frontiers | Research Topics

The important role of the early school years for reading, writing and math development: assessment and intervention at school entry

Edited by
Christoph Weber, Martin Schöfl, Susanne Seifert
and Natalie Förster

Published in
Frontiers in Education
Frontiers in Psychology

ISSN 1664-8714
ISBN 978-2-8325-6989-4
DOI 10.3389/978-2-8325-6989-4

Generative AI statement
Any alternative text (Alt text) provided alongside figures in the articles in this ebook has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

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

The important role of the early school years for reading, writing and math development: assessment and intervention at school entry

Topic editors
Christoph Weber — University of Education Upper Austria, Austria
Martin Schöfl — University of Education Upper Austria, Austria
Susanne Seifert — University of Graz, Austria
Natalie Förster — University of Wuppertal, Germany

Citation
Weber, C., Schöfl, M., Seifert, S., Förster, N., eds. (2025). *The important role of the early school years for reading, writing and math development: assessment and intervention at school entry*. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-8325-6989-4

Table of contents

frontiers | Frontiers in Education

TYPE Editorial
PUBLISHED 28 October 2025
DOI 10.3389/feduc.2025.1720577

Check for updates

OPEN ACCESS

EDITED AND REVIEWED BY
Mary Frances Rice,
University of New Mexico, United States

*CORRESPONDENCE
Christoph Weber
✉ christoph.weber@ph-ooe.at

RECEIVED 08 October 2025
ACCEPTED 14 October 2025
PUBLISHED 28 October 2025

CITATION
Schöfl M, Weber C, Seifert S and Förster N (2025) Editorial: The important role of the early school years for reading, writing and math development: assessment and intervention at school entry.
Front. Educ. 10:1720577.
doi: 10.3389/feduc.2025.1720577

Editorial: The important role of the early school years for reading, writing and math development: assessment and intervention at school entry

Martin Schöfl[1,2], Christoph Weber[1]*, Susanne Seifert[3] and Natalie Förster[4]

[1]University of Education Upper Austria, Linz, Austria, [2]Johannes Kepler Universitat Linz, Linz, Austria, [3]Universitat Graz, Graz, Austria, [4]Bergische Universitat Wuppertal, Wuppertal, Germany

KEYWORDS

data-based decision making (DBDM), early literacy, early numeracy, screening, implementation fidelity, assessment mode effects, equity, kindergarten

Editorial on the Research Topic
The important role of the early school years for reading, writing and math development: assessment and intervention at school entry

Introduction

Children start formal schooling in reading, writing, and mathematics with substantially varying learning prerequisites (Davis-Kean et al., 2022; Merrell, 2023). While some children barely know any letters or numbers, others can already read, write, and calculate with great fluency. Lack of letter or number knowledge at entry does not mean that children lack foundational experiences with literacy and numeracy. They bring rich, practice-based resources, including ways of engaging with language, quantity, and space that are often undervalued in school (Moll et al., 1992; Nunes and Bryant, 2015; Rosa and Orey, 2016). Children with less favorable learning conditions are at higher risk for disadvantageous learning outcomes and lack educational success (Sparks et al., 2014; Stanley et al., 2018). As early literacy and numeracy, as well as their precursors, are associated with family background variables, these unequal starting conditions set the stage for later social and ethnic-cultural educational disparities (e.g., Bradley and Corwyn, 2002; Davis-Kean et al., 2022; Skopek and Passaretta, 2021).

Addressing these differences between children requires instruction that builds on sound diagnostic information on their learning prerequisites. The availability of objective, reliable, and valid information regarding students' learning prerequisites should enable teachers to implement targeted interventions and adapt their instruction to meet the individual needs of young learners (Mandinach, 2012). This should result in optimal learning support, compensate for unequal starting conditions, and counteract educational inequalities. These assumptions correspond to the concept of data-based decision making (DBDM; Mandinach, 2012; Souvignier and Förster, 2025). In short, DBDM is a process that covers several central steps, starting with the assessment of data on students' learning prerequisites, then analyzing data to derive instructional interventions, and finally evaluating the interventions' effects. The 14 articles of this Research Topic address aspects of the DBDM process around the time of school entry (from the last preschool year to grade 2), primarily focusing on assessment and assessment-based interventions.

Generally, assessments must meet standard quality criteria (objectivity, reliability, and validity) to be effectively used for instructional adaptations. In this context, the question arises as to what aspects of learning prerequisites should be assessed to optimize the prediction of future learning outcomes. Moreover, assessments must be economical so that they can be used in the classroom context. Therefore, technology-enhanced assessments that facilitate test administration, scoring, and reporting of results are increasingly being developed. Furthermore, the simple provision of assessment data may not be sufficient to ensure that teachers can derive instructional interventions with manageable effort (Hebbecker et al., 2022). Finally, interventions should be effective. This editorial summarizes the findings of the articles in this Research Topic along the DBDM steps: (1) providing basic knowledge for the development of assessments, (2) providing evidence that assessments are sound, (3) providing evidence on how to support teachers in interpreting and using data for instruction, and (4) providing evidence on effective interventions. Unless otherwise justified by content, we present the studies in each of the following sections according to the domains of reading, writing, and mathematics.

Basic knowledge for the development of assessments

The following articles lay the conceptual groundwork for the design of assessments by examining which early skills predict later outcomes and how these indicators are associated with sociodemographic factors.

In their article "*Morphological awareness predicts reading comprehension in first-grade students*," Sparks and Metsala focus on a specific predictor of reading comprehension. Their study of English-speaking children from Canada contributes to the understanding of the role of morphological awareness in predicting reading success. They investigated whether morphological awareness uniquely contributes to the prediction of reading after controlling for oral language skills. Their findings indicate that morphological awareness assessed in grade 1 accounts for unique variance in reading comprehension in grade 2.

Sigmund et al. investigate "*Early cognitive predictors of spelling and reading in German-speaking children*." Their article contributes to the understanding of early spelling and reading skills predictors by explicitly investigating whether predictors assessed before school entry show differential predictive validity for reading and spelling skills in grade 1. Specifically, their results based on a German sample underscore the importance of phonological processing, intelligence, and letter knowledge for predicting grade 1 spelling after accounting for preschool spelling and grade 1 reading skills. In contrast, only rapid automatized naming (RAN) emerged as a significant predictor of reading once preschool reading and grade 1 spelling were controlled for.

The article by Banfi et al. titled "*Longitudinal predictors of reading and arithmetic at different attainment levels*" stands out by addressing both reading and mathematics, examining not only children at risk for academic failure but also high achievers. Their study advances the understanding of the cognitive foundations underlying the association between reading and arithmetic fluency. Drawing on longitudinal data of over 300 students in Grades 1, 2, and 3 from the UK and Austria, the authors identify both domain-general predictors (e.g., nonverbal IQ) and domain-specific predictors (e.g., phonological awareness for reading and magnitude processing for arithmetic). Notably, most predictors proved to be equally important for high and low achievers.

Concluding this section, the study by Schöfl, Weber et al. titled "*Language abilities and phonological information processing mediate the association of spelling with bilingualism and socioeconomic status*" is placed last because it extends the focus to the socioeconomic and migration-related factors underlying differences in spelling achievement and examines how these effects are mediated by language abilities and phonological information processing. Using an Austrian sample, the authors found that language skills assessed at the very beginning of grade 1 mediate the association between migration-related bilingual status and spelling skills in grade 2, and that the link between socioeconomic status and spelling is partly mediated by language skills and phonological information processing. These findings highlight the importance of promoting language skills and phonological information processing in preschool and at the beginning of primary school to reduce social and ethnic-cultural disparities.

Evidence that assessments are sound

The following articles examine the reliability and validity of assessments, as well as the potential effects of test administration mode, to determine whether the studied instruments provide consistent and accurate measures on which instructional decisions can be based.

Schöfl, Steinmair et al. report in their article "*Using an app-based screening tool to predict deficits in written word spelling at school entry*" results on the validity and diagnostic accuracy of an efficient and cost-free screening tool for identifying spelling problems in community school settings. They used an Austrian sample to examine the predictive validity of different precursors of word spelling, measured at school entry, for predicting spelling difficulties at the end of grade 2. Their results indicate that a screening consisting of the two measures letter knowledge (an aspect of phonological information processing) and sentence repetition (a measure of grammatical knowledge) has an acceptable diagnostic accuracy.

Chatzaki et al. explore the assessment of early mathematical competencies through a playful, game-based approach in their article "*Exploring the potential of a game-based preschool assessment of mathematical competencies using a linear board game called the House of Numbers*." Drawing on data from a sample of German preschoolers, they examine the reliability and validity of the game-based test. Their results suggest that the House of Numbers is a valid and reliable tool for assessing the mathematical abilities of preschoolers. In addition, they found that the game-based test is an appropriate screening tool with high diagnostic validity for identifying children at risk of low math performance.

Cruz et al. also focus on assessing mathematics skills in the preschool years. In their article "*Assessment of math abilities before school entry: a tool development*," using a Portuguese sample, the authors provide support for the reliability and validity of a developed screening tool that covers several precursors of mathematics skills assessed during the last year before school entry. They reported high internal consistency of the screening measures and found support for the validity of the screening for predicting grade 1 mathematics achievement.

While these studies validate or develop new assessment instruments for early literacy and numeracy, Seifert et al. focus on a methodological aspect that cuts across assessment context, namely, the potential influence of test administration mode on assessment outcomes. In their article "*Unveiling mode effects in grade 1 vocabulary assessment: the intriguing influence of test mode*," the authors investigate whether the digital administration of a well-established paper-and-pencil test of receptive vocabulary produces comparable results to the traditional format. In their study with Austrian grade 1 students, they found positive effects of the digital mode, which highlights the need to consider the test mode in the field application of the test and emphasizes the need for further investigations to better understand mode effects. They conclude that ensuring comparability across paper and digital forms will require cross-mode score linking and measurement invariance testing.

Evidence on how to support teachers in interpreting and using data for instruction

The following articles address questions regarding how assessment results can be reported and how they can be effectively aligned with instruction to support data-based decision-making in education.

Extending beyond the validation of assessments, the contributions by Stuhr et al. and Ufer et al. also explore how assessment outcomes might be translated into instructional guidance. In their article "*Exploration of latent early literacy profiles in German kindergarten children using a newly developed app*," Stuhr et al. use app-assessed data from a German preschool sample to identify latent literacy profiles based on precursors of reading, such as word awareness, phonological awareness, and alphabet knowledge, and suggest profile-specific recommendations on dialogic book reading. This should facilitate the transformation from assessment results to instructional activities. They emphasize the need for experimental and implementation-focused studies to evaluate the effectiveness of the tailored recommendations.

In their article, "*I have three more than you, you have three less than me? Levels of flexibility in dealing with additive situations*," Ufer et al. assess the flexibility in dealing with additive situations in a sample of second and third-grade German students and develop a model distinguishing discrete levels of flexibility. The authors discuss that assigning students to levels may help teachers make use of assessment results, especially when these are linked to suggestions for instructional adaptations.

Providing evidence on effective interventions

Finally, a total of four articles in this Research Topic deal with the next step of the DBDM process, namely, interventions and what is necessary for interventions to be effective.

In their article "*A brief research report on the efficacy of a RAN training in elementary school age children*," Wolfsperger and Mayer evaluate an adaptive, software-based RAN training program designed to increase naming speed for children with RAN deficits. Using a German sample of 57 children, their study showed that the naming speed for letters and numbers increased during the training phase. Additionally, they found evidence of a transfer effect of the training on reading speed for words and reading comprehension on the word and sentence levels.

The article by Ennemoser et al. titled "*From developmental theory to effective training: long-term and transfer effects of promoting the quantity–to–number word linkage in grade 1 students at risk for mathematical difficulties*" focuses on a targeted intervention based on an assessment of risk for mathematical learning problems. In detail, they investigate long-term and transfer effects of a quantity–to–number word linkage training in a German sample of grade 1 students. They found that the training had transfer effects on mathematics achievement up to 15 months after the training ended. Therefore, their research provides support for an effective and successful alignment of assessment and intervention.

Powell et al. also focus on an intervention to promote mathematics skills, but address the preschool age. In their article "*Investigation of the initial feasibility of extended mathematics read-alouds used by kindergarten teachers*," they investigate the feasibility and effects of a mathematics-focused read-aloud program in a US sample. They found no significant effect of the training, but they found evidence that implementation fidelity is significantly associated with outcomes. This highlights the need for high program fidelity when specific programs are used within instructional interventions to achieve desired effects.

While the preceding studies focus on evaluating the effects of specific interventions, Lembke et al. address a central prerequisite for their sustainable impact in practice, namely, implementation fidelity. In their article "*Data-based individualization in early writing: the importance and measurement of implementation fidelity*," they focus on implementing a data-based individualization approach to promote early writing. Using a US sample, the authors monitor the implementation of DBDM, describe challenges for teachers in implementing DBDM in their classroom, and provide evidence on how to improve implementation fidelity to ensure that evidence-based practices can take effect in everyday school life.

Conclusion

A close examination of the articles in this Research Topic reveals three messages that are largely aligned. Firstly, it is essential to assess early and precisely. Brief, reliable screeners of phonological, letter, language, and early numerical competencies can forecast later literacy and numeracy, and identify both risk and high attainment. Secondly, the data must be made actionable in an instructional context. While technological advancements and profiling tools can enhance the feasibility of this process, it is imperative to obtain explicit validation for the translation of results into concrete teaching methodologies. Thirdly, alignment of assessment and intervention within sustained data-based decision-making cycles has been demonstrated to engender promising long-term and transfer effects. Implementing well-specified instructional responses with high fidelity and prioritizing language and phonological appears to support for socio-demographically disadvantaged learners has been shown to engender equity gains. Complementarily, assessment and intervention should be guided by an assets-based stance that values children's existing literacies and everyday numerical–spatial competencies, ensuring that data use supports responsive instruction rather than reproducing deficit framings (Moll et al., 1992; Nunes and Bryant, 2015).

To facilitate the implementation of this approach, we advocate integrating low-stakes screening procedures at school entry into the regular curriculum, in accordance with the principles of the Multi-Tiered System of Supports (MTSS) and Response to Intervention (RTI) frameworks (Fuchs and Fuchs, 2006; McIntosh and Goodman, 2016). MTSS/RTI represents a tiered model of early identification and support that aligns universal screening, progress monitoring, and data-based instructional adaptations. These screening procedures should be complemented by the implementation of DBDM routines that follow the same tiered principles of the MTSS/RTI framework and are supported by professional learning opportunities focusing on the interpretation and use of assessment data.

In the context of research, three priorities emerge: firstly, the experimental evaluation of profile-based recommendations; secondly, the establishment of cross-mode equivalence in assessments; and thirdly, the design of studies that explicitly target implementation fidelity and equity impacts. Overall, the reviewed studies collectively illustrate that there is a clear and direct path from early assessment to measurable learning gains in the early school years, provided that the loop from measurement to instruction is effectively closed.

We invite future research to combine scalable assessments with empirically validated instructional routines to demonstrate durable gains at school entry.

Author contributions

MS: Writing – review & editing, Writing – original draft. CW: Writing – original draft, Writing – review & editing. SS: Writing – original draft, Writing – review & editing. NF: Writing – review & editing, Writing – original draft.

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.

Generative AI statement

The author(s) declare that no Gen AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

References

Bradley, R. H., and Corwyn, R. F. (2002). Socioeconomic status and child development. *Annu. Rev. Psychol.* 53, 371–399. doi: 10.1146/annurev.psych.53.100901.135233

Davis-Kean, P. E., Domina, T., Kuhfeld, M., Ellis, A., and Gershoff, E. T. (2022). It matters how you start: early numeracy mastery predicts high school math course-taking and college attendance. *Infant Child Dev.* 31:e2281. doi: 10.1002/icd.2281

Fuchs, D., and Fuchs, L. S. (2006), Introduction to response to intervention: what, why, and how valid is it? *Read. Res. Q.* 41, 93–99. doi: 10.1598/RRQ.41.1.4

Hebbecker, K., Förster, N., Forthmann, B., and Souvignier, E. (2022). Data-based decision-making in schools: examining the process and effects of teacher support. *J. Educ. Psychol.* 114, 1695–1721. doi: 10.1037/edu0000530

Mandinach, E. B. (2012). A perfect time for data use: using data-driven decision making to inform practice. *Educ. Psychol.* 47, 71–85. doi: 10.1080/00461520.2012.667064

McIntosh, K., and Goodman, S. (2016). *Integrated Multi-Tiered Systems of* Support*: Blending RTI and PBIS*. New York, NY: Guilford Press.

Merrell, C. (2023). "Children's developmental levels at the start of school," in *The First Year at School: An International Perspective*, eds. P. Tymms, T. Bartholo, S. Howie, E. Kardanova, M. C. Koslinski, C. Merrell, and H. Wildy (New York, NY; Abingdon: Springer International Publishing), 125–147.

Moll, L. C., Amanti, C., Neff, D., and González, N. (1992). Funds of knowledge. *Theory Pract.* 31, 132–141. doi: 10.1080/00405849209543534

Nunes, T., and Bryant, P. (2015). *Children's Reading and Mathematics: The Cognitive Foundations*. Oxford: Wiley.

Rosa, M., and Orey, D. C. (2016). State of the art in ethnomathematics. *Rev. Latinoam. Etnomatemát.* 9, 1–9. doi: 10.1007/978-3-319-30120-4_3

Skopek, J., and Passaretta, G. (2021). Socioeconomic inequality in children's achievement from infancy to adolescence: the case of Germany. *Social Forces* 100, 86–112. doi: 10.1093/sf/soaa093

Souvignier, E., and Förster, N. (2025). "Data-based decision-making," in *Handbuch Bildungsarmut*, eds. G. Quenzel, K. Hurrelmann, J. Groß Ophoff, and C. Weber (Wiesbaden: Springer). doi: 10.1007/978-3-658-4469 8-7_39-1

Sparks, R. L., Patton, J., and Murdoch, A. (2014). Early reading success and its relationship to reading achievement and reading volume: replication of '10 years later'. *Read. Writ.* 27, 189–211. doi: 10.1007/s11145-013-9439-2

Stanley, C. T., Petscher, Y., and Catts, H. (2018). A longitudinal investigation of direct and indirect links between reading skills in kindergarten and reading comprehension in tenth grade. *Read. Writ.* 31, 133–153. doi: 10.1007/s11145-017-9777-6

frontiers | Frontiers in Education

TYPE Original Research
PUBLISHED 27 December 2023
DOI 10.3389/feduc.2023.1304688

Check for updates

OPEN ACCESS

EDITED BY
Susanne Seifert,
University of Graz, Austria

REVIEWED BY
Christoph Helm,
University of Teacher Education Zug, Switzerland
Reinhard Kargl,
Independent Researcher, Graz, Austria

*CORRESPONDENCE
Jamie L. Metsala
✉ Jamie.Metsala@msvu.ca

RECEIVED 02 October 2023
ACCEPTED 07 December 2023
PUBLISHED 27 December 2023

CITATION
Sparks E and Metsala JL (2023) Morphological awareness predicts reading comprehension in first grade students.
Front. Educ. 8:1304688.
doi: 10.3389/feduc.2023.1304688

Morphological awareness predicts reading comprehension in first grade students

Erin Sparks[1] and Jamie L. Metsala[2]*

[1]Department of Psychology and Neuroscience, Dalhousie University, Halifax, NS, Canada,
[2]Faculty of Education, Mount Saint Vincent University, Halifax, NS, Canada

Research examining a role for morphological awareness in first grade students' reading comprehension is scarce, although it is a well-established predictor for students in mid-to-late elementary school. One question that remains is whether morphological awareness explains unique variance in these young readers' comprehension after accounting for other oral language skills. In this longitudinal study, we assessed Grade 1 students' inflectional morphological awareness as a predictor of their concurrent ($n = 58$) and Grade 2 ($n = 55$) reading comprehension. When controlling for decoding and vocabulary, Grade 1 morphological awareness explained unique variance in concurrent and subsequent reading comprehension (4% and 5%, respectively). In novel analyses that controlled for decoding, vocabulary, and syntactic awareness, morphological awareness explained unique variance in Grade 2 reading comprehension (5%), but not in concurrent reading comprehension. This unique contribution only in second grade may be because decoding skills accounted for less of the overall variance in second than in first grade comprehension or due to the expectation that polymorphemic words are more frequent in second grade texts. Overall, morphological awareness emerged as the strongest oral language predictor in all models. These results support morphological awareness' relevance to reading comprehension from early in children's reading development and highlight the need for research to further explore the effects of targeting English morphological awareness with young students.

KEYWORDS

reading comprehension, reading development, morphological awareness, longitudinal research, early elementary

Introduction

Comprehending written texts is central to success in educational settings and to navigating the demands of a literate society (e.g., Murnane and Levy, 1996; Wilson, 2016). The foundations of skilled reading comprehension begin from the earliest days of children's reading acquisition—reading skills in Grade 1 predict reading comprehension and language abilities ten years later (Cunningham and Stanovich, 1997; Sparks et al., 2014), and preschool language skills are associated with reading comprehension into the elementary school years (Hjetland et al., 2017). Such longitudinal findings highlight the importance of a strong start in oral language and reading comprehension for young students and the need to understand the early phases of reading comprehension development. In the current study, we explore whether Grade 1 students' inflectional morphological awareness—a metalinguistic facility with units of meaning in oral language—predicts their concurrent and subsequent reading comprehension.

Comprehending what one reads refers to making meaning from the text—understanding, for example, the gist, message, information, or scene represented in a text. The act of comprehending connected sentences requires the reader to "…retrieve the meanings of individual words, compute the sense of each sentence, integrate the meanings of successive sentences, and incorporate background knowledge to construct a representation of the state of affairs described in the text" (Cain and Oakhill, 2009, p. 144). When we consider reading comprehension in beginning readers, word-level decoding stands out as a key contributor. A child who cannot decode the words in a text will be unable to comprehend that text, regardless of their oral language skills (Gough and Tunmer, 1986; Lonigan and Burgess, 2017), and thus decoding has traditionally been given primacy in the context of early reading comprehension (e.g., Hoover and Gough, 1990; Gough et al., 1996; Storch and Whitehurst, 2002). Yet even for young readers with somewhat simple texts, on top of decoding the written words they need to know what the words mean, understand the relations between those words in the sentences, and integrate what they are reading with what they know about the world. Indeed, multiple studies have confirmed that individual differences in oral language abilities predict variance in reading comprehension, beyond decoding, from early in children's reading development (e.g., Vellutino et al., 2007; Foorman et al., 2015; *cf.* Storch and Whitehurst, 2002). A question remains about the developmental time-frame for when individual oral language skills make unique contributions to comprehending texts. Answering this question addresses the practical problem of knowing which oral language skills may be candidates for targeted instruction in the earliest years in elementary classrooms. Both theoretical and practical issues highlight the need to understand the various oral language skills that are relevant to young students' developing reading comprehension.

In recent years, morphological awareness has received attention as potentially important in children's reading comprehension (see Levesque et al., 2021 for a review). Morphological awareness is one's ability to recognize and manipulate morphemes, which are the smallest meaningful units of language (Carlisle, 2000). English has a morphophonemic orthography – meaning that our spelling system represents both the sounds and morphemes in spoken English; this deep orthography can be contrasted with orthographies that are more consistent in representing the sounds in the spoken language (e.g., Finnish and Spanish). As explained by Moats (2020, p.8), in English "…the spellings of morphemes are often stable even when pronunciation varies in words with a common root" (e.g., spelling of *jumped* maintains the past-tense morpheme "ed" rather than representing the final /t/ sound). Morphological awareness has been proposed to influence literacy outcomes in multiple ways. According to the Reading Systems Framework (Perfetti and Stafura, 2014), morphology exists as part of the broader linguistic and lexical systems, feeding into text comprehension processes through direct and indirect means. More specifically, the Morphological Pathways Framework proposes that morphological awareness affects comprehension of text through both decoding of morphologically complex words and analysis of word meanings (Levesque et al., 2021). Within this framework, two aspects of morphology can be delineated—inflections and derivations. In English, inflectional morphology involves adding small word parts, suffixes and prefixes, to change the form of a word without altering its basic meaning or category. These changes typically indicate grammatical aspects such as tense, number, possession, or comparison (e.g., signaling the tense and aspect of a verb, as in *walked* or *walking*). Derivational morphology involves adding affixes to a base morpheme and changes the word's meaning or grammatical category (e.g., *walkable*). Both aspects of morphology are represented in the English writing system, allowing readers to make sense of polymorphemic words and determine their grammatical functions in sentences. English-speaking children show a clear awareness of inflections by school entry (e.g., Berko, 1958; de Villiers and de Villiers, 1973), and their derivational awareness shows a particular increase after age 8 (Anglin, 1993; Berninger et al., 2010). Both forms of morphological awareness improve across the elementary school years (Kuo and Anderson, 2006). Given its early development, we focus on children's inflectional morphological awareness in the current study.

Research with children in the mid- to late-elementary grades has demonstrated that morphological awareness is a predictor of reading comprehension (e.g., Carlisle, 2000; Deacon and Kirby, 2004; Cunningham and Carroll, 2015). This relationship persists after accounting for the effects of word reading (e.g., Deacon et al., 2014; Levesque et al., 2017; James et al., 2021), vocabulary (e.g., Kirby et al., 2012), and/or a general language measure (e.g., Kieffer et al., 2016; Metsala et al., 2021). Yet research examining morphological awareness and reading comprehension with younger students is sparse by comparison (and as reviewed below, some inconsistent findings have emerged; Carlisle, 1995; Carlisle and Fleming, 2003; Muter et al., 2004; Kirby et al., 2012; Kruk and Bergman, 2013; Apel and Henbest, 2016; James et al., 2021). Notably, the existing studies of morphological awareness as a predictor of first graders' reading comprehension have not simultaneously controlled for decoding, vocabulary, or syntactic awareness—skills that are each implicated in the reading comprehension of somewhat older students (e.g., Quinn et al., 2015; Deacon and Kieffer, 2018) and that relate to morphological awareness (e.g., Kuo and Anderson, 2006; Sparks and Deacon, 2015). In the current study, we address this gap by evaluating morphological awareness as a predictor of Grade 1 students' concurrent and later reading comprehension after first controlling for decoding and vocabulary, and then also controlling for syntactic awareness. We thus distinguish effects of morphological awareness from decoding and a set of related oral language skills.

Vocabulary has long been recognized as an important predictor of reading comprehension (e.g., Beck and McKeown, 1991; Quinn et al., 2015; *cf.* Ouellette and Beers, 2010) and is frequently included in studies that focus on morphological awareness. For instance, James et al. (2021) found that a comprehensive measure of morphological awareness predicted concurrent reading comprehension among 6- to 8-year-olds after controlling for vocabulary and word reading, among other controls (age; nonverbal reasoning; phonological awareness). Similarly, Apel and Henbest (2016) found that, among children in Grades 1–3, knowledge of affix meaning predicted concurrent reading comprehension beyond age, vocabulary, and phonological awareness. Kim (2023) found that morphological awareness in first grade, largely explained the shared variance between word reading and listening comprehension, both major contributors to reading comprehension. These concurrent findings suggest a relationship between early morphological awareness and reading comprehension. They serve as an impetus for further exploring this relationship among Grade 1 students specifically, while controlling for variables such as decoding and vocabulary skill.

Syntactic awareness is another oral language skill relevant to reading comprehension (e.g., Cain, 2007; Deacon and Kieffer, 2018). It captures children's metalinguistic ability to reflect on and manipulate syntax (Nagy, 2007) and may facilitate the parsing of sentences—a skill that appears important to establishing the meanings of texts (e.g., Sorenson-Duncan et al., 2021). Conceptually, syntactic awareness is a valuable skill to include when isolating the effects of morphological awareness on reading, as morphemes serve grammatical functions (e.g., signaling word class; contributing to subject–verb agreement in sentences). Research with somewhat older children suggests that effects of morphological awareness emerge when syntactic awareness and vocabulary are included as controls, uniquely predicting gains in reading comprehension among children in Grades 2 and 3 (Metsala et al., 2021; see Kieffer et al., 2016 for concurrent effects among older students; *cf.* Proctor et al., 2012).

Longitudinal studies have also examined beginning readers' morphological awareness as a predictor of subsequent reading comprehension. Early work found that Grade 1 students' morphological awareness predicted their later reading comprehension when considered as the sole predictor (Carlisle and Fleming, 2003) and when controlling for phonological awareness (Carlisle, 1995). Yet mixed findings have emerged in studies that include additional controls. For example, Kruk and Bergman (2013) found that first graders' morphological awareness—as assessed in a morphological generation task—predicted their Grade 3 reading comprehension beyond variance explained by phonological awareness, vocabulary, and word reading. This aligns with evidence from students in later elementary grades, for whom morphological awareness uniquely predicts gains in reading comprehension beyond vocabulary, word reading, and other relevant controls (e.g., Deacon et al., 2014). Together, this research suggests that morphological awareness' relationship with reading comprehension extends beyond children's knowledge of word meanings. Kirby et al. (2012) reported similar findings at Grades 2 and 3—morphological awareness (measured using a word analogy task with inflected and derived items) predicted Grade 3 reading comprehension beyond vocabulary, phonological awareness, and nonverbal ability. However, morphological awareness at Grade 1 did not uniquely predict students' later reading comprehension. The authors noted floor effects in Grade 1 students' performance on their morphological awareness task, which may have contributed to their nonsignificant results (Kirby et al., 2012). These sparse and conflicting longitudinal findings at Grade 1 point to a need for additional research into morphological awareness as a predictor of beginning reading comprehension beyond both decoding and vocabulary.

Whether morphological awareness predicts reading comprehension in beginning readers after controlling for syntactic awareness is an open question. One study to date has measured both syntactic and morphological awareness as predictors of beginning reading comprehension (Muter et al., 2004). This study—which started at children's school entry at age 4—combined the two constructs into a composite measure of grammatical awareness. The researchers found that kindergarten vocabulary and Grade 1 grammatical awareness predicted unique variance in reading comprehension at Grade 2 (controlling for word reading, phonological awareness, and letter knowledge). This provides evidence for the importance of these metalinguistic skills in predicting early reading comprehension, but the combined measure of grammatical awareness cannot isolate the effect of morphological awareness from that of syntactic awareness. We address this gap in the current study.

In summary, the current study builds on the small body of evidence around the relationship between morphological awareness and reading comprehension among English-speaking readers in Grade 1. Specifically, we examined whether children's Grade 1 morphological awareness would uniquely predict their concurrent reading comprehension and their reading comprehension one year later in Grade 2. We first assessed this question while controlling for decoding and vocabulary. Many prior studies with this age group have controlled for phonological awareness (Carlisle, 1995; Kirby et al., 2012; Apel and Henbest, 2016), but fewer have controlled for the word-level reading skill that are more proximally related to reading comprehension (but see Kruk and Bergman, 2013; James et al., 2021). By controlling for decoding and vocabulary, we isolated morphological awareness' effect on reading comprehension from two widely prioritized predictors of reading comprehension in this age group (e.g., García and Cain, 2014; Quinn et al., 2015). Next, we considered whether morphological awareness would predict concurrent and/or subsequent reading comprehension while controlling for decoding, vocabulary, and syntactic awareness, skills which have not been controlled together in studies of morphological awareness with this age group. We focused on inflectional morphological awareness using task formats that are suited to young children (Muter et al., 2004) to minimize the risk of floor effects in our key predictor (Kirby et al., 2012). Given past research with first-grade students (e.g., Carlisle, 1995; Muter et al., 2004; Kruk and Bergman, 2013), and similar research with older students (e.g., Kieffer et al., 2016; Metsala et al., 2021), we hypothesized that Grade 1 morphological awareness would predict a small, but unique and significant, amount of variance in concurrent (Grade 1) and subsequent (Grade 2) reading comprehension beyond both sets of controls.

Materials and Methods

Participants

At Time 1 (T1), we recruited 75 students from five schools in a largely suburban school district in eastern Canada. Of the 75 students recruited, 8 students were excluded from analyses because they were English language learners, according to parent report, and another 9 students were excluded due to a labor dispute which meant they were unavailable for the second session of T1. Thus, concurrent analyses at T1 included 58 participants (28 females and 30 males; mean age 6 yrs. 9 mos., $SD = 3.55$ months), all of whom were native English speakers. Approximately one year later, 55 of these students were available for the Time 2 (T2) follow-up in the winter of Grade 2 (28 females and 27 males; mean age 7 yrs. 8 mos., $SD = 3.26$ months; 5.1% attrition).

Students who attended each of the five schools were mostly from working- and middle-class families. According to school district data, families served by these five schools did not differ from the larger school district (about 90 elementary schools) in the proportion of households with post-secondary education or the proportion categorized within the lowest-income bracket (Ms = 49 and 7% across the school district, respectively).

Procedure

Trained research assistants conducted all testing in quiet rooms at the participants' schools. T1 testing involved two individual sessions in the winter of Grade 1, each lasting approximately 45 min. T2 testing occurred approximately one year later (mean interval between time points = 334 days; *SD* = 27.1). All testing was completed individually, aside from the Gates-MacGinitie Comprehension subtest at T2, which was administered in small groups of 2–3 students.

Instruments

Oral language and decoding measures were administered at T1; reading comprehension was measured at both T1 and T2. Table 1 shows maximum scores, descriptive statistics, and reliabilities for each task.

Vocabulary (T1)

Receptive vocabulary was measured with a modified version of the Peabody Picture Vocabulary Test (PPVT–IV; Dunn and Dunn, 2007) in which every third item was administered. All manualized procedures were otherwise followed. The shortened task maintains the structure of the original while reducing testing duration to minimize participants' missed class time. Similar shortened versions of the PPVT have been used previously (e.g., Deacon et al., 2014) and validated with children in Grade 1 (Deacon et al., 2013). Children are asked to point to one of four pictures representing the target spoken word (e.g., point to the ball; shoulder; meadow).

Syntactic awareness (T1)

Syntactic awareness was measured with 15 items that required sentence judgment and/or correction. The task was modelled after similar experimenter-constructed measures that assess children's judgment (e.g., Wulfeck, 1993; Geva and Farnia, 2012) and correction (e.g., Tunmer et al., 1988; Cain, 2007; Deacon and Kieffer, 2018) of sentences. All incorrect items involved word order errors, which is an important cue for English sentence comprehension (e.g., Wulfeck, 1993) and none involved morphological violations (*cf.* Geva and Farnia, 2012)—this ensures a focus on syntax. Prior research has shown that children in our study's age range can detect word-order violations, though not at the ceiling levels seen in adults (e.g., Wulfeck, 1993).

For 10 items, participants judged whether an orally-presented sentence was syntactically correct (e.g., correct: "*The goats run and the squirrels climb*," "*The teacher stays and Tim leaves*"; incorrect: "*The*

TABLE 1 Descriptive statistics and task reliabilities.

	Maximum possible score	Mean	*SD*	Reliability
Dependent variables				
Time 1 Reading comp: WJ-Achievement[a,g]				0.96[c]
Raw score	47	16.12	5.94	
Standard score	–	101.26	15.70	
Time 2 Reading comp: WJ Rdg Mastery[a,h]				0.94[c]
Raw score	38	12.42	3.98	
Standard score	–	99.27	13.82	
Time 2 Reading comp: Gates-MacGinitie[b]				0.93[d]
Raw score	48	32.04	10.01	
Standard score	–	48.52	8.48	
Time 1 – predictor variables				
Vocabulary	76	42.45	5.43	0.77[e]
Syntactic awareness	15	8.19	3.29	0.78[f]
Morphological awareness	20	12.75	3.47	0.78[f]
Time 1 – control variables				
Phonological decoding: word attack[a]				0.94[f]
Raw score	45	8.38	6.13	
Standard score	–	102.67	12.88	

[a]Norm referenced test with a mean of 100 and SD of 15.
[b]Norm referenced test with a mean of 50 and SD of 10.
[c]Publisher-reported split-half reliability coefficient.
[d]Publisher-reported Kuder–Richardson coefficient.
[e]Split half-reliability.
[f]Cronbach's alpha reliability coefficient.
[g]Woodcock-Johnson III Tests of Achievement.
[h]Woodcock Reading Mastery Tests – 3rd edition.

baby cries and laughs Jenny"; "*The dog sits and stands the person*"). The remaining five items involved sentence correction (e.g., "*The sisters teach and learn the brothers,*" which can be corrected to "*The sisters teach and the brothers learn.*"; for similar tasks, see Siegel and Ryan, 1988; Deacon and Kieffer, 2018). Both question formats included practice items with feedback.

Morphological awareness (T1)

The morphological awareness task assessed inflectional morphology. English-speaking children are adept at the use and manipulation of inflections by first grade (e.g., Berko, 1958; de Villiers and de Villiers, 1973), though individual differences in inflectional morphological awareness remain (Muter et al., 2004; Robertson and Deacon, 2019). By comparison, first grade students' understanding of derivational morphology is limited (Anglin, 1993; Kuo and Anderson, 2006). By focusing on inflectional morphology, we were able to assess metalinguistic awareness of morphological processes for which participants have a strong foundational knowledge and avoid the floor effects that have been seen in some prior work with this age group (Kirby et al., 2012).

Fifteen items, from the Word Structure subtest of the Clinical Evaluation of Language Fundamentals, 5th edition (CELF–5; Wiig et al., 2013), required students to complete a spoken sentence stem while referring to a picture (e.g., *Here is one child. Here are two* ____). Items captured the following range of inflectional processes: regular and irregular plural (e.g., cat; *cats* and child; *children,* respectively); past tense (e.g., cook: *cooked*); future tense (e.g., Here she is baking; tomorrow she will *bake*); third person singular (e.g., Here the girl plays; here the girl *skips*); superlative (e.g., This ball is bigger; this ball is the *biggest*); present progressive (e.g., Here the girl walks; here she is *walking*). Responses were scored as correct if they involved the appropriate morphological form for the prompt. The final five items involved correcting the errors in a spoken sentence and were included to emphasize metalinguistic awareness by requiring participants to manipulate morphemes. Errors were in subject–verb agreement (e.g., *The tree grow and the flowers dies;* correct responses: *The tree grows and the flowers die* or *The trees grow and the flower dies*). Both question formats included practice items with feedback.

To validate the separability of the three oral language measures, we ran a confirmatory factor analysis using item-level data to create latent factors for syntactic and morphological awareness (see also Deacon and Kieffer, 2018), and a single indicator of vocabulary (based on total raw scores). Wald tests confirmed that all resulting correlations between factors were significantly lower than 1.0. Specifically, the correlation between the latent syntactic and morphological awareness factors was 0.60 (Wald $\chi^2 = 16.6$, $df = 1$, $p < 0.001$), the correlation between morphological awareness and vocabulary was 0.36 (Wald $\chi^2 = 27.4$, $df = 1$, $p < 0.001$), and the correlation between syntactic awareness and vocabulary was 0.36 (Wald $\chi^2 = 28.1$, $df = 1$, $p < 0.001$). These results provide evidence that the oral language measures represent distinguishable skills for the purpose of our analyses.

Pseudoword decoding (T1)

Children completed the Word Attack subtest of the Woodcock-Johnson III Tests of Achievement (Woodcock et al., 2001). This measure involved reading a list of pronounceable pseudowords of increasing difficulty, assessing participants' ability to apply decoding skills to pronounce unfamiliar words. For example, the child would see and read pseudowords like dat, jat, heg, chrobe, and derine. Testing was discontinued after six consecutive errors.

Passage comprehension (T1 and T2)

Participants completed the Passage Comprehension subtest of the Woodcock-Johnson III Tests of Achievement (Woodcock et al., 2001) at T1 and the Woodcock Reading Mastery Tests, 3rd edition, at T2 (Woodcock, 2011). In these tasks, children read short passages to themselves and provided a missing word in each. For young children, items start out as simple sentences with an accompanying picture and the test progresses to longer passages without pictures. The child is required to read the sentence or passage and to supply a word that fits in the blank. Initial items are akin to "The boys are swimming in the ________" or "The book was placed on the _________." The task was administered and scored according to standardized procedures, and testing was discontinued after six consecutive errors.

Gates-MacGinitie reading comprehension (T2)

As an additional measure of T2 reading comprehension, participants completed the Comprehension subtest of the Gates-MacGinitie Reading Tests, 2nd Canadian edition (Level C, Form 3; MacGinitie and MacGinitie, 1992). The grade two students had 35 min to read silently through a series of stories, answering 3–5 multiple choice questions after each story. There are 14 stories in all, ranging from 4 to 11 sentences in length.

In our longitudinal analyses, children's T2 reading comprehension was captured by factor scores reflecting their combined performance on the T2 Woodcock and Gates-MacGinitie tasks. The factor scores were created using a least squares regression approach.

Analytic procedures

We used hierarchical regression analyses to address whether first grade morphological awareness, entered as the final step, predicted unique variance in concurrent and in later reading comprehension. This analytic strategy allowed an increasingly strict set of predictor variables to be accounted for to test our hypotheses. In our first set of hierarchical regression, pseudoword decoding and vocabulary were entered before morphological awareness. The third regression examined if morphological awareness predicted second grade reading comprehension after accounting for first grade reading comprehension. In the second set of hierarchical regressions, syntactic awareness was also entered into the equation before morphological awareness.

Prior to analysis, we examined variable distributions and checked data for outliers. All variables were normally distributed except for pseudoword decoding, which we square-root transformed to address positive skew (Tabachnick and Fidell, 2007). This yielded normally distributed residuals for all models reported below. There were no univariate outliers (defined as 3.0 times the median absolute distance; Leys et al., 2013) or multivariate outliers (based on Mahalanobis-minimum covariance determinant (MCD) distance computed from a central 75% of observations, using $p = 0.001$ as a cutoff for outlier detection; Leys et al., 2018).

One participant was missing a score on the morphological awareness task (0.29% of data). We dealt with this data point through multiple imputation (*mice* package; van Buuren and

Groothuis-Oudshoorn, 2011), imputing 100 data sets using predictive mean matching. The imputation models included all predictor variables used in the regressions reported below (decoding; vocabulary; syntactic awareness; morphological awareness). As a robustness check, we conducted a complete case analysis by excluding the participant with missing data. Doing so yielded the same pattern of regression results that we obtained when using the imputed data. We therefore report results from the imputed data below.

Results

Descriptive statistics

Table 1 shows descriptive data for the reading and oral language measures collected at T1 and T2. All mean standard scores on the normed reading measures fell within the average range (see Table 1), suggesting the sample is representative in their reading achievement. We used raw scores in our analyses (aside from the transformed decoding scores), and Table 2 shows the zero-order correlations between the major variables in this study. All three oral language skills were significantly correlated with reading comprehension at each time point, with the exception that syntactic awareness was not significantly correlated with T2 reading comprehension. Collinearity diagnostics for the regressions reported below indicate that multicollinearity was not an issue in our models (all variance inflation factors <1.68; tolerances >0.59; Freund et al., 2006). Given labor disruptions within the schools, the number of participants was somewhat lower than initially planned; however, the number meets the lower limits, with the total number of participants equal to or exceeding 50 plus the number of predictor variables (Harris, 1985; Green, 1991). Furthermore, the number of participants is similar to previous investigations examining the contribution of young students' oral language skills to reading outcomes (e.g., Lam et al., 2011). We return to this issue in the section on limitations.

Does morphological awareness predict reading comprehension beyond decoding and vocabulary?

We first examined whether Grade 1 morphological awareness predicted concurrent (T1) and/or longitudinal (T2) reading comprehension after controlling for decoding and vocabulary (both measured at T1). Results from these hierarchical regression analyses are shown in Table 3.

Taken together, the predictors accounted for 65.1% of the variance in concurrent (T1) reading comprehension. When entered at Step 1, decoding accounted for 58.8% of the variance ($p < 0.001$). Vocabulary did not account for significant unique variance at Step 2 ($p = 0.083$). When entered at Step 3, morphological awareness predicted an additional 4.1% unique variance ($p = 0.015$). In the final equation, Grade 1 decoding and morphological awareness were both significant predictors of unique variance in concurrent reading comprehension.

A similar pattern emerged in our longitudinal analysis predicting T2 reading comprehension. Together, T1 decoding, vocabulary, and morphological awareness accounted for 42.2% of the variance in reading comprehension one year later. Decoding accounted for 33.3% of the variance at Step 1 ($p < 0.001$), vocabulary's contribution was nonsignificant at Step 2 ($p = 0.069$), and morphological awareness predicted 4.7% unique variance at Step 3 ($p = 0.046$). In the final equation, Grade 1 decoding and morphological awareness were both significant predictors of Grade 2 reading comprehension.

Since morphological awareness was a significant contributor to grade 2 reading comprehension, we tested whether this remained significant when accounting for the autoregressive variable. This final analysis helps to determine if morphological awareness contributed to gains in reading comprehension that students made across this one-year period. As can be seen in the final regression reported in Table 3, morphological awareness was not a significant predictor of gains in reading comprehension.

Does morphological awareness predict reading comprehension beyond decoding, vocabulary, and syntactic awareness?

Next, we added syntactic awareness to our models as an additional control, allowing for a more stringent analysis of Grade 1 morphological awareness as a unique predictor of concurrent (T1) and subsequent (T2) reading comprehension. Results from these hierarchical regression analyses are shown in Table 4. The first two steps of each regression were identical to those of the models reported above.

In the concurrent analysis, syntactic awareness predicted a unique and significant 3.1% of the variance in T1 reading comprehension when entered at Step 3 ($p = 0.034$). At Step 4, morphological awareness did not make a unique contribution to T1 reading comprehension ($p = 0.058$). Together, the Grade 1 predictors accounted for 66.5% of

TABLE 2 Zero-order correlations among oral language and reading variables.

		1	2	3	4	5
1.	Reading comprehension (Time 1)	–				
2.	Reading comprehension (Time 2)	0.80***	–			
3.	Morphological awareness	0.58***	0.52***	–		
4.	Vocabulary	0.37**	0.37**	0.42***	–	
5.	Syntactic awareness	0.28*	0.13	0.39**	0.31*	–
6.	Phonological decoding	0.77***	0.58***	0.48***	0.30*	0.09

*$p < 0.05$; **$p < 0.01$; ***$p \leq 0.001$.

TABLE 3 Hierarchical regressions predicting Time 1 concurrent ($n = 58$) and Time 2 longitudinal ($n = 55$) reading comprehension, controlling for decoding and vocabulary.

Step	Predictor	ΔR^2	Initial β [95% CI]	Final β [95% CI]
Concurrent analysis: predicting Time 1 reading comprehension				
1	Pseudoword decoding	0.588	0.767*** [0.599, 0.935]	0.629*** [0.449, 0.809]
2	Vocabulary	0.022	0.156^ [−0.017, 0.329]	0.076 [−0.101, 0.253]
3	Morphological awareness	0.041	0.245* [0.054, 0.436]	0.245* [0.054, 0.436]
Longitudinal analysis: predicting Time 2 reading comprehension				
1	Pseudoword decoding	0.333	0.583*** [0.361, 0.805]	0.418*** [0.176, 0.659]
2	Vocabulary	0.041	0.211^ [−0.012, 0.435]	0.132 [−0.099, 0.363]
3	Morphological awareness	0.047	0.257* [0.010, 0.503]	0.257* [0.010, 0.503]
Longitudinal analysis: predicting gains in Time 2 reading comprehension				
1	T1 reading comp	0.645	0.803*** [0.669, 1.017]	0.760*** [0.583, 1.012]
2	Morphological awareness	0.004	0.074 [−0.133, 0.284]	0.074 [−0.133, 0.284]

*$p < 0.05$; ***$p < 0.001$; ^$p < 0.10$ (ns).

TABLE 4 Hierarchical regressions predicting Time 1 concurrent ($n = 58$) and Time 2 longitudinal ($n = 55$) reading comprehension, controlling for decoding, vocabulary, and syntactic awareness.

Step	Predictor	ΔR^2	Initial β [95% CI]	Final β [95% CI]
Concurrent analysis: predicting Time 1 reading comprehension				
1	Pseudoword decoding	0.588	0.767*** [0.599, 0.935]	0.644*** [0.468, 0.826]
2	Vocabulary	0.022	0.156^ [−0.017, 0.329]	0.054 [−0.127, 0.229]
3	Syntactic awareness	0.031	0.187* [0.019, 0.355]	0.130 [−0.043, 0.305]
4	Morphological awareness	0.024	0.196^ [−0.003, 0.395]	0.196^ [−0.003, 0.395]
Longitudinal analysis: predicting Time 2 reading comprehension				
1	Pseudoword decoding	0.333	0.583*** [0.361, 0.805]	0.406** [0.157, 0.655]
2	Vocabulary	0.041	0.211^ [−0.012, 0.435]	0.143 [−0.095, 0.382]
3	Syntactic awareness	0.001	0.039 [−0.194, 0.271]	−0.054 [−0.297, 0.189]
4	Morphological awareness	0.048	0.279* [0.011, 0.546]	0.279* [0.011, 0.546]

*$p < 0.05$; ***$p \leq 0.001$; ^$p < 0.10$ (ns).

the variance in concurrent reading comprehension; in the final equation only decoding made a significant unique contribution.

Turning to the longitudinal analysis, syntactic awareness did not uniquely predict T2 reading comprehension when entered at Step 3 ($p = 0.744$), whereas morphological awareness predicted a significant 4.8% unique variance when entered at Step 4 ($p = 0.047$). Together, the Grade 1 predictors accounted for 42.4% of the variance in reading comprehension one year later; in the final equation, decoding and morphological awareness made significant unique contributions.

Discussion

The relationship between morphological awareness and reading comprehension has been widely explored among children in the mid-elementary grades and beyond (e.g., Deacon et al., 2014; Kieffer et al., 2016; Levesque et al., 2017), but research with younger readers is less prevalent. Our study adds to this sparse literature by evaluating morphological awareness as a predictor of Grade 1 students' concurrent reading comprehension, and prospective reading comprehension one year later, after controlling for decoding and two important oral language skills—vocabulary and syntactic awareness. We found that Grade 1 students' morphological awareness predicted unique variance in both concurrent and subsequent reading comprehension beyond decoding and vocabulary, two widely prioritized aspects of early reading skill (e.g., Garcia and Cain, 2014; Quinn et al., 2015). In analyses that also controlled for syntactic awareness, Grade 1 morphological awareness predicted significant unique variance in children's reading comprehension one year later, but not in their concurrent reading comprehension. Furthermore, morphological awareness was not a significant predictor of gains in reading comprehension across the year. Together, these findings suggest that individual differences in morphological awareness are relevant to reading comprehension from early in children's formal reading instruction.

Our finding that morphological awareness uniquely predicted reading comprehension when controlling for decoding and vocabulary builds on the small body of similar research with beginning readers. The concurrent results align with those of two prior studies (Apel and Henbest, 2016; James et al., 2021). Both studies grouped together children in Grades 1–3 and found that morphological awareness predicted unique variance in children's concurrent reading comprehension (beyond age, phonological awareness, and vocabulary, and in James et al., 2021, word reading and nonverbal reasoning). The current study affirms this pattern of results while focusing on a narrower age range, increasing our confidence that morphological awareness uniquely predicts reading comprehension among Grade 1 readers specifically.

Past longitudinal results with this age range have been more conflicting. Kirby et al. (2012) found no effect of Grade 1 morphological awareness on Grade 3 reading comprehension, whereas in Kruk and Bergman's (2013) examination of the same grade levels, morphological awareness did uniquely predict subsequent reading comprehension. Our longitudinal findings align with the latter study—in our sample, first graders' morphological awareness uniquely predicted their Grade 2 reading comprehension beyond decoding and vocabulary (see also Carlisle, 1995; Carlisle and Fleming, 2003 for similar patterns without these controls). Notably, these converging findings occurred despite the different morphological awareness task formats across studies (generating inflectional forms and correcting morphological errors in the current study; (de) composition of morphological forms in Kruk and Bergman, 2013). This suggests that the pattern is robust to differences in the measurement of morphological awareness, provided the task is sufficiently sensitive to individual differences among children (*cf.* Kirby et al., 2012, who found floor effects in Grade 1 students' morphological awareness). Morphological awareness did not contribute specifically to gains in reading comprehension over the year. First grade morphological awareness does contribute to second grade readers' comprehension, but is not restricted to the association between these two variables over this one year of time. In other words, morphological awareness was not observed to be a driver of gains in reading comprehension from first to second grade.

Our second set of analyses added syntactic awareness as a control variable (along with decoding and vocabulary). Inflectional morphemes serve grammatical functions in sentences, making syntactic awareness a conceptually valuable control that isolates children's metalinguistic awareness of morphemes from that of other aspects of grammar. However, syntactic awareness has not been incorporated into prior work with this age group (but has been with older children; see Proctor et al., 2012; Kieffer et al., 2016; Deacon and Kieffer, 2018; Metsala et al., 2021). Our concurrent results underscore the importance of doing so—once shared variance between morphological and syntactic awareness was accounted for, the unique effect of morphological awareness on Grade 1 reading comprehension was nonsignificant.

Interestingly, a different pattern emerged longitudinally—Grade 1 morphological awareness explained unique variance in Grade 2 reading comprehension beyond decoding, vocabulary, and syntactic awareness. Broadly, this aligns with prior work by Muter et al. (2004), who found that Grade 1 grammatical awareness—a composite capturing awareness of morphology and syntax—predicted reading comprehension one year later beyond vocabulary, word reading, and other controls. Our results built on this finding by isolating morphological awareness from syntactic awareness. Indeed, the pattern we found supports distinguishing the two predictors; in our sample, first graders' syntactic awareness did not predict their subsequent reading comprehension, whereas their morphological awareness did.

Why might morphological awareness have predicted later, but not concurrent, reading comprehension when controlling for syntactic awareness? One explanation is that decoding accounted for a larger proportion of variance in concurrent than subsequent reading comprehension (see also Garcia and Cain, 2014), leaving less variance to be explained by other skills in the concurrent analysis. Here, we note that morphological awareness was more strongly related to concurrent reading comprehension than vocabulary or syntactic awareness, as shown in the bivariate correlations and the magnitudes of their regression coefficients. Thus, although decoding explained the bulk of variance in Grade 1 reading comprehension, morphological awareness outperformed the other oral language skills. We should also note that our relatively small sample size (see limitations section), may have contributed to morphological awareness not meeting traditional levels of statistical significance in this more stringent concurrent analysis. It may also be that students' morphological awareness may have accounted for unique variance in second versus first grade as the number of morphologically complex words increases with higher text levels (e.g., Dawson et al., 2023). An additional explanation concerns the negligible relationship between syntactic awareness and Time 2 reading comprehension. The reason for this is unclear—there were no apparent issues with the syntactic awareness task or the distribution of scores in our sample, its association with Time 1 reading comprehension was significant, and similar measures have been significantly associated with reading comprehension among slightly older children (e.g., Deacon and Kieffer, 2018; Metsala et al., 2021). Although syntactic awareness was not an especially powerful control when predicting Grade 2 reading comprehension in our sample, the longitudinal results suggest that children's early morphological skills were more relevant than their syntactic skills to the reading comprehension demands they faced in Grade 2.

From a theoretical standpoint, the Morphological Pathways Framework (Levesque et al., 2021) sheds light on our findings by proposing three paths to explain the effects of morphology on reading comprehension—one direct path, and two indirect paths mediated by children' use of morphemes to read, and determine the meaning of, morphologically complex words (morphological decoding and morphological analysis, respectively). The framework has empirical support from research with older elementary-school students. For instance, Levesque and colleagues (2019) found that Grade 3 morphological awareness predicted subsequent gains in children's morphological analysis for derived words—a skill that allows children to infer the meaning of morphologically complex words (e.g., using knowledge of the base word *question* and suffix *–able* understand the meaning of *questionable*). Morphological analysis, in turn, predicted gains in reading comprehension. Although this indirect path might plausibly contribute to our findings, we suspect that morphological analysis of inflections may be less useful than analysis of derived words in English, which has a limited system of inflections (Kuo and Anderson, 2006).

The direct path of the Morphological Pathways Framework posits a link between morphological awareness and text comprehension processes, through which readers parse and interpret texts to build a situation model (Kintsch and van Dijk, 1978). One proposed explanation for this direct path is that morphological awareness captures general metalinguistic and/or oral language skills (Levesque et al., 2021). Consistent with our longitudinal findings, prior research suggest that morphological awareness predicts reading comprehension after controlling for other metalinguistic skills (e.g., Carlisle, 1995; Kirby et al., 2012) and general language abilities (e.g., Kieffer et al., 2016; Metsala et al., 2021). A second, yet-to-be-explored possibility is that children's awareness of inflections might facilitate text comprehension directly by signaling the temporal and causal relationships involved in building strong situation models (Morrow, 1986; Zwaan et al., 1995). As one example, grammatical tense and aspect indicate when an action took place (past, present, future) and whether that action has been completed or is ongoing (e.g., *Sally walked down the street* vs. *Sally was walking down the street*). Adult readers are sensitive to grammatical aspect's role in signaling temporal information (Magliano and Schleich, 2000) and use it to define narrative event boundaries (Feller et al., 2019). Indeed, some have argued that morphology provides a set of processing instructions for situation model construction (Givón, 1992). Our findings of a unique relationship between Grade 1 morphological awareness and reading comprehension highlight the need to further explore the role of inflectional morphology in these comprehension processes in general, and in young readers specifically.

Limitations

As with all research, limitations need to be considered when interpreting our findings. Central among them is the modest sample size, which was largely a result of labor negotiations that disrupted our planned participant numbers. Our achieved sample size was sufficient for our purposes, but constrained the analyses we could run. We present our findings as preliminary evidence that beginning readers' morphological awareness predicts later reading comprehension beyond decoding and a comprehensive set of oral language controls, but encourage future research that reproduces these analyses in larger samples. Longitudinal research with these controls in young children is limited; this study contributes to initial understandings and to inform further research. Another limitation is our morphological awareness task, which only captures inflections—one aspect of English's wider morphological system. This focus was motivated by a desire to target morphological skills that are well-developed among first-grade students, though we fully support calls to assess morphological awareness in a comprehensive way (Apel et al., 2013). Recent work by James et al. (2021) found that awareness of inflections, derivations, and compounds comprised a single factor among 6- to 8-year-old children, which suggests that inflectional awareness taps into the broader morphological awareness construct. Even so, we must be clear that our morphological awareness task does not capture the full scope of English morphology.

Conclusion

In summary, our study speaks to morphological awareness' role in predicting early reading comprehension. We took the important step of including vocabulary and syntactic awareness—oral language skills linked with both morphological awareness and reading comprehension—as controls, alongside decoding. In doing so, we were able to isolate morphological awareness' contribution to early reading comprehension. Grade 1 students' morphological awareness predicted their concurrent and prospective reading comprehension beyond decoding and vocabulary, and the longitudinal effect remained when further controlling for syntactic awareness. Indeed, morphological awareness emerged as the strongest oral language predictor of reading comprehension in our models. Although more work remains to be done with this age group, this study points to the relevance of morphological awareness to reading comprehension from early in children's reading development—a finding that adds to our understanding of the oral language contributors to beginning reading comprehension. Our findings also point to the need for further research to guide assessment and instruction of morphology in the earliest elementary grades.

Data availability statement

The datasets presented in this article are not readily available because the consent forms signed for this study preclude making the data for this manuscript openly available. Requests to access the datasets should be directed to Jamie.Metsala@msvu.ca.

Ethics statement

The studies involving humans were approved by Mount Saint Vincent University's University Research Ethics Board. The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation in this study was provided by the participants' legal guardians/next of kin.

Author contributions

ES: Formal analysis, Investigation, Methodology, Writing – original draft. JM: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Resources, Writing – review & editing, Project administration.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This research was supported by an Inter-University Research Network grant from the Nova Scotia Department of Education and Early Childhood Development and by the Gail and Stephen Jarislowsky Chair in Learning Disabilities Endowment to Mount Saint Vincent University.

Acknowledgments

The authors wish to thank participating students, their teachers and principals, and the project team, including research assistants and individual collaborators, for their contributions to this research.

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

Anglin, J. M. (1993). Vocabulary development: a morphological analysis. *Monogr. Soc. Res. Child Dev.* 58, i–186. doi: 10.2307/1166112

Apel, K., Diehm, E., and Apel, L. (2013). Using multiple measures of morphological awareness to assess its relation to reading. *Topics in Language Disorders*, 33, 42–56. doi: 10.1097/TLD.0b013e318280f57b

Apel, K., and Henbest, V. S. (2016). Affix meaning knowledge in first through third grade students. *Lang. Speech Hear. Serv. Sch.* 47, 148–156. doi: 10.1044/2016_LSHSS-15-0050

Beck, I., and McKeown, M. (1991). "Conditions of vocabulary acquisition" in *Handbook of reading research, vol. 2*. eds. R. Barr, M. L. Kamil, P. B. Mosenthal and P. D. Pearson (Mahwah, NJ: Lawrence Erlbaum Associates, Inc), 789–814.

Berko, J. (1958). The child's learning of English morphology. *Word* 14, 150–177. doi: 10.1080/00437956.1958.11659661

Berninger, V. W., Abbott, R. D., Nagy, W., and Carlisle, J. (2010). Growth in phonological, orthographic, and morphological awareness in grades 1 to 6. *J. Psycholinguist. Res.* 39, 141–163. doi: 10.1007/s10936-009-9130-6

Cain, K. (2007). Syntactic awareness and reading ability: is there any evidence for a special relationship? *Appl. Psycholinguist.* 28, 679–694. doi: 10.1017/S0142716407070361

Cain, K., and Oakhill, J. (2009). "Reading comprehension development from 8 to 14 years. Beyond decoding: the behavioral and biological foundations of reading comprehension" in *Beyond decoding: the behavioral and biological foundations of reading comprehension*. eds. R. K. Wagner, C. Schatschneider and C. Phythian-Sence (New York, NY: Guilford Press), 143–175.

Carlisle, J. F. (1995). "Morphological awareness and early reading achievement" in *Morphological aspects of language processing*. ed. L. B. Feldman (Hillsdale, NJ: Erlbaum), 189–209.

Carlisle, J. F. (2000). Awareness of the structure and meaning of morphologically complex words: impact on reading. *Read. Writ.* 12, 169–190. doi: 10.1023/A:1008131926604

Carlisle, J. F., and Fleming, J. (2003). Lexical processing of morphologically complex words in the elementary years. *Sci. Stud. Read.* 7, 239–253. doi: 10.1207/S1532799XSSR0703_3

Cunningham, A. J., and Carroll, J. M. (2015). Early predictors of phonological and morphological awareness and the link with reading: evidence from children with different patterns of early deficit. *Appl. Psycholinguist.* 36, 509–531. doi: 10.1017/S0142716413000295

Cunningham, A. E., and Stanovich, K. E. (1997). Early reading acquisition and its relation to reading experience and ability 10 years later. *Dev. Psychol.* 33, 934–945. doi: 10.1037/0012-1649.33.6.934

Dawson, N., Hsiao, Y., Wei Ming Tan, A., Banerji, N., and Nation, K. (2023). Effects of target age and genre on morphological complexity in children's reading material. *Sci. Stud. Read.* 27, 529–556. doi: 10.1080/10888438.2023.2206574

de Villiers, J. G., and de Villiers, P. A. (1973). A cross-sectional study of the acquisition of grammatical morphemes in child speech. *J. Psycholinguist. Res.* 2, 267–278. doi: 10.1007/BF01067106

Deacon, S. H., Benere, J., and Pasquarella, A. (2013). Reciprocal relationship: children's morphological awareness and their reading accuracy across grades 2 to 3. *Dev. Psychol.* 49, 1113–1126. doi: 10.1037/a0029474

Deacon, S. H., and Kieffer, M. J. (2018). Understanding how syntactic awareness contributes to reading comprehension: evidence from mediation and longitudinal models. *J. Educ. Psychol.* 110, 72–86. doi: 10.1037/edu0000198

Deacon, S. H., Kieffer, M. J., and Laroche, A. (2014). The relation between morphological awareness and reading comprehension: evidence from mediation and longitudinal models. *Sci. Stud. Read.* 18, 432–451. doi: 10.1080/10888438.2014.926907

Deacon, S. H., and Kirby, J. R. (2004). Morphological awareness: just "more phonological"? The roles of morphological and phonological awareness in reading development. *Appl. Psycholinguist.* 25, 223–238. doi: 10.1017/S0142716404001110

Dunn, L. M., and Dunn, D. M. (2007). *Peabody picture vocabulary test* (*4th ed.*). London: Pearson Education.

Feller, D. P., Eerland, A., Ferretti, T. R., and Magliano, J. P. (2019). Aspect and narrative event segmentation. Collabra: *Psychology*, 5, 12. doi: 10.1525/collabra.182

Foorman, B. R., Herrera, S., Petscher, Y., Mitchell, A., and Truckenmiller, A. (2015). The structure of oral language and reading and their relation to comprehension in kindergarten through grade 2. *Read. Writ.* 28, 655–681. doi: 10.1007/s11145-015-9544-5

Freund, R. J., Wilson, W. J., and Sa, P. (2006). *Regression analysis: statistical modeling of a response variable* (*2nd ed.*). Cambridge, MA: Academic Press.

García, J. R., and Cain, K. (2014). Decoding and reading comprehension: A meta-analysis to identify which reader and assessment characteristics influence the strength of the relationship in English. *Review of Educational Research*, 84, 74–111. doi: 10.3102/0034654313499616

Geva, E., and Farnia, F. (2012). Developmental changes in the nature of language proficiency and reading fluency paint a more complex view of reading comprehension in ELL and EL1. *Read. Writ.* 25, 1819–1845. doi: 10.1007/s11145-011-9333-8

Givón, T. (1992). The grammar of referential coherence as mental processing instructions. *Linguistics*, 30, 5–56. doi: 10.1515/ling.1992.30.1.5

Gough, P. B., Hoover, W. A., and Peterson, C. L. (1996). "Some observations on a simple view of reading" in *Reading comprehension difficulties: processes and intervention*. eds. C. Cornoldi and J. Oakhill (Mahwah, NJ: Erlbaum Publishers), 1–13.

Gough, P. B., and Tunmer, W. E. (1986). Decoding, reading, and reading disability. *Remedial Spec. Educ.* 7, 6–10. doi: 10.1177/074193258600700104

Green, S. B. (1991). How many subjects does it take to do a regression analysis? *Multivar. Behav. Res.* 26, 499–510. doi: 10.1207/s15327906mbr2603_7

Harris, R. J. (1985). *A primer of multivariate statistics* (*2nd ed.*). New York: Academic Press.

Hjetland, H. N., Brinchmann, E. I., Scherer, R., and Melby-Lervåg, M. (2017). Preschool predictors of later reading comprehension ability: a systematic review. *Campbell Syst. Rev.* 13, 1–155. doi: 10.4073/csr.2017.14

Hoover, W. A., and Gough, P. B. (1990). The simple view of reading. *Read. Writ.* 2, 127–160. doi: 10.1007/BF00401799

James, E., Currie, N. K., Tong, S. X., and Cain, K. (2021). The relations between morphological awareness and reading comprehension in beginner readers to young adolescents. *J. Res. Read.* 44, 110–130. doi: 10.1111/1467-9817.12316

Kieffer, M. J., Petscher, Y., Proctor, C. P., and Silverman, R. D. (2016). Is the whole greater than the sum of its parts? Modeling the contributions of language comprehension skills to reading comprehension in the upper elementary grades. *Sci. Stud. Read.* 20, 436–454. doi: 10.1080/10888438.2016.1214591

Kim, Y. -S. G. (2023). Executive functions and morphological awareness explain the shared variance between word reading and listening comprehension. *Sci. Stud. Read.* 27, 451–474. doi: 10.1080/10888438.2023.2195112

Kintsch, W., and van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review*, 85, 363–394. doi: 10.1037/0033-295X.85.5.363

Kirby, J. R., Deacon, S. H., Bowers, P. N., Izenberg, L., Wade-Woolley, L., and Parrila, R. (2012). Children's morphological awareness and reading ability. *Read. Writ.* 25, 389–410. doi: 10.1007/s11145-010-9276-5

Kruk, R. S., and Bergman, K. (2013). The reciprocal relations between morphological processes and reading. *J. Exp. Child Psychol.* 114, 10–34. doi: 10.1016/j.jecp.2012.09.014

Kuo, L., and Anderson, R. C. (2006). Morphological awareness and learning to read: a cross-language perspective. *Educ. Psychol.* 41, 161–180. doi: 10.1207/s15326985ep4103_3

Lam, K., Chen, X., Geva, E., Luo, Y. C., and Li, H. (2011). The role of morphological awareness in reading achievement among young Chinese-speaking English language learners: a longitudinal study. *Read. Writ.* 25, 1847–1872. doi: 10.1007/s11145-011-9329-4

Levesque, K. C., Breadmore, H. L., and Deacon, S. H. (2021). How morphology impacts reading and spelling: advancing the role of morphology in models of literacy development. *J. Res. Read.* 44, 10–26. doi: 10.1111/1467-9817.12313

Levesque, K. C., Kieffer, M. J., and Deacon, S. H. (2017). Morphological awareness and reading comprehension: examining mediating factors. *J. Exp. Child Psychol.* 160, 1–20. doi: 10.1016/j.jecp.2017.02.015

Levesque, K. C., Kieffer, M. J., and Deacon, S. H. (2019). Inferring meaning from meaningful parts: The contributions of morphological skills to the development of children\u0027s reading comprehension. *Reading Research Quarterly*, 54, 63–80. doi:10.1002/rrq.219

Leys, C., Klein, O., Dominicy, Y., and Ley, C. (2018). Detecting multivariate outliers: use a robust variant of the Mahalanobis distance. *J. Exp. Soc. Psychol.* 74, 150–156. doi: 10.1016/j.jesp.2017.09.011

Leys, C., Ley, C., Klein, O., Bernard, P., and Licata, L. (2013). Detecting outliers: do not use standard deviation around the mean, use absolute deviation around the median. *J. Exp. Soc. Psychol.* 49, 764–766. doi: 10.1016/j.jesp.2013.03.013

Lonigan, C. J., and Burgess, S. R. (2017). Dimensionality of reading skills with elementary-school-age children. *Sci. Stud. Read.* 21, 239–253. doi: 10.1080/10888438.2017.1285918

MacGinitie, W. H., and MacGinitie, R. K. (1992). *Gates-MacGinitie reading tests* (*2nd Canadian ed.*). Toronto: Nelson Canada.

Magliano, J. P., and Schleich, M. C. (2000). Verb aspect and situation models. *Discourse Processes*, 29, 83–112. doi: 10.1207/S15326950dp2902_1

Metsala, J. L., Sparks, E., David, M., Conrad, N., and Deacon, S. H. (2021). What is the best way to characterise the contributions of oral language to reading comprehension: listening comprehension or individual oral language skills? *J. Res. Read.* 44, 675–694. doi: 10.1111/1467-9817.12362

Moats, L. C. (2020). Teaching Reading Is Rocket Science: What Expert Teachers of Reading Should Know and Be Able to Do. *American Educator*, 44, 4–39.

Morrow, D. G. (1986). Grammatical morphemes and conceptual structure in discourse processing. *Cognitive Science*, 10, 423–455. doi: 10.1207/s15516709cog1004_2

Murnane, R., and Levy, F. (1996). *Teaching the new basic skills: principles for educating children to thrive in a changing economy*. New York: Free Press.

Muter, V., Hulme, C., Snowling, M. J., and Stevenson, J. (2004). Phonemes, rimes, vocabulary, and grammatical skills as foundations of early reading development: evidence from a longitudinal study. *Dev. Psychol.* 40, 665–681. doi: 10.1037/0012-1649.40.5.665

Nagy, W. E. (2007). "Metalinguistic awareness and the vocabulary–comprehension connection" in *Vocabulary acquisition: implications for reading comprehension*. eds. R. K. Wagner, A. E. Muse and K. R. Tannenbaum (New York City: Guilford Press), 52–77.

Ouellette, G., and Beers, A. (2010). A not-so-simple view of reading: how oral vocabulary and visual-word recognition complicate the story. *Read. Writ.* 23, 189–208. doi: 10.1007/s11145-008-9159-1

Perfetti, C. A., and Stafura, J. (2014). Word knowledge in a theory of reading comprehension. *Sci. Stud. Read.* 18, 22–37. doi: 10.1080/10888438.2013.827687

Proctor, C. P., Silverman, R. D., Harring, J. R., and Montecillo, C. (2012). The role of vocabulary depth in predicting reading comprehension among English monolingual and Spanish–English bilingual children in elementary school. *Read. Writ.* 25, 1635–1664. doi: 10.1007/s11145-011-9336-5

Quinn, J. M., Wagner, R. K., Petscher, Y., and Lopez, D. (2015). Developmental relations between vocabulary knowledge and reading comprehension: a latent change score modeling study. *Child Dev.* 86, 159–175. doi: 10.1111/cdev.12292

Robertson, E. K., and Deacon, S. H. (2019). Morphological awareness and word-level reading in early and middle elementary school years. *Appl. Psycholinguist.* 40, 1051–1071. doi: 10.1017/S0142716419000134

Siegel, L. S., and Ryan, E. B. (1988). Development of grammatical-sensitivity, phonological, and short-term memory skills in normally achieving and learning disabled children. *Dev. Psychol.* 24, 28–37. doi: 10.1037/0012-1649.24.1.28

Sorenson-Duncan, T., Mimeau, C., Crowell, N., and Deacon, S. H. (2021). Not all sentences are created equal: evaluating the relation between children's understanding of basic and difficult sentences and their reading comprehension. *J. Educ. Psychol.* 113, 268–278. doi: 10.1037/edu0000545

Sparks, E., and Deacon, S. H. (2015). Morphological awareness and vocabulary acquisition: a longitudinal examination of their relationship in English-speaking children. *Appl. Psycholinguist.* 36, 299–321. doi: 10.1017/S0142716413000246

Sparks, R. L., Patton, J., and Murdoch, A. (2014). Early reading success and its relationship to reading achievement and reading volume: replication of '10 years later'. *Read. Writ.* 27, 189–211. doi: 10.1007/s11145-013-9439-2

Storch, S. A., and Whitehurst, G. J. (2002). Oral language and code-related precursors to reading: evidence from a longitudinal structural model. *Dev. Psychol.* 38, 934–947. doi: 10.1037/0012-1649.38.6.934

Tabachnick, B. G., and Fidell, L. S. (2007). *Using multivariate statistics* (*5th ed.*). Boston, MA: Allyn & Bacon/Pearson Education.

Tunmer, W. E., Herriman, M. L., and Nesdale, A. R. (1988). Metalinguistic abilities and beginning reading. *Read. Res. Q.* 23, 134–158. doi: 10.2307/747799

van Buuren, S., and Groothuis-Oudshoorn, K. (2011). Mice: multivariate imputation by chained equations in R. *J. Stat. Softw.* 45, 1–67. doi: 10.18637/jss.v045.i03

Vellutino, F. R., Tunmer, W. E., Jaccard, J. J., and Chen, R. (2007). Components of reading ability: multivariate evidence for a convergent skills model of reading development. *Sci. Stud. Read.* 11, 3–32. doi: 10.1080/10888430709336632

Wiig, E. H., Semel, E., and Secord, W. A. (2013). *Clinical evaluation of language fundamentals —Fifth edition (CELF-5)*. London: Pearson.

Wilson, R. T. (2016). Literacy and its significance in modern life. *Colleagues* 13, 1–8. Available at: http://scholarworks.gvsu.edu/colleagues/vol13/iss1/5

Woodcock, R. W. (2011). *Woodcock reading mastery tests* (*3rd ed.*). Bloomington, MN: Pearson

Woodcock, R. W., McGrew, K. S., and Mather, N. (2001). *Woodcock-Johnson III tests of achievement*. Boston, MA: Riverside Publishing.

Wulfeck, B. B. (1993). A reaction time study of grammaticality judgments in children. *J. Speech Lang. Hear. Res.* 36, 1208–1215. doi: 10.1044/jshr.3606.1208

Zwaan, R. A., Langston, M. C., and Graesser, A. C. (1995). The construction of situation models in narrative comprehension: An event-indexing model. *Psychological Science*, 6, 292–297. doi: 10.1111/j.1467-9280.1995.tb00513.x

frontiers | Frontiers in Education

TYPE Original Research
PUBLISHED 11 January 2024
DOI 10.3389/feduc.2023.1347143

Check for updates

OPEN ACCESS

EDITED BY
Martin Schöfl,
University of Education Upper Austria, Austria

REVIEWED BY
Ann Dowker,
University of Oxford, United Kingdom
Belinda Blevins-Knabe,
University of Arkansas at Little Rock,
United States

*CORRESPONDENCE
Irene Cadime
irenecadime@ie.uminho.pt

†These authors have contributed equally to this work

RECEIVED 30 November 2023
ACCEPTED 28 December 2023
PUBLISHED 11 January 2024

CITATION
Cruz J, Alves D, Carvalho M, Mendes SA, Rodrigues B and Cadime I (2024) Assessment of math abilities before school entry: a tool development. *Front. Educ.* 8:1347143. doi: 10.3389/feduc.2023.1347143

Assessment of math abilities before school entry: a tool development

Joana Cruz[1†], Diana Alves[2†], Marisa Carvalho[3], Sofia Abreu Mendes[1], Bruna Rodrigues[1] and Irene Cadime[4*]

[1]Centro de Investigação em Psicologia para o Desenvolvimento, Instituto de Psicologia e Ciências das Educação, Universidade Lusíada, Porto, Portugal, [2]Faculdade de Psicologia e Ciências da Educação, Universidade do Porto, Porto, Portugal, [3]Centro de Investigação para o Desenvolvimento Humano, Faculdade de Educação e Psicologia, Universidade Católica Portuguesa, Porto, Portugal, [4]Centro de Investigação em Estudos da Criança, Instituto de Educação, Universidade do Minho, Braga, Portugal

Introduction: The years preceding school entry are pivotal for children to develop fundamental skills that bolster their learning in literacy and math in the following years. Skills like counting and subitizing stand as key predictors of future math abilities. Hence, the availability of universal screening tools is vital to promptly identify children who are lacking in these skills, allowing for tailored classroom practices to meet their needs. This study aimed to develop a mathematical screening tool and assess its psychometric properties.

Methods: The tool encompasses both group and individual tasks, to evaluate counting, cardinality, subitizing, quantity discrimination, addition, and subtraction. The tool was administered twice to 257 children, aged 5-6, in their final year of kindergarten in Portugal, at both the start and end of the school year.

Results: The findings reveal that the tasks generally present a low level of difficulty, especially by the end of kindergarten. Factor analysis demonstrated an adequate structure, and all dimensions showed high internal consistency (>0.80). There was a medium-to-high intercorrelation among the various tasks. Correlations between task scores at the beginning and end of the school year were high. Furthermore, scores across all tasks correlated with math achievement in the first grade of primary education.

Discussion: These outcomes suggest that the tool is both valid and reliable, making it a useful resource for assessing and intervening in math skills during the critical years before school entry.

KEYWORDS
screening tool, math, validity, preschool, reliability

Introduction

Early mathematical skills have consistently been shown to predict subsequent performance in both mathematics and reading (Duncan et al., 2007; Claessens et al., 2009; Claessens and Engel, 2013; Watts et al., 2014). Moreover, empirical studies have revealed significant individual differences in the basic mathematical skills of children even before they receive formal instruction in primary school (Starkey et al., 2004; Jordan and Levine, 2009).

The diversity in early mathematical skill profiles appears to lead to distinct learning trajectories in formal education (Sarama and Clements, 2019). Children who enter primary school with solid foundational mathematical skills (such as counting, cardinality, subitizing, quantity manipulation, recognizing, and writing numbers) tend to commence formal learning in this domain successfully and maintain strong performance throughout their school journey (Jordan et al., 2009; Ehlert and Fritz, 2013). Conversely, children who begin formal learning with weaker mathematical skills often continue to be among the lowest performers (Aunola et al., 2004; Toll and van Luit, 2014). This consistent empirical evidence supports the need for investment in educational policies aimed at preventing learning difficulties. Preschool is an ideal setting for fostering early mathematical skills and for identifying children who may benefit from additional intervention before starting primary school (Cruz et al., 2023a). Based on the cumulative evidence of the significant impact of early mathematical skills on predicting learning trajectories across the lifespan (Leyva et al., 2018; Dumas et al., 2019), there has been ongoing discussion about integrating mathematics skills enhancement into early childhood Multi-Tiered Systems of Support (MTSS) models (Hardy et al., 2017). However, most empirical studies in preschool settings have focused on the implementation of MTSS targeting language and literacy skill promotion (Shepley and Grisham-Brown, 2019).

Addressing the specific needs of each child effectively poses a significant challenge for preschool teachers due to the diverse range of individual differences, cultural backgrounds, and socioeconomic statuses in their classrooms (Greenwood et al., 2019). This diversity requires that teachers not only distinguish between high and low-performing children through informal observation but also conduct more thorough assessments. Thus, the need for valid and reliable tools becomes evident, particularly for quick and accurate evaluation of skills and concepts, including those in mathematics (Kilday et al., 2012). The implementation of a multi-tiered system of support for early mathematical skills emerges as a feasible solution to this challenge, such a system would rely on assessment tools with robust psychometric properties, enabling teachers to effectively assess the development of a wide array of basic skills and concepts in children (Purpura and Lonigan, 2015). In this context, universal screening assumes a critical role. By assessing all children at a given grade level and predicting their risk status, it facilitates the selection of targeted and validated interventions; these interventions aim to enhance skills from the earliest stages, thus averting severe learning difficulties (Fuchs et al., 2008; Seethaler and Fuchs, 2010; Gilbert et al., 2012). Identifying at-risk children and determining the appropriate type, intensity, and frequency of intervention is essential in this proactive approach (Fuchs and Fuchs, 2006).

Mathematical knowledge appears to progress in a hierarchical manner, beginning with the acquisition of simpler skills and advancing to more complex ones (Casey et al., 2018). From an early age, children develop informal mathematical knowledge and skills through continuous interaction with significant figures like parents and teachers (Niklas et al., 2016; Trawick-Smith et al., 2016), often facilitated by playing games (Cohrssen and Niklas, 2019; Ompok et al., 2021). Children first begin by learning the ordinal principle (verbal counting), understanding of how to count objects (one-to-one counting, cardinality, subitizing), and even manipulating quantities (numerical operations) through story problems (Purpura and Lonigan, 2015). A fundamental yet crucial skill in this learning process is verbal counting, defined as the ability to count objects out loud (Passolunghi et al., 2007; Lorena et al., 2013; Cuder et al., 2022). Mastery of this basic skill leads to the acquisition of cardinality (Fritz et al., 2013; Paliwal and Baroody, 2018; Cuder et al., 2022), which involves applying oral counting to objects and understanding that the last number in a count represents the total number of objects (Purpura and Lonigan, 2015; Raghubar and Barnes, 2017). As children become adept at these two skills, they learn to compare numerical quantities, discerning which is larger or smaller—for instance, a child understands that the number 2 is greater than the number 1 (Elliott et al., 2019; Zhang et al., 2020; Xu et al., 2023). Proficiency in counting and cardinality also paves the way for the development of subitizing, the ability to instantly recognize the number of objects in a group without needing to count them (Jansen et al., 2014; Clements et al., 2019). Hannula-Sormunen et al. (2015) found that numerosity and verbal counting skills predict math performance and subitizing skills have an indirect effect on math performance in middle school. The development of these skills will influence the ability to manipulate quantities and foster the development of a number sense. Mazzocco et al. (2011) suggest that an intuitive sense of number measured at preschool serves as a foundational mental system for initial formal math skills.

In a preschool setting, children's initial numerical operations include counting forward and backward, representing addition and subtraction, respectively (Raghubar and Barnes, 2017; Starkey and Gelman, 2020). Children are also developing knowledge and skills related to mathematical conventions such as recognizing and writing numbers. Number recognition involves knowing and identifying numbers, and combining different number symbols with their written words, e.g., linking the numerical symbol "5" with the word "five," while number writing refers to the written representation of numbers (Byrge et al., 2014; Göbel et al., 2014; Leyva et al., 2018).

The importance of early numeracy is increasingly recognized, leading to the development of multi-tiered interventions aimed at enhancing early mathematics skills in preschoolers. This is evident in the efforts to create and validate measurement tools suitable for preschool environments (Purpura et al., 2015). These tools serve two key purposes: first, to identify children who require additional support, and second, to determine the specific mathematical skills and concepts each child needs to learn more about (Purpura and Lonigan, 2015). However, different types of measurement tools may be needed for each purpose. For instance, discrete measures, which can be either fluency-based (Hojnoski et al., 2009; Polignano and Hojnoski, 2012) or non-fluency-based (Reid et al., 2006; Lei et al., 2009), focus on specific mathematical skills. In contrast, brief, untimed broad-content measures assess a range of mathematical skills (Weiland et al., 2012; Purpura et al., 2015). The latter, used up to three times a year, can gauge general math ability and might be more effective than discrete skill measures for screening preschoolers at risk of future difficulties (Foegen et al., 2007; Purpura et al., 2015). Despite the value of broad measures for assessing relative performance, there is a need for tools that can also evaluate specific skills. This is crucial for identifying strengths and weaknesses in a child's mathematical abilities, thereby guiding the planning or adjustment of interventions at various levels—whole-class, small-group, and individual (Seethaler and Fuchs, 2010; Purpura and Lonigan, 2015). Hence, there is a need for reliable and valid tools designed to assess individual early mathematics skills.

Multiple-skill screeners usually evaluate computation fluency and number sense skills of kindergarteners (Seethaler and Fuchs, 2010). While computation fluency is group-administered, taking about 5 min to assess counting, addition, and subtraction through 25 items, number sense is individually administered and evaluates numerical knowledge (counting, patterns, magnitude comparison, arithmetic calculation) through the answers to 30 items (Seethaler and Fuchs, 2010). The screener developed by Jordan et al. (2010) was designed integrating multiple skills. It includes 33 items, assessing counting knowledge and principles, number recognition, number knowledge, non-verbal addition/subtraction calculations, addition/subtraction story problems, and addition/subtraction number combinations. Brendefur et al. (2018) also developed a multiple-skill screener, an 8-min web-based assessment tool for identifying K-2 students at risk for poor math outcomes, providing information in six domains: concepts of number, relationships, context, measurement, and spatial reasoning.

Nevertheless, single-skill screeners exist, such as the one developed by Bramlett et al. (2000), which assesses the ability to name numerals between 1 and 20 within 1 min, and the task designed by Chard et al. (2005), where quantity discrimination is evaluated in a similar 1-min time probe. As Fuchs et al. (2008) suggest, math knowledge involves various abilities, and it's important to assess several skills to make informed decisions about children's progress. According to Seethaler and Fuchs (2010), both single-skill and multiple-skill measures yield similar outcomes, especially in terms of conceptual math knowledge. Purpura et al. (2015) proposes an alternative approach to assessing individual math skills, suggesting that early mathematics should be evaluated through brief but broad-content measures. They developed the Early Numeracy Skills Screener-Brief Version, a concise screening tool with 24 items that assess math ability on a continuum of increasing complexity. The Test of Early Mathematics Ability—Third Edition (TEMA-3), developed by Ginsburg and Baroody (2003), operates within the same conceptual framework. This tool measures both informal and formal numeracy skills and, like the previous one, provides a total score of math ability. Starkey et al. (2004) developed the Child Math Assessment which is a measure that not only covers numeracy, but also assesses other math skills, such as space/geometry, measurement, patterns, and logical relations. Research-based Early Mathematics Assessment (REMA) is another instrument that allows the assessment of children's mathematical knowledge and skills, considering the developmental progressions for all important skills in early mathematics (Sarama and Clements, 2019). The measure was created to assess mathematical knowledge and skills acquired along developmental progressions of levels of thinking.

In Portugal, the implementation of tiered models of service delivery was recommended in 2018 (Decree-law no 54/2018, 2018). Despite this, few studies have sought to describe and evaluate the effectiveness of tiered interventions for promoting mathematical skills in preschool education, particularly through universal screening and data-based decision-making (Cruz et al., 2023b). A possible reason for the scarcity of Portuguese studies could be the lack of screening tools to assess math skills in kindergarten. The Test of Early Mathematics Abilities–TEMA 3 (Ginsburg and Baroody, 2003) was adapted for the Portuguese population (Cadima et al., 2008). This test, with 72 items, assesses mathematical competence in children aged from 3 years and zero months to 8 years and 11 months. It has been used in various Portuguese studies with preschool children (e.g., Abreu-Lima et al., 2012; Amaral et al., 2017), but there is evidence of its limitation due to the existence of only one total score, which does not allow for the identification of individual strengths and weaknesses in math skills (Seethaler and Fuchs, 2010).

TABLE 1 Participant characteristics.

		Age	
	N (%)	Mean	Std. Dev.
Female	127 (49.4)	4.85	0.40
Male	115 (44.8%)	4.84	0.41
Missing information	15 (5.8%)	4.80	0.41

Addressing this identified need, the primary goal of this study was to create a math screening tool tailored for Portuguese children aged 5–6 years. The "Universal Screening of Math Skills" was developed within a multi-tiered system of support framework and aims to assess all children, as the first component of a decision-making process and a differentiated intervention system. The aim of the present study was to examine its psychometric properties and to gather robust evidence of its reliability and validity, particularly in relation to other established variables. This newly developed multiple-skill screening tool considers skills that are considered key predictors of academic success in formal learning, as identified in studies by Ehlert and Fritz (2013), Leyva et al. (2018), and Dumas et al. (2019). Designed as a concise measure for use before formal education begins, this tool aligns with the Portuguese Orientations for Preschool Education (Silva et al., 2016) and adheres to evaluation guidelines for preschool education (Cardona et al., 2021). It supports Portugal's Curricular Guidelines for Pre-School Education's emphasis on nurturing "Interest and Curiosity in Mathematics" among young learners, fostering a holistic approach in early mathematics education.

Materials and methods

Participants

The sample comprised 257 kindergarten attendees, aged between 4 and 6 years (aged 4–6 years old, mean age = 4.84, std. dev. = 0.405), in their final year before entering formal schooling. These children were enrolled in public schools situated in the northern (*N* = 120, 46.7%) and central (*N* = 137, 53.3%) regions of Portugal. The distribution of boys and girls in the sample was approximately equal, as detailed in **Table 1**.

Measures

Universal screening of math skills (DUCMa–Despiste Universal de Competências Matemáticas)

With the aim of evaluating children's mathematical skills, a screening was carried out consisting of a collective application (cardinality, number recognition, and number writing) and an individual application (counting, cardinality, subitizing, discrimination of quantity, addition, and subtraction). The collective application was performed with groups of 5 children, and the individual application was performed with each child individually. Structured presentations are used. The tasks used are described below:

Task 1. Cardinality: Cardinality is assessed both collectively and individually. In the collective test ('Jogo das Pintas'), children are instructed to draw the requested number of dots next to a specific image. For example, they are asked to draw 5 dots next to the image of a star. In individual tasks, objects are placed on the table and the child is asked to give the evaluator a specific number of objects, for example, "Give me 5 pencils." Subsequently, various objects are placed on the table, and the child is instructed to count them and report the total number. In cardinality tests, the child's ability to make 1-to-1 correspondences and identify the last object counted as corresponding to the total number of objects counted is assessed. Each correct response is recorded for both collective and individual tasks. A total sum score is obtained by adding the correct answers in the different cardinality tasks.

Task 2. Number Recognition: This is a collective test composed of nine items, where children need to identify numbers from 1 to 9, randomly. Children are required to circle the correct number from three options based on the examiner's question (e.g., "Which one is number 3?"). The total score is determined by the number of correct responses.

Task 3. Number Writing: This task is administered collectively and consists of nine items in which children are asked to write the requested number (from 1 to 9), randomly. The total score is calculated by the sum of correct answers.

Task 4. Counting: In an individual setting, children are prompted to count aloud to the highest number they can recall. The last number children count in the correct order is recorded.

Task 5. Subitizing: This task includes five items and is applied individually. In each item, the child is presented with a slide with a certain number of dots (between 3 and 7 dots). The child is expected to quickly verbalize the number of dots on each slide, without counting them. Each correct answer is equivalent to one point. The total score is calculated by the sum of correct answers.

Task 6. Quantity Discrimination: Individually, children are asked to identify the larger of two numbers presented (ranging from 0 to 10). This task lasts one 1 min, with a maximum of 28 items presented. For example, the child is asked to indicate the larger number between 4 and 5. This task assesses the ability to quickly discriminate magnitudes within pairs of numerals. The total number of correct answers within 1 min is taken as the total score, so that accuracy and speed are simultaneously considered in children's performance.

Task 7. Addition: This is an individual task with four items. Everyday situations are presented to children and questions are asked that involve sum, such as: "Pedro had 1 gum. Mom gave him 2 more gummies. How many gummies did he get?" Children can use materials for manipulation (such as pencils, bottle caps) to assist in answering. The addition tasks responses range from 3 to 11.

Task 8. Subtraction: This is an individual task with four items. Everyday situations are presented to children and questions are asked that involve subtraction, such as: "Inês was playing on the beach and had four shells. A wave came and took two shells. How many shells did Inês end up with?" Children can use materials for manipulation (such as pencils, bottle caps) to assist in answering. The subtraction tasks responses range from 2 to 8.

The test materials (instructions, screening tool, and scoring) are available as **Supplementary material** to this paper. The comprehension of instructions by the children was verified in a preceding pilot study. Except for the discrimination of quantity (task 6), all other tasks have no time limit.

Academic achievement in the subject math (test grades)

When children were in the first grade, children's academic grades in Math tests performed along the school year (reflecting performance in this subject), were collected. These grades were collected at the end of the school year and are expressed on a scale ranging from 1 (insufficient) to 4 (very good).

Counting

Individually, at the end of first grade each child was asked to count orally as far as they could remember. The last number children counted in the correct order was recorded.

Procedures

The study received approval from the Ethics Committee of the Psychology of Development Research Centre (CIPD/2122/DEED/5). Necessary legal permissions were obtained from the school boards and the parents/legal guardians of the participants. Parents/legal guardians provided informed consent for their children's participation, in accordance with the Declaration of Helsinki and the Oviedo Convention.

The evaluators received a 3-h training before the assessment. Individual assessments were scheduled with school teachers at times that did not disrupt the children's daily routines. The collective tasks of the screening tool were carried out in small groups of four children. Data collection took place from October to December 2021, at the end of the school year (May and June 2022), and upon completion of the first grade (May and June 2023). The same screening tasks were used in the first two data collection periods. The screening tool was designed to be completed without a time limit. At the end of the first grade, students' grades in Math and their verbal counting abilities were also gathered.

Statistical analyses

For tasks 1, 2, 3, 5, 7, and 8, confirmatory factor analysis was performed to check their dimensionality. For each one of these tasks a one-factor solution was tested, using Mplus version 7 (Muthén and Muthén, 2012). As the items are dichotomic, the weighted least squares estimator with mean and variance adjusted (WLSMV) was chosen. An acceptable model fit was achieved when the Root Mean Square Error of Approximation (RMSEA) was lower than 0.08 and the Comparative Fit Index (CFI) and Tucker-Lewis Index (TLI) were higher than 0.90 (Browne and Cudeck, 1993; Hoyle and Panter, 1995). The evidence of unidimensionality was followed by the computation of Item Response Theory (IRT)

TABLE 2 Model fit for each subscale.

	χ 2 (df)	RMSEA [90% CI]	CFI	TLI
Cardinality	120.595 (35)***	0.098 [0.079, 0.117]	0.970	0.962
Subitizing	4.098 (5)	0.000 [0.000, 0.078]	1.000	1.001
Addition	2.558 (2)	0.033 [0.000, 0.133]	1.000	0.999
Subtraction	0.306 (2)	0.000 [0.000, 0.066]	1.000	1.004
Number recognition	68.615 (27)***	0.077 [0.055, 0.100]	0.980	0.973
Number writing	73.997 (27)***	0.082 [0.060, 0.105]	0.989	0.985

***$p < 0.001$.

parametrization, namely item discrimination and item difficulty (Asparouhov and Muthén, 2020). The item discrimination is related to the rate at which the probability of responding correctly to an item changes given the aptitude level of the child in the construct that is being measured. Thus, higher values indicate better item discrimination. We classified item discrimination using the Baker (2001) guidelines, according to which values between 0.01 and 0.34 are very low, between 0.35 and 0.64 are low, between 0.65 and 1.34 are moderate, between 1.35 and 1.69 are high, and values higher than 1.70 are very high. Regarding item difficulty, this parameter is on a logit scale with a mean value of zero, with items with values higher than zero indicating more difficult items and values below zero indicating easier items. The items were also analyzed in terms of proportion of correct answers and corrected item-total correlation. Kuder-Richardson formula 20 (KR-20) was used as an estimator of internal consistency, with a minimum of 0.70 being required. For this set of analyses, only the data obtained in the first moment of data collection were used.

Next, descriptive statistics of the total scores in each subscale in both moments were computed. Values of skewness and kurtosis above 2 indicated violations to the assumption of normality in the distribution of the data. Therefore, Spearman correlation coefficients were computed to assess the intercorrelations among the scores in the eight subscales of the measure, as well as between these scores in the subscales and the math achievement and counting in primary school. Correlations were classified as following: negligible, when lower than 0.10; small, when those were between 0.10 and 0.30; medium, when those were between 0.30 and 0.50; and large, when higher than 0.50 (Cohen, 1992). These statistics were computed by using IBM® SPSS Statistics 28.

Results

Table 2 presents the model fit for the one-factor model in each subscale. Although in the cardinality and number writing subscales, the RMSEA slightly exceeded the reference value of 0.08, all other fit indicators suggested an adequate fit for the one-factor model.

Table 3 presents the results for the item analysis. All factor loadings were higher than 0.70. Corrected item-total correlations ranged between 0.462 and 0.772. IRT item discrimination values were mostly high or very high in all subscales. KR-20 values were higher than 0.70 ranging between 0.803 and 0.901. These findings indicate a high discrimination power and high internal consistency of the items for all subscales. Regarding difficulty, the proportion of correct answers, as well as the IRT item difficulty values, suggest that, overall, the items are easy for the children. This is particularly the case in the items of the subscales of cardinality, addition, subtraction and number recognition, where all items were correctly responded by more than 60% of the sample and all items had a negative IRT item difficulty value. The subitizing subscale contained the most difficult items (Table 3).

Table 4 presents the descriptive statistics for the total scores obtained in each subscale at the beginning of the last year of kindergarten (moment 1) and last year of kindergarten (moment 2), as well as the results for counting in the primary school. Percentiles for the scores obtained in kindergarten can also be consulted in Appendix A. For counting, the mean values increased across the three moments: on average, children were able to count up to 24 at the beginning of the kindergarten, up to 35 at the end of the kindergarten and up to 264 at the end of the first grade. However, the variance was quite high, especially in the first grade where the standard deviation was around 320 and some children were able to count up to 1006. Cardinality, addition, subtraction and number recognition had some ceiling effect in both moments, as the means were close to the maximum values of the subscales and there were high and negative values of skewness. Quantity discrimination had low values of skewness and kurtosis and the mean number of correct responses increased from around 11 to around 15 from the beginning to the end of kindergarten, thus showing no ceiling effects. The mean scores in subitizing and number writing also had an increase between both moments, showing only a slight ceiling effect at the end of kindergarten (Table 4).

Table 5 presents the Spearman correlation coefficients among the scores in each subscale and counting and math scores in primary school. Math scores were collected for 177 participants and had the following distribution: 5 children (2.8%) had a score of 1 (insufficient), 30 children (16.9%) had a score of 2 (sufficient), 79 children (44.6%) had a score of 3 (good), and 63 children (35.6%) had a score of 4 (very good). The correlations between the scores in the same task in both moments were large in the subscales of counting, subtraction, quantity discrimination, subitizing and number writing, and medium in the subscales of cardinality, addition and number recognition. Addition and subtraction were highly correlated in both moments. At the beginning of kindergarten, cardinality, number recognition and number writing were particularly highly correlated (correlations ranging between 0.504 and 0.647), but the size of the correlation among these subscales was not so high at the end of kindergarten (ranging between 0.296 and 0.356). Counting at the beginning of the kindergarten was also highly correlated with the cardinality at the same moment, and with quantity discrimination at both time points. Regarding counting at the end of kindergarten, the highest correlations were with cardinality, quantitative discrimination, subitizing and number writing. Quantitative discrimination was the skill in kindergarten with the highest correlation with counting at the end of primary school. Addition, subtraction and number recognition in kindergarten had low or negligible correlations with counting at the end of primary school. When considering

TABLE 3 Results for the item analysis and internal consistency.

	Proportion correct answers	Corrected item-total correlation	Factor loadings	IRT parametrization		KR-20
				Discrimination	Difficulty	
Cardinality						0.870
Ind. Item 1	0.92	0.553	0.904	2.119	−1.541	
Ind. Item 2	0.78	0.573	0.824	1.452	−0.930	
Ind. Item 3	0.91	0.540	0.884	1.889	−1.494	
Ind. Item 4	0.75	0.608	0.858	1.669	−0.790	
Col. Item 1	0.80	0.674	0.867	1.739	−0.977	
Col. Item 2	0.84	0.623	0.878	1.831	−1.154	
Col. Item 3	0.92	0.572	0.968	3.844	−1.440	
Col. Item 4	0.77	0.711	0.916	2.291	−0.808	
Col. Item 5	0.63	0.563	0.782	1.256	−0.425	
Col. Item 6	0.90	0.549	0.878	1.835	−1.427	
Subitizing						0.811
Item 1	0.65	0.689	0.948	2.982	−0.417	
Item 2	0.36	0.645	0.924	2.410	0.383	
Item 3	0.82	0.528	0.952	3.098	−0.966	
Item 4	0.51	0.661	0.878	1.836	−0.039	
Item 5	0.24	0.483	0.787	1.277	0.908	
Addition						0.822
Item 1	0.92	0.462	0.828	1.475	−1.716	
Item 2	0.76	0.738	0.941	2.791	−0.759	
Item 3	0.77	0.738	0.946	2.927	−0.796	
Item 4	0.69	0.694	0.934	2.616	−0.527	
Subtraction						0.803
Item 1	0.91	0.552	0.868	1.747	−1.549	
Item 2	0.87	0.623	0.875	1.804	−1.276	
Item 3	0.83	0.716	0.952	3.116	−0.981	
Item 4	0.68	0.637	0.963	3.558	−0.489	
Number recognition						0.875
Item 1	0.84	0.688	0.914	2.254	−1.073	
Item 2	0.88	0.629	0.865	1.721	−1.333	
Item 3	0.81	0.686	0.929	2.519	−0.927	
Item 4	0.80	0.503	0.728	1.062	−1.145	
Item 5	0.76	0.703	0.909	2.186	−0.772	
Item 6	0.84	0.606	0.855	1.650	−1.165	
Item 7	0.82	0.594	0.808	1.373	−1.119	
Item 8	0.89	0.575	0.848	1.601	−1.453	
Item 9	0.68	0.593	0.840	1.551	−0.547	
Number writing						0.901
Item 1	0.60	0.665	0.861	1.689	−0.304	
Item 2	0.69	0.640	0.882	1.870	−0.571	
Item 3	0.55	0.772	0.943	2.829	−0.140	
Item 4	0.61	0.740	0.921	2.362	−0.306	

(Continued)

TABLE 3 (Continued)

	Proportion correct answers	Corrected item-total correlation	Factor loadings	IRT parametrization		KR-20
				Discrimination	Difficulty	
Item 5	0.41	0.700	0.897	2.027	0.247	
Item 6	0.44	0.626	0.821	1.436	0.173	
Item 7	0.49	0.752	0.921	2.362	0.016	
Item 8	0.65	0.564	0.764	1.184	−0.517	
Item 9	0.37	0.565	0.785	1.269	0.437	

Ind, individual task; Col., collective task. All factor loadings $p < 0.001$.

TABLE 4 Descriptive statistics of the scores in each subscale and counting in primary school.

	N	Min.	Max.	Mean	Std. Dev.	Skewness	Kurtosis
Counting M1	255	0	103	23.878	15.200	1.636	4.910
Cardinality M1	257	0	10	8.214	2.535	−1.556	1.521
Addition M1	257	0	4	3.148	1.297	−1.253	0.119
Subtraction M1	257	0	4	3.284	1.186	−1.542	1.180
Quant. discrimination M1	257	0	28	11.062	6.077	0.641	0.670
Subitizing M1	257	0	5	2.588	1.721	−0.080	−1.250
Number recognition M1	257	0	9	7.299	2.467	−1.548	1.610
Number writing M1	257	0	8	4.136	2.966	−0.084	−1.478
Counting M2	210	4	199	35.105	23.229	3.624	20.785
Cardinality M2	211	0	10	9.479	1.468	−4.068	19.272
Addition M2	211	0	4	3.777	0.739	−3.845	14.869
Subtraction M2	211	0	4	3.777	0.758	−3.783	13.999
Quant. discrimination M2	211	0	28	15.261	5.904	0.562	0.234
Subitizing M2	210	0	5	3.481	1.412	−0.600	−0.593
Number recognition M2	211	0	9	8.630	1.153	−5.112	31.808
Number writing M2	211	0	9	6.649	2.720	−1.124	0.158
Counting (primary school)	156	19	1006	264.173	320.845	1.720	1.294

M1, moment 1; M2, moment 2.

the scores obtained at the beginning of kindergarten, the skills most correlated with math achievement in primary school were quantitative discrimination, followed by counting and cardinality. If considering the scores at the end of kindergarten, the skills most correlated with math achievement in primary school were quantitative discrimination, subitizing and counting.

Discussion

This study aimed to explore the psychometric properties of the Universal Screening of Math Skills, within a framework of multi-tiered systems of support, and to collect evidence of reliability and validity based on the relationship with other variables.

The results show adequate dimensionality and reliability of the subscales. Regarding item difficulty values, the screening tool presents items that are easy for the children, specifically in the subscales of cardinality, addition, subtraction, and number recognition, with slight ceiling effects observed toward the end of kindergarten. The fact that the screening tool includes items that are easy for children in these subscales suggests that these skills are generally well-developed in the target age group (Raghubar and Barnes, 2017; Starkey and Gelman, 2020). Ceiling effects, where a significant portion of children achieve the highest level, could indicate that the tool may not be highly effective in differentiating among children who have mastered the skills. However, it can identify those who present lower results and may need additional support, which is the main purpose of the screeners (Purpura and Lonigan, 2015). The high scores observed for a high portion of children can be due to the characteristics of the Portuguese educational system. Although it is expected that formal math instruction only occurs in primary education, the Portuguese Orientations for Preschool Education (Silva et al., 2016) highlight the need to create deliberate opportunities to develop numeracy skills in preschool, due to their effect on future math skills acquisition. More heterogeneity in the scores would

10.3389/feduc.2023.1347143

TABLE 5 Correlations among the scores in each subscale and counting and math scores in primary school.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1. Counting M1	1	0.555***	0.448***	0.363***	0.538***	0.439***	0.363***	0.458***	0.679***	0.352***	0.345***	0.343***	0.503***	0.390***	0.253***	0.433***	0.353***	0.411***
2. Cardin. M1		1	0.511***	0.496***	0.559***	0.561***	0.504***	0.615***	0.481***	0.403***	0.358***	0.397***	0.452***	0.425***	0.274***	0.408***	0.343***	0.386***
3. Addit. M1			1	0.748***	0.388***	0.313***	0.194**	0.261***	0.384***	0.407***	0.392***	0.492***	0.333***	0.227***	0.221**	0.146*	0.245**	0.334***
4. Subtract. M1				1	0.399***	0.307***	0.195**	0.302***	0.331***	0.446***	0.400***	0.534***	0.329***	0.234***	0.256***	0.204**	0.171*	0.334***
5. Qt. Dis. M1					1	0.564***	0.417***	0.521***	0.449***	0.285***	0.355***	0.354***	0.698***	0.487***	0.240***	0.337***	0.443***	0.449***
6. Subitizing M1						1	0.458***	0.592***	0.410***	0.231***	0.236***	0.283***	0.421***	0.621***	0.219**	0.409***	0.331***	0.266***
7. Num. Rec. M1							1	0.647***	0.404***	0.119	0.282***	0.229***	0.458***	0.416***	0.351***	0.393***	0.262***	0.277***
8. Num. Wr. M1								1	0.515***	0.256***	0.229***	0.228***	0.532***	0.558***	0.281***	0.715***	0.355***	0.329***
9. Counting M2									1	0.374***	0.396***	0.308***	0.482***	0.459***	0.255***	0.521***	0.378***	0.316***
10. Cardin. M2										1	0.522***	0.516***	0.306***	0.222**	0.356***	0.296***	0.257**	0.216**
11. Addit. M2											1	0.592***	0.340***	0.245***	0.460***	0.293***	0.067	0.190*
12. Subtract. M2												1	0.379***	0.214**	0.402***	0.238***	0.164*	0.187*
13. Qt. Dis. M2													1	0.513***	0.310***	0.422***	0.445***	0.397***
14. Subitizing M2														1	0.294***	0.424***	0.435***	0.352***
15. Num. Rec. M2															1	0.353***	0.225**	0.279***
16. Num. Wr. M2																1	0.292***	0.163*
17. Counting (1st grade)																	1	0.518***
18. Math ach. (1st grade)																		1

M1, moment 1; M2, moment 2. *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

probably be observed if other math skills, such as geometry, patterns, or logical relations, were included in the assessment tool. Future studies should address whether assessing these skills in preschool would lead to a better discrimination of the individual differences in children that score at the middle and top of the distribution.

Still considering the item difficulty values, the subitizing subscale contained the most difficult items. This difficulty in subitizing items can be attributed to the developmental progression of numerical abilities. Typically, children progress from simpler numerical tasks involving counting and cardinality for small sets to more complex tasks with larger sets. The challenging nature of subitizing items may be consistent with the natural advancement of math abilities as children's skills develop (Jansen et al., 2014; Clements et al., 2019). Regardless of the difficulty, subitizing scores are a good predictor of math achievement in first grade.

In our analysis of the correlations between skills assessed in kindergarten, such as math grades and counting abilities at the end of primary school, we found that quantitative discrimination was the kindergarten skill most strongly correlated with counting skills at the end of primary school. Furthermore, evidence indicated that both quantitative discrimination and counting skills, assessed at the beginning and end of kindergarten, were highly associated with math achievement by the end of the first grade. Quantitative discrimination and counting are essential foundational numeracy skills, pivotal for grasping more advanced mathematical concepts. Initially, at the start of kindergarten, these skills lay the groundwork for a child's understanding of numbers and sets. By kindergarten's end, these skills typically develop and refine, equipping the child for more sophisticated mathematical reasoning (Fritz et al., 2013; Paliwal and Baroody, 2018; Cuder et al., 2022).

These findings imply that the Universal Screening of Math Skills yields reliable and valid scores useful for assessing early mathematics skills, identifying children at risk in kindergarten, and facilitating a proactive approach in the development of math skills. For benchmarking purposes, percentile ranks against a norm-referenced sample are provided in **Appendix A**. This allows practitioners to identify students performing below the expected level in certain skills, specifically those scoring at least one standard deviation below the mean (Stuckey and Albritton, 2020; Cruz et al., 2023a). Additionally, we propose that low scores in skills such as counting and cardinality might signal a cumulative risk (Paliwal and Baroody, 2018; Cuder et al., 2022), guiding decisions for further assessment or necessary intervention support.

An interesting observation regarding the performance of children by the end of kindergarten is the average counting number, with children typically able to verbally count to around 35. As first grade introduces more complex mathematical concepts and skills within a broader numerical range, this expanded scope may pose challenges for students who are still consolidating basic skills. These findings imply that educators could effectively utilize this foundational ability of verbal counting by integrating interactive and engaging activities in preschool context, such as counting the number of children attending school each day, the number of shoes and sneakers children have, the number of children who have lunch at school and so on. Such activities should build upon the students' existing knowledge, thereby not only improving their verbal counting, but also facilitating the development of more advanced math skills (Purpura and Lonigan, 2015; Elliott et al., 2019; Xu et al., 2023). Furthermore, our results indicate that there was a significant variance in counting abilities both in kindergarten and at the end of first grade. This variability might be attributed to multiple factors, reflecting the diverse ways in which children progress in understanding and mastering counting concepts. Differences in counting skills could arise from variations in children's exposure to mathematical experiences, both in and out of the classroom. For example, children benefit from frequent opportunities to practice counting in various contexts, such as during playtime, group activities, or structured lessons. Students who engage in a rich variety of counting activities are likely to exhibit more advanced skills than those with less exposure, such as the ability to count items mentally without relying solely on visual aids or physical objects, and determine quantities through mental processes, enhancing their efficiency and accuracy (Fuchs et al., 2008; Purpura et al., 2015).

Additionally, cultural and socioeconomic factors might influence the home environment and the types of experiences children encounter before starting school. These elements could be contributing factors to the variance in counting skills, as children from diverse backgrounds may have varying levels of exposure to numerical concepts (Purpura et al., 2015; Greenwood et al., 2019). Specifically, the exposure to diverse counting experiences, cultural activities, and educational resources can shape counting skills. Also, variations in socioeconomic status can impact access to educational opportunities, resources, and support, leading to differences in counting skills. Recognizing these differences and implementing effective instructional strategies can help children build and enhance counting skills for future mathematical learning (Starkey et al., 2004; Greenwood et al., 2019).

This study provides evidence that the Universal Screening of Math Skills can assess a broad range of early math skills. As a brief screening tool that includes both collective and individual tasks, it offers an opportunity to identify children's strengths and weaknesses in mathematics (Seethaler and Fuchs, 2010). This is instrumental in facilitating decision-making for tiered interventions. Specifically, the screener enables the early identification of students who may be at risk of lagging in math skills. Within a MTSS framework, this tool is valuable for customizing interventions to meet individual student needs. Such a personalized approach is key for providing effective, targeted support and early intervention, which are essential in addressing learning gaps before they widen. Additionally, the screening tool yields important data that informs educators about the general numeracy proficiency of the kindergarten cohort. This data-driven strategy is crucial for informed decision-making regarding resource allocation and the evaluation of intervention effectiveness (Fuchs and Fuchs, 2006; Purpura and Lonigan, 2015). Moreover, the results from this tool can be shared with parents, promoting a collaborative relationship between home and school. This enables parents to be well-informed about their child's progress and actively engage in supporting their child's mathematical development (Greenwood et al., 2019).

The use of this screening tool in Portuguese schools offers numerous opportunities to enhance educational practices, aligning with the national Curricular Guidelines for Pre-School Education's emphasis on developing skills and interest in mathematics.

It is important to regularly assess the resources allocated to various support tiers to ensure their alignment with identified needs. Regular evaluations can help determine if adjustments are necessary to optimize the impact of interventions (Fuchs et al., 2008; Seethaler and Fuchs, 2010; Gilbert et al., 2012). Additionally, further professional development for educators and specialized professionals like school psychologists and special education teachers is vital. Training should focus on effectively interpreting screening results, implementing tailored interventions, modifying strategies based on continuous assessments (Silva et al., 2016; Cardona et al., 2021). This holistic approach to training embodies the Curricular Guidelines for Pre-School Education's principles and maximizes the utility of the screening tool in addressing students' math skills needs, while nurturing positive attitudes toward mathematics.

While the integration of a screening tool for math skills in kindergarten within a multi-tiered system of support framework offers significant advantages, it is imperative to recognize and address its inherent limitations.

One limitation is that only scores in math tests during the first grade were collected to collect evidence of validity for the screening test scores. It is unclear whether the scores can also predict long-term academic success. Future studies should focus this predictive power across the children's school path. Moreover, the math tests administered in first grade were developed by each classroom teacher and were not standardized. Thus, future studies should include standardized measures of math abilities, which can provide more valid and accurate estimates of the students' math abilities. Future research should also include divergent validity studies, by studying the association between the results obtained in each task and the results on a test that measures a theoretically unrelated construct.

Another limitation refers to the lack of data on the children's socioeconomic status and parental educational levels, as well as lack of data on children's linguistic and domain-general cognitive abilities, such as working memory, attention, and inhibition, that research has pointed out has being involved in math performance (Soltész et al., 2010; Chu et al., 2016; Coolen et al., 2021). Future studies should include the assessment of these abilities, in order to enhance the knowledge on how cognitive and linguistic skills are implicated in math acquisition (Lefevre et al., 2010). As the quantitative discrimination task also requires processing speed, future studies should include standardized measures of this cognitive ability to collect further evidence of validity.

Regardless of these limitations, the screening tool presents several strengths, including early identification and the possibility to develop personalized interventions, and providing a systemic and comprehensive approach to supporting students (Elliott et al., 2019; Xu et al., 2023). This ensures that interventions are not isolated but are part of a coordinated effort to meet the diverse needs of all children. Moreover, the tool can be easily adapted for other languages and cultures. Future studies could also consider the possibility of developing a digital version of the tool. As mentioned earlier, the screening tool does not distinguish individual differences in children at the middle and higher levels of performance, but offers the possibility to identify those who may need additional support, accordingly to Multi-Tiered Systems of Support framework (Greenwood et al., 2019). However, in future studies, other measures could be developed to differentiate math performance at multiple levels, to guide instruction to middle and higher performances.

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 humans were approved by the Ethics Committee of the Psychology of Development Research Centre (CIPD/2122/DEED/5). The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation in this study was provided by the participants' legal guardians/next of kin.

Author contributions

JC: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Project administration, Resources, Writing – original draft, Writing – review and editing. DA: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Project administration, Resources, Writing – original draft, Writing – review and editing. MC: Conceptualization, Writing – original draft, Writing – review and editing. SM: Conceptualization, Writing – original draft, Writing – review and editing. BR: Conceptualization, Writing – original draft, Writing – review and editing. IC: Data curation, Formal Analysis, Methodology, Validation, Writing – original draft, Writing – review and editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was financially supported by the Portuguese Foundation for Science (FCT) and Technology and the Portuguese Ministry of Science, Technology, and Higher Education through national funds within the framework of the Psychology of Development Research Centre – CIPD (grant number UIDB/04375/2020) and the Research Centre on Child Studies (UIDB/CED/00317/2020). IC was supported by an individual grant (reference CEECINST/00018/2021).

Acknowledgments

We would like to thank the school boards, parents, school psychologists and educators who facilitated the data collection, and to thank the children who participated in 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

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.2023.1347143/full#supplementary-material

References

Abreu-Lima, I., Coelho, V., Lobo, C., Castro, C., Gomes, V., and Monteiro, A. F. (2012). "Promoção da competência matemática pré-escolar: resultados de um programa de intervenção," in *12° Colóquio Psicologia e Educação - Educação, Aprendizagem e Desenvolvimento: Olhares Contemporâneos Através da Investigação e da Prática Actas*, eds L. Mata, F. Peixoto, J. Morgado, J. C. Silva, and V. Monteiro (Portugal: ISPA – Instituto Universitário).

Amaral, J., Cruz, J., Constante, P., Pinto, P., Almeida, M., Lopes, E., et al. (2017). Competências de matemática e de literacia emergente: estudo correlacional. *Rev. Portuguesa Educ.* 30, 85–105.

Asparouhov, T., and Muthén, B. (2020). *IRT in Mplus*. Available online at: https://www.statmodel.com/download/MplusIRT.pdf (accessed September 12, 2023).

Aunola, K., Leskinen, E., Lerkkanen, M. K., and Nurmi, J. E. (2004). Developmental dynamics of math performance from preschool to grade 2. *J. Educ. Psychol.* 96, 699–713. doi: 10.1348/000709905X51608

Baker, F. (2001). *The Basics of Item Response Theory*. Park, MD: University of Maryland College.

Bramlett, R. K., Rowell, R. K., and Mandenberg, K. (2000). Predicting first grade achievement from kindergarten screening measures: a comparison of child and family predictors. *Res. Schools* 7, 1–9.

Brendefur, J. L., Johnson, E. S., Thiede, K. W., Strother, S., and Severson, H. H. (2018). Developing a multi-dimensional early elementary mathematics screener and diagnostic tool: the primary mathematics assessment. *Early Childh. Educ. J.* 46, 153–157. doi: 10.1007/s10643-017-0854-x

Browne, M. W., and Cudeck, R. (1993). "Alternative ways of assessing model fit," in *Testing Structural Equation Models*, eds K. A. Bollen and J. S. Long (Thousand Oaks, CA: Sage), 136–162.

Byrge, L., Smith, L. B., and Mix, K. S. (2014). Beginnings of place value: how preschoolers write three-digit numbers. *Child Dev.* 85, 437–443. doi: 10.1111/cdev.12162

Cadima, S., Abreu-Lima, I., Gomes, V., Coelho, V., Lobo, C., and Ramalho, C. (2008). "Avaliação de competências de matemática dos 4 aos 7 anos de Idade," in *Proceedings of the Actas da XIII Conferência Internacional de Avaliação Psicológica: Formas e Contextos (CD-ROM)*, ed. A. P. Noronha (Braga: Psiquilibrios).

Cardona, M. J., Silva, I. L., Marques, L., and Rodrigues, P. (2021). *Planear e avaliar na educação pré-escolar. Ministério da Educação/Direção-Geral da Educação (DGE)*. Available online at: https://www.dge.mec.pt/sites/default/files/EInfancia/documentos/planearavaliar.pdf (accessed October 10, 2023).

Casey, B. M., Lombardi, C. M., Thomson, D., Nguyen, H. N., Paz, M., Theriault, C. A., et al. (2018). Maternal support of children's early numerical concept learning predicts preschool and first-grade math achievement. *Child Dev.* 89, 156–173. doi: 10.1111/cdev.12676

Chard, D. J., Clarke, B., Baker, S., Otterstedt, J., Braun, D., and Katz, R. (2005). Using measures of number sense to screen for difficulties in mathematics: preliminary findings. *Assess. Effect. Intervention* 30, 3–14.

Chu, F. W., Van Marle, K., and Geary, D. C. (2016). Predicting children's reading and mathematics achievement from early quantitative knowledge and domain-general cognitive abilities. *Front. Psychol.* 7:775. doi: 10.3389/fpsyg.2016.00775

Claessens, A., Duncan, G., and Engel, M. (2009). Kindergarten skills and fifth-grade achievement: evidence from the ECLS-K. *Econ. Educ. Rev.* 28, 415–427.

Claessens, A., and Engel, M. (2013). How important is where you start? early mathematics knowledge and later school success. *Teach. Coll. Rec.* 115, 1–29.

Clements, D. H., Sarama, J., and MacDonald, B. L. (2019). "Subitizing: the neglected quantifier," in *Constructing Number: Merging Perspectives From Psychology and Mathematics Education*, eds A. Norton and M. W. Alibali (Berlin: Springer), 13–45.

Cohen, J. (1992). A power primer. *Psychol. Bull.* 112, 155–159. doi: 10.1037/0033-2909.112.1.155

Cohrssen, C., and Niklas, F. (2019). Using mathematics games in preschool settings to support the development of children's numeracy skills. *Int. J. Early Years Educ.* 27, 322–339.

Coolen, I., Merkley, R., Ansari, D., Dove, E., Dowker, A., Mills, A., et al. (2021). Domain-general and domain-specific influences on emerging numerical cognition: contrasting uni-and bidirectional prediction models. *Cognition* 215:104816. doi: 10.1016/j.cognition.2021.104816

Cruz, J., Mendes, S., Marques, S., Alves, D., and Cadime, I. (2023a). Development of a group emergent literacy screening tool. *Children* 10:306.

Cruz, J., Bóia, A. C., Teixeira, B., Ferreira, C., Faria-Magalhães, F., Cardoso, J., et al. (2023b). Efeitos de uma intervenção de promoção de competências matemáticas na educação pré-escolar. *Rev. Psicol. Educ. Cult.* 27, 6–22.

Cuder, A., Vidoz, M., De Vita, C., Pellizzoni, S., and Passolunghi, M. C. (2022). Numerical training videos and early numerical achievement: a study on 3-year-old preschoolers. *Brain Sci.* 12:88. doi: 10.3390/brainsci12010088

Decree-law no 54/2018 (2018). Diário da República: I série, n.o 129 2918.

Dumas, D., McNeish, D., Sarama, J., and Clements, D. (2019). Preschool mathematics intervention can significantly improve student learning trajectories through elementary school. *AERA Open* 5:233285841987944.

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., and Klebanov, P. (2007). School readiness and later achievement. *Dev. Psychol.* 43, 1–36.

Ehlert, A., and Fritz, A. (2013). Evaluation of math's training program for children with learning difficulties. *South African J. Childh. Educ.* 3, 117–141.

Elliott, L., Feigenson, L., Halberda, J., and Libertus, M. E. (2019). Bidirectional, longitudinal associations between math ability and approximate number system precision in childhood. *J. Cogn. Dev.* 20, 56–74.

Foegen, A., Jiban, C., and Deno, S. (2007). Progress monitoring measures in mathematics: a review of the literature. *J. Special Educ.* 41, 121–139.

Fritz, A., Ehlert, A., and Balzer, L. (2013). Development of mathematical concepts ANNEMARIE FRITZ. *South African J. Childh. Educ.* 3, 38–67.

Fuchs, D., and Fuchs, L. S. (2006). Introduction to response to intervention: what, why, and how valid is it? *Reading Res. Quart.* 41, 93–99.

Fuchs, D., Fuchs, L. S., and Vaughn, S. (2008). *Response to Intervention: a Framework for Reading Educators*. Washington, DC: International Reading Association.

Gilbert, J. K., Compton, D. L., Fuchs, D., and Fuchs, L. S. (2012). Early screening for risk of reading disabilities: recommendations for a four-step screening system. *Assess. Effect. Interv.* 38, 6–14. doi: 10.1177/1534508412451491

Ginsburg, H. P., and Baroody, A. J. (2003). *Test of Early Mathematics Ability*, 3rd Edn. Austin, TX: Pro-Ed.

Göbel, S. M., Watson, S. E., Lervåg, A., and Hulme, C. (2014). Children's arithmetic development: it is number knowledge, not the approximate number sense, that counts. *Psychol. Sci.* 25, 789–798.

Greenwood, C. R., Carta, J. J., Schnitz, A. G., Irvin, D. W., Jia, F., and Atwater, J. (2019). Filling an information gap in preschool MTSS and RTI decision making. *Except. Children* 85, 271–290.

Hannula-Sormunen, M., Lehtinen, E., and Räsänen, P. (2015). Preschool children's spontaneous focusing on numerosity, subitizing, and counting skills as predictors of their mathematical performance seven years later at school. *Math. Thinking Learn.* 17, 155–177.

Hardy, J. K., Hawkins-Lear, S., Hemmeter, M. L., and Grisham-Brown, J. (2017). "Blended practices for promoting early math skills," in *Blended practices for teaching young children in inclusive settings*, eds J. Grisham-Brown and M. L. Hemmeter (Baltimore, MD: Brookes), 311–337.

Hojnoski, R. L., Silberglitt, B., and Floyd, R. G. (2009). Sensitivity to growth over time of the preschool numeracy indicators with a sample of preschoolers in head start. *School Psychol. Rev.* 38, 402–418.

Hoyle, R. H., and Panter, A. T. (1995). "Writing about structural equation models," in *Structural Equational Modeling: Concepts, Issues, and Applications*, ed. R. H. Hoyle (Thousand Oaks, CA: Sage Publications), 158–176.

Jansen, B. R. J., Hofman, A. D., Straatemeier, M., van Bers, B. M. C. W., Raijmakers, M. E. J., and van der Maas, H. L. J. (2014). The role of pattern recognition in children's exact enumeration of small numbers. *Br. J. Dev. Psychol.* 32, 178–194. doi: 10.1111/bjdp.12032

Jordan, N. C., Glutting, J., Ramineni, C., and Watkins, M. W. (2010). Validating a number sense screening tool for use in kindergarten and first grade: prediction of mathematics proficiency in third grade. *School Psychol. Rev.* 39, 181–195.

Jordan, N. C., Kaplan, D., Ramineni, C., and Locuniak, M. N. (2009). Early math matters: kindergarten number competence and later mathematics outcomes. *Dev. Psychol.* 45, 850–867. doi: 10.1037/a0014939

Jordan, N. C., and Levine, S. C. (2009). Socioeconomic variation, number competence, and mathematics learning difficulties in young children. *Dev. Disabil. Res. Rev.* 15, 60–68.

Kilday, C. R., Kinzie, M. B., Mashburn, A. J., and Whittaker, J. V. (2012). Accuracy of teacher judgments of preschoolers' math skills. *J. Psychoeducat. Assess.* 30, 148–159. doi: 10.3758/s13415-021-00906-9

Lefevre, J. A., Fast, L., Skwarchuk, S. L., Smith-Chant, B. L., Bisanz, J., Kamawar, D., et al. (2010). Pathways to mathematics: longitudinal predictors of performance. *Child Dev.* 81, 1753–1767.

Lei, P. W., Wu, Q., DiPerna, J. C., and Morgan, P. L. (2009). Developing short forms of the EARLI numeracy measures: comparison of item selection methods. *Educ. Psychol. Measure.* 69, 825–842.

Leyva, D., Davis, A., and Skorb, L. (2018). Math intervention for latino parents and kindergarteners based on food routines. *J. Child Family Stud.* 27, 2541–2551.

Lorena, A. B., Castro-Caneguim, J. F., and Carmo, J. S. (2013). Habilidades numéricas básicas: algumas contribuições da análise do comportamento. *Estudos de Psicologia* 18, 439–446.

Mazzocco, M. M. M., Feigenson, L., and Halberda, J. (2011). Preschoolers' precision of the approximate number system predicts later school mathematics performance. *PLoS One* 6:e23749. doi: 10.1371/journal.pone.0023749

Muthén, L. K., and Muthén, B. O. (2012). *Mplus User's Guide*, 7th Edn. Los Angeles: Muthén & Muthén.

Niklas, F., Cohrssen, C., and Tayler, C. (2016). Improving preschoolers' numerical abilities by enhancing the home numeracy environment. *Early Educ. Dev.* 27, 372–383. doi: 10.7189/jogh.09.010403

Ompok, C. C., Mei Teng, L., and Sapirai, J. (2021). Effect of games towards children's mathematics performance. *Southeast Asia Early Childh. J.* 10, 1–17.

Paliwal, V., and Baroody, A. J. (2018). How best to teach the cardinality principle? *Early Childh. Res. Quart.* 44, 152–160. doi: 10.1037/dev0001660

Passolunghi, M. C., Vercelloni, B., and Schadee, H. (2007). The precursors of mathematics learning: working memory, phonological ability and numerical competence. *Cogn. Dev.* 22, 165–184.

Polignano, J. C., and Hojnoski, R. L. (2012). Preliminary evidence of the technical adequacy of additional curriculum-based measures for preschool mathematics. *Assess. Effect. Intervention* 37, 70–83.

Purpura, D. J., and Lonigan, C. J. (2015). Early numeracy assessment: the development of the preschool early numeracy scales. *Early Educ. Dev.* 26, 286–313.

Purpura, D. J., Reid, E. E., Eiland, M. D., and Baroody, A. J. (2015). Using a brief preschool early numeracy skills screener to identify young children with mathematics difficulties. *School Psychol. Rev.* 44, 41–59.

Raghubar, K. P., and Barnes, M. A. (2017). Early numeracy skills in preschool-aged children: a review of neurocognitive findings and implications for assessment and intervention. *Clin. Neuropsychol.* 31, 329–351. doi: 10.1080/13854046.2016.1259387

Reid, E. E., Morgan, P. L., DiPerna, J. C., and Lei, P. W. (2006). Development of measures to assess young children's early academic skills: preliminary findings from a Head Start-university partnership. *Insights Learn. Disabil.* 3, 25–38.

Sarama, J., and Clements, D. (2019). "Learning trajectories in early mathematics education," in *Mathematical Cognition and Learning, Researching and Using Learning Progressions (trajectories) in Mathematics Education*, eds D. Siemon, T. Barkatsas, and R. Seah (Rotterdam: Sense Publishers), 1–29.

Seethaler, P. M., and Fuchs, L. S. (2010). The predictive utility of kindergarten screening for math difficulty. *Except. Children* 77, 37–59. doi: 10.1186/s12913-016-1423-5

Shepley, C., and Grisham-Brown, J. (2019). Multi-tiered systems of support for preschool-aged children: a review and meta-analysis. *Early Childh. Res. Quart.* 47, 296–308.

Silva, I., Marques, L., Mata, L., and Rosa, M. (2016). *Orientações curriculares para a Educação Pré-Escolar. Ministério da Educação (DGE)*. Available online at: http://www.dge.mec.pt/ocepe/sites/default/files/Orientacoes_Curriculares.pdf (accessed October 13, 2023).

Soltész, F., Szücs, D., and Szücs, L. (2010). Relationships between magnitude representation, counting and memory in 4- to 7-year-old children: A developmental study. *Behav. Brain Func.* 6, 1–14. doi: 10.1186/1744-9081-6-13

Starkey, P., and Gelman, R. (2020). "The development of addition and subtraction abilities prior to formal schooling in arithmetic," in *Addition and Subtraction: a Cognitive Perspective*, eds T. P. Carpenter, J. M. Moser, and T. A. Romberg (Milton Park: Routledge), 99–116.

Starkey, P., Klein, A., and Wakeley, A. (2004). Enhancing young children's mathematical knowledge through a pre-kindergarten mathematics intervention. *Early Childh. Res. Quart.* 19, 99–120.

Stuckey, A., and Albritton, K. (2020). Exploring the use of a multiple-gating screening process to identify preschool-age children for multitiered instructional support. *Top. Early Childh. Special Educ.* 39, 238–249.

Toll, S. W. M., and van Luit, J. E. H. (2014). The developmental relationship between language and low early numeracy skills throughout kindergarten. *Except. Children* 81, 64–78.

Trawick-Smith, J., Swaminathan, S., and Liu, X. (2016). The relationship of teacher-child play interactions to mathematics learning in preschool. *Early Child Dev. Care* 186, 716–733.

Watts, T. W., Duncan, G. J., Siegler, R. S., and Davis-Kean, P. E. (2014). What's past is prologue: relations between early mathematics knowledge and high school achievement. *Educ. Res.* 43, 352–360. doi: 10.3102/0013189X14553660

Weiland, C., Wolfe, C. B., Hurwitz, M. D., Clements, D. H., Sarama, J. H., and Yoshikawa, H. (2012). Early mathematics assessment: validation of the short form of a prekindergarten and kindergarten mathematics measure. *Educ. Psychol.* 32, 311–333.

Xu, X., Chen, C., Wang, L., Zhao, M., Xin, Z., and Liu, H. (2023). Longitudinal relationship between number line estimation and other mathematical abilities in Chinese preschool children. *J. Exp. Child Psychol.* 228:105619. doi: 10.1016/j.jecp.2022.105619

Zhang, T., Chen, C., Chen, C., and Wei, W. (2020). Gender differences in the development of semantic and spatial processing of numbers. *Br. J. Dev. Psychol.* 38, 391–414. doi: 10.1111/bjdp.12329

Appendix A

Universal screening of math skills' percentile ranks.

TABLE 1 Percentiles for each subscale at the beginning of the last year of kindergarten.

	Percentiles						
	5	10	25	50	75	90	95
Counting	5	10	11	20	30	39	49
Cardinality	3	4	7	9	10	10	10
Addition	0	1	3	4	4	4	4
Subtraction	0	1	3	4	4	4	4
Quant. discrimination	1	4	7	11	14	18	23
Subitizing	0	0	1	3	4	5	5
Number recognition	2	3	6	9	9	9	9
Number writing	0	0	1	5	7	8	8

TABLE 2 Percentiles for each subscale at the end of the last year of kindergarten.

	Percentiles						
	5	10	25	50	75	90	95
Counting	11	15	20	30	40	59	64
Cardinality	7	8	10	10	10	10	10
Addition	2	3	4	4	4	4	4
Subtraction	2	3	4	4	4	4	4
Quant. discrimination	7	9	12	14	18	27	28
Subitizing	1	1	2	4	5	5	5
Number recognition	7	8	9	9	9	9	9
Number writing	0	2	5	7	9	9	9

frontiers | Frontiers in Education

TYPE Original Research
PUBLISHED 12 March 2024
DOI 10.3389/feduc.2024.1337716

Check for updates

OPEN ACCESS

EDITED BY
Martin Schöfl,
University of Education Upper Austria, Austria

REVIEWED BY
Christian Kißler,
Technical University Dortmund, Germany
Erica Lembke,
University of Missouri, United States

*CORRESPONDENCE
Maria-Aikaterini Chatzaki
✉ mchatzaki@ku.de

RECEIVED 13 November 2023
ACCEPTED 13 February 2024
PUBLISHED 12 March 2024

CITATION
Chatzaki M-A, Skillen J, Ricken G and Seitz-Stein K (2024) Exploring the potential of a game-based preschool assessment of mathematical competencies.
Front. Educ. 9:1337716.
doi: 10.3389/feduc.2024.1337716

Exploring the potential of a game-based preschool assessment of mathematical competencies

Maria-Aikaterini Chatzaki[1]*, Johanna Skillen[1], Gabriele Ricken[2] and Katja Seitz-Stein[1]

[1]Department of Psychology, Catholic University of Eichstätt-Ingolstadt, Eichstätt, Germany, [2]Department of Primary and Secondary Education, Social Pedagogy as well as Special Needs Education, University of Hamburg, Hamburg, Germany

Background: Early mathematical competencies are foundational for later academical development. There is a need for valid and resource-saving approaches to assess those skills. The House of Numbers (HoN) is a newly developed linear board game that allows the assessment of preschool early mathematical competencies. This article aims to examine aspects of this 24-item screening such as its reliability and validity, and whether it can successfully identify children at risk of developing math difficulties. It also aims to explore children's perceptions of the game-based HoN compared to a typical preschool math test.

Methods: A sample of 147 German preschoolers (M_{age} = 5 years 10 months, SD_{age} = 5 months) was evaluated with the HoN and with a standard instrument for assessing early mathematical competencies (MARKO-D). Additionally, a subsample of n = 47 children rated their perception of both tools.

Results: The results speak against an effect of the game-setting on the children's performance. Regarding the aspects of the reliability and the validity of the HoN, both tools were sensitive to age differences between 5- and 6-year-old children. The high correlation between the two assessments speaks for the convergent validity of the HoN. Furthermore, an item analysis based on the Rasch model showed excellent results for all items of the new game-based approach. The distribution of the items on the logit measurement ruler of the Person-Item Map confirms, with only a few explainable exceptions, the developmental levels of the model the HoN is based on. A high person and item reliability confirm the internal consistency of the HoN. Regarding the diagnostic validity of the HoN, running a receiver operating characteristic curve resulted in a superior discrimination area under the curve. A sample relevant cut-off z-score was specified. Using this score as an indicator of low math performance resulted in high sensitivity, specificity and a high relative improvement over chance index. In addition, children's explorative assessment of perception speaks in favor of the game-based assessment.

Conclusion: All in all, the findings suggest, that the game-based measurement HoN can be a reliable, valid, time-saving, and attractive option for assessing early mathematical competencies in preschool settings.

KEYWORDS

math development, assessment, screening, preschool, board game, validity

1 Introduction

Early mathematical competencies are the basis for mastering more complex mathematical understanding (Fritz et al., 2018). Yet, despite the large amount of research on early mathematical competencies, there is not always consensus on the definition of these skills (Bakker et al., 2022). Most of the researchers seem to agree that apart from counting, early mathematical competencies include components such as the understanding of cardinality and the ability to compare magnitudes (Purpura and Lonigan, 2013; Purpura et al., 2017; Bakker et al., 2022). Children intuitively develop such concepts through personally meaningful activities that facilitate the learning process (Fisher et al., 2012). They seek meaning in what they do and actively construct their knowledge through exploration and discovery. Such activities are found in everyday situations, including play (Hirsh-Pasek et al., 2009; Fisher et al., 2012). For example, they estimate how many blocks they need to build the tallest skyscraper or translate the number on the dice in a board game and move that many spaces on the board (Ginsburg et al., 2001; Fisher et al., 2012).

Accordingly, different trajectories of mathematical competence begin to emerge before school entry (Stern, 2009). Research suggests that developmental trajectories of mathematical development are cumulative, meaning that individual differences in children's math performance increase over time (Aunola et al., 2004). Therefore, children with low mathematical competencies in the kindergarten are most likely to have difficulties with mathematics in school. Difficulties with mathematics at school may include the developmental learning disorder with impairment in mathematics (dyscalculia). Dyscalculia is defined in ICD-11 as "an impairment in mathematical skills such as number sense, memorization of number facts, accurate calculation, fluent calculation, and accurate mathematic reasoning" (ICD-11, 6A03.2). Children with dyscalculia show mathematical performance well below that expected for their age and intellectual functioning (World Health Organization, 2019/2021). Similarly, dyscalculia is defined in DSM-V as the difficulties in processing numbers, acquiring arithmetic factual knowledge and performing arithmetic operations quickly and accurately (American Psychiatric Association, 2013). These impairments are manifested by more errors and longer solution times on mathematical tasks in comparison to people with no such disorder. They are also usually accompanied by reduced performance particularly in visual-spatial working memory and/or reduced performance in inhibitory control. Dyscalculia can be reliably diagnosed from the end of the second year of primary school, when math performance becomes more stable (Morgan et al., 2009). An earlier assessment of math competencies in kindergarten and the first year of primary school can identify a risk of developing dyscalculia (Deutsche Gesellschaft für Kinder- und Jugendpsychiatrie, Psychosomatik und Psychotherapie e.V, 2023).

In the recent years, a series of studies have highlighted the importance of early mathematical competencies as specific precursors of later mathematical performance during formal education (i.a., Krajewski and Schneider, 2006, 2009; Chu et al., 2018). Specifically, according to the study of Stern (2009) referring to mathematical development, even after controlling for intelligence at the ages 6, 8, 12, and 18, the correlations between mathematical reasoning at those ages and at the age of 23 remained statistically significant. The results of Krajewski and Schneider (2006) show that preschool quantity-numeracy skills could explain about 25% of the variance in primary school mathematics skills at the end of grades 1 and 4, while intelligence as a nonspecific cognitive predictor explained only up to 10% of the variance in basic numerical skills. Thus a lack of fostering of relevant skills during the preschool years (before enrollment in first elementary grade) and even an overestimation of individual skills at school entry can lead to persisting discrepancies in mathematical competencies during the school years (Krajewski and Schneider, 2006; Morgan et al., 2009; Stern, 2009; Anders et al., 2013). That being so, difficulties in mathematical understanding during the school years are not only due to school-related reasons (Wittmann, 2001).

Early mathematical competencies are highly heterogeneous among preschool children (Mähler et al., 2017; Cahoon et al., 2021; Bakker et al., 2022) but research shows that these competencies can be effectively promoted in kindergarten before school enrolment (Ehlert and Fritz, 2016; Moraske et al., 2018; Skillen et al., 2018a). The educational staff stimulates children's interest in further engaging with new mathematical concepts in the daily kindergarten routine, and potential developmental risks can be examined more intensively (Ministerium für Kultus, Jugend und Sport Baden-Württemberg, 2022). Characteristics that seem to foster the success of preschool intervention studies are shown to be the promotion of support of the social environment, children being viewed as active, experiential learners, and educational staff adapting the curriculum to children's individual needs (Hirsh-Pasek et al., 2009).

Playful approaches are a familiar, spontaneous, and engaging opportunity for learning math. Hirsh-Pasek et al. (2009) make a distinction between ordinary children's play and guided play. The former is usually unstructured, whereas guided play is structured and guided by adults with a set of learning objectives in mind. Guided play allows educators to be goal oriented but also sensitive and responsive to children's behavior. They can find out what children already know and build on their prior math knowledge. Developmentally appropriate and guided play can provide rich context for children's learning by motivating children to participate (Hirsh-Pasek et al., 2009). The effectiveness of play-based methods using guided play, both at home and in the classroom, in promoting mathematical skills in preschool children has been well documented (Ramani et al., 2012; Jörns et al., 2013; Hauser et al., 2014; Vogt et al., 2018). In particular, the potential of linear board games has often been shown to be quite effective in fostering early mathematical competencies (Ramani et al., 2012; Laski and Siegler, 2014; Elofsson et al., 2016; Skillen et al., 2018b; Gasteiger and Moeller, 2021). Linear board games, such as the commercial board game *Chutes and Ladders* (Milton Bradley Company, 1978), have linearly arranged, consecutively numbered and equally sized spaces (Siegler and Ramani, 2009). The distinctive property of linear board games is that they support the mental linear number representation. Such games provide multiple cues that help understand more about the order and the magnitude of the displayed numbers (Siegler and Ramani, 2009; Laski and Siegler, 2014).

Effective assessment tools are needed to enable educational staff to assess children's developmental status and identify their individual needs in order to adapt the fostering. Yet, since time and personnel resources of educators are limited in the kindergarten, there is a need for resource-saving approaches that can be integrated into everyday life (Textor, 2006; Jörns et al., 2013). Moreover, effective early identification of cases at risk is necessary in order to provide targeted support and attempt to close interindividual developmental gaps (Gerlach et al., 2013; Ricken et al., 2013). Screening tests are suitable for this purpose. They are economic, as they are simple, easy to carry out, time and cost-effective procedures (Moosbrugger and Kelava, 2020). Screening tests are used to look for the first signs of a disorder and provide a general statement about a child's developmental status according to age as to whether there are any developmental abnormalities. If there are indications of developmental deviations, the screening result is to be objectified and differentiated by means of more comprehensive diagnostics. They thus, have a filtering function, as they aim to identify individuals at risk in a larger population. Apart from these characteristics, screenings must also fulfill the general quality criteria for tests (reliability, validity, and objectivity) (Tröster, 2009; Moosbrugger and Kelava, 2020). The current assessment of early mathematical competencies in German-speaking countries is mainly carried out with standardized preschool math tests like the *Test mathematischer Basiskompetenzen im Kindergartenalter* [Test of basic math skills at kindergarten age] (MBK 0; Krajewski, 2018b), the *Würzburger Vorschultest* [Würzburg preschool test] (WVT; Endlich et al., 2016) or the *Mathematik- und Rechenkonzepte im Vorschulalter – Diagnose* [Preschool mathematics and numeracy concepts – Diagnosis] (MARKO-D; Ricken et al., 2013). Nonetheless, these are all typical test procedures with interactive material but no game elements.

Despite much research on the potential of linear board games in promoting early mathematical competencies, so far comparatively few attempts have been made to develop and examine the potential of game-based approaches to assessing these competencies. Most research on game-based assessment concentrates on the digital gamification of assessment of cognitive tasks in research (Lumsden et al., 2016) or in education (e.g., Ninaus et al., 2017). Gamification in this context is the transformation of simple test tasks by enhancing them with game-like features such as competition, narrative, and other game elements (Lumsden et al., 2016). The results show that in-game measures can reliably and valid assess students math knowledge (Ninaus et al., 2017). This finding speaks in favor of the applicability of a game-based approach as a diagnostic and research tool. One non-digital game-based approach is the *House of Numbers* (HoN), a screening instrument which allows the assessment of early mathematical competencies (Skillen et al., 2023). The HoN has been developed in the form of a linear board game. There are several reasons for developing a game-based assessment in the form of an analog linear board game rather than a digital assessment. Although preschool children grow up in a media-influenced environment, the findings on young children's exposure to digital media are controversial and recommendations suggest limited screen time (Linebarger et al., 2014; World Health Organization, 2019; Bochicchio et al., 2022; Deutsche Gesellschaft für Kinder- und Jugendmedizin e.V, 2022; Thorell et al., 2022; Paulus and Gerstner, 2023; Radesky et al., 2023). Whereas board games are fun to play, they offer an open space for creative expression and have an unpredictable course because it is not clear from the start who will win. They are a social activity, which can be easily integrated into everyday life. On top of that, children who may refrain from participating in a performance measurement, may be motivated to participate in a game-based evaluation (Ramani et al., 2012; Bayeck, 2020). Board games are also a familiar process, which adds to the content validity of the assessment. Children are usually familiar with the game process and the equipment used. In contrast, with unfamiliar tools, there is a risk that the respective mathematical competences are not tested, but the way in which the tools are used (Moser Opitz and Ramseier, 2012). Similarly, it could be argued that by digital tests, task presentation can be an impediment to content validity. In these tests, for example, the assessment of processing time depends on how the children use the keyboard or tablet. Finally, the board game format adds ecological validity to the assessment by ensuring that the tasks are related to reality and are not just hypothetical in nature (Moser Opitz and Ramseier, 2012).

The items of the HoN screening were developed in the manner of conventional assessment tools for preschool children (Ricken et al., 2013; Krajewski, 2018b) but are integrated in a linear board game. The screening HoN consists of a board depicting a house, probably a familiar motif for many preschool children. The house consists of five accelerating floors. There are 10 numbered doors on each floor. This alignment of numbers on each floor always progressing from left to right allows the representation of decades on the rows and units on the columns. Players use one token per person, two 10-sided dice with the Arabic numerals one to five evenly distributed on the 10 sides, task cards and further accompanying material such as colorful sticks. The screening HoN is used in a single setting with a child and a trained experimenter. Both players start at the entrance ("0") and work their way up the floors to the finish (door number 50). During a game session, the experimenter provides standardized instructions on how to play the game, sets standardized test items and documents the child's responses. Some items are set at specific points during the playing progress and others are masked by game tasks that must be solved when a task card is drawn (for item examples see Skillen et al., 2023). Although the game starts with easy items as "icebreakers," the difficulty of the items varies throughout the whole game. This helps explore all levels according to the developmental model the screening is based on. At the same time, this arouses the interest of children at all levels of competence to participate and finish the game. The standardization of items and procedures for administering and scoring the screening contributes to the objectivity of the HoN.

Previous research on the applicability, the psychometric quality (reliability and validity) of the HoN screening and its successful identification of children with below average math performance in a sample of 275 4–6-year-old children has demonstrated promising results (Skillen et al., 2023). The game-based setting allows the implementation of items while playing the linear board game. Yet, after conducting a pilot study (Skillen et al., 2023), further validation aspects need to be examined. Moreover, the item characteristics ought to be further analyzed to decide whether they remain in the final test version and considerations on the theoretical model need to be controlled (Moosbrugger and Kelava, 2020). Therefore, this study focuses on the game-based approach HoN with the primary aim to further investigate its quality aspects.

The HoN examined in this study is a 24-item version based on the model for the development of numerical concepts of Ricken et al. (2013). This empirically confirmed model (Fritz et al., 2018) describes the development of mathematical competencies through five stages extending from the age of about 4–8 years. In the first level (*count number*), children are able to compare two quantities, to name number words in the correct order, and to count out small quantities. However, they are not yet able to associate number words to individual entities. At the second level (*mental number line*) children mentally construct a successive alignment of numbers. That is, they can understand that numbers have a definite position and that some numbers proceed other numbers, which means that some numbers are also smaller than others. At this level children are also able to perform simple addition and abstraction tasks in smaller number ranges. The next level (*cardinality and decomposability*) describes the cardinal understanding of numbers. The understanding that a word number represents a number of elements. Furthermore, children at this level can understand that numbers as cardinal entities are composed of smaller quantities. Level four (*class inclusion and embeddedness*) describes the concrete understanding of the relationship between the total quantity and its partial quantities. That is, after understanding that a given quantity is composed of a number of elements, the child can understand the connection between the partial quantities. Moreover, it can understand the fixed relationship between them. As a result, if two partial quantities are known, a third can be determined. For example, having understood that the number five as a total amount consists of five elements, the child can now understand that a five can also be the total of two plus three or even one plus four elements. This leads on to recognizing differences between quantities as well as part-whole relationships. Children at this level can solve tasks that ask for the final quantity, an exchange quantity or the initial quantity. At the fifth level of this model (*relationality*) children deepen their understanding of the part-whole concept as well of the relational number concept and can conduct calculations modeling the difference between two quantities. Nonetheless, this model does not describe the development of mathematical competencies as a linear process. Rather, mathematical development up to the early school years is described as a process of overlapping waves (Siegler, 1995), with no sharp boundaries between the stages, but rather transitions (Fritz and Ricken, 2008).

Finally, as one of the goals of developing the HoN is to create an innovative assessment tool for children's competencies, one of its major features should be age-appropriateness. Previous research showed that the HoN can be reliably conducted with young children as young as 4 years of age (Skillen et al., 2023). However, further research is needed to investigate how children perceive this instrument and whether they find the game-based assessment of mathematical competencies an attractive procedure. To this end, children's views should be taken into account.

2 Study goals

The purpose of the present study was to investigate the potential of a linear board game with mathematical tasks for the assessment of early mathematical competencies. To the best of our knowledge, there is no such game-based instrument other than the HoN. However, there is a need for such valid, resource-saving and innovative approaches. The properties of the HoN have not yet been extensively examined. Therefore, four questions were investigated.

(1) Do children perform equivalently on a game-based test as on a standard test? To investigate this question, we compared the results of the HoN with the results of an established math test. We expected that the game-based setting would not affect the children's performance.

(2) Can the game-based approach be used to assess early mathematical competencies in a valid and reliable way? In particular, with regard to the construct validity of the HoN, we expected that younger children would score lower on the HoN than older children. We also expected the HoN to show convergent validity, that is, the performance on the HoN would correlate positively with other test scores of mathematical competence. Furthermore, we expected the items of the HoN to show good indices of model fit and reflect the levels of the theoretical model the HoN is based on. In terms of reliability, our expectation was that the HoN would show good item and person reliability.

The next question concerned the diagnostic validity of the HoN. More specifically, (3) can the game-based approach successfully identify children who show below-average performance in individual diagnostics and therefore children at risk of developing math difficulties? We evaluated the diagnostic performance and accuracy of the HoN as a screening tool by comparing its performance to an established diagnostic test. Previous research has shown good screening indices for the HoN, thus we expected to find equally good results for the 24-item version of the HoN.

(4) Is this game-based approach attractive for young children? The fourth aim of this study was to explore how children perceive and evaluate the game-based assessment setting and whether they prefer it to a standard procedure. To do so we explored children's opinions about the HoN and a standard test for this age group separately and in comparison to one another.

3 Materials and methods

3.1 Participants

A sample of $N = 147$ preschoolers was recruited from 15 kindergartens in Germany. The age of children ranged from 5 years to 6 years and 8 months (M_{age} = 5 years 10 months, SD_{age} = 5 months). Girls (71) and boys (76) were represented in roughly equal numbers. Kindergartens had been contacted by telephone. Those agreeing to participate received further information material via e-mail. Participation was voluntary. Ethical committee approval was obtained prior to testing. Only children to whom parents had given written consent were allowed to participate, and only if these children had given their verbal consent. From these cases, all children attended the last year of kindergarten before formal schooling and could understand German as a spoken language. One hundred twenty-four of the distributed parent questionnaires were returned fully completed. Only 2.4% of these children were not born in Germany. A total of 80.6% of them speak only German at home, 16.9% speak German and another language and 2.4% speak no German at home. The mother of 17.7% of these children was born in a country other than Germany. The highest German school leaving certificate (Abitur) is held by 49.2% of the mothers and 54% of the fathers. In addition, a subsample of $n = 47$ children (M_{age} = 5 years 9 months, SD_{age} = 5 months; 18 girls and 29 boys) participated in a survey asking about their opinions of the instruments in which they had taken part.

3.2 Instruments

3.2.1 Assessment of early mathematical competencies

The linear board game HoN (Skillen et al., 2023) that allows assessment of early mathematical competencies was administered by trained experimenters to all participating children. The HoN was administered in a single session and lasted on average 21:32 min ($SD_{duration}$ = 04:48 min; $minimum_{duration}$ = 08:00 min; $maximum_{duration}$ = 42:11 min). The version of the HoN used in this study consisted of 24 dichotomous items based on the model for the development of numerical concepts of Ricken et al. (2013). The experimenter documented one point for each correctly solved item. The result was a raw score out of a maximum of 24 points for each child. This raw score was then converted into a percentage score.

Early mathematical competencies of all participating children were also examined using the MARKO-D (Ricken et al., 2013). The MARKO-D is a standardized preschool mathematics test that consists of 55 dichotomous items based on the same model mentioned above (Ricken et al., 2013). As with the HoN, the MARKO-D takes place in a single setting with a trained experimenter and a child (Ricken et al., 2013). In this study, the MARKO-D took on average 27:23 min ($SD_{duration}$ = 04:58 min; $minimum_{duration}$ = 18:00 min, $maximum_{duration}$ = 45:00 min) to complete. The item difficulty is varied throughout the test. To interpret the test results, a raw total score is obtained for each of the 55 items achieved and a percentage score can be calculated from the raw score. The excellent test quality characteristics of the MARKO-D and the common theoretical construct it shares with the HoN (Ricken et al., 2011, 2013) makes it a reasonable option for testing the HoNs quality characteristics. In this study person reliability of the MARKO-D was 0.92 and item reliability 0.98.

3.2.2 Assessment of the perception of the instruments used

The subsample ($n = 47$) was also surveyed individually by trained experimenters about their perceptions of the math instruments in which they had participated. A questionnaire consisting of five items was used for each of the two instruments. The first item was a dichotomous question about whether the instrument was a game or not. The next three items were in a four-point Likert scale format and asked how much fun the instrument was, whether the child would like to do it again and whether it is a pity that it is over. The fifth item was a preference question between the two instruments and was therefore only asked once after the child had participated in both instruments. The answering format of the first four items was based on the format of the picture scale *Bildskala zur Erfassung von Lernfreude und Selbstkonzept in Mathematik und Schriftsprache bei Vorschulkindern* [Picture scale for the assessment of learning enjoyment and self-concept in preschool mathematics and language abilities] (BLSL-MS; Roux et al., 2010). The BLSL-MS is a further development of the original picture scale of Harter and Pike (1984) and is used mainly in the self-concept research. The picture-based approach makes it easier for preschool children to respond on a multi-level scale. In this study, the experimenter presented a picture of two children for each item. While pointing to one child, the experimenter said, for example, that this child enjoyed the HoN. The experimenter would then point to the other child and say that this child did not enjoy the HoN. The child would then be asked to point to the child who was like them. Depending on which child was chosen, a further distinction was made between the two options no fun at all and somewhat fun or quite a lot of fun and a lot of fun by pointing to a smaller or larger circle under each child (four circles in total). For a better understanding see **Figure 1**. No scala was used for the fifth item, as the child was only asked about its preference. The interview took approximately 4 min to complete for each child.

3.3 Design

This study uses a cross-sectional design. Data were collected as part of larger studies during three survey periods between March 2019 and June 2022. Each child was tested individually with the HoN and the MARKO-D in a quiet kindergarten room during a normal kindergarten day. For each math assessment there was one session on a different day. A random part of the sample was tested first with the HoN and the other part first with the MARKO-D. After participating in both instruments, the subsample was administered the survey on the perception of the assessment tools.

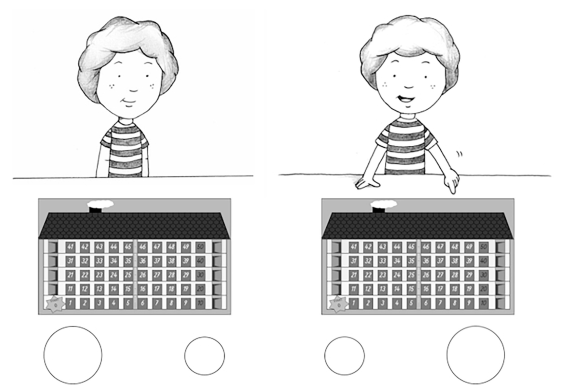

FIGURE 1
Example of the picture scale response format. Here is item number four about whether it is a pity that the HoN is over. The picture scale of this study was developed based on the scale of Roux et al. (2010). The original graphics were created by Caroline Reusch.

3.4 Analytic strategy

There were 1.99% of missing values for the MARKO-D and 0.2% of missing values for the HoN. All missing values were examined individually. Most missing values occurred due to performance-related reasons. The experimenter skipped one or more items because the child could not answer equivalent items correctly or could not answer related items correctly. For example, a child would be asked the same question in succession but with different numbers. In another example, a child would be asked which number his pawn is on. The next two items would ask the child to name the number before and the number after that number. The next item would then be the question "How many do you need to get to 10?" If the child could not answer the previous questions correctly, he or she will most likely not be able to answer this question. Missing values due to such performance-related problems were defined as "item not achieved" and 0 points were awarded for this item. Finally, we compared the results with and without imputed data. They remained the same.

Bayesian inference has the advantage over frequentist methods of being able to consider evidence for, as well as against the null hypothesis (van Doorn et al., 2021), and was therefore used for some analyses in this study. The Bayesian approach uses the Bayes factor (BF) to evaluate the degree of probability of one hypothesis over the other. When testing for hypotheses, the subscript of the BF indicates which hypothesis is being addressed. The BF_{10} signifies the BF in favor of the alternative hypothesis (H_1) over the null hypothesis (H_0), whereas the BF_{01} indicates the BF in favor of the H_0 over the H_1. A BF can range from 0 to ∞. According to van Doorn et al. (2021), a BF of 10 or larger (or 0.1% and smaller) indicates strong evidence for the respective hypothesis (i.e., in favor of the H_1 in case of a BF_{10} and in favor of H_0 in case of a BF_{01}). Correspondingly, a BF between 3 and 10 (or between 0.1 and 0.33) indicates moderate evidence in favor of the respective hypothesis. A BF of 3 is typically used as a cut-off criterion to determine the existence (BF_{10}) or non-existence (BF_{01}) of an effect (Jeffreys, 1998; Dienes, 2014). Thus, BF $\geq$ 3 will also serve as the cut-off criterion of this study. As there was no basis for a more informed choice of prior, default priors (JASP Team, 2024) were used. Preliminary analyses of the math performance were conducted using the Bayesian approach to examine possible differences by period of assessment, kindergarten, experimenter, gender, or test taken first. Specifically, Bayesian analyses of variance and Bayesian independent samples t-tests were performed to examine the evidence against a difference, with the total sum of points achieved in MARKO-D as the dependent variable. All preliminary analyses resulted in $BFs_{01} \geq 3$, which indicates support for the H_0: $\delta = 0$ against the one- or two sided H_1: $\delta > 0$, H_1: $\delta < 0$ or H_1: $\delta \neq 0$. Thus, indicating that it is more likely that there are no differences between the respective conditions given the data than that there are differences (van Doorn et al., 2021).

To examine the first research question and thus confirm the equivalence of the performance results in both tests, an equivalence Bayesian paired-samples t-test comparing the total percentage scores in both tests was performed. This BF approach to equivalence testing (Morey and Rouder, 2011) uses alternative BFs. According to this approach, the H_1 describes that the effect size (δ) is defined by an unrestricted prior distribution. Two restricted versions of the H_1 describe that the δ falls inside the equivalence interval (I) ($H_{\in}$), or that the δ falls outside the I ($H_{\notin}$). The so-called overlapping hypotheses (OH) BF examines the probability of the H_1 against the two restricted $H_{\in}$ and $H_{\notin}$. The non-overlapping hypotheses (NOH) BF shows the probability of the $H_{\in}$ against the $H_{\notin}$.

Independent samples t-tests (Student and Mann–Whitney) were used to test whether both instruments were sensitive to age differences in total percentage performance. After checking for sources of bias, a correlation was performed to investigate convergent validity using Pearson's correlation coefficient and Spearman's rho between the total percentage scores on each instrument. Due to the limited sample size sensitivity power analyses were performed. These analyses were tested for the independent t-tests between the age groups and for the correlation analysis to determine the minimum effect sizes that the tests were sufficiently sensitive to detect in the context of this study (statistical power; Faul et al., 2007). The power of the $1 - \beta$ probability error was set at the 95% confidence level for both power analyses. Based on the item response theory, Rasch analysis to examine the model fit, the item characteristics and the internal consistency of the HoN was run (Boone, 2016).

To investigate the screening characteristics of the HoN, a receiver operating characteristic (ROC) analysis and a crosstabs procedure were carried out. According to Tröster (2009) the quality indices sensitivity (SEN) and specificity (SPE) should not be viewed as invariant quality screening characteristics because they also depend on the selection and basic quota. Therefore, the relative improvement over chance (RIOC) index (Marx, 1992) was also used.

To explore the children's perception of the two instruments, the percentage frequencies of their responses were examined. Chi-square tests were calculated to check whether the children's answers to the interview items differed between the established and the game-based method. Finally, sensitivity power analyses were calculated for the Chi-square tests.

These calculations were operated using SPSS 29.0.0 (IBM Corp, 2022), JASP 0.18.3.0 (JASP Team, 2024) and Winsteps 5.7.0 (Linacre, 2024a). Power analyses were performed using G*Power

TABLE 1 Total percentage scores and percentage scores for each model level in both math instruments.

	M	SD	Min.	Max.	Skewness	Kurtosis
HoN total score	61.96	18.86	8.33	100	−0.30	−0.46
HoN Level 1	90.14	14.28	25.00	100	−1.79	3.69
HoN Level 2	65.65	34.27	0	100	−0.59	−0.93
HoN Level 3	61.90	29.64	0	100	−0.46	−0.70
HoN Level 4	41.72	35.72	0	100	0.29	−1.18
HoN Level 5	17.18	24.45	0	100	1.47	1.67
MARKO-D total score	59.85	18.09	12.73	94.55	−0.15	−0.63
MARKO-D Level 1	93.92	9.54	37.50	100	−2.80	11.04
MARKO-D Level 2	71.43	26.59	0	100	−0.84	−0.12
MARKO-D Level 3	53.12	29.78	0	100	−0.14	−1.21
MARKO-D Level 4	39.18	32.26	0	100	0.32	−1.08
MARKO-D Level 5	20.12	25.52	0	100	1.29	0.74

N = 147. Five levels according to the model of Ricken et al. (2013).

TABLE 2 Equivalence Bayesian paired samples *t*-test.

	Model comparison	BF	Error %
MARKO-D – HoN	$\delta \in \mathrm{I}$ vs. H_1	6.782	6.015×10^{-6}
	$\delta \notin \mathrm{I}$ vs. H_1	0.728	5.605×10^{-5}
	$\delta \in \mathrm{I}$ vs. $\delta \notin \mathrm{I}$	9.318	8.756×10^{-6}
	$\delta \notin \mathrm{I}$ vs. $\delta \in \mathrm{I}$	0.107	7.603×10^{-4}

δ = effect size; I = equivalence interval. I ranges from −0.05 to 0.05.

TABLE 3 Equivalence mass table.

	Section	Mass
MARKO-D – HoN	$p(\delta \in \mathrm{I} \mid \mathrm{H}_1)$	0.045
	$p(\delta \in \mathrm{I} \mid \mathrm{H}_1, \text{data})$	0.305
	$p(\delta \notin \mathrm{I} \mid \mathrm{H}_1)$	0.955
	$p(\delta \notin \mathrm{I} \mid \mathrm{H}_1, \text{data})$	0.695

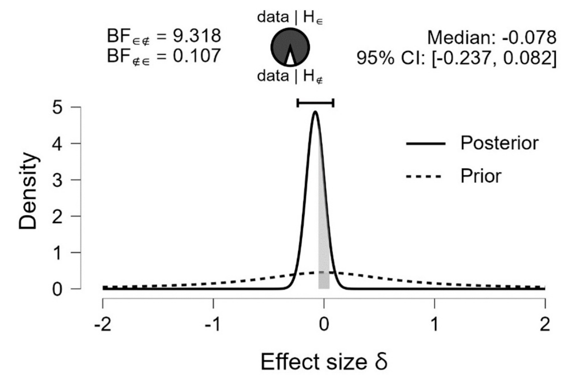

FIGURE 2
Prior and posterior plot. The dashed line represents the prior distribution. The solid line represents the posterior distribution. The gray area represents the specified equivalence region.

3.1.9.7 (Faul et al., 2009). All programs used in this study were run with Windows 11 operating system.

4 Results

4.1 Effect of the setting

Descriptive statistics for the HoN and the MARKO-D are shown in Table 1. The total percentage scores both for the HoN and the MARKO-D appear to be approximately symmetric and platykurtic in distribution, suggesting that most of the children scored average, but in a flat distribution. The mean total percentage score for the MARKO-D was 59.85 (SD = 18.09) and for the HoN 61.96 (SD = 18.86). As the descriptive statistic show, the children scored slightly better in the game-based than in the standardized assessment. An equivalence Bayesian paired samples *t*-test was used to examine whether the total percentage scores of the two tests were equivalent. The OH BF in favor of the $\mathrm{H}_{\in}$ was 6.782 (error 6.015×10^{-6} %), and the NOH BF in favor of the $\mathrm{H}_{\in}$ was 9.318 (error 8.756×10^{-6} %) (see Tables 2, 3 for further results). Both BFs show moderate evidence that the δ falls in the I, and that the scores are not meaningfully different from each other. As Figure 2 also shows, the median of the resulting posterior distribution for the standardized effect size δ equals −0.08 with a central 95% CI [−0.237, 0.082]. According to the same figure, the mass inside the I has increased from prior to posterior, which also speaks in favor of a parameter estimate inside the I. That is, there is evidence to support the equivalence of the results.

4.2 Reliability and validity

As far as construct validity is concerned, the 6-year-old children scored higher than the 5-year-old children in both instruments. The independent samples *t*-tests revealed significant math performance differences between the 5 and the 6-year-old children in both instruments (see Table 4). This suggests that both tests could detect

TABLE 4 Percentage math performance by age group in each test and comparison for each test.

	5-year-olds		6-year-olds							
	n	*M* (SD)	*n*	*M* (SD)	Test	Statistic	df	*p*	Effect size	SE effect size
HoN	93	58.78 (18.74)	54	67.44 (17.94)	Student	−2.74	145	0.007	−0.47	0.175
					Mann–Whitney	1836.00		0.007	−0.27	0.099
MARKO-D	93	55.62 (17.89)	54	67.13 (16.11)	Student	−3.90	145	<0.001	−0.67	0.178
					Mann–Whitney	1593.00		<0.001	−0.37	0.099

For the Student's t-test, effect size is given by Cohen's d. For the Mann–Whitney test, effect size is given by the rank biserial correlation.

age-related differences in performance. However, a sensitivity power analysis showed that the independent samples t-test with the given N could reliably detect effect sizes of $d = 0.62$ or larger. The limitation of the medium effect size $d = -0.47$ for the HoN must be considered in this case. To evaluate the convergent validity of the HoN, the degree of agreement between the total percentage scores of the MARKO-D and of the HoN was examined. There was a significant correlation between the two assessments $r = 0.85$, 95% CI [0.79, 0.89], $p < 0.001$. Non-parametric analysis showed the same result $\rho = 0.84$, 95% CI [0.88, 0.78], $p < 0.001$. The sensitivity analysis against the null hypothesis of a null correlation resulted in a correlation parameter of $\rho = 0.29$, 95% CI [−0.16, 0.16]. That is, the results of the high correlation between the two tests can be treated as a clear result.

Regarding further aspects of construct validity and the internal consistency of the HoN, the results of the Rasch analysis provide information on the model fit and the reliability of the 24 items of the HoN. As shown in **Table 5**, the Rasch analysis indicated Mean-square (MNSQ) item Infit values of 1 ± 0.20 for all HoN items. The distribution of the HoN items on the logit measurement ruler of the Person-Item Map corresponds, with the exception of three items, to the five levels of the Ricken et al. (2013) model based on which these items were developed (see **Figure 3**). The HoN displayed a person reliability of 0.82, an item separation of 6.51 and an item reliability of 0.98. Person reliability is similar to reliability indices of the classical test theory (Linacre, 2024b). According to this, the test reliability of 0.82 is considered good, since values close to 1 suggest a more internally consistent instrument (Boone et al., 2014). The item reliability depends on the range of item measures and the size of the sample (Linacre, 2024b). According to Linacre (2024b) a low item separation (<3) and an item reliability <0.90 suggests that the sample size was inadequate to confirm the construct validity of the measure. In this study, the high item separation and reliability suggest that the sample was large enough to confirm the construct validity of the HoN.

TABLE 5 HoN Rasch analysis item statistics.

		MNSQ	
Model level	Item	Infit	Outfit
Level 1	Item 1	1.20	0.62
	Item 3	1.08	1.28
	Item 4	1.09	1.18
	Item 7	0.89	0.70
	Item 17	0.82	0.84
	Item 19	0.99	0.85
	Item 23	1.10	0.89
	Item 5	0.92	0.80
Level 2	Item 12	0.89	0.78
	Item 13	1.01	0.86
	Item 21	0.89	0.73
	Item 9	1.04	1.02
Level 3	Item 6	1.17	1.33
	Item 18	0.90	0.72
	Item 20	1.06	1.08
	Item 24	0.92	0.90
	Item 11	1.06	1.04
Level 4	Item 2	0.88	0.80
	Item 22	0.95	1.03
	Item 14	1.05	1.03
Level 5	Item 8	0.91	0.64
	Item 10	0.95	1.01
	Item 15	1.03	1.57
	Item 16	1.11	1.24

N = 147. Model levels according to the model of Ricken et al. (2013) for the development of numerical concepts. MNSQ = mean-square of weighted responses.

4.3 Screening quality indices

To examine the screening diagnostic quality characteristics of the HoN, a ROC analysis was performed using a MARKO-D t-value of <40 as the state variable. The results show an area under the curve (AUC) of 0.97, $p < 0.001$, SE = 0.01, 95% CI [0.94, 1.00]. Due to the splitting of the ROC curve in the upper left quadrant, there are three possible cut-off values, which can be seen in **Figure 4**. A comparison of the quality indices of the three possible cut-off scores can be seen in **Table 6**. The sample relevant cut-off z-score

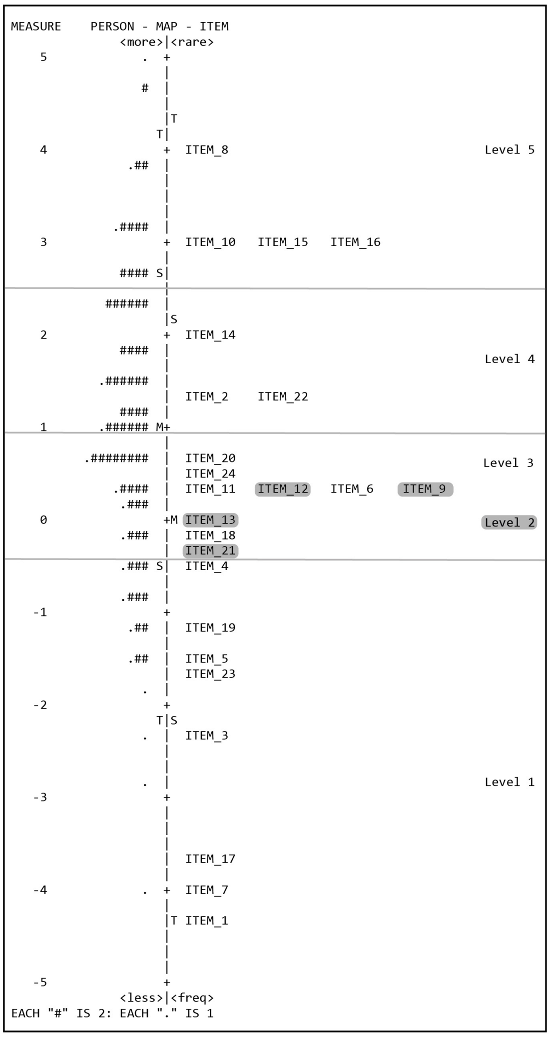

FIGURE 3
HoN Person-Item Map. N = 147. On the left side of the logit span the participants are shown according to their performance (the higher the more items scored). On the right side the items are displayed according to their difficulty (the higher the more difficult the item). The five levels of the Ricken et al. (2013) model are marked on the distribution of items on the logit span.

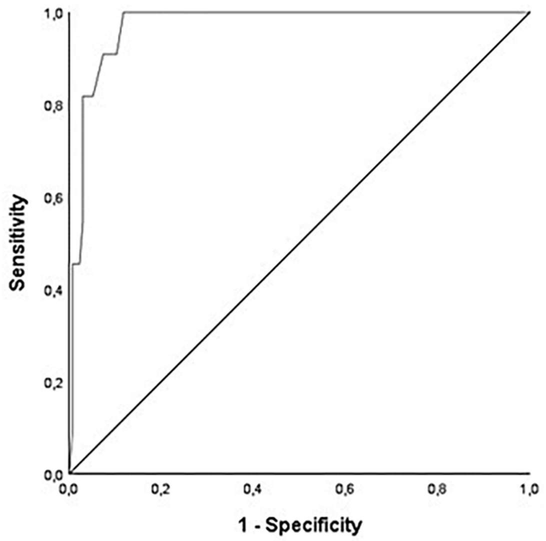

FIGURE 4
Receiver operating characteristic (ROC) curve of the HoN. N = 147. The diagonal line represents the case of random prediction. The ROC curve (on the left) represents the diagnostic performance of the test. The area under the ROC curve represents the predictive quality of the test.

TABLE 6 Comparison of the quality indices of the potential cut-off values.

HoN cut-off *z*-scores	SEN	SPE	YI	RIOC index
−1.06	90.9	92.6	0.84	0.84
−0.90	100	88.2	0.88	1.00
−0.81	100	84.6	0.85	1.00

N = 147. SEN = sensitivity; SPE = specificity; YI = Youden index; RIOC index = relative improvement over chance index.

of −0.90 with the highest Youden Index (YI = SEN + SPE−1) was chosen. With this cut-off score the HoN demonstrated a SEN of 100% and a SPE of 88.2%. The RIOC index was 100%. See Table 7 for further results.

4.4 Children's perceptions of the assessment

According to the results of the survey on children's perception of the instruments used, and as can be seen from Table 8, 100% of the children in the subsample (n = 47) thought that the HoN was a game. By contrast, only 66% of the children also thought the MARKO-D to be a game. We further calculated a Chi-square test to see if there was a significant association between the instrument and whether or not it was perceived as a game. The result was significant χ^2 (4) = 19.28, $p < 0.001$ with a medium effect $\varphi = -0.45$ ($p < 0.001$). A sensitivity power analysis for df = 4 was calculated. The result showed that the Chi-square test with a 95% power ($1 - \beta$ error probability) would be able to detect a critical $\chi^2 = 9.49$. The majority of the children (63.80%) said that they had a lot of fun with the HoN. In contrast only 42.60% said they had a lot of fun doing the MARKO-D. Regarding the questions about repetition and whether it's a pity that it's over, the children gave similar answers in both instruments. Most of the children would very much like to do them again. As to whether it's a pity that it's over, the children gave evenly distributed answers at all levels of the Likert scale. Further Chi-square analyses were calculated for the three Likert scale items to test whether there was a relationship between the instrument and the child's response on the Likert scale. None of them showed significant results (see Table 9 for more details). A sensitivity power analysis for df = 3, showed that the 95% power Chi-square test would be able to detect a critical $\chi^2 = 7.81$. That is, the results of the Chi-square tests for the Likert scale items did not show efficient power to reliably detect effects smaller than the critical χ^2. Finally, 74.5% of the children said that they would rather do the HoN again than the MARKO-D.

TABLE 7 Crosstabs of children identified as low-achieving and therefore at risk by the two instruments.

At risk cases as identified by the MARKO-D	At risk cases as identified by the HoN		
	Negative	Positive	Sum
Negative	TN = 120 (100%)	FP = 16 (59.3%)	136 (92.5%)
Positive	FN = 0 (0%)	TP = 11 (40.7%)	11 (7.5%)
Sum	120 (100%)	27 (100%)	147 (100%)

N = 147. HoN cut-off *z*-score = −0.90. TN, true negative; FN, false negative; TP, true positive; FP, false positive.

TABLE 8 Results of the children survey for the HoN and the MARKO-D.

Item	Answer frequency, %							
	HoN				MARKO-D			
1. Game vs. no game	No game		Game		No game		Game	
	0%		100%		34%		66%	
2. It was fun	No fun at all	It was somewhat fun	It was quite fun	It was a lot of fun	No fun at all	It was somewhat fun	It was quite fun	It was a lot of fun
	2.10%	6.40%	27.70%	63.80%	0%	19.10%	38.30%	42.60%
3. Would like to do it again	Not at all	To some extent	Quite like	Very much	Not at all	To some extent	Quite like	Very much
	12.80%	17%	23.40%	46.80%	17%	21.30%	27.70%	34%
4. It's a pity that it's over	Not at all	Quite	Somewhat	Very	Not at all	Quite	Somewhat	Very
	29.80%	23.40%	25.50%	21.30%	25.50%	23.40%	23.40%	27.70%
5. Repeat	HdZ				MARKO-D			
	74.50%				25.50%			

n = 47.

5 Discussion

Although a considerable amount of research has been carried out on the game-based promotion of early mathematical competencies with linear board games (Hauser et al., 2014; Skillen et al., 2018a; Gasteiger and Moeller, 2021; Lange et al., 2021), only few studies have examined their potential for the preschool assessment of these competencies (Skillen et al., 2023). Thus, the purpose of the present study was to examine the potential of the 24-item screening HoN for the assessment of early mathematical competencies.

The first question of this study was whether the game-setting would affect the children's performance in a preschool math test. To investigate this question, we evaluated the equivalence in the performance outcomes between the newly developed game-based assessment HoN and the established standardized test for mathematical competencies in preschool MARKO-D. The MARKO-D is based on the same theoretical model on the development of mathematical competencies as the HoN (Ricken et al., 2013). We expected that the game-based setting would not affect children's performance. The results show that children scored slightly higher in the HoN than in the MARKO-D. The small difference in the mean performance between the two assessments is in line with the findings of Skillen et al. (2023) and also with studies showing higher performance on cognitive tasks compared to the same cognitive tasks when gamification features have been added (Mekler et al., 2013; Ninaus et al., 2015). However, in this study, although there is a slight difference between the two scores, the results also suggest that the performance scores on the two tests are probably equivalent. In any case, the experimental treatment can be ruled out as a possible explanation as the children were randomly divided into two groups and preliminary analyses showed a probability of no differences depending on whether they attended the HoN or the MARKO-D appointment first. The task format of an assessment is crucial for the reliability of a preschool test. Non-mathematical requirements during the test, such as the strain on working memory due to exclusively verbal tasks or the motor demands of handling objects, must not interfere with the results (Gloor, 2023). It could be expected that the game material and the actions of the HoN might influence the reliability of the test, because they demand attention and processing capacity, or require greater involvement of the central executive, but this result is not given. Against the expectation that the game context might impose additional cognitive demands that could negatively affect performance, the performance in the HoN was higher than in the MARKO-D. This finding is particularly relevant with regard to comorbid attention deficits of dyscalculia. Children with learning deficits, among them dyscalculia, show an increased risk of developing attention/hyperactivity disorders (Schuchardt et al., 2015). Hence, the characteristics of the test approach may influence the performance. One further example of such a characteristic, is the duration of the test. The HoN took approximately 21:32 min to complete. This is an excellent duration for a test for 5–6-year-olds (Gloor, 2023). According to the recommendation of Wyschkon

TABLE 9 Frequencies and Chi-square results for the first four items of the survey.

Item	Response	HoN	MARKO-D	χ^2	df	*p*
1. Game vs. no game	Yes	47	31	19.28	4	<.001
	No	0	16			
2. It was fun	No fun at all	1	0	6.80	3	0.078
	It was somewhat fun	3	9			
	It was quite fun	13	18			
	It was a lot of fun	30	20			
3. Would like to do it again	Not at all	6	8	1.62	3	0.654
	To some extent	8	10			
	Quite like	11	13			
	Very much	12	16			
4. It's a pity that it's over	Not at all	14	12	0.59	3	0.899
	Quite	11	11			
	Somewhat	12	11			
	Very	10	13			

$n = 47$.

(2015), test duration for this population, even with interactive test procedures, should not exceed 60 min, because preschool children are not used to working for so long on a task determined by others and often have short attention spans. One plausible explanation of why children scored higher in the HoN than in the MARKO-D, could lie in the nature of the setting of the HoN. Mathematics assessment is often associated with preschoolers' experiences of negative affect and worry (Lu et al., 2021). It is possible, that children who do not perceive the process of the HoN as a test performance process, may show better performance. On top of that, the exciting game-setting, which allows the engaging in social interaction and the use of interesting materials such as moving game tokens, rolling dice, and drawing task cards, may promote the motivation to participate, which in turn also may have a positive effect on the performance of the children. As Hirsh-Pasek et al. (2009) put it, playful children are motivated children. Similarly, Gelman (2006) observed that children as young as 2 years and 6 months could successfully perform tasks of prediction of items and counting when they were embedded in a play task, but not when they were bluntly asked to count in a count-only task. Further research is needed to investigate whether the small performance difference is due to the HoN setting characteristics or due to other unknown factors.

Regarding the second question of this study and according to our expectations, the results show that both instruments were able to reliably detect differences in performance between the groups of the 5- and the 6-year-olds. This finding is also consistent with the results of Skillen et al. (2023) who investigated whether mathematical performance in the game-based approach differed between the groups of 4-, 5-, and 6-year old children. Such age-related differences in performance are expected according to the developmental model, which posits that older children excel their skills and develop further competencies that allow them to deal with more complex mathematical understanding (Ricken et al., 2013). To determine the convergent validity of the HoN, the degree of agreement between the mean percentage scores of the HoN and the MARKO-D was examined. In line with our hypothesis, the positive correlation between the new and the established tool of 0.85 is high according to Cohen (1988) and it shows that probably both tools evaluate the same mathematical concepts. This result is also in accordance with the findings of earlier evaluation of the psychometric characteristics of the HoN (Skillen et al., 2023) and the findings of studies on gamification of cognitive testing (Lumsden et al., 2016). The latter found relatively good correlations between gamified assessments and their non-gamified counterparts, which provide evidence that cognitive tests can be gamified and remain useful research tools. Furthermore, as expected, in the results of the Rasch analysis all items showed a model fit between 0.82 and 1.20, which according to Wright and Linacre (1994) indicates a reasonable model fit. These MNSQ values suggest that there is neither too much unexplained variance in the data nor that the model overpredicts the data causing inflated reliability statistics (Boone et al., 2014). The plotting of the items on the logit range of the Person-Item Map confirmed with only three exceptions the levels of the model of Ricken et al. (2013). The three overlapping items can be assigned to the level two (mental number line) and three (cardinality and decomposability). In line with the overlapping theory (Siegler, 1995), skills acquired at level two are also required to solve tasks that address level three. This could be one possible interpretation of this overlap, which also appears in established tests (see MARKO-D; Ricken et al., 2013). Finally, the high values of item and person reliability for the HoN suggest that item parameters calculated from the current data are generalizable for similar person and item samples (Boone et al., 2014).

The third question in this study was whether the HoN can successfully predict later difficulties in math. To evaluate the diagnostic validity of the HoN screening, its performance in detecting potential "at risk" cases was compared with the performance of the MARKO-D. The sample of this study was randomly selected. A total of 7.5% of the children were identified by the MARKO-D as having below average math performance compared to the expected performance for their age and therefore at risk of developing dyscalculia. This finding reflects the prevalence rates of dyscalculia, which lie within approximately 2%–8%, making the sample representative (Wyschkon et al., 2009; Fischbach et al., 2013). According to our hypothesis, the results of the ROC analysis are overall promising. The high AUC of 0.97 manifests the ability of the HoN to successfully distinguish between children with age-appropriate expected math skills and children possibly at risk. A sample related cut-off value was defined. The resulting high rates of SEN and SPE indicate a high likelihood that the HoN will be able to detect a possible impairment in early mathematics and a good likelihood that the absence of an impairment will be also successfully detected by the HoN. In particular, the highest possible SEN is desirable for a screening test. A high proportion of false positives is accepted in view of the concept of a screening test as a coarse screening procedure, which should be followed by a fine diagnosis for all "at-risk" children (Marx and Lenhard, 2011). The RIOC index above 66%, according to Jansen et al. (2002), can be described as a very good classification. The high psychometric results of this study are similar to the high results of Skillen et al. (2023). Testing the 23-item version of the HoN based on the model of Ricken et al. (2013), they found a high AUC value of 0.86, $p < 0.001$, 95% CI [0.78, 0.93]. After defining two possible sample-relevant cut-off z-scores the same authors found good screening indices. In detail, using the z-scores of -0.63 (percentile rank 26%) and -0.24 (percentile rank 41%) as resulting cut-off scores, they found a SEN of 79% and 95%, a SPE of 76% and 65% and high RIOC indices of 71% and 91%, respectively.

The fourth question of this study was whether the game-based HoN is attractive for young children. The aim of this question was to shed light on how children perceive this newly developed instrument. To examine this question the preschool children's perception of the HoN was investigated with a short survey. The explorative findings show that all surveyed children perceived the HoN as a game. In contrast, only 66% of the children also perceived the MARKO-D as a game. This result shows that the children perceived a difference between the HoN and the MARKO-D. In addition, 74.5% of the children chose the HoN over the standard assessment. This finding is also consistent with findings from gamification studies that have compared gamified and non-gamified tasks, showing that intrinsic motivation is enhanced by the use of the gamified tasks (Lumsden et al., 2016). The above results support further development of the game-based HoN as a tool with potential in the preschool population.

5.1 Implications

Additional research is required to examine further (validation) aspects of the screening HoN. For example, it would be interesting to evaluate the test-retest reliability of the HoN, the extent to which the HoN also overlaps with other competence tests for preschool mathematics, or with tests of other competence areas (divergent validity). In future studies, further information could also be obtained to better determine the influence of nonspecific cognitive or environmental predictors of later mathematical performance. Such predictors include intelligence, home numeracy environment and working memory. In addition, to further investigate the potential of the HoN screening, it could be applied to different populations. The focus of this study was the age group attending the last year of kindergarten (5–6-year-olds). In future studies, the test could also be administered to younger or older children (e.g., first graders) to evaluate whether the HoN still responds to expected developmental difficulties. Alternatively, it could be tested in different kinds of populations, such as primary school children who have already been diagnosed with dyscalculia. In this way, it could be demonstrated whether the HoN correctly identifies cases who show difficulties with mathematics. Another possibility would be to examine the HoN properties based on other theoretical models (e.g., Krajewski, 2018a).

Once multiple aspects of the screening have been thoroughly reviewed and all characteristics of the HoN have been clarified, the next step would be to collect sufficient data on performance in several age groups and to establish population-based reference scores for the HoN (norming). Our aim for the future is to develop the HoN as an instrument that can be used by (preschool) educators. Further research is therefore important to investigate how these educators would apply the HoN and what their specific needs are with regard to this instrument.

5.2 Limitations

An apparent limitation of this study is the small sample size. The limited sample size has an impact on the power of some of the statistical analyses in this study. The first reason for this limitation is the nature of this type of assessment. Data collection in kindergartens can be very time consuming. At the same time, many kindergartens in Germany are understaffed and overburdened. As a result, some kindergartens did not agree to participate. Furthermore, even among those that did participate, we often faced the problem of missing assessment sessions, because preschool children were often sick and stayed at home. In addition, the almost 2 years of the COVID-19 pandemic were a great loss of potential data, as external persons were not allowed access to educational institutions in Germany for longer periods of time. Even so, the fact that the HoN already shows good results despite the small sample size is very promising.

Another limitation of this study is that other variables, such as intelligence, were not measured for the whole sample and therefore could not be documented in this article. This limits the description of the sample and the comparison of various aspects of performance. Moreover, there were no fidelity controls other than comparing the average children math performance between experimenters for differences. However, all experimenters received training and had to obtain supervisor approval after demonstrating the testing procedure to be allowed to go into the field. In addition, a detailed documentation of the data, including the dates and start and end times of the testing for each measure can help to reconstruct the conditions of the procedure.

Lastly, as far as the children's survey is concerned, and as the results of the Chi-square tests show, the Likert scale items were probably difficult for the young children to understand, despite the age-appropriate adapted format. This may also explain why the children responded similarly between the HoN and the MARKO-D to these items. In addition, the exploratory findings on children's perception of the tools should be treated as tentative, as preschool children tend to exhibit a response bias when questioned by giving affirmative answers (Okanda and Itakura, 2010). Similarly, the children in this study may have given more positive responses. In future studies, the use of mixed methods with mainly open-ended question formats should be the preferred method of questioning in a preschool sample. Nevertheless, our aim in this study was achieved, as it was only a first attempt to describe in an exploratory way what children think about the HoN.

6 Conclusion

Overall, the results so far suggest that the game-based approach can be used to assess early mathematical competencies. Moreover, they suggest that the HoN is a game-based assessment, which can reliably and effectively evaluate early mathematical skills and detect possible cases at risk for developing math difficulties. The HoN represents an attractive option for preschool children in comparison to standard preschool testing procedures. It is an innovative screening tool that combines requirements of screening procedures with the positive characteristics of board games. Screening procedures provide identification of a possible risk, which is linked to a recommendation for support. According to the findings, the HoN screening could provide reliable support to early childhood professionals in a time and resource efficient, and also enjoyable naturalistic way. The HoN could be easily integrated into the work that preschool educators already do. By eliminating or reducing the deficits in the precursor skills successfully identified by the screening, the chances of at-risk children to later acquire mathematical competencies can be increased through selective prevention for those at risk.

Data availability statement

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

Ethics statement

This study involves human subjects and has been approved by the Ethics Committe of the Catholic University of Eichstätt-Ingolstadt – Faculty of Philosophy and Education: Prof. Dr. Marco Steinhauser (Chair), Dr Valérie Berner, Dr. Martin E. Maier, Jun.-Prof. Dr. Christina Pfeuffer, and Prof. Dr. Michael Zehetleitner. The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation in this study was provided by the participants' legal guardians.

Author contributions

M-AC: Conceptualization, Data curation, Formal analysis, Funding acquisition, Methodology, Project administration, Validation, Visualization, Writing – original draft, Writing – review & editing. JS: Conceptualization, Data curation, Methodology, Project administration, Writing – review & editing. GR: Writing – review & editing. KS-S: Conceptualization, Methodology, Resources, Supervision, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project number 512640851.

Acknowledgments

We would like to thank all children, parents, and early childhood professionals who participated in our studies. A special thank you goes to all students involved in our research, as well as to work colleagues and fellow researchers for their support and collegial advice.

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

American Psychiatric Association (2013). *Diagnostic and statistical manual of mental disorders*, 5th Edn. Washington, DC: American Psychological Association.

Anders, Y., Grosse, C., Rossbach, H., Ebert, S., and Weinert, S. (2013). Preschool and primary school influences on the development of children's early numeracy skills between the ages of 3 and 7 years in Germany. *Sch. Effect. Sch. Improv.* 24, 195–211. doi: 10.1080/09243453.2012.749794

Aunola, K., Leskinen, E., Lerkkanen, M., and Nurmi, J. (2004). Developmental dynamics of math performance from preschool to grade 2. *J. Educ. Psychol.* 96, 699–713. doi: 10.1037/0022-0663.96.4.699

Bakker, M., Torbeyns, J., Verschaffel, L., and DeSmedt, B. (2022). Longitudinal pathways of numerical abilities in preschool: Cognitive and environmental correlates and relation to primary school mathematics achievement. *Dev. Psychol.* 59, 442–459. doi: 10.1037/dev0001491

Bayeck, R. Y. (2020). Examining board gameplay and learning: A multidisciplinary review of recent research. *Simul. Gam.* 51, 411–431. doi: 10.1177/1046878119901286

Bochicchio, V., Keith, K., Montero, I., Scandurra, C., and Winsler, A. (2022). Digital media inhibit self-regulatory private speech use in preschool children: The "digital bubble effect". *Cogn. Dev.* 62:101180. doi: 10.1016/j.cogdev.2022.101180

Boone, W. J. (2016). Rasch analysis for instrument development: Why, when, and how? *CBE Life Sci. Educ.* 15:rm4. doi: 10.1187/cbe.16-04-0148

Boone, W. J., Staver, J. R., and Yale, M. S. (2014). *Rasch analysis in the human sciences*. Dordrecht: Springer Netherlands.

Cahoon, A., Gilmore, C., and Simms, V. (2021). Developmental pathways of early numerical skills during the preschool to school transition. *Learn. Instruct.* 75:101484. doi: 10.1016/j.learninstruc.2021.101484

Chu, F. W., vanMarle, K., Rouder, J. N., and Geary, D. C. (2018). Children's early understanding of number predicts their later problem-solving sophistication in addition. *J. Exp. Child Psychol.* 169, 73–92. doi: 10.1016/j.jecp.2017.12.010

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*, 2nd Edn. Hillsdale, NJ: Routledge.

Deutsche Gesellschaft für Kinder- und Jugendmedizin e.V (2022). *Leitlinie zur Prävention dysregulierten Bildschirmmediengebrauchs in der Kindheit und Jugend [SK2-guideline: Guideline for the Prevention of dysregulated screen media use in childhood and adolescence]*. Berlin: Deutsche Gesellschaft für Kinder- und Jugendmedizin e.V. Available online at: https://register.awmf.org/de/leitlinien/detail/027-075.

Deutsche Gesellschaft für Kinder- und Jugendpsychiatrie, Psychosomatik und Psychotherapie e.V (2023). *S3-Leitlinie: Diagnostik und behandlung der rechenstörung [S3-guideline: diagnosis and treatment of dyscalculia]*. München: Deutsche Gesellschaft für Kinder- und Jugendpsychiatrie, Psychosomatik und Psychotherapie. Available online at: https://register.awmf.org/de/leitlinien/detail/028-046

Dienes, Z. (2014). Using bayes to get the most out of non-significant results. *Front. Psychol.* 5:781. doi: 10.3389/fpsyg.2014.00781

Ehlert, A., and Fritz, A. 2016. "MARKO-T" - Ein mathematisches Förderprogramm evaluiert an Kindern mit dem Förderschwerpunkt Lernen. ["MARKO-T" - a mathematical support programme evaluated on children with learning difficulties]." In *Förderprogramme für Vor- und grundschule*, eds M. Hasselhorn and W. Schneider (Göttingen: Hogrefe), 29–48.

Elofsson, J., Gustafson, S., Samuelsson, J., and Träff, U. (2016). Playing number board games supports 5-year-old children's early mathematical development. *J. Math. Behav.* 43, 134–147. doi: 10.1016/j.jmathb.2016.07.003

Endlich, D., Berger, N., Küspert, P., Lenhard, W., Marx, P., Schneider, W., et al. (2016). *Würzburger Vorschultest: Erfassung schriftsprachlicher und mathematischer (Vorläufer-)Fertigkeiten und sprachlicher Kompetenzen im letzten Kindergartenjahr: [Würzburg preschool test: Assessment of written and mathematical (precursor) skills and language competencies in the last year of kindergarten]*. Göttingen: Hogrefe.

Faul, F., Erdfelder, E., Buchner, A., and Lang, A. (2009). Statistical power analyses using g*power 3.1: Tests for correlation and regression analyses. *Behav. Res. Methods* 41, 1149–1160. doi: 10.3758/BRM.41.4.1149

Faul, F., Erdfelder, E., Lang, A., and Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav. Res. Methods* 39, 175–191. doi: 10.3758/BF03193146

Fischbach, A., Schuchardt, K., Brandenburg, J., Klesczewski, J., Balke-Melcher, C., Schmidt, C., et al. (2013). "Prävalenz von Lernschwächen und Lernstörungen: Zur Bedeutung der Diagnosekriterien." [Prevalence of poor learners and children with learning disorders: Investigating the role of diagnostic criteria]. *Lern. Lernstörungen* 2, 65–76. doi: 10.1024/2235-0977/a000035

Fisher, K., Hirsh-Pasek, K., and Golinkoff, R. M. (2012). "Fostering mathematical thinking through playful learning," in *Contemporary debates in childhood education and development*, eds S. Suggate and E. Reese (London: Routledge), 95–106.

Fritz, A., and Ricken, G. (2008). *Rechenschwäche [Dyscalculia]*, 1st Edn. München: Ernst Reinhardt Verlag.

Fritz, A., Ehlert, A., and Leutner, D. (2018). Arithmetische Konzepte aus kognitiv-entwicklungspsychologischer Sicht [Arithmetic concepts from a cognitive-developmental psychological perspective]. *J. Math. Didaktik* 39, 7–41. doi: 10.1007/s13138-018-0131-6

Gasteiger, H., and Moeller, K. (2021). Fostering early numerical competencies by playing conventional board games. *J. Exp. Child Psychol.* 204:105060. doi: 10.1016/j.jecp.2020.105060

Gelman, R. (2006). Young natural-number arithmeticians. *Curr. Direct. Psychol. Sci.* 15, 193–197. doi: 10.1111/j.1467-8721.2006.00434.x

Gerlach, M., Fritz, A., and Leutner, D. (2013). *Mathematik- und Rechenkonzepte im Vor- und Grundschulalter-Training (MARKO-T) [Mathematics and numeracy concepts in pre-school and primary school age training (MARKO-T)]*, 1st Edn. Göttingen: Hogrefe.

Ginsburg, H. P., Pappas, S., and Seo, K. H. (2001). "Everyday mathematical knowledge: Asking young children what is developmentally appropriate," in *Psychological perspectives on early childhood education: Reframing dilemmas in research and practice. The Rutgers invitational symposium on education series*, ed. S. L. Golbeck (New Jersey, Lawrence Erlbaum Associates Inc), 181–219.

Gloor, N. (2023). Erfassung von numerischen Kompetenzen im Kindergarten. [Assessment of numerical skills in kindergarten]. *Lernen und Lernstörungen* 12, 175–187. doi: 10.1024/2235-0977/a000409

Harter, S., and Pike, R. (1984). The pictorial scale of perceived competence and social acceptance for young children. *Child Dev.* 55, 1969–1982. doi: 10.2307/1129772

Hauser, B., Vogt, F., Stebler, R., and Rechsteiner, K. (2014). "Förderung früher mathematischer Kompetenzen: Spielintegriert oder trainingsbasiert." [Promotion of early mathematical competences: Game-integrated or training-based]. *Frühe Bild.* 3, 139–145. doi: 10.1026/2191-9186/a000144

Hirsh-Pasek, K., Michnick, R., Berk, L. E., and Singer, D. G. (2009). *A mandate for playful learning in preschool: Presenting the evidence*. Oxford: Oxford University Press.

IBM Corp (2022). *IBM SPSS statistics*. New York, NY: IBM Corp.

Jansen, H., Mannhaupt, G., Marx, H., and Skowronek, H. (2002). *Bielefelder Screening zur Früherkennung von Lese-Rechtschreibschwierigkeiten (BISC) [Bielefeld screening for the early detection of reading and spelling difficulties (BISC)]*, 2nd Edn. Göttingen: Hogrefe.

JASP Team (2024). *JASP*. Amsterdam: J Team.

Jeffreys, H. (1998). "The theory of probability," in *Oxford classic texts in the physical sciences*, 3rd Edn. Oxford: Oxford University Press.

Jörns, C., Schuchardt, K., Mähler, C., and Grube, D. (2013). "Alltagsintegrierte Förderung numerischer Kompetenzen im Kindergarten." [Everyday integrated promotion of numerical competences in kindergarten]. *Frühe Bild.* 2, 84–91. doi: 10.1026/2191-9186/a000088

Krajewski, K. (2018a). *Vorhersage von Rechenschwächen in der Grundschule. Studien zur Kindheits- und Jugendforschung 29* [Predicting dyscalculia in primary school]. 2nd Edn. Hamburg: Dr. Kovac.

Krajewski, K. (2018b). *MBK-0: Test mathematischer Basiskompetenzen im Kindergartenalter [MBK-0: Test of basic mathematical competencies at kindergarten age]*, 1st Edn. Göttingen: Hogrefe.

Krajewski, K., and Schneider, W. (2006). "Mathematische Vorläuferfertigkeiten im Vorschulalter und ihre Vorhersagekraft für die Mathematikleistung bis zum Ende der Grundschulzeit." [Preschool mathematical precursor skills and their predictive power for mathematics achievement by the end of the primary school years.]. *Psychol. Erziehung Unterricht* 53, 246–262.

Krajewski, K., and Schneider, W. (2009). Early development of quantity to number-word linkage as a precursor of mathematical school achievement and mathematical difficulties: Findings from a four-year longitudinal study. *Learn. Instruct.* 19, 513–526. doi: 10.1016/j.learninstruc.2008.10.002

Lange, A. A., Brenneman, K., and Sareh, N. (2021). Using number games to support mathematical learning in preschool and home environments. *Early Educ. Dev.* 32, 459–479. doi: 10.1080/10409289.2020.1778386

Laski, E. V., and Siegler, R. S. (2014). Learning from number board games: You learn what you encode. *Dev. Psychol.* 50, 853–864. doi: 10.1037/a0034321

Linacre, J. M. (2024a). *Winsteps® Rasch measurement computer program*. Portland, OR: Winsteps.com. Available online at: https://www.winsteps.com/ (accessed on January 20, 2024).

Linacre, J. M. (2024b). Winsteps® *Rasch Measurement Computer Program User's Guide: Version 5.7.0*. Portland, OR: Winsteps.com.

Linebarger, D. L., Barr, R., Matthew, A., and Piotowski, J. (2014). Associations between parenting, media use, cumulative risk and children's executive functioning. *J. Dev. Behav. Pediatr.* 35, 367–377. doi: 10.1097/DBP.0000000000000069

Lu, Y., Li, Q., Patrick, H., and Mantzicopoulos, P. (2021). Math gives me a tummy ache!" Mathematics anxiety in kindergarten. *J. Exp. Educ.* 89, 362–378. doi: 10.1080/00220973.2019.1680518

Lumsden, J., Edwards, E. A., Lawrence, N. S., Coyle, D., and Munafò, M. R. (2016). Gamification of cognitive assessment and cognitive training: A systematic review of applications and efficacy. *JMIR Ser. Games* 4:e11. doi: 10.2196/games.5888

Mähler, C., Grube, D., and Schuchardt, K. (2017). Interindividuelle Unterschiede kognitiver Kompetenzen als Herausforderung für die frühkindliche Bildung. [Inter-individual differences in cognitive competences as a challenge for early childhood education]. *Pädagogische Rundschau* 71, 349–366. doi: 10.3726/PR2017-3/4_349

Marx, H. (1992). "Methodische und inhaltliche Argumente für und wider eine frühe Identifikation und Prädiktion von Lese-Rechtschreibschwierigkeiten." [Methodological and substantive arguments for and against early identification and prediction of reading and spelling difficulties]. *Diagnostica* 38, 249–268.

Marx, P., and Lenhard, W. (2011). "Diagnostische Merkmale von Screening-Verfahren zur Früherkennung möglicher Probleme beim Schriftspracherwerb [Diagnostic features of screening procedures for the early detection of possible problems in the acquisition of written language]," in *Frühprognose schulischer Kompetenzen*, eds H. Marcus and W. Schneider (Göttingen: Hogrefe Verlag), 68–84.

Mekler, E. D., Brühlmann, F., Opwis, K., and Tuch, A. N. (2013). "Do Points, levels and leaderboards harm intrinsic motivation?" in *Gamification 2013: Proceedings of the 1st international conference on gameful design, research, and applications: October 2-4, 2013, Stratford, Ontario, Canada*, eds H. Kevin, N. Randall, and L. E. Nacke (New York, NY: ACM), 66–73.

Milton Bradley Company (1978). *Chutes and ladders*. Pawtucket, RI: Hasbro.

Ministerium für Kultus, Jugend und Sport Baden-Württemberg (2022). *ALLES ZÄHLT! Mathe im Kiga-Alltag: Begleitheft. Wissenschaftlicher Hintergrund für Fachkräfte, Multiplikatorinnen und Multiplikatoren im Kindheitspädagogischen Arbeitsfeld. [EVERYTHING COUNTS! Maths in everyday life at the kindergarten]*. Stuttgart: Ministerium für Kultus, Jugend und Sport Baden-Württemberg.

Moosbrugger, H., and Kelava, A. (2020). *Testtheorie und Fragebogenkonstruktion. [Test theory and questionnaire construction]*. Heidelberg: Springer.

Moraske, S., Penrose, A., Wyschkon, A., Kohn, J., Rauscher, L., von Aster, M., et al. (2018). "Prävention von Rechenstörungen: Kurz- und mittelfristige Effekte einer Förderung der mathematischen Kompetenzen bei Risikokindern im Vorschulalter." [Prevention of dyscalculia: Short– and medium-term effects of promoting mathematical competencies in at-risk preschool children]. *Kindheit Entwicklung* 27, 31–42. doi: 10.1026/0942-5403/a000242

Morey, R. D., and Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychol. Methods* 16, 406–419. doi: 10.1037/a0024377

Morgan, P., Farkas, G., and Qiong, W. (2009). Five-year growth trajectories of kindergarten children with learning difficulties in mathematics. *J. Learn. Disabil.* 42, 306–321. doi: 10.1177/0022219408331037

Moser Opitz, E., and Ramseier, E. (2012). "Rechenschwach oder nicht Rechenschwach?" [Poor at maths or not poor at maths?]. *Lern. Lernstörungen* 1, 99–117. doi: 10.1024/2235-0977/a000013

Ninaus, M., Kiili, K., McMullen, J., and Moeller, K. (2017). Assessing fraction knowledge by a digital game. *Comput. Hum. Behav.* 70, 197–206. doi: 10.1016/j.chb.2017.01.004

Ninaus, M., Pereira, G., Stefitz, R., Prada, R., Paiva, A., Neuper, C., et al. (2015). Game elements improve performance in a working memory training task. *IJSG* 2:156. doi: 10.17083/ijsg.v2i1.60

Okanda, M., and Itakura, S. (2010). When do children exhibit a "Yes" bias? *Child Dev.* 81, 568–580. doi: 10.1111/j.1467-8624.2009.01416.x

Paulus, F. W., and Gerstner, I. (2023). Vorschulkinder und digitale Medien in der zweiten Moderne." [Preschool children and digital media in the second modern age]. *Pädiatr. Pädol.* 2, 99–102. doi: 10.1007/s00608-023-01111-2

Purpura, D. J., and Lonigan, C. J. (2013). Informal numeracy skills: The structure and relations among numbering, relations, and arithmetic operations in preschool. *Am. Educ. Res. J.* 50, 178–209. doi: 10.3102/0002831212465332

Purpura, D. J., Day, E., Napoli, A. R., and Hart, S. A. (2017). Identifying domain-general and domain-specific predictors of low mathematics performance: A classification and regression tree analysis. *J. Num. Cogn.* 3, 365–399. doi: 10.5964/jnc.v3i2.53

Radesky, J. S., Kaciroti, N., Weeks, H. M., Schaller, A., and Miller, A. L. (2023). Longitudinal associations between use of mobile devices for calming and emotional reactivity and executive functioning in children aged 3 to 5 years. *JAMA Pediatr.* 177, 62–70. doi: 10.1001/jamapediatrics.2022.4793

Ramani, G. B., Siegler, R. S., and Hitti, A. (2012). Taking it to the classroom: Number board games as a small group learning activity. *J. Educ. Psychol.* 104, 661–672. doi: 10.1037/a0028995

Ricken, G., Fritz, A., and Balzer, L. (2011). Mathematik und Rechnen - Test zur Erfassung von Konzepten im Vorschulalter. Ein Beispiel für einen niveauorientierten Ansatz. [Mathematics and arithmetic –test for the acquisition of concepts in preschool age (MARKO-D). An example of a level-oriented approach]. *Empirische Sonderpädagogik* 3, 256–271. doi: 10.25656/01:9327

Ricken, G., Fritz, A., and Balzer, L. (2013). *Mathematik- und Rechenkonzepte im Vorschulalter – Diagnose (MARKO-D) [Preschool mathematics and numeracy concepts - Diagnosis (MARKO-D)]*, 1st Edn. Göttingen: Hogrefe.

Roux, S., Kammermeyer, G., Donie, C., and Kinnunen, M. (2010). *Bildskala zur Erfassung von Lernfreude und Selbstkonzept in Mathematik und Schriftsprache bei Vorschulkindern (BSLS-MS) [Picture scale for the assessment of learning enjoyment and self-concept in preschool mathematics and language abilities]*. Koblenz: University of Koblenz-Landau.

Schuchardt, K., Fischbach, A., Balke-Melcher, C., and Mähler, C. (2015). "Die Komorbidität von Lernschwierigkeiten mit ADHS-Symptomen im Grundschulalter." [The comorbidity of learning difficulties and ADHD symptoms in primary-school-age children]. *Z. Kinder Jugendpsychiatr. Psychother.* 43, 185–193. doi: 10.1024/1422-4917/a000352

Siegler, R. S. (1995). How does change occur: A microgenetic study of number conservation. *Cogn. Psychol.* 28, 225–273. doi: 10.1006/cogp.1995.1006

Siegler, R. S., and Ramani, G. B. (2009). "Playing linear number board games–but not circular ones–improves low-income preschoolers' numerical understanding. *J. Educ. Psychol.* 101, 545–560. doi: 10.1037/a0014239

Skillen, J., Berner, V.-D., and Seitz-Stein, K. (2018b). The rule counts! Acquisition of mathematical competencies with a number board game. *J. Educ. Res.* 111, 554–563. doi: 10.1080/00220671.2017.1313187

Skillen, J., Berner, V.-D., Ricken, G., and Seitz-Stein, K. (2018a). Förderung basisnumerischer Kompetenzen mit einem regelbasierten Spiel im Einzel– und Gruppensetting." [Promoting basic numerical skills with a rule-based game in individual and group settings]. *Frühe Bild.* 7, 144–151. doi: 10.1026/2191-9186/a000381

Skillen, J., Ricken, G., and Seitz-Stein, K. (2023). Game-based assessment of early mathematical competencies –a pilot study. *Empirische Sonderpädagogik* 1, 38–60. doi: 10.25656/01:27183

Stern, E. (2009). "The development of mathematical competencies: Sources of individual differences and their developmental trajectories," in *Human development from early childhood to early adulthood*, eds S. Wolfgang and M. Bullock (Milton Park: Taylor and Francis), 221–237.

Textor, M. R. (2006). *Bildung im Kindergarten: Zur Förderung der kognitiven Entwicklung. [Education in the kindergarten: On promoting cognitive development]*. Münster: Monsenstein und Vannerdat.

Thorell, L. B., Burén, J., Wiman, J. S., Sandberg, D., and Nutley, S. B. (2022). Longitudinal associations between digital media use and ADHD symptoms in children and adolescents: A systematic literature review. *Eur. Child Adolesc. Psychiatry* doi: 10.1007/s00787-022-02130-3 [Epub ahead of print].

Tröster, H. (2009). *Früherkennung im Kindes- und Jugendalter: Strategien bei Entwicklungs-, Lern- und Verhaltensstörungen [Early diagnosis in childhood and adolescence: Strategies for developmental, learning and behavioural disorders]*, 1st Edn. Göttingen: Hogrefe.

van Doorn, J., van den Bergh, D., Böhm, U., Dablander, F., Derks, K., Draws, T., et al. (2021). The JASP guidelines for conducting and reporting a Bayesian analysis. *Psychon. Bull. Rev.* 28, 813–826. doi: 10.3758/s13423-020-01798-5

Vogt, F., Hauser, B., Stebler, R., Rechsteiner, K., and Urech, C. (2018). Learning through play – pedagogy and learning outcomes in early childhood mathematics. *Eur. Early Child. Educ. Res. J.* 26, 589–603. doi: 10.1080/1350293X.2018.1487160

Wittmann, E. C. (2001). "Ein alternativer ansatz zur förderung "rechenschwacher" Kinder, [An alternative approach to supporting "maths-deficient" children]" in *Beiträge zum mathematikunterricht*, ed. K. Gabriele (Berlin: Franzbecker), 660–663.

World Health Organization (2019). *Guidelines on physical activity, sedentary behaviour, and sleep for children under 5 years of age*. Geneva: World Health Organization.

World Health Organization (2019/2021). *International classification of diseases, eleventh revision (ICD-11)*, 11th Edn. Geneva: World Health Organization.

Wright, B. D., and Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Meas. Trans.* 8:370.

Wyschkon, A. (2015). "Erfassung und prognostische Bedeutung der nonverbalen Intelligenz im Vorschulalter: Analyse vorliegender Testverfahren und Kurzvorstellung des PIT-VA". [Assessment and prognostic significance of non-verbal intelligence at pre-school age: analysis of existing test procedures and brief presentation of the PIT-VA], in *Diagnostik im vorschulalter*, eds E. Günter, M. Hasselhorn, and W. Schneider (Göttingen: Hogrefe), 1–11.

Wyschkon, A., Kohn, J., Ballaschk, K., and Esser, G. (2009). Sind Rechenstörungen genau so häufig wie Lese-Rechtschreibstörungen?" [Is a specific disorder of arithmetic skills as common as reading/spelling disorder?]. *Z. Kinder Jugendpsychiatr. Psychother.* 37, 499–510. doi: 10.1024/1422-4917.37.6.499

frontiers | Frontiers in Education

TYPE Original Research
PUBLISHED 11 April 2024
DOI 10.3389/feduc.2024.1378313

OPEN ACCESS

EDITED BY
Martin Schöfl,
University of Education Upper Austria, Austria

REVIEWED BY
Timo Gnambs,
Leibniz Institute for Educational Trajectories (LG), Germany
Isaías Martín-Ruiz,
Universidad de Málaga, Spain
Barbara Maria Schmidt,
University of Cologne, Germany

*CORRESPONDENCE
Jan Luis Sigmund
Jan.Sigmund@med.uni-muenchen.de

RECEIVED 29 January 2024
ACCEPTED 01 April 2024
PUBLISHED 11 April 2024

CITATION
Sigmund JL, Mehlhase H, Schulte-Körne G and Moll K (2024) Early cognitive predictors of spelling and reading in German-speaking children. *Front. Educ.* 9:1378313.
doi: 10.3389/feduc.2024.1378313

Early cognitive predictors of spelling and reading in German-speaking children

Jan Luis Sigmund*, Heike Mehlhase, Gerd Schulte-Körne and Kristina Moll

Department of Child and Adolescent Psychiatry, Psychosomatics, and Psychotherapy, University Hospital, Ludwig-Maximilians-University Munich, Munich, Germany

Theoretical background: While reading and spelling skills often are interconnected in models of literacy development, recent research suggests that the two skills can dissociate and that reading and spelling are associated with at least partly different cognitive predictors. However, previous research on dissociations between reading and spelling skills focused on children who have already mastered the first phases of literacy development. These findings suggest that dissociations are due to distinct deficits in orthographic processing (i.e., unprecise orthographic representations vs. inefficient serial processing). It is therefore unclear whether dissociations already become apparent during the initial stages, or rather emerge later in development. This study aims to enhance the understanding of the predictors of early spelling and reading skills, investigating potential variations, by considering various cognitive factors beyond well-established ones.

Methods: Data were collected at two time points: cognitive predictors and early reading and spelling skills were assessed at the end of kindergarten (T1) before formal literacy instruction started, and reading and spelling skills were again assessed in Grade 1 (T2). The data analysis included 353 first-grade participants. Linear regression analyses assessed predictive patterns, while logistic regression analyses explained children's likelihood of belonging to different proficiency groups (at-risk or typical skills).

Results: Results revealed phonological processing, letter knowledge, and intelligence, as significant predictors for Spelling in grade 1 (T2), even after adding the autoregressor (Spelling in kindergarten at T1) and the respective other literacy skill (Reading T2). For Reading in grade 1 (T2), phonological processing, and rapid automatized naming (RAN) surfaced as significant predictors after adding the autoregressor (Reading T1). However, only RAN surfaced as a significant predictor for Reading T2 after adding the respective other literacy skill (Spelling T2). In line with these findings, logistic regression analyses revealed that phonological processing predicted group allocation for Spelling T2 and RAN predicted group allocation for Reading T2.

Conclusions: Overall, the study underscores the importance of phonological processing and letter knowledge as early predictors of spelling and reading skills in Grade 1. Moreover, intelligence is identified as a predictor for early spelling, while rapid automatized naming (RAN) emerges as a predictor for early reading.

KEYWORDS

spelling development, reading development, predictors, longitudinal, kindergarten, German orthography

1 Introduction

Mastering literacy skills is a crucial necessity for active participation in today's society. Children must accomplish two essential literacy skills in the initial years of literacy development: reading, which involves accurate and fluent word recognition, and spelling, which requires precisely transcribing words. Proficiency in these basic literacy skills forms the basis for more advanced literacy abilities, such as understanding and creating written content. Difficulties in acquiring spelling and reading skills are remarkable developmental hurdles, significantly impacting children's academic paths, career opportunities, and overall wellbeing (Carroll et al., 2005; Schulte-Körne, 2010; Ritchie et al., 2013). This is concerning, considering the diminished spelling proficiency observed in recent research conducted on German representative school samples. For instance, findings from the IGLU 2007 (Internationale Grundschul-Lese-Untersuchung [International reading assessment in primary schools]) indicate that only 80% of 4th graders were able to reliably apply spelling conventions expected to be mastered by the end of 2nd grade (Bos et al., 2007).

Numerous investigations have evidenced a strong correlation between reading and spelling skills (Ehri, 1986; Landerl and Wimmer, 2008; Caravolas et al., 2012). Spelling and reading in alphabetic orthographies rest upon understanding the alphabetic principle, mastering grapheme-phoneme correspondences for reading, along with phoneme-grapheme correspondences for spelling, and the ability to recognize, manipulate, and understand the individual phonemes (phoneme awareness). As expounded in influential developmental models by Ehri (1986) and Frith (1986), this signifies the alphabetic phase of literacy acquisition, enabling children to decode words for reading and break down spoken words into their constituent phonemes and pair them with the appropriate graphemes for spelling.

Nevertheless, spelling and reading skills do not always progress simultaneously, and deficits in spelling and reading can dissociate, resulting in poor spelling but adequate reading skills and vice versa (Frith, 1980; Wimmer and Mayringer, 2002; Moll and Landerl, 2009; Moll et al., 2020). Furthermore, research suggests that the cognitive predictors associated with spelling differ, to some extent, from those linked to reading (Wimmer et al., 2000; Cirino et al., 2005; Moll et al., 2014). This discrepancy is not only supported by empirical data but is also recognized within the classification systems of the ICD-11 (World Health Organization, 2022). Despite the distinct nature of these two literacy skills, previous research has predominantly focused on reading development and reading interventions, giving relatively limited attention to spelling (Treiman and Kessler, 2005). Furthermore, research directly comparing spelling and reading development in order to identify common as well as unique predictors is still rare. Previous research suggests that both spelling and reading proficiency advance in tandem during early literacy development and are heavily influenced by phonological processing skills and letter knowledge. As literacy development progresses, the acquisition of orthographic knowledge becomes pivotal for both spelling and reading. Based on findings in an orthographic learning experiment, Mehlhase et al. (2019) suggested that dissociations might occur as a consequence of distinct orthographic deficits. Isolated reading disorder is characterized by a deficiency in efficiently processing stored words despite a lack of deficit in storing orthographic representations. Conversely, isolated spelling disorder involves no deficit in efficient processing, but a deficit in storing orthographic representations, resulting in imprecise word representations and in turn many spelling errors. These imprecise representations are still sufficient for word recognition during reading, especially in relatively consistent orthographies like German which are characterized by high consistency between graphemes and phonemes. Compared to other languages like English, characterized by a vast number of syllables and high phonological complexity with simpler inflectional morphology, German exhibits medium phonological complexity, predominantly with closed syllables, and possesses rich derivation and inflectional morphology (Verhoeven and Perfetti, 2022). In contrast to reading, the relationship between phonemes and graphemes in the spelling direction is rather inconsistent in German, as in many other alphabetic orthographies. For instance, the long vowel/o:/in German can be represented in three different ways: "o" as in "Brot" (bread), "oo" as in "Boot" (boat), and "oh" as in "Sohn" (son). Correctly spelling such words demands children to progress beyond the alphabetic phase by forming word-specific representations in their long-term memory. Progression begins through the acquisition of metacognitive skills involving learning orthographic spelling rules and applying morphological principles (Steffler, 2001). As a result of the increasingly diverging mechanisms associated with spelling versus reading (problems), the hypothesis emerges, suggesting that the correlation between spelling and reading is more pronounced in early developmental phases but gradually diminishes as orthographic processing becomes more prominent in later stages of development. However, existing studies on dissociations related to spelling and reading have predominantly employed cross-sectional designs, with the majority conducted at a later stage of development, when children were in Grade 3 or beyond.

This study aims to fill this research gap by delving into the factors influencing early spelling and reading in Grade 1 within the specific context of German orthography. What sets this study apart is its distinctive approach, examining both spelling and reading at an early developmental stage, with a particular emphasis on spelling. This longitudinal study begins in kindergarten and extends through grade 1, allowing for a comprehensive assessment of early spelling and reading development. By assessing spelling and reading before formal literacy instruction starts and in Grade 1, predictive patterns for each skill in isolation can be examined (by controlling for the respective other literacy skill) and together, unraveling the impact of distinct predictors on spelling, reading, and both. Furthermore, the aim is to identify children at risk of spelling and/or reading problems in kindergarten, determining predictors that categorize them into either at-risk or typical proficiency groups in these skills. By integrating these approaches, the goal is to provide a holistic understanding of the intricate interplay between early spelling and reading.

In the following, the included predictors will be discussed in more detail. As mentioned before, the importance of phonological processing skills and grapheme-phoneme/phoneme-grapheme correspondence in spelling and reading proficiency is well-recognized. A third well-known predictor is rapid automatized naming (RAN), which turned out to be especially important for reading proficiency. However, there is a gap in studying additional cognitive factors that might be crucial to early literacy skills within a single study, especially in predicting early spelling proficiency. Thus, visual memory, and visual-spatial/graphomotor skills were included as they are hypothesized to contribute to different subskills of spelling proficiency such as the encoding of orthographic representations, precision and speed of letter formation, and understanding of intricate orthographic rules. Furthermore, intelligence was included as a control variable as it likely affects learning in general.

1.1 Phonological processing skills

The strongest and most consistently reported phonological predictor of literacy skills is phoneme awareness (i.e., the ability to identify and manipulate individual sounds) (phonemes) in spoken words. Numerous empirical studies underscore the robust relationship between phoneme awareness and literacy skills (Cataldo and Ellis, 1988; Landerl and Wimmer, 2008; Diamanti et al., 2017; Schmitterer and Schroeder, 2018), especially in the early phases of development (Babayigit and Stainthorp, 2010; Nielsen and Juul, 2016). The significance of phoneme awareness as a predictor of literacy skills transcends linguistic and writing system boundaries (Furnes and Samuelsson, 2009). Nevertheless, the direct contribution of phoneme awareness to reading may exhibit variability influenced by factors such as task characteristics, developmental stage, and orthographic complexity (Landerl et al., 2019). Therefore, the predictive power might diminish over time, especially in transparent orthographic systems (Landerl and Wimmer, 2000; Leppänen et al., 2006).

Another frequently assessed phonological skill is verbal/phonological memory, which involves the cognitive capacity to perceive and store verbal information, predominantly through the reception and encoding of speech stimuli, thus demonstrating a close connection to phoneme awareness. Indeed, many phoneme awareness tests also draw on verbal memory skills as they require to hold information in memory during manipulation of phonemes (e.g., phoneme deletion tests). Several studies have therefore investigated these two abilities jointly (Vellutino et al., 2004). Its role as an independent predictor appears to be somewhat limited. A large European study on developmental dyslexia suggested that verbal memory played a relatively minor role in predicting dyslexia status (Landerl et al., 2013). Similarly, a cross-linguistic analysis on cognitive mechanisms underlying reading and spelling development revealed verbal memory's relatively limited significance, however, it appears to carry more weight in spelling tasks compared to reading evaluations due to its critical role in storing word-specific knowledge (Moll et al., 2014). Compared to phoneme awareness, verbal memory generally plays a subordinate role (Ziegler et al., 2010; Caravolas et al., 2012).

Vowel length distinction has garnered less research attention, yet studies reveal its intricate relationship with literacy skills. A study by Post et al. (1999) explored the relationship between vowel perception and reading skill, finding a linear association between vowel identification errors and reading. However, in English orthography, the distinction among vowels encompasses not only variations in length but also involves differences in grapheme-phoneme-mapping (e.g., grapheme a: cat vs. car). In contrast, German vowels mainly vary in length, which can be marked in several ways (e.g., by doubling the vowel as in "Saal" hall). Still, there is an anticipation that vowel length distinction does influence reading, given that previous research has shown that German dyslexic children are impaired in phonological, temporal, and spectral vowel length distinction (Groth et al., 2011; Steinbrink et al., 2012, 2014). However, a more pronounced correlation between vowel length distinction and spelling is expected compared to reading, as careful attention to vowel length is essential for many spelling rules and in turn for orthographic spelling (e.g., if there is only one consonant sound after a short vowel the consonant is doubled. Examples are words like Hammer [hammer] and Schiff [ship]). A study involving German-speaking 10-year-old children revealed that the skill to differentiate vowel length was connected to accurately marking vowel length in spelling (Landerl, 2003). However, it still needs to be determined whether the association between vowel length distinction and spelling is already evident at the beginning of literacy instruction or rather emerges when orthographic processing becomes more prominent. Furthermore, studies assessing the role of vowel length distinction for both spelling and reading are still rare.

1.2 Letter knowledge

Extensive research conducted has spotlighted the importance of the alphabetic principle, which is firmly rooted in letter-sound knowledge (Caravolas et al., 2005; Otaiba et al., 2010; Byrne, 2014). These findings align with those of Caravolas et al. (2001) and Furnes and Samuelsson (2010), which highlight the role of letter knowledge as a predictor for both spelling and reading skills, elucidating its significance in explaining group differences between individuals with reading disorders and those with spelling disorders. Furthermore, several studies underscored the significance of letter knowledge as a prerequisite for early reading and/or spelling, emphasizing its role across various stages of literacy development (Ellis and Large, 1988; Lervåg and Hulme, 2010; Schmitterer and Schroeder, 2019). Additionally, studies showed that the influence of letter knowledge is evident in both consistent and inconsistent alphabetic orthographies (Ziegler et al., 2010; Moll et al., 2014).

1.3 Rapid automatized naming (RAN)

The role of rapid automatized naming (RAN) has garnered increasing attention, especially in the development of reading abilities. Numerous studies have demonstrated the predictive power of RAN on reading skills (Moll et al., 2009; Araújo et al., 2014; Papadopoulos et al., 2016; McWeeny et al., 2022). Especially in children at the kindergarten age, non-alphanumeric RAN is especially relevant, as not all children have acquired letter and number knowledge at this stage of development (Lervåg and Hulme, 2009). In both inconsistent and consistent orthographies, RAN has been linked to reading, and especially to reading fluency (Kirby et al., 2003). Mechanisms that are discussed to explain the RAN-reading relationship are verbal access and retrieval of stored information together with efficient serial processing (Protopapas et al., 2018). However, the role of RAN for spelling is less clear. While some studies have found that RAN predicts spelling skills, even after controlling for other variables like phonological awareness and letter-sound knowledge (Furnes and Samuelsson, 2010; Stainthorp et al., 2013), other studies have failed to establish a connection between RAN and early spelling development, especially in longitudinal studies (Landerl and Wimmer, 2008; Vaessen and Blomert, 2013). The inconsistent findings are likely to be due to the fact that the impact of RAN varies depending on factors such as orthographic consistency, the developmental stage, and linguistic characteristics.

In contrast to phonological processing skills, letter knowledge, and RAN, the impact of additional cognitive skills such as visual processing skills and intelligence on predicting literacy (i.e., reading) skills appear to be relatively restrained. While these factors might contribute to explaining variance beyond the primary predictors mentioned, their influence is generally modest, especially in the early stages of reading development. Nonetheless, these factors have the potential to impact spelling, by contributing to various subskills of spelling proficiency, including the encoding of orthographic representations, precision and speed in letter formation, understanding complex orthographic rules, hand-eye coordination, and spatial awareness. However, the extent to which these factors significantly affect spelling (and/or reading), and if so, whether noticeable effects already manifest early in literacy development, still needs to be determined.

1.4 Visual processing skills

Visual processing skills comprise a variety of constructs. With respect to literacy skills, the two constructs that have been predominantly mentioned to affect literacy skills are visual memory and visual-spatial skills.

While the role of *visual memory* for reading seems rather limited, a central question revolves around whether visual memory holds a more substantial role in spelling. This speculation arises from the notion that visual memory might be pivotal in encoding, storing, and retrieving orthographic representations, an important subskill of spelling. Limited research suggests that general working memory (Preßler et al., 2014), and visuo-spatial memory (Bourke et al., 2013) play a significant role in early spelling, over and above the influence of other predictors. Complementing this, Atkins and Tierney (2021) reveal that lower scores on auditory and visual sequential memory are linked to lower spelling and reading scores. However, Giles and Terrell (1997), find that visual memory does not differentiate between poor spellers and controls.

Existing research predominantly focused on visual attentional aspects rather than visual-spatial skills. These studies suggest that there is a predictive pattern between visual attentional skills and performance in spelling and/or reading, though the evidence is limited, and findings are mixed (Bosse and Valdois, 2009; Liu et al., 2016). In a study involving German children, Banfi et al. (2018) found no significant differences in visual attention among typically developing children, those with dyslexia, and those with isolated spelling deficits. Research of *visual-spatial skills* on the prediction of spelling and reading skills is comparatively small. This is unfortunate, given that visual-spatial skills may play a role at the beginning of literacy acquisition, when learners need to distinguish between visually similar yet distinct letters, such as/d/and/b/or/p/and/b/. These skills could impact the precision and speed of letter recognition influencing reading proficiency. With respect to spelling, visual-spatial skills might be especially relevant in combination with *graphomotor skills* when learning to write letters and words as they affect letter formation. Graphomotor skills encompass psychomotor abilities, coordinating cognitive processes with fine-motor skills, and are essential prerequisites for engaging in spelling activities. Therefore, good graphomotor skills release resources important for spelling performance. Investigations on graphomotor skills showed that these skills explain variance in spelling (Pontart et al., 2013; Mohamed and O'Brien, 2022).

1.5 Intelligence

In the current study, intelligence is included as a control variable because it affects learning more generally, rather than being specifically related to literacy acquisition. With respect to literacy, several studies have underscored the noteworthy contribution of intelligence to spelling performance at an early stage of literacy development (Ennemoser et al., 2012; Zarić et al., 2021). Furthermore, a longitudinal German study positions intelligence as a pivotal predictor for academic achievement in early and later stages of development, accentuating its enduring impact (Schneider and Niklas, 2017). On the other hand, the study of Ningrum and Wibowo (2017) illustrates a substantial contribution of intelligence to predicting more complex literacy skills later in development, such as reading comprehension and writing achievement. In addition, several studies introduce the notion that intelligence, lacks detailed correlations or predictive values especially for basic spelling and/or reading outcomes (Caravolas et al., 2001; Yeung et al., 2013). Therefore, the expectation is that the impact of intelligence on literacy skills is relatively small at the beginning of literacy development and increases during development. Furthermore, it is anticipated that intelligence has a more discernible impact on spelling, due to its inherent complexity, than on reading.

The objective of this study is to enhance our understanding of early spelling and reading skills by identifying both common and distinct predictors of reading vs. spelling skills. In this context, spelling's importance is highlighted, bringing attention to the limited literature and less-explored predictors in this domain, which are often overshadowed by research primarily concentrated on reading. This study, to the best of our knowledge, is one of few to investigate the relationship between spelling and reading, considering various cognitive factors beyond well-established ones, during the early stages of literacy development in German orthography. Therefore, the current study goes beyond previous investigations by integrating further potentially relevant cognitive predictors (here defined as secondary cognitive factors) in addition to predictors routinely reported in the literature (primary cognitive factors), and by investigating the differences in their influence on reading vs. spelling. By evaluating spelling and reading performance before formal literacy instruction and in Grade 1, the predictive patterns for each skill can be analyzed independently and collectively. This facilitates the deciphering of the influence of distinct predictors on spelling, reading, and both skills simultaneously, in order to establish if the patterns of predictions for the two literacy skills already dissociate at this early stage in literacy development. Moreover, the aim is to identify children who may face challenges in spelling and/or reading. This involves determining specific predictors that categorize them into either at-risk or typical proficiency groups in these skills.

Regression analyses are employed to identify common and distinct predictors of spelling and reading. Regarding early literacy development, previous findings indicate that phonological processing skills and letter knowledge emerge as the most pivotal predictors. Therefore, the expectation is that both spelling and reading skills will be predicted by these measures. In line with previous research, the anticipation is that the kindergarten measure of RAN will predict later reading skills, but not spelling skills. However, it is unclear, if the association between RAN and reading is already strong at the beginning of reading acquisition when reading is not fluent yet. Concerning secondary cognitive factors, predictive associations are anticipated between visual memory and spelling, as well as between visual-spatial/graphomotor skills and spelling. Furthermore, it is expected that intelligence, due to the inherent complexity of the task in German orthography, will emerge as a substantial predictor of spelling. Furthermore, correlation and logistic regression analyses were used to explain children's likelihood of belonging to different proficiency groups in spelling and reading (at-risk skills or typical skills). It is anticipated that measures of letter knowledge and phoneme awareness will strongly predict allocation to different proficiency groups in spelling and reading. Precise predictions about the role of RAN are avoided at this point, as the precise role in predicting group allocation remains uncertain at this early stage of literacy development. The same applies for the secondary factors.

2 Method

2.1 Participants and procedure

The present study presents findings from the first two time points of an ongoing longitudinal study aiming to assess the early cognitive predictors of spelling skills. At T1 (end of kindergarten) and T2 (mid of Grade 1), each child was tested individually on spelling and reading skills, and on cognitive predictors of literacy skills. A total of 372 students participated at the first two time points. The analyses are based on 353 students with complete data sets for the outcome measures. The final sample consisted of 47% boys; the mean age was 75.83 months (SD = 4.25) at T1 and 83,6 months (SD = 4.36) at T2. Out of the 353 children, 64.3% were brought up monolingually in German, while 35.7% were raised with multiple languages. For T1, all measures were administered individually by trained assistants in a quiet room at the respective kindergartens (n = 301) or at the University Hospital, Munich, Germany (n = 52). The testing period for T1 started at the end of April 2022 and lasted until the end of the summer holidays of 2022 before children were enrolled in schools. Thus, all children have been tested before formal spelling and reading instructions as well as teaching of letter sounds and names started, as it is not systematically taught to children in kindergarten in Germany. However, early reading and spelling skills were assessed not only at T2 but also in kindergarten to account for the fact that some children might already be able to read or spell simple syllables or words. For T2, all measures were administered individually by trained assistants in a quiet room at the University Hospital, Munich, Germany. The assessment period started at the beginning of January 2023 and covered 3 months to the beginning of April. The participants visit a total of 135 different schools and 233 different classes in the Munich district. Due to regional regulations all participants in this study are being taught using "Fibels," which are specifically designed to blend analytical and synthetic approaches. Inclusion criteria were an average scaled score of the two subtests Block Design and Vocabulary from the Wechsler Intelligence Test (WISC-V; Petermann, 2017) of at least 4 at T2 (IQ $\geq$ 70), German language skills for at least 3 years, normal or corrected-to-normal vision, and the absence of neurological deficits. The study was approved by the institutional ethics committee University Hospital, Munich, Germany and was performed in accordance with the latest version of the Declaration of Helsinki and in compliance with national legislation. Parents and children were informed in detail about the study and parents gave their written consent. Children received vouchers in return for their participation. The distinction between children with at-risk spelling and children with typical spelling skills was based on the performance in the standardized spelling test HSP 1 Plus (May et al., 2019) at T2. Children with spelling performance at or below the 25th percentile were assigned to the at-risk spelling group (n = 86, 44.2% girls), and children belonging to the typical spelling group had a spelling score above the 25th percentile (n=267, 55.4% girls). The distinction between children with at-risk reading and children with typical reading skills was based on their performance in the SLRT-III reading test, which is an updated version of the standardized reading test SLRT-II (Moll and Landerl, 2010) at T2. Children with reading performance at or below the 25th percentile were assigned to the at-risk reading group (n = 87, 52.9% girls), and children belonging to the typical reading group had a spelling score above the 25th percentile (n = 266, 52.6% girls). In our study, we opted to use the 25th percentile as the cutoff score for defining risk, considering the balance of sensitivity and specificity of risk identification. This decision was informed by relevant literature (Gersten et al., 2020; Hall et al., 2023), which indicated the applicability of this cutoff in similar contexts.

2.2 Measures

Table 1 provides a descriptive summary table with the characteristics of the measurements.

2.2.1 Phonological processing skills

2.2.1.1 Phoneme awareness

Phoneme awareness was assessed by an experimental test. In the first part of the test (phoneme identification) children had to identify the first or last phoneme of four words and

TABLE 1 Descriptive statistics for the measurements.

Variables	Min.	Max.	*M*	*SD*	Skewness	Kurtosis	*r*
Phoneme awareness[a]	0	100	37.04	29.84	0.42	−1,22	0.93
Verbal memory[b]	3	19	9.19	2.24	0.50	1.16	0.75
Vowel length distinction[b]	1	10	6.23	2.01	0.10	−0.70	0.61
Letter knowledge[a]	6	100	74.24	22.13	−0.75	−0.38	0.93
RAN[c]	0	1	0.82	0.18	0.40	0.21	
Visual-spatial and graphomotor skills[b]	0	30	9.20	6.11	0.62	0.09	0.74
Visual memory[b]	9	30	21.76	4.10	−0.57	0.04	0.83
Intelligence[d]	5	17	11.29	2.16	−0.48	0.29	0.84–0.86
Spelling T1[2]	0	22	9.27	7.84	0.28	−1.38	0.94
Spelling T2[2]	0	39	31.73	6.95	−2.48	6.87	0.94
Reading T1[2]	0	8	2.97	3.05	0.52	−1.31	0.93
Reading T2[2]	0	118	13.98	13.03	3.13	15.68	0.93–0.98

[a]percent correct.
[b]raw scores.
[c]items per second.
[d]scaled scores.

four pseudowords. The maximum score that can be achieved is 8. In the second part (phoneme deletion task) children had to pronounce four words and eight pseudoword without a specified phoneme. The maximum score that can be achieved is 12. The phoneme awareness test has a sample-based Cronbach's alpha of 0.93. Children had two practice trials each and the average accuracy in percent of the two parts was measured.

2.2.1.2 Verbal memory

Verbal memory was assessed using the standardized digit span subtest of the Kaufman-Assessment Battery for children–2nd Ed. (Melchers and Melchers, 2015) with maximum 21 items. After one practice trial, children had to repeat a series of digits of increasing length (from 2 to 9 digits) that were verbally presented by the experimenter. The task is terminated after three consecutive errors (discontinuation rule) so that the number of provided items differs according to the child's performance. The digit span test has a sample-based Cronbach's alpha of 0.75.

2.2.1.3 Vowel-length distinction

Vowel-length distinction was assessed using an adapted version of the vowel length distinction task from a screening tool (Witzel et al., 2023). The screening instrument has a sample-based split-half reliability of 0.61. Children had to decide whether the vowel in a heard word is long or short (e.g., "is the/i/in Schiff (ship) long or short?"). The maximum score that can be achieved is 10.

2.2.2 Letter-knowledge

Letter knowledge was assessed based on two subtests (grapheme-phoneme knowledge and phoneme-grapheme knowledge) from the standardized LRS-Screening (Endlich et al., 2019). the grapheme-phoneme subtest children were shown nine graphemes and were asked to name the corresponding phoneme. Additionally, a list of the 17 remaining graphemes of the German alphabet was presented and children were again asked whether they could name any of them, resulting in a maximum score of 26. The order of letter presentation was random to avoid systematic effects due to alphabetical order. In the phoneme-grapheme subtest children were shown nine separate lists with four graphemes on them each. The experimenter pronounced a phoneme and asked the child to point to the corresponding grapheme. The maximum score for the phoneme-grapheme task that can be achieved is 9. The letter knowledge test has a sample-based Cronbach's alpha of 0.93. The average accuracy in percent of the two letter-knowledge subtests was calculated.

2.2.3 Rapid automatized naming (RAN)

A standard RAN-objects paradigm (Denckla and Rudel, 1976) was presented. Children had to name a matrix of 30 visual stimuli (5 items: fish, clock, glass, leaf, dog) as quickly and accurately as possible. Due to the age of participants, we used a non-alphanumeric paradigm and accounted for accuracy, given the slightly higher error rates (mean = 0.4 errors) observed in this age group compared to older children or adults. The matrix was arranged in six rows with five items each. The item order was randomized. Each item was presented once in each line. The time needed to name the full item set was recorded and any occurring errors were marked. For the analysis, the number of correctly named items per second was computed.

2.2.4 Visual processing skills

2.2.4.1 Visual memory

Visual memory was assessed using the standardized Corsi Block-Tapping-Test from the Working memory test battery for children (Pickering and Gathercole, 2001). A board with nine blocks on it was presented. A series of blocks was tapped by the experimenter and after three practice trials the children had to reproduce the sequence by tapping the same blocks in the correct order. The span of the sequences that the children had to tap gradually increased from 1 to a maximum of 9 items, with 6 sequences each. The score was the number of correctly tapped sequences. The task is terminated after three errors within one sequence (discontinuation rule) so that the number of provided items differs according to the child's performance. The maximum score that can be theoretically achieved is 54. The test has a sample-based Cronbach's alpha of 0.83.

2.2.4.2 Visual-spatial and graphomotor skills

Visual-spatial/graphomotor skills were assessed using a subtest of the standardized German version of the Developmental Test of Visual Perception 2 (Büttner et al., 2008). Four different geometric shapes were presented with two of them having a specific mark. After two practice trials, children received a sheet of paper with the items on it. Within 1 min they had to draw the correct mark in the corresponding shape as quickly and accurately as possible. The maximum score that can be achieved is 63. The test has a retest reliability of $r_{tt} = 0.74$, according to manual.

2.2.5 Intelligence

To get an estimate of general cognitive abilities a non-verbal and verbal subtest from the German version of the Wechsler Intelligence Scale for Children (WISC-V; Petermann, 2017) were administered: Block Design and Vocabulary. During Block Design, the children viewed a model and/or picture and used two-colored blocks to re-create the design within a certain time limit. The subtest belongs to the visual spatial index and reflects the ability to understand visual-spatial relationships. The maximum score that can be achieved is 58. The subtest has a split-half reliability of $r_{tt} = 0.84$, according to manual. Vocabulary required children to name depicted objects and/or define words and is assigned to the verbal comprehension index. The subtest captures children's ability to verbalize meaningful concepts and to access word knowledge. The maximum score that can be achieved is 54. The subtest has a split-half reliability of $r_{tt} = 0.86$, according to manual. The two subtests were combined to have an estimate of intelligence including nonverbal and verbal skills. The average of the scaled scores of the two subtests are reported.

2.2.6 Spelling

2.2.6.1 Spelling T1

In kindergarten, early spelling skills were assessed by an experimental spelling test (Cronbach's alpha sample-based: 0.94). Children spelled a list of 5 two-letter syllables (e.g., la, me) and a list of 3 short high frequency words. The score was the number of correctly spelled graphemes. The maximum score that can be achieved is 22.

2.2.6.2 Spelling T2

Spelling at T2 was assessed using the standardized spelling test HSP 1 Plus (May et al., 2019). Children had to write 4 high frequency words and one sentence consisting of 6 words dictated by the experimenter. The score was the number of correctly spelled graphemes. The maximum score that can be achieved is 40. The test has an internal consistency of 0.94, according to manual.

2.2.7 Reading

2.2.7.1 Reading T1

In kindergarten, early reading skills were assessed by an experimental reading test (Cronbach's alpha sample-based: 0.93). Children read a list of 5 two-letter syllables (e.g., ma, le) and a list of 3 short high-frequency words. Number of items read correctly was scored. The maximum score that can be achieved is 8.

2.2.7.2 Reading T2

At T2, reading was assessed using the word reading subtest of the SLRT-III reading test, which is an updated version of the standardized reading test SLRT-II (Moll and Landerl, 2010). Children were given a list of words and were instructed to read aloud as many words as quickly and correctly as possible within one minute. The number of correctly read items within the time limit was scored. The maximum score that can be theoretically achieved is 156. The test has a parallel-test reliability of $r_{tt} = 0.93$–0.98, according to manual.

3 Results

3.1 Data pre-processing

To identify any differences in performance due to developmental and schooling effects in the three-month assessment period of T2, the performance of children tested at the beginning and at the end of the period for all variables were compared. Differences were found for both outcome variables Spelling T2 and Reading T2 due to the time point of assessment. Therefore, these values were z-standardized separately for each month of testing. Next, the distributional properties of the two spelling and reading measures were examined, in order to meet the assumptions of parametric tests. Due to skewness in both outcome variables, log transformation was performed to normalize their distribution [Spelling (Skewness: -1.05; Kurtosis: 1.45); Reading (Skewness: 1.12; Kurtosis: 2.29)] (Tabachnick et al., 2013). The subsequent analyses are reported based on transformed data but were also confirmed using raw data.

We initially explored reducing the number of predictors by combining theoretically related constructs, like phonological predictors and visual predictors, based on their correlations. An analysis showed a moderate correlation of $r = 0.40$ between phoneme awareness and verbal memory, that was higher than with vowel length discrimination and the two visual measures. Additionally, the phoneme deletion task in our study necessitates verbal short-term and working memory, so that the two constructs are interrelated. Therefore, the two tests (verbal memory and phoneme awareness) were z-standardized separately and merged into a composite variable of phonological processing. Despite the potential theoretical inclusion of vowel length in phonological processing, its correlation of $r = 0.18$ with verbal memory led us to maintain it separately. We further assessed whether visual memory and visual-spatial/graphomotor skills can be combined to

TABLE 2 Means (SD) at T1 and T2 and group comparisons (spelling).

Variables	Typical spelling (267)		At-risk spelling (86)		*t* (351)	*p*	Cohen's *d*	95% confidence interval	
	M	*SD*	*M*	*SD*				Lower	Upper
Intelligence[a]	11.58	2.04	10.39	2.26	4.56	<0.001	0.57	0.32	0.81
Letter knowledge[b]	79.57	19.59	57.68	21.48	8.80	<0.001	1.09	0.83	1.35
Vowel length distinction[c]	6.42	2.09	5.65	1.61	3.12	<0.001	0.39	0.14	0.63
Phonological processing[d]	0.18	0.82	−0.54	0.63	7.43	<0.001	0.92	0.67	1.17
RAN[e]	0.84	0.18	0.76	0.16	3.20	<0.001	0.40	0.15	0.64
Visual-spatial and graphomotor skills[c]	9.85	6.20	7.21	5.37	3.54	<0.001	0.44	0.19	0.68
Visual memory[c]	22.30	3.90	20.08	4.25	4.49	<0.001	0.56	0.31	0.80
Spelling T1[c]	11.06	7.81	3.69	4.68	8.28	<0.001	1.03	0.77	1.28
Spelling T2[d]	−0.23	0.09	−0.52	0.13	23.94	<0.001	2.97	2.64	3.29
Reading T1[c]	3.71	3.05	0.67	1.48	8.89	<0.001	1.10	0.85	1.36
Reading T2[d]	0.35	0.14	0.17	0.08	10.91	<0.001	1.35	1.09	1.62

[a]scaled scores.
[b]percent correct.
[c]raw scores.
[d]z-scores.
[e]items per second.

a single visual processing measure by calculating the correlations between the two measures. The correlation analysis for the two visual processing predictors revealed only a small correlation of $r = 0.23$. Therefore, the two visual predictors were kept separate.

3.2 Design and statistical analysis

Descriptive statistics for age, intelligence, literacy-specific predictors, and literacy outcomes are displayed separately for children with typical skills and those at-risk of spelling and reading disorders separately in Tables 2, 3.

Pearson correlation coefficients between predictor variables at T1 and the two literacy measures at the two time points were calculated and are presented in Table 4. Next, the predictive pattern for Spelling and Reading at T2 was examined separately in a series of stepwise regression analyses. In step 1, individual differences in intelligence were controlled for. The literacy-specific predictors were added simultaneously in step 2, and the unique variance for each predictor variable and the variance shared between predictors were calculated. In step 3 the respective autoregressor was added (Reading T1 for Reading T2; Spelling T1 for Spelling T2). To partial out the mechanisms that overlap between the two outcome measures Spelling and Reading at T2, the respective other literacy variable of the dependent variable (Reading T2 for Spelling T2; Spelling T2 for Reading T2) was included in step 4 of the respective regression analyses. This enabled an examination of the predictive patterns that are unique for spelling or reading, without the influence of the respective other literacy skill. The results are shown in Tables 5, 6. Finally, two logistic regression analyses were run in order to model children's likelihood of belonging to different proficiency groups, both for seplling and reading. The results are shown in Tables 7, 8.

3.3 Descriptive statistics

Tables 2, 3 shows that children with typical spelling and reading skills outperformed the children in the at-risk groups in all predictor and outcome measures. In the domains of both spelling and reading, children deemed at-risk consistently manifest notable deficiencies in primary cognitive risk factors, while differences in secondary risk factors were small to moderate. Biggest effect sizes were found for phonological processing (Cohen's $d_s = 0.92$ and 0.88) and letter knowledge (Cohen's $d_s = 1.09$ and 0.95). This underscores the importance of these two skills for literacy acquisition. Rapid automatized naming (RAN) demonstrates moderate effect sizes (Cohen's $d_s = 0.40$ and 0.61). In line with the important role of RAN for reading, effect sizes were higher in the reading domain (i.e., when comparing children at risk of reading problems with typical readers) than in the spelling domain (i.e., when comparing children at risk of spelling problems with typical spellers).

3.4 Correlations

Results in Table 4 showed that there was a significant, strong, positive correlation between almost all the variables ($p_s < 0.01$). Only the correlations between vowel length distinction and RAN and between vowel length distinction and visual memory were slightly lower ($p_s < 0.05$). Correlations for literacy measures at both time points were particularly high with letter knowledge and phonological processing. In line with our predictions, correlations between the spelling and reading measures were significantly higher for the first time point in kindergarten (0.88) than for the second time point in grade 1 (0.65) ($z = 8.84$; $p < 0.001$; Dunn and Clark, 1969).

TABLE 3 Means (SD) at T1 and T2 and group comparisons (reading).

Variables	Typical reading (266)		At-risk reading (87)		*t* (351)	*p*	Cohen's *d*	95% confidence interval	
	M	*SD*	*M*	*SD*				Lower	Upper
Intelligence[a]	11.57	2.09	10.42	2.14	4.44	<0.001	0.55	0.30	0.79
Letter knowledge[b]	79.02	20.15	59.61	21.61	7.66	<0.001	0.95	0.69	1.20
Vowel length distinction[c]	6.41	2.02	5.68	1.88	3.00	0.001	0.37	0.13	0.61
Phonological processing[d]	0.17	0.84	−0.52	0.59	7.11	<0.001	0.88	0.63	1.13
RAN[e]	0.84	0.18	0.74	0.16	4.92	<0.001	0.61	0.36	0.85
Visual-spatial and graphomotor skills[c]	9.91	6.30	7.06	4.92	3.85	<0.001	0.48	0.23	0.72
Visual memory[c]	22.35	3.89	19.95	4.20	4.90	<0.001	0.60	0.36	0.85
Spelling T1[c]	11.04	7.88	3.84	4.56	8.08	<0.001	1.00	0.74	1.25
Spelling T2[d]	−0.25	0.11	−0.47	0.18	13.88	<0.001	1.71	1.44	1.99
Reading T1[c]	3.65	3.10	0.90	1.60	7.92	<0.001	0.98	0.72	1.23
Reading T2[d]	0.36	0.13	0.14	0.05	15.11	<0.001	1.87	1.59	2.14

[a]scaled scores.
[b]percent correct.
[c]raw scores.
[d]z-scores.
[e]items per second.

TABLE 4 Correlations for study variables.

Variable	1	2	3	4	5	6	7	8	9	10	11
1. Intelligence	–										
2. Letter knowledge	0.26**	–									
3. Vowel length distinction	0.23**	0.30**	–								
4. Phonological processing	0.35**	0.54**	0.31**	–							
5. RAN	0.16**	0.23**	0.12*	0.23**	–						
6. Visual-spatial and graphomotor skills	0.19**	0.30**	0.17**	0.22**	0.15**	–					
7. Visual memory	0.22**	0.31**	0.12*	0.27**	0.21**	0.23**	–				
8. Spelling T1	0.24**	0.79**	0.32**	0.65**	0.26**	0.35**	0.25**	–			
9. Spelling T2	0.35**	0.51**	0.18**	0.51**	0.24**	0.24**	0.29**	0.53**	–		
10. Reading T1	0.26**	0.73**	0.31**	0.63**	0.28**	0.34**	0.24**	0.88**	0.50**	–	
11. Reading T2	0.23**	0.48**	0.20**	0.51**	0.37**	0.21**	0.26**	0.58**	0.65**	0.58**	–

**p < 0.01.
*p < 0.05.

3.5 Regression analyses

The regression results for Spelling T2 as dependent variable are shown in Table 5. Step 1 (intelligence) explained a significant amount (12%) of the variance in spelling ($p < 0.001$). The specific predictors in step 2 explained additional 25.7% of the variance in spelling ($p < 0.001$), with intelligence (2%), letter knowledge (5%), and phonological processing (5%) explaining significant unique variance. However, vowel length distinction, RAN, visual-spatial/graphomotor skills, and visual memory did not significantly predict spelling. The unique variance of the predictors was 13% and the shared variance between the predictors was 25%. In step 3 (autoregressor) the model explained additional 0.9% of the variance in Spelling ($p < 0.001$). Intelligence (2%), letter knowledge (1%), and phonological processing (2%) still explained significant unique variance in spelling, even after adding the autoregressor (Spelling T1) to the model. The unique variance of the predictors was 8% and the shared variance between the predictors was 31%, suggesting a large overlap between predictors at this age. In step 4 (Reading T2), the model explained additional 12.8% of the variance in spelling ($p < 0.001$). The cognitive predictors intelligence (2%), letter knowledge (1%), and phonological processing (1%), continued to explain a significant but small amount of unique variance in spelling after adding Reading T2 to the model. The unique variance of the predictors was 17% and the shared variance between the predictors was 34%.

TABLE 5 Four step regression analysis for spelling as dependent variable (z-standardized values).

	B	$SE\ B$	β	t	p	R	R^2 (%)	ΔR^2 (%)
Step 1 (intelligence)						0.35	12.10	12.10
Constant	−0.30	0.01		−38.35	<0.001			
Intelligence	0.05	0.01	0.35	6.95	<0.001		12.10	
Step 2 (specific predictors)						0.62	37.83	25.74
Constant	−0.30	0.01		−45.21	<0.001			
Intelligence	0.02	0.01	0.16	3.35	0.001		2.02	
Letter knowledge	0.04	0.01	0.28	5.34	<0.001		5.15	
Vowel length distinction	−0.01	0.01	−0.05	−1.20	0.232		0.26	
Phonological processing	0.05	0.01	0.27	5.15	<0.001		4.78	
RAN	0.01	0.01	0.07	1.62	0.106		0.47	
Visual-spatial and graphomotor skills	0.01	0.01	0.05	1.07	0.284		0.21	
Visual memory	0.01	0.01	0.07	1.58	0.115		0.45	
Unique variance step 2							13.33	
Shared variance step 2							24.50	
Step 3 (autoregressor)						0.62	38.73	0.90
Constant	−0.30	0.01		−45.48	<0.001			
Intelligence	0.03	0.01	0.16	3.55	<0.001		2.25	
Letter knowledge	0.03	0.01	0.18	2.59	0.010		1.19	
Vowel length distinction	−0.01	0.01	−0.06	−1.36	0.176		0.33	
Phonological processing	0.04	0.01	0.22	3.74	<0.001		2.49	
RAN	0.01	0.01	0.06	1.43	0.153		0.36	
Visual-spatial and graphomotor skills	0.00	0.01	0.03	0.67	0.500		0.08	
Visual memory	0.01	0.01	0.08	1.74	0.083		0.54	
Spelling T1	0.03	0.01	0.18	2.25	0.025		0.90	
Unique variance step 3							8.14	
Shared variance step 3							30.59	
Step 4 (reading T2)						0.72	51.57	12.83
Constant	−0.45	0.02		−27.15	<0.001			
Intelligence	0.02	0.01	0.15	3.59	<0.001		1.82	
Letter knowledge	0.03	0.01	0.18	2.82	0.005		1.13	
Vowel length distinction	−0.01	0.01	−0.05	−1.24	0.215		0.22	
Phonological processing	0.02	0.01	0.13	2.45	0.015		0.85	
RAN	−0.01	0.01	−0.04	−0.88	0.380		0.11	
Visual-spatial and graphomotor skills	0.01	0.01	0.04	0.95	0.344		0.13	
Visual memory	0.01	0.01	0.05	1.18	0.240		0.20	
Spelling T1	0.00	0.01	0.00	−0.06	0.949		0.00	
Reading T2	0.49	0.05	0.47	9.53	<0.001		12.83	
Unique variance step 4							17.28	
Shared variance step 4							34.29	

The regression results for Reading T2 as dependent variable are shown in Table 6. Step 1 (intelligence) explained a significant amount (5%) of the variance in reading ($p < 0.001$) before adding the specific predictors. In step 2 (specific predictors) the model explained additional 32.2% of the variance in reading ($p < 0.001$). Letter knowledge (4%), phonological processing (6%), and RAN

TABLE 6 Four step regression analysis for reading as dependent variable (z-standardized values).

	B	*SE B*	*β*	*t*	*p*	*R*	R^2 (%)	ΔR^2 (%)
Step 1 (intelligence)						0.23	5.45	5.45
Constant	0.30	0.01		38.36	<0.001			
Intelligence	0.04	0.01	0.23	4.50	<0.001		5.45	
Step 2 (specific predictors)						0.61	37.70	32.25
Constant	0.30	0.01		46.85	<0.001			
Intelligence	0.00	0.01	0.01	0.31	0.754		0.02	
Letter knowledge	0.04	0.01	0.24	4.43	<0.001		3.54	
Vowel length distinction	0.00	0.01	−0.01	−0.19	0.849		0.01	
Phonological processing	0.06	0.01	0.31	5.83	<0.001		6.14	
RAN	0.04	0.01	0.23	5.17	<0.001		4.83	
Visual-spatial and graphomotor skills	0.00	0.01	0.02	0.49	0.624		0.04	
Visual memory	0.01	0.01	0.05	1.12	0.262		0.23	
Unique variance step 2							14.81	
Shared variance step 2							22.89	
Step 3 (autoregressor)						0.65	42.24	4.54
Constant	0.30	0.01		48.59	<0.001			
Intelligence	0.00	0.01	0.02	0.44	0.661		0	
Letter knowledge	0.01	0.01	0.06	0.95	0.344		0	
Vowel length distinction	0.00	0.01	−0.02	−0.50	0.617		0	
Phonological processing	0.04	0.01	0.20	3.56	<0.001		2	
RAN	0.03	0.01	0.20	4.64	<0.001		4	
Visual-spatial and graphomotor skills	0.00	0.01	−0.01	−0.32	0.749		0	
Visual memory	0.01	0.01	0.07	1.47	0.142		0	
Reading T1	0.05	0.01	0.35	5.20	<0.001		5	
Unique variance step 3							10.88	
Shared variance step 3							31.36	
Step 4 (spelling T2)						0.74	54.81	12.57
Constant	0.44	0.01		29.82	<0.001			
Intelligence	−0.01	0.01	−0.05	−1.27	0.206		0	
Letter knowledge	−0.01	0.01	−0.05	−0.82	0.415		0	
Vowel length distinction	0.00	0.01	0.00	0.11	0.910		0	
Phonological processing	0.02	0.01	0.09	1.78	0.076		0	
RAN	0.03	0.01	0.17	4.48	<0.001		3	
Visual-spatial and graphomotor skills	0.00	0.01	−0.03	−0.79	0.429		0	
Visual memory	0.00	0.01	0.03	0.78	0.437		0	
Reading T1	0.05	0.01	0.30	5.04	<0.001		3	
Spelling T2	0.44	0.04	0.45	9.77	<0.001		13	
Unique variance step 4							19.45	
Shared variance step 4							35.37	

(5%) explained significant unique variance in reading. However, intelligence, vowel length distinction, visual-spatial/graphomotor skills, and visual memory did not significantly predict reading. The unique variance of the predictors was 15% and the shared variance between the predictors was 23%. In step 3 (autoregressor) the model explained additional 4.5% of the variance in reading

TABLE 7 Logistic regression results for spelling (z-standardized values).

	B	*SE B*	Wald	*p*	Exp (*B*)	95% C.I. for Exp (*B*)	
						Lower	Upper
Step 1							
Letter knowledge	0.36	0.21	3.05	0.081	1.44	0.96	2.17
Vowel length distinction	0.02	0.15	0.02	0.884	1.02	0.76	1.38
Phonological processing	0.60	0.25	6.01	0.014	1.83	1.13	2.96
RAN	0.16	0.16	1.05	0.306	1.17	0.86	1.59
Visual-spatial and graphomotor skills	0.08	0.17	0.21	0.649	1.08	0.78	1.49
Visual memory	0.21	0.15	1.90	0.168	1.23	0.92	1.66
Spelling T1	0.54	0.28	3.73	0.053	1.71	0.99	2.96
Constant	1.58	0.18	76.61	0.036	4.83		

TABLE 8 Logistic regression results for reading (z-standardized values).

	B	*SE B*	Wald	*p*	Exp (*B*)	95% C.I. for Exp (*B*)	
						Lower	Upper
Step 1							
Letter knowledge	0.16	0.19	0.76	0.385	1.18	0.82	1.69
Vowel length distinction	0.02	0.15	0.03	0.870	1.03	0.76	1.38
Phonological processing	0.52	0.24	4.60	0.032	1.68	1.05	2.69
RAN	0.44	0.16	7.35	0.007	1.56	1.13	2.15
Visual-spatial and graphomotor skills	0.11	0.17	0.42	0.519	1.12	0.80	1.55
Visual memory	0.27	0.15	3.19	0.074	1.31	0.97	1.75
Reading T1	0.76	0.28	7.38	0.007	2.14	1.24	3.69
Constant	1.64	0.19	72.65	0.003	5.17		

($p < 0.001$). Phonological processing (2%) and RAN (4%) still explained significant unique variance in reading, even after adding the autoregressor (Reading T1) to the model. The unique variance of the predictors was 11% and the shared variance between the predictors was 31%, showing again a large overlap between predictors. In step 4 (Spelling T2), the model explained additional 12.6% of the variance in reading ($p < 0.001$). Only the cognitive predictor RAN (3%) continued to explain significant unique variance in reading after adding Spelling T2 to the model. The unique variance of the predictors was 19% and the shared variance between the predictors was 35%.

The results of the logistic regression analysis for spelling are shown in Table 7. The analysis identified one significant predictor for the group allocation of spelling proficiency: phonological processing ($B = 0.60$, $SE\ B = 0.25$, $p = 0.014$), indicating that weaker phonological processing skills are associated with an increased likelihood of being at-risk of spelling problems at T2. All other predictor variables, including letter knowledge, vowel length distinction, RAN, visual-spatial/graphomotor skills, and visual memory did not significantly predict group allocation of spelling proficiency at T2.

The results of the logistic regression analysis, employed to model the likelihood of participants falling into different reading proficiency groups at T2 are shown in Table 8. The results revealed three factors that emerged as significant predictors for the classification of reading groups at T2: phonological processing ($B = 0.52$, $SE\ B = 0.24$, $p = 0.032$), the corresponding literacy skill (here T1 reading; $B = 0.76$, $SE\ B = 0.28$, $p = 0.007$), and RAN ($B = 0.44$, $SE\ B = 0.16$, $p = 0.007$). The other factors, including letter knowledge, vowel length distinction, visual-spatial/graphomotor skills, and visual memory did not predict group allocation of reading proficiency at T2.

4 Discussion

This study aimed to deepen our understanding of early spelling and reading skill, by identifying both common and distinct predictors of spelling and reading skills and deficits in these domains within a large sample of 353 children from kindergarten to second half of Grade 1. Predictive measures (phonological processing, letter knowledge, RAN, visual-spatial/graphomotor skills, and visual memory) were assessed in kindergarten, before formal literacy instruction, the outcome variables (spelling and reading) were measured in kindergarten and in first grade, and the control variable intelligence in Grade 1. Children with typically developing spelling and reading skills in Grade 1 outperformed their counterparts in at-risk groups across all predictor and literacy measures assessed at T1, with largest effect sizes found for literacy skills at T1, letter knowledge, and phonological processing. The regression analysis for spelling abilities at T2 identified phonological processing, letter knowledge, and intelligence, as significant predictors, even after including step 3 (autoregressor Spelling T1) and step 4 (Reading T2) in the model. For reading abilities, phonological processing and RAN surfaced as significant predictors after step 3 (autoregressor Reading T1). However, only RAN (and the autoregressor) turned out to be a significant predictor for reading after step 4 (Spelling T2) was included in the model.

This research, in line with previous studies (Ehri, 1986; Caravolas et al., 2001; Landerl and Wimmer, 2008), underscores at least partly simultaneous enhancement of spelling and reading proficiency in the initial stages of literacy acquisition. However, the study also shows that the correlation between spelling and reading already decreases early in development, with stronger correlations in kindergarten ($r = 0.88$) compared to Grade 1 ($r = 0.65$). It was hypothesized that the association between reading and spelling skills is gradually diminishing during literacy acquisition as orthographic processing becomes more central, and distinct deficits in orthographic processing emerge. This suggests a shift in the dynamics of the relationship between spelling and reading skills over time. Therefore, it is expected, that the correlation will decrease further in grades 2 and 3 as orthographic processing takes on greater significance, and subprocesses of orthographic processing such as the storage of orthographic representations (associated with spelling) and effective processing of orthographic representations (associated with reading fluency) potentially diverge. This expectation aligns with findings from previous studies based on large representative samples of German speaking children in grades 3 and 4, who showed correlations between spelling and reading abilities of r_s between 0.48 and 0.57 (Moll and Landerl, 2009; Landerl and Moll, 2010; Görgen et al., 2021).

To deepen the understanding of distinct cognitive factors associated with early reading vs. spelling, a diverse set of cognitive factors was explored, extending beyond commonly acknowledged predictors of literacy skills. The results confirm the crucial contribution of well-established cognitive factors, such as phonological processing and letter-knowledge, to proficiency in literacy skills. Furthermore, it highlights the pivotal role of RAN in reading (but not spelling). However, the impact of other cognitive factors appears to be limited, with the notable exception of intelligence, which was associated with spelling proficiency.

In line with previous research (Cataldo and Ellis, 1988; Landerl and Wimmer, 2008; Diamanti et al., 2017; Schmitterer and Schroeder, 2018), the results revealed that kindergarten proficiency in phonological processing skills significantly predicted performance in spelling and reading during Grade 1, even after adding the respective autoregressor. A novel finding of the current study was that when identifying predictors that are specifically related to either reading or spelling by controlling for the respective other literacy skill (step 4), phonological processing turned out to be specifically related to spelling and to the variance shared between reading and spelling, but not to the variance related to reading only. The association between phonological processing and the shared variance between reading and spelling suggests that phonological processing taps into mechanisms underlying reading as well as spelling acquisition, most probably the understanding of the sound structure of words. The specific association between phonological processing and spelling might reflect the role of phonological processing in building-up precise word specific representations, which are crucial for spelling words correctly. Alternatively, the strong association of phonological processing with spelling might be the consequence of the higher demands on phonological processing associated with spelling compared to reading. The strong association between phonological processing and spelling is not restricted to the German orthography, but is also reported in other alphabetic orthographies, such as Norwegian (Lervåg and Hulme, 2010), Danish (Nielsen and Juul, 2016), and English (Caravolas et al., 2001; Moll and Landerl, 2009; Lervåg and Hulme, 2010). With respect to reading, the study identified no relationship between phonological processing and early reading skills, after controlling for Spelling T2. This is in alignment with Landerl et al. (2019), who found no predictive link between phonological awareness and reading in Grade 1 German children. The important role of phonological processing for early spelling was also evident when determining group allocation (at-risk or typical). Phonological processing was the only significant kindergarten measure for predicting spelling group allocation in Grade 1. Phonological processing also predicted reading group allocation, but played a less important role compared to reading at T1 and RAN, who turned out as the most important predictors for reading group allocation.

In addition to phonological processing, vowel length distinction was included as an independent factor. Despite its intricate relationship with literacy skills in previous studies (Groth et al., 2011; Steinbrink et al., 2012, 2014), the study found no significant predictive link between vowel length distinction assessed in kindergarten and Grade 1 spelling or reading abilities. It also did not emerge as a significant predictor for spelling and reading group allocation. These findings do not seem to be the result of a large overlap between vowel length distinction and phonological processing given that the correlation between both measures was relatively low ($r = 0.31$). There are at least two potential explanations for this finding. First, the association between vowel length distinction and literacy skills might depend on the stage of literacy development of the participants. Vowel length distinction is crucial for following spelling rules and, by extension, achieving precision in orthographic spelling (e.g., doubling of a single consonant following a short vowel in the German orthography). This interpretation aligns with the observation that prior research on this subject predominantly focused on participants in more advanced stages of literacy development. For example, the studies by Groth et al. (2011) and Steinbrink et al. (2012) centered on young adults or 3rd and 4th-grade children (Steinbrink et al., 2014). Secondly, vowel length distinction might predominantly develop alongside literacy acquisition and might be fostered by explicit teaching. Thus, vowel distinction might be rather difficult for children in kindergarten, which is supported by our data. Even though the mean score in our sample for the vowel length distinction task was above chance rate ($M = 6.42$), some children did not score above the rate level of 5, reducing sensitivity regarding predictive patterns. Taken together, vowel length distinction assessed in kindergarten does not seem to predict literacy skills in the first year of literacy acquisition but might play a role in later developmental stages when orthographic processing and the acquisition of spelling rules becomes more relevant.

The significance of letter knowledge skills has been firmly established in the context of early literacy development (Ellis and Large, 1988; Lervåg and Hulme, 2010; Schmitterer and Schroeder, 2019). The current findings align, to some extent, with previous studies, as letter knowledge was found to significantly predict spelling but not reading in all four steps of the regression analyses. The predictive power for spelling persists even after introducing the autoregressor (Spelling T1) in step 3 and Reading T2 in step 4, emphasizing the substantial contribution of letter knowledge for spelling abilities. In the prediction of Grade 1 reading abilities, letter knowledge (4%) played a substantial role in explaining significant unique variance in reading abilities, before the introduction of the autoregressor (Reading T1). However, once the autoregressor (Reading T1) was included in the model in step 3, letter knowledge did not continue to explain significant unique variance in reading. A potential explanation for this finding is that the assessment of letter knowledge in kindergarten incorporates a higher number of items aligned with the reading direction (grapheme-phoneme correspondence) in comparison to the spelling direction (phoneme-grapheme correspondence). Thus, by controlling for the autoregressor (Reading T1), the variance assessed by the grapheme-phoneme correspondence measure seems to be accounted for by the initial reading skills in kindergarten. Moreover, letter knowledge in kindergarten did not emerge as a significant predictor in determining group allocation for spelling and reading abilities in Grade 1 in the logistic regression analyses. The findings that letter-knowledge did not predict group allocation is likely to be due to the overlap between letter-knowledge and T1 spelling and reading skills, as evident from the high correlations of 0.73 and 0.79. Still letter-knowledge turned out to be an important predictor, especially in children who cannot read and spell yet. In the current sample, 34.6% of the kindergarten children were unable to decipher a single syllable or word from a possible maximum score of 8, and 21% of the children struggled to spell a single grapheme from a potential maximum score of 22. Among children unable to read or spell, the mean percentage score for letter knowledge was 53.4% and 49.6%, respectively, suggesting they knew about half of the letters. Consequently, it is essential to consider letter knowledge as well as early reading and spelling skills in the early stages of literacy development.

Several studies have emphasized the robust predictive capacity of RAN on reading skills (Moll et al., 2009; Papadopoulos et al., 2016; McWeeny et al., 2022). While the mechanisms connecting RAN and reading involve verbal access, retrieval, and efficient serial processing, its impact on spelling remains inconclusive. Some studies assert a predictive role of RAN in spelling (Furnes and Samuelsson, 2010; Stainthorp et al., 2013), while others do not establish a clear connection (Landerl and Wimmer, 2008; Vaessen and Blomert, 2013). Consistent with previous research, the current study found a significant predictive relationship between RAN and reading performance, with RAN also significantly predicting reading group allocation in Grade 1. This extends existing research by illustrating that RAN predicts reading abilities already in early development, but not spelling skills. The finding that RAN is associated with reading in very early phases when reading still strongly relies on decoding rather than on fluent recognition of whole-words or larger units (e.g., syllables or morphemes) supports the idea that RAN reflects efficient serial processing, encompassing whole words as well as single letters, or smaller units during decoding, rather than focusing on orthographic processing (Moll et al., 2009; Protopapas et al., 2013).

This research not only considered established cognitive predictors of early literacy but also explored secondary factors influencing spelling subskills. Visual memory, potentially crucial for encoding orthographic representations, was examined for its role in spelling proficiency. Despite limited research indicating the significance of visuo-spatial working memory in early spelling (Bourke et al., 2013), this study, aligning with Giles and Terrell (1997), found that visual memory did not significantly predict later spelling or reading skills. It also did not emerge as a significant predictor for spelling and reading group allocation in Grade 1. Thus, the predictive value of visual memory for early spelling and reading skills appears constrained. Furthermore, a task measuring visual-spatial and graphomotor skills was included. Several studies established the notion, that visual-spatial skills (Bosse and Valdois, 2009; Liu et al., 2016) and graphomotor skills (Pontart et al., 2013; Suggate et al., 2018; Mohamed and O'Brien, 2022) explain variance in spelling and reading. However, our results are not in line with these findings. Visual-spatial and graphomotor skills did not significantly predict later spelling or reading skills and did not emerge as a significant predictor of spelling and reading group allocation in Grade 1. A possible reason for the differences in results might lie in the divergence of the specific constructs measured. The studies by Bosse and Valdois (2009) and Liu et al. (2016) primarily concentrated on visual-spatial attention, while the works of Pontart et al. (2013), Suggate et al. (2018), and Mohamed and O'Brien (2022) focused specifically on graphomotor skills. In contrast, the measurement used in the current study assesses the visual and motor speed at which a child can draw predetermined symbols into specific shapes, concurrently involving visual discrimination, fine motor skills, and execution speed. Thus, future studies will have to examine the associations between different visual-spatial and graphomotor constructs and their relation to literacy skills and whether this changes during development.

Moreover, intelligence was integrated as a general cognitive factor. Previous research reported mixed findings, with some studies providing evidence for the pivotal role of intelligence in spelling skills (Ennemoser et al., 2012; Zarić et al., 2021), while others showed limited correlations or predictive values (Caravolas et al., 2001; Yeung et al., 2013). Anticipating a relatively modest impact on early reading skills but a more pronounced effect on spelling due to its complexity in the German orthography, the current results indeed revealed that intelligence significantly predicts Grade 1 spelling skills but not reading. This phenomenon can be attributed to the strong asymmetry of the German orthography (and the majority of other alphabetic orthographies), where spelling is considerably more inconsistent and complex than reading. The findings are in line with previous research based on

German (Ennemoser et al., 2012; Zarić et al., 2021), and English samples (Caravolas et al., 2001), and extend previous research by showing that intelligence predicts spelling abilities already in early development alongside established predictors. Future research is needed to identify the exact mechanisms underlying the relationship between intelligence and spelling in early and later spelling development.

Finally, it is necessary to discuss some limitations and, by doing so, provide suggestions for future research. One limitation is the relatively short period of 1 year. Therefore, future studies should aim to investigate the evolving predictive patterns of spelling and reading skills over an extended period in one and the same sample. Furthermore, in the early phases of development, where formal literacy instruction is yet to be introduced, a notable range of diversity in literacy and precursor skills is due to environmental factors, such as the home learning environment, the extent of parental support, and the training of precursor skills in kindergarten. Therefore, future research should examine the interplay between cognitive and environmental factors in order to enhance our understanding of literacy development in the initial stages. Finally, as a result of examining preliterate children in kindergarten, the use of a standardized spelling or reading test was not feasible at T1. Consequently, it was not possible to utilize the same tests for spelling and reading in both kindergarten and Grade 1, limiting the comparability of spelling and reading performances between the two assessment periods. However, the data demonstrated that the reading and spelling tasks in kindergarten differentiated between at-risk children and successfully predicted performance in Grade 1. Finally, it should be mentioned that any cutoff to define specific risk groups is somehow arbitrary and needs to balance sensitivity and specificity of risk identification. Identifying too many children as at-risk can be economically burdensome and may lead to the stigmatization of these children. Conversely, identifying too few can result in overlooking children who are on the borderline, thus missing the opportunity for early intervention, which is crucial particularly in the early stages.

The outcomes of the present study carry significant practical implications. In line with previous research, phonological processing and letter-knowledge were the most important predictors for reading and spelling when early reading and spelling skills cannot yet be assessed. Once children can read and spell first syllables, these early reading and spelling skills are good predictors of later literacy skills and should therefore be included in kindergarten screening tools. Importantly, the current study also highlighted predictors that were specific to either reading or spelling, already at this early stage in development. The findings particularly underscore the pivotal role of phonological processing in shaping early spelling abilities, suggesting the necessity of targeted interventions for those who struggle with phonological processing and especially with phoneme awareness. In terms of reading abilities, RAN emerged as a critical predictor over and above the autoregressor (Reading T1) and Spelling T2, and differentiated between children with and without risk for reading problems. Thus, RAN plays an important role in the diagnostic process and in the early identification of children at risk of reading problems and should therefore be part of screening and diagnostic reading tools (see for example LRS-Screening; Endlich et al., 2019). The findings support the idea that RAN signifies efficient serial processing, encompassing decoding of whole words, as well as single letters, or smaller units. Enhancing efficient serial processing by training to read frequent syllables, morphemes, and letter clusters has been shown to be effective (Thaler et al., 2004; Hintikka et al., 2008). Whether RAN itself can be effectively trained and whether training effects transfer to reading, is still a matter of debate. The few existing studies on RAN training show very inconsistent findings (Fugate, 1997; de Jong and Vrielink, 2004; Vander Stappen and Reybroeck, 2018).

The distinct predictive pattern of reading and spelling was also reflected in the predictors of Grade 1 group allocation. Phonological processing in kindergarten emerged as the most important predictor for spelling, while RAN in kindergarten was the key predictor for reading. Further investigation is needed to understand the developmental trajectories of various precursor abilities and to identify effective methods for integrating them into early education programs.

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: https://osf.io/pvakt/?view_only=5629362e40344025828fc58a28f860a7.

Ethics statement

The studies involving humans were approved by Ethics Committee of the Medical Faculty of the University of Munich. The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation in this study was provided by the participants' legal guardians/next of kin.

Author contributions

JS: Writing—original draft, Writing—review & editing. HM: Writing—original draft, Writing—review & editing. GS-K: Writing—original draft, Writing—review & editing. KM: Writing—original draft, Writing—review & editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 456267474.

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

Araújo, S., Reis, A., Petersson, K. M., and Faísca, L. (2014). Rapid automatized naming and reading performance: a meta-analysis. *J. Educ. Psychol.* 107. doi: 10.1037/edu0000006

Atkins, M., and Tierney, E. (2021). Memory skills and specific learning difficulties. *REACH: J. Inclus. Educ. Ireland* 17, 81–92.

Babayigit, S., and Stainthorp, R. (2010). Component processes of early reading, spelling, and narrative writing skills in Turkish: a longitudinal study. *Read. Writ.* 23, 539–568. doi: 10.1007/s11145-009-9173-y

Banfi, C., Kemény, F., Gangl, M., Schulte-Körne, G., Moll, K., and Landerl, K. (2018). Visual attention span performance in German-speaking children with differential reading and spelling profiles: No evidence of group differences. *PLoS ONE* 13, e0198903. doi: 10.1371/journal.pone.0198903

Bos, W., Lankes, E.-M., Prenzel, M., Schwippert, K., Valtin, R., and Walther, G. (2007). "Erste Ergebnisse aus IGLU: Schülerleistungen am Ende der vierten Jahrgangsstufe im internationalen Vergleich," in *Sozialwissenschaftlicher Fachinformationsdienst soFid* (Berlin: Bildungsforschung 2007/1), 9–46.

Bosse, M.-L., and Valdois, S. (2009). Influence of the visual attention span on child reading performance: a cross-sectional study. *J. Res. Read.* 32, 230–253. doi: 10.1111/j.1467-9817.2008.01387.x

Bourke, L., Davies, S., Sumner, E., and Green, C. (2013). Individual differences in the development of early writing skills: Testing the unique contribution of visuospatial working memory. *Read. Writ.* 27, 3. doi: 10.1007/s11145-013-9446-3

Büttner, G., Dacheneder, W., Schneider, W., and Weyer, K. (2008). *FEW-2: Frostigs Entwicklungstest der visuellen Wahrnehmung - 2: Deutsche Fassung des Developmental Test of Visual Perception, Second Edition (DTVP-2) von,* D. D. Hammill, N. A. Pearson and J. K. Voress (Newburyport: Hogrefe).

Byrne, B. (2014). *The Foundation of Literacy: The Child's Acquisition of the Alphabetic Principle*. London: Psychology Press.

Caravolas, M., Hulme, C., and Snowling, M. J. (2001). The foundations of spelling ability: Evidence from a 3-year longitudinal study. *J. Mem. Lang.* 45, 751–774. doi: 10.1006/jmla.2000.2785

Caravolas, M., Lervåg, A., Mousikou, P., Efrim, C., Litavsk,ý, M., Onochie-Quintanilla, E., et al. (2012). Common patterns of prediction of literacy development in different alphabetic orthographies. *Psychol. Sci.* 23, 678–686. doi: 10.1177/0956797611434536

Caravolas, M., Volín, J., and Hulme, C. (2005). Phoneme awareness is a key component of alphabetic literacy skills in consistent and inconsistent orthographies: Evidence from Czech and English children. *J. Exp. Child Psychol.* 92, 107–139. doi: 10.1016/j.jecp.2005.04.003

Carroll, J. M., Maughan, B., Goodman, R., and Meltzer, H. (2005). Literacy difficulties and psychiatric disorders: evidence for comorbidity. *J. Child Psychol. Psychiat.* 46, 524–532. doi: 10.1111/j.1469-7610.2004.00366.x

Cataldo, S., and Ellis, N. (1988). Interactions in the development of spelling, reading and phonological skills. *J. Res. Read.* 11, 86–109. doi: 10.1111/j.1467-9817.1988.tb00153.x

Cirino, P. T., Israelian, M. K., Morris, M. K., and Morris, R. D. (2005). Evaluation of the double-deficit hypothesis in college students referred for learning difficulties. *J. Learn. Disabil.* 38, 29–43. doi: 10.1177/00222194050380010301

de Jong, P. F., and Vrielink, L. O. (2004). Rapid automatic naming: Easy to measure, hard to improve (quickly). *Ann. Dyslexia* 54, 65–88. doi: 10.1007/s11881-004-0004-1

Denckla, M. B., and Rudel, R. G. (1976). Rapid "automatized" naming (R.A.N.): dyslexia differentiated from other learning disabilities. *Neuropsychologia* 14, 471–479. doi: 10.1016/0028-3932(76)90075-0

Diamanti, V., Mouzaki, A., Ralli, A., Antoniou, F., Papaioannou, S., and Protopapas, A. (2017). Preschool phonological and morphological awareness as longitudinal predictors of early reading and spelling development in Greek. *Front. Psychol.* 8, 2039. doi: 10.3389/fpsyg.2017.02039

Dunn, O. J., and Clark, V. (1969). Correlation coefficients measured on the same individuals. *J. Am. Stat. Assoc.* 64, 366–377. doi: 10.1080/01621459.1969.10500981

Ehri, L. C. (1986). Sources of difficulty in learning to spell and read. *Adv. Dev. Behav. Pediatr.* 7, 121–195.

Ellis, N., and Large, B. (1988). The early stages of reading: a longitudinal study. *Appl. Cogn. Psychol.* 2, 47–76. doi: 10.1002/acp.2350020106

Endlich, D., Küspert, P., Lenhard, W., Marx, P., and Schneider, W. (2019). *LRS-Screening: Laute, Reime, Sprache – Würzburger Screening zur Früherkennung von Lese-Rechtschreibschwierigkeiten.* Göttingen: Hogrefe Verlag.

Ennemoser, M., Marx, P., Weber, J., and Schneider, W. (2012). Spezifische Vorläuferfertigkeiten der Lesegeschwindigkeit, des Leseverständnisses und des Rechtschreibens. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie* 44, 53–67. doi: 10.1026/0049-8637/a000057

Frith, U. (1980). "Unexpected spelling problems," in *Cognitive Processes in Spelling* (London: Academic Press), 495–515.

Frith, U. (1986). A developmental framework for developmental dyslexia. *Ann. Dyslexia* 36, 67–81. doi: 10.1007/BF02648022

Fugate, M. H. (1997). Letter training and its effect on the development of beginning reading skills. *School Psychol. Quart.* 12, 170–192. doi: 10.1037/h0088957

Furnes, B., and Samuelsson, S. (2009). Preschool cognitive and language skills predicting Kindergarten and Grade 1 reading and spelling: a cross-linguistic comparison. *J. Res. Read.* 32, 275–292. doi: 10.1111/j.1467-9817.2009.01393.x

Furnes, B., and Samuelsson, S. (2010). Predicting reading and spelling difficulties in transparent and opaque orthographies: a comparison between Scandinavian and US/Australian children. *Dyslexia* 16, 119–142. doi: 10.1002/dys.401

Gersten, R., Haymond, K., Newman-Gonchar, R., Dimino, J., and Jayanthi, M. (2020). Meta-analysis of the impact of reading interventions for students in the primary grades. *J. Res. Educ. Eff.* 13, 1–27. doi: 10.1080/19345747.2019.1689591

Giles, D. C., and Terrell, C. D. (1997). Visual sequential memory and spelling ability. *Educ. Psychol.* 17, 245–253. doi: 10.1080/0144341970170301

Görgen, R., De Simone, E., Schulte-Körne, G., and Moll, K. (2021). Predictors of reading and spelling skills in German: the role of morphological awareness. *J. Res. Read.* 44, 210–227. doi: 10.1111/1467-9817.12343

Groth, K., Lachmann, T., Riecker, A., Muthmann, I., and Steinbrink, C. (2011). Developmental dyslexics show deficits in the processing of temporal auditory information in German vowel length discrimination. *Read. Writ.* 24, 285–303. doi: 10.1007/s11145-009-9213-7

Hall, C., Dahl-Leonard, K., Cho, E., Solari, E. J., Capin, P., Conner, C. L., et al. (2023). Forty years of reading intervention research for elementary students with or at risk for dyslexia: a systematic review and meta-analysis. *Read. Res. Q.* 58, 285–312. doi: 10.1002/rrq.477

Hintikka, S., Landerl, K., Aro, M., and Lyytinen, H. (2008). Training reading fluency: is it important to practice reading aloud and is generalization possible? *Ann. Dyslexia* 58, 59–79. doi: 10.1007/s11881-008-0012-7

Kirby, J. R., Parrila, R. K., and Pfeiffer, S. L. (2003). Naming speed and phonological awareness as predictors of reading development. *J. Educ. Psychol.* 95, 453–464. doi: 10.1037/0022-0663.95.3.453

Landerl, K. (2003). Categorization of vowel length in German poor spellers: an orthographically relevant phonological distinction. *Appl. Psycholinguist.* 24, 523–538. doi: 10.1017/S0142716403000262

Landerl, K., Freudenthaler, H. H., Heene, M., De Jong, P. F., Desrochers, A., Manolitsis, G., et al. (2019). Phonological awareness and rapid automatized naming as longitudinal predictors of reading in five alphabetic orthographies with varying degrees of consistency. *Scientif. Stud. Read.* 23, 220–234. doi: 10.1080/10888438.2018.1510936

Landerl, K., and Moll, K. (2010). Comorbidity of learning disorders: prevalence and familial transmission. *J. Child Psychol. Psychiatry* 51, 287–294. doi: 10.1111/j.1469-7610.2009.02164.x

Landerl, K., Ramus, F., Moll, K., Lyytinen, H., Leppänen, P. H. T., Lohvansuu, K., et al. (2013), Predictors of developmental dyslexia in European orthographies with varying complexity. *J. Child Psychol. Psychiat.* 54, 686–694. doi: 10.1111/jcpp.12029

Landerl, K., and Wimmer, H. (2000). Deficits in phoneme segmentation are not the core problem of dyslexia: evidence from German and English children. *Appl. Psycholinguist.* 21, 243–262. doi: 10.1017/S0142716400002058

Landerl, K., and Wimmer, H. (2008). Development of word reading fluency and spelling in a consistent orthography: an 8-year follow-up. *J. Educ. Psychol.* 100, 150. doi: 10.1037/0022-0663.100.1.150

Leppänen, U., Niemi, P., Aunola, K., and Nurmi, J.-E. (2006). Development of reading and spelling Finnish from preschool to grade 1 and grade 2. *Scientif. Stud. Read.* 10, 3–30. doi: 10.1207/s1532799xssr1001_2

Lervåg, A., and Hulme, C. (2009). Rapid automatized naming (RAN) taps a mechanism that places constraints on the development of early reading fluency. *Psychol. Sci.* 20, 1040–1048. doi: 10.1111/j.1467-9280.2009.02405.x

Lervåg, A., and Hulme, C. (2010). Predicting the growth of early spelling skills: Are there heterogeneous developmental trajectories? *Scientif. Stud. Read.* 14, 485–513. doi: 10.1080/10888431003623488

Liu, D., Chen, X., and Wang, Y. (2016). The impact of visual-spatial attention on reading and spelling in Chinese children. *Read. Writ.* 29, 1435–1447. doi: 10.1007/s11145-016-9644-x

May, P., Malitzky, V., and Vieluf, U. (2019). *HSP 1 Plus - Hamburger Schreib-Probe*. Stuttgart: Ernst Klett Verlag.

McWeeny, S., Choi, S., Choe, J., LaTourrette, A., Roberts, M. Y., and Norton, E. S. (2022). Rapid automatized naming (RAN) as a kindergarten predictor of future reading in english: A systematic review and meta-analysis. *Read. Res. Q.* 57, 1187–1211. doi: 10.1002/rrq.467

Mehlhase, H., Bakos, S., Landerl, K., Schulte-Körne, G., and Moll, K. (2019). Orthographic learning in children with isolated and combined reading and spelling deficits. *Child Neuropsychology* 25, 370–393. doi: 10.1080/09297049.2018.1470611

Melchers, P., and Melchers, M. (2015). *KABC-II. Kaufman Assessment Battery for Children–II.* Deutschsprachige Fassung, A. S. Kaufman and N. L. Kaufman (Frankfurt: Pearson Vue Test Center).

Mohamed, M. B. H., and O'Brien, B. A. (2022). Defining the relationship between fine motor visual-spatial integration and reading and spelling. *Read. Writ.* 35, 877–898. doi: 10.1007/s11145-021-10165-2

Moll, K., Fussenegger, B., Willburger, E., and Landerl, K. (2009). RAN is not a measure of orthographic processing. Evidence from the asymmetric German orthography. *Scientif. Stud. Read.* 13, 1–25. doi: 10.1080/10888430802631684

Moll, K., Gangl, M., Banfi, C., Schulte-Körne, G., and Landerl, K. (2020). Stability of deficits in reading fluency and/or spelling. *Sci. Stud. Read.* 24, 241–251. doi: 10.1080/10888438.2019.1659277

Moll, K., and Landerl, K. (2009). Double dissociation between reading and spelling deficits. *Scientif. Stud.Read.* 13, 359–382. doi: 10.1080/10888430903162878

Moll, K., and Landerl, K. (2010). *SLRT-II: Lese- und Rechtschreibtest; Weiterentwicklung des Salzburger Lese- Und Rechtschreibtests (SLRT)* Göttingen: Huber.

Moll, K., Ramus, F., Bartling, J., Bruder, J., Kunze, S., Neuhoff, N., et al. (2014). Cognitive mechanisms underlying reading and spelling development in five European orthographies. *Learn. Instruct.* 29, 65–77. doi: 10.1016/j.learninstruc.2013.09.003

Nielsen, A.-M. V., and Juul, H. (2016). Predictors of early versus later spelling development in Danish. *Read. Writ.* 29, 245–266. doi: 10.1007/s11145-015-9591-y

Ningrum, A., and Wibowo, R. (2017). Intelligence quotient (IQ) as a predictor of reading comprehension and writing achievement of efl learners. *JEELS* 4, 53–79. doi: 10.30762/jeels.v4i1.331

Otaiba, S. A., Puranik, C. S., Rouby, D. A., Greulich, L., Sidler, J. F., and Lee, J. (2010). Predicting kindergarteners' end-of-year spelling ability based on their reading, alphabetic, vocabulary, and phonological awareness skills, as well as prior literacy experiences. *Learn. Disabil. Quart.* 33, 171–183. doi: 10.1177/073194871003300306

Papadopoulos, T. C., Spanoudis, G. C., and Georgiou, G. K. (2016). How is RAN related to reading fluency? A comprehensive examination of the prominent theoretical accounts. *Front. Psychol.* 7, 1217. doi: 10.3389/fpsyg.2016.01217

Petermann, F. (2017). *Wechsler Intelligence Scale for Children - Fifth Edition (Deutsche Bearbeitung)*. Midlothian, VA: Pearson.

Pickering, S. J., and Gathercole, S. E. (2001). *The Working Memory Test Battery for Children*. Midlothian, VA: Pearson.

Pontart, V., Bidet-Ildei, C., Lambert, E., Morisset, P., Flouret, L., and Alamargot, D. (2013). Influence of handwriting skills during spelling in primary and lower secondary grades. *Front. Psychol.* 4, 818. doi: 10.3389/fpsyg.2013.00818

Post, Y., Swank, P., Hiscock, M., and Fowler, A. (1999). Identification of vowel speech sounds by skilled and less skilled readers and the relation with vowel spelling. *Ann. Dyslexia* 49, 161–194. doi: 10.1007/s11881-999-0023-z

Preßler, A.-L., Könen, T., Hasselhorn, M., and Krajewski, K. (2014). Cognitive preconditions of early reading and spelling: a latent-variable approach with longitudinal data. *Read. Writ.* 27, 383–406. doi: 10.1007/s11145-013-9 449-0

Protopapas, A., Altani, A., and Georgiou, G. K. (2013). Development of serial processing in reading and rapid naming. *J. Exp. Child Psychol.* 116, 914–929. doi: 10.1016/j.jecp.2013.08.004

Protopapas, A., Katopodi, K., Altani, A., and Georgiou, G. K. (2018). Word reading fluency as a serial naming task. *Scientif. Stud. Read.* 22, 248–263. doi: 10.1080/10888438.2018.1430804

Ritchie, S. J., Bates, T. C., Der, G., Starr, J. M., and Deary, I. J. (2013). Education is associated with higher later life IQ scores, but not with faster cognitive processing speed. *Psychol. Aging* 28, 515–521. doi: 10.1037/a0030820

Schmitterer, A. M., and Schroeder, S. (2018). The recognition of letters in emergent literacy in German: evidence from a longitudinal study. *J. Res. Read.* 41, 423–437. doi: 10.1111/1467-9817.12116

Schmitterer, A. M., and Schroeder, S. (2019). Effects of reading and spelling predictors before and after school entry: Evidence from a German longitudinal study. *Learn. Instruct.* 59, 46–53. doi: 10.1016/j.learninstruc.2018.09.005

Schneider, W., and Niklas, F. (2017). Intelligence and verbal short-term memory/working memory: their interrelationships from childhood to young adulthood and their impact on academic achievement. *J Intell* 5, 2. doi: 10.3390/jintelligence5020026

Schulte-Körne, G. (2010). Diagnostik und Therapie der Lese-Rechtschreib-Störung. *Deutsches Ärzteblatt* 107, 718—727. doi: 10.3238/arztebl.2010.0718

Stainthorp, R., Powell, D., and Stuart, M. (2013). The relationship between rapid naming and word spelling in English. *J. Res. Read.* 36, 371–388. doi: 10.1111/jrir.12002

Steffler, D. J. (2001). Implicit cognition and spelling development. *Dev. Rev.* 21, 168–204. doi: 10.1006/drev.2000.0517

Steinbrink, C., Groth, K., Lachmann, T., and Riecker, A. (2012). Neural correlates of temporal auditory processing in developmental dyslexia during German vowel length discrimination: an fMRI study. *Brain Lang.* 121, 1–11. doi: 10.1016/j.bandl.2011.12.003

Steinbrink, C., Klatte, M., and Lachmann, T. (2014). Phonological, temporal and spectral processing in vowel length discrimination is impaired in German primary school children with developmental dyslexia. *Res. Dev. Disabil.* 35, 3034–3045. doi: 10.1016/j.ridd.2014.07.049

Suggate, S., Pufke, E., and Stoeger, H. (2018). Do fine motor skills contribute to early reading development? *J. Res. Read.* 41, 1–19. doi: 10.1111/1467-9817.12081

Tabachnick, B. G., Fidell, L. S., and Ullman, J. B. (2013). *Using Multivariate Statistics (Vol. 6).* Boston, MA: Pearson.

Thaler, V., Ebner, E. M., Wimmer, H., and Landerl, K. (2004). Training reading fluency in dysfluent readers with high reading accuracy: Word specific effects but low transfer to untrained words. *Ann. Dyslexia* 54, 89–113. doi: 10.1007/s11881-004-0005-0

Treiman, R., and Kessler, B. (2005). "Writing systems and spelling development," in *The Science of Reading: A Handbook.* (Hoboken, NJ: Blackwell Publishing), 120–134.

Vaessen, A., and Blomert, L. (2013). The cognitive linkage and divergence of spelling and reading development. *Scientif. Stud. Read.* 17, 89–107. doi: 10.1080/10888438.2011.614665

Vander Stappen, C., and Reybroeck, M. V. (2018). Phonological awareness and rapid automatized naming are independent phonological competencies with specific impacts on word reading and spelling: an intervention study. *Front. Psychol.* 9, 320. doi: 10.3389/fpsyg.2018.00320

Vellutino, F. R., Fletcher, J. M., Snowling, M. J., and Scanlon, D. M. (2004). Specific reading disability (dyslexia): What have we learned in the past four decades? *J. Child psychol. Psychiat.* 45, 2–40. doi: 10.1046/j.0021-9630.2003.00305.x

Verhoeven, L., and Perfetti, C. (2022). Universals in learning to read across languages and writing systems. *Scientif. Stud. Read.* 26, 150–164. doi: 10.1080/10888438.2021.1938575

Wimmer, H., and Mayringer, H. (2002). Dysfluent reading in the absence of spelling difficulties: a specific disability in regular orthographies. *J. Educ. Psychol.* 94, 272. doi: 10.1037/0022-0663.94.2.272

Wimmer, H., Mayringer, H., and Landerl, K. (2000). The double-deficit hypothesis and difficulties in learning to read a regular orthography. *J. Educ. Psychol.* 92, 668. doi: 10.1037/0022-0663.92.4.668

Witzel, B., Görgen-Rein, R., Galuschka, K., Huemer, S., Schulte-Körne, G., and Moll, K. (2023). Digital game-based spelling intervention for children with spelling deficits: a randomized controlled trial. *Learn. Instruct.* 89. doi: 10.1016/j.learninstruc.2023.101842

World Health Organization (2022). *International Classification of Diseases Eleventh Revision (ICD-11)*. Geneva: WHO.

Yeung, P.-S., Ho, C. S.-H., Wong, Y.-K., Chan, D. W.-O., Chung, K. K.-H., and Lo, L.-Y. (2013). Longitudinal predictors of Chinese word reading and spelling among elementary grade students. *Appl. Psycholinguist.* 34, 1245–1277. doi: 10.1017/S0142716412000239

Zarić, J., Hasselhorn, M., and Nagler, T. (2021). Orthographic knowledge predicts reading and spelling skills over and above general intelligence and phonological awareness. *Eur. J. Psychol. Educ.* 36, 21–43. doi: 10.1007/s10212-020-00464-7

Ziegler, J. C., Bertrand, D., Tóth, D., Csépe, V., Reis, A., Faísca, L., et al. (2010). Orthographic depth and its impact on universal predictors of reading: a cross-language investigation. *Psychol. Sci.* 21, 551–559. doi: 10.1177/0956797610363406

frontiers | Frontiers in Education

TYPE Original Research
PUBLISHED 12 April 2024
DOI 10.3389/feduc.2024.1350266

Check for updates

OPEN ACCESS

EDITED BY
Martin Schöfl,
University of Education Upper Austria, Austria

REVIEWED BY
Faye Antoniou,
National and Kapodistrian University of Athens, Greece
Susanne Seifert,
University of Graz, Austria

*CORRESPONDENCE
Christina Stuhr
✉ christina.stuhr@uni-rostock.de

RECEIVED 07 December 2023
ACCEPTED 28 March 2024
PUBLISHED 12 April 2024

CITATION
Stuhr C, Yumus M, Meindl M, Jungmann T and Hughes CML (2024) Exploration of latent early literacy profiles in German kindergarten children using a newly developed app.
Front. Educ. 9:1350266.
doi: 10.3389/feduc.2024.1350266

Exploration of latent early literacy profiles in German kindergarten children using a newly developed app

Christina Stuhr[1]*, Melike Yumus[2], Marlene Meindl[1], Tanja Jungmann[2] and Charmayne Mary Lee Hughes[3]

[1]Department of Special Education, Institute for Special Educational Development and Rehabilitation, University of Rostock, Rostock, Germany, [2]Department of Special Needs Education and Rehabilitation, Carl-von-Ossietzky University of Oldenburg, Oldenburg, Germany, [3]Age-Appropriate Human-Machine Systems, Institute of Psychology and Ergonomics, Technische Universität Berlin, Berlin, Germany

Introduction: This study explores the effectiveness of the tablet-based EuLeApp© assessment tool in identifying distinct latent early literacy profiles among German kindergarten children aged between 4.0 and 7.0 years. Additionally, the study investigates how factors such as age, biological sex, and language impairment contribute to the manifestation and prevalence of these latent literacy profiles.

Methods: Utilizing latent class analysis on a diverse sample of 285 children, we administered a prototype of the EuLeApp© assessment tool to measure early literacy skills (i.e., print awareness, concepts of print, word awareness, phonological awareness, alphabet knowledge, and narrative skills). The sample included children aged between 4.0 and 7.0 years from various demographic backgrounds. Analysis involved identifying distinct early literacy profiles based on the assessment results. We also examined the influence of age, biological sex, and language impairment on the manifestation and prevalence of these profiles.

Results: The analysis revealed four stable and interpretable early literacy profiles: Exceptional Performers (17%), Typical Performers (41%), Marginal Performers (35%), and Subpar Performers (7%). These profiles were found to be shaped by the complex interplay of age, biological sex, and language impairment status among the children.

Discussion: These findings underscore the importance of employing person-centered methodologies to characterize early childhood literacy profiles, providing a framework for tailoring intervention programs to suit the unique characteristics and requirements of individual children.

KEYWORDS

early literacy, kindergarten, language impairment, latent class analysis, mobile application

1 Introduction

Although an overwhelming majority of illiterate individuals live in developing countries (World Literacy Foundation, 2018), low literacy and numeracy proficiencies are also observed in developed nations. For example, a 2018 study conducted in Germany reported low literacy skills in 12.1% of the German-speaking adult population (Buddeberg et al., 2021). Additionally, literacy difficulties have been identified among German children, with Wittig and Schneider (2022) reporting that nearly 18.8% of all fourth-grade children in Germany (around 10 years of age) fail to achieve the minimum reading standard, and approximately 30.4% falling short of the minimum spelling standard.

The existing body of literature strongly establishes that the acquisition of reading and writing skills takes place long before formal education begins. Known as emergent or early literacy (Clay, 1966), it involves various foundational skills (e.g., alphabet knowledge, recognizing the link between spoken and written language, grasping the foundational phonological structure of a language, as well as oral linguistic abilities such as vocabulary and grammar), that are necessary prerequisites for future reading success. One view of emergent literacy, forwarded by Lonigan and colleagues (Whitehurst and Lonigan, 1998; Lonigan and Shanahan, 2010) describes the construct of emergent literacy as consisting of two interdependent components: outside-in and inside-out dimensions. Outside-in processes encompass children's understanding of print conventions, their emergent reading (e.g., reading environmental print), vocabulary, and narrative skills. In contrast, the inside-out component refers to children's letter-name and letter-sound knowledge, their phonetic (i.e., invented) spelling, as well as language-based components (e.g., phonological and syntactic awareness). These various factors of early literacy skills influence the development of reading skills in different ways and at different times. While alphabet knowledge, word awareness, and phonological awareness are fundamental for the development of word reading skills (Lonigan et al., 2000; Roth et al., 2002; Lenel and Knopf, 2015; Hjetland et al., 2019), oral language (such as narrative skills and vocabulary) predict later reading comprehension skills (Sénéchal et al., 2001; Roth et al., 2002; Hjetland et al., 2019; Babayiğit et al., 2021). As such, it is argued that kindergarten narrative skills and vocabulary enhances future reading comprehension (Sénéchal et al., 2001; Babayiğit et al., 2021).

To address the early identification and support of children vulnerable to reading and spelling difficulties, it is imperative to transcend the prevailing reactive approach and broaden our efforts by creating and implementing preventative strategies (Murphy et al., 2016). Effective implementation of evidence-based instructional practices and progress monitoring assessments is paramount, particularly in early childhood education, necessitating the use of accurate and valid assessment tools (Gorman et al., 2016; Antoniou et al., 2022). Embracing principles of primary and secondary prevention, it is crucial to detect the underlying causes and initial indicators of future reading challenges at the earliest possible stage, well before these difficulties manifest. Subsequently, a personalized and focused intervention should be employed to improve literacy outcomes (Hudson et al., 2023). Scholarly discourse underscores the initiation of screening and support processes during kindergarten, a pivotal stage for foundational reading skill development (Wanzek and Vaughn, 2007; Gaab and Petscher, 2022). Empirical evidence highlights literacy achievement disparities emerging as early as kindergarten, persisting and widening between typical and atypical readers through 12th grade in the absence of intervention (Ferrer et al., 2015). Consistent with these findings, interventions targeting reading skills in kindergarten yield notably higher effect sizes (across studies Cohen's d: 0.18–0.84) compared to those initiated later in second or third grade (across studies Cohen's d: −0.05 to 0.27) (Wanzek and Vaughn, 2007).

Furthermore, a reciprocal relationship between reading self-concept and competence emerges as early as 5 years of age (Morgan et al., 2008; Walgermo et al., 2018a,b) and persists through middle childhood (e.g., 11 years, Sewasew and Koester, 2019). For example, Morgan et al. (2008) reported that within just 6 months of starting the first grade, students with poorer early literacy skills exhibited less confidence in their reading abilities compared to their peers, a discrepancy that endured for 3 years. These empirical findings suggest that children with weaker reading self-concept often demonstrate poor reading abilities, leading to avoidance of reading-related tasks due to the anticipation of failure. Consequently, they struggle to find ways to develop proficient reading skills, and often remain low-performing readers as they mature. Early identification and intervention targeting both reading and spelling skills, along with precursor skills like early literacy, have the potential to significantly reduce the prevalence of reading challenges among young learners, thereby enhancing adult literacy rates.

Another factor that shapes a child's early literacy skills is their level of language (Cabell et al., 2010; Justice et al., 2015). For instance, compared to their peers who are developing typically, children with language development disorders [e.g., Specific Language Impairment (SLI)] typically exhibit lower proficiency in several aspects of early literacy such as phonological awareness in a broader sense (e.g., rhyming), productive knowledge of letters (including letter names), writing skills, and various facets of storytelling abilities (Cabell et al., 2010; Justice et al., 2015; Pavelko et al., 2018). In addition, children with SLI commonly encounter difficulties with reading fluency, reading comprehension, and other academic tasks that rely heavily on language skills, with difficulties persisting into primary school grades and beyond (Snowling et al., 2016).

Clustering approaches have emerged as a useful analytic tool to identify the linguistic profiles of children with SLI (e.g., Cabell et al., 2010, 2011, 2013; Justice et al., 2015), which may be informative for examining and understanding variability in the literacy skills. For example, Cabell et al. (2010) used K-means cluster analysis to group 59 pre-school aged children (48–60 months, ~4–5 years) with SLI based on their emergent literacy skills and oral language skills. Their analysis revealed three distinct early literacy clusters arising from the data, characterized by a combination of age, language proficiency, and receptive and expressive language competencies. Similar results were obtained by Justice et al. (2015) who employed latent profile analysis to explore early literacy profiles of 218 three-to five-year-old children with language impairment, with four profiles emerging that distinguished by variations in early literacy scores and spoken language skills. These analytical approaches offer the potential to move beyond broad group-based generalizations concerning early literacy skills and ascertain whether distinct profiles exist to delineate early literacy development. This differentiation could, in turn, provide insights into the risk factors and resilience factors shaping the journey of young children as they progress toward becoming proficient readers.

1.1 The evaluation of early literacy

In the Anglo-Saxon (e.g., the United States, the United Kingdom) literature, early literacy knowledge and skills can be comprehensively evaluated using paper and pen assessments, such as the Dynamic Indicators of Basic Early Literacy Skills (DIBELS, Good and Kaminski, 2002), the Test of Preschool Early Literacy (TOPEL, Lonigan et al., 2007), or the Phonological Awareness and Literacy Screening PreK (PALS-PreK, Invernizzi et al., 2004). Until recently, a comprehensive standardized German language early literacy assessment had been lacking, with empirical studies having to adapt existing English tests to measure various early literacy components (Niklas and Schneider, 2013; Fricke et al., 2016). Inspired by the absence of a suitable tool, Meindl and Jungmann (2014) embarked on the creation of a novel German-language assessment tailored to evaluate early literacy and narrative skills among children aged 4 to 5 years (EuLe 4–5; Erzähl-und Lesekompetenzen erfassen bei 4-bis 5-jährigen Kindern). The EuLe 4–5 is aligned with the component model proposed by Whitehurst and Lonigan (1998) and evaluates five early literacy factors: concepts of print, print awareness, word awareness, alphabet knowledge, and narrative skills. The internal consistency of the EuLe 4–5 varies between 0.78 (print knowledge scale) and 0.95 (alphabet knowledge scale) (Meindl and Jungmann, 2015).

The conventional paper-and-pen formats used in early literacy assessment present several challenges, including difficulties in engaging children due to the static nature of paper-based assessments, challenges in achieving standardized administration due to the diverse range of children's abilities and behaviors, and the labor-intensive nature of manual scoring, which can lead to delays in delivering timely feedback and interventions. However, the proliferation of smartphones, tablets, and other portable devices, along with improvements in internet connectivity, cloud computing, and advanced data analytics technologies, has facilitated the implementation of school-based assessments on a large scale. Digital assessment tools offer distinct advantages over traditional paper-and-pen exams (Hipkins and Cameron, 2018). They allow for dynamic adjustment of question difficulty based on students' responses, provide immediate feedback on students' performance, progress, and areas of difficulty, and adhere to established guidelines and criteria, ensuring consistency and reliability across diverse testing conditions.

In response to the identified gap in available resources and in alignment with the objectives outlined in the German Digital Strategy 2025 (Federal Ministry for Economic Affairs and Energy, 2016), we have embarked on the modernization of the pen-and-paper-based EuLe 4–5 assessment into a tablet-based tool, known as the EuLeApp© (Meindl et al., 2022). As a pivotal aspect of our ongoing efforts, our goal is to explore whether a prototype of this newly developed digital assessment possesses the capability to distinguish between various latent early literacy profiles among children enrolled in German Kindergartens. In this study, we utilize a person-centered Latent Class Analysis (LCA) approach to identify profiles and their prevalence in early literacy skills among this sample of children. Furthermore, we endeavor to elucidate how factors such as a child's age, sex, and language impairment relate to the variety and occurrence of these latent literacy profiles. By employing a person-centered clustering technique and considering diverse influencing factors, our aim is to deepen our understanding of early literacy development and inform the design of tailored interventions to support children's literacy skills.

2 The German early childhood education and care system

The considerable disparities in educational frameworks between the Anglo-Saxon world and Germany raise the possibility that the aforementioned outcomes may not extend to children who undergo their initial educational experiences in German early childhood education and care (ECEC) centers.

Within the German ECEC landscape, two primary programs are prominent. Traditionally, facilities designed for children aged 0–3 are referred to as Kinderkrippen or simply Krippen (crèches), while those catering to children aged 3–6 are known as Kindertagesstätte (commonly colloquialized as Kita or Kindergarten). German ECEC services prioritize nurturing children's independence, fostering critical thinking skills, encouraging ownership of the learning process, and promoting creative problem-solving abilities. While formal reading instruction does not occur in kindergarten, children may encounter reading and writing activities during craft projects, story-time sessions, local excursions, and outdoor play. Formal schooling typically commences around the ages of six to seven when children enter Grundschule (equivalent to primary school) and embark on the formal journey of learning to read and write (Mann and Wimmer, 2002).

Governance over German ECEC is a collaborative effort among federal, state, and municipal authorities (Deutscher Bildungsserver, 2023), with the federal government mandating that ECEC services (1) promote the development of the child into a self-determined, independent, and socially competent personality, (2) support and supplement upbringing and education in the family, and (3) help parents to better reconcile employment, child rearing and family care (Section 22 subsection 2 SGB VIII; Bundesministerium der Justiz, 1990). Thus, education in German ECECs can be viewed as providing both education and care, rather than being primarily focused on enhancing school readiness (European Commission, 2023).

The decentralization of ECEC services gives rise to a diverse array of pedagogical curricula and practices (Resa et al., 2016; European Commission, 2023). However, this decentralization has also resulted in varying quality standards, often falling short of evidence-based recommendations (Viernickel et al., 2015), with actual conditions frequently failing to align with evidence-based recommendations (Bock-Famulla et al., 2017). Consequently, despite high attendance rates in kindergartens, significant variations exist in the early literacy skills of children as they start primary school (e.g., Brügelmann, 1998). These differences can be attributed to the inconsistent exposure children receive to early literacy activities in ECEC environments (Tietze et al., 1998; Wirts et al., 2017). For example, Wirts et al. (2017) explored literacy-related activities within 34 German ECEC settings and reported that pedagogical staff incorporated picture book activities into the daily schedule 0.895 times per day ($SD = 1.0$) lasting 14 min on average ($SD = 9.09$), with such activities initiated by one or more children in 53% of instances. Phonological awareness and early writing activities were even more rarely found, occurring only 0.485 times per day ($SD = 0.55$), and lasting 15 min on average ($SD = 15.37$).

In addition, significant regional disparities exist in the availability of early childhood education and care (ECEC) slots across German states, counties, and municipalities. For example, among children aged 3–6 years, 90% of children in Thuringia attended an ECEC institution for more than 7 h a day, whereas only 24% of children in Baden-Wuerttemberg met this criterion (Statistisches Bundesamt, 2018). Furthermore, the inadequate number of ECEC facilities often fails to meet the demand for childcare, resulting in lengthy waiting lists and delays in securing placements. Consequently, it is not uncommon for children to enter German ECEC centers at the age of four or even 5 years old.

3 Methods

3.1 Participants

Kindergartens from the Lower Saxony and Mecklenburg-West Pomerania states of Germany were recruited for the present study. Eligibility for the study required that the children: (1) had parental consent to participate, (b) were between the ages of 3.11 and 6.11 years, (c) demonstrated sufficient German language proficiency to comprehend the instructions in the language and literacy assessments (assessed via a picture search task, Petermann et al., 2016), and (d) had no cognitive and communicative disabilities (e.g., Down syndrome, autism spectrum disorder). To mitigate selection bias and ensure the representativeness of the target population, recruitment strategies employed a two-stage sampling method. Kindergartens in Lower Saxony and Mecklenburg-West Pomerania received invitations via both regular mail and email, accompanied by a study pamphlet and a link to the study webpage for detailed information. Persistent efforts, including telephone follow-ups and personal visits, were made in cases of non-response from kindergartens. Upon obtaining consent from kindergarten directors, recruitment materials, comprising informed consent forms and study pamphlets, were distributed to parents/caregivers by kindergarten teachers, outlining comprehensive study details.

The study sample consisted of 285 children from 14 local Kindergartens in the Lower Saxony and Mecklenburg-West Pomerania states of Germany. Fifty four percent of the participants were males, and the average age of the children was 62.37 months (range 47–78 months, SD = 7.39). The experiment conformed to the declaration of Helsinki and was approved by the local school authorities and the institutional review boards of University of Rostock and Carl-von-Ossietzky University of Oldenburg. Informed consent was obtained from the parents of the children prior to participation in the experiment.

3.1.1 Early literacy measures

To assess early literacy skills, we administered the prototype of a newly developed assessment tool, the EuleApp© (Meindl et al., 2022). It allows for the assessment of six different early literacy skills: concepts of print, print awareness, word awareness, phonological awareness, alphabet knowledge, and narrative skills. The order of sub-tests was consistent across participants [in the order below], with questions presented in ascending order of difficulty within each sub-test. Raw scores from the EuleApp© subtests were saved in the app, but subsequent data analysis was conducted using custom Python scripts.

The assessment of children's print awareness comprised a 40-item subtest, wherein children were tasked with distinguishing words from pictures, numbers, and symbols ("Tap on the letter"; "Tap on writing"). Each child was presented with four pictures, one being the target and the other three serving as distractors, and they were required to tap either on the word or the letter corresponding to the target. The internal consistency for correct items ranged between $\alpha = 0.93$ and $\alpha = 0.94$.

Children's concepts of print were assessed using a 19-item measure involving print-related tasks, such as identifying reading direction ("Where do you begin reading? Touch it") or recognizing the front of a book. The internal consistency for correct items ranged between $\alpha = 0.75$ and $\alpha = 0.77$.

Children's word awareness was measured using a 21-item measure that increased in difficulty across items. Children's word awareness was measured with a 21-item measure that progressively increased in difficulty. During this assessment, children encountered short texts and were directed to tap on specific elements, such as the first word, the second word, or the space between two words ("Tap on the space between two words"). The reliability score for correct items ranged between $\alpha = 0.74$ and $\alpha = 0.75$.

The phonological awareness subtest comprises 21 items and assesses two key components: the synthesis of syllables and phonemes to form words, and the analysis of the syllabic and phonemic structure of words. For example, tasks include identifying the initial sounds of words (e.g., "Here you can see three pictures: Grandma, Mum, Apple. Touch the picture that starts with /m/"). During this subtest, children were presented with three pictures, one being the target and the others serving as distractors, and they were required to select the appropriate picture representing the correct answer. The internal consistency for correct items ranged between $\alpha = 0.67$ and $\alpha = 0.72$.

To assess children's receptive alphabet knowledge, phonetic realizations of letters were presented, and the children's task was to select the corresponding letter from a set of four options (e.g., "Tap the /m/"). In the productive segment, children were asked if they could name the 26 letters of the alphabet. Children's first receptive and productive reading abilities on the syllable and word level were assessed using items such as "Tap on /am/", "Tap on /mama/" (receptive segment), and "What is written here?" (productive segment). The internal consistency for correct items ranged between $\alpha = 0.95$ and $\alpha = 0.96$.

The "Hungry Seagull Marius" story, consisting of seven pictures, was developed to evaluate children's narrative skills. Initially, the story's pictures were presented to the child to provide an overview. Subsequently, the pictures were shown in pairs, with the instruction for the child to tell the story. We analyzed the macrostructure (story grammar) of the narratives, including components such as setting, characters, initiating event, internal response, plan, attempts, direct consequence, and resolution (with a possible score range of 0–21). Inter-rater reliability was confirmed by double scoring a random selection of 30% of the audio samples and resulted in Kendall's Tau of 0.84.

3.1.2 Defining language impairment

To identify children with language impairment, we utilized either the "Language level test for children aged 3–5 years" (Sprachentwicklungstest für Kinder 3–5 [SET 3–5]; Petermann et al., 2016) or the "Language level test for children aged 5–10 years" (Sprachentwicklungstest für Kinder 5–10 [SET 5–10]; Petermann, 2018), depending on the child's age. The SET 3–5 comprises 12 subtests that measure a child's receptive language processing skills (understanding, recording), productive language processing skills (own speech acts), and auditory memory skills (language memory). Similarly, the SET 5–10 consists of 8 subtests that measure a child's vocabulary, semantic relations, processing speed, language comprehension, language production, grammar/morphology, and auditory memory. The internal consistency for administered subtests from the SET 3–5 ranged between $\alpha = 0.70$ and $\alpha = 0.93$ (Petermann et al., 2016), while it ranged between $\alpha = 0.71$ and $\alpha = 0.91$ for the SET 5–10 (Petermann, 2018). Raw scores obtained in each subtest were converted into standardized T-scores (mean: 50, SD: 10). Children scoring at or below the 10th percentile in at least two subtests were classified as language impaired.

3.1.3 Procedure

All children were tested individually in a quiet room at their respective kindergarten. The tasks were administered by a trained experimenter according to the procedures of the EuLe-F and SET manuals and was the same for all children, which served to reduce bias during data collection and analysis. Each child participated in two sessions lasting between 20 and 30 min each. During these sessions, half of the children completed the EuLeApp© early literacy measures initially, followed by the SET 3–5/5–10 assessment, while the remaining children followed the reverse order.

3.2 Data analysis

Data was analyzed using the person-centered latent class analysis (LCA) approach (Lazarsfeld and Henry, 1968; Bergman and Trost, 2006; Collins and Lanza, 2010). LCA is a form of mixture modeling that allows for the identification of unobserved classes based on their response patterns on a set of variables. One of the primary assumptions of LCA is that homogenous sub-groups exist within the dataset, have distinct probability distributions, and that class assignment is mutually exclusive (Bergman and Magnusson, 1997).

Before conducting the LCA, the dataset underwent several preprocessing steps. First, the data were examined for the number and patterns of missing values, and normality tests were performed by calculating skewness and kurtosis. Additionally, potential outliers were identified. Children with missing data on one or more of the predictors were excluded from the analysis, resulting in the removal of 14 students (less than 4.9% of the sample). There were no specific issues regarding normality or outliers.

Class-defining variables for the LCA included the early literacy scores (obtained from the EuleApp©), presence of language impairment (obtained from either the SET 3–5 or SET 5–10), as well as the child's age and biological sex. The modeling process began by estimating the one-class model. Subsequently, we progressively examined model fits by successively increasing the number of classes until we obtained a model where a class comprised less than 5% of the total sample (Nylund-Gibson and Choi, 2018), or model estimation errors occurred.

Consistent with best-practice guidelines, a holistic evaluation of multiple fit statistics was used to determine the best fitting model (Masyn, 2013; Nylund-Gibson and Choi, 2018) and included information criterion, relative fit indices, and substantive interpretability by the project team. The relative model fit for each class model was evaluated using four information theory-based metrics including the Bayesian Information Criterion (BIC; Schwarz, 1978), sample-size adjusted BIC (SABIC; Sclove, 1987), and Consistent Akaike Information Criterion (CAIC; Bozdogan, 1987). In interpreting these criteria, the lower the BIC, AIC, and CAIC values, the better the comparative fit of the model to the k-1 class model. Absolute model fit was assessed using the log-likelihood, where a higher value (closer to 0) represents a better fit. In the case that the information theory-based metrics continued to decrease as each additional class is added (i.e., there was no global minimum), the selected information criteria was plotted and the "elbow" of the plot indicating the point of diminishing returns was selected as the best solution (Nylund-Gibson and Choi, 2018). Model fit was also examined using the Lo-Mendell-Rubin adjusted likelihood ratio test (LMR-LRT; Lo et al., 2001; Nylund-Gibson et al., 2023), with a significant p-value signifying that the estimated model with k classes fits the data better than the simpler k-1 model. In addition to quantitative criteria, the principles of parsimony and substantive interpretability of the latent class solutions were used to guide model criteria (Masyn, 2013). These complementary approaches ensured that the chosen latent class solution balanced statistical complexity with meaningful interpretation.

After identifying the best-fitting class solution, the extent to which the classes could be differentiated was examined using multiple metrics: relative entropy, average posterior probabilities, substantive knowledge and expertise by the project team, and univariate statistics. Relative entropy was used as the primary measure to determine how well the classes are differentiated (Masyn, 2013; Nylund-Gibson and Choi, 2018), and can be considered as an omnibus metric of classification accuracy (range 0–1; Celeux and Soromenho, 1996), where values > 0.80 indicate "good" assignment of individuals into classes (Ram and Grimm, 2009). This was complemented by examining the average posterior probabilities, with values > 0.70, indicating high classification precision (Nagin, 2005). Subsequently, univariate statistics (chi-square for categorical variables, ANOVAs for continuous variables) were employed to detect differences between latent groups in demographic information (e.g., biological sex, age) and language performance. Tukey *post-hoc* analysis was conducted following significant main effects in ANOVAs. Finally, the project team convened to define meaningful child profiles derived from the LCA results. These profiles were informed by the team's practical experiences working with diverse kindergarten children in educational and clinical settings.

Statistical analyses were performed in Python 3.10.9 (Python Core Team, 2019) using the Pandas (McKinney, 2010), NumPy (Harris et al., 2020), and StepMix libraries (Morin et al., 2023), as well as the tidyLPA package in r (Rosenberg et al., 2018) using a Windows 10 operating system.

4 Results

4.1 Determination of latent classes

The model fit indices are shown in Table 1. To decide on the model that best described the heterogeneity in kindergarten literacy performance, the project team holistically considered the information theory-based, classification accuracy, and substantive interpretability.

Although the log likelihood favored a 6-class solution, it was noted that one of the profiles comprised only 3% of the total sample, suggesting over-extraction (Nylund-Gibson and Choi, 2018). The LMR-LRT p-value for the 5-class model was statistically significant, but it had higher Akaike information criterion (AIC), Bayesian information criterion (BIC), consistent Akaike information criterion (CAIC), and lower entropy values compared to the 4-class model, indicating potential instability and over-extraction. The p-value of the Lo-Mendell-Rubin likelihood ratio test (LMR-LRT) for the 3-class

TABLE 1 Goodness of fit indices for the competing LCA models.

Classes	Log likelihood	AIC	BIC	CAIC	SABIC	Entropy	aLMR-LRT	Smallest class size (%)
1	−1839.01	3710.03	3767.66	3783.66	3806.57	−inf	–	1.00
2	−1539.96	3145.92	3264.79	3297.79	3345.03	0.87	**p < 0.001**	0.50
3	−1436.53	2973.05	3153.16	3203.16	**3274.73**	0.85	**p < 0.001**	0.20
4	−1371.54	**2877.08**	**3118.42**	**3185.43**	3281.33	0.86	**p < 0.001**	0.07
5	−1355.53	2879.06	3181.64	3265.64	3385.88	0.83	**p = 0.035**	0.07
6	**−1316.09**	2834.19	3198.00	3299.00	3443.57	0.85	p = 1.000	0.03

Bolded values in the tables indicate the model that was preferred by the given fit index.
aLMR-LRT Lo-Mendell-Rubin adjusted likelihood-ratio test for k-1 versus k classes.

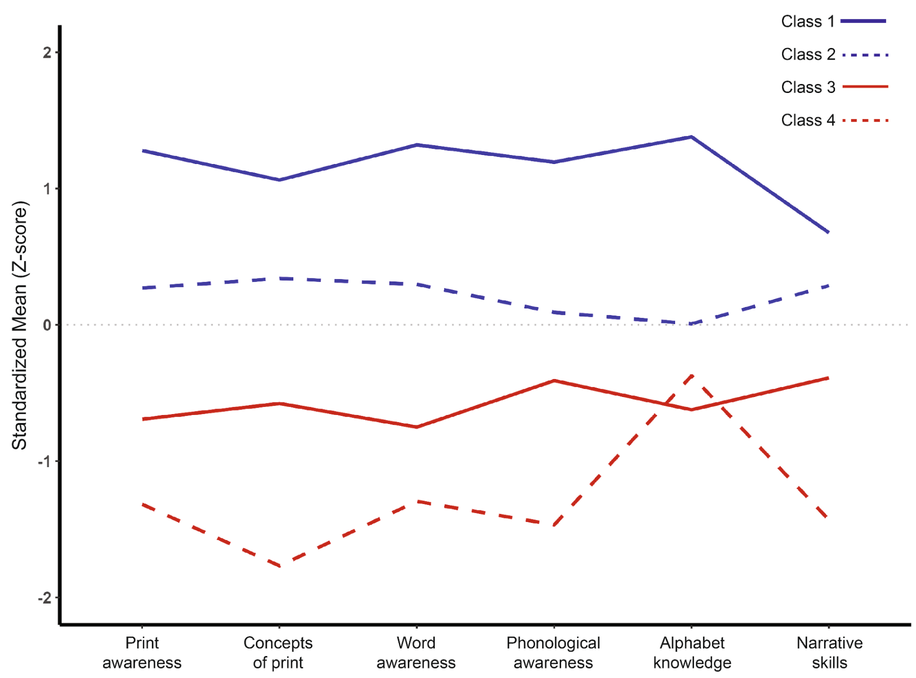

FIGURE 1
Latent profiles of kindergarten children, presented as z-scores.

model was statistically significant, and it exhibited the lowest sample-size adjusted Bayesian information criterion (SABIC) value among all tested models. In contrast, the 4-class solution exhibited lower overall information criteria values (i.e., AIC, BIC, and CAIC values), a statistically significant p-value using the LMR-LRT, while also yielding the most meaningful classes relevant to the research questions being investigated as evaluated by the project team. Considering all factors, the 4-class solution was deemed as the most parsimonious model and explored in further depth.

4.2 Latent class descriptions

The profiles of early literacy ability in the four classes are presented in Figure 1. Within-class classification was good (Masyn, 2013), as indicated by a relative entropy value of 0.86 and average posterior probabilities ranging from 0.65 (Subpar Literacy Performers) to 0.95 (Typical Literacy Performers).

Table 2 provides descriptive statistics for the early literacy skills subtests measured using the EuLeApp© for the entire sample, while Table 3 displays the distribution of sociodemographic characteristics and literacy performance according to latent class membership. There were significant differences across classes with respect to age [$F(3, 267) = 20.35$, $p < 0.001$], with Tukey *Post-Hoc* tests revealing that the age of children was similar for Classes 1 and 2 ($p = 0.95$) versus Classes 3 and 4 ($p = 0.11$). There was also a significant difference in the proportion of boys and girls across classes, with a larger proportion of boys for Class 4 than any other class [χ^2 (3, $N = 271$) = 22.16, $p < 0.001$]. The proportion of children with language impairment also differed between classes, with a far higher proportion of language impaired children in Class 4 (94.74%) and a lower proportion of language impaired children in Class 1 (10.64%) when compared to Classes 2 and 3 (36.94 and 52.13%, respectively) [χ^2 (3, $N = 271$) = 76.16, $p < 0.001$].

One-way ANOVAs revealed significant differences in the scores on the six tested literacy components across classes (all p's < 0.001). These findings indicate that the classes exhibited distinct patterns of

TABLE 2 Descriptive statistics for the early literacy skills subtests measured using the EuLeApp©.

Subtest	Mean	SD	Min	Max
Print awareness	22.78	10.35	4.00	40.00
Print knowledge	11.82	3.77	1.00	20.00
Word awareness	5.13	3.11	0.00	12.00
Phonological awareness	18.42	4.09	8.00	28.00
Letter knowledge	22.88	13.65	0.00	80.00
Narrative skills	6.38	3.69	0.00	15.00

TABLE 3 Model-estimated, class-specific means and standard deviations based on latent class membership.

	Class 1 (exceptional performers)	Class 2 (typical performers)	Class 3 (marginal performers)	Class 4 (subpar performers)
N, %	47 (17%)	111 (41%)	94 (35%)	19 (7%)
Participant characteristics				
Age (mean, SD)	66.74 (5.69)	64.15 (7.12)	58.94 (6.17)	58.05 (8.41)
Biological males (%)	40.43%	52.25%	55.32%	89.47%
Language impairment present (%)	10.64%	36.94%	52.13%	94.74%
Literacy raw scores				
Print awareness	36.02 (4.17)	25.58 (7.89)	15.62 (6.15)	9.16 (4.09)
Print knowledge	15.83 (2.03)	13.11 (2.43)	9.65 (2.65)	5.16 (2.61)
Word awareness	9.23 (1.91)	6.05 (2.17)	2.80 (1.46)	1.11 (0.99)
Phonological awareness	23.30 (2.77)	18.79 (3.21)	16.74 (2.99)	12.42 (2.50)
Letter knowledge	41.70 (15.41)	22.98 (9.15)	14.38 (4.07)	17.79 (16.13)
Narrative skills	8.87 (3.75)	7.44 (3.33)	4.95 (2.62)	1.11 (1.29)

literacy performance levels in kindergarten children and that the classes were correctly clustered according to the specific features of literacy performance.

Data from Figure 1 and Table 3 were reviewed and discussed by the project team. The group agreed on the final descriptive labels for the four kindergarten literature profiles shown in Table 3.

4.2.1 Class 1: exceptional literacy performers (*n* = 47, 17%)

Class 1 comprised 17% of the total sample and had an average age of 66.74 months. Overall, these children excelled in all aspects of early literacy, scoring notably higher in literacy measures compared to the other classes (0.68–1.38 SD above the average). Consequently, they have been categorized as having a low risk of encountering future reading difficulties. This class had the lowest proportion of children with language impairment of all groups (10.64% vs. 61.27%). Differences due to language impairment were apparent for the Alphabet Knowledge measure (difference = 0.65), but comparable for all other measures (mean difference = −0.09).

4.2.2 Class 2: typical literacy performers (*N* = 111, 41%)

Class 2 comprised the largest portion of children, accounting for 41% of the entire sample. The mean age of children in this class was 64.15 months, which was similar to Class 1 (66.74 months), but significantly higher than Classes 3 and 4 (average = 58.50 months). As can be seen in Figure 1, Class 2 was characterized by average achievement on all variables (range = 0.01–0.34). While 36.94% of children in this sample exhibited a language impairment, the presence of impairment did not influence performance on any of the tested literacy measures (mean difference = 0.70).

4.2.3 Class 3: marginal literacy performers (*n* = 94, 35%)

Class 3 encompasses 35% of the overall sample. The children in Class 2 had an average age of 58.94 months, which was significantly younger than those in Classes 1 and 3 (average = 65.45 months), but quite similar to that of Class 4 (58.05 months). Overall, the performance of children in Class 3 exhibited below average performance on all measures, especially for Word Awareness (−0.750), Print Awareness (−0.69), and Alphabet Knowledge (−0.62). There was a nearly equal distribution of non-language impaired children (47.87%) and language-impaired children (52.13%) in this class. The presence of language impairment influenced performance on the Narrative Skills measure (difference = 0.29), but none of the other tested literacy measures (mean difference = 0.002, Figure 2, bottom left panel).

4.2.4 Class 4: subpar literacy performers (*n* = 19, 7%)

The fourth class comprised 7% of the entire sample. As with Class 3, children in this class were younger (58.05 months) than children in the first two classes (average = 65.45 months). Except for Alphabet Knowledge (−0.37), the literacy performance of children in

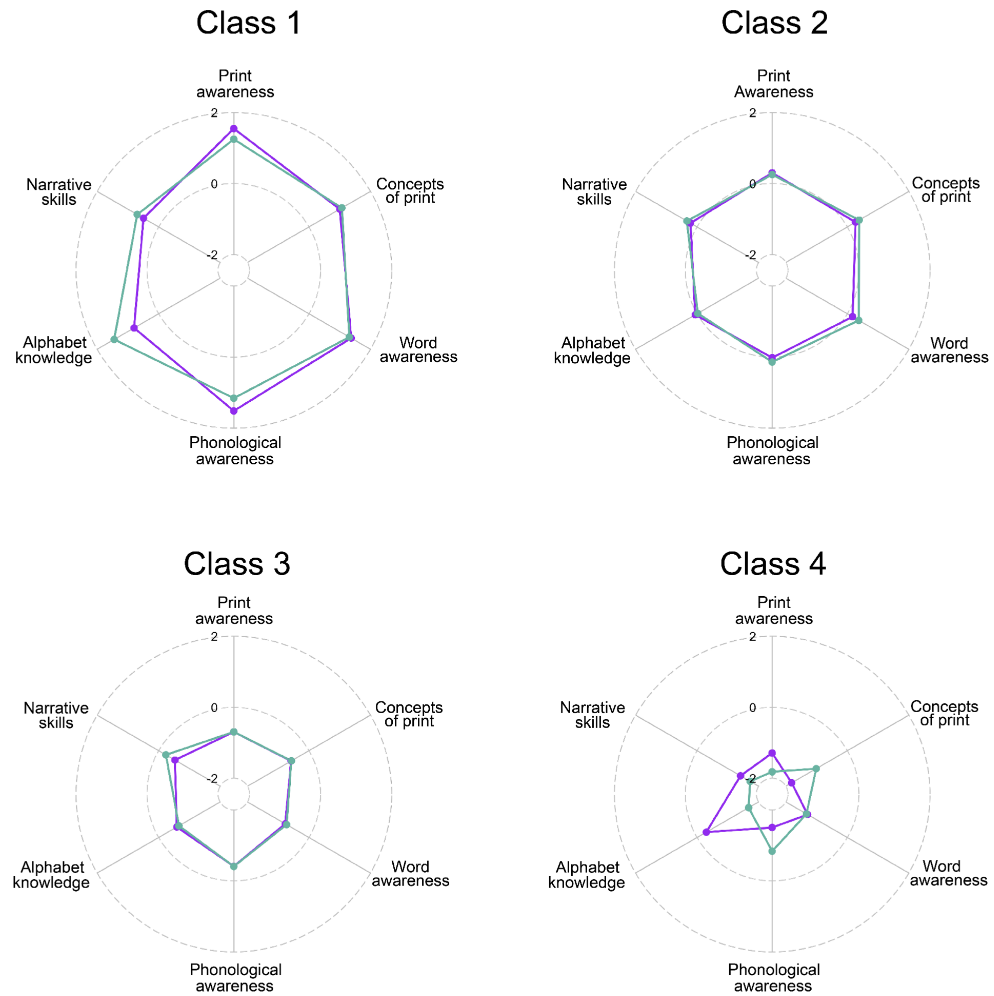

FIGURE 2
Latent profiles of kindergarten children as a function of whether they have a language impairment (purple) or not (green), presented as z-scores.

this class was below average (mean = −1.43). In comparison to the other latent classes, literacy performance was higher than Class 3 for Alphabet Knowledge (difference = 0.250) but lower than Class 3 for all other measures (mean difference = −0.892). It was lower than Classes 1 and 2 on all tested measures (mean difference = −2.43 and −1.49, respectively). Class 4 stood out with a notably greater percentage of language-impaired children when compared to other groups (94.74% vs. 33.24%). As can be seen in Figure 2, performance on the Word Awareness measure was consistent regardless of language impairment (mean difference = −0.036). However, variations between subgroups were evident in all other assessments. Non-language-impaired children exhibited stronger performance on the Concepts of Print and Phonological Awareness subscales (with mean differences of 0.80 and 0.67, respectively) but performed less well on the Print Awareness (−0.53), Alphabet Knowledge (−1.38), and Narrative Skills subscales (−0.32).

5 Discussion

Researchers recognize the period preceding formal school entry as a crucial stage for providing support to children at risk of future literacy challenges. In the context of primary and secondary prevention, it is imperative to proactively identify and address potential literacy difficulties before difficulties become intractable (Wanzek and Vaughn, 2007; Ozernov-Palchik and Gaab, 2016). Digital early literacy screening and assessment tools offer an engaging approach to identify at risk children, while also streamlining the assessment workflow, reducing administrative burdens on educators, and facilitating quicker grading and feedback dissemination. Hence, one of the aims of this current study was to evaluate the capacity of the EuLeApp© prototype to distinguish between different latent profiles among kindergarten children with the person-centered LCA approach. Additionally, given that children's early literacy proficiencies are influenced by various factors (Aaron et al., 2008; Cabell et al., 2010), a second aim was to investigate the impact of age, biological sex, and the presence of language impairment on these profiles (Table 3).

The LCA approach revealed four distinct classes of early literacy performance measured by the EuLeApp© prototype—Class 1: exceptional early literacy performers (prevalence = 17%), Class 2: typical early literacy performers (41%), Class 3: the marginal early literacy performers (35%), and Class 4: subpar early literacy performers (7%). As the name suggests, children in the *Exceptional*

Performers class excelled in all aspects of early literacy and were noteworthy for having fewer children with a language impairment but were older in age. The *Typical Performers* class exhibited average achievement on all variables, had older aged children, approximately equal proportions of boys and girls, and 37% of children classified with a language impairment. The *Marginal Performers* class exhibited below average literacy skills, younger aged children, and were balanced with respect to biological sex and language impairment. The *Subpar Performers* class exhibited well below average literacy skills, was composed of younger aged children, and was noteworthy for having a large proportion of boys and children classified as having a language impairment. Entropy was 0.86, indicating a high level of discrimination among the four latent classes.

The four distinct classes were identified based on their early literacy skill levels, yet they also exhibited unique profiles influenced by the interplay of input variables, including age, the presence of language impairment, and to a lesser degree, biological sex. Prior research studies from English-speaking samples have indicated that age is a predictor of early writing skills group membership (Cabell et al., 2010; McWayne et al., 2012; Guo et al., 2018). Our findings also indicate that age played a role in shaping the latent classes, with the two highest-performing early literacy groups primarily composed of older children, while the two lowest-performing classes consisted mainly of younger children. However, while age may account for some of the observed differences among the classes, the identification of more than two latent classes through LCA suggests that age is unlikely to serve as the sole determinant of the latent structure.

Interestingly, children with language impairment were present in all four groups. That said, it did appear to be a contributing factor to class differentiation. For instance, Class 1 (the *Exceptional Performers*) had only 10.64% language impaired children, whereas Class 4 (the *Subpar Performers*) had a significantly higher proportion (94.74%). This observation aligns with previous studies (Cabell et al., 2010; Justice et al., 2015), suggesting that while language abilities influence early literacy development, they do not solely determine class assignments. Regarding the impact of language impairment on early literacy, it was observed that performance was similar for most literacy measures for the *Exceptional Performers* class, except for Alphabet Knowledge. In contrast, the *Typical Performers* class exhibited uniform performance across the assessed literacy subscales, regardless of the presence of language impairment. The *Marginal Performers* class similarly demonstrated consistent performance in relation to the presence of language impairment, except for Narrative Skills, which exhibited lower early literacy scores among children with language impairment. The distinctions in early literacy related to language impairment categorization were most pronounced for the *Subpar Performers* group. Although Word Awareness performance was similar irrespective of language impairment, Concepts about Print and Phonological Awareness scores were notably lower for children classified as language impaired, but higher for Print Awareness, Alphabet Knowledge, and Narrative Skills scores.

The subsequent section will focus on delineating the distinctions in early literacy skills between language-impaired children and their non-impaired counterparts within each class individually. In class 1, the proportion of children with and without language impairments was unbalanced (10.64% of children with a language impairment). Therefore, the differences based on language impairment observed in the polar plots (Figure 2) should therefore be interpreted with caution. Non-language-impaired children outperformed language-impaired children in the Alphabet Knowledge domain. This might be explained by the moderating variable of environmental factors on language-and early literacy development. Studies have reported that kindergarten knowledge of the alphabet can be explained primarily by underlying environmental factors or the Home Literacy Environment (HLE) (Lemelin et al., 2007; Lehrl et al., 2012). For example, Lehrl et al. (2012) showed that Letter Knowledge can be predicted by the frequency of parental teaching of literacy (e.g., "Using simple words, I encourage my child to recognize or read letters and words"). In the present study, extensive questions on HLE were asked, which unfortunately due to the high rate of missing values, were not included in the analyses. However, it can be assumed that these influencing variables play an important role in the development of early literacy skills, and in particular Alphabet Knowledge (e.g., Cabell et al., 2011; Lehrl et al., 2012).

In Class 3, the presence of language impairment influenced only the Narrative Skills domain. In contrast, Narrative Skill performance was similar across language impairment for Classes 1 and 2. While speculative, this finding may be related to the scoring procedures used to measure performance, which focused on the macro-structure rather than the micro-structure of storytelling. For instance, we assessed the number of verbalized structural elements in the story (e.g., "Does the child mention the place, the people involved, or the attempted solution in the story?"). This emphasis on macro-structure assessment might have influenced the observed differences in narrative skills between language-impaired and non-impaired children. Interestingly, a recent review by Andreou and Lemoni (2020) indicate that language impaired children produce stories of varying strength, such that they either produce high-quality content that is grammatically weak or master the grammatical challenges but then produce short stories with inadequate content. It can be assumed that children with language difficulties in Class 3 may have different or greater weaknesses in vocabulary and/or grammar than language impaired children in Classes 1 and 2. These findings suggest that children in Class 3 may fall below a certain threshold that exceeds the requirements and cumulative load of creating a story (Duinmeijer et al., 2012)

Class 4 also shows differences due to language impairment. However, because there is only one non-language-impaired child in this class we exercise caution in making definitive statements about the specific effects on sub-tests. Nevertheless, this aspect underscores that non-language-impaired children can also be included in this profile. Moreover, all the aspects mentioned above can be used to explain different profile strengths and weaknesses. In addition to the causes and manifestations of language impairment, there are also the influencing factors that affect the development of early literacy skills and can sometimes lie in the child's proximal and distal environment (Hill, 2001; Lemelin et al., 2007; Aaron et al., 2008; Lehrl et al., 2012).

In summary, we posit that both the composition of children within each latent class and the differing early literacy profiles between children with and without language difficulties stem from the inherent heterogeneity within the group of children with language impairment. Conti-Ramsden et al. (1997), for example, identified six different subgroups within a sample of children with SLI, characterized by variations in impaired linguistic domains, presence of receptive and/or expressive difficulties, or different combinations of these parameters. In this context, the profiles and especially the polar plots (Figure 2) might have looked different with a different sample. Future research with larger and more diverse samples of children with various types of SLIs could provide further insights into the interplay between early literacy performance and language impairment.

Finally, in the context of this study, the biological sex distribution was highly skewed in favor of boys in Class 4, characterized by subpar early literacy performance. Previous research examining biological sex disparities in the development of reading skills has consistently noted a female advantage, evident as early as kindergarten (Chatterji, 2006) and becoming more pronounced over childhood and adolescence (Camarata and Woodcock, 2006). However, while our findings suggest that boys may exhibit poorer early literacy performance, this interpretation is nuanced by the observation that the *Exceptional Performers* class displayed a relatively balanced sex distribution. If biological sex exerted a significant influence on latent profiles, one might anticipate a higher proportion of females in this class. Moreover, the *Subpar Performers* class consisted of only 7% of the total sample, emphasizing the need for caution when drawing inferences related to the effects of sex on early literacy profiles. Therefore, while our results hint at potential sex-related differences, further research with larger and more diverse samples is warranted to explore this relationship more comprehensively.

5.1 Class-specific literacy recommendations

The LCA approach utilized in the present study allowed for the classification of children into latent classes with distinct attributes, offering the opportunity to formulate both general and class-specific literacy recommendations (Table 4). In accordance with the CELL Early Literacy Learning Model (Dunst et al., 2006), it is recommended that general literacy activities be incorporated into routines, ideally occurring daily or as frequently as possible. These activities should be varied, taking place in both planned and spontaneous contexts, and should employ adaptive and playful strategies to actively engage children and cultivate a sense of joy and enthusiasm for language learning.

Furthermore, learning must take place in the zone of proximal development (ZPD; Vygotsky, 1987), which in the context of early literacy can be conceptualized as the gap between what a child can achieve independently and what they can accomplish with appropriate support, whether from an adult or a more skilled peer. From an applied perspective, the ZPD serves as a framework that defines both the lower and upper limits within which instruction should be delivered and encompasses activities or tasks that are slightly beyond the child's current abilities. Designing reading and writing tasks that are within reach with minimal support (i.e., giving hints and cues to guide pupils' task responses) empowers children to recognize and appreciate their progressing literacy abilities, enhancing their innate drive to engage in reading and writing activities. As children advance in their literacy proficiency, they become proficient in completing tasks independently, leading to a shift in the complexity of literacy activities aligned with their ZPD.

As such, the ZPD, as well as the specific reading and writing activities used to promote intrinsic motivation and maximum literacy growth, will differ based on the attributes of each latent class (Table 4, column 2). To preface these recommendations, we provide a brief explanation of dialogic reading which is one possible approach to foster early literacy skill or enhance first experiences with a written text. Dialogic reading enhances interactive learning and comprehension through active dialog and engagement and involves a dynamic interaction between a reader and one or more participants, where the reader encourages discussion, asks questions, and prompts reflections on the text being read.

The *Exceptional Performers* (Class 1) and *Typical Performers* (Class 2) should be given age-appropriate reading and writing materials of higher difficulty and encouraged to participate in storytelling. In the context of dialogic reading, the children in these classes should be motivated to continue telling the stories on their own or to present them in a theater format. In addition, individual letters can be discovered and named. In contrast, children from Class 3 (*Marginal Performers*) and class 4 (*Subpar Performers*) require other situational adaptations (Campbell et al., 2008) that enable them to learn in the ZPD. They would probably benefit more from picture books (or books with only a little text) fostering visual literacy as a precursor to literacy, dialogic "reading" of wimmelbooks or hidden picture books with humans, animals, and objects to discover and to talk about (e.g., "My Big Wimmelbook," "Antonia war schon mal da" [Antonia has been there before], "Die Torte ist weg" [the cake is gone]). The emphasis here is not on independent storytelling or letter recognition, but on comprehending language and learning book handling skills. In this context, visual media primarily serves as a linguistic tool for communication, idea exchange, and navigating our complex world. Therefore, books could be used to explore writing in comparison to images, practice rhyming, or enhance receptive and productive vocabulary skills. Consequently, one could pose questions about the pictures and subsequently search for answers within the text or pictures.

5.2 Limitations

Despite the implications of these findings, it is important to acknowledge several limitations of our work. First, our research was confined to a sample of 15 daycare facilities in two German states, deliberately excluding children with limited German language proficiency. Additionally, unlike previous studies that focused on children with language disorders and/or low SES background (Cabell et al., 2010, 2011, 2013; Justice et al., 2015), we aimed for our sample to reflect the heterogeneity present in typical German kindergartens. Consequently, the age range of our sample was broad, and included children with normal language acquisition and educated households. Consequently, it's important to recognize that the results may not be applicable to all children of this age group, those from different geographic regions, or children who exhibit language impairment or varied SES.

Second, we acknowledge that the educational systems of different German states or countries could play a significant role in determining the early literacy profiles. It is important to note that while LCA is a methodologically rigorous and data-driven approach, the derived classes are influenced by the demographic composition

TABLE 4 Proposed literacy recommendations during dialogic book reading for the identified latent classes.

Latent classes	Intervention suggestions	Implementations in dialogic book reading
Exceptional performers	• Cultivate and maintain enthusiasm for books and reading • Nurture and sustain enjoyment and motivation for storytelling • Engage in daily higher difficulty level reading and writing activities that are playful and progressively challenging	• **Phonological Awareness:** Employ dialogic reading to enhance phonological awareness, focusing on tasks such as analyzing sounds within words, manipulating sounds within words, and synthesizing sounds to form words • **Alphabet Knowledge:** Engage children in identifying both familiar and unfamiliar letters within the text, encouraging them to recognize the visual and auditory characteristics of individual letters ("Which letters have a curved shape?") • **Narrative Skills:** Encourage children to create their own stories or contribute to completing existing narratives
Typical performers	• Maintain enthusiasm for books and reading over time • Cultivate and maintain joy and motivation for storytelling • Integrate daily playful language and communication activities to promote language development • Integrate playful daily reading and writing activities at a moderate difficulty level	• **Word Awareness:** Help children recognize differences between short and long words, and identify identical words consistently • **Phonological Awareness:** Implement dialogic reading techniques to enhance phonological awareness, particularly focusing on recognizing initial sounds • **Alphabet Knowledge:** Introduce children to the first letters they learned, including those of their own name, to reinforce alphabet familiarity • **Narrative Skills:** Encourage children to retell stories or speculate on potential story continuations ("How do you think the story might continue?")
Marginal performers	• Cultivate joy and motivation for storytelling • Integrate playful language and communication activities to enhance language development • Include daily playful activities aimed at promoting word awareness, print awareness, and alphabet knowledge at levels ranging from low to medium difficulty	• **Word Awareness:** Explore words within the book, distinguishing between long and short words and discussing how word length relates to meaning (e.g., comparing "honeybee" to "bear") • **Phonological Awareness:** Engage in dialogic reading to develop phonological awareness broadly, including activities such as identifying rhyming words ("What rhymes with mouse?") • **Alphabet Knowledge:** Connect letters to corresponding images and help the child identify the initial letter of their name with the assistance of the teacher or caregiver • **Narrative Skills:** Talk about the story structure (e.g., sequencing story sections) and practice shifting perspectives; prompt discussions by asking questions about the story
Subpar performers	• Implement daily playful language and communication activities to nurture language development • Beyond understanding meaning, also focus on drawing the child's attention to the form and structure of language • Integrate playful daily activities to promote alphabet knowledge at a low level of difficulty • Integrate playful daily activities to enhance print awareness, concepts of print, word awareness, and narrative skills at a low level of difficulty	• Utilize picture books and promote visual literacy • Cultivate a sense of enjoyment and familiarity with reading • **Print Awareness:** Collaboratively search for and identify words and text within a book alongside images • **Concepts about Print**: Introduce writing conventions together, prompting questions such as "Where is the title of the book?" and "Where do we begin reading?" • **Phonological Awareness:** Employ dialogic reading to enhance phonological awareness more broadly, incorporating activities like syllable clapping and identifying rhymes • **Alphabet Knowledge:** Connect letters to corresponding images and assist the child in identifying the initial letter of their name with the support of the teacher or caregiver • **Narrative Skills:** Stimulate discussions about the story or the book cover, asking questions such as "What do you think the story is about?" and "Where do you think the story takes place?"

of the sample. The obtained profiles reflect the abilities of the specific sample and must be interpreted as such. To establish the robustness of our findings, replication with independent sample sharing is necessary.

Third, scholars unanimously agree that the acquisition of emergent literacy skills involves a complex interplay of fundamental literacy elements (Mason and Stewart, 1990; Whitehurst and Lonigan, 1998; Sénéchal et al., 2001), in which each component follows its own development trajectory, supports and influences the development of the other components, and is shaped by the dynamic interplay of sociocultural, environmental, and demographic factors (Rohde, 2015). In this study, we focused on how factors such as age, biological sex, and language impairment influence the variety and occurrence of these latent literacy profiles. Cabell et al. (2010, 2011) mentioned three possible reasons that explain the identified patterns, all of which emphasize the importance of children's literacy learning environment to their growth in literacy. Although we did query the Home Literacy Environment (HLE, Niklas et al., 2020) using a paper-based questionnaire, there was insufficient participation for those particular questions and thus we did not include it in the analysis.

Fourth, it's important to note that the classification accuracy for Class 4 (*Subpar Literacy Performers*), as indicated by the average posterior probabilities, did not meet the threshold for robust classification. Therefore, interpretations regarding this smaller class should be approached cautiously, pending replication and additional validation. It is plausible that the observed patterns within this group may be specific to the sample under study and may not generalize to broader populations.

5.3 Future directions

In the second phase of the project, we will evaluate the effectiveness of the EuLeApp© assessment tool and the adaptive, playful everyday intervention in promoting early literacy skills. Following the baseline assessment, the EuLeApp© will provide an individualized profile outlining each child's strengths and weaknesses across various early literacy components. Subsequently, pedagogical specialists will have the option to manually select exercises targeting specific early literacy components from a collection of 95 games and 28 impulse cards contained within a wooden box. These games, designed for group settings, are categorized into three levels of difficulty to accommodate each child's ZPD. Additionally, the games are seamlessly integrated into the children's daily schedule, with labels indicating suitable times of the day (e.g., morning circle, naptime, mealtime, etc.). At the conclusion of the four-month intervention period, we will employ latent transition analysis (Nylund-Gibson et al., 2023) to study the movement of children between groups across several time points. This will allow us to understand the dynamics and patterns of stability and change after the intervention implementation, especially as it pertains to specific literacy skills (e.g., Print Awareness, Alphabet Knowledge, Narrative Skills), latent class, or language impairment.

In addition, we have refined our data collection protocol to enhance compliance with the completion of HLE questions by using electronic survey software that forces respondents to answer each question before survey progression and tying a portion of the remuneration to survey completion. An additional methodological enhancement slated for the next project phase is the inclusion of a standardized non-verbal intelligence test to assess the cognitive abilities of children [i.e., the Snijders-Oomen Nonverbal Intelligence Test 2.5–7 Revised (Tellegen, 1996)].

Furthermore, research should prioritize examining Germany's most vulnerable populations, such as first-generation migrants and refugees, many of whom enter Germany without prior knowledge of the language. We believe that the EuLeApp© could serve as a valuable tool by which early literacy skills and cultural knowledge can be imparted to kindergarten-aged children. This, in turn, can support the cultivation of social connections and active participation in their new host communities, contributing to their success and integration in society.

6 Conclusion

LCA and the EuLeApp© prototype effectively distinguished between different latent early literacy profiles in German kindergarten children, reflecting the influence of factors such as age, biological sex, and language impairment. These findings provide valuable insights for future research endeavors and underscore the importance of supporting early childhood professionals in adapting their daily practices to meet the unique needs of young learners in the realm of early literacy.

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 humans were approved by Research Ethics Committee of the Carl von Ossietzky University of Oldenburg. The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation in this study was provided by the participants' legal guardians/next of kin.

Author contributions

CS: Investigation, Supervision, Writing – original draft, Writing – review & editing. MY: Methodology, Writing – original draft, Writing – review & editing. MM: Conceptualization, Funding acquisition, Project administration, Writing – review & editing. TJ: Conceptualization, Funding acquisition, Project administration, Writing – original draft, Writing – review & editing. CH: Data curation, Formal analysis, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This work was supported by the German Federal Ministry of Education and Research (BMBF) [grant number 01NV2105A].

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

Aaron, P. G., Joshi, R. M., Gooden, R., and Bentum, K. E. (2008). Diagnosis and treatment of reading disabilities based on the component model of reading. An alternative to the discrepancy model of LD. *J. Learn. Disabil.* 41, 67–84. doi: 10.1177/0022219407310838

Andreou, G., and Lemoni, G. (2020). Narrative skills of monolingual and bilingual pre-school and primary school children with developmental language disorder (DLD): a systematic review. *Open J. Mod. Linguist.* 10, 429–458. doi: 10.4236/ojml.2020.105026

Antoniou, F., Ralli, A. M., Mouzaki, A., Diamanti, V., and Papaioannou, S. (2022). Logometro®: the psychometric properties of a norm-referenced digital battery for language assessment of Greek-speaking 4–7 years old children. *Front. Psychol.* 13:900600. doi: 10.3389/fpsyg.2022.900600

Babayiğit, S., Roulstone, S., and Wren, Y. (2021). Linguistic comprehension and narrative skills predict reading ability: a 9-year longitudinal study. *Br. J. Educ. Psychol.* 91, 148–168. doi: 10.1111/bjep.12353

Bergman, L. R., and Magnusson, D. (1997). A person-oriented approach in research on developmental psychopathology. *Dev. Psychopathol.* 9, 291–319. doi: 10.1017/S095457949700206X

Bergman, L. R., and Trost, K. (2006). The person-oriented versus the variable-oriented approach: are they complementary, opposites, or exploring different worlds? *Merrill-Palmer Q.* 52, 601–632. doi: 10.1353/mpq.2006.0023

Bock-Famulla, K., Strunz, E., and Löhle, A. (2017). Länderreport Frühkindliche Bildungssysteme 2017. Available at: https://www.bertelsmannstiftung.de/fileadmin/files/BSt/Publikationen/imported/leseprobe/LP_978-3-86793-786-3_1.pdf (Accessed November 13, 2023).

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): the general theory and its analytical extensions. *Psychometrika* 52:345. doi: 10.1007/BF02294361

Brügelmann, H. (1998). "Kinder lernen lesen und schreiben [Children Learn Reading and Writing]" in *Kinder lernen anders: vor der Schule – in der Schule [Children Learn Differently: Preschool – in School]*. ed. H. Brügelmann (Lengwil: Libelle), 41–57.

Buddeberg, K., Dutz, G., Heilmann, L., Stammer, C., and Grotlüschen, A. (2021). Participation and Independence with low literacy: selected findings of the LEO 2018 survey on low literacy in Germany. *Adult Liter. Educ.* 3, 19–34. doi: 10.35847/KBuddeberg.GDutz.CStammer.AGrotluschen.3.3.19

Bundesministerium der Justiz. (1990). Sozialgesetzbuch (SGB) - Achtes Buch (VIII) - Kinder-und Jugendhilfe - (Artikel 1 des Gesetzes v. 26. Juni 1990, BGBl. I S. 1163). Available at: https://www.gesetze-im-internet.de/sgb_8/index.html#BJNR111630990BJNE004708377.

Cabell, S. Q., Justice, L. M., Konold, T. R., and McGinty, A. S. (2011). Profiles of emergent literacy skills among preschool children who are at risk for academic difficulties. *Early Child. Res. Q.* 26, 1–14. doi: 10.1016/j.ecresq.2010.05.003

Cabell, S. Q., Justice, L. M., Logan, J. A. R., and Konold, T. R. (2013). Emergent literacy profiles among prekindergarten children from low-SES backgrounds: longitudinal considerations. *Early Child. Res. Q.* 28, 608–620. doi: 10.1016/j.ecresq.2013.03.007

Cabell, S. Q., Lomax, R. G., Justice, L. M., Breit-Smith, A., Skibbe, L. E., and McGinty, A. S. (2010). Emergent literacy profiles of preschool-age children with specific language impairment. *Int. J. Speech Lang. Pathol.* 12, 472–482. doi: 10.3109/17549507.2011.492874

Camarata, S., and Woodcock, R. (2006). Sex differences in processing speed: developmental effects in males and females. *Intelligence* 34, 231–252. doi: 10.1016/j.intell.2005.12.001

Campbell, P. H., Milbourne, S. A., and Wilcox, M. J. (2008). Adaptation interventions to promote participation in natural settings. *Infants Young Child.* 21, 94–106. doi: 10.1097/01.IYC.0000314481.16464.75

Celeux, G., and Soromenho, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. *J. Classif.* 13, 195–212. doi: 10.1007/BF01246098

Chatterji, M. (2006). Reading achievement gaps, correlates, and moderators of early reading achievement: evidence from the early childhood longitudinal study (ECLS) kindergarten to first grade sample. *J. Educ. Psychol.* 98, 489–507. doi: 10.1037/0022-0663.98.3.489

Clay, M. M. (1966). Emergent reading behavior. Doctoral dissertation, University of Auckland, New Zealand.

Collins, L. M., and Lanza, S. T. (2010). *Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences*. New York: John Wiley & Sons.

Conti-Ramsden, G., Crutchley, A., and Botting, N. (1997). The extent to which psychometric tests differentiate subgroups of children with SLI. *J. Speech Lang. Hear. Res.* 40, 765–777. doi: 10.1044/jslhr.4004.765

Deutscher Bildungsserver (2023). Bildungspläne der Bundesländer für die frühe Bildung in Kindertageseinrichtungen. Available at: https://www.bildungsserver.de/Bildungsplaene-fuer-Kitas-2027-de.html (Accessed November 13, 2023).

Duinmeijer, I., de Jong, J., and Scheper, A. (2012). Narrative abilities, memory and attention in children with a specific language impairment. *Int. J. Lang. Commun. Disord.* 47, 542–555. doi: 10.1111/j.1460-6984.2012.00164.x

Dunst, C. J., Trivette, C. M., Masiello, T., Roper, N., and Robyak, A. (2006). Framework for developing evidence-based early literacy learning practices. CELL Papers 1, 1–12.

European Commission. (2023). *Germany: Early childhood education and care*. Available at: https://eurydice.eacea.ec.europa.eu/national-education-systems/germany/educational-guidelines.

Federal Ministry for Economic Affairs and Energy (2016). *Digital strategy 2025* Bonifatius GmbH, Federal Ministry for Economic Affairs and Energy (BMWi) Available at: https://www.de.digital/DIGITAL/Redaktion/EN/Publikation/digital-strategy-2025.pdf?__blob=publicationFile&v=9.

Ferrer, E., Shaywitz, B. A., Holahan, J. M., Marchione, K. E., Michaels, R., and Shaywitz, S. E. (2015). Achievement gap in reading is present as early as first grade and persists through adolescence. *J. Pediatr.* 167, 1121–1125.e2. doi: 10.1016/j.jpeds.2015.07.045

Fricke, S., Szczerbinski, M., Fox-Boyer, A., and Stackhouse, J. (2016). Preschool predictors of early literacy acquisition in German-speaking children. *Read. Res. Q.* 51, 29–53. doi: 10.1002/rrq.116

Gaab, N., and Petscher, Y. (2022). "Screening for early literacy milestones and reading disabilities: the why, when, whom, how, and where" in *Perspectives on Language and Literacy*, vol. 48, 11–18.

Good, R. H., and Kaminski, R. A. (2002). *DIBELS oral reading fluency passages for first through third grades*. Eugene, OR: University of Oregon.

Gorman, B. K., Bingham, G. E., Fiestas, C. E., and Terry, N. P. (2016). Assessing the narrative abilities of Spanish-speaking preschool children: a Spanish adaptation of the narrative assessment protocol. *Early Child. Res. Q.* 36, 307–317. doi: 10.1016/j.ecresq.2015.12.025

Guo, Y., Sun, S., Puranik, C., and Breit-Smith, A. (2018). Profiles of emergent writing skills among preschool children. *Child Youth Care Forum* 47, 421–442. doi: 10.1007/s10566-018-9438-1

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., et al. (2020). Array programming with NumPy. *Nature* 585, 357–362. doi: 10.1038/s41586-020-2649-2

Hill, N. E. (2001). Parenting and academic socialization as they relate to school readiness: the roles of ethnicity and family income. *J. Educ. Psychol.* 93, 686–697. doi: 10.1037/0022-0663.93.4.686

Hipkins, R., and Cameron, M. (2018). *Trends in assessment: An overview of themes in the literature.New Zealand Council for Educational Research*. Available at: https://www.nzcer.org.nz/system/files/Trends%20in%20assessment%20report.pdf.

Hjetland, H. N., Lervåg, A., Lyster, S. A. H., Hagtvet, B. E., Hulme, C., and Melby-Lervåg, M. (2019). Pathways to reading comprehension: a longitudinal study from 4 to 9 years of age. *J. Educ. Psychol.* 111, 751–763. doi: 10.1037/edu0000321

Hudson, A., Bailet, L. L., Piasta, S. B., Logan, J. A. R., Lewis, K., and Zettler-Greeley, C. M. (2023). Latent profile moderation: examining the differential impact of a small-group emergent literacy intervention. *JESPAR*, 1–33. doi: 10.1080/10824669.2023.226418

Invernizzi, M., Meier, J., and Swank, L. (2004). Phonological awareness literacy screening. PreK. Charlottesville, VA: University of Virginia.

Justice, L., Logan, J., Kaderavek, J., Schmitt, M. B., Tompkins, V., and Bartlett, C. (2015). Empirically based profiles of the early literacy skills of children with language impairment in early childhood special education. *J. Learn. Disabil.* 48, 482–494. doi: 10.1177/0022219413510179

Lazarsfeld, P. F., and Henry, N. W. (1968). *Latent structure analysis*. Boston: Houghton Mifflin.

Lehrl, S., Ebert, S., Roßbach, H.-G., and Weinert, S. (2012). Die Bedeutung der familiären Lernumwelt für Vorläufer schriftsprachlicher Kompetenzen im Vorschulalter. *Zeitschrift für Familienforschung* 24, 115–133.

Lemelin, J.-P., Boivon, M., Forget-Dubois, N., Dionne, G., Séguin, J. R., Brendgen, M., et al. (2007). The genetic-environmental etiology of cognitive school readiness and later academic achievement in early childhood. *Child Dev.* 78, 1855–1869. doi: 10.1111/j.1467-8624.2007.01103.x

Lenel, A., and Knopf, M. (2015). The development of a concept of word in kindergarten and its impact on Reading achievement in school. A longitudinal study. *Zeitschrift für Literaturwissenschaft und Linguistik* 45, 43–70. doi: 10.1007/BF03379723

Lo, Y., Mendell, N. R., and Rubin, D. B. (2001). Testing the number of components in a normal mixture. *Biometrika* 88, 767–778. doi: 10.1093/biomet/88.3.767

Lonigan, C. J., Burgess, S. R., and Anthony, J. L. (2000). Development of emergent literacy and early reading skills in preschool children: evidence from a latent-variable longitudinal study. *Dev. Psychol.* 36, 596–613. doi: 10.1037/0012-1649.36.5.596

Lonigan, C. J., Wagner, R. K., Torgesen, J. K., and Rashotte, C. A. (2007). *TOPEL: Test of preschool early literacy*. Austin, TX: Pro-Ed.

Lonigan, C. J., and Shanahan, T. (2010). Developing early literacy skills: Things we know we know and things we know we don't know. *Educational Researcher*. 39, 340–346.

Mann, V., and Wimmer, H. (2002). Phoneme awareness and pathways into literacy: a comparison of German and American children. *Read. Writ.* 15, 653–682. doi: 10.1023/A:1020984704781

Mason, J. M., and Stewart, J. P. (1990). "Emergent literacy assessment for instructional use in kindergarten" in *Assessment for instruction in early literacy*. eds. L. M. Morrow and J. K. Smith (Englewood Cliffs, NJ: Prentice Hall), 155–175.

Masyn, K. E. (2013). "Latent class analysis and finite mixture modelling" in *The Oxford handbook of quantitative methods: Statistical analysis*. ed. T. D. Little (Oxford: Oxford University Press), 551–611.

McKinney, W. (2010). Data structures for statistical computing in python. In: *Proceedings of the 9th Python in science conference* 445, 51–56.

McWayne, C. M., Hahs-Vaughn, D. L., Cheung, K., and Wright, L. E. G. (2012). National profiles of school readiness skills for head start children: an investigation of stability and change. *Early Child. Res. Q.* 27, 668–683. doi: 10.1016/j.ecresq.2011.10.002

Meindl, M., and Jungmann, T. (2014). Erfassung der frühen Erzähl-und Lesekompetenzen im Vorschulalter zur primären Prävention von Schwierigkeiten im Schriftspracherwerb. *Empirische Sonderpädagogik* 6, 211–226. doi: 10.25656/01:9931

Meindl, M., and Jungmann, T. (2015). Kapitel 8 Erfassung der Erzähl-und Lesekompetenzen bei 4-bis 5-jährigen Kindern: EuLe 4-5. *Diagnostik im Vorschulalter* 13:129.

Meindl, M., Stuhr, C., Nicosia, R., and Jungmann, T. (2022). "Adaptive Diagnostik und Förderung der frühen Erzähl-und Lesekompetenzen im Elementar-und Primarbereich mit der EuLeApp©" in *Sprachentwicklung im Dialog – Digitalität – Kommunikation – Partizipation*. eds. M. Spreer, M. Wahl and H. Beek (Idstein: Schulz-Kirchner), 89–95.

Morgan, P. L., Fuchs, D., Compton, D. L., Cordray, D. S., and Fuchs, L. S. (2008). Does early reading failure decrease children's reading motivation? *J. Learn. Disabil.* 41, 387–404. doi: 10.1177/00222194083211

Morin, S., Legault, R., Bakk, Z., Giguère, C. É., de la Sablonnière, R., and Lacourse, É. (2023). StepMix: a Python package for Pseudo-likelihood estimation of generalized mixture models with external variables. *arXiv preprint arXiv:2304.03853*. doi: 10.48550/arXiv.2304.03853

Murphy, K. A., Justice, L. M., O'Connell, A. A., Pentimonti, J. M., and Kaderavek, J. N. (2016). Understanding risk for reading difficulties in children with language impairment. *J. Speech Lang. Hear. Res.* 59, 1436–1447. doi: 10.1044/2016_JSLHR-L-15-0110

Nagin, D. (2005). *Group-based modeling of development*. Cambridge, MA: Harvard University Press.

Niklas, F., and Schneider, W. (2013). Home literacy environment and the beginning of reading and spelling. *Contemp. Educ. Psychol.* 38, 40–50. doi: 10.1016/j.cedpsych.2012.10.001

Niklas, F., Annac, E., and Wirth, A. (2020). App-based learning for kindergarten children at home (Learning4Kids): study protocol for cohort 1 and the kindergarten assessments. *BMC Pediatr.* 20:554. doi: 10.1186/s12887-020-02432-y

Nylund-Gibson, K., and Choi, A. Y. (2018). Ten frequently asked questions about latent class analysis. *Transl. Issues Psychol. Sci.* 4, 440–461. doi: 10.1037/tps0000176

Nylund-Gibson, K., Garber, A. C., Carter, D. B., Chan, M., Arch, D. A. N., Simon, O., et al. (2023). Ten frequently asked questions about latent transition analysis. *Psychol. Methods* 28, 284–300. doi: 10.1037/met0000486

Ozernov-Palchik, O., and Gaab, N. (2016). "Tackling the early identification of dyslexia with the help of neuroimaging" in *Perspectives on Language and Literacy*, vol. 7, 156–176.

Pavelko, S. L., Lieberman, R. J., Schwartz, J., and Hahs-Vaughn, D. (2018). The contributions of phonological awareness, alphabet knowledge, and letter writing to name writing in children with specific language impairment and typically developing children. *Am. J. Speech Lang. Pathol.* 27, 166–180. doi: 10.1044/2017_AJSLP-17-0084

Petermann, F. (2018). *SET 5–10. Sprachstandserhebungstest für Kinder im Alter zwischen 5 und 10 Jahren*. Göttingen: Hogrefe.

Petermann, F., Rißling, J.-K., and Metzer, J. (2016). *SET 3–5. Sprachstandserhebungstest für Kinder im Alter zwischen 3 und 5 Jahren*. Göttingen: Hogrefe.

Python Core Team (2019). *Python: A dynamic, open source programming language*. Wilmington, DE: Python Software Foundation.

Ram, N., and Grimm, K. J. (2009). Growth mixture modelling: a method for identifying differences in longitudinal change among unobserved groups. *Int. J. Behav. Dev.* 33, 565–576. doi: 10.1177/0165025409343765

Rohde, L. (2015). The comprehensive emergent literacy model: early literacy in context. *SAGE Open* 5:215824401557766. doi: 10.1177/2158244015577664

Resa, E., Ereky-Stevens, K., Wieduwilt, N., Penderi, E., Anders, Y., Petrogiannis, K., et al. (2016). Overview of quality monitoring systems and results of moderator analysis. *Curriculum Quality Analysis and Impact Review of European Early Childhood Education and Care (ECEC), (D4. 3).*

Rosenberg, J. M., Beymer, P. N., Anderson, D. J., Van Lissa, C. J., and Schmidt, J. A. (2018). tidyLPA: an R package to easily carry out latent profile analysis (LPA) using open-source or commercial software. *J. Open Source Softw.* 3:978. doi: 10.21105/joss.00978

Roth, F. P., Speece, D. L., and Cooper, D. H. (2002). A longitudinal analysis of the connection between oral language and early reading. *J. Educ. Res.* 95, 259–272. doi: 10.1080/00220670209596600

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Stat.* 6, 461–464. doi: 10.1214/aos/1176344136

Sclove, S. L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika* 52, 333–343. doi: 10.1007/BF02294360

Sénéchal, M., LeFevre, J. A., Smith-Chant, B. L., and Colton, K. V. (2001). On refining theoretical models of emergent literacy the role of empirical evidence. *J. Sch. Psychol.* 39, 439–460. doi: 10.1016/S0022-4405(01)00081-4

Sewasew, D., and Koester, L. S. (2019). The development dynamics of students´ reading self-concept and reading competence: examining reciprocal relations and ethnic-background patterns. *Learn. Individ. Differ.* 73, 102–111. doi: 10.1016/j.lindif.2019.05.010

Snowling, M. J., Duff, F. J., Nash, H. M., and Hulme, C. (2016). Language profiles and literacy outcomes of children with resolving, emerging, or persisting language impairments. *J. Child Psychol. Psychiatry* 57, 1360–1369. doi: 10.1111/jcpp.12497

Statistisches Bundesamt. (2018). *Statistiken der Kinder-und Jugendhilfe*. Available at: https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Soziales/Kindertagesbetreuung/Publikationen/Downloads-Kindertagesbetreuung/tageseinrichtungen-kindertagespflege-5225402187004.pdf?__blob=publicationFile&v=6.

Tellegen, P. J. (1996). Snijders-Omen Non verbaler Intelligenztest. *SON-R* 2:1/2-7.

Tietze, W., Meischner, T., Gänsfuß, R., Grenner, K., Schuster, K. M., Völkel, P., et al. (1998). *Wie gut sind unsere Kindergärten? Eine Untersuchung zur pädagogischen Qualität in deutschen Kindergärten*. Neuwied, Kriftel, Berlin: Beltz.

Viernickel, S., Fuchs-Rechlin, K., Bensel, J., Strehmel, P., Preissing, C., and Haug-Schnabel, G. (2015). *Qualität für alle: Wissenschaftlich begründete Standards für die Kindertagesbetreuung*. Freiburg, Germany: Herder Verlag GmbH.

Vygotsky, L. S. (1987). *Mind in Society: The Development of Higher Psychological Processes*. (Harvard University Press, Cambridge, MA).

Walgermo, B. R., Foldnes, N., Uppstad, P. H., and Solheim, O. J. (2018a). Developmental dynamics of early reading skill, literacy interest and readers' self-concept within the first year of formal schooling. *Read. Writ.* 31, 1379–1399. doi: 10.1007/s11145-018-9843-8

Walgermo, B. R., Frijters, J. C., and Solheim, O. J. (2018b). Literacy interest and reader self-concept when formal reading instruction begins. *Early Child. Res. Q.* 44, 90–100. doi: 10.1016/j.ecresq.2018.03.002

Wanzek, J., and Vaughn, S. (2007). Research-based implications from extensive early reading interventions. *Sch. Psychol. Rev.* 36, 541–561. doi: 10.1080/02796015.2007.12087917

Whitehurst, G. J., and Lonigan, C. J. (1998). Child development and emergent literacy. *Child Dev.* 69, 848–872. doi: 10.1111/j.1467-8624.1998.tb06247.x

Wirts, C., Egert, F., and Reber, K. (2017). Early literacy in deutschen Kindertageseinrichtungen. Eine analyse der Häufigkeit von literacy-Aktivitäten im Kita-Alltag [early literacy in German early child education and care: analysis on the frequency of early literacy activities during a day]. *Forschung Sprache* 2, 96–106.

Wittig, J., and Schneider, R. (2022). "Kompetenzstufenbesetzung im Fach Deutsch" in *IQB Bildungstrend 2021. Kompetenzen in den Fächern Deutsch und Mathematik am Ende der 4. Jahrgangsstufe im dritten Ländervergleich*. eds. P. Stanat, S. Schipolowski, R. Schneider, K. A. Sachse, W. Weirich and S. Henschel (Münster, New York: Waxmann), 41–65.

World Literacy Foundation. (2018). The economic and social cost of illiteracy. A white paper by the world literacy foundation. World Literacy Summit, March 25th-27th. Oxford.

frontiers | Frontiers in Education

TYPE Original Research
PUBLISHED 26 April 2024
DOI 10.3389/feduc.2024.1335957

OPEN ACCESS

EDITED BY
Christoph Weber,
University of Education Upper Austria, Austria

REVIEWED BY
Steffen Zitzmann,
University of Tübingen, Germany
Shelley Shaul,
University of Haifa, Israel

*CORRESPONDENCE
Chiara Banfi
✉ chiara.banfi@uni-graz.at

RECEIVED 14 November 2023
ACCEPTED 16 April 2024
PUBLISHED 26 April 2024

CITATION
Banfi C, Jöbstl V, Göbel SM and Landerl K (2024) Longitudinal predictors of reading and arithmetic at different attainment levels.
Front. Educ. 9:1335957.
doi: 10.3389/feduc.2024.1335957

Longitudinal predictors of reading and arithmetic at different attainment levels

Chiara Banfi[1]*, Viktoria Jöbstl[1], Silke M. Göbel[2,3] and Karin Landerl[1,4]

[1]Department of Psychology, University of Graz, Graz, Austria, [2]Department of Psychology, University of York, York, United Kingdom, [3]Department of Special Needs Education, University of Oslo, Oslo, Norway, [4]BioTechMed-Graz, Graz, Austria

Reading and arithmetic are distinct academic skills that share similarities in skill acquisition and use. Previous research investigated the cognitive basis of associations and dissociations between reading and arithmetic by using either subtyping or dimensional approaches. In the current study, we aim to bridge the gap between these two approaches by investigating common and distinct predictors of reading and arithmetic at different performance levels with quantile regression models. This allowed us to look more closely at the lower tail of the ability distributions, and to test whether predictions for children with low reading and arithmetic fluency differed from the typical performance range. We analyzed longitudinal data of 357 children speaking English or German. Outcome variables were reading and arithmetic fluency assessed at the end of Grade 1, 2, and 3. Predictors were assessed in Grade 1. Results confirmed nonverbal IQ and working memory as domain-general predictors of reading and arithmetic. The association of reading and arithmetic was mainly explained by nonverbal IQ, phonological awareness, RAN and multi-digit transcoding. Across grades and performance levels, phonological awareness and RAN made a specific contribution to reading. Magnitude processing and multi-digit transcoding were specific predictors of arithmetic. Counting also made a specific prediction to arithmetic in Grade 3, but only in the low performance range. Our findings indicate partly distinct underlying cognitive mechanisms for reading and arithmetic. Shared predictors are involved in retrieval efficiency, language processing and cross-format integration. These results have important implications, as they suggest that most predictors are equally relevant for children with low, typical or even excellent reading and arithmetic fluency.

KEYWORDS

arithmetic, development, longitudinal, quantile regression, reading

Introduction

Reading and arithmetic are core academic skills that provide fundamental tools for everyday-life and learning more complex abilities. They may seem like different skill sets. But there are many similarities. To start with, reading and arithmetic are both symbol-based systems in which visual stimuli (letters, Arabic digits) are put in relation with sounds (phonemes, number words; Dehaene, 1992; Coltheart et al., 2001; Barrouillet et al., 2004). In the early stages of skill acquisition, both abilities require reliance on algorithms to obtain a consistent output: For reading, this refers to grapheme-phoneme conversion, which enables reading by serial decoding strategies (Frith, 1985; Coltheart et al., 2001). In the field of arithmetic, young children are able to compute simple additions and subtractions by serial counting (Peters and De Smedt, 2018). Repeated decoding of the same word or computing of the same multiplication enables the build-up of memory traces, i.e., associations between orthographic and phonological representations or between arithmetic problems and solutions. These facts can be retrieved quickly and effortlessly, enabling efficient use of the learnt skills as in lexico-orthographic reading and arithmetic fact retrieval.

Skillful use of reading and arithmetic abilities requires the efficient integration of visual, verbal and semantic representations. There is a parallelism between the cognitive structure of models for reading and number processing, as recently highlighted by Jöbstl et al. (2023) and depicted in Figure 1. The Lexical Quality Hypothesis (Perfetti and Hart, 2001) postulates that interindividual differences in reading are related to the degree in which the three main constituents of a word are integrated, namely orthography (word form), phonology (word sound) and semantics (word meaning). Fluent, skilled reading requires well specified and connected orthographic, phonological and semantic representations of the same word. The Triple Code Model for number processing (Dehaene, 1992) builds on a similar structure. The basic tenet is that numbers are represented by three codes: The visual Arabic code (for Arabic digits; e.g., 3), the verbal word frame (number words; e.g., "three") and the analog magnitude representation (quantities; e.g., ■■■). These three codes are connected and get activated during numerical and arithmetic computations.

Empirical research reported associations of varying strength between reading and arithmetic, with correlation coefficients between 0.14 and 0.77 (Dirks et al., 2008; Landerl and Moll, 2010; Moll et al., 2014; Peterson et al., 2017; Willcutt et al., 2019; Zoccolotti et al., 2021). This variability is driven by different sources: Studies used different operationalizations of reading (word, text, accuracy, speed, fluency, comprehension) and arithmetic (oral/written format, problems, computations, fact accuracy, fact fluency). Note that when reading and arithmetic were consistently assessed as fluency measures, correlation coefficients were moderate to high [$r = 0.44$ in Moll et al. (2014); $r = 0.66$ in Willcutt et al. (2019)]. Participants' age also varies substantially in the reported studies (between 7 and 18 years old) and this may affect the strength of the associations. Furthermore, studies were carried out in countries with different languages (Dutch, German, English, Italian). It is possible that linguistic and orthographic characteristics modulate the associations among academic skills.

In line with the notion that acquired skills are related in typical samples, learning disorders in these two skill domains (namely, dyslexia and dyscalculia) co-occur three to five times more often than what would be expected based on individual prevalence rates (Landerl and Moll, 2010; Moll et al., 2014; Koponen et al., 2018; Joyner and Wagner, 2020). These associations suggest the existence of a common basis of the two skills. Studies investigating common and distinct cognitive underpinnings of reading and arithmetic have mostly implemented two different methodological approaches. In one approach, specific profiles of individuals with either isolated deficits (e.g., only reading vs. arithmetic problems) or combined deficits (co-occurring reading and arithmetic problems) are identified and compared (Siegel and Heaven, 1986). This approach is based on the subtyping classification scheme. Studies based on this approach often reported a selective pattern of impaired cognitive performance specifically related to reading but not to arithmetic problems or vice versa (Willburger et al., 2008; Landerl et al., 2009; Willcutt et al., 2013; Cirino et al., 2015; Moll et al., 2015; Raddatz et al., 2017). Landerl et al. (2009) for instance showed that 8- to 10-year-old children with impaired reading fluency and age-adequate arithmetic skills displayed

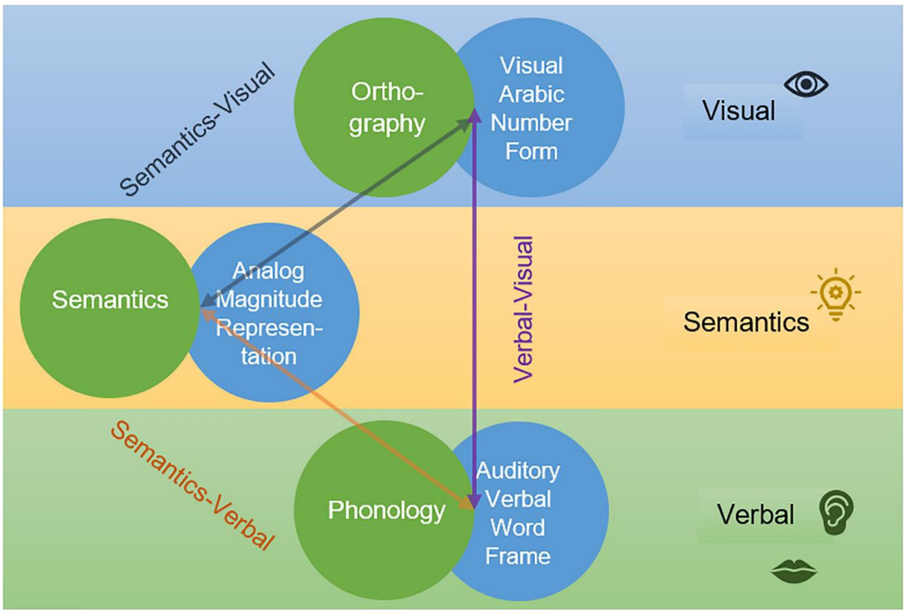

FIGURE 1
Similarity in the cognitive architecture of the lexical quality hypothesis (green circles) and the triple code model (blue circles). From "A-B-3—Associations and dissociations of reading and arithmetic: is domain-specific prediction outdated?," by V. Jöbstl, A. F. Steiner, P. Deimann, U. Kastner-Koller, K. Landerl, 2023, *PlosOne*, 18 (5): e0285437 (doi: 10.1371/journal.pone.0285437). CC BY-NC.

lower performance in phonological processing and rapid naming (RAN) tasks, but not in numerical processing paradigms. Children in the group with impaired arithmetic performance and age-adequate reading skills showed the opposite pattern of impaired number processing performance but age-adequate phonological and rapid naming skills. Importantly, the vast majority of studies in this field consistently reported an additive pattern of cognitive deficits in the group with comorbid reading and arithmetic problems, which means that children with a combined deficit profile manifested the sum of the deficits reported in each single-deficit group and did not show substantial qualitative differences in the cognitive profile compared to single-deficit groups. The subtyping approach is informative, because it provides clear-cut comparisons between matched deficit groups that are specifically selected to show a deficit in one skill domain while controlling for performance in the other skill domain. However, as recently suggested (Peters and Ansari, 2019; Astle and Fletcher-Watson, 2020), this approach has some limitations: (1) Group membership is based on arbitrary cut-offs which differ from study to study, thus impairing replicability and comparability; (2) Participants are selected on the basis of very specific inclusion, exclusion criteria and performance cut-offs. This in turn makes the samples less ecologically valid, in the sense that they do not reflect the full range of variability observed in the general population; (3) Group size is often low, with a clear impact on statistical power.

A more recent, alternative approach assesses reading and arithmetic as continuous dimensions and tests the contribution of different cognitive predictors by means of linear relationships. This dimensional approach has the advantage of being free from the definition of arbitrary cut-offs, as it considers the whole spectrum of possible performance in the skill domain of interest, and is usually implemented in large, unselected samples, thus potentially overcoming the problem of power and generalizability. This approach is in line with the basic tenet of the multiple deficit model (Pennington, 2006; McGrath et al., 2020) and with recent theoretical accounts (Protopapas and Parrila, 2018) that assume quantitative but no qualitative differences in the underlying cognitive skills of individuals with poor vs. typical skill performance. Research using a dimensional approach identified a broad range of skills that similarly predict reading and arithmetic. These include working memory, visuo-spatial memory, phonological awareness, RAN, language skills, processing speed, attention, reasoning, counting and number naming (Durand et al., 2005; Geary, 2011; Fuchs et al., 2016; Cirino et al., 2018; Child et al., 2019; Vanbinst et al., 2020; Zoccolotti et al., 2020; Amland et al., 2021; Bernabini et al., 2021). Longitudinal studies in Finnish children investigated dimensions that predict the co-variation of reading and arithmetic and consistently found that RAN and counting made an important contribution (Koponen et al., 2007, 2020; Korpipää et al., 2017). This evidence was partly replicated in a recent longitudinal study with German-speaking children (Jöbstl et al., 2023). However, uncovering the cognitive underpinnings driving the association of reading and arithmetic, as reported in the above-mentioned studies, is just as relevant as researching the substrates of dissociations (Landerl et al., 2013a). In line with this reasoning, the cross-sectional study by Bernabini et al. (2021) investigated the contribution of a set of cognitive predictors (nonverbal IQ, phonological awareness, magnitude processing and number system knowledge) to reading while controlling for arithmetic performance and vice versa. They collected a sample of 97 Italian speaking children attending 4th and 5th grade of primary school. Results of linear regressions indicated counting as the only dimension predicting both reading and math controlling for the other outcome measures. Magnitude processing, number transcoding and number repetition measures made a specific contribution to arithmetic but not to reading. In a longitudinal study following 885 German-speaking children from kindergarten to Grade 2, Jöbstl et al. (2023) performed a series of fully latent structural equation models. Their results showed that RAN was a significant predictor of reading after controlling for arithmetic skills. In contrast, variance in arithmetic was predicted by magnitude processing independent of reading performance.

To sum up, the available literature indicates distinct as well as common cognitive predictors of reading and arithmetic. Findings are so far controversial and this could be due to different designs, methodological and statistical approaches, languages of participants, age ranges and operationalizations of predictive and outcome measures. There are inconsistencies especially between studies classifying subtypes compared to studies using a dimensional approach. This might be due to the fact that studies using the dimensional approach mostly predicted one skill-domain without controlling for the other skill-domain. Studies that predicted variance in reading controlling for arithmetic and vice versa (Bernabini et al., 2021; Jöbstl et al., 2023) indeed found a more consistent picture which broadly mimics findings by studies with deficit groups.

The current study

The current study investigates the cognitive contributions to low- and high-achievement performance in reading and arithmetic fluency in a longitudinal sample of English- and German-speaking children followed across the first 3 years of elementary school. This developmental period is particularly informative, as it spans the time in which children switch from relying on procedures such as decoding and counting to - at least in part – retrieving verbal facts to read and do arithmetic.

In a previous study (Jöbstl et al., 2024), we showed that there is a degree of specificity in the cognitive predictors of reading and arithmetic fluency. We conducted Ordinary Least Squares (OLS) regression analyses in the same longitudinal sample as in the present study and predicted reading controlling for arithmetic and vice versa. Phonological awareness was a specific predictor of reading fluency, magnitude comparison made a specific contribution to arithmetic fluency. Our results revealed that there are cognitive substrates for associations among skill domains: RAN and multi-digit transcoding explained shared variance between skills. Yet, it is unclear whether these relations hold across the whole continuum of performance. The present study aims to fill this gap by testing whether cognitive predictors of reading and arithmetic have a different importance depending on the skill-level of the criterion variable. We addressed the cognitive basis of associations between reading and arithmetic fluency by modeling their shared variance and, in line with our previous analysis, we inspected specific relations, thus predicting reading after partialling out the influence of arithmetic and vice versa. Importantly, we used the method of quantile regression as a way to overcome the existing controversy related to the methodological inconsistencies of studies with a subtyping vs. dimensional approach. This statistical method will provide us with insight into the question of whether there is a qualitative difference in the pattern of cognitive predictors for high- vs. low-level performance in reading and arithmetic. We aimed to test two competing hypotheses: (1) According to the multiple deficit model, we would expect similar predictive trends in low- and high-achievement ranges. The multiple deficit model theorizes that the distribution of risk factors for a particular disorder is continuous and quantitative (Pennington, 2006), thus implying underlying linear relations between cognitive predictors and outcome variables. (2) Alternatively, we tested whether the predictive pattern differed at different performance levels. It is still possible, that certain predictors are non-linearly related to the criterion variables. This was reported for instance for RAN- and PA-reading associations (de Groot et al., 2015; McIlraith, 2018; Ozernov-Palchik et al., 2022) and also for the association between numerical predictors and arithmetic (Devlin et al., 2022). This research question is particularly relevant in light of literature on low-achieving individuals with dyslexia or dyscalculia. Accