and associations with sex in Irish children. *J. Sports Sci.* 40, 1–8. doi: 10.1080/02640414.2021.1978161

Folio, M. R., and Fewell, R. R. (2000). *Peabody Developmental Motor Scales-Second Edition: Examiner's Manual*. Austin, TX: PRO-ED.

Fransen, J., D'Hondt, E., Bourgois, J., Vaeyens, R., Philippaerts, R. M., and Lenoir, M. (2014). Motor competence assessment in children: convergent and discriminant validity between the BOT-2 Short Form and KTK testing batteries. *Res. Develop. Disabil.* 35, 1375–1383. doi: 10.1016/j.ridd.2014.03.011

Freitas, D. L., Lausen, B., Maia, J. A., Lefevre, J., Gouveia, É. R., Thomis, M., et al. (2015). Skeletal maturation, fundamental motor skills and motor coordination in children 7–10 years. *J. Sports Sci.* 33, 924–9342. doi: 10.1080/02640414.2014.977935

Gallahue, D., Ozmun, J., and Goodway, J. (2012). *Understanding Motor Development: Infants, Children, Adolescents, Adults*, ed D. Patterson (New York: McGraw-Hill International Edition).

Giuriato, M., Biino, V., Bellafiore, M., Battaglia, G., Palma, A., Baldari, C., et al. (2021). Gross motor coordination: we have a problem! A study with the körperkoordinations test für kinder in youth (6–13 years). *Front. Pediat.* 9, 785990. doi: 10.3389/fped.2021.785990

Giuriato, M., Pugliese, L., Biino, V., Bertinato, L., La Torre, A., and Lovecchio, N. (2019). Association between motor coordination, body mass index, and sports participation in children 6–11 years old. *Sport Sci. Health* 15, 463–468. doi: 10.1007/s11332-019-00554-0

Goodway, J. D., Ozmun, J. O., and Gallahue, D. L. (2019). *Understanding Motor Development: Infants, Children, Adolescents, Adults, 8th Edn*. Burlington, MA: Jones and Bartlett Learning, LLC.

Griffiths, A., Toovey, R., Morgan, P. E., and Spittle, A. J. (2018). Psychometric properties of gross motor assessment tools for children: a systematic review. *BMJ open* 8, e021734. doi: 10.1136/bmjopen-2018-021734

Hassan, M. M. (2001). Validity and reliability for the bruininks-oseretsky test of motor proficiency-short form as applied in the United Arab Emirates culture. *Percept. Motor Skills* 92, 157–166. doi: 10.2466/pms.2001.92.1.157

Henderson, S., and Sugden, D. (1992). *The Movement Assessment Battery for Children*. London: The Psychological Corporation.

Henderson, S., Sugden, D., and Barnett, A. (2007). *The Movement Assessment Battery for Children-2*. Minnesota: Pearson Education, Inc. doi: 10.1037/t55281-000

Kiphard, E. J., and Schilling, F. (1974). *Körperkoordinationstest für Kinder: KTK*. Weinheim: Beltz Test GmbH

Kiphard, E. J., and Schilling, F. (2007). *Körperkoordinationstest für Kinder: KTK*. Weinheim: Beltz Test GmbH

Lloyd, R. S., and Oliver, J. L. (2012). The youth physical development model: a new approach to long-term athletic development. *Strength Condit. J.* 34, 61–72. doi: 10.1519/SSC.0b013e31825760ea

Lopes, V. P., Malina, R. M., Maia, J., and Rodrigues, L. P. (2018). Body mass index and motor coordination: Non-linear relationships in children 6-10 years. *Child Care Health Develop.* 44, 443–451. doi: 10.1111/cch.12557

Luz, L. G., Cumming, S. P., Duarte, J. P., Valente-Dos-Santos, J., Almeida, M. J., Machado-Rodrigues, A., et al. (2016). Independent and combined effects of sex and biological maturation on motor coordination and performance in prepubertal children. *Percept. Motor Skills* 122, 610–635. doi: 10.1177/0031512516637733

McCarron, L. (1997). *McCarron Assessment of Neuromuscular Development: Fine and Gross Motor Abilities (revised ed.)*. Dallas, TX: Common Market Press.

Mostaert, M., Deconnick, F., Pion, J., and Lenoir, M. (2016). Anthropometry, physical fitness and coordination of young figure skaters of different levels. *Int. J. Sports Med.* 37, 531–538. doi: 10.1055/s-0042-100280

Nascimento, W., Henrique, N. R., and Marques, M. (2019). KTK motor test: review of the main influencing variables. *Rev. Paulista Pediatr.* 37, 372–381. doi: 10.1590/1984-0462/;2019;37;3;00013

Novak, A. R., Bennett, K. J., Beavan, A., Pion, J., Spiteri, T., Fransen, J., et al. (2017). The Applicability of a short form of the Körperkoordinationstest für Kinder for measuring motor competence in children aged 6 to 11 years. *J. Motor Learn. Dev.* 5, 227–239. doi: 10.1123/jmld.2016-0028

O'Brien-Smith, J., Tribolet, R., Smith, M. R., Bennett, K., Fransen, J., Pion, J., et al. (2019). The use of the Körperkoordinationstest für Kinder in the talent pathway in youth athletes: a systematic review. *J. Sci. Med. Sport* 22, 1021–1029. doi: 10.1016/j.jsams.2019.05.014

Opstoel, K., Pion, J., Elferink-Gemser, M., Hartman, E., Willemse, B., Philippaerts, R., et al. (2015). Anthropometric characteristics, physical fitness and motor coordination of 9 to 11 year old children participating in a wide range of sports. *PLoS ONE* 10, e0126282. doi: 10.1371/journal.pone.0126282

Pion, J. A., Fransen, J., Deprez, D. N., Segers, V. I., Vaeyens, R., Philippaerts, R. M., et al. (2015). Stature and jumping height are required in female volleyball, but motor coordination is a key factor for future elite success. *J. Strength Condit. Res.* 29, 1480–1485. doi: 10.1519/JSC.0000000000000778

Platvoet, S., Faber, I., De Niet, M., Pion, J., Kannekens, R., Elferink- Gemser, M., et al. (2018). Development of a tool to assess fundamental movement skills in applied settings. *Front. Educ.* 3, 75.00075. doi: 10.3389/feduc.2018.00075

Ré, A., Logan, S. W., Cattuzzo, M. T., Henrique, R. S., Tudela, M. C., and Stodden, D. F. (2018). Comparison of motor competence levels on two assessments across childhood. *J. Sports Sci.* 36, 1–6. doi: 10.1080/02640414.2016.1276294

Rudd, J., Butson, M. L., Barnett, L., Farrow, D., Berry, J., Borkoles, E., et al. (2016). A holistic measurement model of movement competency in children. *J. Sports Sci.* 34, 477–485. doi: 10.1080/02640414.2015.1061202

Stodden, D. F., Goodway, J. D., Langendorfer, S. J., Roberton, M. A., Rudisill, M. E., and Garcia, C. (2019). A developmental perspective on the role of motor skill competence in physical activity: an emergent relationship. *Quest* 60, 290–306. doi: 10.1080/00336297.2008.10483582

Urlich, D. A. (2019). *Test of Gross Motor Development, 3rd Edn*. New York, NY: pro-ed.

Valentini, N. C., Rudisill, M. E., Bandeira, P., and Hastie, P. A. (2018). The development of a short form of the test of gross motor development-2 in Brazilian children: validity and reliability. *Child Care Health Dev.* 44, 759–765. doi: 10.1111/cch.12598

van der Fels, I. M., Te Wierike, S. C., Hartman, E., Elferink-Gemser, M. T., Smith, J., and Visscher, C. (2015). The relationship between motor skills and cognitive skills in 4–16 year old typically developing children: a systematic review. *J. Sci. Med. Sport* 18, 697–703. doi: 10.1016/j.jsams.2014.09.007

Vandendriessche, J. B., Vaeyens, R., Vandorpe, B., Lenoir, M., Lefevre, J., and Philippaerts, R. M. (2012). Biological maturation, morphology, fitness, and motor coordination as part of a selection strategy in the search for international youth soccer players (age 15–16 years). *J. Sports Sci.* 30, 1695–1703. doi: 10.1080/02640414.2011.652654

Vandorpe, B., Vandendriessche, J., Lefevre, J., Pion, J., Vaeyens, R., Matthys, S., et al. (2011). The KörperkoordinationsTest für Kinder: reference values and suitability for 6–12-year-old children in Flanders. *Scand. J. Med. Sci. Sports* 21, 378–388. doi: 10.1111/j.1600-0838.2009.01067.x

Viezel, K., Zibulsky, J., Dumont, R., and Willis, J. O. (2014). "Bayley scales of infant and toddler development," in *Encyclopedia of Special Education: A Reference for the Education of Children, Adolescents, and Adults with Disabilities and Other Exceptional Individuals, 3rd Edn* (Hoboken, NJ: John Wiley and Sons). doi: 10.1002/9781118660584.ese0278

Yoon, D. Y., Scott, K., Hill, M. N., Levitt, N. S., and Lambert, E. V. (2006). Review of three tests of motor proficiency in children. *Percept. Mot. Skills* 102, 543–551. doi: 10.2466/pms.102.2.543-551

Zimmer, R., and Volkamer, M. (1987). *Motoriktest für vierbis sechsjährige Kinder: Mot 4-6; Manual.* Weinheim: Beltz-Test.

frontiers | Frontiers in Education

TYPE Original Research
PUBLISHED 26 September 2022
DOI 10.3389/feduc.2022.912837

OPEN ACCESS

EDITED BY
José Sánchez-Santamaría,
University of Castilla-La Mancha, Spain

REVIEWED BY
Ariel Mariah Lindorff,
University of Oxford, United Kingdom
Timothy Fukawa-Connelly,
Temple University, United States

*CORRESPONDENCE
Xin Liu
Xin.Liu@UGent.Be

SPECIALTY SECTION
This article was submitted to Assessment, Testing and Applied Measurement, a section of the journal Frontiers in Education

RECEIVED 04 April 2022
ACCEPTED 24 August 2022
PUBLISHED 26 September 2022

CITATION
Liu X, Valcke M, Yang Hansen K and De Neve J (2022) Exploiting the linked teaching and learning international survey and programme for international student assessment data in examining school effects: A case study of Singapore. *Front. Educ.* 7:912837. doi: 10.3389/feduc.2022.912837

Exploiting the linked teaching and learning international survey and programme for international student assessment data in examining school effects: A case study of Singapore

Xin Liu[1]*, Martin Valcke[1], Kajsa Yang Hansen[2] and Jan De Neve[3]

[1]Department of Educational Studies, Faculty of Psychology and Educational Sciences, Ghent University, Ghent, Belgium, [2]Department of Education and Special Education, Faculty of Education, University of Gothenburg, Gothenburg, Sweden, [3]Department of Data Analysis, Faculty of Psychology and Educational Sciences, Ghent University, Ghent, Belgium

This paper attempts to demonstrate the usefulness of the linkage data from two international large-scale assessment studies, Teaching and Learning International Survey 2013 (TALIS) 2013 and Programme for International Student Assessment (PISA) 2012, in examining the effects of schools. Data from seven educational systems are used to link, and four critical issues with five selection criteria are applied to the data selected. The linking dataset facilitates the investigation of mathematics performance while considering individual learner characteristics, mathematics teacher variables in the classroom environment and the school-level variables. We extend the new avenue of research by developing a linked database geared to the specific mathematics teaching and learning domain to reflect the school mathematics educational environment. The case study using Singapore linkage data demonstrated the feasibility and potential of exploring school effectiveness. In Singapore, schools with teachers of a higher level of education and self-efficacy in teaching mathematics related to a higher level of school mathematics performance. The study offers a guideline and inspiration to the research community to exploit the rich information in both TALIS and PISA studies to facilitate school effectiveness studies.

KEYWORDS

TALIS and PISA, mathematics achievement, educational effectiveness, school climate, multi-level perspectives

Introduction

Educational effectiveness research – Dynamic model of educational effectiveness

Given the growing globalization of education policy and practice, evaluation research focusing on "efficiency" and "effectiveness" of educational outcomes has grown rapidly. Many studies search for the factors playing a role at different levels in the school context (e.g., student background characteristics, quality of instruction, school leadership) as well as at the level of the educational system or regional context (e.g., educational policy). These factors are expected to be associated with students' learning outcomes (e.g., cognitive, affective, psychomotor, and metacognitive) see (Creemers and Scheerens, 1994; Scheerens and Bosker, 1997; Opdenakker and Van Damme, 2000; Creemers and Kyriakides, 2008; Reynolds et al., 2014; Chapman et al., 2015; Kyriakides et al., 2020). Educational Effectiveness Research (EER) takes into account that students are nested within classrooms, that classrooms are nested within schools, and that schools are nested in the region/country context. Student learning outcomes are associated with variables at these multiple levels.

Scholars describe EER as a dynamic process in which multiple levels of the educational system interact, and teaching and learning constantly adapt to changing demands and opportunities, e.g., (Opdenakker and Van Damme, 2006a,b, 2007; Creemers and Kyriakides, 2008; Scheerens, 2013). Over the years, educational researchers tested and developed a more advanced EER model, labeled the "Dynamic Model of Educational Effectiveness" (Creemers and Kyriakides, 2008; Kyriakides et al., 2020).

The Dynamic Model of Educational Effectiveness (DMEE) situates education effectiveness at four nested levels: student, classroom/teacher, school, and system/context. **Figure 1** depicts this DMEE levels hierarchy, which attempts to describe the direct and indirect effects of related factors on a range of student outcomes.

Since teaching and learning are mainly situated at the student and classroom/teacher level, the DMEE also models the interrelationships between student factors (e.g., student background characteristics) and teaching practices. This implies that teachers adjust and apply teaching practices based on the characteristics of students to adapt the teaching to their needs. School factors influence teaching and learning through the implementation of, e.g., a school policy and by the creation of an optimal school learning environment for all. Nonetheless, students, teachers, and schools are agencies within a system or context that is defined by educational policies implemented in their countries, regions, or other functions operating above the school level (Kyriakides et al., 2017). For instance, in highly centralized or decentralized educational systems, the degrees of freedom in defining the learning environment, options for school leaders, or the degrees of freedom in opting for teaching styles, depend on the restrictions imposed by the supra-school level.

Available studies at the national and international levels and meta-analyses tested the validity of the DMEE, with a focus on variables at the different levels, a focus on the measuring dimensions, and a focus on the associations between variables and the learning outcomes, e.g., (White, 1982; Driessen, 2002; Sirin, 2005; Kyriakides et al., 2010, 2013, 2014; Van Damme et al., 2010; Antoniou and Kyriakides, 2013; Scheerens, 2013; Muijs et al., 2014; Panayiotou et al., 2014, 2016). These empirical studies shed light on specific factors that are associated with effective teaching and learning and provide insights to improve educational effectiveness research.

Dynamic Model of Educational Effectiveness (DMEE) highlights that micro-, meso- and macro-level factors are critical when analyzing learning outcomes. Additionally, the model helps in conceptualizing the nature of instructional quality. In this dissertation, DMEE will be applied as the theoretical framework to study mathematical instructional quality and to help in explaining mathematics performance by looking at associated variables at the student level and the school level.

The rise of international large-scale assessments (ILSAs) helped in providing reliable evidence to support policy development and implementation, and can be used to analyze the long-term implications of earlier decisions (Rutkowski et al., 2013; Wagemaker, 2014, 2020). The core feature of ILSAs is the generation of hierarchical data about the home, student, teacher, school, and societal factors to evaluate educational outcomes, develop country profiles, and foster comparison between educational systems (Rutkowski et al., 2013; Wagemaker, 2014). ILSAs provide multiple indicators, covering the student, (teacher) school, and system level of the DMEE, which allow for a decomposition of the variation in outcome measures. The ILSAs contribute to an investigation of educational outcomes both within and across countries and help policymakers learn from other countries (Klieme, 2013). Current ILSAs examples include the Trends in International Mathematics and Science Study (TIMSS), the Progress in International Reading Literacy Study (PIRLS) conducted by the International Association for the Evaluation of Educational Achievement (IEA), and the Program for International Student Assessment (PISA) conducted by the Organization for Economic Co-operation and Development (OECD).

Some large-scale effectiveness studies have already applied the DMEE to measure educational quality and equity at the classroom, school, and system levels. Nilsen and Gustafsson (2016) explained variance linked to school climate when examining the relationship between school climate, teacher quality, and student's learning outcomes in eight-grade

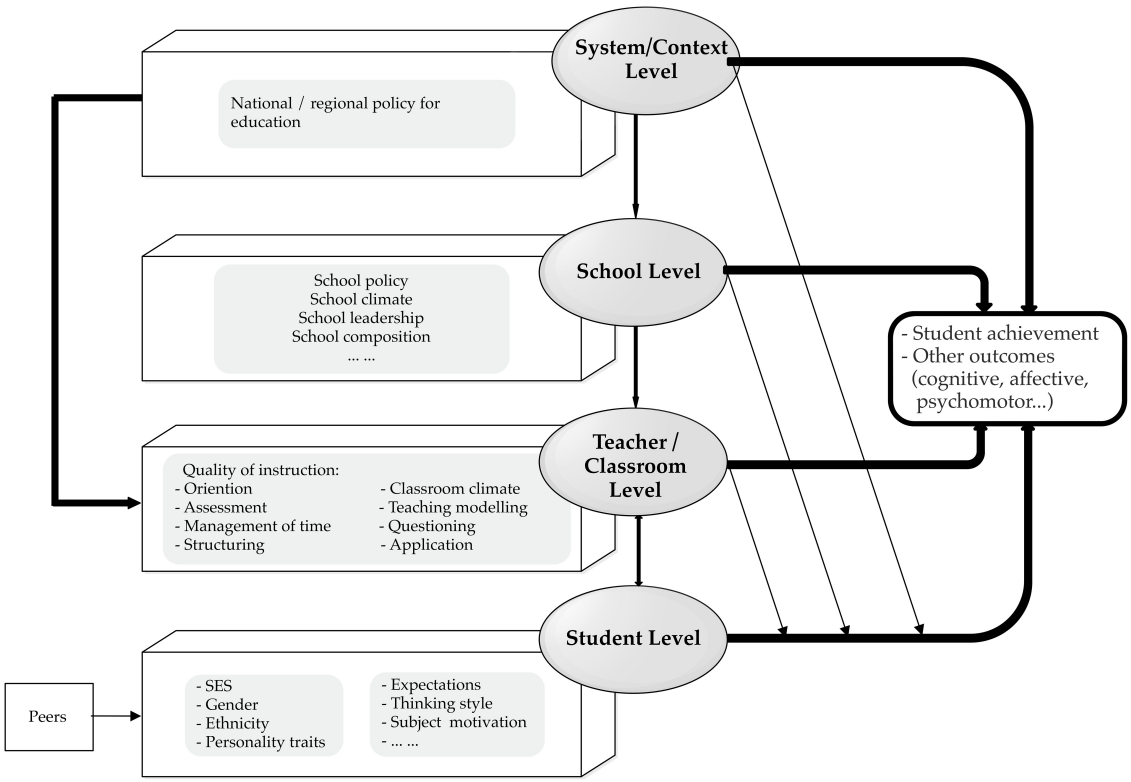


FIGURE 1
The Dynamic Model of Educational Effectiveness (adapted from Creemers and Kyriakides, 2007; Kyriakides et al., 2020).

across 38 countries using TIMSS 2007 and 2011 data. The main findings confirmed a positive and significant relationship between a positive school climate and mathematics outcomes. Meanwhile, teachers' attained education level and professional development were significantly and positively associated with mathematics achievement in grade eight. Other studies did build on PISA data, e.g., (Caro et al., 2016; Martínez-Abad et al., 2020; You et al., 2021). These studies revealed that student-level variables (e.g., socioeconomic status, motivation, enjoyment) and school factors (e.g., school type, school climate, school socioeconomic status) explain a significant proportion of the variation in student achievement.

Connecting mathematics teachers and mathematics performance applying teaching and learning international survey 2013 and programme for international student assessment 2012 linkage data

The PISA data provide insight into the backgrounds, beliefs, attitudes, motivations, mathematics achievement of students, and their perceptions of the learning environment but lack data collected from teachers in their classroom. In addition, the Teaching and Learning International Survey 2013 (TALIS), also set up by the OECD, collects data about the background, characteristics, beliefs, and teaching practices of teachers and their school principals (OECD, 2010). However, the absence of student data and their academic performance does not allow us to measure the association between teacher and teaching characteristics and student performance. This has been solved by the availability of the 2013 PISA-TALIS linkage database. Though a more recent PISA-TALIS linkage database from 2018 is available, the 2013 cycle is still the most recent one focusing on mathematics performance and instruction. Looking at the 2013 linkage database resulting from PISA and TALIS, the single anchor variable to accomplish a link is the school ID (variable "PISASCHOOLID"). This is the sole key variable shared in both TALIS and PISA. This implies that all analyses building on this database has to start from aggregated data at the school level in both TALIS and PISA. The linkage data helps in adding these distinctive teacher-level factors and perspectives (TALIS 2013 data) to the student mathematics performance data from PISA (PISA 2012 data). Moreover, comparisons between countries can center on differences in mathematics instruction, school environments, and education systems. This sounds promising, but much depends on the way we can link the two databases.

Several earlier studies already connected TALIS-PISA data by using the linkage database. These studies can be categorized into three types: (1) examining how school-level profiles of students impact teachers, e.g., (Austin et al., 2015; Sealy et al., 2016), (2) explaining student learning outcomes on the basis of the teacher or school variables at the school level, e.g., (Echazarra et al., 2016; Cordero and Gil-Izquierdo, 2018; Delprato and Chudgar, 2018; Mammadov and Cimen, 2019), and (3) statistical matching and guidelines for data fusion, e.g., (Kaplan and McCarty, 2013; Leunda Iztueta et al., 2017; Gil-Izquierdo and Cordero, 2018; Strietholt and Scherer, 2018). These studies provide – next to empirical evidence about theoretical assumptions – practical information on how to link available TALIS data and PISA data. For instance, a study was conducted by Cordero Ferrera and Gil-Izquierdo (2016). The researchers proposed guidelines for utilizing the original TALIS-PISA Link 2013 data and how this could be further linked to PISA 2012 data. They next studied the relationship between (general) teaching strategies and student mathematics performance in the Spanish context (Cordero and Gil-Izquierdo, 2018; Gil-Izquierdo and Cordero, 2018). Delprato and Chudgar (2018) utilized the linking database to link the variables competitive pressure, school autonomy, and teaching practices when looking at students performing in private and public schools, and this in the context of three countries. Huang et al. (2019) examined the relationships between variables of their school excellence model (e.g., school responsibility, distributed leadership, human resources, material resources) and student achievement in reading, mathematics, and science by applying data from Singapore. Also, three OECD working papers (Austin et al., 2015; Echazarra et al., 2016; Le Donné et al., 2016) focused on the link between student-level factors and teacher variables, between teaching strategies and student's learning strategies, and student PISA mathematics outcomes; in eight countries. An overview of the specific literature using TALIS and PISA linkage data is presented in **Table 1**.

Notwithstanding the availability of these earlier studies, the present study goes further. Firstly, the earlier studies did neglect that the teacher data did originate from different subject teachers. As such, they linked, e.g., data from language teachers to student mathematics outcomes. The PISA TALIS linkage dataset does not differentiate between mathematics and non-mathematics teachers. This raises the question about the adequacy of this choice: Is it possible to use a sample of teachers from other disciplines to convey "mathematical content knowledge" and "mathematical pedagogical content knowledge" to students during instruction? Is it plausible to use students perceived other subject teachers' instructional behaviors to represent their perceptions of "quality of mathematics instruction"? Is it reasonable to use the professional knowledge and instructional behaviors of teachers in other disciplines to explain "mathematical performance"?

Shulman (1986, 1987) highlighted three core categories of teachers' professional knowledge, namely, content knowledge (CK), general pedagogical knowledge (PK), and pedagogical content knowledge (PCK). CK is summarized as a teacher's deep and thorough understanding of the subject matter to be taught, such as the body of knowledge – facts, theories, principles, concepts, and ideas – they should master to be effective. PK refers to the knowledge about teaching and learning that transcends subject matter, such as general theories and principles of classroom behaviors and management, how students are learning, and how best to facilitate that learning in a variety of situations. PCK can be described as the knowledge of specific-subject instructional strategies, the knowledge of representations and explanations, and the knowledge of students' cognitions and (mis)conceptions (e.g., using appropriate strategies to describe ideas, understanding the particular needs of their particular students, providing explanations, making content accessible, setting up tasks to teach subject-matter knowledge). Of course, the three knowledge domains are interconnected. CK leads to teachers knowing what to teach (knowledge of subject matter). PK influences teachers knowing how to teach (general teaching knowledge). Moreover, PCK is the specialized expert kind of knowledge of how to transform subject matter representations "to make content comprehensible to students, combining an understanding of content and pedagogy specifically for instruction (Ball et al., 2005; Ma, 2010; Kleickmann et al., 2017).

In mathematics education, PCK features distinctive subject-specific characteristics. Shulman (1986) and several scholars expanded as such mathematical PCK. This refers to knowledge of the mathematics curriculum, knowledge of the aims of mathematics teaching, and knowledge of the construct of mathematics for teaching and learning (Grossman, 1990; Hill et al., 2004, 2005, 2008; Ball et al., 2008; Blömeke et al., 2012; Senk et al., 2012). Specifically, these components include, for example, conventional mathematical language, mathematical communication, worthwhile mathematical tasks, and making connections links between mathematical topics see (Hunter, 2005; Ainley et al., 2006; Anghileri, 2006; Watson and Mason, 2006; Chapin and O'Connor, 2007). In the case of mathematics teachers, holding a degree in mathematics is expected to ground their solid mathematical professional knowledge. However, also their pedagogical knowledge dimension is to be developed to guarantee that they adopt teaching behavior that leads to the effective delivery of the instructional content.

This critical stance toward the available linking data research in the literature explains the different approaches adopted in the present study. We prefer to interpret mathematics achievement and instructional quality by starting from the unique perspective of mathematics teachers. This implied a redesign of the available linkage dataset by focusing on "mathematics teachers." A second difference with earlier studies building on the PISA TALIS link is that we catered for the bias induced by the time gap between

TABLE 1 Overview of papers using the linkage data from teaching and learning international survey-programme for international student assessment (TALIS-PISA) Link 2013 and PISA 2012.

Category	Author	Brief introduction
School context features impact teachers	Sealy et al., 2016	Examine the relationships between principal job satisfaction, school characteristics, roles of the principal, and student achievement in eight countries.
	Austin et al., 2015	Aggregate student data to the school level to examine how student factors in a school may influence teachers' work, their attitudes, and their perceived needs for support (multilevel regression models).
Teaching strategies and students' learning	Cordero and Gil-Izquierdo, 2018	Examine the different teaching strategies (teacher characteristics, satisfaction of teacher with profession, student management efficacy, school ownership, curriculum, and assessment) on student achievement in Spain. The research is based on an instrumental variable approach.
	Delprato and Chudgar, 2018	Focus on understanding how systemic differences between private and public educational institutions (namely competitive pressure, administrative autonomy, staffing practices, and accountability) can explain differences in students' performance (mathematics, reading, and science) in Australia, Portugal, and Spain.
	Echazarra et al., 2016	Examines how particular teaching and learning strategies are related to student performance on specific PISA test questions, particularly mathematics questions: four teaching strategies– teacher-directed, student orientation, formative assessment, and cognitive activation – and three approaches to learning mathematics – memorization, control, and elaboration strategies.
	Fernández-Díaz et al., 2016	Analyze the relationships between the results from PISA 2012 and those relating to the teaching practice of secondary TALIS 2013, trying to find out the consistencies and discrepancies between the results of both.
	Huang et al., 2019	Investigate the relationships between the key elements of school excellent model variables (e.g., school responsibility, distributed leadership, human resources, material resources) and student achievement in reading, mathematics, and science in Singapore.
	Le Donné et al., 2016	Explore the relationships between mathematics teachers' teaching strategies and student learning outcomes in eight countries: active learning, cognitive activation, and teacher-directed instruction (24 items) at teacher, class, and school levels.
Methodological perspective	Gil-Izquierdo and Cordero, 2018	Guideline of theoretical linkage of TALIS and PISA.
	Leunda Iztueta et al., 2017	Use R software for statistical matching to link the PISA and TALIS studies with Spain's data.

PISA 2012 and TALIS 2013. These time gaps affect the extent to which teachers were teaching in the actual schools sampled in 2013. Some earlier studies neglected teacher mobility and assumed that a one-year time gap did not result in differences in teacher presence at the school level within a country. This assumption might result in less reliable results, and uncontrolled bias. Hence, we added another additional selection criterion to the revised linkage database to ensure that mathematics teachers in our redesigned database did actually work in the schools when the PISA students were studied in 2012. This helped guarantee that the sample of teachers did actually teach PISA 2012 students in the same school, and how their "mathematics professional knowledge" could be associated with a proportion of the variation in "school mathematics performance.

Present study

The above helps to add focus to this study by connecting the topics "mathematics teachers" and "mathematics performance." This brings us to the main focus of the present paper – exploring how to link TALIS 2013 and PISA 2012 data to study the relations between multiple educational effectiveness factors and mathematics achievement as reflected in the dynamic model. The purpose of the linkage is to use school-level data from mathematics teachers' responses in TALIS 2013 to contextualize student performance in PISA 2012 and shed light on how teacher- and school variables explain student achievement. Linking the information from two databases can help identify and explain the relationships between student socioeconomic background, student motivation and attitudes, mathematics teacher background and characteristics, mathematics teaching practices (aggregated at the school level), school compositions, and other school factors (e.g., school leadership, school environment), and school-level profiles of student learning outcomes. This mirrors a multi-level model that might provide insight into what improves student's mathematics learning process and outcomes, how mathematics teachers effectively handle the classroom and motivate their teaching, and how school principals support their teachers and carry out policies in practice. The results of a linked database might additionally be informative for policymakers, school administrators, and teachers themselves (e.g., supporting resources, professional development, teaching quality). Additionally, the linkage allows comparing the results across countries and developing more effective educational policies to improve teaching and student learning.

The general aim of this study is to design a linkage dataset for providing valuable information about multiple mathematics educational factors that potentially infuse future research

about PISA 2012 mathematics performance using a multilevel perspective, especially building on mathematics teacher-related factors. We propose the following research question: Is it feasible to exploit a revised dataset to reflect the school effectiveness in mathematics teaching using the linkage data from TALIS 2013 and PISA 2012?

The current paper is organized as follows. First, we describe the structure of the original TALIS and PISA database and related questionnaires. Secondly, the sample selection criteria are given database redesign, and linkage of the datasets is introduced. Thirdly, a multi-level case study is applied to demonstrate the potential of using this newly designed linked database. Lastly, we address the limitations of linking TALIS and PISA in this way when studying the dynamic model in the context of educational effectiveness research.

Original database: Teaching and learning international survey and programme for international student assessment

TALIS[1] aims to investigate teachers' and school principals' learning environment and working conditions in private and public schools, mainly at the lower secondary education level, by exploring teacher-related factors, examining the roles of school principals, and how they support their teachers (OECD, 2010). PISA involves samples of 15-year-olds from schools – independent of their grade – and focuses on mapping their reading, mathematics, and science literacy. The PISA cycle is repeated every three years and focuses on a different main literacy domain. The PISA measurement framework reflects a skill-orientated and helps to describe mastery of competencies to handle the real-world challenges at the end of – in most countries – the compulsory education cycle (OECD, 2013a, 2017, 2019a; Stacey, 2015).

When implementing the TALIS 2013 cycle, participating countries could apply TALIS to mathematics teachers in a subsample of teachers who participated in the PISA 2012 cycle. This particular option was labeled the TALIS-PISA Link (TPL). The TPL helped start a series of studies examining student mathematics achievement from a multi-level perspective.

The second cycle of TALIS 2013 included 34 countries and economies. Four additional countries and economies administered the survey in 2014, resulting in a total of 38 countries. TALIS 2013 provided data about teachers and school principals, mainly from lower secondary education (ISCED[2] Level 2). Three sampling options were offered: a representative sample of teachers and principals in *option 1* primary education (ISCED Level 1), *option 2* in upper secondary education (ISCED Level 3), and *option 3,* the representative teachers of 15-year-olds and their principals drawn from the schools that already participated in PISA 2012, the so-called TPL mentioned above (OECD, 2009, 2010, 2013a, 2014a, 2019b).

TALIS 2013 collected data based on three questionnaires (see **Figure 2**) filled out by teachers or school principals: the *Teacher General Questionnaire*, the *School General Questionnaire,* and *Mathematics Teacher Questionnaire*. The first covered the teacher background and characteristics (e.g., professional development, teacher self-efficacy, teacher cooperation) and teachers' perspectives about their working environment. The *School General Questionnaire* was filled out by the principals and collected data about the school background and composition, teacher induction and mentoring, formal teacher appraisal, school autonomy, school leadership, a principal's background and job satisfaction, and school climate (e.g., school delinquency and violence, mutual respect).

In countries that signed up for the third sampling option (TPL), after completing the *Teacher General Questionnaire*, all mathematics teachers were additionally asked to complete the *Mathematics Teacher Questionnaire*. This helped identify specific data about their mathematics classes and instructional school climate (OECD, 2013b, 2014c). Sampling *option 3* comprised next to all mathematics teachers of a school, 20 non-mathematics teachers and one school principal of each of the 150 schools in an *option 3*-country (OECD, 2014c). Eight countries opted for the TALIS-PISA Link approach: Australia (AUS), Finland (FIN), Latvia (LVA), Mexico (MEX), Portugal (PRT), Romania (ROU), Singapore (SGP), and Spain (ESP). The TALIS-PISA Link helped to center on teaching practices in the target class[3], mathematics teaching strategies, educational approaches, initial training/education for teaching mathematics, and self-efficacy in teaching mathematics. TALIS-PISA Link offers a school-level perspective on mathematics instructional quality from TALIS 2013 that can be linked to student-level data from PISA 2012.

1 Three cycles of TALIS had been conducted in 2008, 2013, and 2018. The first cycle was conducted in 2008 and involved 24 countries. The second cycle was in 2013 and involved 34 countries and economies. Another four countries and economies were administrated in 2014.The third cycle was in 2018 and involved 48 countries and economies.

2 Classification of levels of education is based on the International Standard Classification of Education 1997: pre-primary education (ISCED level 0), primary education or first basic education (ISCED level 1), lower secondary education or second stage of basic education (ISCED level 2), upper secondary education (ISCED level 3), post-secondary non-tertiary level of education (ISCED level 4), the first stage of tertiary education (ISCED level 5), thee second stage of tertiary education (ISCED level 6).

3 Target class: Considering the teaching practices in the class, TPL selected a necessary "target class" to finish the mathematics module about *Mathematics Teacher Questionnaire*. "Target class" was composited of the majority of PISA-eligible "15-year-old" students in the class and identified as the first-class attended by 15-year-old students teachers taught in the current school year in TPL.

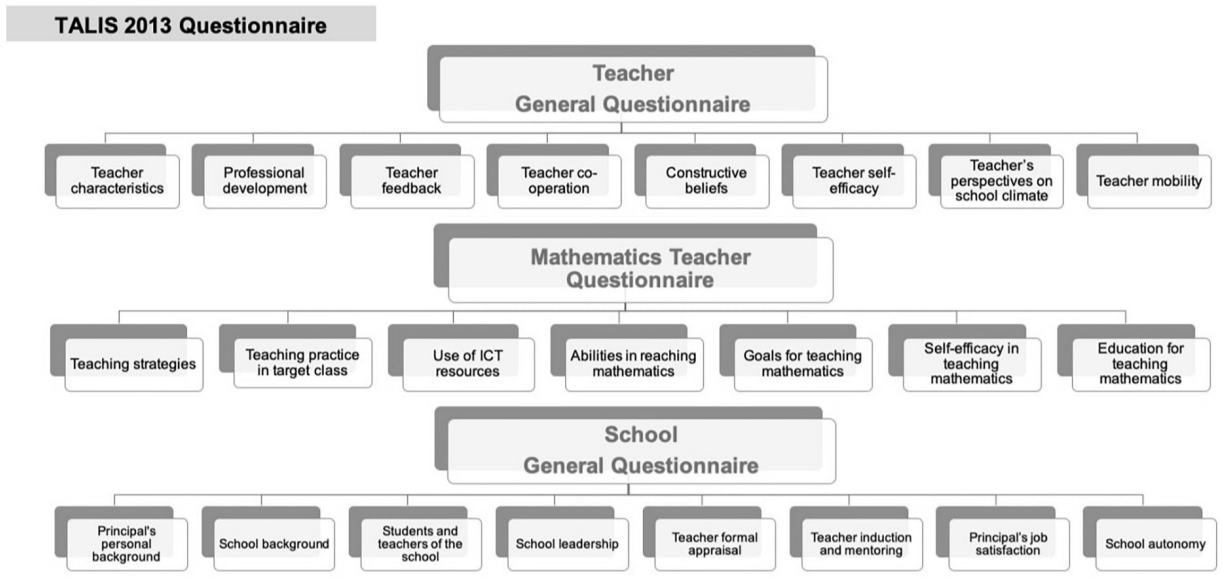


FIGURE 2
The main aspects of the teacher and school questionnaire in teaching and learning international survey 2013.

PISA 2012 was the fifth cycle and covered reading, mathematics, science, problem-solving and financial literacy, with mathematics as the primary domain (OECD, 2013a). PISA 2012 data was collected with three questionnaires: the *Student Questionnaire,* the *School Questionnaire,* and the *Parent Questionnaire*.

The *Student Questionnaire* focused on student characteristics, family background, personal intrinsic factors, student perspectives on the learning environment, teaching practices and school climate. The *School Questionnaire* – filled out by school principals – looked into school background information, school climate, school leadership, school curriculum assessment, school mathematics policies, and instructional practices. In 11 countries, also the *Parent Questionnaire* was administered to collect data about parents' background, their attitudes toward school, parent support for learning in the home, mathematics in the job market, children's past academic performance and academic and professional expectations in the field of mathematics (OECD, 2013a). Around 510,000 students, aged 15 years three months to 16 years two months, from 65 countries participated in PISA 2012: 34 OECD countries and 31 partner countries and economies. The main aspects of the student questionnaire and school questionnaire in PISA 2012 are summarized in **Figure 3**.

In the PISA 2012 Questionnaires, only limited data about teachers are being collected, and therefore large parts of the EER dynamic model about teaching effectiveness cannot be studied directly. TALIS offers a rich database to study the dynamic model in full by focusing on original teacher self-reported information. But this requires linking both separate datasets. The linkage will be established at the school-level since the only anchor variable shared in both databases is the school ID – PISASCHOOLID.

Tables 2, **3** summarize the number of schools and teachers in TALIS-PISA Link and students sampled from schools for PISA 2012 of the eight participating countries.

Redesigning the teaching and learning international survey-programme for international student assessment link database

When considering the linkage of TALIS-PISA Link 2013 and PISA 2012, critical issues need to be addressed. Firstly, the key sampling variable differs in TALIS and PISA. The TALIS-*Teacher General Questionnaire* builds on "grades" (i.e., ISCED Level 1, ISCED Level 2, and ISCED Level 3). However, the PISA *Mathematics Teacher Questionnaire* starts with teachers teaching students "the age of 15 years" (OECD, 2013b, 2014b,c). Linking data from both TALIS and PISA requires focusing on students from the same age group.

Secondly, TALIS teacher data cannot directly be linked to PISA individual student data (OECD, 2013b, 2014b,c; Le Donné et al., 2016). In other words, it is not possible to link a student to her or his personal mathematics teacher. In both databases, there is only one single anchor variable that is shared: the ID of the school (variable "PISASCHOOLID"). In view of linking

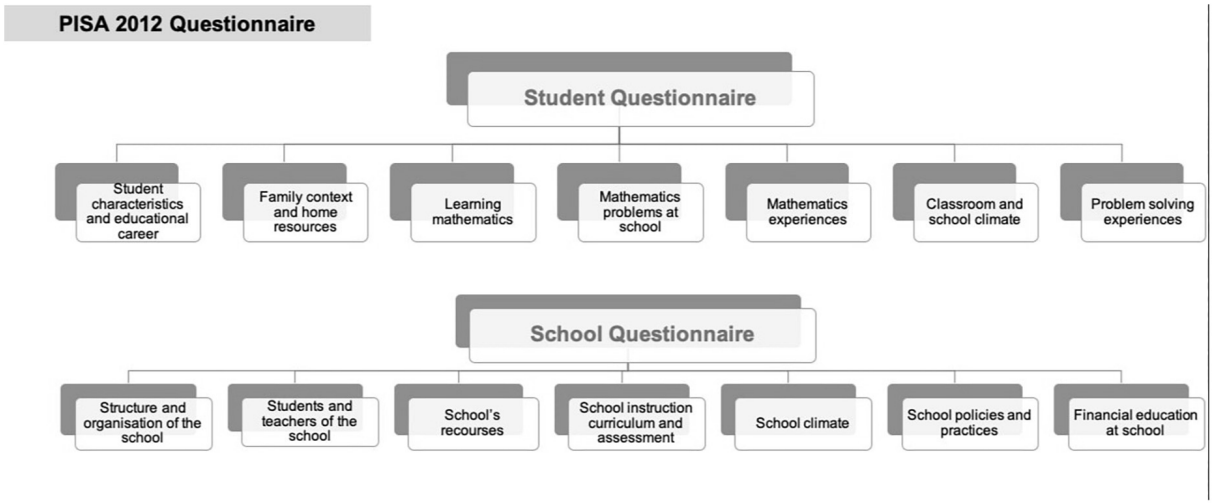


FIGURE 3
The main aspects of and student and school questionnaire in programme for international student assessment (PISA 2012).

TABLE 2 Overview of the original raw data of TALIS-PISA Link 2013 samples.

	AUS	FIN	LVA	MEX	PRT	ROU	SGP	ESP	Total
Number of schools for TALIS-PISA LINK	122	147	118	150	141	147	166	310	1 301
Respondent teachers in schools for TALIS-PISA Link	2 719	3 326	2 123	2 167	3 152	3 275	4 130	6 130	27 022

Source from OECD TALIS 2013 Database. AUS, Australia; FIN, Finland; LVA, Latvia; MEX, Mexico; PRT, Portugal; ROU, Romania; SGP, Singapore; ESP, Spain.

TABLE 3 Overview of the original raw data of and PISA 2012 samples.

	AUS	FIN	LVA	MEX	PRT	ROU	SGP	ESP	Total
Number of schools sampled for PISA 2012	775	311	211	1 471	195	178	172	902	4 215
Participating student sampled for PISA 2012	14 481	8 829	4 306	33 806	5 722	5 074	5 546	25 313	103 077

Source from OECD PISA 2012 Database. AUS, Australia; FIN, Finland; LVA, Latvia; MEX, Mexico; PRT, Portugal; ROU, Romania; SGP, Singapore; ESP, Spain.

the datasets, data have to be aggregated at the school level. This implies that no classroom-level information is available in the new dataset, but the average teacher and student factors in a school.

Thirdly, the administration of TALIS 2013 questionnaires occurred nearly one year after administering the PISA 2012 instruments. TALIS 2013 was conducted from September to December 2012 in Southern Hemisphere countries and from February to June 2012 in Northern Hemisphere countries. Whereas the Southern Hemisphere countries (AUS, SGP) developed PISA 2012 between May and August 2012, the Northern Hemisphere countries (FIN, LVA, MEX, PRT, ROU, ESP) were between March to May 2012 (OECD, 2014c; Echazarra et al., 2016). This resulted in a time gap that could create a misfit between teachers and students within the same school. To cater for this time gap, additional criteria were applied to refine teacher selection in view of a revised link dataset: a teacher should have at least one year of work experience in the Southern hemisphere and at least two years of work experience in the Northern hemisphere. In this way, we increased the probability to map the data from the actual teachers and students who participated in PISA 2012 with the data of teachers who participated in TALIS 2013.

Fourthly, since we focus on student mathematics achievement, the revised link database should solely center on data from mathematics teachers from the TALIS-PISA 2013 study. At the same time, we focused on mathematics literacy performance and related data from the PISA 2012 study.

Considering a linking procedure, Le Donné et al. (2016) proposed two approaches: either (A) PISA student data are aggregated at the school level and next merged with TALIS data; or (B) TALIS teacher data are aggregated at the school level and next merged with PISA data. Also, Gil-Izquierdo and Cordero (2018) see the two databases as potentially different "donor" or "recipient" datasets, and how this reflects a different merging approach: (a) TALIS as the recipient dataset and merging PISA

data into TALIS based on the same PISASCHOOLID; (b) TALIS as the donor dataset and PISA as a recipient dataset that are merging TALIS data into PISA based on the same PISASCHOOLID.

One could state that (A) and (a) is equivalent to examining teacher outcomes (e.g., professional development, beliefs about teaching, self-efficacy) in the learning environment by measuring some constructs based on student's self-reported (e.g., learning motivation, attitudes toward school, teacher and student relation) in PISA (Austin et al., 2015). On the other hand, (B) and (b) can be seen as equivalent to evaluating student achievement depending on teachers' characteristics, teaching practice, and educational approach in the classroom. Making a choice for either approach depends on the nature of the research question being addressed.

Since we aim to use the redesigned linking dataset to analyze student-level data (mathematics literacy) by considering the teacher and school-level data, we opted for the second approach with PISA as a recipient dataset and merge the TALIS donor-data into PISA. The teacher information was aggregated at the school level before being merged into the student dataset. The resulting dataset structure fits the multi-level perspectives as reflected in the Dynamic Model of Educational Effectiveness (Creemers and Kyriakides, 2007). The resulting redesigned dataset consists of data organized at the individual student and school level from PISA 2012, the school profile of teacher factors from TALIS 2013, and school factors from both PISA 2012 and TALIS 2013.

Building on the above rationale, a Redesigned TALIS-PISA Link database (rTPL) was created to link "mathematics teacher" data to "student mathematics achievement" data resulting from the two original data sets. This rTPL reflects the following teacher sampling criteria (see **Figure 4**):

- Teacher data are from teachers with at least one year of work experience in the Southern hemisphere and at least two years of work experience in the Northern hemisphere.
- Teachers did teach mathematics to 15-year-old students in the test administration school year.
- The selected teachers did teach mathematics in the target class: the "target class" contains potential PISA pupils. In this way, teacher factors can be linked to pupils and their math performance.
- The teachers did fill out the *Mathematics Teacher Questionnaire*.
- The teacher was, as such, also a PISA mathematics teacher.

The "redesigned TALIS-PISA Link database" (rTPL) consisted of data from 3473 valid teachers from 1115 valid schools and representing 31,548 students from schools with matching PISASCHOOL ID in the TALIS-PISA Link and PISA 2012 (see **Table 4**).

Feasibility of using the redesigned teaching and learning international survey-programme for international student assessment link database

Next to the design of the rTPL database, the present article explores the feasibility of using the rTPL to test complex EER related multi-level models. Such a model builds on the theoretical assumptions from the Dynamic Model of Educational Effectiveness and the Opportunity-Propensity framework. The Opportunity-Propensity (O-P) framework has been put forward to explain associations with student performance; see **Figure 5**. Three main categorical predictors are presented in the model. These include antecedent factors, opportunity factors, and propensity factors (Byrnes, 2003, 2020; Byrnes and Miller, 2007; Byrnes and Wasik, 2009; Byrnes and Miller-Cotto, 2016). The antecedent factors are related to aspects of a students' home environment and socio-cultural demographics, including socioeconomic status, gender, race, ethnicity, and parental expectations for their children's academic achievement. The opportunity factors comprise aspects of the learning context (i.e., at home and school) that promote learning and development, such as content exposure, teaching strategies, and overall instructional quality. The propensity factors are related to a student's ability and willingness to learn in a particular context (e.g., prior knowledge, academic motivation, cognitive level).

According to the O-P framework, antecedent factors operate earlier and already lead to variations in opportunity factors and propensity factors. For instance, students from high-socioeconomic families are financially able to relocate to neighborhoods with schools that employ more qualified and effective teachers, receiving high-quality instruction (opportunity factor) while being able to mobilize high-level prerequisite knowledge. Hence, the academic achievement and development outcomes vary between students.

The empirical evidence is, for instance, abundant with studies linking teacher characteristics to student achievement. For example, teacher self-efficacy is an essential teacher characteristic and has been found to be strongly associated with the quality of instruction (Holzberger et al., 2013). In turn, effective teaching is a vital characteristic of high-performing schools, mirroring high student achievement and other educational outcomes (Muijs and Reynolds, 2002; Caprara et al., 2006). Meta-analysis studies from Hattie (2008) and others, e.g., (Desimone et al., 2002; Snow-Renner and Lauer, 2005) reiterate consistently that teachers' professional development may have the strongest impact on teachers' learning. Effective professional development seems to increase teacher self-efficacy and their instructional beliefs (Robardey et al., 1994; Rimm-Kaufman et al., 2006; Tschannen-Moran and McMaster, 2009) and with

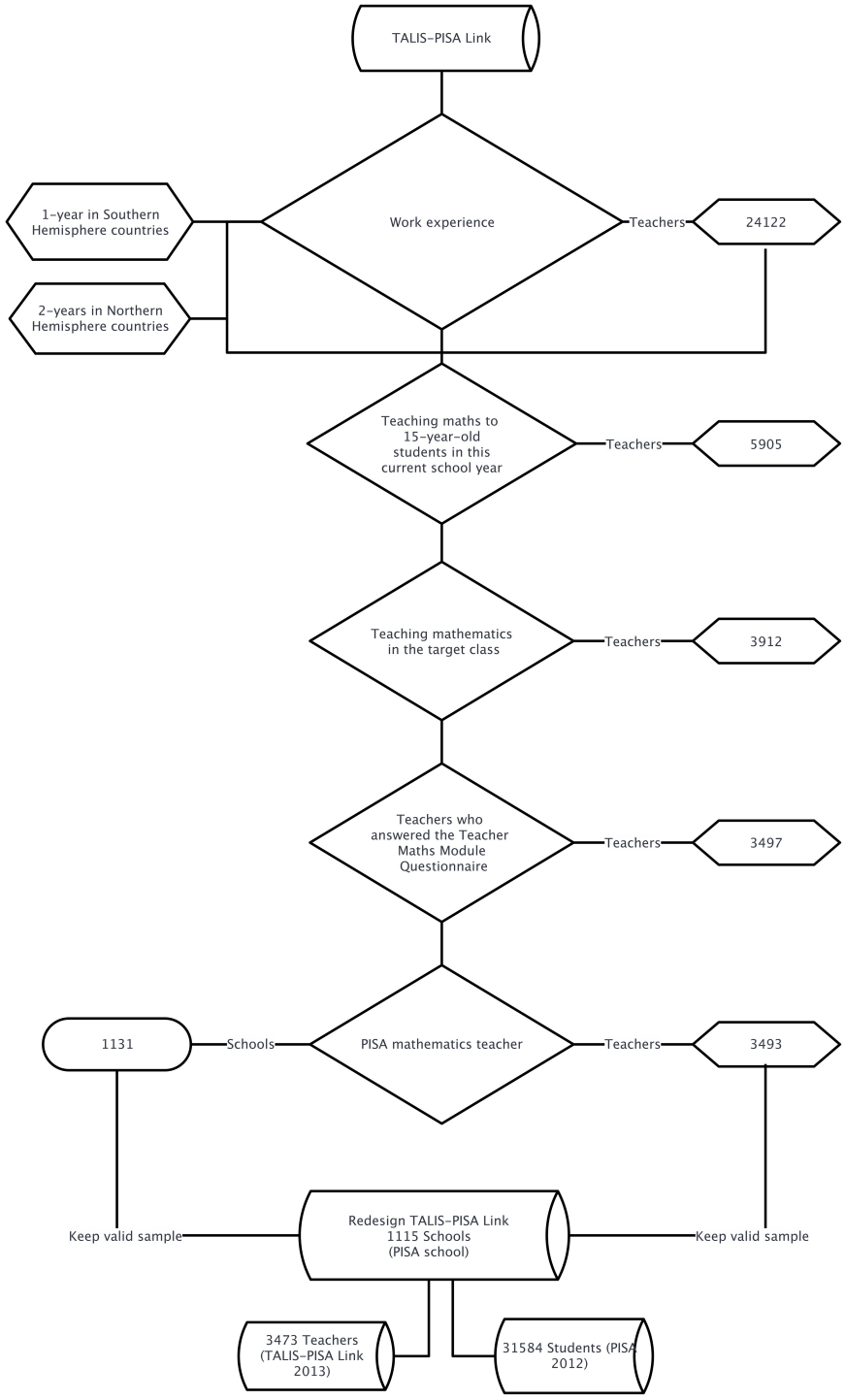


FIGURE 4
The teacher sample selection criteria in the redesigned TALIS-PISA Link database (rTPL).

the identification of strong effects on student achievement (Borko and Putnam, 1995; Timperley et al., 2008). Teacher cooperation seems to be a powerful form of professional development and is regarded as a vital facet of teacher professional practices in the school environment (Goddard et al., 2007; Timperley et al., 2008; Desimone, 2009). In the process of professional communicating and sharing among teachers, improvement-oriented changes seem to develop

TABLE 4 Overview of redesign TALIS-PISA Link database (rTPL).

	AUS	FIN	LVA	MEX	PRT	ROU	SGP	ESP	Total
1-year (S) / 2-year work (N) at same school	2686	2873	1980	1723	2670	2977	4066	5147	24122
Teaching maths to 15-year-old students in this current school year	853	856	315	405	616	516	1163	1181	5905
Teaching maths in the target class	528	370	225	212	569	420	776	812	3912
Teachers who answered the Maths Module Questionnaire	419	332	191	175	537	392	719	732	3497
Pisa Maths Teacher	415	332	191	175	537	392	719	732	3493
Number of teachers in XX Schools	415 in 113	332 in 133	191 in 94	175 in 92	537in 131	392 in 133	719 in 164	732in 271	3 497 in 1 131
Valid number of teachers in the valid same PISASCHOOL	415 in 113	332 in 133	178 in 85	170 in 87	537 in 131	390 in 131	719 in 164	732 in 271	3 473 in 1 115
Number of students in the valid same PISASCHOOL	2 251	4 010	2 013	2 151	3 886	4 103	5 302	7 868	31 584

Source from OECD TALIS 2013 Database and PISA 2012 Database. AUS, Australia; FIN, Finland; LVA, Latvia; MEX, Mexico; PRT, Portugal; ROU, Romania; SGP, Singapore; ESP, Spain.

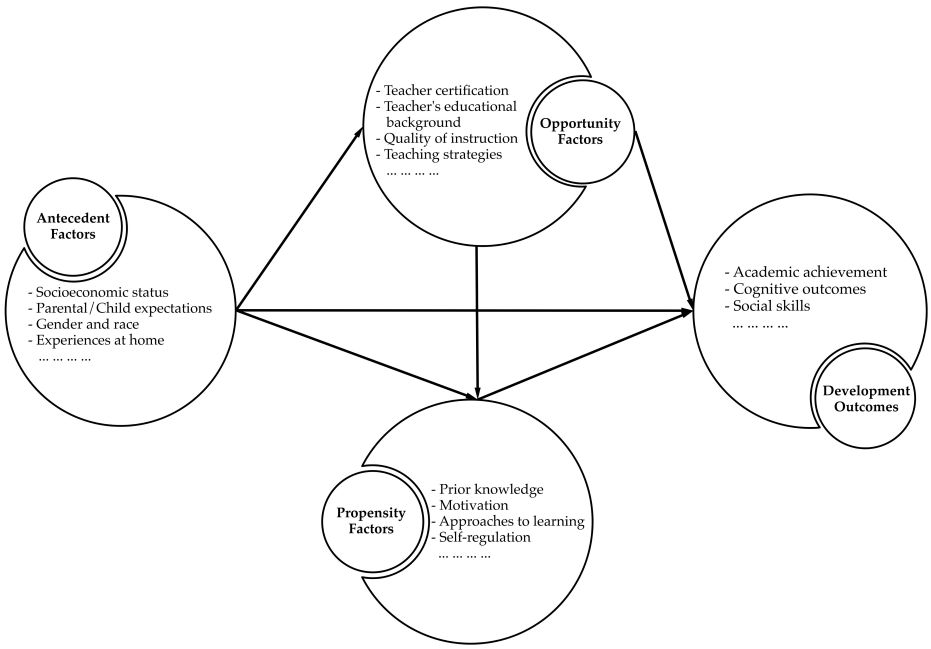


FIGURE 5
The original and the latest combined version of the of Opportunity-Propensity framework (adapted from Byrnes and Miller, 2007; Byrnes and Wasik, 2009; Byrnes and Miller-Cotto, 2016; Byrnes, 2020).

from an evolving knowledge base, professional development, and teacher self-efficacy (Garet et al., 2001; Erickson et al., 2005).

In addition to teacher variables, adding student variables helps look at a more complex way to EER. Considering the socioeconomic status (SES), a meta-analysis study from Sirin (2005) integrated 58 studies published between 1990 and 2000, underpinning the association between SES and academic achievement. A longitudinal study – based on a ten-year window – by Yang Hansen et al. (2011) examined the relations between SES and reading achievement at the individual and school level in Sweden. They found that school differences were highly related to SES differences in 2001, and SES differences did explain more than half of the average reading attainment variation at the school level in 2001, compared to about 30% in 1991. Muijs and Reynolds (2003) analyzed the relationship of SES, classroom social context, classroom organization, teacher behavior and mathematics achievement. Teacher behavior was the strongest performance predictor and was significantly related to student achievement, explaining over 5.6% of the

total variance, while individual student background variables explained 3% of the variance in student academic performance. Opdenakker et al. (2002) applied multi-level analyses to examine the associations between SES, gender, average class SES, learning environment, and mathematics attainment at different levels. They concluded that learning environment factors mediated the relationship between individual variables and mathematics attainment. When researching educational effectiveness, plenty of studies suggest that higher-level factors should be considered, such as at the classroom-, teacher- or school-level (Hattie, 2002; Van Ewijk and Sleegers, 2010; Kelly, 2012; Creemers and Kyriakides, 2015; Hornstra et al., 2015; Verhaeghe et al., 2018).

As explained earlier, the present study tests the feasibility of the rTPL against this background by focusing on identifying the relationships between school-level profiles of teacher characteristics (e.g., self-efficacy, beliefs, cooperation) and how each contributes to student mathematics achievement. Additionally, variables mapping socioeconomic status, teacher qualifications (i.e., years of experience, educational background), and school climate (i.e., school size, mutual respect) are used as control variables at the school-level in the analytical procedure. To illustrate the potential of the rTPL database in testing such a model, we selected the Singaporean rTPL data as a case study.

Variables in testing the use of redesigned teaching and learning international survey-programme for international student assessment link

According to the DMEE and the O-P framework, and the literature supports, we selected the variables of student socioeconomic status (PISA 2021 data) and teacher and school characteristics (TALIS 2013). As explained in the former section, TALIS 2013 indicators were aggregated at the school level: *teacher educational background, teacher work experience*, *teacher self-efficacy*, *self-efficacy in teaching mathematics, teacher cooperation, effective professional development*, and *constructivist beliefs*. Since no variation was found in the variables of *school composition (i.e., public or private school systems, school location)*, or the variables of *teacher gender and age* between schools in Singapore, they were excluded from the study.

Teacher self-efficacy (TSELEFFS) was defined on the base of three subscales with efficacy in classroom management, efficacy in instruction, and efficacy in student engagement. All scales are built on four-point Likert items, with response categories ranging from "not at all" to "a lot."

Self-efficacy in teaching mathematics (TMSELEFFS) was derived from the TPL instruments presented to mathematics teachers. This is different from the indicator *teacher self-efficacy* and built on statements about teachers' ability to teach mathematics. The four scale items were based on a four-point Likert scale, with response categories ranging from "strongly disagree" to "disagree," "agree" and "strongly agree."

The composite scale, *teacher cooperation (TCOOPS)* consisted of two subscales that centered on exchange and coordination in view of teaching and professional collaboration. Eight six-point Likert scale items were presented with response options ranging from "never" to "once a year or less," "2-4 times a year," "5-10 times a year," "1-3 times a month" and "once a week or more."

Teacher effective professional development (TEFFPROS) focused on the opportunities for active learning and collaborative learning activities or research with other teachers. The four four-point items response options ranged from "not in any activities" to "yes, in all activities."

Constructivist beliefs (TCONSBS) were mapped with four four-point scale items, with response categories ranging from "strongly disagree" to "strongly agree." This index concerned teacher personal beliefs on teaching and learning.

The indicator of *mutual respect (PSCMUTRS)* consisted of four items: school staff have an open discussion about difficulties, mutual respect for colleagues' ideas, a culture of sharing success and the relationships between teacher and student. Items required a response on the base of a four-point scale with response categories ranging from "strongly disagree" to "strongly agree."

The PISA 2012 index of *student economic, social and cultural status (ESCS)* was defined at the student and school level and consisted of three subscales: *the highest parental occupation (HISEI), the highest parental education expressed as years of schooling (PARED),* and *the home possessions (HOMEPOS)*. The HISEI index was coded on the base of ISCO-08 and next mapped onto the international socioeconomic index of occupational status (ISEI) (Ganzeboom, 2010), students' responses to PARED were classified using ISCED (United Nations Educational Scientific and Cultural Organization [UNESCO], 2003).

Other variables included *school size (SCHSIZE)* and the first plausible value for *student mathematics achievement (PIVMATH)* in PISA. PISA 2012 datasets include five plausible values (PV1MATH, PV2MATH, PV3MATH, PV4MATH, PV5MATH) in relation to mathematics literacy, computed by administering 34 mathematics items. It is essential to understand that plausible values are not actual test scores. "They are random numbers that were taken from the distribution of scores that could be reasonably assigned to each individual. Plausible values contain random error variance components and are not as optimal as scores to be used as an indicator of individual student performance. Plausible values are rather suited to describe the performance of the population" (OECD, 2014a). The PISA 2012 plausible values were equated to the PISA scale by utilizing common item equating. In our analytical procedure, the five-combined plausible values and the first plausible value have initially been used and compared. When

combined values were used, the separate results of the model parameters across the five datasets were combined using the command TYPE = IMPUTATION in Mplus. The results showed that there was no substantial difference in using multiple plausible values or the first value, either at the individual or the school level. To facilitate the operation of the analysis process in MPlus and its subsequent interpretation, we have used only the first plausible value. Therefore, our further analyses were based on the first plausible value – PV1MATH – as the indicator of individual students' mathematics achievement.

For more detailed information about each scale, see the PISA 2012 technical report (OECD, 2014a) and TALIS 2013 Technical Report (OECD, 2014c).

Analytical methods

Multi-level Path Analysis was applied using Mplus 8.4 (Muthén and Muthén, 2017). The Maximum Likelihood Estimator with robust standard errors (MLR) was used to handle missing and non-normal data. Chi-Square statistics with the degree of freedom and other goodness-of-fit indices (e.g., RMSEA, CFI and SRMR)[4] were used to evaluate whether the model fits the data. When the cut-off value for CFI is greater or equal to 0.95, for RMSEA being less than 0.06, and for SRMR being less than 0.08, the model can be regarded as an acceptable fitting model (Hu and Bentler, 1999).

The interaction correlation coefficient (ICC) is a key tool to check whether the model structure impacts the outcome variable by grouping clusters in multi-level modeling. It also represents the correlation between randomly selected individuals in the same group (Hox et al., 2017). An ICC value exceeding 0.05 indicates that a multi-level structure is needed to model the data (Dyer et al., 2005). R-square represents the proportion of the variance in the dependent variable that is explained by the independent variable and therefore reflects the capability of the model and the predictors to explain a proportion of the variance in the outcome of interest (Finch and Bolin, 2017).

Analytical process

Analysis of variance (ANOVA) model 1 helped estimate the variance within the individuals (σ^2_w) and between the clusters (σ^2_B). These values are used to estimate ICC (ρ), as in Equation (1),

$$\rho = \sigma^2 B/(\sigma^2 w + \sigma^2 B) \quad (1)$$

In the Random Intercept with Level-1 Predictor model (model 2), the subscript *i* refers to the individual in the *j* school-cluster; ε_{ij} and μ_{oj} are error terms at Level-1 and Level-2; β_{oj} is the intercept of achievement for each school; γ_{00} represents an average intercept value across schools. The predictors of student variables in PISA, *student economic, social and cultural status (ESCS)* and *mathematics achievement (PV1MATH)* were added at the student-level. We estimated the values for the two fixed effects of *level-2 PV1MATH* (γ_{00}) and *ESCS* (γ_{10}) as well as for the residual variance of *PVIMATH* (μ_{oj}) and other predictors (ε_{ij}). The equation for model 2 is given by:

$$PV1MATH_{ij} = \gamma_{00} + \gamma_{10} ESCS_{ij} + \mu_{oj} + \varepsilon_{ij} \quad (2)$$

In model 3, we added school-level variables to ascertain how much variation in *PVIMATH* was present across schools. Specific TALIS and PISA data were entered in this model. Student *ESCS* (PISA 2012) was used as a predictor at the student-level. School size (*SCHSIZE*) and school *ESCS* in PISA 2012, teacher characteristics (e.g., *TSELFEFFS, TCOOPS, TEFFPROS, TMSELEFFS, TCONSBS*) and mutual respect (*PSCMUTRS*) in TALIS 2013 were entered as predictors at the school-level.

Results

The ANOVA model results help estimate the variance of student mathematics achievement. This is 0.407 and 0.708 at the individual- and between-level, respectively; thus, the value for ICC is estimated as 0.37 based on Equation 1. The value indicates that the correlation of the mathematics achievement among students within the same schools is 0.37, and about 37% of the variability of student mathematics achievement can be explained by schools' diversity in Singapore.

The goodness-of-fit indices of model 2 are satisfactory: CFI = 0.95, RMSEA = 0.03, and SRMR = 0.04. The estimated slope for *ESCS* is 0.20 and is significantly associated with *PV1MATH,* indicating that as ESCS score increased by 1 point, the mathematics achievement shows an associated increase by an estimated 0.20 points.

Figure 6 presents the results for model 3, with the indices CFI being 1.00, RMSEA = 0.01, within-level SRMR = 0.00, and between-level SRMR = 0.02. At the individual student-level, the indicator of *student economic, social and cultural status (ESCS)* reflects significant and positive associations with mathematics achievement considering an estimated slope of 0.21. The R-square of outcome variable PVIMATH is about 0.05 at the individual student-level, stating that around 4% variation of student mathematics achievement can be explained within the schools, accounting for 1.9% (ICC * R^2 = 0.37*0.04) of the total variance (ICC = 0.37).

At the school level, as shown in **Figure 6**, four indicators do positively and directly contribute to student achievement:

4 RMSEA is an absolute measure of model fit, which stands for Root Mean Square Error of Approximation. CFI is short for Comparative Fit Index. Both RMSEA and CFI pay the penalty for model complexity. SRMR (Standardized Root Mean Square Residual) measures the absolute model fit.

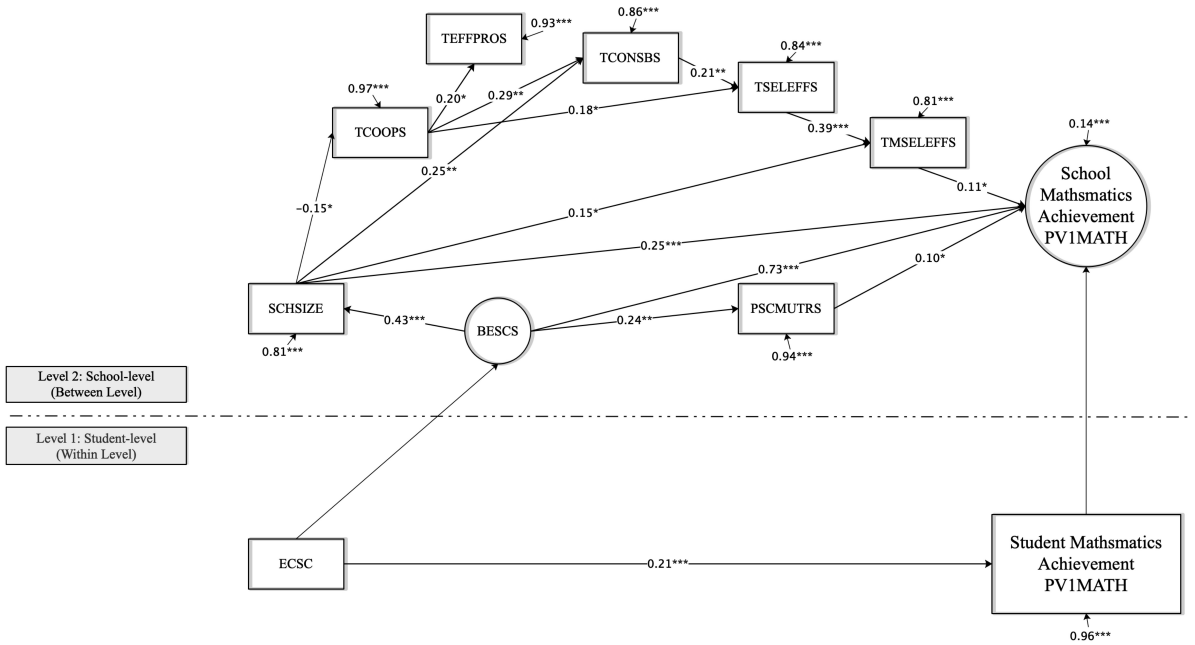


FIGURE 6
The standardized model results of the relationships between teacher characteristics and student mathematics achievement at the student- and school-level. ***$P < 0.001$, **$P < 0.01$, *$P < 0.05$.

school SES *economic, social and cultural status (BESCS, 0.73), self-efficacy in teaching mathematics (TMSELEFFS, 0.11), school size (SCHSIZE, 0.25)* and *mutual respect (PSCMUTRS, 0.10)*. It is interesting to observe that a higher mutual respect working environment *(i.e., school staff have an open discussion about difficulties, mutual respect for colleagues' ideas, a culture of sharing success, and the relationships between teacher and student)* is associated with higher academic performance. The predictors explain 86% of the variation in mathematics achievement between schools, accounting for about 32% (ICC * R^2 = 0.37*0.32) of the total variance (ICC = 0.37).

In Singapore, teachers working in relatively small schools (-0.15) prefer cooperating with other colleagues. In turn, *teacher cooperation (TCOOPS)* is positively correlated with *teacher self-efficacy (TSELEFFS, 0.18), effective professional development (TEFFPRO, 0.20S)* and *constructivist beliefs (TCONSBS, 0.29)*. *Teacher self-efficacy (TSELEFFS)* helps predict *constructivist beliefs (TCONSBS)*, with a standardized coefficient of 0.21. As a result, *teacher self-efficacy (TSELEFFS) is* directly associated with *self-efficacy in teaching mathematics (TMSELEFFS, 0.39)*.

Also, *school size (SCHSIZE)* seems significantly and positively correlated with *constructivist beliefs (TCONSBS, 0.25)* and *self-efficacy in teaching mathematics (TMSELEFFS, 0,15)*. *School economic, social and cultural status (BESCS)* positively contributes to *school size (SCHSIZE, 0.43)* and *mutual respect (PSCMUTRS, 0.24)*.

In summary, the results in **Figure 6** show the direct and indirect factors that are significantly related to mathematics achievement in Singapore. Student mathematics achievement vary according to students with different socioeconomic status. School socioeconomic status, school size, school collective mathematics teachers' teaching self-efficacy, and mathematics teacher mutual respect are positively and significantly related to the school's mathematics performance.

Discussion

The present study aimed to study the educational effectiveness from a multi-level perspective by building on a newly designed linkage database, connecting student and teacher data collected *via* TALIS and PISA. Applying the linkage dataset was expected to help unravel the interconnections between students' mathematics performance while considering individual learner characteristics, mathematics teacher variables in the teaching/classroom environment, and the school-level variables. As stated earlier, this requires new and adequate teacher sampling procedures.

Building on the rTPL based analysis results, our findings help operationalize specific school variables, teaching style elements, and culture-related constructs that play a significant role. These – exemplary – findings could become ingredients to inspire instructional policies to foster quality measures at the different levels in the model. This could also, on the

one hand, promote a school mathematics culture and related instructional approaches in view of improving mathematics teaching effectiveness and student learning outcomes. On the other hand, this could also foster between-country comparison to identify explanatory variables building on differences in mathematics curricula, school context, and educational systems.

The case study analysis results demonstrate the feasibility and potential for linking TALIS and PISA. The findings suggest that, in Singapore, schools with highly educated teachers and higher self-efficacy teachers in teaching mathematics contributed to improving schools' mathematics performance. The results of the case study will not be discussed in-depth. Still, nevertheless, some aspects are noteworthy since they complement previous research and further illustrate the potential of the rTPL database. This is tackled in the next paragraphs.

Available research considers teacher background and teacher characteristics as critical differences between teachers in classrooms (Fraser, 2013; Creemers and Kyriakides, 2015). However, studies rarely focus on looking at the effects of these differences on student learning outcomes. Even the Dynamic Model of Educational Effectiveness primarily concentrates on teaching activities (e.g., classroom management of time, classroom climate, teaching-modeling, assessment) to study student learning outcomes (Kyriakides et al., 2020). Teacher background and teacher characteristics are mostly approached as teacher-level input variables when studying teaching effectiveness/instructional quality, e.g., (Scheerens, 2007; Creemers and Kyriakides, 2015).

International large-scale assessments have the potential to boost multi-level analysis studies that fit state-of-the-art educational effectiveness models. Nevertheless – as tackled in the present paper – this potential is often flawed by methodological constraints in the data available for the studies. The present paper stated solutions, procedures, and strategies to develop overarching databases that link datasets from earlier studies; more specifically, TALIS 2013 and PISA 2012. Using the redesigned TALIS-PISA Link (rTPL) dataset to evaluate student mathematics achievement in Singapore, we provided the feasibility of developing a multi-level perspective from the teacher self-reported data and the student survey. Compared to available studies linking TALIS and PISA data, we extended this new avenue of research by developing a linked database that is geared to the specific mathematics teaching and learning domain. The rTPL considered specific inclusion and exclusion criteria to construct a better fitting database to reflect the school mathematics educational environment.

The further potential of the rTPL is to center between-country comparisons when explaining differences in learning performance. International large-scale assessments studies suggest that the theoretical constructs are "universal" and apply to all countries. This introduces the question of whether relationships put forward in specific national contexts do hold in other countries. International comparison studies might help identify factors associated with differences in mathematics achievement in each country and test measurement invariance to check the comparability in the eight national contexts that are contained in the rTPL dataset. Meanwhile, the three-level model can be conducted to examine which country-level profile of teacher and school factors appear to play a role in predicting or explaining student mathematics achievement. We found that about 22% variation of achievement varies across schools, and around 19% vary across eight participating countries.

Although the present study offers valuable insights into linking TALIS and PISA, the rTPL dataset reflects some apparent limitations. In TALIS, we miss student-level data, and in PISA, we miss specific teacher-level data that can be related to the unique student data. This was tackled by aggregating data at the school level. Therefore, it is not possible to look at the impact of unique characteristics situated within and between classroom settings in a school. Specific teaching style approaches and unique classroom composition effects cannot be identified. Several statistical and conceptual challenges should be taken into account when using the rTPL dataset: the original number of schools, teachers, and pupils participating in TALIS 2013 and PISA 2012 is far larger than the number in the rTPL dataset, and this affects the weights to be used when looking at values in the database.

The next thing to consider is how to solve the time gap in the TALIS 2013 and PISA 2012 administration and the way we selected teachers with at least one or two years of experience, depending on the hemisphere. This resulted in a smaller sample of schools, teachers, and pupils; but could also have harmed the representativeness of the final sample. For instance, the teacher sample of Mexico and Latvia was reduced to less than 200, while in other countries, more teachers could be retained in the rTPL sample. This smaller sample size could result in a loss of statistical power. Since the rTPL dataset will contain only data from eight countries, this also affects the extent to which we can generalize findings.

Additionally, the current study focuses on exploring the linking and possible use of the two databases. Regarding the case study, we emphasize using TALIS data to explain the achievement at the school level but less considering individual factors at the student level. In the subsequent studies, the student-related indicators, such as mathematics self-efficacy, and mathematics anxiety, could be considered.

Conclusion

The current study aimed to develop a linked database geared to mathematics teaching and learning to reflect the school mathematics educational environment. Taking into the subject-specific characteristics of mathematics education, we

extend a recent new avenue of research by (re)developing a linked database geared to the specific mathematics teaching and learning domain to reflect the school mathematics educational environment. The redesigned linkage dataset connects student and teacher data collected *via* TALIS 2013 and PISA 2012. It explores how to link TALIS and PISA data to study the dynamic relations between multiple educational effectiveness factors and student achievement as reflected in the Dynamic Model of Educational Effectiveness and the Opportunity-Propensity framework. Data from seven educational systems are used in this linkage process, and four critical issues related to five selection criteria are considered to address the specific sample of mathematics teachers. A case study, using Singapore linkage data through Multilevel Path Analysis demonstrated the feasibility and potential of exploring school effectiveness on the base of this new data set. Meanwhile, we pointed out that the Redesigned TALIS 2013 and PISA 2012 data presented challenges in terms of identifying a linkage variable, the aggregation of variables, and a sample selection procedure to identify the relevant mathematics teachers.

Student learning outcomes are the product of teachers and teaching, schools, educational systems, and students' diverse background characteristics, e.g., (Kyriakides and Luyten, 2009; Kyriakides et al., 2020). The current study provided new perspectives to understand this complex relationship while using a newly designed database of TALIS 2013 and PISA 2012. The design of the rTPL presented challenges in terms of identifying a linkage variable, the aggregation of variables, and a sample selection procedure to identify the relevant mathematics teachers.

Taken together, the study approach potentially stimulates future research about multi-level perspectives on PISA students' mathematics learning outcomes in various national contexts building on the EER dynamic model. A next avenue was suggested to focus on the international comparison of the relationships in the EER model. Also, the study could inspire future attempts linking data from TALIS 2018 and PISA 2018, with a focus on reading literacy as the primary domain. Nine countries participated in the TALIS-PISA Link 2018. Since both studies were administered in the same year, some drawbacks of the current linking approach do not apply. A collaboration with other researchers in view of this new endeavor is welcomed to tackle the methodological challenges and study the richness of the Dynamic Model of Educational Effectiveness and Opportunity-Propensity framework.

Data availability statement

The original contributions presented in this study are publicly available. The datasets for this study can be found at https://www.oecd.org/education/talis/talis-2013-data.htm (TALIS 2013 data) and https://www.oecd.org/pisa/pisaproducts/pisa2012database-downloadabledata.htm (PISA 2012 data).

Ethics statement

The Organization for Economic Co-operation and Development (OECD) reviewed and approved the studies involving human participants. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

Author contributions

XL was responsible for the theoretical framework and a major contributor in writing the manuscript. MV, KYH, and JDN supervised the research project together, ensured the article's coherence, and offers guidance on misunderstanding text. MV set out the objectives of the project and gave feedback during all phases. KYH had been responsible for the accuracy and logic of any part of the work and provides comments to the overall text. All authors put joint effort for this article, read, and approved the final manuscript.

Funding

This research was fully funded by the China Scholarship Council (CSC), grant number: CSC201807930019 and partially funded by the Fonds Wetenschappelijk Onderzoek (FWO), grant number: V412020N.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Ainley, J., Pratt, D., and Hansen, A. (2006). Connecting engagement and focus in pedagogic task design. *Br. Educ. Res. J.* 32, 23–38. doi: 10.1080/01411920500401971

Anghileri, J. (2006). Scaffolding practices that enhance mathematics learning. *J. Math. Teach. Educ.* 9, 33–52. doi: 10.1007/s10857-006-9005-9

Antoniou, P., and Kyriakides, L. (2013). A dynamic integrated approach to teacher professional development: Impact and sustainability of the effects on improving teacher behaviour and student outcomes. *Teach. Teach. Educ.* 29, 1–12. doi: 10.1016/j.tate.2012.08.001

Austin, B., Adesope, O. O., French, B. F., Gotch, C., Bélanger, J., and Kubacka, K. (2015). *Examining School Context and its Influence on Teachers: Linking TALIS 2013 with PISA 2012 Student Data. Education Working Paper No. 115*. Paris: OECD Publishing.

Ball, D. L., Hill, H. C., and Bass, H. (2005). Knowing mathematics for teaching: Who knows mathematics well enough to teach third grade, and how can we decide? *Am. Educ.* 29, 14–17; 20–22; 43–46.

Ball, D. L., Thames, M. H., and Phelps, G. (2008). Content knowledge for teaching: What makes it special? *J. Teach. Educ.* 59, 389–407. doi: 10.1177/0022487108324554

Blömeke, S., Suhl, U., Kaiser, G., and Döhrmann, M. (2012). Family background, entry selectivity and opportunities to learn: What matters in primary teacher education? An international comparison of fifteen countries. *Teach. Teach. Educ.* 28, 44–55. doi: 10.1016/j.tate.2011.08.006

Borko, H., and Putnam, R. T. (1995). "Expanding a teacher's knowledge base: a cognitive psychological perspective on professional development," in *Professional Development in Education: New Paradigms and Practices*, eds T. R. Guskey and M. Huberman (New York, NY: Teachers College Press), 35–65.

Byrnes, J. P. (2003). Factors predictive of mathematics achievement in white, black, and Hispanic 12th graders. *J. Educ. Psychol.* 95, 316–326. doi: 10.1037/0022-0663.95.2.316

Byrnes, J. P. (2020). The potential utility of an opportunity-propensity framework for understanding individual and group differences in developmental outcomes: A retrospective progress report. *Dev. Rev.* 56:100911. doi: 10.1016/j.dr.2020.100911

Byrnes, J. P., and Miller, D. C. (2007). The relative importance of predictors of math and science achievement: An opportunity-propensity analysis. *Contemp. Educ. Psychol.* 32, 599–629. doi: 10.1016/j.cedpsych.2006.09.002

Byrnes, J. P., and Miller-Cotto, D. (2016). The growth of mathematics and reading skills in segregated and diverse schools: An opportunity-propensity analysis of a national database. *Contemp. Educ. Psychol.* 46, 34–51. doi: 10.1016/j.cedpsych.2016.04.002

Byrnes, J. P., and Wasik, B. A. (2009). Factors predictive of mathematics achievement in kindergarten, first and third grades: An opportunity-propensity analysis. *Contemp. Educ. Psychol.* 34, 167–183. doi: 10.1016/j.cedpsych.2009.01.002

Caprara, G. V., Barbaranelli, C., Steca, P., and Malone, P. S. (2006). Teachers' self-efficacy beliefs as determinants of job satisfaction and students' academic achievement: A study at the school level. *J. Schl Psychol.* 44, 473–490. doi: 10.1016/j.jsp.2006.09.001

Caro, D. H., Lenkeit, J., and Kyriakides, L. (2016). Teaching strategies and differential effectiveness across learning contexts: Evidence from PISA 2012. *Stud. Educ. Eval.* 49, 30–41. doi: 10.1016/j.stueduc.2016.03.005

Chapin, S. H., and O'Connor, C. (2007). "Academically productive talk: Supporting students' learning in mathematics," in *The Learning of Mathematics*, Vol. 69, eds W. G. Martin, M. E. Strutchens, and P. C. Elliott (Reston, VA: National Council of Teachers of Mathematics), 113–128.

Chapman, C., Muijs, D., Reynolds, D., Sammons, P., and Teddlie, C. (2015). *The Routledge International Handbook of Educational Effectiveness and Improvement: Research, Policy, and Practice*. London: Routledge. doi: 10.4324/9781315679488

Cordero Ferrera, J. M., and Gil-Izquierdo, M. (eds) (2016). "TALIS-PISA link: guidelines for a robust quantitative analysis," in *Proceedings of the International Conference on Qualitative and Quantitative Economics Research (QQE)*, (Singapore: Global Science and Technology Forum). doi: 10.5176/2251-2012_QQE16.19

Cordero, J. M., and Gil-Izquierdo, M. (2018). The effect of teaching strategies on student achievement: An analysis using TALIS-PISA-link. *J. Policy Model.* 40, 1313–1331. doi: 10.1016/j.jpolmod.2018.04.003

Creemers, B. P. M., and Kyriakides, L. (2008). *The Dynamics of Educational Effectiveness: A Contribution to Policy, Practice and Theory in Contemporary Schools*. London: Routledge.

Creemers, B. P., and Scheerens, J. (1994). Developments in the educational effectiveness research programme. *Int. J. Educ. Res.* 21, 125–140. doi: 10.1016/0883-0355(94)90028-0

Creemers, B., and Kyriakides, L. (2007). *The Dynamics of Educational Effectiveness: A Contribution to Policy, Practice and Theory in Contemporary Schools*. London: Routledge. doi: 10.4324/9780203939185

Creemers, B., and Kyriakides, L. (2015). Developing, testing, and using theoretical models for promoting quality in education. *Schl Effect. Schl Improv.* 26, 102–119. doi: 10.1080/09243453.2013.869233

Delprato, M., and Chudgar, A. (2018). Factors associated with private-public school performance: Analysis of TALIS-PISA link data. *Int. J. Educ. Dev.* 61, 155–172. doi: 10.1016/j.ijedudev.2018.01.002

Desimone, L. M. (2009). Improving impact studies of teachers' professional development: Toward better conceptualizations and measures. *Educ. Res.* 38, 181–199. doi: 10.3102/0013189X08331140

Desimone, L. M., Porter, A. C., Garet, M. S., Yoon, K. S., and Birman, B. F. (2002). Effects of professional development on teachers' instruction: Results from a three-year longitudinal study. *Educ. Eval. Policy Anal.* 24, 81–112. doi: 10.3102/01623737024002081

Driessen, G. (2002). School composition and achievement in primary education: A large-scale multilevel approach. *Stud. Educ. Eval.* 28, 347–368. doi: 10.1016/S0191-491X(02)00043-3

Dyer, N. G., Hanges, P. J., and Hall, R. J. (2005). Applying multilevel confirmatory factor analysis techniques to the study of leadership. *Leadersh. Q.* 16, 149–167. doi: 10.1016/j.leaqua.2004.09.009

Echazarra, A., Salinas, D., Méndez, I., Denis, V., and Rech, G. (2016). *How Teachers Teach and Students Learn: Successful Strategies for School. OECD Education Working Papers, No 130*. Paris: OECD Publishing.

Erickson, G., Brandes, G. M., Mitchell, I., and Mitchell, J. (2005). Collaborative teacher learning: Findings from two professional development projects. *Teach. Teach. Educ.* 21, 787–798. doi: 10.1016/j.tate.2005.05.018

Fernández-Díaz, M. J., Rodríguez-Mantilla, J. M., and Martínez-Zarzuelo, A. (2016). PISA y TALIS ¿congruencia o discrepancia? *RELIEVE - Rev. Electr. Investig. Eval. Educ.* 22:9. doi: 10.7203/relieve.22.1.8247

Finch, H., and Bolin, J. (2017). *Multilevel Modeling using Mplus*. Boca Raton, FL: CRC Press. doi: 10.1201/9781315165882

Fraser, B. J. (2013). "Classroom learning environments," in *Handbook of Research on Science Education*, (London: Routledge), 117–138.

Ganzeboom, H. B. (ed.) (2010). "A new international socio-economic index (ISEI) of occupational status for the international standard classification of occupation 2008 (ISCO-08) constructed with data from the ISSP 2002–2007," in *Paper Presented at the Annual Conference of International Social Survey Programme*, (Lisbon).

Garet, M. S., Porter, A. C., Desimone, L., Birman, B. F., and Yoon, K. S. (2001). What makes professional development effective? Results from a national sample of teachers. *Am. Educ. Res. J.* 38, 915–945. doi: 10.3102/00028312038004915

Gil-Izquierdo, M., and Cordero, J. M. (2018). Guidelines for data fusion with international large scale assessments: Insights from the TALIS-PISA link database. *Stud. Educ. Eval.* 59, 10–18. doi: 10.1016/j.stueduc.2018.02.002

Goddard, Y. L., Goddard, R. D., and Tschannen-Moran, M. (2007). A theoretical and empirical investigation of teacher collaboration for school improvement and student achievement in public elementary schools. *Teach. Coll. Record.* 109, 877–896. doi: 10.1177/016146810710900401

Grossman, P. L. (1990). *The Making of a Teacher: Teacher Knowledge and Teacher Education*. New York, NY: Teachers College Press.

Hattie, J. (2002). Classroom composition and peer effects. *Int. J. Educ. Res.* 37, 449–481. doi: 10.1016/S0883-0355(03)00015-6

Hattie, J. (2008). *Visible Learning: A Synthesis of Over 800 Meta-Analyses Relating to Achievement*. Abingdon: Taylor & Francis. doi: 10.4324/9780203887332

Hill, H. C., Blunk, M. L., Charalambous, C. Y., Lewis, J. M., Phelps, G. C., Sleep, L., et al. (2008). Mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study. *Cogn. Instruc.* 26, 430–511. doi: 10.1080/07370000802177235

Hill, H. C., Rowan, B., and Ball, D. L. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *Am. Educ. Res. J.* 42, 371–406. doi: 10.3102/00028312042002371

Hill, H. C., Schilling, S. G., and Ball, D. L. (2004). Developing of teachers' measures mathematics knowledge for teaching. *Elem. Schl J.* 105, 11–30. doi: 10.1086/428763

Holzberger, D., Philipp, A., and Kunter, M. (2013). How teachers' self-efficacy is related to instructional quality: A longitudinal analysis. *J. Educ. Psychol.* 105, 774–786. doi: 10.1037/a0032198

Hornstra, L., van der Veen, I., Peetsma, T., and Volman, M. (2015). Does classroom composition make a difference: Effects on developments in motivation, sense of classroom belonging, and achievement in upper primary school. *Schl Effect. Schl Improv.* 26, 125–152. doi: 10.1080/09243453.2014.887024

Hox, J. J., Moerbeek, M., and Van de Schoot, R. (2017). *Multilevel Analysis: Techniques and Applications*. London: Routledge. doi: 10.4324/9781315650982

Hu, L. T., and Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Struct. Equ. Model.* 6, 1–55. doi: 10.1080/10705519909540118

Huang, J. L., Tang, Y. P., He, W. J., and Li, Q. (2019). Singapore's school excellence model and student learning: Evidence from PISA 2012 and TALIS 2013. *Asia Pacif. J. Educ.* 39, 96–112. doi: 10.1080/02188791.2019.1575185

Hunter, R. (2005). "Reforming communication in the classroom: One teacher's journey of change," in *Building Connections: Research, Theory and Practice (Proceedings of the Annual Conference of the Mathematics Education Research Group of Australasia)*, eds P. Clarkson, A. Downton, D. Gronn, M. Horne, A. McDonough, R. Pierce, et al. (Melbourne: MERGA), 451–458.

Kaplan, D., and McCarty, A. T. (2013). Data fusion with international large scale assessments: A case study using the OECD PISA and TALIS surveys. *Large Scale Assess. Educ.* 1:6. doi: 10.1186/2196-0739-1-6

Kelly, A. (2012). Measuring equity' and equitability' in school effectiveness research. *Br. Educ. Res. J.* 38, 977–1002. doi: 10.1080/01411926.2011.605874

Kleickmann, T., Tröbst, S., Heinze, A., Bernholt, A., Rink, R., and Kunter, M. (2017). "Teacher knowledge experiment: conditions of the development of pedagogical content knowledge," in *Competence Assessment in Education*, eds D. Leutner, J. Fleischer, J. Grünkorn, and E. Klieme (Cham: Springer), 111–129. doi: 10.1007/978-3-319-50030-0_8

Klieme, E. (2013). "The role of large-scale assessments in research on educational effectiveness and school development," in *The Role of International Large-Scale Assessments: Perspectives from Technology, Economy, and Educational Research*, eds M. von Davier, E. Gonzalez, I. Kirsch, and K. Yamamoto (Dordrecht: Springer), 115–147. doi: 10.1007/978-94-007-4629-9_7

Kyriakides, L., and Luyten, H. (2009). The contribution of schooling to the cognitive development of secondary education students in Cyprus: An application of regression discontinuity with multiple cut-off points. *Schl Effect. Schl Improv.* 20, 167–186. doi: 10.1080/09243450902883870

Kyriakides, L., Christoforou, C., and Charalambous, C. Y. (2013). What matters for student learning outcomes: A meta-analysis of studies exploring factors of effective teaching. *Teach. Teach. Educ.* 36, 143–152. doi: 10.1016/j.tate.2013.07.010

Kyriakides, L., Creemers, B. P. M., Muijs, D., Rekers-Mombarg, L., Papastylianou, D., Van Petegem, P., et al. (2014). Using the dynamic model of educational effectiveness to design strategies and actions to face bullying. *Schl Effect. Schl Improv.* 25, 83–104. doi: 10.1080/09243453.2013.771686

Kyriakides, L., Creemers, B. P. M., Panayiotou, A., and Charalambous, E. (2020). *Quality and Equity in Education: Revisiting Theory and Research on Educational Effectiveness and Improvement*. London: Routledge. doi: 10.4324/9780203732250

Kyriakides, L., Creemers, B., Antoniou, P., and Demetriou, D. (2010). A synthesis of studies searching for school factors: Implications for theory and research. *Br. Educ. Res. J.* 36, 807–830. doi: 10.1080/01411920903165603

Kyriakides, L., Georgiou, M. P., Creemers, B. P. M., Panayiotou, A., and Reynolds, D. (2017). The impact of national educational policies on student achievement: A European study. *Schl Effect. Schl Improv.* 29, 171–203. doi: 10.1080/09243453.2017.1398761

Le Donné, N., Fraser, P., and Bousquet, G. (2016). *Teaching Strategies for Instructional Quality: Insights from the TALIS-PISA Link Data. OECD Education Working Papers No. 148*. Paris: OECD Publishing, doi: 10.1787/5jln1hlsr0lr-en

Leunda Iztueta, I., Garmendia Navarro, I., and Etxeberria Murgiondo, J. (2017). Statistical matching in practice – An application to the evaluation of the education system from PISA and TALIS. *Rev. Investig. Educ.* 35, 371–388. doi: 10.6018/rie.35.2.262171

Ma, L. (2010). *Knowing and Teaching Elementary Mathematics: Teachers' Understanding of Fundamental Mathematics in China and the United States*. London: Routledge. doi: 10.4324/9780203856345

Mammadov, R., and Cimen, I. (2019). Optimizing teacher quality based on student performance: A data envelopment analysis on PISA and TALIS. *Int. J. Instruc.* 12, 767–788. doi: 10.29333/iji.2019.12449a

Martínez-Abad, F., Gamazo, A., and Rodríguez-Conde, M.-J. (2020). Educational data mining: Identification of factors associated with school effectiveness in PISA assessment. *Stud. Educ. Eval.* 66:100875. doi: 10.1016/j.stueduc.2020.100875

Muijs, D., and Reynolds, D. (2002). Teachers' beliefs and behaviors: What really matters? *J. Classroom Interact.* 37, 3–15.

Muijs, D., and Reynolds, D. (2003). Student background and teacher effects on achievement and attainment in mathematics: A longitudinal study. *Educ. Res. Eval.* 9, 289–314. doi: 10.1076/edre.9.3.289.15571

Muijs, D., Creemers, B., Kyriakides, L., Van der Werf, G., Timperley, H., and Earl, L. (2014). Teaching effectiveness. A state of the art review. *Schl Effect. Schl Improv.* 24, 231–256. doi: 10.1080/09243453.2014.885451

Muthén, L. K., and Muthén, B. (2017). *Mplus User's Guide: Statistical Analysis with Latent Variables*. New York, NY: Wiley.

Nilsen, T., and Gustafsson, J.-E. (2016). *Teacher Quality, Instructional Quality and Student Outcome. Relationships Across Countries, Cohorts and Time*. Berlin: Springer. doi: 10.1007/978-3-319-41252-8

OECD (2009). *Creating Effective Teaching and Learning Environments First Results from TALIS*. Paris: OECD Publishing. doi: 10.1787/9789264068780-en

OECD (2010). *TALIS 2008 Technical Report*. Paris: OECD Publishing. doi: 10.1787/9789264079861-en

OECD (2013a). *PISA 2012 Assessment and Analytical Framework*. Paris: OECD Publishing.

OECD (2013b). *TALIS 2013: Conceptual Framework*. Paris: OECD Publishing.

OECD (2014a). *PISA 2012 Technical Report*. Paris: OECD Publishing.

OECD (2014b). *TALIS 2013 Results: An International Perspective on Teaching and Learning*. Paris: OECD Publishing.

OECD (2014c). *TALIS 2013 Technical Report*. Paris: OECD Publishing.

OECD (2017). *PISA 2015 Assessment and Analytical framework*. Paris: OECD Publishing. doi: 10.1787/9789264281820-en

OECD (2019a). *PISA 2018 Assessment and Analytical Framework*. Paris: OECD Publishing.

OECD (2019b). *TALIS 2018 Technical Report*. Paris: OECD Publishing.

Opdenakker, M. C., and Van Damme, J. (2000). The importance of identifying levels in multilevel analysis: An illustration of the effects of ignoring the top or intermediate levels in school effectiveness research. *Schl Effect. Schl Improv.* 11, 103–130. doi: 10.1076/0924-3453(200003)11:1;1-#;FT103

Opdenakker, M. C., and Van Damme, J. (2006a). Differences between secondary schools: A study about school context, group composition, school practice, and school effects with special attention to public and Catholic schools and types of schools. *Schl Effect. Schl Improv.* 17, 87–117. doi: 10.1080/09243450500264457

Opdenakker, M. C., and Van Damme, J. (2006b). Teacher characteristics and teaching styles as effectiveness enhancing factors of classroom practice. *Teach. Teach. Educ.* 22, 1–21. doi: 10.1016/j.tate.2005.07.008

Opdenakker, M. C., and Van Damme, J. (2007). Do school context, student composition and school leadership affect school practice and outcomes in secondary education? *Br. Educ. Res. J.* 33, 179–206. doi: 10.1080/01411920701208233

Opdenakker, M. C., Van Damme, J., De Fraine, B., Van Landeghem, G., and Onghena, P. (2002). The effect of schools and classes on mathematics achievement. *Schl Effect. Schl Improv.* 13, 399–427. doi: 10.1076/sesi.13.4.399.10283

Panayiotou, A., Kyriakides, L., and Creemers, B. P. (2016). Testing the validity of the dynamic model at school level: A European study. *Schl Leadersh. Manag.* 36, 1–20. doi: 10.1080/13632434.2015.1107537

Panayiotou, A., Kyriakides, L., Creemers, B. P., McMahon, L., Vanlaar, G., Pfeifer, M., et al. (2014). Teacher behavior and student outcomes: Results of a European study. *Educ. Assess. Eval. Acc.* 26, 73–93. doi: 10.1007/s11092-013-9182-x

Reynolds, D., Sammons, P., De Fraine, B., Van Damme, J., Townsend, T., Teddlie, C., et al. (2014). Educational effectiveness research (EER): A state-of-the-art review. *Schl Effect. Schl Improv.* 25, 197–230. doi: 10.1080/09243453.2014.885450

Rimm-Kaufman, S. E., Storrn, M. D., Sawyer, B. E., Pianta, R. C., and LaParo, K. M. (2006). The teacher belief Q-sort: A measure of teachers' priorities in relation to disciplinary practices, teaching practices, and beliefs about children. *J. Schl Psychol.* 44, 141–165. doi: 10.1016/j.jsp.2006.01.003

Robardey, C. P., Allard, D. W., and Brown, D. M. (1994). An assessment of the effectiveness of full option science system training for third-through sixth-grade teachers. *J. Elem. Sci. Educ.* 6, 17–29. doi: 10.1007/BF03170647

Rutkowski, L., von Davier, M., and Rutkowski, D. (2013). *Handbook of International Large-Scale Assessment*. New York, NY: Chapman and Hall. doi: 10.1201/b16061

Scheerens, J. (2007). *Conceptual Framework for the PISA 2009 Context Questionnaires and Thematic Reports*. Oslo: PISA Governing Board.

Scheerens, J. (2013). The use of theory in school effectiveness research revisited. *Schl Effect. Schl Improv.* 24, 1–38. doi: 10.1080/09243453.2012.691100

Scheerens, J., and Bosker, R. (1997). *The Foundations of Educational Effectiveness*. Oxford: Pergamon.

Sealy, K. M., Perry, S. M., and DeNicola, T. (2016). *Relationships and Predictors of Principal Job Satisfaction Across Multiple Countries: A Study using TALIS 2013 and PISA 2012*. Rochester, NY: SSRN. doi: 10.2139/ssrn.2717056

Senk, S. L., Tatto, M. T., Reckase, M., Rowley, G., Peck, R., and Bankov, K. (2012). Knowledge of future primary teachers for teaching mathematics: An international comparative study. *ZDM Math. Educ.* 44, 307–324. doi: 10.1007/s11858-012-0400-7

Shulman, L. (1987). Knowledge and teaching: Foundations of the new reform. *Harv. Educ. Rev.* 57, 1–23. doi: 10.17763/haer.57.1.j463w79r56455411

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educ. Res.* 15, 4–14. doi: 10.3102/0013189X015002004

Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Rev. Educ. Res.* 75, 417–453. doi: 10.3102/00346543075003417

Snow-Renner, R., and Lauer, P. A. (2005). *Professional Development Analysis. McREL Insights*. Aurora, CO: Mid-continent Research for Education and Learning.

Stacey, K. (ed.) (2015). "The international assessment of mathematical literacy: PISA 2012 framework and items," in *Proceedings of the 12th International Congress on Mathematical Education*, (Seoul). doi: 10.1007/978-3-319-17187-6_43

Strietholt, R., and Scherer, R. (2018). The contribution of international large-scale assessments to educational research: Combining individual and institutional data sources. *Scand. J. Educ. Res.* 62, 368–385. doi: 10.1080/00313831.2016.1258729

Timperley, H., Wilson, A., Barrar, H., and Fung, I. (2008). *Teacher Professional Learning and Development*. Wellington, NZ: Ministry of Education.

Tschannen-Moran, M., and McMaster, P. (2009). Sources of self-efficacy: Four professional development formats and their relationship to self-efficacy and implementation of a new teaching strategy. *Elem. Schl J.* 110, 228–245. doi: 10.1086/605771

United Nations Educational Scientific and Cultural Organization [UNESCO] (2003). *International Standard Classification of Education, ISCED 1997*. Paris: UNESCO.

Van Damme, J., Liu, H., Vanhee, L., and Pustjens, H. J. E. E. (2010). Longitudinal studies at the country level as a new approach to educational effectiveness: Explaining change in reading achievement (PIRLS) by change in age, socio-economic status and class size. *Effect. Educ.* 2, 53–84. doi: 10.1080/19415531003616888

Van Ewijk, R., and Sleegers, P. (2010). The effect of peer socioeconomic status on student achievement: A meta-analysis. *Educ. Res. Rev.* 5, 134–150. doi: 10.1016/j.edurev.2010.02.001

Verhaeghe, J. P., Vanlaar, G., Knipprath, H., De Fraine, B., and Van Damme, J. (2018). Can group composition effects explain socioeconomic and ethnic achievement gaps in primary education? *Stud. Educ. Eval.* 57, 6–15. doi: 10.1016/j.stueduc.2017.07.006

Wagemaker, H. (2014). "International large-scale assessments: From research to policy," in *Handbook of International Large-Scale Assessment: Background, Technical Issues, and Methods of Data Analysis*, eds L. Rutkowski, M. von Davier, and D. Rutkowski (London: CRC Press), 11–36.

Wagemaker, H. (2020). *Reliability and Validity of International Large-Scale Assessment: Understanding IEA's Comparative Studies of Student Achievement*. Cham: Springer. doi: 10.1007/978-3-030-53081-5

Watson, A., and Mason, J. (2006). Seeing an exercise as a single mathematical object: Using variation to structure sense-making. *Math. Think. Learn.* 8, 91–111. doi: 10.1207/s15327833mtl0802_1

White, K. R. (1982). The relation between socioeconomic status and academic achievement. *Psychol. Bull.* 91, 461–481. doi: 10.1037/0033-2909.91.3.461

Yang Hansen, K., Rosén, M., and Gustafsson, J. E. (2011). Changes in the multi-level effects of socio-economic status on reading achievement in Sweden in 1991 and 2001. *Scand. J. Educ. Res.* 55, 197–211. doi: 10.1080/00313831.2011.554700

You, H. S., Park, S., and Delgado, C. (2021). A closer look at US schools: What characteristics are associated with scientific literacy? A multivariate multilevel analysis using PISA 2015. *Sci. Educ.* 105, 406–437. doi: 10.1002/sce.21609

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 29 September 2022
DOI 10.3389/fpsyg.2022.925394

OPEN ACCESS

EDITED BY
Aaron Williamon,
Royal College of Music, United Kingdom

REVIEWED BY
Clemens Wöllner,
University of Hamburg,
Germany
Frank Heuser,
University of California,
United States

*CORRESPONDENCE
Elena Alessandri
elena.alessandri@hslu.ch

SPECIALTY SECTION
This article was submitted to Performance Science, a section of the journal Frontiers in Psychology

RECEIVED 21 April 2022
ACCEPTED 05 September 2022
PUBLISHED 29 September 2022

CITATION
Alessandri E, Baldassarre A and Williamson VJ (2022) The critic's voice: On the role and function of criticism of classical music recordings. *Front. Psychol.* 13:925394. doi: 10.3389/fpsyg.2022.925394

The critic's voice: On the role and function of criticism of classical music recordings

Elena Alessandri*, Antonio Baldassarre and Victoria Jane Williamson

School of Music, Lucerne University of Applied Sciences and Arts, Lucerne, Switzerland

In the Western classical tradition music criticism represents one of the most complex and influential forms of performance assessment and evaluation. However, in the age of peer opinion sharing and quick communication channels it is not clear what place music critics' judgments still hold in the classical music market. This article presents expert music critics' view on their role, function, and influence. It is based on semi-structured interviews with 14 native English- and German-speaking critics who had an average of 32 years professional activity in classical music review. We present the first visual model to summarize music critics' descriptions of their role and responsibilities, writing processes, and their influences (on the market and on artists). The model distinguishes six roles (*hats*): *consumer adviser*, *teacher*, *judge*, *writer*, *stakeholder*, and artist *advocate*. It identifies core *principles* governing critical writing for music as well as *challenges* that arise from balancing the above six responsibilities whilst remaining true to an implicit code of conduct. Finally, it highlights the factors that inform critics' writing in terms of the *topics* they discuss and the discursive *tools* they employ. We show that music critics self-identify as highly skilled mediators between artists, producers and consumers, and justify their roles as judge and teacher based on a wealth of experience as against the influx of pervasive amateur reviews. Our research approach also offers occupation-based insights into professional music review standards, including the challenges of maintaining objectivity and resisting commercial pressures. This article offers a new viewpoint on music critics' judgments and recommendations that helps to explain their expectations and reflections.

KEYWORDS

music review, music recording, classical music, expert judgment, performance value

Introduction

This article explores the performance evaluation discourse and its context through the examination of the nature and role of one of the most complex and historically relevant authorities in this domain: professional music criticism. The landscape of critical discourse on art criticism – and music criticism within it – dates back to the 19th century, with seminal works by, e.g., Brendel (1855, pp. 231–240), Buck (1905), Hellouin (1906),

Calvocoressi (1923), Newman (1925), Fox Strangways (1938/1939), French (1948), Becker (1965), Aschenbrenner (1981), Cone (1981) or Ellis (1995). In the past few decades, this theoretical reflection has been expanded through systematic examinations of specific features and conditions of criticism that cover culture and the arts, including surveys on the status, role and function of classical music critics (e.g., Eatock, 2004; McGill et al., 2005; Kristensen and From, 2015a; Verboord and Janssens, 2015). Within this research focus, art critics have been described as "journalists with a difference" (Forde, 2003, p. 113) and as "journalist with that little something extra," (Harries and Wahl-Jorgensen, 2007, p. 623). They deal with "culture" as encapsulated in and expressed through "works and practices of intellectual and especially artistic activity" (Williams, 1985, p. 90) Therefore, art, and specifically, music criticism is broadly held to be a "cultural" and not a "literary" practice in the emphatic sense of the concept (Eagleton, 1984, p. 18) – actually an overly intimate relationship with literature as was broadly practiced in the nineteenth century (e.g., Plantinga, 1967, pp. 59–78; Dahlhaus, 1971, p. 12; Dahlhaus, 1981; Schmitz-Emans, 2015) is considered a dangerous liaison (Kramer, 1989), even against the current popular opinion that "[music] criticism is supposed to be the effort of literary, entertaining, and provocative craftsmanship" (Frederik Hanssen in Diederichs-Lafite, 1996, p. 505). Consequently, critics are regarded as "cultural mediators and gatekeepers" (Janssen and Verboord, 2015) or as "cultural intermediary," to apply a concept coined by Bourdieu (1984), p. 325), defined as, among others, "critics of 'quality' newspapers and magazines and all the writer-journalists and journalist-writers', who have assigned themselves the role of divulging 'legitimate culture' " (Bourdieu, 1984, p. 326). In the wake of Bourdieu's notion, scholars have analyzed criticism to study the constructs through which music is made meaningful by the "quality press" (Shrum, 1991; Cheyne and Binder, 2010).

In music criticism the aforementioned "little something extra" concerns "the intellectual activity of formulating judgments on the value and degree of excellence of individual works of music, or whole groups or genres" (Bujic, 2011). The basis of such activity is "aesthetic appreciation," however music criticism encompasses much "more than spontaneous liking"; it assumes the ability "to judge and to talk about style, technique, originality" thus identifying music critics as "experts" in the state of the art (Barzun, 2001, pp. 71–72). In addition, since the early days of music criticism, critics significantly contributed to a collective knowledge (Becker, 1982) that built the parameters upon which current music reviewers seek to analyze the quality and value of a classical music recording. The practice of talking, evaluating, and judging cultural objects as music is, however, culturally determined, i.e., the institutional embeddedness of music criticism is not a minor or marginal issue but rather a central analytical dimension worthy of examination (Blank, 2007).

The music critic's product, i.e., music criticism or music reviews, is a well-established practice in the history of Western classical music (Schenk-Güllich, 1972; Kirchmeyer, 2017; Dingle, 2019a). In the 18th century music criticism developed into a professional, and, from the 19th century onward, an influential intellectual practice within the European musical discourse (Stuckenschmidt, 1965; Baldassarre et al., 2022). It is important to point out that – starting from early approaches (Mattheson, 1722–1725; Scheibe, 1737/1740) – music criticism was first and foremost developed into a critique of works and compositions (Dahlhaus, 1971; Monelle, 2002), for which Schumann (1854/1985); see also Plantinga (1967) and Hanslick (1870) provide prime examples, rather than an explicit critique of musical performance (Ertelt and von Loesch, 2021). Only "opera criticism offers a striking exception" in this context given its predominant focus on the quality of "opera singers' voices" (Abbate, 2004, p. 508; see also Fenner, 1994; Baldassarre, 2009; Ellis, 2012).

Genuine performance criticism did not emerge until mid/late-nineteenth century, influenced by a modified understanding of the musical artist's persona as shaped by the nineteenth-century concept of and discourse on musical virtuosity (Samson, 2003; Gooley, 2009; Ruprecht, 2013; Strandberg, 2014; Stefaniak, 2016; Doran, 2020). The belated recognition of the music performer's accomplishments is hardly surprising in view of the generally wide-spread dismissive and neglecting stance of music critics toward the role and function of the musical performer that persisted till the beginning of the 20th century. For instance, the famous music critic William James Henderson stated that "the consideration of the performer is the last important office of real criticism; but unfortunately, it is the one on which the public lays the largest attention" (Henderson, 1915, p. 75). During the first half of the 20th century, driven by the innovation in the recording technology (Benjamin, 1980; Siefert, 1984, pp. 114–115; Katz, 2004; Burgess, 2014) and the strengthening of a canon of both the classical music repertoire (Hamer, 2019) and of its auditory appropriation (Nikolsky, 2012; Thorau and Ziemer, 2019), not only was the performer's reputation as an essential agent significantly enhanced, but also a new form of music criticism developed, focussed on recorded music as the result of the interpreter's performative choices (Dingle, 2019b): professional reviews of classical music recordings.

Recording criticism is a complex form of reasoned evaluation that is very different from live performance criticism in terms of its text content, process, and purpose (Schick, 1996, pp. 153–165). During the course of the century, recording reviews started to appear regularly in specialized magazines such as *The Gramophone* in the United Kingdom (from 1923 to present), the US-based *American Record Guide* (founded in 1935) and *Fono Forum* (from 1957 onward) in Germany, and soon, from the 1920s, music recording criticism "became commonplace" (Dingle, 2019b, p. 253), i.e., a familiar form of written response to music, which entails the description, analysis, categorization and evaluation of music with a focus on topics linked to music performance (Carroll, 2009; Alessandri et al., 2016a). These critical writings have potential purpose and impact beyond historical record and reader information; they are supposed to influence consumer choices and affect musicians' careers and the standing of recording labels (Pollard 1998; Alessandri et al., 2014). The significance of

music performance criticism can hardly be overestimated given the fact that most of the music people listen to is first and foremost in a recorded format (Elste, 2009).

Previous work by the authors (Alessandri et al., 2015, 2016a, 2016b), in which hundreds of published recording reviews were text-analyzed, offered a first structured model of the content of reviews of classical music recordings. It showed how the evaluation of music performance lies at the core of critics' writings: the nuanced variety of metaphorical and technical descriptors of the performed sound covers on average over half of the review text and is used by critics to ground and support their judgments of value. Those judgments assess the aesthetic qualities of the performance, but also go beyond that to evaluate the musical output as the result of the artist's achievement and its importance in the wider music market. This work offered us a solid understanding of the topics discussed in published reviews, but not into the critics' intentions and motivations in writing.

Building on this analysis of published review content, in the present study we thus expanded this modelling approach from the written word to the spoken dialogue. Through a series of purpose built semi-structured interviews, we sought classical music critics' opinions in order to understand the motivations and perceived roles behind their self- and situational-descriptions, as well as the narratives they use to justify their methods and compartmentalize their professional identities. This approach allowed us for the first time to move beyond the published critique and contrast critics' intentions about critique with their actual written outputs.

Research in this area is timely given the, for decades now, repeatedly cited 'crisis' regarding a sharp decline not only in music criticism but in all form of art journalism (Boenisch, 2008; Kristensen, 2010; Caduff, 2014; Jaakkola, 2015; Heikkilö et al., 2017; Melnyk, 2019; Widholm et al., 2021) and, not least, also due to the new music consumption behaviors and peer-communication channels in the digital age (Varriale, 2012; Hracs et al. 2016; Baldassarre and Alessandri, 2022). Digital technologies have revolutionized the way we listen to and discuss music, giving artists more direct access to their audiences, creating platforms for peer-opinion, and empowering listeners with new means and resources to facilitate decision making with regards to purchasing and listening (Carboni, 2012; Datta et al., 2017). In a world of peer-opinion, it is reasonable to question the role of professional critics. And yet for the listener, the ease of access to digitalized music, combined with its dematerialization and the displacement of product-ownership (due to streaming services) have combined to create a sense of disconnection to artists and a renewed interest in gathering knowledge about the music and the musicians behind it (Crossley and Bottero, 2015; Arditi, 2018; Hesmondhalgh and Meier, 2018).

To understand the critics' rapidly changing role in the news pantheon – with regard to which Caduff (2014) wonders whether these changes could really be taken as symptoms of decline or whether they are more likely signs of a re-formation of music criticism – we must scrutinize their place in the classical music market as agents in the cultural industries (Debenedetti, 2006), where they seem to face increasing marginalization from alternative reviews and commercial pressures such as online rating systems, PR stunts, and the influences of 'celebrity' classical music artists and fan culture. In the face of this shift, this article focuses on how music critics themselves view their role in today's classical music market, how they value their professional standards, and how they experience and assess the relationships with artists, music producers and the readers of music critique.

Materials and Methods

Participants

We interviewed eight English- and six German-speaking music critics based in UK, Germany, and Switzerland at the time of the study. The critics were recruited *via* social media, radio stations, and specialized communities and all of them had at least five-year experience in reviewing recorded classical music. Besides their extensive practice with record critique, we set no further criteria in terms of quality of their experience, preferring instead to take a wide sample of music critics from print and broadcast media and, for the first time to our knowledge, from different countries (UK, Germany and Switzerland). The fourteen critics (2 women, 12 men; age average 59.14, range 32–76) had an average of 31.71 years activity in major classical music review outlets (range 5–50 yrs) including BBC Music, Gramophone, FonoForum, and Rondo (see Supplementary material 1 for details on the experience of each critic). The gender distribution within the sample reflects the current market, with a clear predominance of male critics (McGill et al., 2005; Reus and Naab, 2014; Reus and Müller, 2017). All had a graduate or postgraduate degree in an art or language related field (7 musicology, 2 German language, 2 music, 1 drama/theatre, 1 English literature, 1 French/German literature). By the time of the interview, our critics had published an average of 40 classical music-recording reviews in the past 12 months. They also had extensive experience as performers, editors, and/or record producers.

The critics completed an online survey prior to their interview in order to collect demographic information and their Goldsmith Musical Sophistication index score (GoldMSI). The GoldMSI is a standardized self-report inventory that measures ability to engage with music in a nuanced, flexible, and effective way (Müllensiefen et al., 2014). As expected, all critics scored far above the population average on this scale (population average: 81.58; music critics: 102.79; range 90–120).

Interview

In-depth, semi-structured interviews were conducted in United Kingdom, Germany, and Switzerland. This form of interview has been described as "conversations with a purpose" (Legard et al., 2003, p. 138) that explores a person's opinions,

feelings and beliefs. The interviews are structured around a leading thread of discourse based on the main themes of enquiry (in this case the nature, role, and influence of music criticism) while allowing conversation to remain flexible, in terms of topic order and new, unexpected topics raised by the interviewee. This method is an ideal way to collect rich data from a small pool of experts (Harries and Wahl-Jorgensen, 2007).

Interviews focused on: (i) the aspects of a recording that were typically reviewed; (ii) the way these aspects are discussed, in terms of language and rhetorical devices and; (iii) the role of professional music criticism in the classical recording market and its influence on key stakeholders such as artists, music producers and the reading public. The development of the interview schedule followed the results of previous work on published music reviews (Alessandri et al., 2014, 2015, 2016a,b). Themes and hypotheses that emerged from these analyses were used to develop questions and prompts. For example, the extended used of comparative judgments evidenced in the analysis of Gramophone reviews (Alessandri et al., 2014, 2015) gave rise to the prompt "*How important is it to compare the recording reviewed with other recordings?*" (for the full interview schedule see Supplementary material 2). Interviews lasted on average 1 h 42′ (range 1 h 12′ – 2 h 57′). The conversations were audio recorded and transcribed verbatim. The interview protocol was approved by the authors' university ethical review board. All critics gave written informed consent in accordance with the Declaration of Helsinki.

Analysis

We used a double-coder inductive thematic analysis, in line with general thematic applied analysis methods (e.g., Guest et al., 2012), to produce a visual map of the topics discussed. The protocol followed Williamson et al. (2012), Williamson and Jilka (2013), and Alessandri et al. (2015, 2016a), with the addition of a third coder and a two-stage procedure to account for the bilingual data (English, German). Three researchers (the authors) were involved; each interview was analyzed by one native speaker (third and second authors) plus one researcher (first author) fluent in both languages, thereby assuring methodological continuity and coherence.

The eight English interviews were analyzed first. The first and third authors examined the transcripts independently using line-by-line open coding, comparing and contrasting quotes and organizing codes to develop a map of emergent themes. These themes were then compared between the researchers. To minimize subjectivity, each researcher in turn explained a theme, justifying it by means of quotes and proposing a definition. Based on the newly developed codebook, all data were re-coded by the two researchers independently. Text parsing in the second stage of coding was performed at minimum close level and multi-layered coding was avoided. If a text fragment encompassed more than one theme, then (i) new ideas were prioritized and (ii) the text was coded for the most salient idea. Avoiding multi-layered coding meant that we could not account for the intricacy of language (e.g., the distance between themes as a proxy to links between concepts). However, this approach allowed us to extract the thematic content in its purest form without being burdened by the nature of language construction and thereby to develop a general model from both an English and a German sample.

The model was then applied to the six German interviews. At this stage the second author, a native German speaker, joined the analysis. Again, all interviews were analyzed independently using the developed theme codes, revising and clarifying definitions where needed. NVivo version 11 was used for the application of codes and for computing agreement level in both stages. Reliability in the application of codes between researchers was high for both English ($\hat{\kappa} = 0.976$) and German ($\hat{\kappa} = 0.959$) interviews. This protocol permitted a structured development of the final visual model. It also allowed us to test the applicability of the model to a different critique sample, in a different language, with a different musical tradition, and with a new coder.

Findings

Five main theme categories emerged from the interviews, which contained a total of 47 themes and sub-themes (Figure 1). Together they described the nature of music criticism through the eyes of critics in terms of their role (**Hats**, **Principles**, and **Challenges**) and strategies (**Topics** and **Tools**). Supplementary material 3 shows all themes with their definitions and example quotes from the interviews. For German quotes, English translations are provided: original German quotes are reported in Supplementary material 4.

Hats – Things I am

In this first theme category, critics described how they see their role in the classical market. The theme family is called 'Hats' to emphasize that critics move between different functions and responsibilities. In interview, they distinguished between six different roles.

Three roles reflect functions usually attributed to cultural intermediaries, mediators or gatekeepers (Bourdieu, 1984; Cheyne and Binder, 2010; Smith Maguire and Matthews, 2012; Kristensen and From, 2015b; Janssen and Verboord, 2015): ascribing value to products, thus setting "*a few reference points in this [music industry] jungle*" (C9) (**Judge**); legitimizing the cultural industry and acting as communication channel between artists and consumers (**Stakeholder**); and acting as creating agents who deliver valuable journalistic products (**Writer**). The role of **Stakeholder** was described by critics as central to their work, encapsulating the nature of criticism as the point of intersection between artists, industry and the public. In the words of critics: "*…the role of the review in the classical recording market is crucial. Without reviews the market would only half-function, because it*

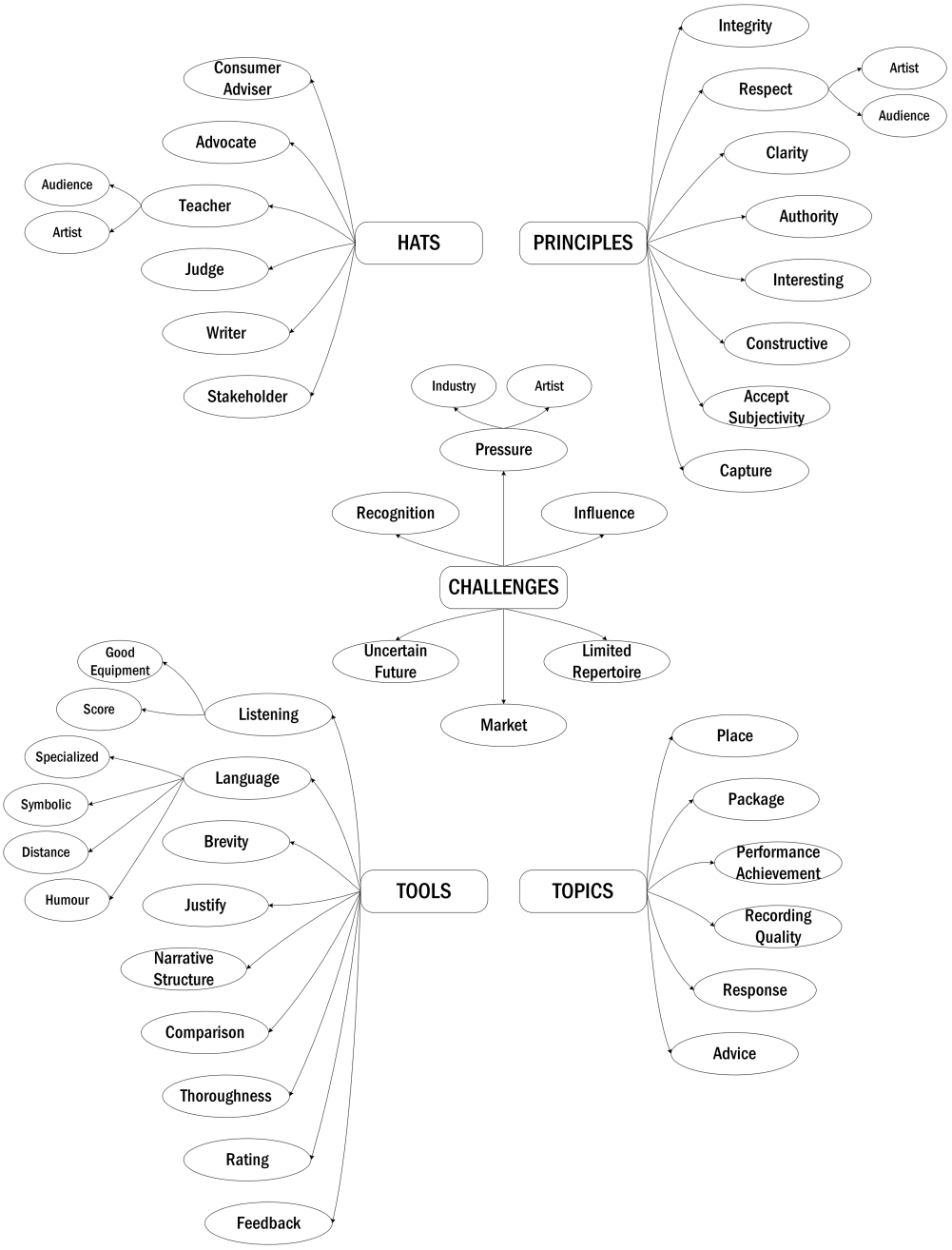


FIGURE 1
Visual model of the themes discussed by critics. Themes are organized hierarchically from rounded rectangles to ovals; arrows reinforce the visualization of this hierarchical structure.

needs to have the critical input, the validation from Critics" [C2] and "*The role today of professional criticism? Well, it is that conduit from the producer to the public. It is that bridge*" [C4].

Beyond commercial interests, **Hats** revealed human-centered dimensions, in line with Cottle (2003) remark on how journalists harbor a genuine desire to serve the public. This is reflected in three further roles that critics described, which focus on the service offered to artists and consumers. Critics saw themselves as musicians' **Advocates**, co-responsible to support the progress of an artist's career, and at the same time as **Consumer Advisors**, pledged to provide guidance to purchasing and listening behavior. The words used by critics emphasize their feeling of responsibility toward both audience and artists, for example: "*…that's the sort of thing I'm very, very aware of. I feel I'm doing it for the musicians. I'm writing for them*" [C2].

One last role that focuses the human-centered dimension of criticism is that of a **Teacher**. Critics saw in their work the potential to inform, illustrate, and educate, thus assisting listeners

to understand and appreciate the music performance, but also offering feedback to musicians on the value of their artistic choices. As such, the music critic "*today is also a social critic, a teacher, a pedagogue*" with "*pedagogical duties to fulfil*" [C11]. The view of the critic as a teacher seems to partially reflect the early 20th century music appreciation movement (e.g., Scholes, 1928; Jorgensen, 1987; Prictor, 1998; Witts, 2011). It also resonates with Cone (1981) distinction between the role of the "reviewer," whose aim is to guide listeners' choices (here this would be the **Consumer Advisor Hat**) and the proper "critic," whose aim is to broaden and deepen the reader's appreciation of music. What seems unique to our critics' description of their role, however, is that they see their pedagogical value addressed not only toward listeners, but also toward the musicians themselves.

One role that was largely absent from the interviews was that of "Journalist." Our critics rarely used this term, preferring instead to define themselves as writers. This stands in contradiction to reports that arts journalists are increasingly seeking solidarity in news organizations as "reporters" (Hellman and Jaakkola, 2011). Our critics' professional self-concept more closely follows an aesthetic paradigm that defines them as "connoisseurs and ultimately experts" (Barzun, 2001, p. 71) and "representative(s) of the artistic field in the newspapers" (Hellman and Jaakkola, 2011, p. 785). This may be a unique feature of classical music critics who have multiple opportunities to write specialized articles for general outlets, offering a critical product that is "less reportage than interpretation" (Muller, 2005, p. 105).

Principles – Things I must be or have

In the second theme category, critics described eight core conventions or moral standards governing their writing. These principles find parallels in the five ideal-typical values of journalism proposed by Deuze (2007) and align with Harries and Wahl-Jorgensen (2007) who interviewed arts critics from a wider genre base and reported a set of rules that represents critics' "code of conduct."

Our critics principles align with the 'Hats' they described. Three principles revolve around critics' main functions as assessors and stakeholders of the music industry: **Integrity**, **Authority**, and **Accepting Subjectivity** set out the grounds upon which critics' judgments build and profile critics' statements as, essentially, an informed opinion. Music critique judgments, according to our critics, should be based on a solid foundation of knowledge and extensive experience in the field, which give critics the **Authority** to command their position. At the same time, the critic should avoid normative statements and present him/herself as a provider of a well-informed, but ultimately personal judgment given at a particular time and place (**Accepting Subjectivity**). In a critic's words: "*Because we do have this, kind of, idea, this false idea I think, that Reviewers are objective. I mean, you are objective to a certain extent, but a lot of it is based on subjective opinion, …you, kind of, have to accept it as part of it and say, well, this is an informed subjective opinion, but it's still a subjective opinion*" [C7]. Critics' call to accept the subjectivity inherent in any aesthetic judgment (Harries and Wahl-Jorgensen, 2007; Hellman and Jaakkola, 2011) counters the value of 'objectivity' in Deuze (2007) and reflects theories in aesthetics that date back to Hume's Standard of Taste (Levinson, 2002; Budd, 2007) as well as current models in economics of information that set music as stereotypical "experience good" (Nelson, 1970; Mudambi and Schuff, 2010). According to our critics, objectivity in music criticism is replaced by expertise (**Authority**) combined with impartiality and truthfulness (**Integrity**). In particular, the **Integrity** principle in our critics' words seems to them arise both values of "ethics" and "autonomy" found in Deuze (2007): Critics should remain true to their own response to music, free from prejudices or conflicts of interest, and open-minded to new ideas and interpretations. Critics define this as an "*element of courage in reviewing*" which requires people *"to stick their neck out"* and *"to be prepared to say what you believe, and what you think*" [4].

Building on these three pillars of critical judgments, two further principles focus on the human-centered dimension of critique, in line with critics' roles as pedagogues, advocates, and consumers' guides. In communicating their judgments, critics should be aware of and understand the expectations, efforts and standpoints of the people involved (**Respect**). Again, critics' feeling of responsibility apply to both the audience and the artist, thus strongly resonating with Deuze (2007) dimension of "public service": critics should have a keen sense of the audience's knowledge and appreciate the readers' perspective. At the same time, they should respect the musician's feeling and sensitivity and actively try to understand what s/he may have tried to achieve. **Respect** toward the artist was described in interview as a fundamental rule of critique: "*The core principle is always …to take the person, who is offering me the recording, seriously. And this means that I have to ask myself, what does s/he want to tell me?*" [C11]. This in turn translates into a form of criticism which ought to be **Constructive**, to offer an evaluation that is potentially beneficial to the musician and avoiding a damning review: "*I do not like, basically, the negative criticism. I think criticism should … be constructive. You should be saying something which could be just possibly helpful*" [C8].

Building on the principle of constructing review, the last three principles described by critics focus on the way the review is written, setting critics' writer role to the front. Interestingly, these principles reflect broadly Beardsley's triadic theory of aesthetic value in the arts (Beardsley, 1962, 1982), which has been found to be reflected also in critics' evaluations of music recordings (Alessandri et al., 2016a). In interviews, critics pledged for reviews to be immediately understandable to the reader, coherent and user-friendly (**Clarity**), to be engaging and pleasurable to read, able to catch the reader's attention and arouse his/her curiosity (**Interesting**), and to represent and share the spirit of and passion for the music as well as a sense of the listening experience in words (**Capture**). Principles like **Capture**, **Interesting**, and **Clarity** accent a further dimension in music critics' values and professional

self-concept; as communicators, translators of knowledge, and sources of inspiration. The fact that these principles roughly align with the criteria of clarity, intensity and complexity that emerged in Alessandri's analysis of published music reviews (2016a) emphasizes critics' role as creative agents and suggests that music reviews – on top of the different functions they fulfill – might be seen as a work of art in its own right, as a piece of art evaluating art. Critics' words in interview well convey the view of review as a creative product: reviews ought not just to be clear and informative, they have to 'captivate' and 'charm' the reader [C10] and even become the written essence of the music. As a critic said: "*I want to …endlessly recreate it in my work, to recreate a spirit of someone's performance …in words*" [C8].

Challenges – Things I feel about my job

After 'Hats' and 'Principles' the third theme category highlights six key **Challenges** that arise from the need to juggle responsibilities toward artists, audience, and the recording industry while remaining true to an implicit code of conduct. **Challenges** are discussed in terms of the circumstances that critics negotiate, and how this makes them feel.

Two **Challenges** highlight conflicts between critics' roles, principles, and the context in which critics act. Critics are aware of the potential impact of their writing, both negative and positive (**Influence**). This is the challenge of having power, and critics discuss this in terms of potentially misleading consumers, damaging a person's career or increasing/decreasing sales and publicity: "*I know that they are liable to use my words to advertise that CD and to advertise the Pianist in general. So, I'm aware of the power and the power of the press*" [C3]. The awareness of the impact of their critique, mixes with a feeling of **Pressure** arising from personal or indirect reports criticizing the critic's work or encouraging them to provide a certain opinion or information in their review, or to use a certain tone or language.

The source of pressure can be the artist or their representatives but also the recording industry in general, including labels, magazines, record producers, and the dynamics between them. Conflicts of interests can arise from entertaining relationships with artists or having personal sympathies, and critics warn about getting "*too close to the people in the business, so close, you cannot be truthful*" [C2]. On the other hand, even in absence of sympathies or relationships, critics are aware of the possible consequences of their writing, in terms of reactions toward the critics themselves: "*…you have to always be thinking about the legal consequences, you do not want to libel anyone …you have to be quite careful with your language to make sure that you do not say anything that they could take you to court over*" [C7].

As stakeholders of the music industry critics feel pulled in different directions, stretched through the "*inextricable link*" between "*the commercial life of the record industry and …how record magazines cover these records*": "*…although it would never say so, the Gramophone has an agenda, which is to promote current recording and the critical faculties will follow from that*" [C1].

Personal interests and biases as well as the consequences of their critique clash with the need for fair and impartial judgments ("**Integrity"** Principle) and the desire to guide and support consumers and artists (Hats "**Consumer Guide**" and "**Advocate**"). While conflicts of interest and pressure from the industry have been reported in other art criticism contexts (Chong, 2017), one challenge that does not find resonance in the literature is the feeling of responsibility that the music critics bear toward classical music artists, nurtured by the awareness of the impact of the press on musicians' self-concept and career.

Besides **Pressure** and **Influence**, one further challenge points at an inner conflict between critics' day-per-day job requirements and their creative and aesthetic needs (Hat "**Writer"**). In interview, critics bemoaned to some extent being asked to review the same music works many times or having limited freedom in the choice of what to review (**Limited Repertoire**): "*…if you want longevity with the magazines …you have to take what they send you. They tend to send you what you have done before, so there's very little renewal in the Reviewer's frame of reference*" [C1].

Finally, the last three challenges them arise the complexity of being a music critic in the context of the current communication and music consumption market: critics operate today in a large field of published opinion and coexist (and compete) with multiple novel channels of communication like blogs, Twitter or Amazon (**Market**): "*…the freeness of the Internet is a great boon in some ways, but it's a disaster in others, because it's overload, information overload …And, you know, it's very difficult to weed out what opinions are worth reading, for readers*" [C4]. Their role as mediators in the music industry is more relevant than ever, and yet critics bemoaned this position in the modern consumer market "*Somewhere in-between a diffuse, heavily changing public*" [C12]. The complexity of this scenario raises questions on the very nature of the critic as consumers' guide: "*I do not know how people are going to consume music. The question is if you have got YouTube and you have got iTunes and, you know, all these massive channels for acquiring music, how do you guide people and do people want to be guided?*" [C1]. This adds to a general feeling of disconnect that critics described, between the expected and actualized response to their work (**Recognition**). This includes issues around payment, misunderstanding of their aims or meaning, as well as a perceived overly negative portray of critics: "*And I've always said to people, 'Look, everybody hates Critics. Get your machine guns out'*" [C8]. These circumstances and the changes in the industry and consumer habits nurture in critics a deep concern for the legitimization and even meaningfulness of critical practice itself. In line with the recent acknowledgement of a 'crisis' within arts journalism (Jaakkola, 2015; Baía Reis, 2018), critics in interview shared thoughts about the **Uncertain Future** of their profession, including the idea that it is losing volume and significance. In the words of critics: "*…I think record criticism is declining. It will probably dissolve*" [C10] and

"*…the social structures and social mores have given people confidence to make their own decisions. So, the role of the Critic's not necessary*" [C1].

The first three theme categories above depicted critics' role, the principles they follow, and the challenges they face. The lower half of the visual model features the two remaining themes of **Topics** and **Tools,** which focus on the strategies and devices critics employ to fulfil their purposes, be true to their principles, and deal with challenges.

Topics – Things I discuss

Topics define the aspects of the recorded performance covered in review. This includes seven subthemes that detail comments on the context, the product, the music, and the critic's reaction. The seven **Topics** that we identified are a good match to those reported previously, and which were based on the analysis of critical review content from one outlet, the Gramophone magazine (Alessandri et al., 2015, 2016b).

In line with those previous findings, critics in interview confirmed that the core content of their writing is the description and evaluation of the musical performance. This includes comments on style, originality, communication, interpretation, as well as comments on musical parameters like tempo or phrasing (**Performance Achievement**). For example: "*…you are saying how the performance is, what was good about the performance, you know, the expression and the phrasing*" [C4] or "*…these are important things to cover. The liveliness of the Gestaltung and of course the faithfulness to the text*" [C9]. As can be seen from these excerpts, in the interviews critics did not go into detail on the discussion of the **Performance Achievement**, limiting themselves instead to a few examples of themes therein. When asked for more details they tended to provide concrete musical examples, instead of venturing into a general categorization of performance. This pattern of behavior makes sense when interpreted in the light of previous findings; discussion of a performance tends to form the largest part of review and is characterized by a complex variety of descriptive and value-laden constructs.

The discussion of the performance merges in review with the description and evaluation of the recorded sound (**Recording Quality**), consistent with Alessandri et al. (2016b). In line with Philip (2004) and Patmore and Clarke (2007) critics thematized within this theme the importance of recreating through the recording the impression of a live performance, "*I will certainly comment on …whether it sounds like a studio recording or whether the artist had been able to transcend the recording studio and give me the impression that it's a live important event, which is just happening*" [C3].

In addition to the performance and the recorded sound the recording product itself is an object of discussion in review (**Package**). This topic clusters different sub-level themes found in Alessandri et al. (2016b), which are all elements *extra to the actual sound*, e.g., the program performed, the composer, the instrument and score edition used, but also sleeve notes, cover art design, comments by the artist or issues of translation.

Two further **Topics** are used to contextualize the recording in terms of its history and its **Place** in the emerging market, and to offer information about the performing **Artist**, their career, school their come from, track record of recordings and general skills: "*I always like to contextualize a record, you know, when was it made? Why was it made? Who was it made for? What were the circumstances around the recording?*" [C1]; "*…I would then offer a few background information. Biographical information of the interpreter, what has s/he done so far, to introduce the musician a bit*" [C12].

The last two **Topics** focus on the critic's affective reaction to the music (**Response**) and his/her recommendation for the reader, in terms of whether to buy or whether and how to engage with the recording (**Advice**). These last **Topics** are the only ones that do not find a direct correspondent in the music review model from our previous work (Alessandri et al., 2015, 2016b). This discrepancy might be explained by the fact that critic's **Response** and **Advice** are always about an element of the recording – e.g., the artist, the performance, or the recorded sound – and thus have been coded in the previous work under such themes. That critics in interviews described these as separate categories offers new insights concerning the motivation behind those statements, which strongly aligns with critics' felt responsibilities as advisors and teachers (Hats). The weight given to the critic's own **Response** to music also resonates with the typical amount of affective evaluative terms (e.g., moving; daunting; cloying) used in reviews (Alessandri et al., 2015) and with findings on the importance of emotional response for the evaluation of art (Chong, 2017).

The high convergence of evidence between what critics told us about their writing and what emerged previously from the text analysis of published reviews adds support to the proposed model structure and construct conceptualization. Moreover, this indicates that the model previously developed from only one specialized magazine reflects well the content from other sources and authors. One area of music critical writing untouched by previous examination of reviews were the writing strategies employed to discuss and structure topics; these would have been impossible to presume from the written source alone. These writing strategies are the focus of the last theme category.

Tools – Things I use when writing

This theme category entails nine subthemes containing comments on devices and strategies that critics may employ whilst reviewing a recording, i.e., from the minute they listen to the music to the final production of the document.

The first two **Tools** concern the act of **Listening** itself and the reliance on colleagues, editor, and/or the artist(s) **Feedback** during the review process. Critics in interview emphasized the importance of using a high-quality reproduction system (at least at some stage of listening) and

having the score to hand as reference material while listening. They also reported that they actively seek discussion with colleagues or artists to clarify questions or just have "*an informed discussion*" about the recording [C7]: "*It also happens that I actually call the agency or the CD label and say: 'I would like to briefly talk to the pianist. I simply have specific questions.' …then we talk*" [C11].

A further block of themes within **Tools** covers structural and literacy devices used in writing. In line with previous findings (Alessandri et al., 2015) and with anecdotical reports (Pollard, 1998) **Comparisons** between the recording reviewed and other recordings or related experiences (e.g., seeing the performance live) are a common device in critics' writing toolbox: "*A bad review considers the work or the CD as an individual object. …But a good one shows that this CD indeed does not stand alone, it is anchored in a wider space, in a repertoire*" [C14]. In addition, the use of a concise writing style was presented as an essential requirement in review practice and in journalism in general (**Brevity**): "*…space is the crucial thing, always, with journalism. …if you are only allowed a small slot, you must keep within that slot and it does limit what you can do*" [C2]. In their study of arts journalism in Finland, Hellman and Jaakkola (2011) interpreted a reduction in review length as a sign of a shift from the aesthetic to the journalistic paradigm. Indeed, all our critics discussed word counts as a challenge to a thorough and insightful music review. However, they framed this issue in terms of the need for concise writing skills and awareness of target audience rather than as an invitation to change their approach. This concise writing style was reflected in the high density of themes per clause found in previous analysis of written text (Alessandri et al., 2015).

A third structural writing tool concerns the use of story elements within the review, like a clear headline, distinct opening or closing statements, and a core message or angle. In reviews critics offered concrete examples of their preferred **Narrative Structure**: "*…begin with a fanfare. So get the reader's attention. And end with a cadence, so you get the feeling at the end that, yes, this is the end of the review. We have come to a conclusion*" [C6].

Critics also commented on the use of different vocabulary and linguistic styles to describe the recording (**Language**), for instance weighing the use of musical terms (e.g., fermata, counterpoint, Leitmotiv; **Specialized**) and figurative speech (like metaphors, similes or personifications; **Symbolic**) according to the target audience, or using wit, satire or irony (**Humor**) as well as first and third person (**Distance**) to shape the character of their writing. In line with the Principles of **Respect** and **Clarity**, critics found that music specialized terms should be used sparely, while metaphors and similes were appreciated as a way to "*color your writing*" [C5]. Mixed feelings were expressed toward the use of quantified evaluations in review, like numbers or stars (**Rating**). Some critics found these could be an added value for consumers, facilitating comparison between recordings, while other warned about the risk of over-reducing the critical appreciation to a quantified value: "*The star system simplifies things at times, but this simplification takes away the possibility to undergo very differentiated experiences*" [C12].

The last two **Tools** that the critics discussed appeal to more abstract but essential strategies in reviewing: the accuracy and diligence in the reporting of details about a recording (**Thoroughness**) and the justification of value judgments by means of factual and rational writing combined with the use of examples to back up assessments (**Justify**). The act of adducing reasons to support evaluative judgments was described by critics as the fundamental nature of reviewing, "*the essence of all of this is about …you have to justify and make clear your process of thinking*" [C5].

The importance of grounding critical judgments of recorded music in reason resonates with a wider philosophical debate on the nature of criticism, which sees the reasoning process behind evaluation as the defining trait of critique, as opposed to a more information and description-based journalism (Cone, 1981; Beardsley, 1982; Carrier, 1986; Davies, 2001; Danto, 2002; Carroll, 2009). This understanding is also grounded in the history of the critical practice (Carroll, 2009, p. 16). We find this assumption for instance already underpinning the two seminal essays on music criticism by Calvocoressi (1923) and Newman (1925), and also, 40 years later, in Walker's *An Anatomy of Musical Criticism*, in which this idea is stated explicitly at the opening (1968, p. xi):

> "The practice of criticism boils down to one thing: making value judgments. The theory of criticism, therefore, boils down to one thing also: explaining them. If you formulate a theory of criticism, it is not enough to know that one work is a masterpiece and another is a mediocrity. You must also explain why they are different."

The previous work done by the authors (Alessandri et al., 2015) supported this view based on a large sample of evidence from published reviews. The analysis showed that critics' texts contained a large variety of descriptors adduced as reasons to support judgments. Descriptors were divided into two major categories, which resonate with the different use of **Specialized** and **Symbolic** language: technical constructs like sound parameters and mechanics of delivery, and abstract constructs like character, structure, or style, where critics made use of metaphors and similes to convey their impressions.

The current interview work further adds to these previous findings, confirming that this very quality of review is intentional and that critics are well aware of this. In the words of one of our critics: "*You argue. You reason, exemplify, and justify. This is critique*" [C11].

The emphasis on reasoning given by our music critics thus supports a professional self-concept distinct from that of a more general journalist or reporter. The use of rhetorical and stylistic tools as first-person or wit, together with assertions on the importance of emotional **Response**, is in line with literature describing the increasing relevance of emotion-related statements in music criticism as a means to achieving a more engaging and

direct form of communication (Wahl-Jorgensen, 2012; Coward, 2013).

Taken together, the **Tools** category again offers a good match with the results of previous analysis (Alessandri et al., 2015, 2016a,b), noting that this has been the first opportunity to verify the contents of written review with verbal confirmation of the intention behind the source output.

Discussion

We interviewed 14 expert music critics from United Kingdom, Germany, and Switzerland to understand how they view their role and practice. The resultant visual model offers a detailed, reflective map of the nature of criticism in the classical music market, comprising music critics' opinions and beliefs regarding their impact on consumers and artists as well as how these thoughts inform their writing process.

Critics in the modern classical market

The model generated from interviews with critics self-identifies them as "cultural intermediaries" (Bourdieu, 1984, p. 325) between classical music producers, artists and consumers: As a bridge that fulfils a variety of purposes for each industry stakeholder.

Following Kristensen and From's (2015b) typology, our music critics can be included under the heading of cultural journalists: passionate professionals, who aim to deliver aesthetic evaluations grounded in clear reason (*Judge*; *Justify*) while offering an engaging literary product (*Writer*; *Interesting*). Their profile defines them as intellectual cultured critics, driven by a sense of responsibility to create accessible and relatable knowledge for all their perceived stakeholders (*Teacher*, *Clarity*, *Respect*; *Constructive*). Our critics remain "devoted to the comparison and analysis, to the interpretation and evaluation" (Cuddon, 1982, p. 207), triggered by the feeling of "fulfilment of a duty toward a matter" (Adorno, 1998, p. 142).

The music critics voiced beliefs in line with those of other arts journalists, highlighting their role beyond the news agenda (Harries and Wahl-Jorgensen, 2007). In line with Kristensen and From's (2015b) typology and Harries and Wahl-Jorgensen's (2007) theory of "arts exceptionalism," our findings clarify the cultural journalist profile and show how it is experienced by seasoned music critics both in terms of responsibility and concerns. In positioning themselves squarely in the aesthetic paradigm of occupational professionalism (Örnebring, 2009), many music critics report struggling with an arts journalism archetype that is shifting toward a media-led organizational standard.

Our critics were aware of their multiple roles within the music market and the potential for controversial consequences (*Stakeholder*; *Consumer Advisor*; *Pressure*; *Influence*; *Market*), and yet they emphasized their drive to be conveyors of culture, advocates of music and of musicians, and teachers (*Teacher*; *Advocate*; *Respect*; *Constructive*; *Capture*; *Interesting*). In a music world that often appears to be dominated by prejudices against them (Brennan, 2006), the critics' pledge to the aesthetic paradigm: their passion for music combined with their desire to share knowledge and serve musicians and listeners alike emerged from the interviews as a call for understanding and acknowledgment (*Recognition*).

Is professional music criticism dying?

The call for recognition amongst music critics gains urgency in the context of new opinion sharing and communication channels. Our music critics identified online blogs and digital magazines as both a resource and a threat to quality criticism (*Market*). This conflict reflects a wider debate on the shifting role of journalism in the digital era (Agarwal and Barthel, 2015). Our music critics observe this shift with concern and scepticism; in line with Deuze (2007), they fear a marginalization of professional critique in the digital media age (*Market*; *Uncertain Future*). However, the opposite position in this debate, that of a constructive integration of professional criticism into a hybrid media system (Chadwick, 2013) was also reflected. Some critics entertained the idea of fusing traditional and new practices to redefine the nature of music critique in the near future.

Of deeper concern to our participants was the perceived *raison d'être* of music criticism in view of modern music consumption. In the age of Spotify and YouTube, the critics questioned what kind of guidance, if any, consumers need when music is low cost (or free) and selected by computer algorithms. This question was accompanied by feelings of resignation and marginalization (*Recognition*, *Uncertain Future*), but was also met by a strong sense of purpose and self-identity: classical music critics emphasized the importance of their autonomy, today more than ever. They outlined a set of norms (**Principles**) and job roles (**Hats**), grounding their critical identity firmly in their unique expertise (*Authority*), aesthetic purpose (*Teacher*), and third-party perspective (*Judge*; *Integrity*; *Comparison*).

Informed judgment as added value

Critics described their ultimate value in terms of a benefaction for the music listening public. Their skill is in taking the aesthetic response that we all experience and transforming it into a public discourse. Only by this transformation does the aesthetic judgment obtain importance: "Through the relationship with the reading public, critical reflection loses its private character." (Eagleton, 1984).

The justification of aesthetic judgments in terms of *Authority* and *Respect* seems at first glance to reflect an elitist image of the cultural critic, which conveys not only knowledge but also actions to consumers (Dahlgren, 2012). However, music critics are clear

that what they offer is only an informed evaluation (*Accept Subjectivity*). Its value resides in their knowledge as well as the principles and journalistic skills embedded in and sparked by – paraphrasing one of the critics – a burning desire to share their lifelong love for music.

In a consumption market characterized by quick and free opinion, stars and thumbs up, classical music critics pleas for deeper engagement with their text and with music listening are challenging. However, the market is expanding in terms of music devices and recordings (Krause et al., 2013) and one immediate consequence of this trend is paralysis of choice (Schwartz, 2008); consumers can be left unsatisfied with the music selected for them and feel mislead by judgments that they perceive as ill-informed or created by artificial means. This situation has led to some advocating random selection of music as the only reasoned approach (Leong et al., 2008). It is an irony that the same market which critics have come to view with suspicion may need them now more than ever.

The job of music critique

The act of music criticism has been defined as "the translation and grading of an aesthetic experience by means of intellectual analysis and imaginative inquiry" (Dean, 1980, p. 44). By asking critics about their practice, we gained insights into the tools of this trade that enable them to produce quality content.

The subheadings within **Topics** and **Tools** provided a good match with the constructs from previous written review analysis (Alessandri et al., 2015, 2016a,b), indicating a high correlation between intent and outcome in critical review. However, within the current model, aspects of recordings and the way these are discussed did not play a central role. Rather, **Hats**, **Principles,** and **Challenges** emerged as dominant, complex theme clusters. This meta-reflection on the job of being a music critic usually remains hidden to the reader. In previous analyses, it was shown that critics can sporadically let "slip" their thoughts about review writing, its processes and challenges (*Meta-Criticism*, Alessandri et al., 2016b). However, this was always a minor point, and it is interesting to note that none of the critics in the present paper mentioned that they wrote about these issues. This suggests that sporadic meta-reflections on the critical practice itself in review reflect an inner need for explanation and understanding.

Conclusion

Critical reviews of music recordings are a common and relevant form of performance evaluation. Building on previous *post hoc* research on the content of critical writing, in this article we report findings from interviews with professional critics that offer insights on the intentions, motivations, and principles behind this well-established form of critical appraisal. Our visual model of music critique brings together many layers and facets of critics' professional self-concept in combination with their experience and practice. As such, it adds a new dimension to the music criticism literature and gives insights into the mechanisms and reasons behind critics' evaluations and into the key elements of critical reviews, which experts see as influential and relevant for consumers. It also shows the challenges critics face, standing as they do in an intermediate position between the producers and consumers of classical music, as well as straddling a complex intersection between artist, journalist and educator.

Ultimately, these findings offer a new interpretative viewpoint on critics' aesthetic judgments and on their perceived place within the digital classical musical world. They bring a message of hope: Although many critics spoke of their fear for the future, the engaged and multifaceted evaluative approach they bring to music gives good reason to believe that their unique abilities will be in increasing demand by the sophisticated music consumer who asks for more and not less informed control over their choices.

Data availability statement

The datasets presented in this article are not readily available because the sharing of the original interview transcripts would mar the anonymity requirement. Relevant excerpts from interviews are included in the manuscript and in the supplementary material. Requests to access the datasets should be directed to the corresponding author.

Ethics statement

The studies involving human participants were reviewed and approved by Research Ethics Committee, Department of Music, University of Sheffield. The patients/participants provided their written informed consent to participate in this study.

Author contributions

VW was responsible for the ethical approval. EA ran the interviews. EA and VW wrote the first draft of the manuscript. All authors contributed to the design of study, participant recruitment, and data analysis, reviewed and edited the manuscript, and approved the final version.

Funding

This study was supported by the Swiss National Science Foundation (grant number 100016M_162819). Open access funding provided by Lucerne University of Applied Sciences and Arts.

Acknowledgments

We wish to thank all our music critics for contributing their time and expertise to this study. We gratefully acknowledge Katrin Szamatulski for her assistance in transcribing the interview data, reviewing literature, and proofreading the manuscript.

Conflict of interest

The authors declare that the submitted work was carried out in the absence of any personal, professional, or financial relationships that could potentially be construed as a conflict of interest.

Publisher's note

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2022.925394/full#supplementary-material

References

Abbate, C. (2004). Music: drastic or gnostic? *Critical Inquiry* 30, 138–153. doi: 10.1086/421160

Adorno, W. (1998). "Late word without late style," in *Beethoven: The philosophy of music*. ed. R. Tiedemann (Stanford: Stanford University Press), 138–153.

Agarwal, S. D., and Barthel, M. L. (2015). The friendly barbarians: professional norms and work routines of online journalists in the United States. *J. Theory Pract. Crit.* 16, 376–391. doi: 10.1177/1464884913511565

Alessandri, E., Eiholzer, H., and Williamon, A. (2014). Reviewing critical practice: an analysis of Gramophone's reviews of Beethoven's piano sonatas, 1923–2010. *Music. Sci.* 18, 131–149. doi: 10.1177/1029864913519466

Alessandri, E., Williamson, V. J., Eiholzer, H., and Williamon, A. (2015). Beethoven recordings reviewed: a systematic method for mapping the content of music performance criticism. *Front. Psychol.* 6, 1–14. doi: 10.3389/fpsyg.2015.00057

Alessandri, E., Williamson, V. J., Eiholzer, H., and Williamon, A. (2016a). A critical ear: analysis of value judgements in reviews of Beethoven's piano sonata recordings. *Front. Psychol.* 7, 1–17. doi: 10.3389/fpsyg.2016.00391

Alessandri, E., Williamson, V. J., Eiholzer, H., and Williamon, A. (2016b). "Evaluating recorded performance: An analysis of critics' judgements of Beethoven piano sonata recordings," in *Proceedings of the 14th International Conference for Music Perception and Cognition*. ed. G. Vokalek (San Francisco, USA: ICMPC14), 19–24.

Arditi, D. (2018). Digital subscriptions: the unending consumption of music in the digital era. *Pop. Music Soc.* 41, 302–318. doi: 10.1080/03007766.2016.1264101

Aschenbrenner, K. (1981). "Music criticism: practice and malpractice," in *On criticizing music. Five philosophical perspectives*. ed. K. Price (Baltimore: Johns Hopkins University Press).

Baía Reis, A. (2018). Is Portuguese theatre criticism still relevant? *Sinais de Cena* II:3

Baldassarre, A. (2009). "'Critiche stupide, ed elogi più stupidi ancora. spropositi e sciocchezze sempre': Konstanten und Besonderheiten in der europäischen Verdi-Rezeption des 19. Jahrhunderts," in *Wie europäisch ist die Oper? Das Musiktheater als Zugang zu einer kulturellen Topographie Europas*. eds. P. Ther and P. Stachel (Munich: Oldenbourg Wissenschaftsverlag), 127–153.

Baldassarre, A., Alessandri, E, and Williamson, V. J. (2022). Forthcoming. Changing Times: The Evolution of the Persona of the Classical Music Critics.

Baldassarre, A., and Alessandri, E. (2022). "Musikkritik in Zeiten der Digitalisierung," in *Musikjournalismus: Radio - Fernsehen - Print - Online*. ed. P. Overbeck (Wiesbaden: Springer Verlag), 35–41.

Barzun, J. (2001). *From Dawn to decadence, 1500 to the present: 500 years of Western cultural life*. Harper Collins.

Beardsley, M. C. (1962). On the generality of critical reason. *J. Philos.* 59, 477–486. doi: 10.2307/2023219

Beardsley, M. C. (1982). "The relevance or reasons in art criticism," in *The aesthetic point of view: selected essays*. eds. M. J. Wreen and D. M. Callen (Ithaca, N.Y: Cornell University Press), 15–34.

Becker, H. (1965). *Beiträge zur Geschichte der Musikkritik*. Regensburg: Gustav Bosse Verlag.

Becker, H. S. (1982). *Art worlds*. Berkley and Los Angeles, CA: University of California Press.

Benjamin, W. (1980). "Das Kunstwerk im Zeitalter seiner technischen Reproduzierbarkeit (1939, 3rd and last authorized version)," in *Walter Benjamin, Gesammelte Schriften*. eds. R. Tiedemann and H. Schweppenhäuser, *Vol. I/2* (Frankfurt/Main: Suhrkamp), 471–508.

Blank, G. (2007). *Critics, ratings and society: The sociology of reviews*. Lanham, MD: Rowman & Littlefield.

Boenisch, V. (2008). *Krise der Kritik? Was Theaterkritiker denken – und ihre Leser erwarten*. Berlin: Theater der Zeit.

Bourdieu, P. (1984). *Distinction: a social critique of the judgement of taste* Cambridge, Mass.: Harvard University Press.

Brendel, F. (1855). "Musikkritik," in *Geschichte der Musik in Italien, Deutschland und Frankreich von den ersten christlichen Zeiten bis auf die Gegenwart. 2nd Edn.* (Leipzig: Matthes).

Brennan, M. (2006). The rough guide to critics: musicians discuss the role of the music press. *Pop. Music* 25, 221–234. doi: 10.1017/S0261143006000870

Buck, P. C. (1905). Prolegomena to musical criticism. *Proc. Music. Assoc.* 32, 155–160. doi: 10.1093/jrma/32.1.155

Budd, M. (2007). The intersubjective validity of aesthetic judgements. *Br. J. Aesthet.* 47, 333–371. doi: 10.1093/aesthj/aym021

Bujic, B. (2011). Criticism of music. in A. Latham (Ed.), *The oxford companion to music*. Oxford: Oxford University Press.

Burgess, R. J. (2014). *The history of music production*. Oxford and New York: Oxford University Press.

Caduff, C. (2014). Kritik – Niedergang oder Neuformierung? *Jahrbuch für Kulturmanagement* 21, 149–160. doi: 10.1515/transcript.9783839419632.149

Calvocoressi, M. D. (1923). *The principles and methods of musical criticism*. London: Oxford University Press.

Carboni, M. (2012). "*Proceedings of 2012 international conference on economics, business and marketing management*," in *The classical music industry and the future that digital innovations can bring to its business models*, vol. *29* (Singapore: IACSIT Press), 343–347.

Carrier, D. (1986). Philosophical art criticism. *Leonardo* 19, 170–174. doi: 10.2307/1578285

Carroll, N. (2009). *On criticism*. New York: Routledge.

Chadwick, A. (2013). *The hybrid media system: Politics and power*. Oxford: Oxford University Press.

Cheyne, A., and Binder, A. (2010). Cosmopolitan preferences: the constitutive role of place in American elite taste for hip-hop music 1991–2005. *Poetics* 38, 336–364. doi: 10.1016/j.poetic.2010.01.001

Chong, P. (2017). Valuing subjectivity in journalism: bias, emotions, and self-interest as tools in arts reporting. *J. Theory Pract. Crit. Adv.* 20, 427–443. doi: 10.1177/1464884917722453

Cone, E. T. (1981). The authority of music criticism. *J. Am. Musicol. Soc.* 34, 1–18. doi: 10.2307/831032

Cottle, S. (2003). *Media organization and production*. London: Sage.

Coward, R. (2013). *Speaking personally: The rise of subjective and confessional journalism*. Baingstoke: Palgrave Macmillan.

Crossley, N., and Bottero, W. (2015). Music worlds and internal goods: the role of convention. *Cult. Sociol.* 9, 38–55. doi: 10.1177/1749975514533209

Cuddon, J. A. (1982). "The penguin dictionary of literary terms and literature," in *Penguin*. 3rd ed

Dahlgren, P. (2012). Public intellectuals, online media, and public spheres: current realignments. *Int. J. Pol. Cult. Soc.* 25, 95–110. doi: 10.1007/s10767-012-9124-5

Dahlhaus, C. (1971). "Probleme der Kompositionskritik," in *Über Musik und Kritik*. ed. R. Stephan (Mainz: B. Schott's Söhne), 9–18.

Dahlhaus, C. (1981). E. T. A. Hoffmanns Beethoven-Kritik und Die Ästhetik des Erhabenen. *Arch. Musikwiss.* 38, 79–92. doi: 10.2307/930602

Danto, A. (2002). From philosophy to art criticism. *Am. Art* 16, 14–17. doi: 10.1086/444655

Datta, H., Knox, G., and Bronnenberg, B. J. (2017). Changing their tune: how consumers' adoption of online streaming affects music consumption and discovery. *Mark. Sci.* 37, 5–21. doi: 10.1287/mksc.2017.1051

Davies, S. (2001). *Musical works and performance: A philosophical exploration*. New York: Oxford University Press.

Dean, W. (1980). "Criticism," in *The new grove dictionary of music and musicians*. ed. S. Sadie, *Vol. 5.* (London: Macmillan), 36–50.

Debenedetti, S. (2006). The role of media critics in the cultural industries. *Int. J. Arts Manag.* 8, 30–42.

Deuze, M. (2007). *Media work*. Cambridge: Polity Press.

Diederichs-Lafite, M. (1996). Prinzipien der Musikkritik. 9. Muiksgespräch mit dem Bertelsmann-Musikkritikseminar. *Österreichische Musikzeitung* 51, 504–513.

Dingle, C. (Ed.) (2019a). *The Cambridge history of music criticism*. Cambridge: Cambridge University Press.

Dingle, C. (2019b). "Comparing notes: recording and criticism," in *The Cambridge history of music criticism*. ed. C. Dingle (Cambridge: Cambridge University Press), 249–271.

Doran, R. (2020). *Liszt and virtuosity*. Rochester, N.Y: University of Rochester Press.

Eagleton, T. (1984). *The function of criticism*. London: Verso.

Eatock, C. (2004). Classical music criticism at the globe and mail: 1936-2000. *Can. Univ. Mus. Rev.* 24, 8–28. doi: 10.7202/1014580ar

Ellis, K. (1995). Music criticism in nineteenth-century France: La revue et gazette musicale de Paris 1834–80. *Cambridge Univ. Press*. doi: 10.1017/CBO9780511470264.010

Ellis, K. (2012). Opera criticism and the Paris periodical press. *Revue Belge de Musicologie/Belgisch Tijdschrift Voor Muziekwetenschap* 66, 127–131.

Elste, M. (2009). "A matter of circumstance: on experiencing recordings," in *The Cambridge companion to recorded music*. eds. N. Cook, E. Clarke, D. Leech-Wilkinson and J. Rink (Cambridge: Cambridge University Press), 116–119.

Ertelt, T., and von Loesch, H. (eds.) (2021). Geschichte der musikalischen Interpretation im 19. und 20. Jahrhundert, 2 vols. *Bärenreiter.* Kassel.

Fenner, T. (1994). *Opera in London: Views of the press, 1785–1830*. Carbondale, IL: Southern Illinois University Press.

Forde, E. (2003). "Journalists with a difference: producing music journalism," in *Media in Focus: Media organization and production*. ed. S. Cottle (London: SAGE Publications Ltd), 113–130. doi: 10.4135/9781446221587.n7

Fox Strangways, A. H. (1938/1939). The criticism of music. *Proc. Music. Assoc.* 65, 1–18. doi: 10.1093/jrma/65.1.1

French, R.F. (1948). *Music and criticism*. Cambridge, MA: Harvard University Press, doi: 10.4159/harvard.9780674332447.

Gooley, D. A. (2009). *The virtuoso Liszt*. Cambridge: Cambridge University Press.

Guest, G., MacQueen, and T., Namey, E. E., (eds.) (2012). Applied Thematic Analysis. Thousand Oaks, CA: SAGE.

Hamer, L. (2019). "Critique the canon: the role of criticism in canon formation," in *The Cambridge history of music criticism*. ed. C. Dingle (Cambridge University Press), 231–248. doi: 10.1017/9781139795425.013

Hanslick, E. (1870). *Aus dem Concertsaal: Kritiken und Schilderungen aus den letzten 20 Jahren des Wiener Musiklebens* W. Braumüller Verlag.

Harries, G., and Wahl-Jorgensen, K. (2007). The culture of arts journalists: elitists, saviors or manic depressives? Journalism: theory. *Pract. Crit.* 8, 619–639. doi: 10.1177/1464884907083115

Heikkilö, R., Lauronen, T., and Purhonen, S. (2017). The crisis of cultural journalism revisited: the space and place of culture in quality European newspapers from 1960 to 2010. *Eur. J. Cult. Stud.* 21, 669–686. doi: 10.1177/1367549416682970

Hellman, H., and Jaakkola, M. (2011). From aesthetes to reporters: the paradigm shift in arts journalism in Finland. *Journalism* 13, 783–801. doi: 10.1177/1464884911431382

Hellouin, F. (1906). Essai critique de la critique musicale. *A. Joanin Cie*

Henderson, W. J. (1915). The function of musical criticism. *Music. Q.* I, 69–82. doi: 10.1093/mq/I.1.69

Hesmondhalgh, D., and Meier, L. M. (2018). What the digitalization of music tells us about capitalism, culture and the power of the information technology sector. *Inf. Commun. Soc.* 21, 1555–1570. doi: 10.1080/1369118X.2017.1340498

Hracs, B. J., et al. (2016). *The production and consumption of music in the digital age* Routledge doi: 10.4324/9781315724003.

Jaakkola, M. (2015). Witnesses of a cultural crisis: representations of media-related meta-processes as professional meta-criticism of arts and cultural journalism. *Int. J. Cult. Stud.* 18, 537–554. doi: 10.1177/1367877913519308

Janssen, S., and Verboord, M. (2015). "Cultural mediators and gatekeepers," in *International encyclopedia of the social and behavioral sciences*. *2nd Edn.* ed. J. D. Wright, *Vol. 5*. (Elsevier), 440–446. doi: 10.1016/B978-0-08-097086-8.10424-6

Jorgensen, E. (1987). Percy Scholes on music appreciation: another view. *Br. J. Music Educ.* 4, 139–156. doi: 10.1017/S0265051700005908

Katz, M. (2004). *Capturing sound: How technology has changed music* University of Califronia Press.

Kirchmeyer, H. (2017). *System- und Methodengeschichte der deutschen Musikkritik vom Ausgang des 18. bis zum Beginn des 20* Jahrhunderts. Franz Steiner Verlag.

Kramer, L. (1989). Dangerous liaisons: the literary text in musical criticism. *19th Century Music* 13, 159–167. doi: 10.2307/746653

Krause, A. E., North, A. C., and Hewitt, L. (2013). Music-listening in everyday life: devices and choice. *Psychol. Music* 43, 155–170. doi: 10.1177/0305735613496860

Kristensen, N. N. (2010). Cultural journalism in the Danish printed press – a history of decline or increasing medai institutional profiling? *Nord. J. Media Stud.* 8, 69–92. doi: 10.1386/nl.8.69_1

Kristensen, N. N., and From, U. (2015a). Cultural journalism and cultural critique in a changing media landscape. *Journal. Pract.* 9, 760–772. doi: 10.1080/17512786.2015.1051357

Kristensen, N. N., and From, U. (2015b). From ivory tower to cross-media personas. *Journal. Pract.* 9, 853–871. doi: 10.1080/17512786.2015.1051370

Legard, R., Keegan, J., and Ward, K. (2003). "In-depth interviews," in *Qualitative research practice: A guide for social science students and researchers*. eds. J. Ritchie and J. Lewis (SAGE Publications), 138–169.

Leong, T., Vetere, F., and Howard, S. (2008). Abdicating choice: the rewards of letting go. *Dig. Creat.* 19, 233–243. doi: 10.1080/14626260802550777

Levinson, J. (2002). Hume's standard of taste. *J. Aesth. Art Crit.* 60, 227–238. doi: 10.1111/1540-6245.00070

Mattheson, J. (1722–1725). *Critica Musica*. Hamburg: Auf Unkosten des Autoris.

McGill, L., Conrad, W. J., Rosenberg, D., and Szántó, P. (2005). "Compilers and editors," in *The classical music critic: A survey of music critics at general-interest and specialized news publications in America. A collaborative project of the music critics Association of North America and the National Arts Journalism Program at Columbia University*, Music Critics Association of North America and National Arts Journalism Program, Columbia University. http://www.columbia.edu/cu/najp/news/pastnews/cmcsfinal.pdf

Melnyk, L. (2019). Who killed classical music criticism: social strategies of music journalism today. *Lietuvos muzikologija* 20, 20–30.

Monelle, R. (2002). "The criticism of musical performance," in *Musical Performance, A Guide to Understanding*. ed. J. Rink, (Cambridge: Cambridge University Press), 231–224.

Mudambi, S. M., and Schuff, D. (2010). What makes a helpful online review? A study of customer reviews on Amazon.com. *MIS Q.* 34, 185–200. doi: 10.2307/20721420

Müllensiefen, D., Gingras, B., Musil, J., and Stewart, L. (2014). The musicality of non-musicians: an index for assessing musical sophistication in the general population. *PLoS One* 9:e101091. doi: 10.1371/journal.pone.0089642

Muller, J. (2005). Music criticism and Adorno. *Int. Rev. Aesthet. Sociol. Music.* 36, 101–116.

Nelson, P. (1970). Information and consumer behavior. *J. Polit. Econ.* 78, 311–329. doi: 10.1086/259630

Newman, E. (1925). in *A musical critic's holyday*. ed. A. K. Alfred

Nikolsky, A. (2012). Listeners' canon in Western music: the secret of conservatism of public taste and its ramifications for the music industry. Academia.Edu (last accessed: 14.02.2022).

Örnebring, H. (2009). *The two professionalisms of journalism: Journalism and the changing contest of work*. University of Oxford, Reuters Institute for the Study of Journalism.

Patmore, D. N. C., and Clarke, E. F. (2007). Making and hearing virtual worlds: John Culshaw and the art of record production. *Music. Sci.* 11, 269–293. doi: 10.1177/102986490701100206

Philip, R. (2004). *Performing music in the age of recording* Yale University Press.

Plantinga, L. B. (1967). *Schumann as critic* Yale University Press.

Pollard, A. (1998). *Gramophone: The first 75 years* Gramophone Publications Limited.

Prictor, M. (1998). To catch the world: Percy Scholes and the English musical appreciation movement 1918–1939. *J. Mus. Res.*, 61–71.

Reus, G., and Müller-Lindenberg, R. (2017). *Die Notengeber: Gespräche mit Journalisten über die Zukunft der Musikkritik* Springer doi: 10.1007/978-3-658-15935-1.

Reus, G., and Naab, T. (2014). Verhalten optimistisch. Wie Musikjournalistinnen und Musikjournalisten ihre Arbeit, ihr Publikum und ihre Zukunft sehen – eine Bestandsaufnahme. *Publizistik* 59, 107–133. doi: 10.1007/s11616-014-0199-z

Ruprecht, L. (2013). The imaginary life of nineteenth-century virtuosity. *Deutsche Vierteljahrsschrift für Literaturwissenschaft und Geistesgeschichte* 87, 323–355. doi: 10.1007/BF03375695

Samson, J. (2003). *Virtuosity and the musical work: The transcendental studies of Liszt* Cambridge University Press doi: 10.1017/CBO9780511481963.

Scheibe, J.A. (1737/1740). Der critische Musicus. *Beneke.*

Schenk-Güllich, S. (1972). *Anfänge der Musikkritik in frühen Periodica*. (Doctoral dissertation). Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany.

Schick, R. D. (1996). *Classical music criticism* Garland.

Schmitz-Emans, M. (2015). "Musikkritik und literarische Schreibwerkstatt bei Robert Schumann," in *Zwischen Gattungsdisziplin und Geamtkunstwerk: Literarische Intermedialität 1815–1848*. eds. T. Keppler and W. G. Schmidt (De Gruyter), 239–262. doi: 10.1515/9783110404128.239

Scholes, P. A. (1928). Musical appreciation as common sense. *Mus. Sup. J.* 14, 9–45. doi: 10.2307/3382781

Schumann, R. (1854/1985). "Gesammelte Schriften über Musik und Musiker," in *Wiesbaden: Breitkopf & Härtel (reprint of the edition of 1854)*

Schwartz, B. (2008). "Can there ever be too many flowers blooming?" in *Engaging art: The next great transformation of America's cultural life*. eds. S. J. Tepper and W. Ivey (Routledge), 239–256.

Shrum, W. (1991). Critics and publics: cultural mediation in highbrow and popular performance. *Am. J. Sociol.* 97, 347–375. doi: 10.1086/229782

Siefert, M. (1984). The dynamics of evaluation: a case study of performance reviews. *Poet. Tod.* 5, 111–127. doi: 10.2307/1772429

Smith Maguire, J., and Matthews, J. (2012). Are we all cultural intermediaries now? An introduction to cultural intermediaries in context. *Eur. J. Cult. Stud.* 15, 551–562. doi: 10.1177/1367549412445762

Stefaniak, A. (2016). *Schumann's virtuosity: Criticism, composition, and performance in nineteenth-century Germany*. Bloomington, IN: Indiana University Press, doi: 10.2307/j.ctt2005sfp.

Strandberg, K. (2014). Art or artifice?: violin virtuosity and aesthetics in Parisian criticism, 1831–1848. PhD thesis Indiana University.

Stuckenschmidt, H. H. (1965). "Prognosen und Irrtümer der Musikkritik," in *Beiträge zur Geschichte der Musikkritik*. ed. H. Becker (Gustav Bosse Verlag), 11–17.

Thorau, C., and Ziemer, H. (Eds.) (2019). *The Oxford handbook of music listening in the. 19th and 20th centuries* Edn Oxford University Press.

Varriale, S. (2012). "Music, journalism, and the study of cultural change," in *East Asia and globalization in comparison, conference proceedings* (Seoul, ChungAng University), 97–107.

Verboord, M., and Janssens, S. (2015). Arts journalism and its packaging in France, Germany, the Netherlands, and the United States 1955-2005. *J. Pract.* 9, 829–852. doi: 10.1080/17512786.2015.1051369

Wahl-Jorgensen, K. (2012). The strategic ritual of emotionality: a case study of Pulitzer prizewinning articles. *Journalism* 14, 129–145. doi: 10.1177/1464884912448918

Widholm, A., Riegert, K., and Roosvall, A. (2021). Abundance or crisis? Transformations in the media ecology of Swedish cultural journalism over four decades. *Journalism* 22, 1413–1430. doi: 10.1177/1464884919866077

Williams, R. (1985). Keywords: a vocabulary of culture and society. Oxford University Press.

Williamson, V. J., and Jilka, S. R. (2013). Experiencing earworms: an interview study of involuntary musical imagery. *Psychol. Music* 42, 653–670. doi: 10.1177/0305735613483848

Williamson, V. J., Jilka, S. R., Fry, J., Finkel, S., Müllensiefen, D., and Stewart, L. (2012). How do "earworms" start? Classifying the everyday circumstances of involuntary musical imagery. *Psychol. Music* 40, 259–284. doi: 10.1177/0305735611418553

Witts, R. (2011). The music appreciation movement. BBC radio 3, the essay. https://www.bbc.co.uk/programmes/b013m49b

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 10 October 2022
DOI 10.3389/fpsyg.2022.934158

OPEN ACCESS

EDITED BY
George Waddell,
Royal College of Music, United Kingdom

REVIEWED BY
João Nunes Prudente,
University of Madeira,
Portugal
Pirkko Markula,
University of Alberta,
Canada

*CORRESPONDENCE
Nahoko Sato
nsato@ngu.ac.jp

SPECIALTY SECTION
This article was submitted to Performance Science, a section of the journal Frontiers in Psychology

RECEIVED 02 May 2022
ACCEPTED 13 September 2022
PUBLISHED 10 October 2022

CITATION
Sato N (2022) Improving reliability and validity in hip-hop dance assessment: Judging standards that elevate the sport and competition.
Front. Psychol. 13:934158.
doi: 10.3389/fpsyg.2022.934158

Improving reliability and validity in hip-hop dance assessment: Judging standards that elevate the sport and competition

Nahoko Sato*

Department of Physical Therapy, Faculty of Rehabilitation Science, Nagoya Gakuin University, Nagoya, Aichi, Japan

This study examined the reliability and validity of judging system scores of past hip-hop dance competitions in Japan. The analysis focused on the scores for each assessment category separately. Judges' scores were obtained from national dance competitions held annually in Japan between 2014 and 2019. In these competitions, five experienced judges evaluated the dancers' performances. The judges scored on a 10-point scale in five categories as follows: creativity, expression and interpretation, impression, technical quality, and synchronisation. This study found that the technical quality category demonstrated good reliability, whilst the impression showed poor reliability. Systematic bias was significant for all categories. There are no levels of difficulty defined for technique, no criteria set for correct movement and no explanation provided for each scoring level, which suggests that each judge may have interpreted the criteria for evaluating hip-hop dance differently. Developing these definitions and identifying the biases that affect evaluation would ensure a reliable evaluation system.

KEYWORDS

hip-hop dance, judging system, aesthetic sport, reliability, validity, competition, Japan

Introduction

Hip-hop dance is freestyle dance that began as street dancing, a part of the hip-hop culture (Craine and Mackrell, 2010), which includes breaking, rocking, popping, house and street jazz dances (Ojofeitimi et al., 2012). It has spread rapidly and many hip-hop dance competitions have been held worldwide. Originally, the impression of the audience was considered to be the most important factor in evaluating hip-hop dance; the winner of a competition was determined based on the audience's extent of excitement. However, in recent years, hip-hop dance has become more competitive. It was first considered an Olympic sport in the 2018 Youth Olympics and will make its debut in the 2024 Olympics (International Olympic Committee, 2021). In this context, clear evaluation criteria must be defined for hip-hop dance to be considered a viable competition so that dancers, judges

TABLE 1 Ten categories used in the most well-known hip-hop dance competitions.

Domain	Category
Performance	Creativity
	Staging, spacing, formations, and level changes
	Showmanship: intensity, confidence, projection and presence
	Style presence and attire
	Entertainment value/audience appeal
Skill	Musicality
	Synchronisation/timing
	Execution/controlled mobility and stabilisation
	Difficulty of execution of authentic dance style
	Variety of dance styles

TABLE 2 Six categories used to evaluate hop-hop dance performance at the 2018 Youth Olympic Games.

Domain	Category
Physical Quality: represents the qualities related to the body	Technique
	Variety
Interpretative Quality: represents the qualities related to the Soul	Performativity
	Musicality
Artistic Quality: represents the qualities related to the Mind	Creativity
	Personality

and audiences share a common understanding of the definition of superior hip-hop dance performance.

In Olympic artistic gymnastics, evaluations are divided into artistic and technical categories. Scores are determined by absolute evaluations that are based on the difficulty and kinematic criteria for all techniques, as defined in the Code of Points (Fédération Internationale de Gymnastique, 2021). Many studies have examined the reliability of this evaluation system using the results of past competitions, and high reliability has been reported (Leskošek et al., 2010; Atiković et al., 2011; Bučar et al., 2012; Pajek et al., 2013, 2014). In figure skating, another Olympic sport, final scores are calculated based on scores for technical elements, programme components and any deductions (International Skating Union, 2021). The reliability of figure-skating judges has also been investigated. Inter-judge correlation has been found to be above 0.9 for both technical and artistic scores (Lockwood et al., 2005). Thus, both artistic gymnastics and figure-skating competitions employ highly reliable evaluation systems.

In Dancesport, competitive ballroom dancing, a new evaluation system based on absolute evaluation, was introduced in 2013; this replaced the previous evaluation system that was found to be relative (World DanceSport Federation, 2021b). In the new evaluation system, as in artistic gymnastics and figure skating, evaluation is divided into artistic and technical aspects. The scoring system is based on a 10-point scale, with a performance description defined for each level. Research on the reliability of the new evaluation system reported that the mean correlation amongst all judges was 0.48 (Premelč et al., 2019), which was lower than correlation scores for artistic gymnastics and figure-skating competitions. Insufficient description of performance at each level was determined to be a reason for poor reliability.

At the biggest hip-hop dance competitions worldwide, multi-member groups compete, and their performances are evaluated across 10 categories in two domains (Table 1; Hip Hop International, 2021). The combined scores of the 10 categories are used to rank the competitors. Although descriptions of each category's evaluations have been publicised, detailed kinematic criteria for techniques and the criteria for assessing each level along a scale have not been described. Thus, judges are likely to score dancers based on their own interpretations and criteria. At the 2018 Youth Olympic Games, all break-dancing (a form of hip-hop dance) matches were set up in a battle format, either individual or group, and the winner was determined by a relative evaluation based on which dancer was better in each of the six categories in three domains (Table 2; World DanceSport Federation, 2021a).

As in figure skating and artistic gymnastics, in hip-hop dance competitions, including break-dancing, performances have been evaluated in categories that include both technical and artistic aspects (Tables 1, 2). For the technical aspect of the assessment, difficulty levels for techniques have not been established, and the correct movements for each technique have not been defined; thus, it is not clear how judges evaluate the technical aspect of performance. Studies have reported that factors such as facial expression (Cunningham et al., 1990) and body shape (Tovée et al., 1999; Pawlowski et al., 2000), as well as movement, affect the judges' evaluation of dance performances. Sato and Hopper (2021) found that the reliability of the judges' scores varied when the actual dancer videos and humanoid animations created from actual dancer movements were evaluated, suggesting that dancer appearance impacted the evaluation of judges. Although several categories exist within the evaluation of the artistic aspect of hip-hop dancing in the current system (Table 1), evaluation categories that consider biases such as those (un) favouring facial expression or body shape have not been developed. To date, the reliability of the evaluation systems currently in use has not been reported based on the results of past competitions in hip-hop dance. To develop an objective evaluation system, the reliability of current evaluation systems must first be examined.

This study analysed judges' scoring of hip-hop dance competitions held in the past, ascertaining each judging category separately and examining the reliability of the scores.

Materials and methods

Judges' scores were obtained from national dance competitions held annually in Japan throughout the years 2014–2019. However, the performances in these competitions were not videotaped. These competitions were open to dancers of elementary to junior high school age, and the results for each year, of the competition

final, performed by the dance teams that won the preliminary rounds, were used for analysis. The dance team consisted of at least 5–40 dancers. Dance genres covered in this competition were hip-hop, which includes rocking, popping, breaking, house and street jazz. This study was approved by the Nagoya Gakuin University Research Ethics Committee. All data used in the analysis were anonymised, and participants were offered opt-out opportunities.

Five experienced judges evaluated the dancers' performances in each competition. They were not the same individuals each year. The judges scored on a 10-point scale in five categories, as follows: creativity, expression and interpretation, impression, technical quality and synchronisation. There were no descriptions of performance for each point level (0–10), and the judges were not allowed to share or discuss their evaluations with each other. The final scores for each of the five categories for individual dance teams were calculated as the mean of the five judges' scores.

Descriptive statistics of all judges' scores for each category were calculated for each year of the competition. The following statistics values were calculated for validity analysis (Bučar et al., 2012). Signed and absolute deviations from the final score for individual judges were calculated as measures of bias. Mean rank and deviation from the expected rank were also assessed for individual judges. The expected rank was calculated as $(m+1)/2$, where m is the number of judges, with reference to Bučar et al. (2012). The reliability of the evaluation was examined and assessed using intra-class correlation coefficients (ICC) for single and mean of five raters for both two-way random (consistency) and fixed (agreement) effects (Premelč et al., 2019). Kendall's W (Kendall's coefficient of concordance) was also calculated. A Kendall's value of $W < 0.40$ was considered poor, 0.40–0.50 moderate, 0.50–0.70 good and greater than 0.70 excellent. ICC values were interpreted as follows: less than 0.40 poor reliability; 0.4–0.75 good reliability; greater than 0.75 excellent reliability (Fleiss et al., 2013). All data were analysed using SPSS Statistics software (version 25.0; SPSS Inc., Chicago, IL, United States).

Results

Amongst the five categories, the highest mean score was 7.35 ± 1.03, for impression, and the lowest was 7.10 ± 1.13, for technical quality (Table 3). Appendix 1 presents the statistics of scores for individual judges, and Table 4 shows values extracted from them, indicating the best and worst deviations in judging. In terms of score bias, the maximum absolute deviation from the final score and mean rank deviation from the expected rank were generally significant for all categories. Regarding the correlation between the scores of the individual judges and the final score, which is the mean of the five judges, technical quality demonstrated the largest maximum correlation coefficient and impression demonstrated the smallest minimum correlation coefficient in most of the competition years.

In terms of score reliability, the Kendall's W values ranged from 0.319 to 0.681 (Table 5). In each year of the competition, the category with the highest reliability was technical quality, with most values indicating good reliability, with scores ranging from 0.576 to 0.681. The category with the lowest reliability was impression, with most values indicating poor reliability, with scores ranging from 0.319 to 0.448. Similar ICC results were obtained; the single-measure ICC coefficients for absolute agreement and consistency for technical quality demonstrated fair to good reliability. The average-measure ICC coefficients for absolute agreement and consistency for almost all categories showed good to excellent reliability.

Discussion

To develop hip-hop dance competition and elevate its competitive status, an evaluation system with high reliability must be developed. This study was the first to examine the reliability of evaluation results of hip-hop dance competitions.

Regarding the reliability, the Kendall's W values ranged from 0.319 to 0.681, which was comparable to the reliability assessments for Dancesport (Premelč et al., 2019), thus indicating that the reliability was not high. In contrast, high reliability has been reported for judging in artistic gymnastics competitions (Leskošek et al., 2010; Atiković et al., 2011; Bučar et al., 2012; Pajek et al., 2013). In artistic gymnastics, the level of difficulty and correct movements for all techniques are defined, and point deductions are described in detail in the Code of Points. However, in hip-hop dance, there are no defined criteria for the difficulty of a technique or a correct movement, and there are no descriptions of each of the 10-point level. This means that each judge may interpret the criteria for evaluation and evaluate the performance differently in hip-hop dance. Various biases also reportedly affect judges' evaluations, including the position of the judges (Dallas et al., 2011), experience of the judges (Flessas et al., 2015), order of the performances (Plessner, 1999) and reputation of the dancers (Findlay and Ste-Marie, 2004). Factors such as the dancers' facial expression and appearance also affect performance evaluations (Cunningham et al., 1990; Tovée et al., 1999; Pawlowski et al., 2000; Sato and Hopper, 2021). These biases may have impacted the low reliability found in this study.

In hip-hop dance, dancers typically perform in groups. Similarly, rhythmic gymnastics involves a group competition, in which five competitors perform, and the judges must evaluate the performances of the five gymnasts simultaneously. The reliability of performance evaluations in artistic gymnastics and figure skating reported in previous studies were all for individual performance competitions, and no studies to date have investigated the reliability of performance evaluations in team competitions such as rhythmic gymnastics. When judges pay attention to one competitor, they lose information about execution to other competitors. Flessas et al. (2015) reported that when evaluating the five-gymnast ensemble routines in rhythmic

TABLE 3 Mean, minimum and maximum values for five categories and the final scores for the 2015–2019 competitions.

			Mean		*SD*			The lowest marks			The highest marks			
Year	***n***	**Category**	**Mean**	**Max**	**Min**	**Mean**	**Max**	**Min**	**Mean**	**Max**	**Min**	**Mean**	**Max**	**Min**
2019	48	Creativity	7.64	8.65	6.69	1.31	1.64	0.98	4.80	6.00	4.00	9.60	10.00	9.00
		Expression and interpretation	7.61	8.67	6.56	1.17	1.75	0.65	5.00	7.00	3.00	9.60	10.00	9.00
		Impression	7.74	8.92	7.02	1.23	1.57	1.00	4.80	6.00	4.00	9.60	10.00	9.00
		Technical quality	7.64	8.27	7.08	1.37	2.05	0.83	4.80	6.00	3.00	9.80	10.00	9.00
		Synchronisation	7.79	9.19	6.83	1.18	1.55	0.81	4.80	6.00	3.00	9.60	10.00	9.00
2018	49	Creativity	7.11	7.73	6.65	0.92	1.19	0.63	5.00	6.00	4.00	8.80	9.00	8.00
		Expression and interpretation	7.11	7.92	6.31	0.91	1.16	0.53	5.20	7.00	4.00	8.80	9.00	8.00
		Impression	7.43	7.94	6.90	0.91	1.26	0.54	5.60	7.00	5.00	9.20	10.00	9.00
		Technical quality	7.19	7.98	6.63	0.98	1.27	0.53	5.20	7.00	4.00	9.00	10.00	8.00
		Synchronisation	7.27	8.18	6.67	0.80	1.07	0.51	5.60	7.00	4.00	8.60	9.00	8.00
2017	53	Creativity	7.12	7.72	6.57	1.07	1.41	0.74	5.20	6.00	5.00	9.20	10.00	8.00
		Expression and interpretation	7.00	7.83	6.38	0.98	1.48	0.64	5.20	7.00	4.00	9.20	10.00	8.00
		Impression	7.35	8.02	6.83	1.03	1.48	0.71	5.80	7.00	5.00	9.20	10.00	8.00
		Technical quality	7.01	7.85	6.06	1.04	1.33	0.60	5.40	7.00	5.00	9.20	10.00	8.00
		Synchronisation	7.22	8.00	6.45	0.88	1.26	0.62	5.80	7.00	5.00	9.00	10.00	8.00
2016	52	Creativity	6.73	6.96	6.40	1.15	1.50	0.96	4.60	5.00	4.00	9.40	10.00	9.00
		Expression and interpretation	6.71	7.25	6.23	0.95	1.33	0.52	4.80	6.00	4.00	8.60	10.00	8.00
		Impression	6.92	7.17	6.83	1.01	1.48	0.83	5.00	6.00	4.00	9.00	10.00	8.00
		Technical quality	6.61	7.50	5.83	1.17	1.49	0.78	4.40	5.00	4.00	9.00	10.00	8.00
		Synchronisation	6.95	7.46	6.42	1.00	1.44	0.73	5.20	6.00	4.00	9.00	10.00	8.00
2015	45	Creativity	6.95	7.60	6.24	1.02	1.48	0.67	5.20	6.00	4.00	9.00	10.00	8.00
		Expression and interpretation	7.14	7.96	6.13	0.99	1.46	0.66	5.20	6.00	4.00	9.20	10.00	8.00
		Impression	7.30	8.18	6.20	0.97	1.63	0.62	5.40	6.00	4.00	9.40	10.00	9.00
		Technical quality	7.05	7.87	6.02	1.11	1.62	0.76	5.00	6.00	4.00	9.40	10.00	9.00
		Synchronisation	6.95	7.78	6.13	0.96	1.31	0.63	5.20	6.00	4.00	9.00	10.00	8.00
Mean 2015–2019	247	Creativity	7.11	7.73	6.51	1.09	1.44	0.80	4.96	5.80	4.20	9.20	9.80	8.40
		Expression and interpretation	7.12	7.92	6.32	1.00	1.44	0.60	5.08	6.60	3.80	9.08	9.80	8.20
		Impression	**7.35**	8.05	6.76	1.03	1.49	0.74	5.32	6.40	4.40	9.28	10.00	8.60
		Technical quality	**7.10**	7.89	6.32	1.13	1.55	0.70	4.96	6.20	4.00	9.28	10.00	8.40
		Synchronisation	7.24	8.12	6.50	0.96	1.32	0.66	5.32	6.40	4.00	9.04	9.80	8.20

gymnastics, international-level judges did not rely on eye fixation to detect errors and may have used other cognitive strategies, as compared to novice and national-level judges. Thus, evaluating performance in the case of group competitions can be considered more challenging, and this may also have affected the reliability results of hip-hop dance.

This study assessed systematic bias in judging to evaluate score validity. For all categories, the values of absolute deviations from the final score and mean rank and deviation from the expected rank were larger than those values for artistic gymnastics (Bučar et al., 2012), suggesting a more significant systematic bias. Fernandez-Villarino et al. (2013) reported that the special circumstances in which judges must evaluate dancers of different ages and skill levels in one competition could create problems, thereby making it difficult for judges to distinguish performances. The competitions analysed in this study were open to students from elementary to junior high school age; thus, a wide range of skill levels was likely observed and incorporated into performance evaluations. This wide range may be one of the reasons for the higher systematic bias that was found. Pajek et al. (2014) suggested that a possible reason for the low validity of artistic scores in artistic gymnastics was poorly defined criteria in the Code of Points. In this study, biases due to judges' perceptions

TABLE 4 The performance of individual judges.

			Absolute deviation		Judge mean rank deviation from the expected mean rank		Corrected category-5 judges mean correlation of individual judges	
Year	n	Category	min	max	min	max	min	max
2019	48	Creativity	0.68	1.35	−1.42	0.77	0.44	0.84
		Expression and interpretation	0.59	1.42	−1.52	0.96	0.57	0.83
		Impression	0.66	1.36	−1.67	0.42	0.42	0.75
		Technical quality	0.58	1.10	−1.25	0.35	0.76	0.90
		Synchronisation	0.57	1.66	−1.75	0.79	0.54	0.79
2018	49	Creativity	0.45	0.74	−1.61	0.18	0.53	0.76
		Expression and interpretation	0.57	0.95	−1.61	0.80	0.50	0.82
		Impression	0.49	0.80	−1.43	0.27	0.42	0.71
		Technical quality	0.56	0.89	−1.69	0.27	0.66	0.84
		Synchronisation	0.51	0.93	−1.88	0.43	0.25	0.85
2017	53	Creativity	0.56	0.82	−1.45	0.34	0.55	0.86
		Expression and interpretation	0.60	0.91	−1.58	0.45	0.46	0.85
		Impression	0.55	0.87	−1.42	0.17	0.41	0.83
		Technical quality	0.50	0.97	−1.49	0.96	0.55	0.88
		Synchronisation	0.64	0.96	−1.51	0.72	0.46	0.84
2016	52	Creativity	0.50	0.87	−0.98	−0.08	0.58	0.81
		Expression and interpretation	0.50	0.83	−1.25	0.17	0.42	0.79
		Impression	0.49	0.85	−1.08	−0.40	0.41	0.82
		Technical quality	0.64	0.94	−1.60	0.50	0.77	0.88
		Synchronisation	0.45	0.76	−1.31	0.23	0.56	0.85
2015	45	Creativity	0.52	1.09	−1.42	0.42	0.29	0.77
		Expression and interpretation	0.55	1.23	−1.56	0.96	0.50	0.85
		Impression	0.50	1.36	−1.62	0.84	0.17	0.82
		Technical quality	0.50	1.27	−1.60	0.98	0.68	0.84
		Synchronisation	0.50	1.08	−1.58	0.80	0.39	0.77

and preferences in evaluating the quality of performance and differences in interpretation of the judging criteria are assumed to contribute to score variability.

Amongst the evaluation categories used in this study, technical quality and synchronisation fall within the technical category, whilst creativity, expression/interpretation and impression fall within the artistic category. Technical quality, on the technical side, demonstrated the highest reliability, whilst impression, on the artistic side, showed the lowest reliability. Similar results were found in figure skating (Lockwood et al., 2005), artistic gymnastics (Pajek et al., 2014) and Dancesport (Premelč et al., 2019). These results implicate that the artistic side of evaluation may be more impacted by factors, including facial expression and body shape, as previous studies have demonstrated (Cunningham et al., 1990; Tovée et al., 1999; Pawlowski et al., 2000). Therefore, a new evaluation system that accounts for this effect would improve reliability on the artistic side of evaluation of hip-hop dance.

To implement a reliable evaluation system in hip-hop dance competitions, a detailed description of each level for each category must be provided as a first step. A clear evaluation system or tool will help judges interpret the criteria in the same way, thus reducing score variability due to differences in interpretation. Second, evaluation categories must also be reconsidered. In hip-hop dance, many factors other than movement are considered to affect performance evaluation. Evaluation categories should be based on the factors that affect performance evaluation. In artistic gymnastics and figure-skating competitions, rankings are determined by the final technical and artistic point scores. In hip-hop dance, the difficulty of the technique is important, but the artistic aspect is also important. The weight of the technical and artistic aspects in the evaluation, including the number of evaluation categories for each of these two aspects, must be considered. Third, biases that have been reported, including the order of performance, the position of the judges and the experience of the judges, should also be verified in hip-hop dance. Fourth, using a video system that is designed to record performances and observe them immediately afterwards would allow judges to observe dances multiple times; the use of such a system should be considered. Fifth, in hip hop dance competitions that are performed as a group competition, the evaluation criteria must be provided separately for individual dancers' performance and group performance. In rhythmic gymnastics, the evaluation criteria are separately defined for the evaluation of individual

TABLE 5 Reliability for the five categories and the final score for the 2015–2019 competitions.

Year	Category	ICC				Kendall's W coefficient	
		Absolute agreement		Consistency			
		Single-measure (95%CI)	Average-measure (95%CI)	Single-measure (95%CI)	Average-measure (95%CI)	*W*	*p*
2019	Creativity	0.255 (0.129–0.409)	0.631 (0.426–0.776)	0.321 (0.193–0.472)	0.702 (0.545–0.817)	0.448	0.000
	Expression and interpretation	0.280 (0.136–0.446)	0.661 (0.440–0.801)	0.382 (0.251–0.531)	0.756 (0.626–0.850)	0.540	0.000
	Impression	0.174 (0.069–0.314)	0.514 (0.271–0.696)	0.229 (0.111–0.379)	0.598 (0.385–0.753)	0.385	0.000
	Technical quality	0.507 (0.366–0.648)	0.837 (0.743–0.902)	0.552 (0.425–0.681)	0.861 (0.787–0.914)	0.659	0.000
	Synchronisation	0.234 (0.096–0.398)	0.604 (0.347–0.768)	0.351 (0.222–0.502)	0.730 (0.587–0.834)	0.503	0.000
2018	Creativity	0.297 (0.172–0.447)	0.678 (0.510–0.801)	0.297 (0.172–0.447)	0.678 (0.51–0.801)	0.432	0.000
	Expression and interpretation	0.223 (0.092–0.381)	0.589 (0.337–0.755)	0.329 (0.202–0.479)	0.710 (0.559–0.821)	0.452	0.000
	Impression	0.173 (0.070–0.309)	0.511 (0.274–0.691)	0.197 (0.085–0.343)	0.551 (0.317–0.723)	0.364	0.000
	Technical quality	0.343 (0.181–0.513)	0.723 (0.525–0.841)	0.454 (0.323–0.595)	0.806 (0.705–0.880)	0.576	0.000
	Synchronisation	0.158 (0.054–0.296)	0.484 (0.222–0.677)	0.238 (0.12–0.387)	0.610 (0.406–0.759)	0.372	0.000
2017	Creativity	0.341 (0.204–0.493)	0.721 (0.561–0.829)	0.410 (0.284–0.549)	0.776 (0.665–0.859)	0.543	0.000
	Expression and interpretation	0.224 (0.100–0.374)	0.591 (0.358–0.749)	0.312 (0.189–0.457)	0.694 (0.539–0.808)	0.449	0.000
	Impression	0.234 (0.114–0.380)	0.604 (0.391–0.754)	0.305 (0.184–0.449)	0.687 (0.531–0.803)	0.448	0.000
	Technical quality	0.340 (0.152–0.526)	0.720 (0.472–0.847)	0.509 (0.383–0.639)	0.838 (0.756–0.898)	0.592	0.000
	Synchronisation	0.195 (0.071–0.347)	0.548 (0.276–0.727)	0.315 (0.194–0.459)	0.697 (0.546–0.809)	0.449	0.000
2016	Creativity	0.385 (0.260–0.527)	0.758 (0.637–0.848)	0.395 (0.268–0.537)	0.765 (0.647–0.853)	0.500	0.000
	Expression and interpretation	0.306 (0.180–0.453)	0.688 (0.524–0.805)	0.353 (0.228–0.498)	0.732 (0.597–0.832)	0.479	0.000
	Impression	0.238 (0.124–0.382)	0.610 (0.415–0.755)	0.239 (0.124–0.383)	0.612 (0.415–0.756)	0.399	0.000
	Technical quality	0.431 (0.232–0.610)	0.791 (0.602–0.887)	0.578 (0.458–0.698)	0.873 (0.808–0.920)	0.681	0.000
	Synchronisation	0.350 (0.215–0.500)	0.729 (0.578–0.833)	0.409 (0.282–0.550)	0.776 (0.663–0.859)	0.522	0.000
2015	Creativity	0.129 (0.035–0.263)	0.426 (0.152–0.641)	0.156 (0.046–0.305)	0.481 (0.195–0.687)	0.319	0.000
	Expression and interpretation	0.198 (0.079–0.351)	0.553 (0.300–0.730)	0.280 (0.152–0.437)	0.660 (0.473–0.795)	0.416	0.000
	Impression	0.117 (0.028–0.244)	0.399 (0.126–0.618)	0.170 (0.058–0.321)	0.507 (0.235–0.702)	0.319	0.000
	Technical quality	0.339 (0.165–0.521)	0.719 (0.497–0.845)	0.472 (0.336–0.617)	0.817 (0.716–0.890)	0.589	0.000
	Synchronisation	0.164 (0.059–0.306)	0.495 (0.238–0.688)	0.224 (0.104–0.379)	0.591 (0.366–0.753)	0.401	0.000

gymnasts and collaborative performances (Fédération Internationale de Gymnastique, 2021).

In this study, the results of hip-hop dance competitions in which multiple dance groups' performances are ranked by performance scores (similar to gymnastics and figure skating) were analysed. However, break dancers will most likely compete in a one-on-one battle format at the Paris Olympics. Break dancing originated in hip-hop culture, and the winner is determined by the extent of the audience's excitement, which is influenced by their preferences and subjective impressions of the performance. However, as hip-hop dance (including breakdancing) has grown in popularity, objective evaluation systems have been developed to combat potential biases such as reputation and style preferences (Fogarty, 2018). Although the competition format for break dancing at the Paris Olympics is unknown, our findings can be used to develop a reliable standard of evaluation in a battle format. As mentioned earlier, multiple evaluation categories (divided into technical and artistic aspects) should be established, as well as a detailed description of each level for each category. Given that the characteristics of break dance are strongly linked to the creative expression of one's identity, emotions and artistic sensibilities, the weightage of technical and artistic aspects should also be considered in the final score (Fogarty, 2018). Competing to determine which dancer is better scored in these evaluation categories allows for a more reliable evaluation.

This study has a few limitations. First, the performances of dancers with a wide range of skill levels were used for evaluation, as the competitions from which the data were pulled and analysed were open to elementary and junior high school-aged participants. Study results may have been different using data from competitions with more skilled adult dance performances. Second, only the scores of judges from competitions in Japan were analysed. Further studies should be undertaken to investigate scores from competitions held in other countries and world competitions. Third, it is not clear how the judges who participated in the competitions analysed in this study varied in their ability to evaluate the performance accurately and consistently. Since judges' experience is an important factor influencing evaluation reliability (Flessas et al., 2015), this factor may have influenced this study's results.

This study was the first to investigate the reliability of the evaluation results of hip-hop dance competitions. The study's results will contribute to the development of a more reliable evaluation system for hip-hop dance competitions. To implement a reliable evaluation system, the reliability of the evaluation must be constantly investigated and feedback must be provided at the same time the system is developed. An evaluation system that can be explained objectively provides not only reliable evaluations but also guidelines for dancers and coaches to use as they work towards achieving high scores in competitions. A new evaluation system will ensure that hip-hop dance continues to develop as an Olympic sport.

Data availability statement

The datasets generated and/or analyzed during the current study are not publicly available due to contract with the organisation that provided the data but are available from the corresponding author on reasonable request.

Ethics statement

Studies involving human participants were reviewed and approved by the Nagoya Gakuin University Research Ethics Committee. Written informed consent from the (patients/participants or patients/participants legal guardian/next of kin) was not required to participate in this study in accordance with the national legislation and the institutional requirements.

Author contributions

NS contributed to the conception and design of the study, organized the database, performed the statistical analysis, and wrote the manuscript.

Funding

This work was supported by the Nagoya Gakuin University Grant (2021–2024).

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2022.934158/full#supplementary-material

References

Atiković, A., Kalinski, S. D., Bijelić, S., and Vukadinović, N. A. (2011). Analysis results judging world championships in men's artistic gymnastics in the London 2009 year. *Sport. Log.* 7, 95–102. doi: 10.5550/sgia.110702.en.095A

Bučar, M., Čuk, I., Pajek, J., Karacsony, I., and Leskošek, B. (2012). Reliability and validity of judging in women's artistic gymnastics at university games 2009. *Eur. J. Sport Sci.* 12, 207–215. doi: 10.1080/17461391.2010.551416

Craine, D., and Mackrell, J. (2010). *The Oxford Dictionary of Dance*. New York: Oxford University Press.

Cunningham, M. R., Barbee, A. P., and Pike, C. L. (1990). What do women want? facialmetric assessment of multiple motives in the perception of male facial physical attractiveness. *J. Pers. Soc. Psychol.* 59, 61–72. doi: 10.1037/0022-3514.59.1.61

Dallas, G., Mavidis, A., and Chairopoulou, C. (2011). Influence of angle of view on judges' evaluations of inverted cross in men's rings. *Percept. Mot. Skills* 112, 109–121. doi: 10.2466/05.22.24.27.PMS.112.1.109-121

Fédération Internationale de Gymnastique (2021). Rules. Available at: https://www.gymnastics.sport/site/rules (Accessed November 9, 2021).

Fernandez-Villarino, M. A., Bobo-Arce, M., and Sierra-Palmeiro, E. (2013). Practical skills of rhythmic gymnastics judges. *J. Hum. Kinet.* 39, 243–249. doi: 10.2478/hukin-2013-0087

Findlay, L. C., and Ste-Marie, D. M. (2004). A reputation bias in figure skating judging. *J. Sport Exerc. Psychol.* 26, 154–166. doi: 10.1123/jsep.26.1.154

Fleiss, J. L., Levin, B., and Paik, M. C. (2013). *Statistical Methods for Rates and Proportions* New Jersey: John Wiley & Sons.

Flessas, K., Mylonas, D., Panagiotaropoulou, G., Tsopani, D., Korda, A., Siettos, C., et al. (2015). Judging the judges' performance in rhythmic gymnastics. *Med. Sci. Sports Exerc.* 47, 640–648. doi: 10.1249/MSS.0000000000000425

Fogarty, M. (2018). "Why are breaking battles judged? The rise of international competitions" in *The Oxford Handbook of Dance and Competition*. ed. S. Dodds (New York: Oxford University Press), 409–428.

Hip Hop International (2021). Rules Regul. Available at: http://www.hiphopinternational.com/officialrules/ (Accessed November 9, 2021).

International Olympic Committee (2021). Breaking. Available at: https://olympics.com/en/sports/breaking/ (Accessed November 9, 2021).

International Skating Union (2021). ISU judging system. Available at: https://www.isu.org/figure-skating/rules/fsk-judging-system (Accessed November 9, 2021).

Leskošek, B., Čuk, I., Karácsony, I., Pajek, J., and Bučar, M. (2010). Reliability and validity of judging in men's artistic gymnastics at the 2009 university games. *Sci. Gymnast J.* 2, 25–34.

Lockwood, K. L., McCreary, D. R., and Liddell, E. (2005). Evaluation of success in competitive figure skating: an analysis of interjudge reliability. *Avante* 11, 1–9.

Ojofeitimi, S., Bronner, S., and Woo, H. (2012). Injury incidence in hip hop dance. *Scand. J. Med. Sci. Sports* 22, 347–355. doi: 10.1111/j.1600-0838.2010.01173.x

Pajek, M. B., Cuk, I., Pajek, J., Kovač, M., and Leskošek, B. (2013). Is the quality of judging in women artistic gymnastics equivalent at major competitions of different levels? *J. Hum. Kinet.* 37, 173–181. doi: 10.2478/hukin-2013-0038

Pajek, M. B., Čuk, I., Pajek, J., Kovač, M., and Leskošek, B. (2014). The judging of artistry components in female gymnastics: a cause for concern? *Sci. Gymnast J.* 6, 5–12.

Pawlowski, B., Dunbar, R. I., and Lipowicz, A. (2000). Tall men have more reproductive success. *Nature* 403:156. doi: 10.1038/35003107

Plessner, H. (1999). Expectation biases in gymnastics judging. *J. Sport Exerc. Psychol.* 21, 131–144. doi: 10.1123/jsep.21.2.131

Premelč, J., Vučković, G., James, N., and Leskošek, B. (2019). Reliability of judging in dance sport. *Front. Psychol.* 10:1001. doi: 10.3389/fpsyg.2019.01001

Sato, N., and Hopper, L. S. (2021). Judges' evaluation reliability changes between identifiable and anonymous performance of hip-hop dance movements. *PLoS One* 16:e0245861. doi: 10.1371/journal.pone.0245861

Tovée, M. J., Maisey, D. S., Emery, J. L., and Cornelissen, P. L. (1999). Visual cues to female physical attractiveness. *Proc. Biol. Sci.* 266, 211–218. doi: 10.1098/rspb.1999.0624

World DanceSport Federation (2021a). Buenos Aires 2018 youth Olympic games rules and regulations. Available at: https://www.worlddancesport.org/News/BreakingForGold/BAYOG_Rules_and_Regulations-2667 (Accessed November 9, 2021).

World DanceSport Federation (2021b). Judging systems. Available at: https://www.worlddancesport.org/Rule/Competition/General/Judging_Systems (Accessed November 9, 2021).

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 17 November 2022
DOI 10.3389/fpsyg.2022.954261

OPEN ACCESS

EDITED BY
Aaron Williamon,
Royal College of Music, United Kingdom

REVIEWED BY
Caroline Palmer,
McGill University,
Canada
Bruno Gingras,
University of Vienna,
Austria

*CORRESPONDENCE
Yuki Morijiri
morijiri@u-gakugei.ac.jp

SPECIALTY SECTION
This article was submitted to Performance Science, a section of the journal Frontiers in Psychology

RECEIVED 27 May 2022
ACCEPTED 20 October 2022
PUBLISHED 17 November 2022

CITATION
Morijiri Y and Welch GF (2022) Decisions on the quality of piano performance: Evaluation of self and others.
Front. Psychol. 13:954261.
doi: 10.3389/fpsyg.2022.954261

Decisions on the quality of piano performance: Evaluation of self and others

Yuki Morijiri[1]* and Graham F. Welch[2]

[1]Graduate School of Teacher Education, Tokyo Gakugei University, Tokyo, Japan, [2]UCL Institute of Education, University College London, London, United Kingdom

In common with other professional musicians, self-evaluation of practise and performance is an integral part of a pianist's professional life. They will also have opportunities to listen to and evaluate the performances of others based on their own criteria. These self-constructed perspectives towards to a piano performance will have an influence on both self-evaluation and external evaluation, but whether differently or similarly is not known. Consequently, this research study aimed to explore how judgements on the perceived quality of a performance are undertaken by professional standard pianists and what criteria are applied, both with regards their own performances as well as the performance of others. Participants were six professional pianists (3 men, 3 women) who were based in the United Kingdom (Mean age=31.5years old. SD=5.1). They were asked to play individually six trials of a piece of R. Schumann's "Träumerei" Op. 15 No. 7 in a hired hall for recordings. Then, within 2months, each participant was asked to come to a self-evaluation session to listen to and evaluate their own six recordings, using a Triadic method as a Repertory Grid. For the external evaluation focused session, the participants were asked to return again to evaluate a further six recordings made up of 'best' recordings as selected by each participant from their own individual self-evaluations. Analyses of the resultant data suggest that there was no significant difference between the participants in their overall ratings in the external phase, but that self-evaluation showed significant individual differences amongst several participants. The performance criteria in both self-evaluation and external evaluation predominately overlapped with each other in terms of musical factors, such as tone quality, phrasing, and pedalling. The ranking of the performances was highly correlated with perceptions of overall flow, tone quality and pedalling. It appears that pianists apply similar criteria to decide performance quality when evaluating their own performances as well as others.

KEYWORDS

performance criteria, pianists, self-evaluation, external-evaluation, piano performance

Introduction

Musical performances can vary greatly between individuals, and performers are reported to create an individual mental construction of their own performances (e.g., Sloboda, 2000; Repp and Knoblich, 2004). Performers have personal rules about performances and their own uniqueness of interpretation. Nevertheless, although different artists might express a piece of music differently from one another, the characteristics and distinguishing features are reported to be relatively usually maintained and stable in personal performances at the level of the individual (Jung, 2003; Lehmann et al., 2007). Performances by the same performer also tend to have an identity and oneness in terms of memory, action and performance parameters (Chaffin and Imereh, 2002; Gingras et al., 2011, 2013; Van Vugt et al., 2013). From the perspectives of listeners, performance characteristics are identifiable (Palmer et al., 2001), even by non-musicians (Koren and Gingras, 2014) and performer themselves (Keller et al., 2007). Therefore, it could be argued that individual differences, personal tendencies and personal rules become established through practise and experiences on the basis of their own sense of music and standards of music performance within the conventions of an established performance culture.

Earlier, Seashore (1938) reported that, in spite of each performer's expression being different, the expressive parameters and variations are likely to be reproduced and maintained within the same performer. Moreover, Repp and Knoblich (2004) demonstrated that pianists were able to recognise their own recording amongst a set of several performances. In their research, they recorded 12 pianists playing 12 musical excerpts. Several months later, the researchers played these performances back and asked the pianists whether they thought that they were the person playing each excerpt. Participants gave their own performances significantly higher ratings than any other pianist's performances. Furthermore, although they were presented with edited performances, which were different in tempo, overall dynamic level, and with dynamic nuances removed, the pianists' accuracy ratings did not change significantly. This suggests that the remaining information was sufficient for self-recognition (Repp and Knoblich, 2004). The researchers concluded that "pianists seem to recognise their own performances because those performances create a stronger resonance in their action system than other performances do; this stronger resonance implies that there is a closer match between anticipated and perceived action effects" (Repp and Knoblich, 2004, p. 607). The results of this study suggested that although both playing and recognition of performance varied greatly between individuals, it seemed that there were individual rules on performance and inherent cognitive constructions in the performance evaluation. Furthermore, the implication of individuality in music performance is that performers have their own criteria in terms of performance and judge themselves and others on the basis of these criteria.

In terms of the nature of musical criteria used to evaluate a piano performance, these have been suggested in various ways (e.g., Abeles, 1973; Jones, 1986; Nichols, 1991; Palmer, 1996; Bergee and Cecconi-Roberts, 2002; Stanley et al., 2002; Zdzinski and Barnes, 2002; Juslin, 2003; Wapnick et al., 2005; Russell, 2010; Alessandri et al., 2016). For example, Duerksen (1972) suggested eight criteria for piano performance evaluation: rhythmic accuracy, pitch accuracy, tempo, accent, dynamics, tone quality, interpretation and overall quality. Russell (2010) suggested two broad categories, technical and musical, and five subordinate criteria for each: technical (tone, intonation, rhythmic accuracy, articulation and technique) and musical (tempo, dynamics, timbre, interpretation and musical expression). In other research, Thompson et al. (1998) explored how adjudicators rated the piano performance of different individuals by using Personal Construct Theory (Kelly, 1955), and in which these criteria were included: right-hand expression, phrasing, dynamics, rubato, form/structure, tone balance, pedalling, attention to rhythm and meter, articulation, technical competence, tempo, expression of several parts. In the same manner, other research studies have adopted other criteria (Duerksen, 1972; Saunders and Holahan, 1997). Several elements of classification used by one researcher appear to be redundant for another. In addition to the diversity of performance criteria, recent studies also underlined that there are difficulties and uncertainness in determining criteria of performance assessment (P. Johnson, 1997; McPherson and Thompson, 1998; McPherson and Schubert, 2004). Nevertheless, there has not been a consensus on performance criteria agreed by the research community, as each research study has used a different classification in order to evaluate performance. Moreover, the diversity of criteria and factors for performance evaluation have been discussed from the perspective of the comments by judges (Wrigley, 2005). It would be worthwhilst to explore the criteria that pianists themselves apply towards to their own performances to see if their viewpoints are different.

In the field of music performance, it has been argued that self-evaluation is one of the important processes in the development of performance skill from the perspective of self-regulation (Zimmerman, 2000; Zimmerman and Schunk, 2001; McPherson and Zimmerman, 2002). However, self-evaluations of music performances can often be inconsistent and biassed (Bergee, 1993; Kostka, 1997). In the process of self-evaluation, performers make judgements about whether their playing is good or needing improvement on their own terms and consider which elements that they might change and how (Brändström, 1996; Daniel, 2001). Even whilst playing, it is reported that performers will listen to and know their sense of the music, such as in terms of the stresses and phrasing, and be able to feed this knowledge back into the ongoing development of their own performances (Schmidt and Lee, 2010).

Self-assessment is a process of a formative assessment or evaluation of oneself or one's action including performance, work, attitude and learning to an objective standard (Andrade and Du, 2007). Boud (1991) proposed the definition of self-assessment as "identifying standards and/or criteria to apply to their work and making judgements about the extent to which they have met these

criteria and standards" (p.5). Andrade and Du (2007) provide a helpful definition of self-assessment that focuses on the formative learning that it can promote "during which people reflect on and evaluate the quality of their work and judge the degree to which they reflect explicitly stated goals or criteria, identify strengths and weaknesses in their work, and revise accordingly" (p.160). According to Boud (1995), all assessment, including self-assessment, comprises two key stages. The first is to develop knowledge, make decisions about the standards and criteria and apply them to a given work. The second is to assess critically the quality of the performance in relation to these criteria to see if it satisfies these standards or not. An engagement with setting the standards and their criteria are considered to underlie the process of learning (Boud, 1995).

Self-evaluation itself is known as a process of self-reflection and a potential enhancement of learning (McPherson and Zimmerman, 2002). Self-assessment can help to develop the skills effectively to monitor own performances and learning. Through a process of self-evaluation, a learner should be in a position to become more knowledgeable about how learning could be undertaken, what was learnt, how it would be judged and how it progressed, and also be able to utilise the outcomes into making a plan of how to improve further learning. Self-evaluation is considered an important part of ways to improve and enhance self-regulated learning (Boud, 1995). In other words, self-regulated learners are perceived as being more capable of monitoring themselves and of understanding the feedback that they receive and also engage in self-evaluation (McPherson and Zimmerman, 2002).

Whilst positive aspects of self-evaluation have been reported, some research studies have questioned the reliability and validity of self-evaluation (Gordon, 1991; Ross, 2006). For example, it has been demonstrated that self-evaluation does not always agree with instructors' nor externals' evaluations (Bergee and Cecconi-Roberts, 2002). Particularly, students' evaluation and those by expert evaluators do not often match (Bergee, 1993; Kostka, 1997). Bergee (1997) demonstrated that the outcomes of students' self-evaluation were less consistent, compared with faculty or peer evaluation, whilst also noting that there was no significant difference between the students' level of self-evaluation performance and the type of instruments that they played. Several research studies have shown that there might be consistency, but also inconsistency with self-evaluation (Blatchford, 1997; Ross, 1998). For example, Kostka (1997) reported that self-assessment by piano students enrolled at the university compared to the assessment by their teachers showed relatively low agreement (students' self-ratings were lower). Also, this research highlighted that self-assessment was influenced by students' perceptions of what they "know," namely self-perceptions of knowledge.

One of the reasons why self-evaluation can be difficult in terms of its reliability would be related to a feature of music evaluation. For self-evaluation, music performers are aware of their own performances during playing. Enhanced auditory feedback during performances can affect improvements in their performances (Repp, 1999; Finny and Palmer, 2003; Mathias et al., 2017). On the other hand, however, this kind of feedback might be problematic because the performers cannot listen to their own performance in the same ways as their audience (Daniel, 2001). In addition, it can be difficult for performers to evaluate their performance appropriately during the act of playing.

Moreover, complexities of performance evaluation itself include who makes the assessment. Some research studies have demonstrated that more experienced evaluators are likely to be more reliable in evaluation, whilst it might also be difficult to delineate between 'more experienced' or 'more skilful' and 'less good'. The outcomes of research studies have been diverse. Whilst some research studies demonstrated that that more musically trained people are likely to have more reliable evaluation skills for music performances (Johnson, 1996; Shimosako and Ohgushi, 1996; Ekholm, 1997), several researches reported contrary findings that there is no substantial evidence to suggest that higher skilled musicians have more reliable assessment skills (Mills, 1987; Schleff, 1992; Bergee, 1993; Doerksen, 1999). However, it is agreed that the reliability and consistency on performance evaluation by trained musicians, such as professional musicians and faculty members, has been evidenced by research studies (Abeles, 1973; Thompson et al., 1998; Bergee, 2003; Ciorba and Smith, 2009). Wapnick et al. (1993) concluded that the relationship to the instruments which were the evaluator's major study and which were related to the performance being assessed was not necessarily influential in reliable evaluation. This research study also suggested that "performers who excel in any one area of performance may excel in other areas as well" (p.283). It has been evident that non-musicians, who have not received formal musical training in higher education, or have very little prior experience in music, have a different perception and way of evaluating music performance (Geringer and Madsen, 1995/1996; Johnson, 1996). Therefore, it could be argued that higher musical and performance skills support the quality and reliability of musical assessment.

It has been suggested also that how people listen to and perceive music (e.g., Lerdahl and Jeckendoff, 1983; Madsen, 1990; Madsen et al., 1997) and how people evaluate performance as audiences (Johnson, 1996; Thompson et al., 1998; Hentschke and Del Ben, 1999) may be different from self-assessment in performance (e.g., Bergee, 1997; Daniel, 2001; Bergee and Cecconi-Roberts, 2002; Hewitt, 2011). It could be said that performers' perspectives towards their own performances may be different from how others evaluate them. If performers have their own criteria and perspectives for self-evaluation, it would be worthwhile to investigate whether these same personal criteria are also used to judge the performances of others, or whether the self is a special case in terms of expectations. So far, the topic of whether each performer has two perspectives of performance evaluation, namely as a performer and as an audience when they listen to a performance, has been unexplored.

Therefore, this research study aimed to explore how the perceived quality of performance might be decided by professional standard pianists and what criteria were applied, both with regards

their own performances as well as the performance of other peers. In particular, if each performer has a framework of criteria for performance evaluation, it is worthwhile to explore how each musical framework element contributes to the decision concerning the quality of piano performance. Also, how their own constructs could affect both self-evaluation and an evaluation of the performances by others.

Methodology

Personal construct theory and a triadic method

In order to identify personal constructs as an interpretation of person's experience, a large range of research studies, including in clinical settings, education and the arts, have applied a Repertory Grid Technique which was suggested originally by Kelly (1955) (e.g., Beail, 1985; Saúl et al., 2012) in his personal construct theory (PCT). This was based on a concept that "a person's processes are psychologically channelized by the way in which he anticipates events" (Kelly, 1955, p. 46).

In the field of music education and music psychology, the application of PCT has been researched in relation to how people recognise and listen to music (e.g., Hargreaves and Colman, 1981; Thompson et al., 1998). Gilbert (1990) valued PCT as a way of eliciting people's insights as individuals and also of how the elicitation of constructs could have a prospective value for researching how people recognised and developed their own musical perceptions. Thompson et al. (1998) demonstrated constructs of piano performance criteria by six adjudicators with six recordings of a Chopin's Etude. The researchers suggested that, by using a repertory grid technique, adjudicators could develop and refine their skills in evaluation by recognising insights and comparing their personal constructs. Consequently, the method of PCT is seen as not only a useful way to elicit personal constructs from an individual, but can also be beneficial in the development of the person who is involved in the process of using the technique in terms of understanding their own inner vision of a certain world.

Probably the most widely used PCT tool is the Repertory Grid Technique which is a method of eliciting constructs by asking participants to compare three elements and then stating how two are similar and different from the third. For example, in a musical context, two performances, A and B would be allocated a "slow" label and the other performance C could be "fast" on a construct called "tempo." This procedure is called the "Triadic Method." Answers are recorded in a matrix, which can then be analysed to produce a construct map. Regarding such a triadic method, Kelly (1955) originally suggested six ways that this could be applied: 1. The minimum context card form; 2. The full context form; 3. The sequential form; 4. The self-identification form; 5. The personal role form; and 6. Full context form with the personal role featured. The minimum context card form is the most widely used. This form provides three elements that are selected by the participants. The participants need to specify the important features in which two of the elements are similar and subsequently different from the third (Bannister and Mair, 1968; Fransella and Bannister, 1977). The pair of features given by the participant becomes a set of two construct poles, which is used in the next stage, namely completing a grid. In the current research study, the minimum context card form, which appeared to be the most common approach in the literature studies, was used as a basis for eliciting the personal constructs of participants.

Repertory Grid Technique is flexible as a methodology. However, it generally has five procedural stages: 1. Eliciting elements; 2. Eliciting constructs; 3. Completing the grid; 4. Analysis; and 5. Interpretation (Beail, 1985). "Elements can simply be provided by the investigator" or the researcher (Beail, 1985, p. 3). These can be places, people, and also can be generated by descriptions of a situation, unspecified acquaintances or giving roles (c.f. Fransella and Bannister, 1977; Beail, 1985). However, elements should "be representative of the area to be investigated" and "be within a particular range as constructs apply to only a limited number of people, events or things" (Beail, 1985, p. 4). In the current research, elements were recordings of piano performances by participants (c.f., Thompson et al., 1998).

Regarding the choice of constructs, there is another concern of whether or not these should be provided (Fransella and Bannister, 1977). From a practical perspective, to provide constructs can be vital, for example, when it is the purpose of the study to compare the relationship between verbal labels. From another point of view, supplied constructs can be given "a personal meaning by being related to those elicited" from the participant (Fransella and Bannister, 1977, p. 19). It cannot be always said that elicited constructs are more meaningful than provided constructs (Fransella and Bannister, 1977). A researcher needs to acquire a clear idea by understanding the participants' recognition (Yorke, 1978). Kelly (1955) warned that verbal labels provided by the participants might not always reflect their innermost thoughts. Therefore, the researcher should know that the participants will attach their own meaning to the researcher's label if a provided verbal label is used (Beail, 1985). Beail (1985) also added, "what is important is that a supplied verbal label be meaningful to the subject" (p.6). In this research, 13 musical criteria (overall flow, tone quality, interpretation of music, tempo, dynamics, rhythm, melodic accuracy, style, rubato, pedalling, technique, musical expression, phrasing) were listed as the suggested perspectives. These were elicited from previous music-focused research studies (e.g., Abeles, 1973; Jones, 1986; Nichols, 1991; Bergee and Cecconi-Roberts, 2002; Stanley et al., 2002; Zdzinski and Barnes, 2002; Juslin, 2003; Wapnick et al., 2005; Russell, 2010). However, the participants were also allowed to add their own ideas of criteria if they so wished. Providing some suggested construct options enabled the research to have a clear performance-focused context with appropriate musical criteria and acted as a framework for participants to understand their personal viewpoint in which they could also have the option of adding their own ideas of constructs.

In adopting this particular triadic method approach from repertory grid technique in the current research study, participants' recordings of a selected piano piece were used as elements. Participants were tasked with choosing three recordings (as elements) and were asked to identify two similar features and a different feature. The participants then had to explain their choices – as an application of the triadic method – using a minimum context card form. Each participant was asked to name their construct, for example, with comments about "fast" versus "slow" on a construct called "tempo." Figure 1 shows the sample grid with two construct poles for the six recordings. The participant could choose a construct from the provided list of 13 musical criteria, which were elicited from the previous research study, or to add another.

Analyses of the research literature suggests that, as well as there being different ways in eliciting constructs, there are also various ways of completing a grid. The current research study adopted a rating grid because "this method allows the person greater flexibility of response than does the rank grid" (Fransella and Bannister, 1977, p. 40). Based on the two construct poles given by a participant, each element was rated by how close it was to the description of the construct pole, for example, by using a seven point or nine-point scale. In this research study, a nine-point scale was adopted as it can be more precise as a measurable rating. In psychological research, having more scale points is thought to be better. However, there is a diminishing return after around 11 points (Nunnally, 1978).

Participants

Participants were six professional pianists (3 men, 3 women, identified as Performers A–F) who were based in the United Kingdom (Mean age = 31.5 years, SD = 5.1 years). The mean duration of playing the piano was 27.7 years (SD = 4.9 years). In order to guarantee the professional expertise of the participants, the following conditions were required: to be a professional, active musician and concert pianist.

Piece of music

In this research study, all participants were asked to play a common piece of music chosen by the researcher and to evaluate recordings of it for a research session. Robert Schumann's "Träumerei" Op. 15, No. 7 was selected as the piece of music used. This piece is the one of the most famous and lyrical of Schumann's piano pieces (Ostwald, 1985; Magrath, 1993; Gordon, 1996; Kapilow, 2011). This piece has been employed in several music psychological research studies regarding acoustic analysis and performance research (e.g., Repp, 1992; Friberg, 1995; Repp, 1995; Repp, 1996; Beran and Mazzola, 2000; Cambouropoulos and Widmer, 2000; Almansa and Delicado, 2009). It was also expected to encourage various individual differences in performances, as this feature has been reported in several previous research studies (e.g., Repp, 1992, 1995, 1996; Mazzola, 2011). From the perspective of its relative technical difficulty, this piece was selected for the syllabus of the 2007–2008 ABRSM Graded 7 Piano Exam (ABRSM, 2006). This suggests that Träumerei is not technically and musically easy; however, it not too difficult for professional pianists.

Procedure

The research session had three phases: (i) Recording, (ii) Self-evaluation and (iii) External evaluation. The participants were asked to take part in all three sessions and to make sure that they

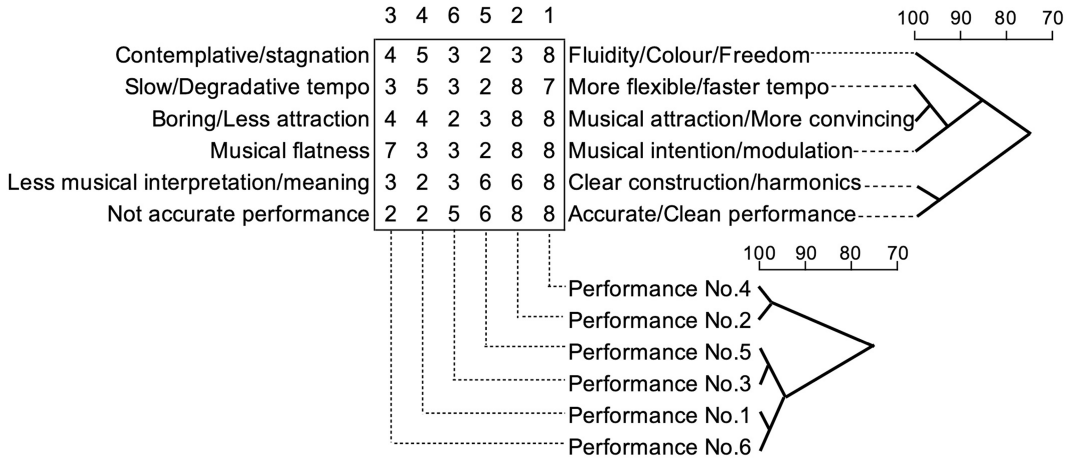


FIGURE 1
Sample grid from performer A.

understood all requirements and the schedule. They were informed of the research procedure by reading a specially designed research leaflet which also explained about confidentiality and related ethical concerns, as approved under the university's ethical procedures for informed consent. The details of the three sessions were as follows.

The first session: Recording

The six participants were asked to practise a piece of Schumann's "Träumerei," Op. 15 No. 7 with repeats (c.f., Repp, 1992, 1994, 1995, 1996) before coming to the recording session. The participants were given a copy of the music score in advance, which was published originally by Breitkopf and Härtel (1839).

For the recording session, participants were asked to come to a hall at the University in London. The participants were asked to play six trials each of Schumann's Träumerei with a tuned YAMAHA grand piano G5. Under the agreed ethical procedures, at the start of the session, each individual participant again received an explanation of the aims of the research and procedure and were asked to sign a research consent form. Before recording, the participants had 10 min to practise and to become familiar with the piano and to check the recording conditions for the sound recordist. With permission, six performances were recorded using an audio recording system (not video) with a professional recording engineer, and all recordings were made under the same conditions. Also, the participants were provided with a sheet to record their feelings about each trial (if they wished) and, at the end of the session, to report which performance that they thought was the 'best'. The whole recording session for each person lasted ~45 min.

The second session: Self-evaluation

Within 2 months after making their original set of six recordings, each participant came to the Music Technology Suite at the University and listened to, evaluated, and made a comparison between their own six recordings. To do this, the participants again used a Triadic method with a minimum context card form (Fransella and Bannister, 1977; Wapnick et al., 1993). They received an explanation of the procedure and the purpose of the session before undertaking the evaluation.

Firstly, each participant was allowed to adjust the comfortable listening volume with headphones (Sennheiser HD650) using a VLC media player (version 2.0.6). The order of playback was from their original first performance to their sixth performance, and which was technically labelled as performance No. 1 to No. 6 in the session. There are some research studies which have showed that piano performances are highly likely to be receive higher scores in the latter order of playing (e.g., Duerksen, 1972; Flôres and Ginsburgh, 1996). The order of performance in both recording and self-evaluation was kept the same so that it could be determined if performance order could affect their decision on the quality of performance differently in both recording time and self-evaluation. Whilst listening to the six recordings, they were asked to write some notes on a comments sheet to be utilised later.

Secondly, after listening to all six versions, each participant was asked to choose three recordings (as elements) randomly, to compare them, identify two similar features and one different feature and then to explain their choices. Each participant was asked to name their features (label), write these down and to indicate each of three performances for each label. For example, two performances, No. 2 and No. 6, from the three could be allocated a "slow" label and the other performance, No. 4, could be labelled "fast" on a construct called "tempo." The participant could choose a construct from the list of 13 musical criteria (overall flow, tone quality, interpretation of music, tempo, dynamics, rhythm, melodic accuracy, style, rubato, pedalling, technique, musical expression, phrasing), which were elicited from the previous research study, or add another of their choice. Each participant completed six sets of constructs. During this stage of evaluation, the participants were allowed to play back and listen to any recordings as many times as they wanted.

Thirdly, the participant rated each of their six recordings (the whole set) using a nine-point scale in terms of each of the six constructs that they had previously given. For example, if a participant named "slow" and "fast" in the construct for tempo, each of the six recordings was required to be rated as 1 = the fastest and 9 = the slowest. Because each participant arranged their six sets of constructs in this manner, all recordings were rated according to six sets of criteria. They were allowed to listen to the recordings again if they wanted. Finally, the participants ranked their six performances and were asked to choose the 'best' one (following Thompson et al., 1998). This whole session lasted ~1.5 h.

The third session: External evaluation

Within 2 months after this self-evaluation, the participants returned to the Music Technology Suite and evaluated a further six recordings. These were the six performers' choices of their 'best' performance, including their own 'best' recording – although this was not disclosed until the end of the third session. Each performers' personal information was not disclosed to participants.

Each participant was tasked with listening to the six recordings based on a randomised order (however, the adjudicator's own recording was always placed as the 3rd), as indicated in Table 1. Listening took place under the same condition as in the second session in terms of the headphones (Sennheiser HD650) using VLC media player (version 2.0.6). The participant's own recording was played as the third in the listening sequence for all participants. After listening to all recordings, the participants were asked to choose three recordings randomly, and to compare and apply the constructs as in the second session. They rated the six recordings based on the six sets of constructs. They also ranked the performances and chose the best performance amongst these six. These processes were the same as followed in the second session.

At the end of the third session, the participant was informed that one of the recordings was their own performance and then was asked to identify if they knew which one this was, and to give

a reason for their decision. Overall, this third and final session lasted ~1.5 h.

Results

Performance time

Table 2 shows the results of the length of each performance, which was measured from onset to offset of the performance sounds. The left column displays the trial number of each performance from the first to sixth performance. The total average length of performance was 2:35 min (range 2:03–3.33 min). The results of a repeated measures analysis of variance (ANOVA) showed a significant difference in terms of each performer's mean time duration in performing the selected piece, $F(5, 25) = 4.06$, $p = 0.008$, $\eta_p^2 = 0.45$. This result indicated that strong individual differences were evident in performers' playing (and, by implication, conception) of the music, even though all the performers played the same piece.

Performers E and F had some diversities regarding performance time for each of their trials. Performer E played with a range of 2:04–3:33 min and Performer F played with a range of 2:16–3:19 min. Both performers mentioned that they intentionally played each performance differently, based on what they wanted to express throughout each performance. In contrast, the other performers each tended to perform within a relatively more uniform time length.

Mauchly's test indicated that the assumptions of sphericity had been taken violated, $X(6) = 41.9$, $p < 0.001$, therefore degrees of freedom were corrected using Greenhouse–Geisser estimates of sphericity ($\varepsilon = 0.26$). The results show that there was a significant effect in different time length of performance, $F(1.31, 6.55) = 7.66$, $p = 0.039$. These results suggested that there are individual differences amongst performers in terms of performance time. A pair-wise comparison between participants was undertaken to investigate which performers had mean time differences of performances. Significant differences are evidenced between performers D and A, B, C each ($p < 0.01$). Performer D had an average performance time of 2:04, which was the lowest average (the fastest tempo performance) amongst the six performers. Other pairs do not show significant differences. Performer D's chosen recording was the shortest and Performer F's recording was the longest, with a time difference of more than 1 min.

TABLE 1 The order of playback at the third session.

Adjudicator	Order of playback					
	1st	2nd	3rd	4th	5th	6th
Performer A	B	E	A	C	F	D
Performer B	C	F	B	D	A	E
Performer C	D	A	C	E	B	F
Performer D	E	B	D	F	C	A
Performer E	F	C	E	A	D	B
Performer F	A	D	F	B	E	C

A = the performance by Performer A. Performer's own recording = shaded text.

Decision for the best performance as self-evaluation

At the recording-focused session (the first session), all participants were asked to think back over their examples and to decide which trial they thought to be the best. At the self-evaluation session (the second session) they listened to all of their own six recordings and ranked these from the best to least best recording. Table 3 shows performers' choices of the best performance at the original recording session and also the subsequent self-evaluation session. At the recording, all performers reported that their best performance was in the latter half of their playing sequence, especially the fifth and sixth versions. In the self-evaluation session with audio playback, their choice of best performance might be the same or different. Overall, the matching of the choice of best performance between the initial session of recordings and the self-evaluation session was 50%. Participants were likely to choose the best performance from the latter half at both sessions.

The results of the self-evaluation

At the self-evaluation (second) session, each participant listened to their own six recordings, evaluated these by using a

TABLE 2 The length of each performance.

Trials	Performer A	Performer B	Performer C	Performer D	Performer E	Performer F
1st	02:36	02:39	02:36	02:03	02:11	02:29
2nd	02:37	02:39	02:41	02:03	02:24	03:01
3rd	02:42	02:38	02:44	02:07	03:33	02:20
4th	02:44	02:42	02:44	02:02	02:04	***03:19***
5th	02:54	02:39	***02:40***	***02:09***	02:56	02:16
6th	***02:55***	***02:39***	02:38	02:03	***02:33***	03:12
Mean	02:44	02:39	02:40	02:04	02:36	02:46

Best perceived performance at the 1st recording session = italic and bold. Best ranked performance at the self-evaluation (2nd session) = shaded.

TABLE 3 Choice of the best performance.

	At the 1st session (Recording)	At the 2nd session (Self-evaluation)*	Matching
Performer A	No. 6	No. 4	No
Performer B	No. 6	No. 6	Yes
Performer C	No. 5	No. 5	Yes
Performer D	No. 5	No. 3	No
Performer E	No. 6	No. 6	Yes
Performer F	No. 4	No. 6	No

*Recordings of the best performances at the 2nd session were actually used in 3rd session (external evaluation).

Triadic method for creating six sets of constructs, and rated each performance with a nine-point scale based on their six sets of constructs. Each recording (element) and construct was subjected to hierarchal cluster analyses with Ward's method (1963).[1] The analyses were undertaken by using SPSS.

Figure 2A illustrates the result of the cluster analyses for both the constructs and the performances. This provides an illustration of the degree of association between constructs and between performances in a tree diagram. The top dendrograms shows the degree of association between constructs (poles). The 6 × 6 matrix of numbers presents the rating of each performance in each construct. Under the matrix, the dotted line indicates the number of the performance for each column. Above the matrix, the numbers show the ranking of performance, which displays as 1 (the highest rank) to 6 (the lowest rank).

As exampled in Figure 2A, each set of two performances which were of a closer ranking are strongly associated, No. 4 (rank 1) and No. 2 (rank 2) at 98%, No. 5 (rank 5) and No. 3 (rank 6) at 99% and then, No. 1 (rank 4) and No. 6 (rank 3) at 99%. The set of No. 5 and No. 3 is associated with the set of No. 1 and No. 6 at 94%. Performer A ranked Performance No. 4 as the best, which has the highest score in all constructs, apart from the sets of criteria 'slow/degradative tempo' and 'more flexible/faster tempo'.

In particular, this 'best' performance attracted the highest score on the construct of 'fluidity/colour/freedom'. Performance No. 3 was ranked as the lowest, which has the lowest rating in the construct pole of 'boring/less attraction' and 'musical attraction/more convincing'. However, Performance No. 5 (rank 5) has more low-rated constructs than Performance No. 3. The constructs of 'more flexible/faster tempo' and 'musical attraction/more convincing' are most strongly associated at 99%. These two constructs are also associated with 'musical intention/modulation' at 93%. Similarly, the constructs of 'less musical interpretation/meaning and clear construct/harmonics' and 'not accurate performance and accurate/clear performance' are strongly associated at 98%. These two constructs show a low degree of association with other constructs at 75%.

These clusters were created for each participant in the same way as Performer A's results. All illustrations are presented in Figures 2A–F. Figure 2B represents the result of a cluster analysis of the evaluation by Performer B. Performance No. 6 was chosen as the best as it achieved the highest total score. The top three performances (Performance No. 6, No. 4 and No. 3) were highly associated with each other at 99%. Performance No. 1 and No. 2 were also rated lower on the construct regarding tempo, which was associated with all other construct at 75%. The constructs of tone quality and phrasing were associated each other as the highest at 99%. These two constructs were also associated with the quality of pedalling at 98%. The degree of association with accuracy of note and musical expression was 97%.

Figure 2C displays the result of Performer C's self-evaluation. On the constructions of criteria, balanced tone and pedalling were the most associated at 99%, as well as the combination of sound colour and musical expression, which were also strongly associated with being musically convincing at 98%. The construction of tempo was associated with all other constructs at 75%. Performance No. 5 ranked as the best overall and No. 4 ranked as the second best. These were strongly associated at 99%, as was Performance No. 3 which was ranked as the lowest with No. 1 ranked as the second lowest and associated at 99%. The best performance was the only one that got the highest rating amongst six performances on the construct of 'Un/Well-balanced tone' and 'tempo'. Also, the top two performances, namely No. 5 and No. 4, achieved the highest rating on the criterion of 'musically convincing'.

Figure 2D indicates the self-evaluation by Performer D. The top ranked Performance No. 3 achieved the highest score on all constructs and was associated with the second ranked Performance No. 2 at 98%. The lowest ranked Performance No. 6 got the lowest rating on the construction of 'dream character', which was associated with 'clear melodic lines' at 99%. The constructs of 'melodic freedom' and 'clear pedalling' were strongly associated at 99%, which were also related to 'dynamic contrast' at 96%. Performance No. 6 was the most associated with the third ranked Performance No. 4 at 94%. However, this third ranked performance recorded a lower score than the fourth ranked Performance No. 5, apart from the construct of tone quality.

Figure 2E shows the result of a cluster analysis of Performer E's evaluation. Performance No. 6 was ranked the best and was strongly associated with Performance No. 4, ranked the second lowest at 99%. Surprisingly, the best performance did not have the highest ratings on all constructs. Compared with each rating on performances and the rank, it was not always likely to be consistent. The construct for rating, which is the most consistent in the ranking, seems to be sound quality. Tempo and the effect of 'Agogik' (which is a German word indicating the way of tempo changes) were highly associated with each other at 99%,

1 Ward's method, namely 'Ward's minimum variance method', was originally proposed by Joe H. Ward Jr. in 1963. In hierarchical cluster analysis, this method is one of most commonly used approaches that combines objects "whose merger increases the overall within-cluster variance to the smallest possible degree, are combined" (Mooi and Sarstedt, 2011, p. 252).

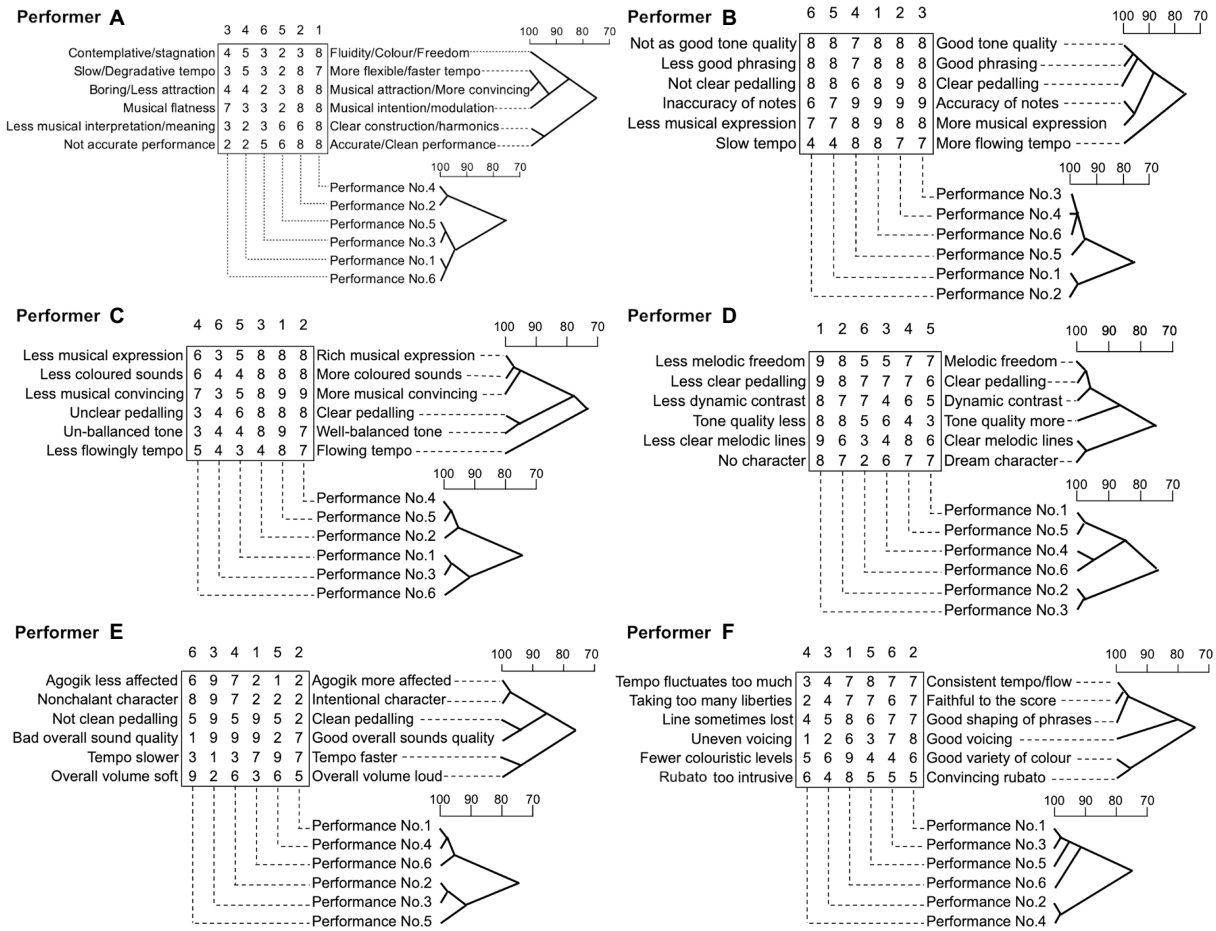


FIGURE 2
Clusters of performers' self-evaluations of their own performances.

as well as volume and pedalling being associated at 99%. 'Character' was associated with tempo and agogik at 83%. As performer E mentioned that he intentionally tried to perform differently on each trial, all recordings had various time ranges. It could be thought that tempo and agogik would be decided by what kind of character he would like to express. From the rating score, it could be said that more nonchalant character accompanies with the musical expression with less agogik (less tempo changes) in faster tempo.

Figure 2F displays the result of self-evaluation by Performer F. The best performance No. 6 marked the highest score, apart from the construction of 'good voicing' which gave the best score to the second ranked performance, No. 1. In particular, Performance No. 6 received the best score on the construct of 'variety of colour' which was consistent in the ranking. The constructions of 'tempo' and 'faithful to the score' were the most strongly associated at 99% and those of which were also associated with phrasing at 97%. The constructions of variety of colour and convincing rubato were associated with each other at 97%. The second-best performance No. 1 was more associated with Performance No. 3 ranked the lowest (99%) and No. 5 (95%) ranked the second lowest than the best performance (92%). The two lowest performances No. 3 and No. 5 did not overall record lower scores than the performances ranked the third and fourth. However, these two lower ranked performances were more deficient in 'variety of colour' which gave the highest score to the best performance No. 6. It seems that the construct of good voicing affects less on the overall ranking.

From all the results from the six pianists, there were three main conspicuous findings. The first finding was that the highest ranked recordings from each participant were likely to have obtained higher scores from certain criteria than others, such as in terms of musical expression, sound quality, phrasing and musical characteristics. The participants' interviewed wording was varied; for example, to express tone quality, several criteria were revealed, such as more coloured sound, well-balanced tone and sound quality. In other words, the lowest ranked recordings had lower scores on these particular criteria even though the recordings obtained higher scores on other criteria.

The second finding was that time-related elements, namely tempo and rubato, were set as criteria in order to identify the characteristics of the performance. However, these did not always

seem to affect a decision on the quality of performance. The focus piece of music, namely "Träumerei," could be characterised as 'slow and calm' as it is the slowest piece of "Kinderszenen" (op.15) and demands the utmost in legato passagework. Nevertheless, several participants used words to describe the tempo as 'flowing', 'consistent' or 'faster'.

The third finding was with regards to pedalling. The scores for pedalling were generally consistently related to the ranking of the recordings. However, several recordings had inconsistent scores in the rankings. It could be thought that pedalling would be an important component in deciding the quality of performance. However, other criteria such as musical expression, tone quality, phrasing and musical characteristics, were sometimes more dominant to determine the rankings.

Table 4 shows the results of overall self-evaluation ratings on their own six recordings by each participant. The rating range was from 1 to 9. The table indicates the mean rates and standard deviations. A one-way, between subjects, ANOVA was conducted to compare the difference of rating points on their own performances. There was a significant difference, $F(5, 210) = 8.11$, $p < 0.001$. *Post hoc* comparisons using the Turkey HSD test indicated that the mean score for the rating was significantly different between Performer B and all the others: B and A ($p < 0.001$), B and C ($p = 0.03$), B and D ($p = 0.11$) B and E ($p < 0.001$) and B and F ($p < 0.001$). Also, there was a significant difference between Performer A and Performer D ($p = 0.09$). It can be said that Performer B rated her own performances significantly higher than other performers did.

The results of the external evaluation

At the end of the external evaluation session (the third session), all participants were informed that one of the recordings was their own performance and that this was the third in the sequence. Apart from Performer A, all other performers were able to identify their own recording amongst the six recordings used in the third session. The stated reasons why they could identify their own best recording were reported as following (as multiple answers): tone colour (2 reports), phrasings (2), dynamics (1), tempo (1) and no idea (1). Performer A declined to answer.

Figure 3A illustrates the result of cluster analyses for both the constructs and the performances elicited by Performer A. This provides an illustration of the degree of association between constructs and between performances in a tree diagram. The top dendrograms shows the degree of association between constructs. The 6×6 matrix of numbers presents the rating of each performance in each construct. Under the matrix, the dotted line indicates the performer of the recording for each column. Above the matrix, the numbers show the ranking of the performance, which displays as 1 (the highest rank) to 6 (the lowest rank). Each performer's best recordings used in the third session are displayed at the bottom of the matrix with dotted lines.

This Figure 3A also displays the result of a cluster analysis of the third session by Performer A. The recording performed by Performer F, which received the best score on the construct of "Convincing interpretation" and "Good overall flow & story telling", was chosen as the best performance. The performance by Performer E, received the best score on the construct of 'Phrase-related rubato', and was ranked as the second-best performance. The constructions of 'Convincing interpretation' and "Good overall flow & story telling" were the most strongly associated at 99%. This cluster was also associated with 'phrase-related rubato' and 'range of dynamics' at 98%. The construction of 'tempo' was the least associated with others at 75%. Apart from 'tempo' and 'pedalling', other constructs were seen to be mostly consistent with the ranking. Performances played by Performer B (ranked fourth) and Performer C (ranked third) were the most strongly associated at 99% and were both associated with the performance by Performer A (ranked fifth) at 90%. The best-chosen performance by Performer F was the most strongly associated with the second-best version by Performer E at 99%. The performance by Performer D, which was the lowest ranked recording, was the least associated with the others at 75%. Performer A ranked her own recording as the fifth under the condition of disclosed information in which her recording was included. These clusters were created for each participant as same as Performer A's results (Figures 3B–F).

The attribution of criteria seemed to be partly similar to the criteria in the self-evaluation. There were several noticeable findings as follows. For several participants, 'interpretation' seemed to be an important key feature in deciding the performance quality. As the participants listened to other pianists' recordings, differences in interpretation were noticed and subject to comment. Yet, in the second phase self-evaluation, the 'interpretation of music' did not appear as a criterion, presumably because the listener (namely the performer) already knew their interpretation of the focused musical piece.

Overall, each participant's data attracted different constructs in the evaluation and ranking of the performances. And each pianist had different perspectives and prioritisation in deciding which performance would be the best or highly ranked. However, the ratings did not always agree with ranking. This suggests that some criteria were being prioritised in the decision to choose a better/best performance. Although the participants ranking of the six performances were different, there were some underlying relationships evident amongst these six recordings in terms of their evaluations. Figure 4 illustrates the results of a cluster analysis using Ward's method of all the constructs provided by all participants for the six recordings. The index of capital letters in the leftmost column is linked to which performer provided the set

TABLE 4 Overall rating on self-evaluation.

Performer	A	B	C	D	E	F
Mean	4.78	7.64	6.11	6.36	5.28	5.50
SD	2.31	1.16	2.07	1.75	2.91	1.95

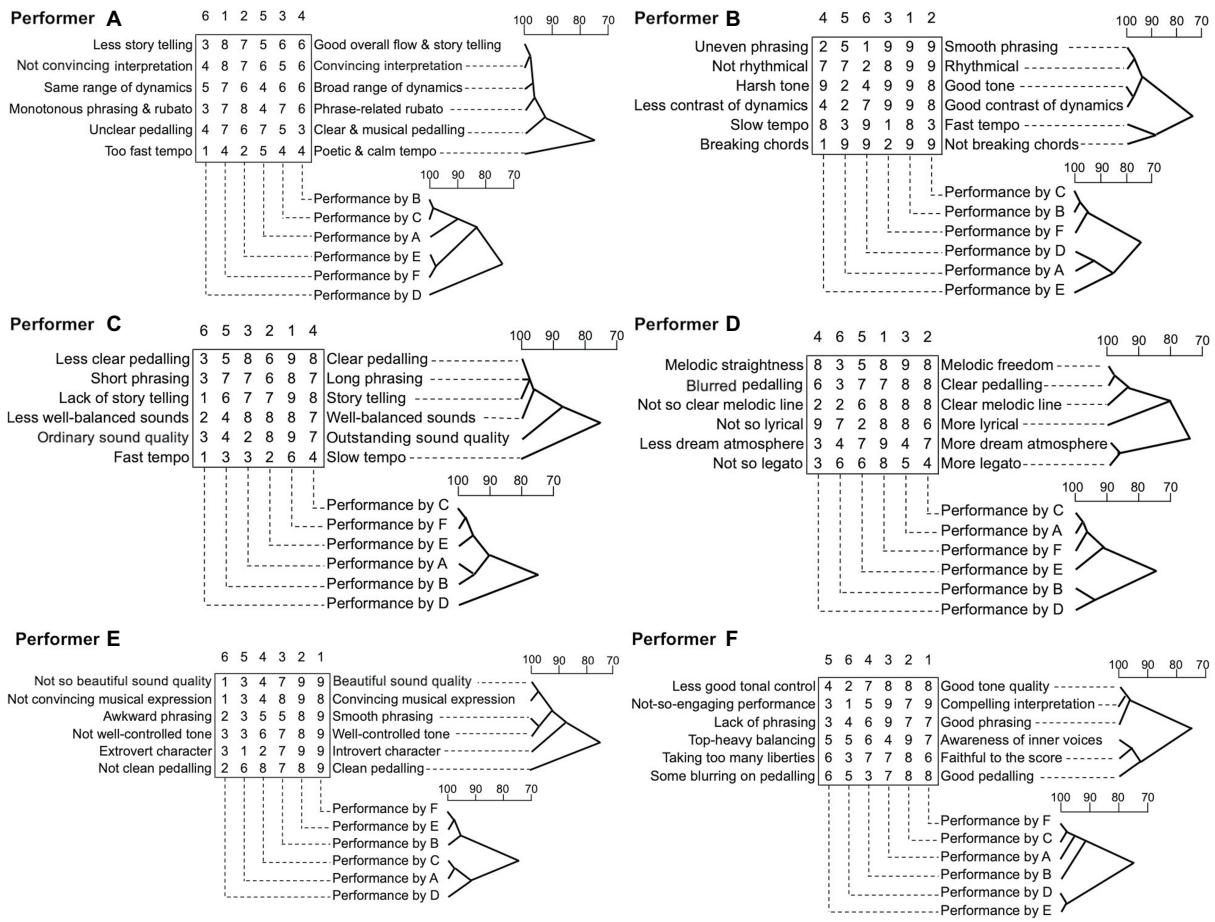


FIGURE 3
Clusters of performers' evaluations of the performances of others.

of constructs directly horizontal to it. For example, the first set of constructs, namely 'not so beautiful sound quality' and 'beautiful sound quality', were provided by Performer E. Numbers 1–9 inside the rectangular box display ratings for each performance as played by Performer A, B, C, D, E, and F in terms of the horizontal constructs. Above and below the ratings box, the capital letters indicate the performer of the recording for each column. The numbers with a round bracket above the capital letter indicates how the performer was ranked overall by all performers. For example, the leftmost column in the box indicates ratings for all constructs by all performers (evaluators) for the lowest ranked (the sixth rank) performance which was by Performer D.

The top dendrograms show the degree of associations amongst all the constructs provided by all performers (acting as the role of evaluators). The lower dendrograms indicate the degree of associations amongst all recordings used in the third session. The top dendrograms with all the constructs are divided into two large branches, which are associated with each other at 75%. The top branch consists of two sub-branches, which are connected with each other at 93%. These sub-branches contain lower sub-branches including several different musical perspectives as constructs; mainly, interpretation, tone quality, musical expression, and dynamics. The lower branch, which is associated with the top branch at 75%, shows a difference structure from the top one. In Figure 4 the branches of [1] and [2] contain the element of phrasing, and both of these are associated with another branch [3] at 91%, which includes the element of pedalling. These three branches are connected to the next branch [4] consisting of tempo at 89%. And then, they are associated with the branch [5] of 'not breaking chords' and 'more lyrical' at 85%. Compared with the two primitive branches divided at 75%, the top branch is likely to contain the constructs of more musical expression, tone quality and dynamics. On the other hand, the lower branch contains more focus on phrasing, tempo and pedalling. The construct related to interpretation appeared in both branches as well as musical atmosphere, for example 'dream atmosphere' and 'story telling'.

The leftmost capital letters indicate the performers who provided the construct in a horizontal line, and also provides some ideas of the relationships amongst the performers giving the constructs. By focusing on the two large branches separated at 75%, the top branch contains all the constructs provided by Performer E. Also, this includes the four constructs provided by Performer A. Correspondingly; the lower branch comprises all the constructs assigned by Performer F and embraces the five

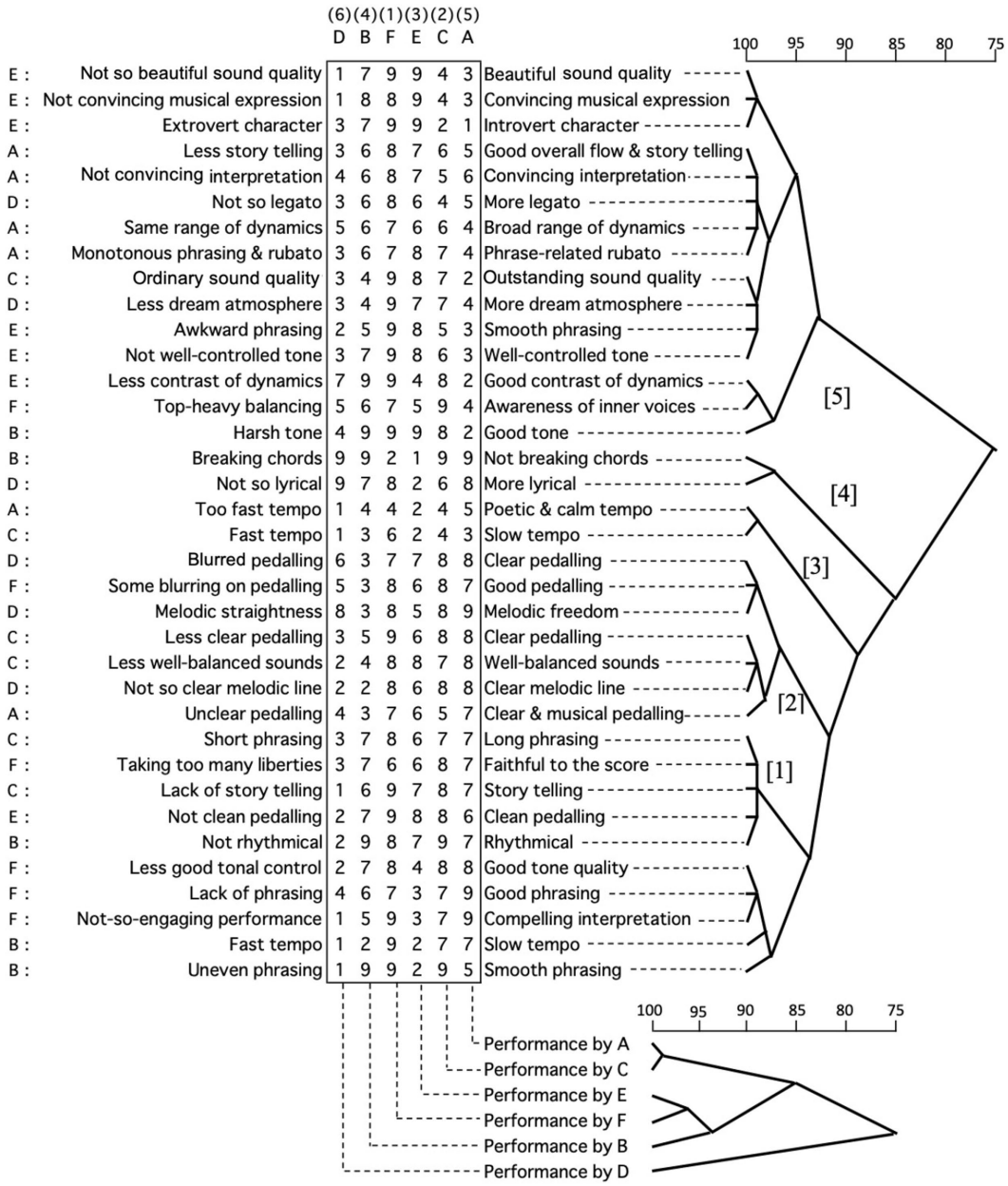


FIGURE 4
Cluster of all performance and constructs from external evaluation.

constructs given by Performer C. In the top branch, the top three constructs were given by Performer E and were strongly associated with each other at 99%. Similarly, below these three constructs by Performer E, the four constructs assigned by Performer A were associated with each other at 99% in a branch. Also in the lower branch, the three constructs given by Performer F were highly associated with each other at 99%. Consequently, it could be said that several constructs provided by the same person were likely to be highly associated, although the constructs provided by Performers B and D seem to be separated more into both branches.

Regarding the recordings, the strongest association was between the performances by Performers A and C at 99%. The top ranked performance by Performer F was the most associated with the performance by Performer E at 96%, both of which were associated with the one by Performer B at 93%. The set of performances by Performers A and C is associated with the cluster constructed of performances by Performers E, F and B at 85%. The lowest ranked performance by Performer D was the least associated with all others at 75%.

The Supplementary Table 1 indicates the results of overall rating on external evaluation by each participant in terms of how

they evaluated all six recordings acting as an adjudicator. The ratings range was from 1 to 9 and the table indicates the mean ratings and standard deviations. A one-way between subjects Analysis of Variance was conducted to compare the difference of the rating points in the external evaluation. There was no significant difference amongst the participants ($F(5,210) = 0.779$, p = n.s.). It could be said that their ratings behaviours were not affected by their individual differences and tendencies.

The Supplementary Table 2 indicates the results of the ranking of each performance at the external evaluation session. The rankings were converted as following: Ranking No. 1 = 6 points, No. 2 = 5 points, No. 3 = 4 points, No. 4 = 3 points, No. 5 = 2 points and No. 6 = 1 point. The left row shows each performer's best performance used in the session and the column header shows evaluators, for example 'A' indicates Performer A. The matrix illustrates the converted points based on their rankings. The numbers in bold and underlined indicate the evaluator's own performance. The mean points of each performance for the rankings are, ordered from the highest points to the lowest points: Performer F: 5.7, Performer C: 4.2, Performer E: 3.7, Performer B: 3.2, Performer A: 3.0, Performer D: 1.3. Overall, the performance by Performer F was ranked the best and that by Performer D was ranked as the lowest. The performance by Performer B was assigned the largest standard deviation. The agreement of ranking amongst evaluators was subjected to a Kendall Coefficient of Concordance analysis. The result suggested that the null hypothesis should be rejected at the 0.05 significance level ($w = 0.584$, $p = 0.004$). This implies that the evaluations of the selected recordings, including the evaluator's own recording, could be concordant to some degree. Performer B and D evaluated their own performances relatively higher compared to the ratings by the other participants.

The participants provided six sets of criteria in order to evaluate the recordings in both the self-evaluation session and the subsequent external evaluation session. All constructions provided by the six pianists were categorised into the 13 original criteria (overall flow, tone quality, interpretation of music, tempo, dynamics, rhythm, melodic accuracy, style, rubato, pedalling, technique, musical expression, phrasing), which were elicited from previous research studies and provided to the participants as potential constructs. All constructs elicited by all the participants were categorised into these criteria and an analysis was undertaken using Spearman's correlations to find the relationships amongst each element and ranking. The results indicated that there were significant correlations between: *Overall flow* and tone quality ($r = +1.000$), musical expression ($r = +0.829$), rubato ($r = +0.899$); *Tone quality* and musical expression ($r = +0.829$), rubato ($r = +0.899$); *Pedalling* and phrasing ($r = +0.943$), ranking ($r = +0.829$); and *Rubato* and ranking ($r = +0.812$).

Comparing the results from the self-evaluation and external evaluation, some criteria highly overlapped for each performer. The criteria used in these external evaluations included tone quality, phrasing, pedalling, tempi, and overall musical expression. An analysis using Kendall's Coefficient of Concordance was undertaken in order to compare with both self and external evaluation criteria. The results revealed that they significantly overlap ($w = 0.746$, $p < 0.001$). It could be said that pianists in this research have similar constructs of criteria for the evaluation of piano performances, whether by themselves or by other pianists.

Regarding their decision concerning their best recording in both the recording session and self-evaluation session, the participants partly showed different decisions. At the recording session, Performer A, Performer B and Performer E decided that their sixth (final) recordings were thought to be the best, whilst Performers C and D chose their fifth performances. Performer F decided that the fourth one was their best. Data analyses showed that the participants chose their best recordings from their latter trials. In the self-evaluation session, Performers B, E and F chose their sixth recordings as the best. For other participants, Performer A chose the fourth recording and Performer C's choice was the fifth one. Performer D decided that the third was the best. Matching of the decisions between the recording and the self-evaluation sessions was 50%. In each session, participants were likely to choose the best performance from the latter half.

Discussion and conclusions

This research study aimed to explore how the perceived quality of performance might be decided by professional standard pianists and what criteria might be applied, both with regards their own performances as well as the performance of others. In terms of ratings of their own performance, overall self-ratings showed a significant difference between several participants. Their evaluation behaviour in both the self-evaluation and the external evaluation were likely to be consistent in terms of how they perceived their own performances. For example, the participants who rated their own performance overall highly in the self-evaluation were likely to rank their own selected recording in the external evaluation more highly than the others did. Correspondingly, a participant who was perhaps a bit strict in the self-evaluation was likely to rank their own performance lower in the external evaluation than the others did.

In the self-evaluation phase, the criteria that performers used to evaluate performances were mainly the musical and performance elements related to tone quality, phrasing, pedalling, tempi and overall musical expression, such as storytelling and having a 'dream character'. Even though a performance may have received higher scores on other criteria, the criteria related to musical expression were likely to be more dominant, or could be an element to raise the ranking. Also, five of the six participants gave a construct of tempo, for example slow or fast tempo. The constructs of tempo tended to be associated with interpretation. It could be thought that tempo was an important feature to describe the characteristics of the performance.

In the external evaluation, as well as self-evaluation, the most influential factor in deciding the performance quality was related to tone quality, phrasing and musical expression. Compared with

the self-evaluation, none of the constructs was related to technical precision. As all participants in this study were professional pianists, it could be thought that their fundamental skill of performance did not create any concerns with technical issues. Therefore, their judgements could be focused more at a musical level, rather than on basic mechanical or technical aspects (*cf* Chaffin and Imereh, 2001). In particular, as the items for their role as external evaluators were made of their best performances, it is assumed that the quality of these particular recordings was relatively high. It would seem that technical precision was not the main focus of attention in their evaluation. It seems likely that the reason why there was no judgement evident on technical precision was the relatively high quality of performances.

Also, five of the six participants gave the construct of pedalling, for example 'clear pedalling' and 'musical pedalling'. The criterion of pedalling would be piano specific as an influence on perceived performance quality and was related to phrasing or interpretation. Statistically, the criteria for self-evaluation and for in external evaluation highly overlapped for each performer (Kendall Coefficient of Concordance, $w = 0.746$, $p < 0.001$). It could be said that pianists in this research have similar constructs of criteria for the evaluation of piano performances, whether performed by themselves or by other pianists. In addition, several criteria given by the same performer in the external evaluation were likely to be classified closely in the cluster analysis, such as tone quality, musical expression and the quality of pedalling.

Evidence of participants' evaluations being considered as consistent and reliable may be drawn from an agreement that the performance by Performer F was assigned high ratings and chosen as the best overall, whereas the performance by Performer D was ranked as the lowest. One of the characteristics of the performance by Performer F was its slowest tempo, which may be considered to well-reflect the characteristics of the piece "Träumerei," which generally means "dreaming" and is sometimes translated by the term "Reverie." Their evaluation of their own performances, namely the hidden self-evaluation in the external evaluation, could be different from the evaluation by others. This suggests that their own perspectives could be influential on their priorities and the preferences in deciding the nature of a better performance.

In this external-evaluation session, one of the six recordings was their own recording previously chosen by them as the best from the self-evaluation session. Their own performances were placed third in the set of the six recordings to be evaluated by all participants. Apart from one pianist (who declined to make a judgement), all the other performers were able to identify their own recording amongst the six used in the external evaluation session, having been informed at the end of the blind judging session that one of recordings was their own one. This finding also supports the outcomes of the research study conducted by Repp and Knoblich (2004), which demonstrated that pianists were able to recognise their own recording amongst several performances by others. According to the participants' own reports in this current study, the reasons why they could identify their own recordings were mainly related to tone quality, phrasing, and dynamics. It would be thought that several specific features of the performance can be kept in mind by the pianist and they can feel and perceive these whilst listening Repp and Knoblich (2004) reported that the identification of self-performance was successful in their study despite of editing of tempo and dynamics. This current research suggests that sense of tone colour and phrasing could be the potentially important features in the self-identification of performance quality. Interestingly, these two elements were also demonstrated as important elements in an external evaluation session.

The performance criteria, namely viewpoints of the decision of performance quality, in both self-evaluation and external evaluation predominately overlapped in terms of musical factors. Comparing the criteria in both sessions, more than half overlapped within the same person. Performers C, D and F provided the four same criteria in both their self-evaluation and external evaluation. On the one hand, some of their written observations are slightly different in terms of wording to the provided criteria; on the other hand, some were exactly the same. Regarding Performers B and E, half of the criteria in both sessions overlapped. Performer A showed a little variety, however, with two criteria being the same in self-and external-evaluations. The result of the analysis using Kendall's Coefficient of Concordance also showed that the categories for both criteria significantly overlapped. It can be inferred that performers' constructed criteria for both self-evaluation and external evaluation are relatively associated.

The results from the current research empirically demonstrated that criteria related to tone quality and musical expression appear to be the most dominative components in deciding the overall quality of performance in both the self-evaluation and the external evaluation phases. Focused on the external evaluation, the most assigned element was phrasing (7/36 items) and then tone quality (6/36 items). Particularly when in the role of external evaluator, tone quality and overall flow were the most associated in the decision to award higher rankings. These results are supported, at least in part, by extant literature, such as the studies by Russell (2010) and Thompson et al. (1998). Russell (2010) found that the component of musical expression had a significantly direct effect on the overall perception of quality. Thompson et al. (1998) found that an overall assessment was strongly related to the evaluation of musical expression, phrasing and right-hand expression. Tempo could be important to identify the quality of performance in this study; however, it was not the main element to emerge in determining the quality of the performances. This finding agreed with the outcome of the study by Thompson et al. (1998) which reported that tempo could be important for identifying performance quality; however, it was not highly associated with overall preferences.

At the opening recording session, all performers reported a perception that the 'best' performance was in the latter half of their playing set, especially the fifth and sixth versions. In the self-evaluation session with playback, the participants were still likely to choose the best performance from the latter half of the session.

These results support other research suggesting that the order of performance can be influential in evaluation (Duerksen, 1972; Flôres and Ginsburgh, 1996). However, it may be said that even these professional pianists did not always make a decision of their best performance concordantly in both recording time (just after the performances) and later at the time of self-evaluation.

This research study suggests that the participant professional pianists did not always consistently evaluate their own performance as others did. However, in terms of the relationship between the roles of self-evaluation and external evaluation by the same performer, the tendencies evidenced within self-evaluation could be found in the context of the role of external evaluator. These interactions indicated that a self-constructed tendency of evaluation is the basis of specific and individual attitudes when deciding the comparative quality of musical performances, by self and others.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding author.

Ethics statement

The studies involving human participants were reviewed and approved by The Institute of Education, formerly University of London, now University College London. The patients/participants provided their written informed consent to participate in this study.

Author contributions

The research cited originally formed part of the doctoral studies of YM. Supervised and supported by GW. The methodology was designed collaboratively and the fieldwork was undertaken by YM. All authors contributed to the article and approved the submitted version.

Funding

This research study was partly supported by The Arnold Bentley New Initiatives Fund by SEMPRE.

Acknowledgments

The authors would like to thank all the participants for their contributions to this study.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2022.954261/full#supplementary-material

References

Abeles, H. F. (1973). Development and validation of a clarinet performance adjudication rating scale. *J. Res. Music. Educ.* 21, 246–255. doi: 10.2307/3345094

ABRSM (2006). *Piano Exam Pieces 2007–2008: Grade 7 Complete Syllabus*. London: ABRSM Publishing.

Alessandri, E., Williamson, V. J., Eiholzer, H., and Williamon, A. (2016). A critical ear: analysis of value judgments in reviews of Beethoven's piano sonata recordings. *Front. Psychol.* 7:391. doi: 10.3389/fpsyg.2016.00391

Almansa, J., and Delicado, P. (2009). Analysing music performance through functional data analysis: rhythmic structure in Schumann's Träumerei. *Connect. Sci.* 21, 207–225. doi: 10.1080/09540090902733848

Andrade, H., and Du, Y. (2007). Student responses to criteria-referenced self-assessment. *Assess. Eval. High. Educ.* 32, 159–181. doi: 10.1080/02602930600801928

Bannister, D., and Mair, J. M. M. (1968). *The Evaluation of Personal Constructs*. London: Academic Press.

Beail, N. (1985). "An introduction to repertory grid technique" in *Repertory Grid Technique and Personal Constructs: Application in Clinical and Educational Settings*. ed. N. Beail (Kent: Croom Helm), 1–26.

Beran, J., and Mazzola, G. (2000). Timing microstructure Schumann's "Träumerei" as an expression of harmony, rhythm, and motivic structure in music performance. *Comput. Mathmatics Appl.* 39, 99–130. doi: 10.1016/S0898-1221(00)00049-3

Bergee, M. J. (1993). A comparison of faculty, peer, and self-evaluation of applied brass jury performances. *J. Res. Music. Educ.* 41, 19–27. doi: 10.2307/3345476

Bergee, M. J. (1997). Relationship among faculty, peer and self-evaluations of applied performances. *J. Res. Music. Educ.* 45, 601–612. doi: 10.2307/3345425

Bergee, M. J. (2003). Faculty interjudge reliability of music performance evaluation. *J. Res. Music. Educ.* 51, 137–150. doi: 10.2307/3345847

Bergee, M. J., and Cecconi-Roberts, L. (2002). Effects of small-group peer interaction on self-evaluation of music performance. *J. Res. Music. Educ.* 50, 256–268. doi: 10.2307/3345802

Blatchford, P. (1997). Students' self-assessment of academic attainment: accuracy and stability from 7 to 16 years and influence of domain and social comparison group. *Educ. Psychol.* 17, 345–359. doi: 10.1080/0144341970170308

Boud, D. (1991). Implementing student self-assessment: issue 5 of green guide. 2nd edn. Sydney: Higher Education Research and Development Society of Australasia.

Boud, D. (1995). *Enhancing Learning Through Self-Assessment*. London: Routledge.

Brändström, S. (1996). Self-formulated goals and self evaluation in music education. *Bull. Counc. Res. Music. Educ.* 127, 16–21. Available at: http://www.jstor.org/stable/40318760

Cambouropoulos, E., and Widmer, G. (2000). "Melodic clustering: Motivic analysis of Schumann's Träumerei," in *Paper presented at the III Journées d' Informatique Musicale, Bordeaus, France*.

Chaffin, R., and Imereh, G. (2001). A comparison of practice and self-report as sources of information about the goals of expert practice. *Psychol. Music* 29, 39–69. doi: 10.1177/0305735601291004

Chaffin, R., and Imereh, G. (2002). Practicing perfection: piano performance as expert memory. *Psychol. Sci.* 13, 342–349. doi: 10.1111/j.0956-7976.2002.00462.x

Ciorba, C. R., and Smith, N. Y. (2009). Measurement of instrumental and vocal undergraduate performance juries using a multidimensional assessment rubric. *J. Res. Music. Educ.* 57, 5–15. doi: 10.1177/0022429409333405

Daniel, R. (2001). Self-assessment in performance. *Br. J. Music Educ.* 18, 215–226. doi: 10.1017/S0265051701000316

Doerksen, P. F. (1999). Aural-diagnostic and prescriptive skills of preservice and expert instrumental music teachers. *J. Res. Music. Educ.* 47, 78–88. doi: 10.2307/3345830

Duerksen, G. L. (1972). Some effects of expectation on evaluation of recorded musical performance. *J. Res. Music. Educ.* 20, 268–272. doi: 10.2307/3344093

Ekholm, E. (1997). Evaluation of solo vocal performance: a comparison of students with experts. Update: applications of research in music. *Education* 16, 3–7. doi: 10.1177/875512339701600102

Finny, S. A., and Palmer, C. (2003). Auditory feedback and memory for music performance: sound evidence for an encoding effect. *Mem. Cogn.* 31, 51–64. doi: 10.3758/BF03196082

Flôres, R. G., and Ginsburgh, V. A. (1996). The Queen Elisabeth musical competition: how fair is the final ranking? *Statistician* 45, 97–104. doi: 10.2307/2348415

Fransella, F., and Bannister, D. (1977). *A Manual for Repertory Grid Technique*. London: Academic Press.

Friberg, A. (1995). Matching the rule parameters of phrase arch to performance of "Träumerei": a preliminary study. *Speech Music Hearing Q. Progr. Status Rep.* 36, 63–70.

Geringer, J. M., and Madsen, C. K. (1995/1996). Focus of attention to element: listening patterns of musicians and nonmusicians. *Bull. Counc. Res. Music. Educ.* 127, 80–87. Available at: https://www.jstor.org/stable/40318770

Gilbert, L. (1990). Aesthetic development in music: an experiment in the use of personal construct theory. *Br. J. Music Educ.* 7, 173–190. doi: 10.1017/S0265051700007762

Gingras, B., Asselin, P.-Y., and McAdams, S. (2013). Individuality in harpsichord performance: disentangling performer-and piece-specific influences on interpretive choices. *Front. Psychol.* 4:895. doi: 10.3389/fpsyg.2013.00895

Gingras, B., Lagrandeur-Ponce, T., Giordano, B. L., and McAdams, S. (2011). Perceiving musical individuality: performer identification is dependent on performer expertise and expressiveness, but not on listener expertise. *Perception* 40, 1206–1220. doi: 10.1068/p6891

Gordon, M. (1991). A review of the validity and accuracy of self-assessment in health professions training. *Acad. Med.* 66, 762–769. doi: 10.1097/00001888-199112000-00012

Gordon, S. (1996). *Robert (Alexander) Schumann a History of Keyboard Literature: Music for the Piano and its Forerunners* Belmont: Schirmer 258.

Hargreaves, D. J., and Colman, A. M. (1981). The dimensions of aesthetic reactions to music. *Psychol. Music* 9, 15–20. doi: 10.1177/030573568100900103O1

Hentschke, L., and Del Ben, L. (1999). The assessment of audience-listening: testing a model in the educational setting of Brazil. *Music. Educ. Res.* 1, 127–146. doi: 10.1080/1461380990010202

Hewitt, M. P. (2011). The impact of self-evaluation instruction on student self-evaluation, music performance, and self-evaluation accuracy. *J. Res. Music. Educ.* 59, 6–20. doi: 10.1177/0022429410391541

Johnson, C. M. (1996). Musicians' and nonmusicians' assessment of perceived rubato in musical performance. *J. Res. Music. Educ.* 44, 84–96. doi: 10.2307/3345415

Johnson, P. (1997). Performance as experience: the problem of assessment criteria. *Br. J. Music Educ.* 14, 271–282. doi: 10.1017/S0265051700001248

Jones, H. (1986). *An Application of the Facet-Factorial Approach to Scale Construction in the Development of a Rating Scale for High School Vocal Solo Performance. PhD thesis*. University of Oklahoma, Norman, OK

Jung, E. (2003). "Pianists interpretative intentions in expressive performance," in *Paper presented at the 5th triennial ESCOM conference, Hanover University of Music and Drama, Germany*.

Juslin, P. N. (2003). Five facets of musical expression: a psychologist's perspective on music performance. *Psychol. Music* 31, 273–302. doi: 10.1177/03057356030313003

Kapilow, R. (2011). *Robert Schumann (1810–1956) "Träumerei" from Kinderszenen What Makes it Great: Short Masterpieces, Great Composers* 156–165. Hoboken, NJ: John Wiley & Sons.

Keller, P. E., Knoblich, G., and Repp, B. H. (2007). Pianists duet better when they play with themselves: on the possible role of action simulation in synchronization. *Conscious. Cogn.* 16, 102–111. doi: 10.1016/j.concog.2005.12.004

Kelly, G. A. (1955). *The Psychology of Personal Constructs*. New York: Norton.

Koren, R., and Gingras, B. (2014). Perceiving indivivuality in harpsichord performance. *Front. Psychol.* 5:141. doi: 10.3389/fpsyg.2014.00141

Kostka, M. J. (1997). Effects of self-assessment and successive approximations on "knowing" and "valuing" selected keyboard skills. *J. Res. Music. Educ.* 45, 273–281. doi: 10.2307/3345586

Lehmann, A. C., Sloboda, J., and Woody, R. H. (2007). *Psychology for Musicians: Understanding and Acquiring the Skills*. New York: Oxford University Press.

Lerdahl, F., and Jeckendoff, R. S. (1983). *A Generative Theory of Tonal Music*. Cambridge, MA: The MIT Press.

Madsen, C. K. (1990). Measuring musical response. *Music. Educ. J.* 77, 26–28. doi: 10.2307/3397835

Madsen, C. K., Geringer, J. M., and Fredrickson, W. E. (1997). Focus of attention to musical elements in Haydn's 'symphony #4'. *Bull. Counc. Res. Music. Educ.* 133, 57–63. Available at: http://www.jstor.org/stable/40318840

Magrath, J. (1993). "Träumerei, Op. 15, No. 7" in *Melodious Masterpieces: Standard Early Advanced Literature for Expressive Performance*. ed. J. Magrath, vol. *3* (Van Nuys: California: Alfred Music Publishing), 4–34.

Mathias, B., Gehring, W. J., and Palmer, C. (2017). Auditory N1 reveals planning and monitoring processes during music performance. *Psychophysiology* 54, 235–247. doi: 10.1111/psyp.12781

Mazzola, G. (2011). *Schumann's Träumerei: The First Performance Experiment with RUBATO. Music Performance: A Comprehensive Approach: Theory, Analytical Tools, and Case Studies* 201–205. Berlin; Heidelberg: Springer.

McPherson, G. E., and Schubert, E. (2004). "Measuring performance enhancement in music" in *Musical Excellence: Strategies and Techniques to Enhance Performance*. ed. A. Williamon (New York: Oxford University Press), 61–83.

McPherson, G. E., and Thompson, W. F. (1998). Assessing music performance: issues and influences. *Res. Stud. Music Educ.* 10, 12–24. doi: 10.1177/1321103X9801000102

McPherson, G. E., and Zimmerman, B. J. (2002). "Self-regulation of musical learning: a social cognitive perspective" in *The New Handbook of Research on Music Teaching and Learning*. eds. R. Colwell and C. Richardson (New York: Oxford University Press), 327–347.

Mills, J. (1987). Assessment of solo musical performance – a preliminary study. *Bull. Counc. Res. Music. Educ.*, 91, 119–125. Avaialbel at: http://www.jstor.org/stable/40318071

Mooi, E., and Sarstedt, M. (2011). A Concise Guide to Market Research: The Process, Data, and Methods Using IBM SPSS Statistics. Heidelberg: Springer. p. 308. doi: 10.1007/978-3-642-12541-6

Nichols, J. P. (1991). A factor analysis approach to the development of a rating scale for snare drum performance. *Dialogue Instrumental Music Educ.* 15, 11–31.

Nunnally, J. C. (1978). *Psychometric Theory*. New York, NY: McGraw-Hill.

Ostwald, P. F. (1985). *Symbolic Union with Clara, 1837–1839 Schumann: The Inner Voices of a Musical Genius* (pp. 133–150). Westford, MA: Northeastern University Press.

Palmer, C. (1996). On the assignment of structure in music performance. *Music. Percept.* 14, 23–56. doi: 10.2307/40285708

Palmer, C., Jungers, M. K., and Jusczyk, P. W. (2001). Episodic memory for musical prosody. *J. Mem. Lang.* 45, 526–545. doi: 10.1006/jmla.2000.2780

Repp, B. H. (1992). Diversity and commonality in music performance: an analysis of timing microstructure in Schuman's "Träumere". *J. Acoust. Soc. Am.* 92, 2546–2568. doi: 10.1121/1.404425

Repp, B. H. (1994). On determining the basic tempo of an expressive music performance. *Psychol. Music* 22, 157–167. doi: 10.1177/0305735694222005

Repp, B. H. (1995). Expressive timing in Schuman's "Träumerei": an analysis of performances by graduate student pianists. *J. Acoust. Soc. Am.* 98, 2413–2427. doi: 10.1121/1.413276

Repp, B. H. (1996). The dynamics of expressive piano performance: Schumans Träumerei revisited. *J. Acoust. Soc. Am.* 100, 641–650. doi: 10.1121/1.415889

Repp, B. H. (1999). Effects of auditory feedback deprivation on expressive piano performance. *Music. Percept.* 16, 409–438. doi: 10.2307/40285802

Repp, B. H., and Knoblich, G. (2004). Perceiving action identity: how pianists recognize their own performances. *Psychol. Sci.* 15, 604–609. doi: 10.1111/j.0956-7976.2004.00727.x

Ross, S. (1998). Self-assessment in second language testing: a meta-analysis and analysis of experiential factors. *Lang. Test.* 15, 1–20. doi: 10.1177/026553229801500101

Ross, J. A. (2006). The reliability, validity, and utility of self-assessment. *Pract. Assess. Res. Eval.* 11, 1–10. doi: 10.7275/9wph-vv65

Russell, B. E. (2010). *The Empirical Testing of a Musical Performance Assessment Paradigm. PhD.* University of Miami, Coral Gables, Florida.

Saúl, L. A., López-González, M. A., Moreno-Pulido, A., Corbella, S., Compañ, V., and Feixas, G. (2012). Bibliometric review of the repertory grid technique: 1998-2007. *J. Constr. Psychol.* 25, 112–131. doi: 10.1080/10720537.2012.651065

Saunders, T. C., and Holahan, J. M. (1997). Criteria-specific rating scales in the evaluation of high school instrumental performance. *J. Res. Music. Educ.* 45, 259–272. doi: 10.2307/3345585

Schleff, J. S. (1992). Critical judgements of undergraduate music education students in response to recorded music performances. *Contrib. Music. Educ.* 19, 60–74. Available at: https://www.jstor.org/stable/24127397

Schmidt, R. A., and Lee, T. D. (2010). *Motor Control and Learning: A Behavioral Emphasis* (*5th Edn*). Champaign, IL: Human Kinetics.

Seashore, C. E. (1938). *Psychology of music.* New York: McGraw Hill Book Company Inc.

Shimosako, H., and Ohgushi, K. (1996). Interaction between auditory and visual processing in impressional evaluation of a piano performance. *J. Acoust. Soc. Am.* 100:2779. doi: 10.1121/1.416429

Sloboda, J. A. (2000). Individual differences in music performance. *Trends Cogn. Sci.* 4, 397–403. doi: 10.1016/S1364-6613(00)01531-X

Stanley, M., Brooker, R., and Gilbert, R. (2002). Examiner perceptions of using criteria in music performance assessment. *Res. Stud. Music Educ.* 18, 46–56. doi: 10.1177/1321103X020180010601

Thompson, W. F., Diamond, C. T. P., and Balkwill, L.-L. (1998). The adjudication of six performance of a Chopin etude: a study of expert knowledge. *Psychol. Music* 26, 154–174. doi: 10.1177/0305735698262004

Van Vugt, F. T., Jabusch, H.-C., and Altenmüller, E. (2013). Individuality that is unheard of: systematic temporal deviations in scale playing leave an inaudible pianistic fingerprint. *Front. Psychol.* 4:134. doi: 10.3389/fpsyg.2013.00134

Wapnick, J., Flowers, P. J., Alegant, M., and Jasinskas, L. (1993). Consistency in piano performance evaluation. *J. Res. Music. Educ.* 41, 282–292. doi: 10.2307/3345504

Wapnick, J., Ryan, C., Campbell, L., Deek, R., Lemire, R., and Darrow, A.-A. (2005). Effects of excerpt tempo and duration on musicians' ratings of high-level piano performance. *J. Res. Music. Educ.* 53, 162–176. doi: 10.1177/002242940505300206

Wrigley, W. J. (2005). *Improving Music Performance Assessment. PhD.* Griffith University, Queensland, Australia.

Yorke, D. M. (1978). Repertory grid in educational research: some methodological considerations. *Br. Educ. Res. J.* 4, 63–74. doi: 10.1080/0141192780040205

Zdzinski, S. F., and Barnes, G. V. (2002). Development and validation of a string performance rating scale. *J. Res. Music. Educ.* 50, 245–255. doi: 10.2307/3345801

Zimmerman, B. J. (2000). "Attaining self-regulation: a social cognitive perspective," in *Handbook of Self-Regulation*. eds. M. Boekaerts, P. Pintrich and M. Zeidner (San Diego, CA: Academic Press), 13–39.

Zimmerman, B. J., and Schunk, D. H. (2001). *Self-regulated Learning and Academic Achievement: Theoretical Perspectives* (*2nd ed*). Mahwah, NJ: Erlbaum Associates.

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 17 November 2022
DOI 10.3389/fpsyg.2022.976218

OPEN ACCESS

EDITED BY
George Waddell,
Royal College of Music, United Kingdom

REVIEWED BY
Piera Centobelli,
University of Naples Federico II, Italy
Arash Akhshik,
Jagiellonian University,
Poland

*CORRESPONDENCE
Yuting Zhang
zhyt2003@163.com

SPECIALTY SECTION
This article was submitted to Performance Science, a section of the journal Frontiers in Psychology

RECEIVED 23 June 2022
ACCEPTED 20 October 2022
PUBLISHED 17 November 2022

CITATION
Duan W, Hu H, and Zhang Y (2022) What determines the performance of small and medium-sized enterprises supply chain financing? A qualitative comparative analysis of fuzzy sets based on the technology–organization–environment framework.
Front. Psychol. 13:976218.
doi: 10.3389/fpsyg.2022.976218

What determines the performance of small and medium-sized enterprises supply chain financing? A qualitative comparative analysis of fuzzy sets based on the technology–organization–environment framework

Weichang Duan[1], Hanzhou Hu[1] and Yuting Zhang[2]*

[1]School of Management, Guangzhou University, Guangzhou, China, [2]School of Business Administration, Anhui University of Finance and Economics, Bengbu, China

With the COVID-19 pandemic sweeping the globe, small and medium-sized enterprises' (SMEs') survival space has been increasingly constrained, and their financing challenges and costly concerns have also become more evident. With the emergence of the supply chain financing model, the problem of difficult financing for SMEs has been effectively alleviated. How to effectively improve the performance of supply chain financing for SMEs is a hot issue of common concern for both business and academic circles. This paper used 90 SMEs involved in supply chain financing business for a case study based on the "Technology–Organization–Environment" (TOE) framework, using fuzzy set qualitative comparative analysis (fsQCA), and explored the linkage effects of technology, organization, and environmental conditions on SMEs' performance in improving supply chain financing and their path choices. The study found that: (1) individual conditions do not constitute a necessary condition for high/low supply chain financing performance. However, technical level preconditions play a more important role in shaping the high supply chain financing performance of firms. (2) The three preconditions at the technological, organizational, and environmental levels work together to form a diverse set of conditions that drive the high supply chain financing performance of firms. That is, the paths driving the high supply chain financing performance of SMEs are characterized by "different routes to the same destination." There are three different models: "technology-supply chain capability-driven," "IT-supply chain capability-driven," and "IS-supply chain capability-driven." (3) The degree of application of corporate information technology, information sharing capability, and supply chain capability, and the lack of environmental competitiveness are the reasons for the generation of the low supply chain financing performance of SMEs. The above research findings can provide a direct theoretical basis for enterprise supply chain

financing practice and are of great practical significance in guiding Chinese SMEs on how to improve their supply chain financing performance.

KEYWORDS

supply chain financing, TOE, fsQCA, medium-sized and small enterprises, environmental competitiveness

Introduction

China's economic development has been greatly hampered by the new coronavirus outbreak. According to public data from the National Bureau of Statistics, private enterprises achieved a total profit of 120.83 billion CNY in January–February 2020, down 36.6% year on year due to the impact of the epidemic. During the epidemic, the weaknesses of SMEs and their dependence on cash flow gradually became apparent, causing them to face shortages of working capital and breaks in cash flow (Qin, 2021). SMEs are an integral part of our economic map. SMEs play an important role in providing employment for our population and in regulating the structure of the national economy (Xing, 2018). However, enterprises in development often encounter the problem of capital shortages, which has seriously restricted the rapid development of SMEs. SMEs are limited by their size, lack of collateral assets, lack of credit, and other factors, resulting in banks and other financial institutions being reluctant to lend to them due to risk control considerations.

Academics have conducted active research on how to help SMEs solve their financing dilemmas. Among them, the supply chain financing model is unanimously recognized by most scholars, who believe that this model can solve the financing problems of SMEs (Xin, 2007; Zhao, 2020; Pan et al., 2021). Supply chain finance has significant advantages in reducing transaction costs, weakening information asymmetry, and enhancing risk control, and plays an important role in solving the problem of difficult financing for SMEs (Xia and Jin, 2011). This financing model focuses on the authenticity and continuity of the transaction background, the closed nature of the business operation, and the self-repayment of the loan. Its main role is to solve the problem of financing difficulties caused by the lack of credit and collateral assets of SMEs, and to obtain funds from banks and other financial institutions with the credit of core enterprises to support their own development. The research on supply chain financing has found that not every supply chain enterprise can make use of the credit support obtained by the supply chain financing model, even if they are in the same supply chain. It is of strong practical guidance to explore those conditions that can affect the supply chain financing performance of enterprises, so that SMEs in the supply chain can build on their strengths and avoid their weaknesses to better obtain financial support from financial institutions.

As the research on supply chain financing continues, it is found that even for enterprises in the same supply chain, not every supply chain enterprise can make use of the credit support obtained by the supply chain financing model (Wei and Liu, 2012; Song et al., 2017a). In order to help SMEs better access supply chain financing and relieve their financial pressure, it is necessary to further dissect the mechanisms affecting the performance of supply chain financing for SMEs. In reality, the supply chain financing performance of SMEs is influenced by many factors, which are inevitably interrelated and work together to form a high supply chain financing performance. However, considering the limited number of previous research perspectives and analytical frameworks, it is difficult to deeply elaborate on the mechanisms affecting the supply chain financing performance of SMEs, leading to difficulties in determining the driving paths to improve the supply chain financing performance of SMEs. Thus, the research question of this paper is posed: what are the factors that affect the supply chain financing performance of firms? And how will these factors work together to affect the firm's supply chain financing performance?

Most of the existing studies have explored the relationship between individual factors such as firm network embeddedness and improved innovation capacity and supply chain financing performance. Based on principal-agent theory, found that intra-firm financial collaboration and financial collaboration between buyers and suppliers have a significant contribution to supply chain financing performance (Wandfluh et al., 2016). Combining information asymmetry theory and network theory, some scholars suggest that strong and bridge connections in supply chain networks can effectively promote the quality of financing for SMEs. Martin (2017), on the other hand, sorted out the role of supplier dependence, buyer bargaining power, and buyer–supplier trust in facilitating supply chain financing based on social exchange theory. Martin and Hofmann (2019) proposed based on transaction cost theory through an exploratory multi-case analysis that the effectiveness of supply chain financing can be judged based on two dimensions: the timing of financing and the source of funds. In addition, other scholars have explored the mechanism of the intrinsic influence of a firm's internal and external capabilities on supply chain financing performance based on capability theory using supply chain financing solution adoption as a mediating variable. These studies are mostly a power variable perspective, exploring the linear relationship between a single factor or moderating variable on firms' supply chain financing performance, which limits the understanding of the synergistic matching effects among multiple factors—such as technology,

organization, and environment—behind the differences in SMEs' supply chain financing performance.

To address the shortcomings of the above study (Wei and Liu, 2012; Wandfluh et al., 2016; Martin, 2017; Martin and Hofmann, 2019), this paper finds that the key point for SMEs to obtain supply chain financing is the willingness of the core enterprises in the supply chain to guarantee for them through the analysis of the supply chain financing model. In this paper, according to the logic that the characteristics of SMEs affect the willingness of core enterprises to guarantee their supply chain financing performance, five factors affecting the performance of supply chain financing are selected from the three levels of "technology, capability, and environment" to build a theoretical model framework affecting the performance of supply chain financing of SMEs. The fuzzy set qualitative comparative analysis (fsQCA) method is used to analyze the grouping of these five factors and derive the conditional grouping and the mechanism of action that cause the difference of enterprise supply chain financing performance; then, by exploring the linkage effect of many factors that affect performance, it helps to establish a systematic analysis idea for constructing the development path of SME supply chain financing performance. Secondly, it enables SMEs in the supply chain to build on their strengths and avoid their weaknesses, increasing the willingness of core enterprises to guarantee for them, and thus better access to financial institutions' financial support.

Literature review and research framework

Application of fsQCA method in the field of supply chain finance

In order to explore the impact mechanism of blockchain technology on supply chain finance, Hong et al. (2022) established the basic theoretical model of "digital credit co-governance-network embeddedness-supply chain finance performance" from the perspective of network embeddedness, used partial least squares to verify the structural equation model, and further used fuzzy set qualitative comparison analysis to analyze four financing performance dimensions, namely financing cycle, financing amount, financing availability, and financing cost, as the outcome variables for group analysis. Based on the basic theoretical logic of "firm capability-competitive advantage—firm performance" in the theory of firm capability, Lu et al. (2019a) uses fuzzy set qualitative comparative analysis to explore the impact of SMEs' innovation capability and market responsiveness as well as the adoption of supply chain financing solutions on the performance of supply chain financing. The impact on supply chain financing performance is explored by using fuzzy set qualitative comparative analysis. Li and Sun (2022) used "financing capability-financing intermediary-financing performance" as the theoretical research framework to construct the cooperative capability of core enterprises, and tested the research hypotheses based on Discriminant Analysis (DA) and Fuzzy set Qualitative Comparison Analysis (Fs-QCA) through the theoretical model that supply chain network integration affects supply chain financing performance. Yu and Wang (2022) used an empirical approach combined with the fuzzy set qualitative fixed ratio analysis (fsQCA) method to investigate the core firms' willingness to participate in supply chain finance and their financing model orientation.

In summary, the application of fuzzy set qualitative comparative analysis method in the field of supply chain amount has been widely paid attention to, and the advantage of fuzzy set qualitative comparative method which has both qualitative and quantitative analysis in studying the problem of causal complexity is obvious. Most of the existing studies, however, use fsQca as a complement to test the hypothesis.

Study on influactors of supply chain financing performance

A review of the existing papers reveals that studies on the factors influencing the performance of supply chain financing can be grouped into the following areas:

A lot of theoretical analysis has been described in these papers on how these ITs can be applied in supply chain finance, and specific application architectures have been provided. Didi (2019) suggested that supply chain finance itself is a high-tech business that fits well with blockchain technology, IoT technology, and other information technologies. Further, Yanni (2020) proposed that one of the important reasons for the difficulty of financing SMEs is due to the information asymmetry between banks and enterprises, while big data can solve the problem of information asymmetry, suggesting that SMEs can use big data technology to enhance their financing ability. Most of the existing papers analyze the application of information technology in supply chain financing from a theoretical perspective, which lacks persuasive power. There is little research in the literature that empirically examines the impact of information technology use on the supply chain financing performance of SMEs.

Some scholars have considered the impact of the characteristics of the organizational level of the firm on the performance of supply chain financing. Numerous scholars have studied the impact of firms' capabilities on supply chain financing performance, including individual entrepreneurs' capabilities, firms' core competencies, and firms' supply chain capabilities. Song et al. (2017b) research found that a firm's distribution operations capability and demand management capability have a significant positive effect on financing performance. These characteristics of a company's capabilities are not easily perceived by the financing provider during the SME financing process and are not as intuitive as reviewing the "hard information" of the company. Scholars have concluded that there is a significant correlation between firm size and the number of years in business

and financing performance. For example, the study by Petersen and Rajan (1997) concluded that the longer the enterprise is established, the closer the connection between the enterprise and upstream and downstream enterprises in the industry chain, the easier it is to form a cooperative relationship of interest, and the easier it is to get a loan. Zhao (2008), on the other hand, from the perspective of firm size, suggested that the larger the company is, the more collateral it has for its assets when facing business risks; further, its solvency will be stronger than that of smaller companies, and the core companies will have fewer concerns when making credit guarantees for them. As a result, larger companies are more reliable to financial institutions in terms of business credit compared to smaller companies.

Some scholars have also included firm-environment-level characteristics in their analysis of supply chain financing performance. In this paper, we focus on SMEs in the supply chain, and refer to their interactions with other firms in the supply chain as the internal supply chain environment and to those outside the entire supply chain as the external environment. From the perspective of supply chain networks, some scholars study the impact of factors such as the health of supply chain networks, the embeddedness of supply chain networks, and the strong and weak connections between supply chain networks on financing performance. Chen and Song (2020) argue that SMEs joining a healthy business network helps SMEs to improve their dual competence, which in turn improves their position in the network and ultimately contributes to their financing performance. Regarding the influence of the external environment on financing performance, scholars mainly consider factors such as external financing policies as well as the external financing environment. Li and Wang (2017), on the other hand, found that the external environment did not have a significant effect on access to Internet supply chain financing for micro and small enterprises.

Construction of a supply chain financing performance model for SMEs

Based on the Technology–Organization–Environment (TOE) analysis framework, this paper constructs a theoretical model framework, which analyzes the supply chain financing performance of SMEs by combining the external environmental characteristics of SMEs with the internal characteristics of firms, as shown in Figure 1.

Technical level

With the "Internet +" background, a new generation of information technology continues to emerge, such as blockchain technology, big data technology, and the Internet of Things (IoT). The use of information technology not only enhances the management of SMEs themselves and improves their efficiency in production, but also plays a vital role in reducing information asymmetry. This increases the level of trust of core enterprises, making them more willing to provide credit guarantees for SMEs, which ultimately improves the availability of supply chain financing for SMEs.

The model specifically includes the degree of information technology application and enterprise information sharing capabilities, which are two secondary conditions. In supply chain management, information technology is a general term for all the technologies used in management as well as information processing (Zhou and Wang, 2016). In business management, the application of information technology usually takes the form of various information systems, such as MRP, ERP, Logistics Management Systems, etc. The resource-based view considers that a firm's competitive advantage comes from the use of the resources it has. If companies are able to transform the IT resources they have into the ability to use IT, it will give them a long-term competitive advantage (Bharadwaj, 2000). In addition, the application of information technology to all processes of enterprise products can lead to the effective management of enterprise information flow, logistics, capital flow, and business flow, improving the productivity of enterprises and reducing their operating costs, thus helping them to achieve better business performance. Higher performance will reduce the default risk of companies and can help them obtain supply chain financing at a lower cost.

In addition, the degree of IT application is an important basis for enterprises to obtain information-sharing capabilities. From the perspective of information asymmetry, many scholars believe that the difficulty of financing SMEs is due to the information asymmetry between banks and enterprises. In traditional lending theory, banks and other financial institutions mostly rely on "hard information" such as financial statements uploaded by enterprises when reviewing loans, which makes it difficult to truly understand the business situation of enterprises. The consideration of avoiding a moral hazard leads to banks and other financial institutions being afraid or unwilling to lend. The application of information technology by enterprises can break the information blockage between upstream and downstream enterprises in the supply chain, enabling enterprises to share information across organizational boundaries, while also improving the efficiency of data analysis and processing. Information sharing has a significant direct positive impact on enterprise performance (Yang et al., 2022a). It significantly reduces the degree of information asymmetry between supply chain companies, and improves the information-sharing ability of companies (Kathuria et al., 2018). When external fund providers know more about an enterprise, they are also more willing to provide funds for its development, which in turn increases the possibility of SMEs obtaining supply chain financing.

Organization level

Enterprise qualification

This level mainly considers enterprise qualifications and enterprise supply chain capability. Enterprise qualifications specifically include the number of employees, the number of years in business, and the annual sales of the enterprise. Akben-Selcuk

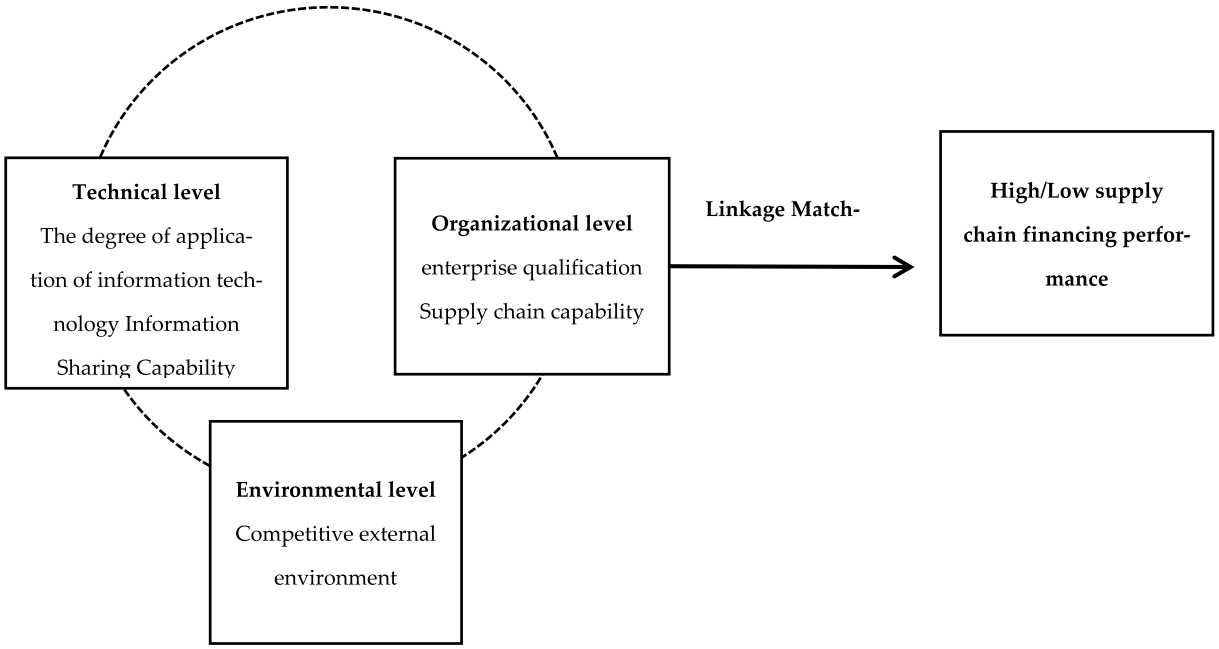


FIGURE 1
Research model.

(Selcuk and Altiok-Yilmaz, 2017) pointed out that it is often the larger firms that have more competitive and more profitable opportunities. According to the theory of enterprise capability, enterprise capability can contribute to the business performance of an enterprise. Similarly, for supply chain financing for SMEs, companies with stronger market competitiveness tend to have higher operating performance, and higher operating performance indicates stronger debt servicing ability. This can bring a higher level of credit to the enterprise, thus reducing the high costs caused by the lack of credit in its financing process and improving the availability of financing to a certain extent. On the other hand, the larger the scale of operation, the longer the operating life, and the larger the annual sales, the more stable the production, supply, and sales capacity of the enterprise, which can support the business activities of other companies in the supply chain. In turn, the company's position in the supply chain network is enhanced, and it is more easily recognized by the core enterprises and incorporated into their own business network, thus helping them to obtain funds through the use of supply chain financing.

Enterprise supply chain capability

Supply chain finance is a financing activity that relies on the supply chain management context, in which the traditional financial attributes are diluted (Caniato et al., 2016; Gelsomino et al., 2016). The most important feature of supply chain financing that is different from traditional enterprise financing is that when banks examine the financing needs of enterprises and decide whether to give them credit support, they do not consider the financial information and other "hard information" of enterprises as the basis for lending, but focus on the capabilities and development prospects of enterprises. In supply chain financing, corporate capacity has become a more important indicator than assets, and it is believed that, the stronger the capacity, the better the development prospects of the company, which tends to also have better solvency. According to signaling theory, SMEs should take the initiative to send signals that reflect the strengths of their enterprises to banks and other financial institutions when financing the supply chain, so that quality SMEs can be identified by the fund providers. Further, the supply chain capacity can reflect the business quality and management level of the companies in the supply chain itself, which is the premise that the funds can be repaid in time (Xiong et al., 2009; Dong et al., 2013). This originates from a series of practices and complex interactions in the supply chain practices of companies (Liu et al., 2016).

The measurement of the supply chain capability of an enterprise mainly considers the distribution operation capability and demand management capability of the enterprise. Some scholars have conducted empirical studies on the relationship between firms' supply chain capability and financing performance. Song et al. (2017b) found that distribution operational capability and demand management capability have a significant positive effect on financing performance through an empirical study of 150 SMEs' supply chain finance activities. Distribution operation capability mainly refers to the enterprise's ability to coordinate and optimize all aspects of product production, processing, and distribution in accordance with the process of supply chain operation, to achieve control and management of the entire business chain, to reduce distribution operation costs, and to achieve basic profitability goals. For banks, a company's ability to operate in distribution sends a signal that it has stable operating

income. Demand management capability mainly refers to the ability of enterprises to respond to customers' needs. With the improvement of people's living standards, the demand for goods also shows diversified characteristics. How companies can meet customers' needs for special goods and customization, and achieve their own profitability goals by creating customer value, are the issues to be considered for companies to gain competitive advantage in the market. In general, companies with demand management capabilities are able to maintain close relationships with their customers and provide the necessary guarantees for long-term profitability. For banks, a firm's demand management capabilities convey information about higher solvency.

Environmental level

The environmental dimension mainly considers the competitiveness of the external environment of the supply chain network. Although environmental-level antecedent conditions do not directly affect corporate supply chain financing performance, in supply chain finance, as it is an activity nested in the supply chain, the influence of the external environment on the supply chain inevitably affects the behavior of the companies in the supply chain. Environmental competitiveness mainly reflects the number of competitors in the external environment as well as the number of competitive fields (Hill and Matusik, 1998).

Previous studies have shown that environmental contestability can significantly contribute to the degree of IT uptake by firms (Gong and Ding, 2014). When the external environment of an enterprise is more competitive, the survival pressure on the enterprise will be greater. In order to obtain a stable market share and maintain its own competitive advantage, the enterprise will be forced to make technological changes to adapt to the strong competitiveness of the external environment. In other words, in a competitive environment, companies will tend to use information technology to improve their information-sharing capabilities and thus avoid problems that hinder the performance of supply chain financing, such as operational instability.

From the perspective of the funding provider, when the external environment becomes more competitive, the funding provider needs to be more aware of the true operational information of the firm and reduce the opportunistic behavior of the firm (Lima et al., 2020). Further, the higher the degree of information technology application, the better the information-sharing ability and the higher the information transparency. This will enable the supply chain capability signal to be more effectively communicated and identified by the capital provider, thus reducing the risk level of the capital provider in providing supply chain financing. This enables companies to obtain supply chain financing with lower financing costs and higher financing efficiency, thus improving their supply chain financing performance.

Research review

This paper considers that among the available studies on the performance of corporate supply chain financing, there are more in-depth studies and richer conclusions on the connotations, models, performance measurement methods, and influencing factors of supply chain financing. These research results are important guidance and inspiration for this paper, and enrich its theoretical foundation. After combing through the historical literature, the following shortcomings were found in previous studies, which are summarized below.

Variable selection

Most scholars who have studied the performance of supply chain financing in the past have considered the influence of the internal factors of enterprises (Lu et al., 2019b; Sun et al., 2021; Zhang, 2021; Yang et al., 2022b), and fewer have considered the influence of the characteristics of the external environments of enterprises on the performance of supply chain financing, and the influence of the combination of the internal factors and external environments of enterprises on the performance of supply chain financing. If studies are detached from the external environment, and do not take into account the influence of the real situation of the competitive nature of the external environment of the enterprise, such studies do not provide practical advice to SMEs in the supply chain on how to access supply chain financing. Paying attention to and studying the external environmental factors of the enterprise will help the enterprise to identify the risks from the market, improve the application of information technology, better form its core competitiveness, and finally obtain supply chain financing solutions to solve their own capital shortage problems.

Research perspective and research methodology

Previous studies have mostly been based on a power-variance perspective, constructing theoretical models based on specific research contexts and formulating causal hypotheses (Petersen and Rajan, 1997; Zhao, 2008; Song et al., 2017b; Chen and Song, 2020). To explore the simple linear relationship of an individual factor or moderating variable on a firm's supply chain financing performance, the study of the synergistic matching effect among multiple factors behind the variation in SME supply chain financing performance is neglected. In addition, there are limitations inherent in traditional research methods, such as regression analysis. The current study could only consider the effects of two or three variables, by setting moderating variables to analyze the interaction term. To observe changes in the direction or the extent of the main effect in various conditions, and less frequently, the joint effect of the simultaneous presence of multiple factors on supply chain financing performance is considered. In the real situation, the difference in supply chain financing performance is the result of multifactor coupling, and research should be closer to the real operating environments of enterprises.

In summary, since the influence of different factors on the performance of corporate supply chain financing is not independent, they will affect the performance of corporate supply chain financing by linkage matching to generate different

combinations among them. Therefore, a study with the "grouping perspective" can help deepen the understanding of the complex mechanisms behind the performance of corporate supply chain financing (Tan et al., 2019). Based on the TOE framework, this study selects five antecedent conditions at the "technology–organization–environment" level, and then conducts a comprehensive analysis of the factors affecting the performance of supply chain financing of SMEs. Using fuzzy-set qualitative comparative analysis (fsQCA) to explore the conditional groupings and mechanisms of action that contribute to differences in SME supply chain financing performance, the driving paths that generate high supply chain financing performance is derived. This helps to establish systematic analysis ideas for the development path of improving the supply chain financing performance of SMEs, and enriches the research on supply chain financing performance.

Research design and data collection

Variable design

1. Drawing on the scales used by Bruque Camara et al. (2015) and Subramani (2004), the research design emphasized the use of information technology by enterprises for resource acquisition and analysis and in support of business processes. The extent to which information technology was used within enterprises to support business processes, the extent to which enterprises used information technology to share resources with partners in the supply chain network, and the extent to which enterprises used information technology to access organizational resources outside the supply chain network were all analyzed. Three dimensions measured the degree of IT adoption in SMEs, including four specific measurement questions.
2. Enterprise information sharing capability refers to the ease and extent to which financing companies can obtain information about SME operations from the supply chain network. Its measurement was mainly based on the studies of Calanni et al. (2015) and Formentini and Taticchi (2014), among others, and consisted of three specific measurement questions.
3. This paper argues that the level of enterprise qualification can affect the repayment ability and willingness of enterprises, which in turn can affect the willingness of core enterprises in the supply chain to provide credit guarantees for them. The measurement of enterprise qualification was mainly based on four aspects: enterprise size (number of employees), years of operation, annual sales, and total assets of the enterprise.
4. For the measurement of enterprise supply chain capabilities, two indicators were selected, namely, enterprise distribution operation capability and customer demand management capability. Drawing mainly on the study by Zhang et al. (2010), the distribution operation capability was measured in terms of the level of management of raw materials by the company's internal transportation system, the timeliness of goods delivery, the speed and accuracy of order picking, etc. Specifically, there were six measurement questions. Drawing primarily from Schary and Coakley (1991), the company's ability to respond to customer needs in terms of product post-production and for customer relationship management, etc., was considered. Four measurement questions were specifically included to measure the customer demand management capability of the company.
5. The measurement of environmental competitiveness mainly reflects the intensity of competition, price competition, the situation of competitors, etc. The measurement scale was mainly based on the research results of Bengtsson and Sölvell (2004) and Claro et al. (2003). Three questions were included, such as "The price competition in our market is fierce."
6. The measurement of supply chain financing performance mainly draws on the research results of Song and Lu (2017). It was measured through four aspects: financing amount, financing cost, financing cycle, and financing availability, and included four specific topics.

Data collection

In this paper, data were collected using a seven-point Likert scale. Respondents rated the questions on a scale of 1–7 (from "strongly disagree" to "strongly agree") according to the actual situation of the company. This paper focused on exploring what antecedent conditions affect the supply chain financing performance of SMEs. The main research target was to meet two conditions: first, to meet the "small and medium-sized enterprises classified as standard regulations" of small and medium-sized enterprises, mainly with reference to the provisions of the research enterprises in the industry, the size of enterprises, business assets, etc., to screen; second, the company was required to have experience in supply chain financing in the past year, including but not limited to warehouse receipt pledge financing, accounts receivable/payable pledge financing, inventory pledge financing, etc. The data collection for this time period mainly used online questionnaires. The online questionnaire was mainly distributed through the Credamo platform, and the questionnaire included two parts: first, the basic information of the enterprise, including the name of the enterprise, whether the enterprise has had supply chain financing experience in the past year, the scale of the enterprise, and other information; the second was a specific question item for variable measurement. A total of 250 questionnaires were distributed online. A total of 171 were returned, and 90 valid questionnaires were finally obtained

through screening, with an efficiency rate of 52.6%. Some characteristics of the sample are listed in the Table 1 (The paper is based on a Chinese context and the data collected are based on companies in China).

Research method and data pre-processing

Reasons for choosing the fsQCA

In this paper, we mainly applied the fsQCA method based on set theory. QCA was developed in the late 1980s by Charles C. Ragin in 1987, and takes a holistic view of comparative case-level analysis, with each case considered as a "grouping" of conditional variables (Rihoux and Ragin, 2009). The aim is to identify complex causal relationships between conditional groupings and outcomes through comparisons between cases. Based on technical tools such as ensemble and Boolean algebra, it aims to blend the advantages of qualitative and quantitative research methods. An attempt was made to analyze the multifaceted and complex relationships behind the supply chain financing performance of SMEs based on a histological perspective. This was mainly due to the following considerations: First, to derive the path to improve the performance of supply chain financing for SMEs, which requires exploring the degree of information technology application, enterprise information sharing capability, enterprise qualification, supply chain capability, environmental competitiveness, and other factors together, the method can break through the limitations of traditional regression analysis methods for overlapping effects among variables, and provide a systematic and comprehensive view to deal with the conformational problem of multiple interrelated factors acting on the results simultaneously; Second, fsQCA analysis suggests that the interdependence and different combinations of causal conditions can constitute multiple concurrent causal relationships, which contributes to a deeper understanding of the differential driving mechanisms of SME supply chain financing performance; Third, the qualitative comparative analysis of fuzzy sets is closer to reality than clear sets and multi-valued sets in terms of calibration, because it requires the data handled by continuous values to be between 0 and 1. In summary, this paper argues that the fsQCA approach is more suitable for exploring the mechanisms of the roles of many factors in the supply chain financing performance of SMEs from a holistic perspective.

Reliability testing

Before conducting the fsQCA analysis, in order to ensure the high reliability and accuracy of the collected data, it was necessary to test the reliability and validity of the data, mainly with the help of software R for each antecedent variable. The supply chain capability at the organization level was mainly measured by the two indicators, the distribution operation capability and demand management capability, so the reliability of the two indicators is reported here. The results are shown in Table 2. The Cronbach's α for the six antecedent variables and one outcome variable were all greater than 0.7, indicating that the question items had high reliability. The combination reliability (CR) values of the combined reliability of all variables were greater than 0.7, indicating the high internal consistency of the model. The average extracted variance (AVE) of each variable was higher than 0.5, indicating that the data scale had good convergent validity.

TABLE 1 Structure of the sample distribution (N =90).

Variable symbols	Variable meaning	Indicators	Number of frequency	Frequency (%)
EQ1	Enterprise size	Less than 100 people	18	16.9%
	(Number of employees)	101–500 people	38	35.8%
		501–2,000 people	28	26.4%
		More than 2,000 people	6	5.6%
EQ2	Years of business operation	Less than 1 year	2	1.8%
		1–5 years	18	16.9%
		5–10 years	25	23.5%
		More than 10 years	45	42.4%
EQ3	Annual corporate sales	Under 1 million	6	5.6%
		1–10 million	28	26.4%
		10–50 million	24	22.6%
		50 million or more 5,000万以上	32	30.1%
EQ4	Total corporate assets	Under 1 million	8	7.5%
		1–10 million	14	13.2%
		10–50 million	22	20.7%
		50 million or more	46	43.3%

TABLE 2 Reliability test of each variable (*N*=90).

Variable symbols	Antecedent conditions	Cronbach's α	CR	AVE
IT	The Degree Of Application Of Information Technology	0.871	0.874	0.635
IS	Information Sharing Capability	0.869	0.871	0.692
EQ	Enterprise Qualification	0.874	0.886	0.668
DO	Distribution Operations Capability	0.926	0.927	0.680
RM	Demand Management Capability	0.844	0.846	0.580
EC	Environmental Competitiveness	0.874	0.876	0.639
SCP	Supply Chain Financing Performance	0.918	0.919	0.739

Variable calibration

Calibration means the process of assigning a set affiliation score to a case (Schneider and Wagemann, 2012). In the qualitative comparative analysis of fuzzy sets, its requirement for the data to be processed is a continuous set between 0 and 1, called the degree of membership. After that, three anchor points are selected based on theoretical and practical experience, and the conditions and results of each case are classified as fully affiliated, semi-affiliated, and fully unaffiliated according to the selected anchor points. The calibration process in this paper was as follows: first, the scores of the measured question items of each variable were summed and averaged, and the mean value was used as the score of each variable to avoid the influence of too large a value of one variable on the whole data; secondly, the key to calibration lay in the choice of anchor points, although the Likert scale already differentiates between the degree and level of specific conditions (variables) in the design phase (Zhang and Du, 2019). However, since most of the surveyed companies belonged to a large scale and the knowledge level of the questionnaire respondents varied, there may be a large subjective will in scoring. Therefore, this paper drew on existing research and applied the direct calibration method on the basis of existing theoretical and empirical knowledge. Taking into account, the actual situation and drawing on the study of Fiss (2011), the three anchor points of the five conditional and outcome variables were set to corresponding values of 25% (fully unaffiliated), 50% (intersection), and 75% (fully affiliated), and the final calibration results are shown in Table 3.

TABLE 3 Calibration values for each variable.

Variable	IT	IS	EQ	SC	EC	SCP
Fully affiliated	6.75	6.67	3.89	6.66	6.50	6.75
Intersection	6.00	5.67	3.00	5.80	5.25	5.50
Fully unaffiliated	4.11	4.00	1.36	4.05	3.50	3.50

Necessity analysis

Before the group analysis, the necessity of the results needed to be checked with respect to each antecedent condition. Necessity analysis refers to whether the appearance of a certain result is necessarily accompanied by the appearance of a certain condition. However, the occurrence of that condition does not necessarily lead to the occurrence of that result. In general, the minimum criterion for determining the necessary condition is greater than 0.9. As shown in Table 4, no single conditional variable had an effect on high/low supply chain financing performance that exceeded 0.9, that is, the results suggest that no single factor can constitute a necessary condition for high or low supply chain financing performance. This shows that the variation in the performance of firms' supply chain financing is the result of a combination of multiple factors.

Research findings

Construction of truth table

After the necessity analysis of each antecedent condition, the data needed to be imported into the fsQCA 3.0 software for the construction of the truth table. The truth table mainly reflects the distribution of different sets of antecedent condition combinations on the results. The truth table gives a clear idea of how many antecedent condition combinations produce high/low supply chain financing performance, as well as the logical residuals. The key to the construction of the truth table is the selection of the case frequency threshold and the determination of the original consistency. For the selection of the case frequency threshold, the suggestion of Ragin (Rihoux and Ragin, 2009) was referred to: when the sample size is small, set the frequency to 1; additionally, the frequency threshold should be selected to retain at least 75% of the total number of cases. In terms of the original consistency threshold setting, Ragin recommends a corresponding minimum threshold of 0.8. In summary, in this paper, the case frequency was set to 1, the original consistency threshold was set at 0.8, and the PRI consistency was set to 0.7. Different groups of high supply chain financing performance and non-high supply chain financing performance were generated in the form of 0 and 1.

Configuration analysis

Unlike the analysis of necessary conditions, group analysis attempts to reveal the sufficiency of different groupings consisting

of multiple conditions to cause the generation of results (Tao et al., 2021). When the truth table is constructed, the software performs operations and then generates three solutions (complex, intermediate, and parsimonious). Since the intermediate solution is produced using an easy counterfactual analysis that includes core and edge conditions for the grouping, the results are more likely to reflect the actual results. In addition, given that existing studies generally report intermediate solutions, this paper mainly reports the intermediate solution. The results of the standardized analysis of high/low supply chain financing performance are shown in Table 5. It can be seen that there were three antecedent condition groups that generated high supply chain financing performance, while there was only one antecedent condition group that generated low supply chain financing performance. Among them, the consistency values of the high supply chain financing performance group (Z1, Z2, and Z3) were 0.931, 0.969, and 0.963, and the consistency of their overall solutions was 0.928. It is shown that these three conditions were sufficient to generate high supply chain financing performance. The overall solution coverage was 0.779, indicating that these three conditional groupings explained 77.9% of the cases. The consistency value was 0.927 for the low supply chain financing performance grouping (Z4) and 0.927 for the overall solution. The overall solution coverage was 0.543, indicating that the article grouping was able to explain 54.3% of the cases. The results of these antecedent condition groupings suggest that the degree of information technology application, information sharing capability, firm qualification, and supply chain capability, and the presence or absence of a competitive external environment can lead to differences in the supply chain financing performance of SMEs.

TABLE 4 Results of necessity tests for individual conditions.

Conditional variables	Result Variables	
	High supply chain financing performance	Low supply chain financing performance
IT	0.826	0.567
~IT	0.484	0.826
IS	0.819	0.617
~IS	0.496	0.782
EQ	0.709	0.629
~EQ	0.530	0.674
SC	0.868	0.629
~SC	0.496	0.832
EC	0.713	0.637
~EC	0.571	0.722

Path analysis

The conditional grouping of high supply chain financing performance

From Table 5, it can be seen that the three grouping paths leading to high supply chain financing performance were different and can be considered as a sufficient condition grouping for high supply chain financing performance.

Configuration Z1: the degree of application of information technology * information sharing capability * supply chain capability. The path suggests that the presence of these three conditions plays a central role for SMEs to generate high supply chain financing performance that can break through the environmental conditions. Regarding the research topic of supply chain financing, companies with high supply chain financing performance invest more resources and incorporate technology level activities as a way to improve their information technology and keep up with the developments of the times. The information-sharing ability of enterprises can better solve the problem of information asymmetry among supply chain enterprises through a high degree of information technology application, thus making

TABLE 5 Group analysis of high and low supply chain financing performance.

Variable category		A grouping that generates high supply chain financing performance			Generating low supply chain finance performance groups
		Z1	Z2	Z3	Z4
Technical level	**IT**	●	●		⊗
	IS	●		●	⊗
Organizational level	**EQ**		●	●	
	SC	●	●	●	⊗
Environmental level	**EC**		⊗	⊗	⊗
Consistency		0.931	0.969	0.963	0.927
Unique coverage		0.334	0.036	0.034	0.543
Original coverage		0.708	0.410	0.409	0.543
Result coverage			0.779		0.543
Results Consistency			0.928		0.927

Referring to the formulation of Ragin (Claro et al., 2003), ● indicates that the variable exists, ⊗ indicates that the variable does not exist among them. Large circles denote core conditions, small circles denote edge conditions, and a blank space indicates that the presence or absence of the variable is irrelevant.

the core enterprises in the supply chain more aware of the operation of enterprises. In addition, the existence of supply chain capabilities indicates that the company has a greater advantage in distribution operations and the management of customer demand, can signal to the outside world that the company has stable operating profits, and can increase the willingness of core firms to provide supply chain financing solutions for them, which in turn helps SMEs to obtain higher supply chain financing performance. For example, Tongchuang Mastery Technology Co. Without regard to corporate qualifications and competitive pressures in the external environment, the company has expanded inward and focused on the development of its own technical capabilities, with integrity being its core business. Further, the company is deeply committed to the improvement of corporate supply chain capabilities in the development process, focusing on the growth and development of corporate partners, which ultimately manifests itself in high supply chain financing performance. As this driving path is composed of the degree of IT adoption, information sharing capabilities (technology), and supply chain capabilities (organization), in this paper, we named it "technology-supply chain capability ". The consistency of this grouping was 0.931, the unique coverage was 0.334, and the original coverage was 0.708. This path can explain about 70.8% of the corporate supply chain financing cases. In addition, about 33.4% of corporate supply chain financing cases can be explained by this path only.

Configuration Z2: The degree of application of information technology * enterprise qualification * supply chain capability *~environmental competitiveness. This shows that SMEs have certain strengths when they have a higher degree of IT application and a stronger supply chain capability, and when faced with weaker competition from the external environment, firms are still able to achieve higher supply chain financing performance regardless of whether they have high or low information-sharing capabilities. Guangdong Wanfang Construction Co. can be taken as an example: although its size is small and it is under certain competitive pressures from the external environment, it still presents a high supply chain financing performance, because of its strong information technology utilization and supply chain capabilities, which stand out from the crowd. The consistency of this grouping was 0.969, the unique coverage was 0.036, and the original coverage was 0.410. This path explained about 41% of corporate supply chain financing cases, but only 3.6% of corporate supply chain financing cases can be explained by this path only.

Configuration Z3: information sharing capability * corporate qualification * supply chain capability *~environmental competitiveness. Among them, the presence of two antecedent conditions, information sharing capability, and supply chain capability, plays a central role in generating high supply chain financing performance for firms, and the presence of firm qualifications plays a supporting role. Liaoning Photoelectric Technology Co. can be taken as an example of this. Despite the company's shortcomings in the use of information technology, the company has been able to communicate and cooperate frequently with its partners, increasing mutual understanding and trust, and has been able to focus on improving its own information-sharing ability and supply chain capability to increase the willingness of core companies to guarantee them, ultimately helping them obtain a high supply chain financing performance. The consistency of this grouping was 0.963, the unique coverage was 0.034, and the original coverage was 0.409. This path explained about 40.9% of corporate supply chain financing cases, but only 3.4% of the sample cases can be explained by this path only.

By comparing the above high supply chain financing performance condition groupings (Z1, Z2, and Z3), it was found that supply chain capability appeared in each path as a core condition, indicating that strong supply chain capability is a strong driver for SMEs to generate high supply chain financing performance. Secondly, by comparing Z2 with Z3, it was found that there is a substitution between IT application and information sharing capability when SMEs have strong supply chain capability. In other words, when SMEs have distribution operation capability and customer demand management capability, they can present high supply chain financing performance when they reach a high level of information technology application or meet the information-sharing capability requirement. Finally, by comparing Z1 with Z2 and Z1 with Z3, it was found that, when SMEs face weaker competitive pressures and have a better survival environment, SMEs with certain strengths will obtain high supply chain financing performance after mastering supply chain capabilities, provided that one of the technical-level conditions is met.

Conditional grouping of low supply chain financing performance

Configuration Z4:~Information technology application degree*~Information sharing capability*~Supply chain capability*~Environmental competitiveness. When firms face weaker competition in the external environment, a low level of utilization of information technology, as well as a lower information sharing capability and lower supply chain capability, will inhibit the performance of supply chain financing for SMEs, regardless of the strength of the company. For example, Suzhou Centennial Legend Food Co. This company belongs to a traditional food company, which does not pay enough attention to the application of information technology and the development of corporate supply chain capabilities, and it is difficult for this company to be identified by core companies, thus generating a lower supply chain financing performance. Among them, the degree of information technology utilization and information sharing capability belongs to the technical level, the supply chain capability belongs to the organizational level factor, and the competitiveness of the external environment belongs to the environmental level factor. This triple dimension has a synergistic inhibitory effect on the supply chain financing performance of SMEs. This path explained 54.3% of the low supply chain financing performance sample among SMEs, and also 54.3% was explained by this path.

In summary, supply chain capability appeared as a core variable in Z1, Z2, and Z3, which had high supply chain financing

performance, while a lack of supply chain capability was missing as a core variable in the Z4 group state, which had low supply chain financing performance. Second, both IT application and information sharing capability appeared as core variables in Z2 and Z3, respectively, while both were missing as core variables in Z4. Therefore, it can be determined that the application of corporate information technology, information sharing capability, and supply chain capability has a significant positive effect on the ability of SMEs to generate high supply chain financing performance.

Robustness tests

In the fsQCA method, the robustness test is a crucial part to check whether the condition set is stable and can guide the enterprise practice. This paper focused on robustness testing by adjusting the original consistency and PRI consistency, as well as the case frequency thresholds. First, the original consistency threshold of 0.80 was increased to 0.85, using a test with a variable consistency threshold. Second, the PRI consistency was increased from 0.7 to 0.75. Finally, the case frequency threshold was adjusted by increasing the original threshold from 1 to 2. The three methods produced consistent groupings, it can be concluded that the findings of this paper were robust.

Discussion

Theoretical contribution

First, the research enriches the study of factors influencing the performance of corporate supply chain financing, as most of the existing studies are based on a single variable or are bivariate studies on the impact of supply chain financing performance. Then, when verifying the correlations between variables, the research perspectives are generally more homogeneous. A deeper analysis of the reasons affecting the supply chain financing performance of SMEs is needed, as well as enriching the scope of supply chain financing performance research. This paper attempted to analyze how the three dimensions of technology, organization, and environment interact with each other in SME supply chain financing performance based on the TOE research framework, as well as how they work together in SME supply chain financing performance.

Second, this research was process-oriented in order to increase the practicality of performance research. The research formed the path of high/low supply chain financing for SMEs, which helps to enhance the practicality of supply chain financing performance, and is more instructive for companies. Is the path to forming a firm's high supply chain financing performance is unique? and can SMEs choose a high supply chain financing performance path that suits their own development according to their actual situation? Further, what are the differences between the paths that lead to high supply chain financing performance and those that lead to low supply chain financing performance? Do the antecedent conditions in the pathway have a substitution relationship with each other? Studying the factors influencing the performance of low supply chain financing is more relevant for companies.

Practical enlightenment

First, SMEs should pay high attention to the issue of variability and diversity in the grouping of supply chain financing performance. There is a linkage matching effect between the various influencing factors. Therefore, in order to obtain supply chain financing provided by core enterprises and solve their own financing constraints, enterprises should pay attention to the synergistic development of multiple links of enterprises. Since SMEs in the supply chain are in different situations and face different external competitive environments, the causes of low supply chain financing performance should be determined according to the actual situation and the right remedy should be prescribed so as to effectively improve supply chain financing performance.

Secondly, SMEs should attach great importance to the development of corporate capabilities. In the supply chain financing model, the various capabilities of SMEs influence the willingness of core enterprises to provide credit guarantees, which in turn affects whether SMEs can obtain credit support from financial institutions. Therefore, the capacity of SMEs is particularly important and critical. Furthermore, the analysis of the above grouping results shows that the supply chain capability of firms plays a key role in improving the supply chain financing performance of SMEs. SMEs should pay attention to the development of their own supply chain capabilities so that they can be recognized by core enterprises or financial institutions and increase the availability of supply chain financing, thus effectively improving supply chain financing performance.

Conclusion and outlook

Conclusion

In this paper, based on the TOE framework model, a sample of 90 SMEs involved in the supply chain financing business was studied. Using the fsQCA method, the grouping paths that generate high versus low supply chain financing performance of SMEs were analyzed to explore the grouping paths of conditional variables affecting supply chain financing performance at three levels. Three grouping paths with high supply chain financing performance, and one with low supply chain financing performance were generated. The main findings are as follows.

First, through a single-condition necessity analysis, it was found that none of the five factors in the three dimensions of

"technology + organization + environment" is necessary to produce high supply chain financing performance. Among them, there are three different models of high supply chain financing performance: "technology-supply chain capability-driven," "IT-supply chain capability-driven," and "IS-supply chain capability-driven." These three models are driving the performance of SME supply chain financing in "different routes to the same destination."

Second, the two antecedent conditions at the technical level are interchangeable. This implies that the application of corporate information technology and the ability to share information can be interchangeable in a given situation, producing an equivalent effect of effectively improving the supply chain financing performance of SMEs. In addition, the application of corporate information technology and the ability to share information and supply chain capabilities play a critical role in facilitating the performance of supply chain financing for SMEs.

Research limitations and outlook

The pathway of the role of SME supply chain financing performance derived in this paper has the following limitations, which need to be further explored in subsequent studies: (1) Based on the existing research results, this paper constructed a TOE analysis model with three levels and five conditional variables, without considering the influence of other factors on the performance of supply chain financing. However, the possibility of the existence of other factors or groupings that have a role in the performance of supply chain financing cannot be excluded. (2) The cases studied in this paper were 90 SMEs identified after screening, and further research is needed to determine whether the findings are generalizable to all SMEs. (3) The data of this paper were obtained from the questionnaire survey. Due to the uneven quality of the respondents, the understanding of the questionnaire items was influenced by subjective will, which to a certain extent affects the quality of the conclusions obtained in this paper.

With the continued decay of the global economy, the heavy pressure on SMEs. The procurement phase, the rise in raw material prices, further led to the increase in production costs of SMEs. Production stage, along with the continued impact of the epidemic, some areas are still in the stage of shutdown, enterprise plant rent, equipment depreciation, personnel wages, continued to compress the survival of small and medium-sized enterprises. During the sales phase, the domestic market is shrinking, people are afraid to consume, and enterprises cannot collect funds, leading to serious cash flow break problems for many enterprises. At this time, enterprises should establish the concept of sustainable development, accelerate the digital transformation of enterprises, shrink production lines, specialize in the strengths of the enterprise products, digital production, improve production capacity, retain cash flow, and live as the primary strategy of the enterprise. Future research on the topic of supply chain finance could focus on whether the efforts made by SMEs on sustainability will increase the willingness of core firms in the supply chain to guarantee them and thus improve the supply chain financing performance of SMEs. And how supply chain finance can empower the digital transformation of enterprises.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

HH: methodology, software, investigation, data curation, writing—original draft preparation, and visualization. HH and YZ: validation. YZ: formal analysis and project administration. WD and YZ: resources and writing—review and editing. WD: supervision. All authors contributed to the article and approved the submitted version.

Funding

This study was supported by the National Social Science Foundation of China "Influencing Factors of Liability of Foreignness in the Internationalization of Chinese Firms" (No. 19BGL024).

Acknowledgments

The authors thank the editor and reviewers for their numerous constructive comments and encouragement that have improved our paper greatly.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Bengtsson, M., and Sölvell, Ö. (2004). Climate of competition, clusters and innovative performance. *Scand. J. Manag.* 20, 225–244. doi: 10.1016/j.scaman.2004.06.003

Bharadwaj, A. S. (2000). A resource-based perspective on information technology capability and firm performance: an empirical investigation. *MIS Q.* 24, 169–196. doi: 10.2307/3250983

Bruque Camara, S. J., Moyano-Fuentes, J. M., and Marín, M. (2015). Cloud computing, web 2.0, and operational performance. *Int. J. Logist. Manag.* 26, 426–458. doi: 10.1108/IJLM-07-2013-0085

Calanni, J. C., Siddiki, S. N., Weible, C. M., and Leach, W. D. (2015). Explaining coordination in collaborative partnerships and clarifying the scope of the belief Homophily hypothesis. *J. Public Adm. Res. Theory* 25, 901–927. doi: 10.1093/jopart/mut080

Caniato, F., Gelsomino, L. M., Perego, A., and Ronchi, S. (2016). Does finance solve the supply chain financing problem? *Supply Chain Manag.* 21, 534–549. doi: 10.1108/SCM-11-2015-0436

Chen, S., and Song, H. (2020). The impact of firm networks and firm capabilities on SME financing performance from a supply chain finance perspective—a chain intermediation model. *J. Bus. Econ.* 4, 18–28. doi: 10.14134/j.cnki.cn33-1336/f.2020.04.002

Claro, D. P., Hagelaar, G., and Omta, O. (2003). The determinants of relational governance and performance: how to manage business relationships? *Ind. Mark. Manag.* 32, 703–716. doi: 10.1016/j.indmarman.2003.06.010

Didi, X. (2019). Study on the application of blockchain technology in supply chain finance. *Southwest Finance* 2, 74–82.

Dong, K., Qian, Z., ZhiX, Z., and Hao, Z. (2013). Case study of supply chain finance model and risk control. *J. Minzu Univ. Chin.* 22, 36–43.

Fiss, P. (2011). Building better causal theories: a fuzzy set approach to typologies in organization research. *Acad. Manag. J.* 54, 393–420. doi: 10.5465/amj.2011.60263120

Formentini, M., and Taticchi, P. (2014). Corporate sustainability approaches and governance mechanisms in sustainable supply chain management. *J. Clean. Prod.* 112, 1920–1933. doi: 10.1016/j.jclepro.2014.12.072

Gelsomino, L. M., Mangiaracina, R., Perego, A., and Tumino, A. (2016). Supply chain finance: a literature review. *Int. J. Phys. Distrib. Logist. Manag.* 46, 348–366. doi: 10.1108/IJPDLM-08-2014-0173

Gong, J., and Ding, K. (2014). A study of the relationship between technological advantage, environmental competitiveness, and organizational information density and information technology uptake in a local context. *Sci. Sci. Manag.* 35, 19–26.

Hill, C. W. L., and Matusik, S. F. (1998). The utilization of contingent work, knowledge creation, and competitive advantage. *Acad. Manag. Rev.* 23, 680–697. doi: 10.2307/259057

Hong, X., Yang, J., and Chen, S. (2022). Blockchain technology, network embeddedness and supply chain finance performance—a qualitative comparative analysis of fuzzy sets. *J. Dalian Univ. Technol.* 43, 13–23. doi: 10.19525/j.issn1008-407x.2022.02.003

Kathuria, A., Mann, A., Khuntia, J., Saldanha, T. J. V., and Kauffman, R. J. (2018). A strategic value appropriation path for cloud computing. *J. Manag. Inf. Syst.* 35, 740–775. doi: 10.1080/07421222.2018.1481635

Li, S. H., and Sun, Q. (2022). Mechanisms of core firm cooperation ability affecting supply chain financing performance. *Finance Theor. Pract.* 1, 49–55.

Li, H., and Wang, T. (2017). Study on the impact factors of supply chain financing for small and micro enterprises under internet finance. *Mod. Bus.* 4, 72–74. doi: 10.14097/j.cnki.5392/2017.04.033

Lima, P. F., Crema, D. M., and Verbano, C. (2020). Risk management in SMEs: a systematic literature review and future directions. *Eur. Manag. J.* 38, 78–94. doi: 10.1016/j.emj.2019.06.005

Liu, Y., Srai, J. S., and Evans, S. (2016). Environmental management: the role of supply chain capabilities in the auto sector. *Supply Chain Manag.* 21, 1–19. doi: 10.1108/SCM-01-2015-0026

Lu, Q., Liu, B., and Song, H. (2019a). The impact of SMEs' capabilities on supply chain financing performance: an information-based perspective. *Nankai Manag. Rev.* 22, 122–136.

Lu, Q., Liu, B., and Song, H. (2019b). The impact of SME's capability on its supply chain financing performance: a study based on information perspective. *Nankai Bus. Rev.* 22, 122–136.

Martin, J. (2017). Suppliers participation in supply chain finance practices: predictors and outcomes. *Int. J. Integrat. Supply Manag.* 11:193. doi: 10.1504/IJISM.2017.086242

Martin, J., and Hofmann, E. (2019). Towards a framework for supply chain finance for the supply side. *J. Purch. Supply Manag.* 25, 157–171. doi: 10.1016/j.pursup.2018.08.004

Pan, A., Ling, R., and Li, B. (2021). How does supply chain finance serve the real economy? Evidence from adjustment of capital structure. *Bus. Manag. J.* 43, 41–55.

Petersen, M. A., and Rajan, R. G. (1997). Trade credit: theories and evidence. *Rev. Financ. Stud.* 10, 661–691. doi: 10.1093/rfs/10.3.661

Qin, J. (2021). Current situation and countermeasures for the development of supply chain finance in China. *Acad. Exchang.* 5, 103–115.

Rihoux, B., and Ragin, C. C. (2009). "Configurational Comparative Methods: Qualitative Comparative Analysis (QCA) and Related Techniques" in *Applied Social Research Series*. Los Angeles, CA: SAGE Publications, Inc. p. 240.

Schary, P., and Coakley, J. (1991). Logistics organization and the information system. *Int. J. Logist. Manag.* 2, 22–29.

Schneider, C, and Wagemann, C. (2012). *Set-Theoretic Methods for the Social Sciences: A Guide to Qualitative Comparative Analysis*. Cambridge: Cambridge University Press. p. 367.

Selcuk, A., and Altiok-Yilmaz, E. A. (2017). "Determinants of Corporate Cash Holdings: Firm Level Evidence From Emerging Markets" in *Global Business Strategies in Crisis*. eds. Ü. Hacioğlu, H. Dinçer and N. Alayoğlu (Cham: Springer), 417–428.

Song, H., and Lu, Q. (2017). What kind of SMEs can benefit from supply chain finance?—network and capacity-based perspective. *Manag. World* 6, 104–121. doi: 10.19744/j.cnki.11-1235/f.2017.06.009

Song, H., Lu, Q., and Yu, K. (2017a). A comparative study on the impact of supply chain finance and bank lending on the financing performance of SMEs. *J. Manag.* 14, 897–907. doi: 10.3969/j.issn.1672-884x.2017.06.013

Song, H., Yang, X., and Yu, K. (2017b). How SMEs get financing performance under information asymmetry - an empirical analysis based on supply chain finance. *Chin. Circulat. Economy* 31, 89–99. doi: 10.14089/j.cnki.cn11-3664/f.2017.09.011

Subramani, M. (2004). How do suppliers benefit from information technology use in supply chain relationships? *MIS Q.* 28, 45–73. doi: 10.2307/25148624

Sun, C., Wang, H., and Wang, P. (2021). The impact of corporate core competencies on supply chain financing: financial support or appropriation? *China Soft Sci.* 6, 120–134. doi: 10.3969/j.issn.1002-9753.2021.06.012

Tan, H., Fan, Z., and Du, Y. (2019). Technology management capability, attention allocation, and local government website development—a TOE framework-based histogram analysis. *Manag. World* 35, 81–94. doi: 10.19744/j.cnki.11-1235/f.2019.0119

Tao, K., Zhang, S., and Zhao, Y. (2021). What determines government public health governance performance?--a study of linkage effects based on the QCA approach. *Manag. World* 37, 128–138.

Wandfluh, M., Hofmann, E., and Schoensleben, P. (2016). Financing buyer–supplier dyads: an empirical analysis on financial collaboration in the supply chain. *Int J Log Res Appl* 19, 200–217. doi: 10.1080/13675567.2015.1065803

Wei, B. Z., and Liu, K. (2012). Can supply chain finance development reduce SMEs' financing constraints?--an empirical analysis based on small and medium-sized listed companies. *Econom. Sci.* 3, 108–118.

Xia, T., and Jin, X. (2011). Analysis of the advantages of supply chain finance in solving the financing difficulties of SMEs. *Commer. Res.* 6, 128–133. doi: 10.13902/j.cnki.syyj.2011.06.023

Xin, X. (2007). Interpreting the dual model of supply chain financing. *Logist. Technol.* 7, 69–73.

Xing, Z. (2018). An introduction to the difficulties of financing small and medium-sized enterprises. China. *Dent. Econ.* 11, 285–287.

Xiong, X., Ma, J., Zhao, W. J., and Zhang, J. (2009). Credit risk evaluation under the supply chain finance model. *Nankai Bus. Rev.* 12, 92–98.

Yang, Y. L., Zheng, Y., Guojie, X., and Tian, Y. (2022a). The influence mechanism of strategic partnership on Enterprise performance: exploring the chain mediating role of information sharing and supply chain flexibility. *Sustain. For.* 14:4800. doi: 10.3390/su14084800

Yang, Y. L., Zheng, Y., Xie, G., and Tian, Y. (2022b). The influence mechanism of learning orientation on new venture performance: the chain-mediating effect of absorptive capacity and innovation capacity. *Front. Psychol.* 13:818844. doi: 10.3389/fpsyg.2022.818844

Yanni, X. (2020). Research on using big data to improve financing of small and medium Enterprises in Quanzhou. *J. Yan'an Vocation. Technic. College'* 34, 20–24. doi: 10.13775/j.cnki.cn61-1472/g4.2020.01.005

Yu, H., and Wang, S. (2022). Core enterprises' willingness to participate in supply chain finance and financing model orientation. *China Circulat. Econ.* 36, 22–34. doi: 10.14089/j.cnki.cn11-3664/f.2022.03.003

Zhang, M. (2021). Management incentives, innovation capability and supply chain financing performance. *Commun. Financ. Account.* 4, 54–57. doi: 10.16144/j.cnki.issn1002-8072.2021.04.010

Zhang, M., and Du, Y. (2019). Application of QCA methods in organization and management research: orientation, strategy and direction. *Chin. J. Manag.* 16, 1312–1323.

Zhang, Q., Vonderembse, M., and Lim, J. S. (2010). Value chain flexibility: a dichotomy of competence and capability. *Int. J. Prod. Res.* 40, 561–583. doi: 10.1080/00207540110091695

Zhao, Y. (2008). A study of factors influencing business credit: evidence from bank credit. *Financ. Theor. Pract.* 6, 38–42.

Zhao, X. (2020). Analysis of the problems and countermeasures in supply chain financing of SMEs. *Manag. Technol. SME* 9, 80–81.

Zhou, S., and Wang, G. (2016). The impact of information technology capabilities on supply chain performance: an information integration-based perspective. *J. Syst. Manag.* 25, 90–102.

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 22 November 2022
DOI 10.3389/fpsyg.2022.951779

OPEN ACCESS

EDITED BY
George Waddell,
Royal College of Music, United Kingdom

REVIEWED BY
Nuno Mateus,
Research Centre in Sports Sciences, Health Sciences and Human Development (CIDESD), Portugal
Maghsoud Nabilpour,
University of Mohaghegh Ardabili, Iran
Ángel Custodio Mingorance-Estrada,
University of Granada,
Spain

*CORRESPONDENCE
Shaoliang Zhang
zslinef@mail.tsinghua.edu.cn

SPECIALTY SECTION
This article was submitted to Performance Science, a section of the journal Frontiers in Psychology

RECEIVED 24 May 2022
ACCEPTED 04 November 2022
PUBLISHED 22 November 2022

CITATION
Lu P, Zhang S, Ding J, Wang X and Gomez MA (2022) Impact of COVID-19 lockdown on match performances in the National Basketball Association. *Front. Psychol.* 13:951779. doi: 10.3389/fpsyg.2022.951779

Impact of COVID-19 lockdown on match performances in the National Basketball Association

Peng Lu[1], Shaoliang Zhang[1]*, Jie Ding[1], Xing Wang[2] and Miguel Angel Gomez[2]

[1]Division of Sport Science and Physical Education, Tsinghua University, Beijing, China, [2]Facultad de Ciencias de la Actividad Física y del Deporte (INEF), Universidad Politécnica de Madrid, Madrid, Spain

This study aimed to compare differences in the match performances between home and away games during pre- and post-COVID-19 lockdown and to identify the key factors to match success with and without spectators. The sample consisted of 1,549 basketball matches including 971 games of the 2019–2020 regular season before the COVID-19 lockdown and 578 ghost matches of the 2020–2021 regular season after the COVID-19 pandemic. The independent *t*-test was used to explore the differences before and after COVID-19 while univariate and multivariable logistic regression models were used to identify the key factors to match success between matches with and without spectators. Our study identified that offensive rebounds were the only indicator differentiating between home and away games after the COVID-19 lockdown. Furthermore, home teams won more matches than away matches before the COVID-19 whereas home advantage had no impact on winning matches after the COVID-19. Our study suggested that crowd support may play a key role in winning games in the NBA. Furthermore, independently of the pre-and post-COVID19 pandemic, free throws made, three-point field goals made, defensive rebounds, assists, steals, personal fouls, and opponent quality were key factors differentiating between win and loss. Coaches and coaching staff can make informed decisions and well prepare for basketball match strategies.

KEYWORDS

team sports, basketball performance analysis, match-related statistics, National Basketball Association, COVID-19

Introduction

The consistently better performance seen by teams in various sporting contexts when playing at home is known as the "Home Advantage" (HA) that has a clear impact on winning basketball matches in the available research (Sampaio et al., 2006; Gómez and Pollard, 2011; García et al., 2014; Higgs, 2021; Mateus et al., 2021). Home teams have a better performance in terms of assists (García et al., 2014; Bustamante-sánchez et al., 2022), block shots (Sampaio et al., 2008; Garcia-Rubio et al., 2009; Bustamante-sánchez et al.,

2022), and personal fouls (Sampaio et al., 2008; Bustamante-sánchez et al., 2022) than away teams. In particular, defensive rebounds seem to be the most common performance indicator that was influenced by HA in the NBA (Zhang et al., 2019b). Although research into the impact of HA on match performances during basketball match-play has produced equivocal results, it is one of the most important contextual variables that should be taken into consideration in basketball science.

Factors that affected the phenomenon have been paid constant attention over the past years (Nevill and Holder, 1999; Neave and Wolfson, 2003; Pollard, 2008; Gómez and Pollard, 2011; Ribeiro et al., 2016; Goumas, 2017; Ponzo and Scoppa, 2018; Fischer and Haucap, 2021; Tilp and Thaller, 2020). Crowd support, territoriality, familiarity with the stadium, and travel fatigue (Ponzo and Scoppa, 2018) are believed to be the key factors of the HA phenomenon. Furthermore, the available research showed that crowd support and density might be the two most important factors that contributed to the HA (Inan, 2020). Ponzo and Scoppa (2018) identified that HA leads to the improvement of team performance and biased decisions of referees. For example, Fioravanti et al. (2021) found that HA has dropped by approximately 5% and the point difference in favor of home teams was reduced from approximately 6 to 4 points on average when a ghost game took place in rugby competitions. Similarly, a study in professional basketball demonstrated that HA affects the microscopic dynamics of the game by increasing the scoring rates and decreasing the time intervals between scores (Ribeiro et al., 2016). However, there is still no common consensus about the relative importance and interactive impact of different factors (Wunderlich et al., 2021).

The lockdowns due to the Covid-19 pandemic provided a unique opportunity to test a natural experiment in terms of team performances that could be analyzed during matches with and without the presence of an audience (McCarrick et al., 2021). Tilp and Thaller (2020) reported that the Covid-19 lockdown caused a decreased trend for HA in Germany's top football league, Bundesliga. Furthermore, this study pointed out that the ambiguity in previous studies' findings may result from different ways of proxying home support (e.g., occupancy rate or absolute attendance) or various degrees of control for covariates (Tilp and Thaller, 2020). Also, the other study examined the impact of crowd support on match performances in the three German men's professional football divisions, they found that there was a reduced HA in the first division in the ghost games, whereas no change was observed in the second and third divisions (Fischer and Haucap, 2021). Indeed, Arboix-Alió et al. (2022) reported that the effect of HA did not disappear despite playing without spectators but decreased from 63.99 to 57.41% while playing with spectators benefited local teams' performance, especially in the Portuguese and Italian Hockey leagues (Arboix-Alió et al., 2022). To our knowledge, four studies have examined changes in HA due to the COVID-19 epidemic in the NBA where the presence of crowds was associated with rebounds and points differential and accounted for a 15.91% increase in terms of winning percentage in comparison with the absence of crowds (Leota et al., 2021; Alonso et al., 2022; Bustamante-sánchez et al., 2022; Gong, 2022; Szabó, 2022). However, these studies fail to consider the impact of situational variables and the variability from season to season. To increase the statistical power of the analysis and the accuracy of the outcomes, our study uses a larger sample size controlling for situational variables to explore the changes in HA before and after the COVID-19 pandemic.

The drivers and mechanisms of HA remain equivocal, yet HA is a robust and reliable phenomenon (Leota et al., 2021). Given the significance of understanding which team performances may be more affected by crowd support in professional basketball, the purpose of this study was to compare differences in the match performances during pre- and post-COVID-19 lockdown and to identify the key factors to winning matches with and without spectators. We hypothesized that there might be a decreased trend after COVID-19 compared to the period before the COVID-19.

Materials and methods

Sample

Data were collected from the official NBA website (and www.basketball-reference.com). An observational case series study design was used to compare match performances before and after COVID-19. A total of 1,549 basketball matches included 971 games of the 2019–2020 NBA regular season before the COVID-19 lockdown and 578 ghost matches of the 2020–2021 NBA regular season after the COVID-19 pandemic. Additionally, the game-related statistics included two-and three-point field goals (both made and missed), free-throws (both made and missed), defensive and offensive rebounds, assists, blocks, fouls, steals, turnovers, and personal fouls. Based on the previous research (Sampaio et al., 2006; García et al., 2014; Zhang et al., 2017), a total of 13 variables were selected to quantify the technical performances.

Procedures

Furthermore, to control for the situational conditions during different matches, match location, opponent quality, and match type were considered in our study, and the detailed explanation is as follows:

- Match location: This was defined as the match being played at home or away (Gómez et al., 2010).
- Opponent quality: This was defined using the team's winning match percentage (Gómez et al., 2013a). A k-means cluster analysis identified two clusters: weak teams (before the COVID-19 lockdown: winning = 37.3 ± 7.6%, after the COVID-19 lockdown: winning = 33.7 ± 7.5%) and strong teams (before the COVID-19 lockdown: winning = 66.4 ± 6.8%, after the COVID-19 lockdown: winning = 59 ± 7.6%).

- Match type: According to the previous studies (Zhang et al., 2019b), a k-means cluster analysis was performed on the entire sample with the aim of creating and describing maximal different groups of match type (balanced and unbalanced matches). The cubic clustering criterion, together with Monte Carlo simulations, was used to identify the optimal number of clusters, thereby avoiding using subjective criteria. A k-means cluster analysis identified a threshold for scoring differences of a match with balanced (cluster 1, 1–14 points difference) and unbalanced (cluster 2, >15 points difference) matches identified.

To control for game rhythm, all variables were then normalized according to game ball possessions and multiplied by 100 (Kubatko et al., 2007). Additionally, possessions are the most important value of advanced statistics, as they are the basis for comparing the indicators that are generated. All calculations such as offensive efficiency, defensive efficiency, rebounding rate or percentage of assists, and shooting accuracy, are normalized on the basis of possessions played. In this way we can compare different games or leagues in future studies. Briefly, ball possessions were calculated using the following equation: 0.976 × (field-goal attempts + (0.4 × free-throw attempts) – "offensive rebounds" + "turnovers") (Kubatko et al., 2007). To assess the validity of data sets, a sub-sample of 10 games was randomly selected and observed by two experienced analysts (basketball coaches with more than 5 years of experience in basketball performance analysis) who recorded key performance indicators. First, two basketball experts were interviewed separately and answered the following question: "In your opinion, which information (technical and tactical actions) can we extract from the match is the more relevant current study?" The basketball experts have the following profiles: Expert 1 – professor in basketball science at a local university; Expert 2 – Assistant coach in a professional basketball club. Then, the experts' answers were compiled and analyzed by the authors of this study (Santos et al., 2022). These results were contrasted with those gathered within the official website, and perfect Intra-class Correlation Coefficients (ICC = 1.0) were obtained for free-throws, two-and three-point field-goals (both made and missed), offensive and defensive rebounds, turnovers, steals, blocked shots, personal fouls. A lower but very acceptable (ICC = 0.93) was obtained for the final performance indicator, assists. All procedures were approved by the local Institutional Research Review Board.

Statistical analysis

Data normality assumptions were verified by using the Kolmogorov–Smirnov test and homogeneity of variance was testified by the Levene test. Data were presented as mean ± standard deviation. An independent t-test was used to identify the difference in the game performance-related variables of the home and away teams before and after the COVID-19 epidemic. To clarify the meaningfulness, Cohen's *d* effect sizes and 95% confidence intervals (CI) were calculated (Cohen, 1988; Fritz et al., 2012). Effect sizes (ES) were interpreted as follows: ≤0.2 trivial, >0.2–0.6 small, >0.6–1.2 moderate, >1.2–2.0 large, >2.0–4.0 very large, and > 4.0 extremely large (Hopkins et al., 2009).

Then, binary logistic regression model was used to identify the key winning factors for both 971 games of the 2019–2020 NBA regular season before the COVID-19 lockdown and 578 ghost games of the 2020–2021 NBA regular season after the COVID-19 pandemic, specifically. Univariate analysis was used to identify individual predictors. Variables with a univariate significance of $p < 0.01$ were entered into a multiple stepwise regression analysis to determine the independence of these predictors (Fairbairn et al., 2012). A significance level of $p < 0.05$ was considered statistically significant. All statistical analyses were performed in R (version 3.5.3; Boston, MA).

Results

The differences between home and away matches before and after COVID-19

Table 1 and Figure 1 illustrate the ES and confidence intervals (95%CI) between home and away teams before and after COVID-19. Before COVID-19, home teams had a clear advantage over away games in terms of defensive rebounds ($p < 0.001$; ES, 0.21), assists ($p < 0.001$; ES, 0.21), two-point field goals made ($p < 0.01$; ES, 0.14), offensive rebounds ($p < 0.01$; ES, 0.12), and blocks ($p < 0.05$; ES, 0.11). By contrast, away teams missed more three-point field goals ($p < 0.05$; ES, 0.1), stole more ($p < 0.05$; ES, 0.1), and committed more personal fouls ($p < 0.05$; ES, 0.09) than home teams. In the ghost games after COVID-19, away teams secured more offensive rebounds than home teams which was the only indicator with statistical significance ($p < 0.05$; ES, 0.11).

The key factors determined between winning and losing matches before and after COVID-19

The inclusion of these 16 variables in a univariate binary logistic regression model resulted in 10 variables that were independently statistically significant winning factors before COVID-19. These variables were further analyzed by multivariable analysis (Table 2; Figure 2). After multivariable analysis, these 10 variables were still statistically significant which included free-throws made (OR, 1.24; 95%CI, 1.21–1.28; $p < 0.001$), two-point field goals made (OR, 1.44; 95%CI, 1.38–1.51; $p < 0.001$), both three-point field goals made (OR, 1.75; 95%CI, 1.64–1.86; $p < 0.001$) and missed (OR, 0.95; 95%CI, 0.92–0.98; $p < 0.01$), defensive rebounds (OR, 1.43; 95%CI, 1.37–1.49; $p < 0.001$), steals (OR, 1.46; 95%CI, 1.38–1.55; $p < 0.001$), blocks (OR, 1.14; 95%CI, 1.07–1.21; $p < 0.001$), personal fouls (OR, 0.95; 95%CI, 0.92–0.98; $p < 0.01$), match location (OR, 1.34; 95%CI, 1.02–1.78; $p < 0.05$), and opponent quality (OR, 2.35; 95%CI, 1.77–3.14; $p < 0.001$).

10.3389/fpsyg.2022.951779

TABLE 1 Descriptive statistics of match performances before and after COVID-19 in the NBA.

	Before COVID-19 (with spectators)				After COVID-19 (without spectators)			
	Home	Away	*p*-value	ES (95% CI)	Home	Away	*p*-value	ES (95% CI)
FT Made	17.67 ± 6.04	17.24 ± 5.74	0.097	−0.08 (−0.16, 0.01)	17.24 ± 5.88	16.9 ± 5.75	0.291	−0.06 (−0.18, 0.05)
FT Missed	5.25 ± 2.57	5.13 ± 2.71	0.294	−0.05 (−0.14, 0.04)	4.82 ± 2.69	4.76 ± 2.56	0.703	−0.02 (−0.14, 0.09)
FG2 Made	28.9 ± 5.3	28.15 ± 5.22	<0.01**	−0.14 (−0.23, −0.06)	28.31 ± 5.22	28.72 ± 5.24	0.167	0.08 (−0.03, 0.20)
FG2 Missed	25.86 ± 5.77	26.12 ± 5.81	0.324	0.04 (−0.04, 0.13)	25.11 ± 5.79	25.6 ± 5.69	0.147	0.09 (−0.03, 0.20)
FG3 Made	12.08 ± 3.84	12.01 ± 3.75	0.689	−0.02 (−0.11, 0.07)	12.86 ± 4.06	12.66 ± 4.12	0.404	−0.05 (−0.16, 0.07)
FG3 Missed	21.37 ± 5.06	21.87 ± 5.04	<0.05*	0.10 (0.01, 0.19)	21.93 ± 5.15	22 ± 5.13	0.797	0.02 (−0.10, 0.13)
OReb	10.28 ± 3.63	9.86 ± 3.6	<0.01**	−0.12 (−0.21, −0.03)	9.65 ± 3.61	10.05 ± 3.42	<0.05*	0.11 (0.00, 0.23)
DReb	35.01 ± 5.09	33.93 ± 5.39	<0.001***	−0.21 (−0.30, −0.12)	34.54 ± 4.88	34.56 ± 5.27	0.972	0.00 (−0.11, 0.12)
Ast	24.72 ± 4.92	23.7 ± 4.89	<0.001***	−0.21 (−0.30, −0.12)	25.04 ± 4.95	24.82 ± 4.87	0.473	−0.04 (−0.16, 0.07)
Stl	7.44 ± 2.82	7.72 ± 2.9	<0.05*	0.10 (0.01, 0.19)	7.49 ± 2.94	7.6 ± 2.88	0.551	0.04 (−0.08, 0.15)
Blk	5.05 ± 2.55	4.77 ± 2.41	<0.05*	−0.11 (−0.20, −0.02)	4.89 ± 2.45	4.8 ± 2.34	0.531	−0.04 (−0.15, 0.08)
To	14.46 ± 3.95	14.31 ± 3.93	0.423	−0.04 (−0.13, 0.05)	14.22 ± 3.8	13.95 ± 3.77	0.211	−0.07 (−0.19, 0.04)
PF	20.2 ± 4	20.57 ± 4.14	<0.05*	0.09 (0.00, 0.18)	19.55 ± 3.91	19.62 ± 4.08	0.757	0.02 (−0.10, 0.13)

*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$. FT, free-throw; FG2, two field-goals; FG3, three field-goals; OReb, offensive rebounds; DReb, defensive rebounds; Ast, assists; Stl, steals; Blk, blocks; To, turnovers; PF, personal fouls.

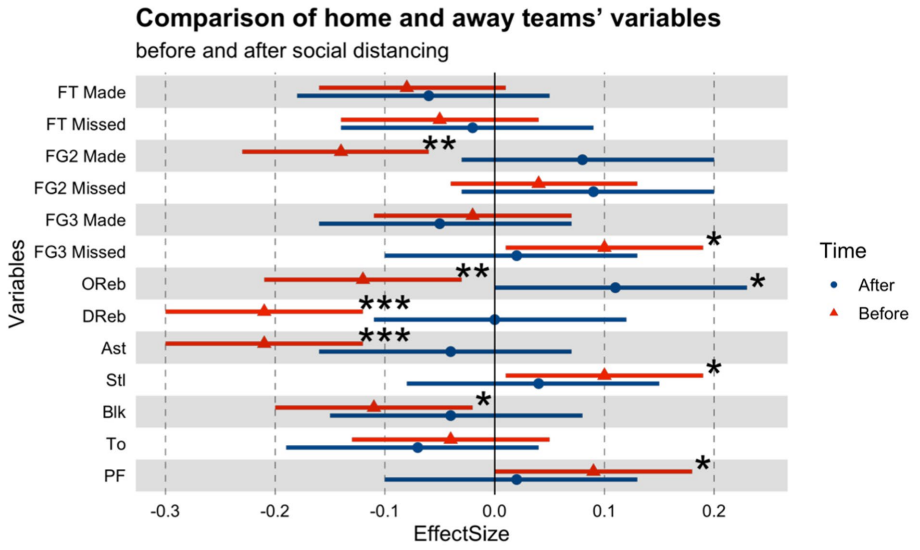


FIGURE 1
Comparison of home and away teams' variables before and after social distancing. *p<0.05, **p<0.01, ***p<0.001. FT, free-throw; FG2, two field-goals; FG3, three field-goals; OReb, offensive rebounds; DReb, defensive rebounds; Ast, assists; Stl, steals; Blk, blocks; To, turnovers; PF, personal fouls.

After the epidemic, nine variables were independently statistically significant winning factors by univariate binary logistic regression model and still statistically significant after multivariable analysis. These variables included both free-throws made (OR, 1.12; 95%CI, 1.08–1.15; $p < 0.001$) and missed (OR, 0.91; 95%CI, 0.86–0.97; $p < 0.01$), both three-point field goals made (OR, 1.27; 95%CI, 1.22–1.33; $p < 0.001$) and missed (OR, 0.86; 95%CI, 0.83–0.89; $p < 0.001$), defensive rebounds (OR, 1.28; 95%CI, 1.24–1.33; $p < 0.001$), steals (OR, 1.31; 95%CI, 1.24–1.39; $p < 0.001$), blocks (OR, 1.08; 95%CI, 1.01–1.15; $p < 0.05$), personal fouls (OR, 0.94; 95%CI, 0.91–0.98; $p < 0.01$), and opponent quality (OR, 2.52; 95%CI, 1.84–3.36; $p < 0.001$).

Discussion

The aim of this study was to compare differences in the match performances during pre- and post-COVID-19 lockdown and to identify the key factors to ultimate success between matches with and without spectators. First, our study found that offensive rebounds were the only indicator differentiating between home and away games after the COVID-19 lockdown. Second, the game location was a key factor differentiating between winning and losing games before the COVID-19 lockdown, whereas it failed to be highlighted after the COVID-19 lockdown. Therefore, crowd support may have a significant impact on winning games in the

TABLE 2 Results relating to the logistic regression models run (dependent variable is "match outcome=WIN").

	Before COVID-19 (with spectators)				AFTER (without spectators)			
	Univariate analysis		Multivariable analysis		Univariate analysis		Multivariable analysis	
	p-value	OR (95%CI)	*p*-value	OR (95%CI)	*p*-value	OR (95%CI)	*p*-value	OR (95%CI)
FT made	<0.001****	1.18 (1.09, 1.28)	<0.001****	1.24 (1.21, 1.28)	<0.01***	1.16 (1.05, 1.28)	<0.001****	1.12 (1.08, 1.15)
FT missed	0.450	0.97 (0.88, 1.06)			<0.1*	0.90 (0.80, 1.02)	<0.01***	0.91 (0.86, 0.97)
FG2 made	<0.01***	1.30 (1.10, 1.55)	<0.001****	1.44 (1.38, 1.51)	0.120	1.18 (0.96, 1.46)		
FG2 missed	0.112	0.87 (0.74, 1.03)			0.106	0.84 (0.68, 1.04)		
FG3 made	<0.001****	1.58 (1.32, 1.88)	<0.001****	1.75 (1.64, 1.86)	<0.01***	1.38 (1.11, 1.72)	<0.001****	1.27 (1.22, 1.33)
FG3 missed	<0.1*	0.85 (0.71, 1.00)	<0.01***	0.95 (0.92, 0.98)	<0.05**	0.80 (0.65, 0.99)	<0.001****	0.86 (0.83, 0.89)
OReb	0.108	1.15 (0.97, 1.37)			0.120	1.19 (0.96, 1.47)		
DReb	<0.001****	1.42 (1.37, 1.48)	<0.001****	1.43 (1.37, 1.49)	<0.001****	1.40 (1.33, 1.47)	<0.001****	1.28 (1.24, 1.33)
Ast	0.486	0.99 (0.95, 1.02)			0.148	1.03 (0.99, 1.08)		
Stl	<0.001****	1.46 (1.37, 1.55)	<0.001****	1.46 (1.38, 1.55)	<0.001****	1.43 (1.33, 1.54)	<0.001****	1.31 (1.24, 1.39)
Blk	<0.001****	1.13 (1.07, 1.20)	<0.001****	1.14 (1.07, 1.21)	<0.001****	1.14 (1.06, 1.23)	<0.05**	1.08 (1.01, 1.15)
To	0.430	0.93 (0.79, 1.10)			0.156	0.86 (0.69, 1.06)		
PF	<0.001****	0.94 (0.91, 0.97)	<0.01***	0.95 (0.92, 0.98)	<0.01***	0.94 (0.90, 0.99)	<0.01***	0.94 (0.91, 0.98)
Location	<0.1*	1.31 (0.99, 1.74)	<0.05**	1.34 (1.02, 1.78)	0.647	1.08 (0.77, 1.53)		
Opponent	<0.001****	2.39 (1.79, 3.19)	<0.001****	2.35 (1.77, 3.14)	<0.001****	1.99 (1.38, 2.87)	<0.001****	2.52 (1.84, 3.36)
Competition	0.841	0.95 (0.57, 1.57)			0.616	0.88 (0.55, 1.43)		

*$p < 0.1$; **$p < 0.05$; ***$p < 0.01$; ****$p < 0.001$. FT, free-throw; FG2, two field-goals; FG3, three field-goals; OReb, offensive rebounds; DReb, defensive rebounds; Ast, assists; Stl, steals; Blk, blocks; To, turnovers; PF, personal fouls; Location (the reference group is away games); Opponent (the reference group is the strong teams); Competition (the reference group is balanced competitions).

NBA. Third, free-throws made, three-point field goals made, defensive rebounds, assists, steals, personal fouls, and opponent quality are in line with the previous studies (Ibáñez et al., 2003; Zhang et al., 2017) suggesting that key factors discriminated between win and loss whatever the period (pre- and post-COVID-19 lockdown) of analysis.

The differences between home and away matches before and after COVID-19

Our study about the differences in match performances between home and away games before the COVID-19 pandemic is in line with the previous studies. Ehrlich and Potter (2022) found that the presence of fans matters to home team performance; in fact, "ghost games" eliminated HA in totality. In particular, Leicht et al. (2017) identified that home teams displayed better performance in terms of shooting efficiency, and offensive and defensive rebounds whereas away teams often make fouls to disturb the game pace of home teams and attempted more aggressive techniques such as steals combined with long-distance shooting to overcome game unexpected factors (e.g., dynamic tactics from home teams, self-negative psychological and behavioral states, crowd pressure, or less protection by the referees). After the COVID-19 pandemic, our study only highlighted offensive rebounds that discriminated between home and away games. Indeed, consistent with prior research (White and Sheldon, 2014), crowd attendance was associated with an improvement in home team rebounding differential (a measurement of effort). Rebounds are widely considered as a "hustle" and "grunt work" statistic since it requires players to fight for optimal position, where rough and physical contact is inevitable (Maheswaran et al., 2012). Offensive rebounds, which means to secure their own missed shot attempts, are considered a particularly robust measurement of effort because offensive players are often further from the rim when a shot is attempted, and they have a lower probability of securing the ball. Additionally, a substantial increase in attention both by the performer i.e., heightened self-focus as well as others at home in view (i.e., players, coaches, referees, and primarily the crowd), places a significant psychological inspiration on the performer for securing offensive rebounds.

The key factors determined between winning and losing matches before and after COVID-19

The impact of game-related statistics on match outcome did not change much before and after COVID-19 according to the logistic regression models. Specifically, the field goal made, defensive rebounds, steals, and blocks were positively correlated with the winning games whereas the missed free-throws, missed three-point field goals, and personal fouls were negatively correlated with winning games. The highlighted positive variables were supported by Sampaio et al. (2010) who

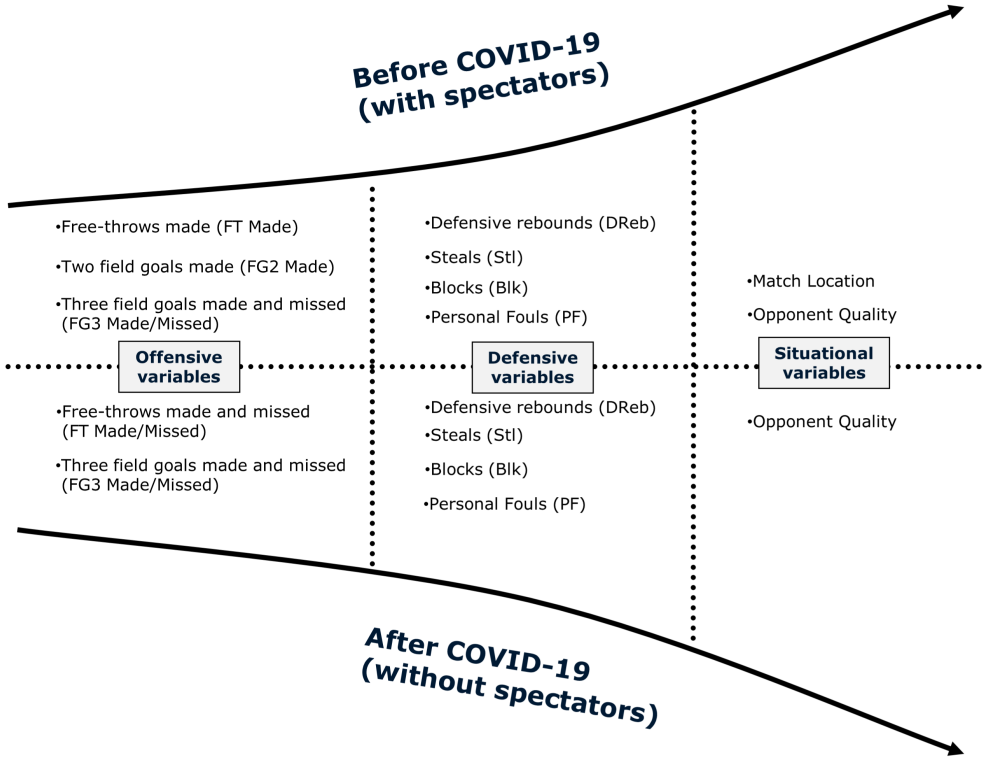


FIGURE 2
The key factors determined between winning and losing matches before and after COVID-19.

suggested that maintaining high shooting efficiency in offense and preventing a team from scoring with defensive pressure (e.g., steals, defensive rebounds, blocks) in defense can be a key determinant of the success of a team. In addition, Gómez et al. (2013b) and Paulauskas et al. (2018) noted that home teams perform better in terms of the mentioned positive variables than away teams. These studies speculated that crowd support was deemed to be critical, due to the spectators' proximity to the playing area and the more constant, loud, inspiring sounds from the crowd, where enthusiastic cheers and chants can inspire initiative, and aggressiveness and encourage home players to try harder. However, our study found that these variables are still key factors associated with match outcome when playing without spectators. Therefore, the team should build up an effective of playing style to win basketball matches, whether or not crowds support matters. Outside players are required to have perimeter shooting skills, including three-point shooting, as well as to guard the opposition with aggressive pressure on the perimeter while inside players can prevent shooting from opponents and secure more defensive rebounds to organize fast breaks (Zhang et al., 2019b). The recent emergence of "small-ball" appears to be a critical factor in the NBA, as this style was more common in dominant teams during the "current" evolution of the NBA (Zhang et al., 2019a). In addition, our study also mentioned coaches who pay more attention to free throws when playing without crowds, especially for away teams, should seize the opportunity to improve the free-throw efficiency without being disturbed by home fans (Sampaio et al., 2006). It is worth noting that game location is the key factor determining between win and loss before the COVID-19 lockdown whereas it failed to be highlighted after the COVID-19 lockdown. Thus, this result appears to identify crowd support plays a key role in winning matches in the NBA which is in line with the previous studies (Huyghe et al., 2021) found that crowd support leads to HA in the NBA is a well-documented phenomenon that has been identified in over 7,000 games spanning 14 seasons (2004–2018) altogether.

There are limitations in the current study that should be considered. Our study only takes advantage of the natural experiment to consider the impact of crowd support on match performances, but match performances may be affected by referee bias, coaches' tactics, and travel fatigue, so future studies are recommended to consider the interactive effect of these factors based on the current study. Additionally, the selected variables about match performance for our study were only based on traditional statistics, limiting to explanation of the key factors differentiating between win and loss during the period of pre-and-post COVID-19. A possible solution is to utilize each quarter's data integrated with tracking and event data to make a spatio-temporal analysis to explore the impact of space control on match performance.

Conclusion

In this study, researchers presented findings from a natural experiment caused by the COVID-19 to examine HA and its drivers and mechanisms in the NBA, especially the factor of crowd support. Our study found that offensive rebounds were the only indicator that presented the difference between home and away games after the COVID-19 lockdown. Second, the match location was the key factor determining between win and loss before the COVID-19 lockdown, whereas it failed to be highlighted after the COVID-19 lockdown. Free-throws made, three-point field goals made, defensive rebounds, assists, steals, personal fouls, and opponent quality were the common key factors discriminating between win and loss whatever pre -and post-COVID-19 lockdown.

Although only a descriptive case series design, our results offer some opinions that might be of interest to coaches and practitioners. Coaches may increase more practice in relation to the skills of box-out, thus allowing players to secure more offensive rebounds in the offense. Furthermore, coaches should adapt to the change in terms of HA by adjusting game strategies and player rotation to game success.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found at: www.nba.com/stats/ and www.basketball-reference.com.

Author contributions

PL, SZ, and MG contributed to the conception and design of the study. PL, XW, and SZ collected and organized the data. PL and JD performed the statistical analysis. PL wrote the first draft of the manuscript. All authors contributed to the article and approved the submitted version.

Acknowledgments

This research was conducted under the Sport Sciences Network (2022): 25/UPB/22 SPAA. Sports Performance Analysis Association.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Alonso, E., Lorenzo, A., Ribas, C., and Gómez, M. Á. (2022). Impact of COVID-19 pandemic on HOME advantage in different European professional basketball leagues. *Percept. Mot. Skills* 129, 328–342. doi: 10.1177/00315125211072483

Arboix-Alió, J., Trabal, G., Peña, J., Arboix, A., and Hileno, R. (2022). The behaviour of home advantage during the COVID-19 pandemic in European rink hockey leagues. *Int. J. Environ. Res. Public Health* 19:228. doi: 10.3390/ijerph19010228

Bustamante-Sánchez, Á., Gómez, M. A., Jiménez-Saiz, S. L., Gómez, M. A., and Jiménez-Saiz, S. L. (2022). Game location effect in the NBA: a comparative analysis of playing at home, away and in a neutral court during the COVID-19 season. *Int. J. Perform. Anal. Sport* 22, 370–381. doi: 10.1080/24748668.2022.2062178

Cohen, J. (1988). Statistical power analysis for the behavioral sciences. *Technometrics* 31, 499–500.

Ehrlich, J., and Potter, J. (2022). Estimating the effect of attendance on home advantage in the National Basketball Association. *Appl. Econ. Lett.*, 1–12. doi: 10.1080/13504851.2022.2061898

Fairbairn, T. A., Mather, A. N., Bijsterveld, P., Worthy, G., Currie, S., Goddard, A. J., et al. (2012). Diffusion-weighted MRI determined cerebral embolic infarction following transcatheter aortic valve implantation: assessment of predictive risk factors and the relationship to subsequent health status. *Heart* 98, 18–23. doi: 10.1136/heartjnl-2011-300065

Fioravanti, F., Delbianco, F., and Tohmé, F. (2021). Home Advantage and Crowd Attendance: Evidence from Rugby during the Covid 19 Pandemic. arXiv preprint ar xiv: 2105.01446.

Fischer, K., and Haucap, J. (2021). Does crowd support drive the home advantage in professional football? Evidence from German ghost games during the COVID-19 pandemic. *J. Sports Econ.* 22, 982–1008. doi: 10.1177/15270025211026552

Fritz, C. O., Morris, P. E., and Richler, J. J. (2012). Effect size estimates: current use, calculations, and interpretation. *J. Exp. Psychol. Gen.* 141, 2–18. doi: 10.1037/a0024338

García, J., Ibáñez, J. S., Gómez, A. M., and Sampaio, J. (2014). Basketball game-related statistics discriminating ACB league teams according to game location, game outcome and final score differences. *Int. J. Perform. Anal. Sport* 14, 443–452. doi: 10.1080/24748668.2014.11868733

Garcia-Rubio, J., Saez, J., Ibanez, S. J., Parejo, I., and Canadas, M. (2009). Home advantage analysis in ACB league in season 2007-2008. *Rev. Psicol. Deporte* 18, 331–335.

Gómez, M. A., Lorenzo, A., Ibáñez, S. J., Ortega, E., Leite, N., and Sampaio, J. (2010). An analysis of defensive strategies used by home and away basketball teams. *Percept. Mot. Skills* 110, 159–166. doi: 10.2466/pms.110.1.159-166

Gómez, M. A., Lorenzo, A., Ibañez, S. J., and Sampaio, J. (2013b). Ball possession effectiveness in men's and women's elite basketball according to situational variables in different game periods. *J. Sports Sci.* 31, 1578–1587. doi: 10.1080/02640414.2013.792942

Gómez, M. A., and Pollard, R. (2011). Reduced home advantage for basketball teams from capital cities in Europe. *Eur. J. Sport Sci.* 11, 143–148. doi: 10.1080/17461391.2010.499970

Gómez, M.-Á., Prieto, M., Pérez, J., and Sampaio, J. (2013a). Ball possession effectiveness in men's elite floorball according to quality of opposition and game period. *J. Hum. Kinet.* 38, 227–237. doi: 10.2478/hukin-2013-0062

Gong, H. (2022). The effect of the crowd on home bias: Evidence from NBA games during the COVID-19 pandemic. *J. Sports Econ.* 23, 950–975. doi: 10.1177/15270025211073337

Goumas, C. (2017). Modelling home advantage for individual teams in UEFA champions league football. *J. Sport Health Sci.* 6, 321–326. doi: 10.1016/j.jshs.2015.12.008

Higgs, N. (2021). *Home Advantage in North American Professional Sports before and during COVID-19: A Bayesian Perspective,* University of Saskatchewan.

Hopkins, W. G., Marshall, S. W., Batterham, A. M., and Hanin, J. (2009). Progressive statistics for studies in sports medicine and exercise science. *Med. Sci. Sports Exerc.* 41, 3–12. doi: 10.1249/MSS.0b013e31818cb278

Huyghe, T., Alcaraz, P. E., Calleja-González, J., and Bird, S. P. (2021). The underpinning factors of NBA game-play performance: a systematic review (2001–2020). *Phys. Sportsmed.* 50, 94–122. doi: 10.1080/00913847.2021.1896957

Ibáñez, S., Sampaio, J., Sáenz-López, P., Giménez, J., and Janeira, M. (2003). Game statistics discriminating the final outcome of junior world basketball championship matches (Portugal 1999). *J. Hum. Mov. Stud.* 45, 1–20.

Inan, T. (2020). The effect of crowd support on home-field advantage: evidence from European football. *Ann. Appl. Sport Sci.* 8, 7–16. doi: 10.29252/aassjournal.806

Kubatko, J., Oliver, D., Pelton, K., and Rosenbaum, D. T. (2007). A starting point for analyzing basketball statistics. *J. Quant. Anal. Sports* 3:1070. doi: 10.2202/1559-0410.1070

Leicht, A. S., Gómez, M. A., and Woods, C. T. (2017). Explaining match outcome during the men's basketball tournament at the olympic games. *J. Sports Sci. Med.* 16, 468–473.

Leota, J., Hoffman, D., Mascaro, L., Czeisler, M. É., Nash, K., Drummond, S., et al. (2021). Home Is Where the Hustle Is: The Influence of Crowds on Effort and Home Advantage in the National Basketball Association. Available at SSRN: https://ssrn.com/abstract=3898283

Maheswaran, R., Chang, Y.-H., Henehan, A., and Danesis, S. (2012). Deconstructing the Rebound with Optical Tracking Data. Proceedings of the 6th Annual MIT SLOAN Sports Analytics Conference.

Mateus, N., Gonçalves, B., and Sampaio, J. (2021). "Home advantage in basketball" in *Home Advantage in Sport: Causes and the Effect on Performance* (Routledge), 211–219.

Mccarrick, D., Bilalic, M., Neave, N., and Wolfson, S. (2021). Home advantage during the COVID-19 pandemic: analyses of European football leagues. *Psychol. Sport Exerc.* 56:102013. doi: 10.1016/j.psychsport.2021.102013

Neave, N., and Wolfson, S. (2003). Testosterone, territoriality, and the 'home advantage'. *Physiol. Behav.* 78, 269–275. doi: 10.1016/S0031-9384(02)00969-1

Nevill, A. M., and Holder, R. L. (1999). Home advantage in sport. *Sports Med.* 28, 221–236. doi: 10.2165/00007256-199928040-00001

Paulauskas, R., Masiulis, N., Vaquera, A., Figueira, B., and Sampaio, J. (2018). Basketball game-related statistics that discriminate between european players competing in the nba and in the euroleague. *J. Hum. Kinet.* 65, 225–233. doi: 10.2478/hukin-2018-0030

Pollard, R. (2008). Home advantage in football: a current review of an unsolved puzzle. *Open Sports Sci. J.* 1, 12–14. doi: 10.2174/1875399X00801010012

Ponzo, M., and Scoppa, V. (2018). Does the home advantage depend on crowd support? Evidence from same-stadium derbies. *J. Sports Econ.* 19, 562–582. doi: 10.1177/1527002516665794

Ribeiro, H. V., Mukherjee, S., and Zeng, X. H. T. (2016). The advantage of playing home in NBA: microscopic, team-specific and evolving features. *PLoS One* 11:e0152440. doi: 10.1371/journal.pone.0152440

Sampaio, J., Ibanez, S. J., Gomez, M. A., Lorenzo, A., and Ortega, E. (2008). Game location influences basketball players' performance across playing positions. *Int. J. Sport Psychol.* 39, 205–216.

Sampaio, J., Janeira, M., Ibáñez, S., and Lorenzo, A. (2006). Discriminant analysis of game-related statistics between basketball guards, forwards and centres in three professional leagues. *Eur. J. Sport Sci.* 6, 173–178. doi: 10.1080/17461390600676200

Sampaio, J., Lago, C., Casais, L., and Leite, N. (2010). Effects of starting score-line, game location, and quality of opposition in basketball quarter score. *Eur. J. Sport Sci.* 10, 391–396. doi: 10.1080/17461391003699104

Santos, R. A., Idiakez, J. A., Ajamil, D. L., and Argilaga, M. T. A. (2022). Analysis of efficiency in under-16 basketball: a log-linear analysis in a systematic observation study. *Cult. Cienc. Deporte* 17, 105–112. doi: 10.12800/ccd.v17i51.1736

Szabó, D. Z. (2022). The impact of differing audience sizes on referees and team performance from a North American perspective. *Psychol. Sport Exerc.* 60, –102162. doi: 10.1016/j.psychsport.2022.102162

Tilp, M., and Thaller, S. (2020). Covid-19 has turned home-advantage into home-disadvantage in the German soccer Bundesliga. *Front. Sports Active Living* 2:165. doi: 10.3389/fspor.2020.593499

White, M. H., and Sheldon, K. M. (2014). The contract year syndrome in the NBA and MLB: a classic undermining pattern. *Motiv. Emot.* 38, 196–205. doi: 10.1007/s11031-013-9389-7

Wunderlich, F., Weigelt, M., Rein, R., and Memmert, D. (2021). How does spectator presence affect football? Home advantage remains in European top-class football matches played without spectators during the COVID-19 pandemic. *PLoS One* 16:e0248590. doi: 10.1371/journal.pone.0248590

Zhang, S., Lorenzo, A., Gómez, M.-A., Liu, H., Gonçalves, B., and Sampaio, J. (2017). Players' technical and physical performance profiles and game-to-game variation in NBA. *Int. J. Perform. Anal. Sport* 17, 466–483. doi: 10.1080/24748668.2017.1352432

Zhang, S., Lorenzo, A., Woods, C. T., Leicht, A. S., and Gomez, M.-A. (2019a). Evolution of game-play characteristics within-season for the National Basketball Association. *Int. J. Sports Sci. Coach.* 14, 355–362. doi: 10.1177/1747954119847171

Zhang, S., Lorenzo, A., Zhou, C., Cui, Y., Gonçalves, B., and Angel Gómez, M. (2019b). Performance profiles and opposition interaction during game-play in elite basketball: evidences from National Basketball Association. *Int. J. Perform. Anal. Sport* 19, 28–48. doi: 10.1080/24748668.2018.1555738

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 13 December 2022
DOI 10.3389/fpsyg.2022.1023697

OPEN ACCESS

EDITED BY
George Waddell,
Royal College of Music,
United Kingdom

REVIEWED BY
Rui Cruz,
European University of Lisbon,
Portugal
Min Wu,
Sichuan University, China

*CORRESPONDENCE
Jianing Lv
lvjianingimu@163.com
Xiaoyang Zhao
zxyresearch2021@163.com

SPECIALTY SECTION
This article was submitted to Performance Science, a section of the journal Frontiers in Psychology

RECEIVED 20 August 2022
ACCEPTED 18 November 2022
PUBLISHED 13 December 2022

CITATION
Wang F, Lv J and Zhao X (2022) How do information strategy and information technology governance influence firm performance?
Front. Psychol. 13:1023697.
doi: 10.3389/fpsyg.2022.1023697

How do information strategy and information technology governance influence firm performance?

Fanlin Wang[1], Jianing Lv[1]* and Xiaoyang Zhao[2]*

[1]School of Accounting, Capital University of Economics and Business, Beijing, China, [2]School of Business Administration, Huaqiao University, Quanzhou, China

Organizations today engage in turbulent competition to seize opportunities and cope with challenges by making strategy planning, increasing information technology (IT) investment, and other means. Based on survey data through questionnaires, this paper constructs models to explore the synergistic effects of information strategy (IS) and IT governance (ITG) on firm performance. The results show that, first, ITG and IS as explanatory variables have a significant influence on firm performance. Second, ITG has a positive effect on the relationship between IS and firm performance. This study extends existing research on IS and ITG by exploring the synergistic effects of IS and ITG on firm performance. The conclusion provides management insight and practical guidance for enterprises by actively implementing IS to improve firm performance to transform from the inherent pattern of traditional governance to the new technology governance.

KEYWORDS

information strategy, IT governance, firm performance, synergistic effects, IT investment

1 Introduction

Information technology (IT) performance has always been the focus of information system (IS) research (Ravichandran and Liu, 2011). Companies have invested a lot in IT investment over the past decades. However, IT performance varies greatly from company to company (Mithas et al., 2012; Ilmudeen et al., 2022). Earlier studies have debated the so-called "productivity paradox" concerning the ambiguous impact of IT investments on firm performance (Takeda et al., 2021), but the intricate relationship between IT investment and corporate performance remains to be explored. For less-developed economies, in particular, several recent studies have shown a less positive relationship between IT investment and performance in developing countries (Lee and Kim, 2006). According to the research data of the international data corporation

(IDC), many enterprises have increased their investment in IT. Whether the increasing investment in IT projects has a positive or negative effect on the improvement of firms' performance is academically controversial. Part of the related literature has proven a significantly positive effect of IT investment on firms' financial performance because of the improvement of scientific decisions by IT systems (Ji et al., 2019; Dong et al., 2021; Alfalah et al., 2022). While other studies on the impact of IT investments show weak or non-existent links between IT investment and firm productivity. For example, Ru et al. (2018) find that the IT investment for vendor-managed inventory is not obvious based on anecdotal evidence and empirical studies. Soh and Markus (1995) state that IT investment cannot be directly transformed into corporate performance, only when IT investment is first transformed into IT assets can IT have an impact on corporate performance. From the perspective of resource input and output, based on different research situations, different conclusions have been obtained. Therefore, relevant empirical evidence needs to be further enriched.

To achieve organizational goals, IT in an organization is not enough if it is only regulated, but IT must be managed professionally. Earlier studies found that many organizations have taken note of the concept of IT governance (ITG) in order to justify IT investments (Mikalef et al., 2021; Ali et al., 2022). Because ITG can balance the interests of all parties and enables IT investments to be effective, the matching of IT goals and enterprise goals is a part of corporate governance (Simonsson et al., 2010). Lunardi et al. (2014) found that ITG adaptation had a positive impact on IT investment performance by measuring the pre-adaptation and post-adaptation performance of Brazilian organizations. Several researchers have done works that are associated with the relationship between ITG and corporate performance. For instance, Prasad et al. (2012) suggest that ITG structures contribute to firm performance through IT-related capabilities that improve the effectiveness and efficiency of internal business processes. Prasad and Green (2015) found evidence of a positive association between these ITG considerations and overall firm performance. In addition, Wu et al. (2015) find that strategic alignment can mediate the effectiveness of ITG on organizational performance. Meanwhile, several practical cases show that firms with stronger ITG are more likely to possess business and IT knowledge to nurture organizational learning.

In addition, ITG can promote IT resources to support enterprise performance according to the established goals and improve the company's cost structure and revenue level, which will increase with IT inputs' increasing (Aubert and Rivard, 2020). Based on the resource-based view (RBV), IT investment is explicitly included as the resource that strategic IT alignment as a capability can inherently help leverage (Sabherwal et al., 2019). In other words, the essence of ITG is the mechanism of information strategy (IS) and ITG (Sheppard, 1990). Specifically, IS determines the direction and goal of the enterprise's informatization, while ITG promotes the effective implementation of IS. Companies invest in IT mainly for strategic and efficiency considerations (Ahmad and Arshad, 2014; Takeda et al., 2021). First, managers need to decide how to allocate spending on IT, such as advertising and research and development investment (Mithas et al., 2012; Witra and Subriadi, 2022). IT investment can produce a sustainable competitive advantage, and along with the governance of IT investment, enterprise performance is improved. Specifically, IS and ITG jointly contribute to the realization of corporate governance goals. Second, performance improvement needs effective resource management and process supervision, otherwise, the input–output effect of IT cannot be guaranteed. The effective use of IT, however, relies heavily on good ITG (Wu et al., 2015). When IT and corporate governance go awry, the goal of the company can be devastating. Therefore, people should consider the consistency of IT resources and enterprise goals as well as the synergy of daily management activities from the perspective of ITG. Given that the desired outcome of effective ITG is to achieve the congruence between information strategies and corporate objectives, however, few research papers have theoretically and empirically examined the effect of ITG and IS on firm performance.

Prior literature (Ahmad and Arshad, 2014; Wu et al., 2015; Ru et al., 2018; Ji et al., 2019; Aubert and Rivard, 2020) generally discusses the economic consequences of IT investment behavior or investment of technical factors, which agree that IT as a resource improves firm performance. Furthermore, the literature on ITG generally shows that ITG plays a positive role in improving firm performance (Wu et al., 2015; Aubert and Rivard, 2020). Only a few studies discuss business and IT strategy alignment (Dairo et al., 2021), the relationship between IT investment and strategic alignment (Sabherwal et al., 2019), and the mediation of strategic alignment in mechanisms that ITG affects organizational performance (Wu et al., 2015). However, the related literature lacks the macro-level thinking of strategy's impact on corporate governance. Specifically, existing research has not explored the impact of ITG and IS on corporate performance. Moreover, the data from the first-hand data directly obtained from the enterprise to consider the IS and ITG are scarce. In order to develop a richer understanding of the relationship between ITG, IS, and firm performance, this study focuses on how ITG influences firm performance and how the alignment effect of ITG and IS affects firm performance.

This study provides research contributions in several primary ways. First, this study extends prior research on firm strategy and firm performance. Existing research mainly focuses on business and IT strategy alignment (Dairo et al., 2021), the relationship between IT investment and strategic alignment (Sabherwal et al., 2019), discusses the mediation of strategic alignment in mechanisms that ITG affects organizational performance (Wu et al., 2015), and lacks research that explores the impact of IS on firm performance from an independent

perspective. This study constructs the measure of IS and tests the impact of IS on firm performance. Furthermore, we explore the influence of IS on enterprise performance from different aspects. This paper focuses on the IS of enterprises, which provides a certain reference value for the application of IS research in organizational performance management. Second, this study adopts the questionnaire data from Chinese enterprises to empirically analyze the impact of informatization strategy and ITG on enterprise performance. Specifically, the data from the first-hand data directly obtained from the enterprises to construct the measure of the IS and ITG are scarce. To fill this research gap, we construct empirical measures for ITG and IS. Third, although most studies related to ITG have explored the relationship between ITG and organizational behavior (Wu et al., 2015; Tawafak et al., 2020), the study still rarely connects IT strategy, ITG, and corporate performance in the same research framework. We also formulate a framework to connect together ITG, IS, and firm performance. Given that ITG and IS have rarely been studied together, their joint relationship in promoting firm performance remains theoretically underdeveloped. The results can provide evidence of the economic consequences of IS under the regulation of ITG, which would inspire Chinese enterprises to adjust current strategies and governance to realize sustainable development.

The remainder of this paper proceeds as follows. The section following reviews the theoretical literature, employing prior literature on ITG, IS, and firm performance, and this section also develops the hypotheses to test the main idea of this study. Subsequently, the research methodology, data collection procedures, variable measurement, and the respondent sample are described. The results of construct validation and model testing are then reported. The paper concludes with a summary of the study findings, highlighting contributions, implications, limitations, and directions for future research.

2 Literature review and hypotheses

2.1 Information strategy and firm performance

Information for an organization is indispensable, which can be used as input in decision-making in order to solve problems faced by the organization (Astuti, 2018). With the upgrading of IT systems, the company will adjust its strategy, which is a basic change in strategic orientation. For example, an ERP is an IS that brings about radical changes within organizations, changing both the IS environment and overall corporate business process, which may influence the organization's performance (Park, 2018). Thus, the strategy is crucial in digital time (Proksch et al., 2021). Because corporate strategy is related to IT mechanisms, a good IS can improve corporate performance (Dong et al., 2021). According to the theory of management experts, such as Porter and Mintzberg, IS should be mentioned together with the corporate general strategy. Thus, many scholars have verified their conclusions through case studies or empirical methods. For example, Jeffers et al. (2008) found that IT can contribute to enhancing the value of the firm *via* its strategic role. Mikalef and Pateli (2017) also suggested that IT-enabled dynamic capabilities facilitate two types of agility, market capitalizing and operational adjustment agility, which in sequence enhance firm performance. Dairo et al. (2021) indicate the strategic alignment of IT and business brings many advantages, including enhanced operational efficiencies, business innovativeness, and additional competitive advantage, which together lead to improved performance. In addition, some scholars believe the complementary system of IT resources has significant effects on corporate performance (Cohen and Olsen, 2013). Lin (2022) expands the understanding of IT value by adding a customer-based view (CBV) to the prevalent RBV and indicates that value from IT investments can have direct or indirect effects on firm performance. The overall IS can guarantee the realization of enterprise performance goals in the aspects of organizational design, resource allocation, and management improvement (Cohen, 2008). However, in practice, many enterprises believe that IT behavior is only a matter of the technical department and that only needs to be considered at the functional department level, just like procurement, production, and other activities. According to relevant management practices, this paper believes that IS which is decided by the management level is of great importance. Compared with enterprises without IS, enterprises with IS orientation have better information effects and a better guarantee of enterprise performance by reforming organization structure design, improving resource allocation, and boosting management efficiency. Hence, this paper proposes the following hypothesis:

> H1: Information strategy has a positive impact on firm performance.

How to use technology strategically to access and apply information quickly is a serious and profound question when IT is highly pervasive. From the perspective of strategy hierarchical structure, IS includes the company's general strategy, competitive strategy, and functional strategy (Lederer et al., 1997). Therefore, when formulating IS, enterprises also have three corresponding choices to match up with the three kinds of strategy types. In order to get closer to the actual situation of enterprises and obtain objective research conclusions and further explore how the IS formulation level affects enterprise performance, this paper designs the following sub-hypotheses from three perspectives.

First, IS has a significant effect on competitive advantage (Astuti, 2018). However, ISs could be a source of sustainable competitive advantage only if the IS will be implemented in

the enterprises. From the perspective of corporate governance structure, the higher the level of strategy formulation, the more secure the resources to implement the strategy, and the strategy is more likely to succeed. Therefore, if the IS is put forward by the top management of the enterprise, the IS will be in line with the goal of managers, and the strategic goal will be vigorously promoted. It will maximize the important role of information resources and be conducive to the improvement of enterprise performance. Therefore, the first sub-hypothesis is proposed as follows:

> H1a: The IS made by the top management will have a positive effect on firm performance.

Second, if a company has a reasonable IS, it must have a thorough and rigorous implementation process, including a comprehensive IT system procurement plan, sufficient talented personnel, technical support, and scientific control of cost, risk, and quality. The control of the process of strategy implementation is as important as the scientific formulation of strategic objectives. As Mohamad et al. (2017) conclude, the enterprise must control the corresponding process in order to improve enterprise performance. According to Nolan's information implementation model theory (Nolan, 1973), the strategy implementation process is the quantitative change process of informatization, but the strategic goal is the qualitative change process of informatization. Only when the two processes are highly unified, can the information resources be used efficiently, the decision-making quality be upgraded highly and overall performance be improved greatly. Therefore, the second sub-hypothesis is proposed:

TABLE 1 Information strategy evaluation index.

Overall index	Specific index	Index code
Information strategy formulation level (SI1)	Formulated and approved by the board of directors	SI11
	Integrate the needs of all departments	SI12
	Information strategy implement 3 years above	SI13
	Detailed planning and implementation of the program	SI14
	Strategic resource planning at company level	SI15
Information strategy process supervision (SI2)	Reasonable IT system procurement plan	SI21
	Adequate human, technical and management resources	SI22
	Reasonable outsourcing or delegation plan	SI23
	Reasonable implementation process plan	SI24
	Reasonably control the change of management	SI25
	Effectively control the process and results	SI26
Information strategy results maintenance (SI3)	Follow-up technical support is incompetence	SI31
	Independent evaluation of the operation effect	SI32
	Operation risk and safety risk is controllable	SI33

> H1b: Information strategy with a clear implementation path and reasonable control has a strong effect on firm performance.

Third, according to the theory of corporate governance, the professional background of board members is conducive to understanding and promoting the implementation of strategies (Turedi, 2020; Shreeve et al., 2022). As the IS is more professional and continuous, the implementation is necessary to need managers who have IT professional background. Qualified managers with IT backgrounds can ensure the implementation of IS from a professional perspective and guarantee the maintenance of information results and the use of information resources, so as to improve enterprise performance. Therefore, the third sub-hypothesis proposes the following:

> H1c: The IT background of board members has a positive effect on firm performance.

2.2 Information technology governance and enterprise performance

The ITG literature provided us with a theoretical basis to investigate how firms effectively build the alignment between IT resources and other resources in creating competitive advantages (Prasad et al., 2012; Wu et al., 2015; Matta et al., 2022). Many companies have spent their resources in order to increase their competitive advantage by controlling their internal processes (Chatzoglou et al., 2011; Wolf and Floyd, 2017). Neff et al. (2013) found ITG is associated with firm performance through IT-relatedness and business process-relatedness. IT-related capabilities also relate to measuring business value at the process and firm levels. This makes it possible to infer that collaborative organizations' ITG efforts contribute to business value (Prasad et al., 2012). Several studies defined ITG. For example, Benaroch and Chernobai (2017) indicate that "ITG is. . .an integral part of enterprise governance and consists of the leadership, organizational structures and processes that ensure that the organization's IT sustains and extends the organization's strategies and objectives." The organization defined ITG as the process of controlling IT resources and balancing the interest demands of various stakeholders under the framework of corporate governance (Huang et al., 2010), so as to make the operational goals of information resources consistent with the overall goals of the enterprise (Williams and Karahanna, 2013).

Information technology governance forms an important and integral part of an organization's corporate governance (Lunardi et al., 2014; De Haes et al., 2019). ITG functions in the same way as enterprise governance to enable an enterprise

to more effectively address major business issues such as enterprise resource planning (Lainhart, 2000). Hence, ITG has received increased attention from business practitioners and researchers (De Haes et al., 2019; Matta et al., 2022). Specifically, ITG includes the management mode of technical resources, stakeholder balancing mechanism, technical level, and the alignment of IT and business systems. From the perspective of economics, ITG includes the implementation of IT software, hardware, and other resources, which could lead to the low efficiency and high cost of the company's short-term business process, and the temporary increase in financial pressure including training costs and error correction cost, but theoretical deduction and practical observation can find that ITG can ultimately improve the overall performance of enterprises (Raymond et al., 2019; Matta et al., 2022).

Information technology governance can improve the company's performance in two ways. On the other hand, ITG cost, which contains the hardware and software investment, security investment, and technology upgrading investment, belongs to environmental investment. This kind of investment can ensure the normal operation of the enterprise's IT system, which indirectly improves the company's performance and directly consumes enterprise resources. At the same time, the integrated IS's overall function, network bearing capacity, the capability of output loading, and other technical resources play an important role in the enterprise's process execution and department collaboration. Therefore, ITG will undoubtedly improve the company's performance. On the other hand, clear and reasonable IT planning is the embodiment of ITG in system service capability. Board-level ITG can improve organizational performance (Matta et al., 2022). More importantly, enterprises can make full use of big data, cloud computing, data mining, and other means to allocate information resources and serve for the improvement of enterprise performance. When an enterprise makes a large IT investment, it is necessary to set up a management team and attach importance to the training of IT skills of employees. Therefore, this process can also improve the professional level and treatment of employees, which is conducive to enhancing the competitiveness of enterprises. In addition, during the construction and operation of the IT system, the business department and IT department need to keep close cooperation. The goal is to embed business or management requirements into IT systems, so as to avoid inadequate implementation in which business and technology are disconnected from each other. Only in this way, the efficiency of business processing, management decision-making, and risk control can be improved, in this process, enterprise performance will also be improved. To sum up, this paper proposes the following hypothesis:

> H2: The implementation of ITG has a positive effect on firm performance.

TABLE 2 Information technology (IT) governance evaluation index.

Process index	Specific index	Index code
IT investment (IV)	Information technology investment quantity	IV1
	The total investment proportion of IT budget	IV2
	The ratio of IT assets to total assets	IV3
IT technology resource (IR)	Technical performance of independent system	IR1
	Overall function of integrated system	IR2
	Network bearing capacity and output load	IR3
IT human resource (IHR)	The proportion of IT staff in the total staff	IHR1
	The salary level of IT staff	IHR2
	The technical level of IT staff	IHR3
	The business level of IT staff	IHR4
IT relationship resource (IRR)	Relationship with business department	IRR1
	Relationship with partnerships	IRR2
	Degree of information sharing	IRR3
IT capability (IC)	IT planning ability	IC1
	Information utilization ability	IC2
	System maintenance ability	IC3

2.3 The synergistic effects of information technology governance

Through the analysis of IS and ITG, it can be found that no matter the ITG mechanism in the context of IS (Schlosser et al., 2015), or the formulation of reasonable IS after analyzing the characteristics of ITG, ITG and IS are always the relationship between strategy and tactics. On the one hand, IS provides direction and targets for IT resource governance. Clear goal setting will make ITG more efficient. On the other hand, in order to maintain competitive edges, organizations try to adapt rapidly to digitalization through the transformation of operations and processes (Ho et al., 2022). ITG basically places an outline around how a firm's IT plan supports a firm plan. This IT-firm configuration will make sure that the business maintains to accomplish its plans and objectives and apply techniques to estimate its performance (Al Romaihi and Hamdan, 2021). Therefore, the depth and breadth of ITG provide implementation support for IS. Strategic alignment and planning have been a top managerial concern since the beginning of the IS profession (Taylor et al., 2010), and prior studies show that ITG mechanisms on organizational performance are fully mediated by strategic alignment (Wu et al., 2015). Therefore, the implementation of an IS can more effectively affect enterprise performance under the influence of ITG. In other words, ITG promotes the promotion of IS to corporate performance, and ITG has synergistic effects on

IS's performance improving effect. Therefore, the following hypothesis is proposed:

> H3: Information technology governance has a positive moderating effect on the relationship between IS and firm performance.

3 Research methodology

3.1 Data collection

Aiming at the influence of IS formulation level and ITG level on enterprise performance, this paper adopted a large-sample questionnaire survey. In the process of questionnaire collection, we made efforts in the following aspects. First, in terms of the subject of investigation, we rely on the China Association of Chief Financial Officers, which is a cross-regional, cross-departmental, and cross-industry national non-profit organization in China to gain the access to distribute questionnaires to 59 group companies and their subsidiaries affiliated to State-Owned Assets Supervision and Administration Commission of China (SASAC). A total of 405 questionnaires were distributed, and 317 valid questionnaires were received during 2016–2018. In 2019, we made a return visit to the relevant companies and added relevant data. The paper questionnaires were distributed to the targeted firms, and then, we collected them later. These firms in Beijing, Shanghai, Guangzhou, Chengdu, Shandong, and other places, the research objects involving managers or directs in manufacturing, Internet, B2B logistics, pharmaceuticals industry, utility industry, and other industries. Second, before distributing questionnaires, we first limited the scope of distributing questionnaires. Specifically, in order to measure the level of ITG and get more information about IS, we choose the firms that have ITG expression in relevant company governance documents. Furthermore, we try to ensure the authenticity and representativeness of the data source; the subjects of the questionnaire are limited to enterprise informatization managers including the CEO, managers of the technical department, or directors of the information center; when a company does not have a technology department, we send questionnaires to managers who are in charge of IT, and we also send questionnaires to financial managers who evaluate enterprise performance. We distributed at least three questionnaires in one firm (the firm refers to group companies or subsidiaries). At the same time, we recheck some financial indexes *via* the China Accounting Informatization Committee and other professional institutions. We also manually checked the relevant financial data with the company's financial statement. Third, each questionnaire inquired about the financial performance data and the IT investment budget data of enterprises in the past 5 years, and the IT investment budget data are based on the average of the 5 years which is due to the firms usually planning their investments over a 5-year horizon. Finally, the data processing is performed on SPSS 19.0.

3.2 Variables

3.2.1 Information strategy

Information strategy stipulates the direction, target, and resource allocation principle of the information scheme (Astuti, 2018). In terms of the decision-making level of the company, IT investment and IS operations are planned and arranged from a higher level. Therefore, the connotation of an IS is abundant and the measurement of an IS is more flexible. In order to accurately grasp the different attributes of IS, this study designs the indicators from the IS formulation level, IS process supervision, and IS result maintenance, specifically including whether the strategy is formulated or approved by the senior leaders of the company (SI1), whether the process supervision of IS touches each important process (SI2), and whether the operation and maintenance of information results are good or not (SI3). All items are expressed on a 10-point Likert scale (where "1" means totally disagree and "10" means totally agree). The specific descriptions are provided in **Table 1**.

3.2.2 Information technology governance

The definition of ITG from the IT Governance Institute (ITGI) indicates that ITG is an integral part of enterprise governance and consists of leadership, organizational structures, and processes. Moreover, Benaroch and Chernobai (2017) show that ITG consists of the leadership and organizational structures and processes that ensure that the organization's IT sustains and extends the organization's strategies. Furthermore, ITG is seen as the organizational capacity exercised by the board, executive management, and IT management to control the formulation and implementation of IT strategy and in this way ensure the fusion of business and IT (Neff et al., 2013). For constructing the measurement of ITG, this study refers to the prior literature. We will evaluate ITG from five dimensions: IT investment, IT technology resource, IT human resource (IHR), IT relationship resource, and IT capability. The index of IT investment (IV) and the sub-item index of IHR are the real value of the subjective firm, and other indexes are expressed on a 10-point Likert scale (where "1" means the level is lowest and "10" means the level is highest). The total indicator (ITG) is taken as a simple average of the sub-indicators (respectively, IV, IR, IHR, IRR, and IC). The specific descriptions are shown in **Table 2**.

For the evaluation of IT investment management, the secondary indexes are designed as "IT investment quantity" to reflect the total investment directly used to construct IT during the investigation period. "The total investment proportion of IT budget" reflects the investment intensity of

enterprises in IT projects. "The ratio of IT assets to total assets" reflects whether an enterprise is a general enterprise or an IT enterprise. For the evaluation of IT technology resources, this paper chooses the performance and load degree of ISs, including "technical performance of independent system," "overall function of integrated system," and "network bearing capacity and output load." The above indicators can comprehensively reflect the capability provided by the enterprise's technical resources. For the IT talent indicators, this paper makes a comprehensive survey from the aspect of IHRs, including "the proportion of IT staff in the total staff," "the salary level of IT staff," "the technical level of IT staff," and "the business level of IT staff." It basically covers the structure of the IT department and examines the comprehensive business and financial capabilities. The evaluation of "IT relationship resource" can examine the ability of various departments to cooperate and share resources in the process of ITG, and its level can reflect the environment and foundation of enterprise ITG. Finally, the measure of IT capacity is measured from three comprehensive indicators, namely "IT planning ability," "information utilization ability," and "system maintenance ability." The IT planning capability represents the planning capability and coverage of the technical framework for the strategy. The operational capability of IT is the implementation carrier of IT planning, and enterprises should have operational organization, job responsibilities, and the capability of system maintenance. IT maintenance ability is the safeguard of administrative effect. The maintenance capability is vital to daily practice, thus, it should also be taken as the evaluation index.

3.2.3 Firm performance

Since the mid-1990s, the SASAC has emerged as a key institution governing firm ownership in China (Wang et al., 2012), SASAC started to use enterprise value-added (EVA) assessment as a performance measure on central enterprises in 2010, to guide the way how enterprises develop. Considering that the subject of the questionnaire is not all listed companies, non-listed companies are also included in the sample, and these firms are affiliated with the SASAC. Therefore, referring to the SASAC's firm performance evaluation method and according to Du et al. (2012), we construct evaluation indexes from the three dimensions of financial ability, development ability, and social responsibility. The specific definition is shown in **Table 3**.

We weighted firm performance from three dimensions, financial ability indicators include return on total assets and total asset turnover, with a total weight of 30%. The development capability indexes are used to evaluate the growth of an enterprise, including the solvency guarantee ability, sustainability ability, and the company scale expansion, accounting for 50% of the weight. Social responsibility measures a company's responsibility for environmental protection, public welfare, and other aspects of social expenditure. This index accounts for 20% of the weight.

3.3 Model design

In order to explore the association among ITG, IS, and firm performance using the regression model, according to Hypothesis H1's each sub-hypothesis and Hypothesis H2, the empirical model was constructed as follows:

$$GP = \beta_0 + \beta_1 SI1 + \beta_2 SI2 + \beta_3 SI3 + \beta_4 DOA + \beta_5 SIZE + \beta_6 \triangle GP + YEAR + INDUSTRY + \varepsilon \quad \text{Model (1)}$$

$$GP = \beta_0 + \beta_1 ITG + \beta_2 DOA + \beta_3 SIZE + \beta_4 \triangle GP + YEAR + INDUSTRY + \varepsilon \quad \text{Model (2)}$$

In the above equation, GP represents the dependent variable firm performance, ITG represents the level of ITG, and SI was represented by the sub-item indexes in order to further explore the different aspects of IS. SI1 represents the IS formulation level, SI2 represents the intensity of IS process supervision, and SI3 represents the capability of IS result maintenance. According to Wu et al. (2015), firm size is usually treated as one of the antecedents to organizational performance, and we model it as a control variable directly affecting firm performance as our main focus. Size is the scale of the enterprise, measured by the logarithm of the number of employees (Wu et al., 2015). We include DOA, which is measured as the leverage of firms as total liabilities to total asset ratio (Mahmood et al., 2019). We also consider the potential influence that IT investment and enterprise performance are mutually influenced, as firms with better performance in the past may have more resources to devote to IT function (Turedi, 2020), thus, we controlled ΔGP. We also include the year (Year) which is a dummy variable to control the implicit influence that can change over time. Xue et al. (2012) also found that the impact of IT asset portfolios on organizational efficiency varies in different industrial environments. We control industries classified by the China Securities Regulatory Commission industry in 2012. Furthermore, due to the limited variable measures that can be obtained from the questionnaire, we do some efforts when selecting the sample of the questionnaire. For example, as mentioned above, the respondents of the questionnaire survey are the companies that mention ITG in their corporate governance documents, in order to control more corporate characteristics that affect firm performance. β_0 is a constant term. ε is the error term. In Model (1), if β_2, β_3, and β_4 are significant, Hypothesis H1 and the sub-hypotheses are supported. β_1 represents the impact of ITG on corporate

TABLE 3 Enterprise performance evaluation index.

Evaluation dimensions	Evaluation index	Weight (%)	Calculation formula
Financial capacity	Return on total assets	20	(Net profit + income tax + financial expenses)/Total annual assets
	Total asset turnover	10	Operating revenue/Total annual assets
Development capacity	Earnings multiples	10	(Net profit + income tax + financial expenses)/Financial expenses
	Technology investment proportion	20	Total technical expenses/Operating revenue
	Growth rate of operating revenue	20	Revenue growth of current year/Revenue of last year
Social responsibility	Social expenditure ratio	20	Social expenditure/Operating revenue

performance. In Model (2), if β_1 is significant, Hypothesis H2 is supported.

To test the Hypothesis H3 and to explore whether ITG has a positive moderating effect on the relationship between IS and firm performance, the empirical model was constructed as follows:

$$\begin{aligned} GP = {} & \beta_0 + \beta_1 ITG + \beta_2 SI1 + \beta_3 SI2 + \beta_4 SI3 + \beta_5 SI4 \\ & + \beta_6 ITG \times SI1 + \beta_7 ITG \times SI2 + \beta_8 ITG \times SI3 \\ & + \beta_9 DOA + \beta_{10} SIZE + \beta_{11} \triangle GP + YEAR \\ & + INDUSTRY + \varepsilon \end{aligned} \quad \text{Model(3)}$$

In Model (3), the control variables are as same as in Model (1). ITG × IS1, ITG × IS2, and ITG × IS3 are interactive terms. If β_2–β_5 are significant, coefficients of estimation are positive, and if β_6–β_8 are significant, coefficients are also positive; it indicates that ITG has a positive moderating effect on the relationship between IS and firm performance. Hence, Hypothesis H3 is supported.

3.4 Index reliability and validity analysis

In order to measure the reliability of the evaluation system, the method is to calculate the Cronbach reliability coefficient. The formula is as follows:

$$Cro\alpha = \frac{K}{K-1}\left(1 - \frac{\sum S_i^2}{S^2}\right) \quad (1)$$

where K represents the number of scale questions; $\sum S_i^2$ is the sum of the variance of items in the scale; S^2 is the variance of the items added together.

According to the estimation experience of previous scholars, the closer the Cronbach coefficient is to 1, the higher the representativeness and stability of the scale and the more reliable the evaluation value will be. When the coefficient exceeds 0.9, the scale is considered to have high internal reliability. If it is between 0.8 and 0.9, it could be accepted in a higher range. If it is between 0.7 and 0.8, the scale design is defective. However, it also has a certain reference value. If the Cronbach coefficient is less than 0.7, the scale design is unsuccessful and needs to be redesigned. The formula shows that the Cronbach coefficient will increase with the increase in K-value. In order to achieve objective results, more indicators are designed for the same project. According to the above-mentioned formula, the Cronbach coefficient can be calculated to be 0.839, which is in an acceptable high range, so the scale system is reliable and stable.

We also do some effort to verify the validity of this questionnaire. First, in the process of developing this questionnaire, the professionals in the China Association of Chief Financial Officers from different provinces help us to assess the items of the questionnaire. We have revised the statement of the corresponding question according to the opinions of experts. Second, we tested the face validity of the test. The enterprises taking the questionnaire test were asked to participate in a study at the beginning of the test to complete a short questionnaire regarding the face validity of the test. The test results show that the questionnaire is valid.

In order to further investigate how IS and ITG affect firm performance, we follow the research framework as is shown in the elements of **Figure 1**. As shown in **Figure 1**, IS is divided into three categories according to the implementation process, which are represented by SI1, SI2, and SI3, respectively. ITG is decomposed into five categories, represented by IV, IR, IC, IHR, and IRR, respectively. At the same time, the relations are shown. For example, H1 represents the impact of IS on enterprise performance. H2 represents the impact of ITG on enterprise performance. H3 represents the synergy impact of ITG and IS on enterprise performance.

4 Empirical analysis

4.1 Descriptive statistics

The mean, minimum value, maximum value, and standard deviation of each study variable are shown in **Table 4**. The mean return on total assets in the sample is 3.21%, which is close to the statistics of Chinese firms' return on total assets in the previous study (Giannetti et al., 2021). SI has a mean of 3.68, which indicates that the sample firms' implementation of the IS level is not mature enough. ITG has a mean of 4.06, and the

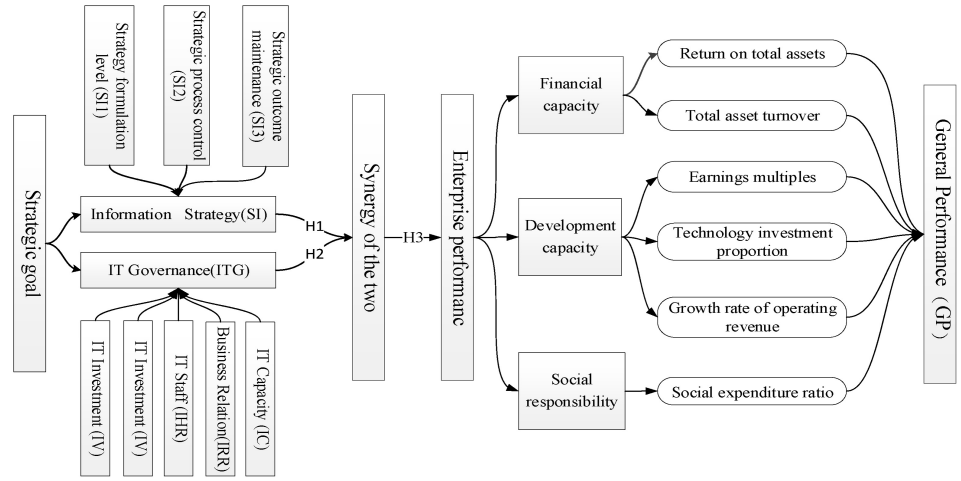


FIGURE 1
A research framework.

standard deviation is 0.31, which means that the difference in ITG level among sample firms is less.

4.2 Correlations

First, we conduct a correlation analysis on the hypothesis and the correlation results are shown in Table 5. There is a significant positive correlation between IS formulation level (SI1) and firm performance (r = 0.807, p < 0.01), between IS process supervision (SI2) and firm performance (r = 0.850, p < 0.01), and between IS result maintenance (SI3) and firm performance (r = 0.836, p < 0.01). This finding suggests that IS has a positive impact on firm performance. The IS made by the top management has a positive effect on firm performance and the IS which has a clear implementation path and reasonable control has a strong effect on firm performance. The IT background of board members also has a positive effect on firm performance. In other words, the three sub-hypotheses of H1a–H1b are preliminarily supported.

Second, this paper mainly focuses on the indicators of ITG which cover a firm's IT investment, IT technology resources, IHRs, IT relationship resources, and IT execution ability. The correlation results between ITG and firm performance are shown in Table 6. There is a significant positive correlation between information governance (ITG) and firm performance (r = 0.911, p < 0.01). Therefore, Hypothesis H2 is supported. It shows that technology, talent, business integration, and technical ability can promote enterprise performance.

TABLE 4 Descriptive statistics.

	Min	Max	Mean	SD
General performance	0.37	4.91	1.88	1.17
Return on total assets	0.16%	11.90%	3.21%	0.03
Total asset turnover	0.05	0.86	0.2816	0.29
Earnings multiples	1.49	30.79	8.7000	7.77
Technology investment proportion	4.62%	909.39%	142.58%	2.20
Growth rate of operating revenue	21.78%	87.79%	58.23%	0.20
Social expenditure ratio	0.01%	4.07%	0.68%	0.01
IT governance (ITG)	3.54	4.63	4.06	0.31
IT investment (IV)	0.08	0.30	0.18	0.07
IT technology resource (IR)	3.17	4.32	3.81	0.24
IT human resource (IHR)	3.01	4.82	3.86	0.63
IT relationship resource (IRR)	3.86	4.74	4.19	0.27
IT capability (IC)	3.33	4.67	3.92	0.34
Effect of information strategy (SI)	2.69	5.07	3.68	0.73
Information strategy formulation level (SI1)	2.47	4.60	3.32	0.63
Information strategy process supervision (SI2)	2.33	4.33	3.18	0.63
Information strategy results maintenance (SI3)	3.27	6.27	4.55	0.92

TABLE 5 Information strategy and corporation performance.

Index		Strategy level (SI1)	Process supervision (SI2)	Results maintenance (SI3)
General performance (GP)	Pearson correlation	0.807**	0.850**	0.836**
	Significance (bilateral)	0.001	0.002	0.000
	N	317	317	317

**Correlation is significant at the 0.01 level (two-tailed).

TABLE 6 Information strategy, IT governance, and corporation performance.

		General performance (GP)	Information strategy (SI)	IT governance (ITG)	SI × ITG
GP	Pearson correlation	1	0.910**	0.911**	0.9430**
	Significance (2-tailed)	0.000	0.000	0.000	0.000
	N	317	317	317	317
SI	Pearson correlation	0.910**	1	0.961**	0.999**
	Significance (2-tailed)	0.000	0.000	0.000	0.000
	N	317	317	317	317
ITG	Pearson correlation	0.911**	0.961**	1	0.999**
	Significance (2-tailed)	0.001	0.000	0.000	0.000
	N	317	317	317	317
SI × ITG	Pearson correlation	0.943**	0.999**	0.999**	1
	Significance (2-tailed)	0.000	0.001	0.000	0.000
	N	317	317	317	317

**Correlation is significant at the 0.01 level (two-tailed).

Third, we also conduct a correlation analysis to analyze the moderating role of ITG. We conduct a correlation analysis on IS, ITG, and corporation performance. The results are shown in Table 6. We mainly focus on the moderating role of ITG on the impact of IS on firm performance; the cross-multiplication term of SI and ITG is used to express the impact of the joint action of IS and ITG on enterprise performance. There is a significant positive correlation between the interaction term (SI × ITG) and firm performance ($r = 0.943$, $p < 0.01$), and this correlation is significant at the 1% level, with a positive correlation and a correlation coefficient above 0.9. Therefore, this result shows that ITG can strengthen the positive relationship between ITG and firm performance. Hypothesis H3 is preliminarily supported.

4.3 Regression analysis

4.3.1 Information strategy and firm performance

Through the correlation analysis of enterprise performance (GP), SI, and ITG in the sample, the correlation coefficient is found to be greater than 0.8. Although it can be explained that they have a correlation in statistical attributes, it is impossible to deeply understand the causal relationship behind them and the specific performance of various factors. Therefore, the ordinary least square linear (OLS) regression model is used in this paper to reveal the causal relationship and logical cause between variables.

We conduct an empirical analysis on Hypotheses H1a–H1b and the regression results are shown in Table 7. Model (1) shows that the coefficients of SI1, SI2, and SI3 are 0.357, 0.430, and 0.625 and are significant at the 1% level or 5% level. This finding suggests that the IS made by the top management, an IS that has a clear implementation path and IT background of board members, has a positive effect on firm performance, which means that the information strategy has a positive effect on firm performance. Therefore, Hypothesis H1 and three sub-hypotheses are supported.

4.3.2 Information technology governance and firm performance

We conduct regression analysis according to Model (2) to test the relationship between ITG and IT performance, and the results are shown in Table 7. In the first column of the regression results, the result shows that ITG has a significant positive impact on firm performance, and the coefficient of ITG is 0.723 and significant at the 5% level. This result shows that ITG contributes to firm performance positively. Our empirical regression results support Hypothesis 2.

4.3.3 The synergistic effects of information technology governance

A regulatory regression Model (3) is constructed to further test the regulatory effect of ITG on the relationship between IS and corporation performance. We add the interactive term of explanatory variable and moderator variable in the regression. By testing the coefficient of the interaction term, we can test the moderating role of ITG. The results are shown in Table 8. The coefficients of the interaction items between ITG and IS formulation level (SI1) and information result maintenance capability (SI3) are 1.032, 0.109, and 1.625 and are significant at the 1, 10, and 5% levels. The coefficient of the ITG is 0.436 and significant at the 5% level. It shows ITG has a moderating effect on IS. The higher the strategic level and the more attention to the maintenance of strategic results and the later assessment, the more conducive to improving enterprise performance. Hypothesis 3 is supported.

4.3.4 Discuss the endogenous

The relationship between the implementation level of IS and firm performance or the relationship between ITG and

TABLE 7 Information strategy and firm performance.

Variables	Dependent variable: Firm performance
	Model 1
C	0.961*** (0.017)
IS1	0.357*** (0.131)
IS2	0.430** (0.195)
IS3	0.625** (0.303)
DOA	−0.104** (0.052)
Size	−0.310** (0.132)
ΔGP	−0.042* (0.024)
Year	Yes
Industry	Yes
R^2	0.208
N	317
F	18.547***

The data are the SPSS processing output of the questionnaire data. ***Significant at the level of 0.01. **Significant at the level of 0.05. *Significant at the level of 0.10. The numbers in parentheses are standard errors.

TABLE 8 Information strategy, IT governance, and firm performance.

Variables	Dependent variable: Firm performance	
	Model 2	Model 3
C	0.626*** (0.170)	0.725*** (0.216)
ITG	0.723** (0.313)	0.436** (0.220)
IS1		0.329** (0.149)
IS2		0.235* (0.139)
IS3		0.316** (0.158)
ITG × IS1		1.032*** (0.357)
ITG × IS2		0.109* (0.063)
ITG × IS3		1.625** (0.674)
DOA	−0.098** (0.045)	−0.102** (0.051)
Size	−0.234* (0.134)	−0.36** (0.181)
ΔGP	−0.019** (0.009)	−0.021** (0.009)
Year	Yes	Yes
Industry	Yes	Yes
R^2	0.226	0.224
N	317	317
F	19.225***	16.013**

The data are the SPSS processing output of the questionnaire data. ***Significant at the level of 0.01. **Significant at the level of 0.05. *Significant at the level of 0.10. The numbers in parentheses are standard errors.

firm performance may not be caused by the impact of IS or ITG, but the other important omitted variables or they have a mutual cause-and-effect relationship. Specifically, we take some efforts to alleviate omitted variables problems. First, in terms of avoiding omitted variable bias, some considerations have been made in the early stage of the study. On the one hand, we randomly selected the interviewee from the alternative companies which are restricted to having certain characteristics. According to the above description of the selection process of respondents for the questionnaire, we randomly select the firm in which corporate governance documents have some ITG-related statements. The inclusion of ITG in corporate governance documents indicates that these companies have a deeper understanding of the concepts of IT investment, ITG, informatization, and so on. Therefore, these companies have more similar characteristics in IT. We randomly selected subjects from these representative firms and handed out questionnaires anonymously. We try to use this method to mitigate the problem of important omitted variables. On the other hand, in the stage of questionnaire design, we seek the opinions of authoritative experts to ensure the comprehensiveness of the index system, and then we can avoid the negative impact of omitted variables in the model design. Second, we conducted interviews with enterprise professionals to confirm whether there is a mutual cause-and-effect relationship between investment in IT (including IT investment, the formulation of informatization strategy, and setting up a corresponding IT department in the organization) and firm performance. In fact, in the first survey, we focused on the problem. Those surveyed believe that if IT investment does not contribute to enterprise performance, the companies will not continue to invest for 3 or even 6 years in IT even if they have more strength. Therefore, the empirical results are credible. It shows that IT investment and enterprise performance are not cause-and-effect relationships.

5 Discussion

5.1 Conclusion

Through the above theoretical analysis and empirical discussion, the following conclusions and enlightenment are obtained. First, when the total amount of information resource input of an organization remains unchanged, different levels of IS have different impacts on firm performance. The higher the level of strategy formulation, the greater the impact on enterprise performance. The principle of "top management" is more suitable for the information work of modern enterprises. Second, the higher the level of ITG, the more conducive to the implementation of IS excluding the control factors such as enterprise scale and organizational structure. The IS approved by the company's top management ensures the authority of investment from the organizational level. However, it also needs to be bound with the process participation of the units and functional departments. The synergy effect between IS and ITG has been well verified in the above empirical processes. Third, in view of the important role of ITG in ensuring the implementation of IS, it is suggested that some board members of IT enterprises should have corresponding IT backgrounds,

and the companies should add ITG departments and posts to ensure the effective management of core IT resources.

5.2 Theoretical contributions

Information technology investment, ITG behavior, and IS in the era of the digital economy have become a hot research topics. Existing research mainly discusses the impact of IT investment and ITG on firm performance. However, the impact of IS on firms' performance is still rare, and how ITG and IS influence firm performance has not caught attention. Therefore, the theoretical contributions of this study are mainly in two aspects: First, existing research mainly focuses on the relationship between IT investment and strategic alignment (Sabherwal et al., 2019) or discusses the mediation of strategic alignment in mechanisms that ITG affects organizational performance (Wu et al., 2015), prior literature not discussed separately from the aspect of IS perspective. This study constructs the measure of IS and tests the impact of IS on firm performance which extends prior research on firm information strategies. Furthermore, we further explore the influence of IS on enterprise performance from different aspects. This paper focuses on the IS of enterprises, which provides a certain reference value for the application of IS research in organizational performance management. Second, this paper emphasizes the interaction effect of ITG and IS on firm performance from the perspective of strategy formulation and control efficiency instead of considering the impact of the two elements on corporate performance separately as most previous studies do. Although most studies related to ITG have explored the relationship between ITG and organizational behavior (Wu et al., 2015; Tawafak et al., 2020), the study still rarely discusses the moderating role of ITG, and we find that ITG can strengthen the positive relationship between IS and firm performance by monitoring strategy formulation and improving control efficiency. This paper provides empirical evidence for corporate governance from the perspective of ITG. Third, this study adopts the questionnaire data from Chinese enterprises to empirically construct the measure of the IS and ITG. This provides a reference for further discussion of the economic consequences of IS and ITG.

5.3 Practical contributions

This study provides new insights and solutions for promoting the firm performance. Strategy is the target setting for an enterprise to develop and maintain its competitive advantage. As an important factor in the enterprise's development, the impact of ITG and IS can play a vital role in improving firm performance. First, with the development of the digital economy and dramatic changes in organizational structure, IT investment performance becoming more and more important, as a strategic goal and means to achieve IT performance, the formulation of IS and ITG is vital for firms to realize high quality development. Second, this paper shows that the IT background of board members has a positive effect on firm performance. Organizations should adjust the structure of management teams according to the development of the IT environment, so as to play a beneficial role in improving enterprise performance. The strategy formulated by the senior management of the organization can ensure the effective implementation of the strategy that inspires the formulation of an IS that is more suitable for the way from "top to bottom" in the strategic decision hierarchy. Third, this paper shows that the higher the level of ITG, the more benefits to IS's implementation effect. ITG is helpful for the company to realize the goal of IS. Therefore, the importance of ITG should be fully realized when improving the level of corporate governance. Overall, the realization of IS needs the coordination of technology and management wisdom, and at the same time, the idea of system planning should be infiltrated into organizational behavior in order to help organizations achieve the state of modern information management.

5.4 Limitations and future research

This study also has some limitations. First, since ITG and IS are difficult to measure, data can only be obtained through questionnaires. In the future, the ITG index and IS implementation index should be further constructed to conduct a more in-depth large-sample test on this issue. Second, this paper only examines the moderating effects of IS and ITG on firm performance but has not yet examined the mechanism of how IS affects firm performance. Future research should explore this issue, so as to open the black box of how corporate strategy setting and corporate governance affect firm performance and firm's sustainable development in the digital time, so as to provide a useful reference for enterprises to improve IT performance and financial performance.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent from the patients/participants or patients/participants

legal guardian/next of kin was not required to participate in this study in accordance with the national legislation and the institutional requirements.

Author contributions

FW and JL conceived and designed the experiments, collected and interpreted the data. XZ analyzed the data, examined and critically contributed to the study, and finally approved the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This study was supported by grants from the Social Science Foundation of Beijing (Grant No. 19GLB020), the National Social Science Foundation of China (Grant No. 21AGL005), and the Academic Newcomer Program of Capital University of Economics and Business (Grant No. 2021XSXR02).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Ahmad, F., and Arshad, N. H. (2014). Value delivery of information technology investment: a conceptual framework. *Int. J. Comp. Theory Eng.* 6, 150–154. doi: 10.7763/IJCTE.2014.V6.854

Al Romaihi, N. A., and Hamdan, A. (2021). The relationship between IT governance and firm performance: a review of the literature. *Sustainable Finance Digital. Role Technol.* 487, 299–312. doi: 10.1007/978-3-031-08954-1

Alfalah, A. A., Muneer, S., and Hussain, M. (2022). An empirical investigation of firm performance through corporate governance and information technology investment with mediating role of corporate social responsibility: evidence from Saudi Arabia telecommunication sector. *Front. Psychol.* 13:959406. doi: 10.3389/fpsyg.2022.959406

Ali, S., Green, P., Robb, A., and Masli, A. (2022). Governing information technology (IT) investment: a contingency perspective on organization's IT investment goals. *Australian J. Manag.* 47, 3–23. doi: 10.1177/03128962211009578

Astuti, E. (2018). The influence of information technology strategy and management support to the internal business process, competitive advantage, financial and non-financial performance of the company. *Int. J. Web Inform. Systems* 14, 317–333. doi: 10.1108/ijwis-11-2017-0079

Aubert, B. A., and Rivard, S. (2020). "The outsourcing of IT governance," in *Information Systems Outsourcing*, eds R. Hirschheim, A. Heinzl, and J. Dibbern (Berlin: Springer), 43–59.

Benaroch, M., and Chernobai, A. (2017). Operational IT failures, IT value-destruction, and board-level IT governance changes. *MIS Quarterly* 41, 729–762.

Chatzoglou, P. D., Diamantidis, A. D., Vraimaki, E., Vranakis, S. K., and Kourtidis, D. A. (2011). Aligning IT, strategic orientation and organizational structure. *Bus. Process Manag. J.* 17, 663–687. doi: 10.1108/14637151111149474

Cohen, J. F. (2008). Contextual determinants and performance implications of information systems strategy planning within South African firms. *Inform. Manag.* 45, 547–555. doi: 10.1016/j.im.2008.09.001

Cohen, J. F., and Olsen, K. (2013). The impacts of complementary information technology resources on the service-profit chain and competitive performance of South African hospitality firms. *Int. J. Hospital. Manag.* 34, 245–254. doi: 10.1016/j.ijhm.2013.04.005

Dairo, M., Adekola, J., Apostolopoulos, C., and Tsaramirsis, G. (2021). Benchmarking strategic alignment of business and IT strategies: opportunities, risks, challenges and solutions. *Int. J. Inform. Technol.* 13, 2191–2197. doi: 10.1007/s41870-021-00815-7

De Haes, S., Huygh, T., Joshi, A., and Caluwe, L. (2019). National corporate governance codes and it governance transparency in annual reports. *J. Global Inform. Manag.* 27, 91–118. doi: 10.4018/jgim.2019100105

Dong, S., Yang, L., Shao, X., Zhong, Y., Li, Y., and Qiao, P. (2021). How can channel information strategy promote sales by combining ICT and blockchain? evidence from the agricultural sector. *J. Cleaner Production* 299:126857. doi: 10.1016/j.jclepro.2021.126857

Du, F., Tang, G., and Young, S. M. (2012). Influence activities and favoritism in subjective performance evaluation: evidence from Chinese state-owned enterprises. *Account. Rev.* 87, 1555–1588. doi: 10.2139/ssrn.1445333

Giannetti, M., Liao, G., You, J., and Yu, X. (2021). The externalities of corruption: evidence from entrepreneurial firms in China. *Rev. Finance* 25, 629–667.

Ho, W. R., Tsolakis, N., Dawes, T., Dora, M., and Kumar, M. (2022). *A Digital Strategy Development Framework for Supply Chains*. Piscataway, NJ: IEEE, doi: 10.1109/TEM.2021.3131605

Huang, R., Zmud, R. W., and Price, R. L. (2010). Influencing the effectiveness of IT governance practices through steering committees and communication policies. *Eur. J. Inform. Systems* 19, 288–302. doi: 10.1057/ejis.2010.16

Ilmudeen, A., Bao, Y. K., and Zhang, P. L. (2022). Investigating the mediating effect of Business-IT alignment between management of IT investment and firm performance. *Inform. Systems Manag.* 1, 1–12. doi: 10.1080/10580530.2022.2107740

Jeffers, P. I., Muhanna, W. A., and Nault, B. R. (2008). Information technology and process performance: an empirical investigation of the interaction between IT and non-IT resources. *Decision Sci.* 39, 703–735. doi: 10.1111/j.1540-5915.2008.00209.x

Ji, P., Yan, X., and Yu, G. (2019). The impact of information technology investment on enterprise financial performance in China. *Chinese Manag. Stud.* 14, 529–542. doi: 10.1108/cms-04-2019-0123

Lainhart, J. W. IV (2000). Why IT governance is a top management issue. *J. Corp. Account. Finance* 11, 33–40.

Lederer, A. L., Mirchandani, D. A., and Sims, K. (1997). The link between information strategy and electronic commerce. *J. Organ. Comp. Electron. Commerce* 7, 17–34. doi: 10.1207/s15327744joce0701_2

Lee, S., and Kim, S. H. (2006). A lag effect of IT investment on firm performance. *Inform. Resources Manag. J.* 19, 43–69. doi: 10.4018/irmj.2006010103

Lin, H.-F. (2022). IT resources and quality attributes: the impact on electronic green supply chain management implementation and performance. *Technol. Soc.* 68:101833. doi: 10.1016/j.techsoc.2021.101833

Lunardi, G. L., Becker, J. L., Gastaud Macada, A. C., and Dolci, P. C. (2014). The impact of adopting IT governance on financial performance: an empirical analysis among Brazilian firms. *Int. J. Account. Inform. Systems* 15, 66–81. doi: 10.1016/j.accinf.2013.02.001

Mahmood, F., Han, D., Ali, N., Mubeen, R., and Shahzad, U. (2019). Moderating effects of firm size and leverage on the working capital finance-profitability relationship: Evidence from China. *Sustainability* 11:2029. doi: 10.3390/su11072029

Matta, M., Cavusoglu, H., and Benbasat, I. (2022). Understanding the board's involvement in information technology governance. *Inform. Systems Manag.* 1, 1–21. doi: 10.1080/10580530.2022.2074580

Mikalef, P., and Pateli, A. (2017). Information technology-enabled dynamic capabilities and their indirect effect on competitive performance: findings from PLS-SEM and fsQCA. *J. Bus. Res.* 70, 1–16. doi: 10.1016/j.jbusres.2016.09.004

Mikalef, P., Pateli, A., and Van de Wetering, R. (2021). IT architecture flexibility and IT governance decentralisation as drivers of IT-enabled dynamic capabilities and competitive performance: the moderating effect of the external environment. *Eur. J. Inform. Systems* 30, 512–540. doi: 10.1080/0960085x.2020.1808541

Mithas, S., Tafti, A., Bardhan, I., and Goh, J. M. (2012). Information technology and firm profitability: mechanisms and empirical evidence. *MIS Quarterly* 36, 205–224. doi: 10.2307/41410414

Mohamad, A., Zainuddin, Y., Alam, N., and Kendall, G. (2017). Does decentralized decision making increase company performance through its information technology infrastructure investment? *Int. J. Account. Inform. Systems* 27, 1–15. doi: 10.1016/j.accinf.2017.09.001

Neff, A. A., Hamel, F., Herz, T. P., Uebernickel, F., and Brenner, W. (2013). "IT governance in multi-business organizations: performance impacts and levers from processes, structures, and relational mechanisms," in *Paper Presented at the 2013 46th Hawaii International Conference on System Sciences*, (Hawaii).

Nolan, R. L. (1973). Managing the computer resource: a stage hypothesis. *Commun. ACM* 16, 399–405. doi: 10.1145/362280.362284

Park, K. O. (2018). The relationship between BPR strategy and change management for the sustainable implementation of ERP: an information orientation perspective. *Sustainability* 10:3080. doi: 10.3390/su10093080

Prasad, A., and Green, P. (2015). Governing cloud computing services: reconsideration of IT governance structures. *Int. J. Account. Inform. Systems* 19, 45–58. doi: 10.1016/j.accinf.2015.11.004

Prasad, A., Green, P., and Heales, J. (2012). On IT governance structures and their effectiveness in collaborative organizational structures. *Int. J. Account. Inform. Systems* 13, 199–220. doi: 10.1016/j.accinf.2012.06.005

Proksch, D., Rosin, A. F., Stubner, S., and Pinkwart, A. (2021). The influence of a digital strategy on the digitalization of new ventures: the mediating effect of digital capabilities and a digital culture. *J. Small Bus. Manag.* 1–29. doi: 10.1080/00472778.2021.1883036

Ravichandran, T., and Liu, Y. (2011). Environmental factors, managerial processes, and information technology investment strategies. *Decision Sci.* 42, 537–574. doi: 10.1111/j.1540-5915.2011.00323.x

Raymond, L., Bergeron, F., Croteau, A. M., and Uwizeyemungu, S. (2019). Determinants and outcomes of IT governance in manufacturing SMEs: a strategic IT management perspective. *Int. J. Account. Inform. Systems* 35:100422. doi: 10.1016/j.accinf.2019.07.001

Ru, J., Shi, R., and Zhang, J. (2018). When does a supply chain member benefit from vendor-managed inventory? *Product. Operat. Manag.* 27, 807–821. doi: 10.1111/poms.12828

Sabherwal, R., Sabherwal, S., Havakhor, T., and Steelman, Z. (2019). How does strategic alignment affect firm performance? the roles of information technology investment and environmental uncertainty. *MIS Quarterly* 43, 453–474. doi: 10.25300/MISQ/2019/13626

Schlosser, F., Beimborn, D., Weitzel, T., and Wagner, H.-T. (2015). Achieving social alignment between business and IT–an empirical evaluation of the efficacy of IT governance mechanisms. *J. Inform. Technol.* 30, 119–135. doi: 10.1057/jit.2015.2

Sheppard, J. (1990). The strategic management of IT investment decisions: a research note. *Br. J. Manag.* 1, 171–181. doi: 10.1111/j.1467-8551.1990.tb00005.x

Shreeve, B., Hallett, J., Edwards, M., Ramokapane, K. M., Atkins, R., and Rashid, A. (2022). The best laid plans or lack thereof: security decision-making of different stakeholder groups. *IEEE Trans. Software Eng.* 48, 1515–1528. doi: 10.1109/tse.2020.3023735

Simonsson, M., Johnson, P., and Ekstedt, M. (2010). The effect of IT governance maturity on IT governance performance. *Inform. Systems Manag.* 27, 10–24. doi: 10.1080/10580530903455106

Soh, C., and Markus, M. L. (1995). "How IT creates business value: a process theory synthesis," in *Proceedings of the International Conference on Information Systems 1995 Proceedings*, (Whistler). doi: 10.1002/jps.24311

Takeda, F., Takeda, K., Takemura, T., and Ueda, R. (2021). The impact of information technology investment announcements on the market value of the Japanese regional banks. *Finance Res. Lett.* 41:101811. doi: 10.1016/j.frl.2020.101811

Tawafak, R., Romli, A., Malik, S., and Shakir, M. (2020). IT Governance impact on academic performance development. *Int. J. Emerg. Technol. Learn. (IJET)* 15, 73–85. doi: 10.3991/ijet.v15i18.15367

Taylor, H., Dillon, S., and Van Wingen, M. (2010). Focus and diversity in information systems research: meeting the dual demands of a healthy applied discipline. *MIS Quarterly* 34, 647–667. doi: 10.2307/25750699

Turedi, S. (2020). The interactive effect of board monitoring and chief information officer presence on information technology investment. *Inform. Systems Manag.* 37, 113–123. doi: 10.1080/10580530.2019.1696589

Wang, J., Guthrie, D., and Xiao, Z. (2012). The rise of SASAC: asset management, ownership concentration, and firm performance in China's capital markets. *Manag. Organ. Rev.* 8, 253–281. doi: 10.1111/j.1740-8784.2011.00236.x

Williams, C. K., and Karahanna, E. (2013). Causal explanation in the coordinating process: a critical realist case study of federated IT governance structures. *MIS Quarterly* 37, 933–964. doi: 10.25300/MISQ/2013/37.3.12

Witra, W. P. P., and Subriadi, A. P. (2022). Gender and information technology (IT) investment decision-making. *Proc. Comp. Sci.* 197, 583–590.

Wolf, C., and Floyd, S. W. (2017). Strategic planning research: toward a theory-driven agenda. *J. Manag.* 43, 1754–1788. doi: 10.1177/0149206313478185

Wu, S. P.-J., Straub, D. W., and Liang, T.-P. (2015). How information technology governance mechanisms and strategic alignment influence organizational performance: insights from a matched survey of business and it managers. *MIS Quarterly* 39, 497–518. doi: 10.25300/misq/2015/39.2.10

Xue, L., Ray, G., and Sambamurthy, V. (2012). Efficiency or innovation: how do industry environments moderate the effects of firms' IT asset portfolios? *MIS Quarterly* 36, 509–528. doi: 10.2307/41703465

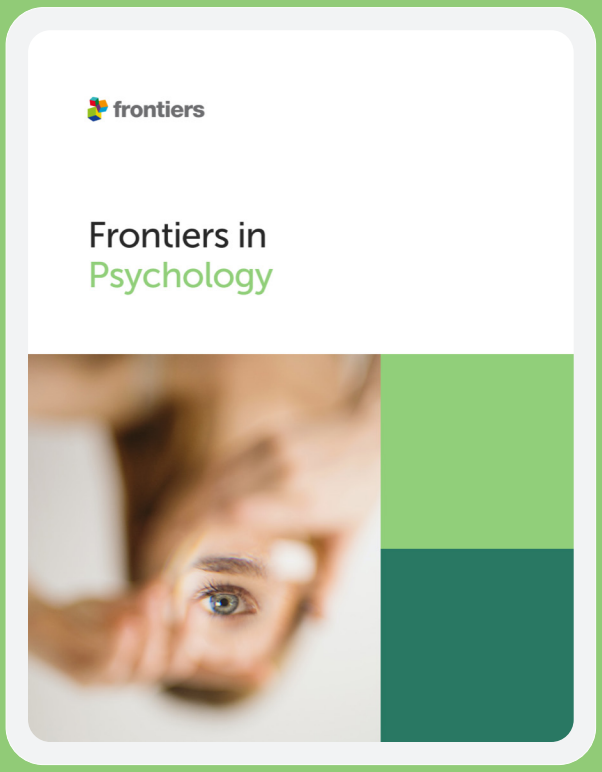
Frontiers in Psychology

Paving the way for a greater understanding of human behavior

The most cited journal in its field, exploring psychological sciences - from clinical research to cognitive science, from imaging studies to human factors, and from animal cognition to social psychology.

Discover the latest Research Topics

See more →



Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact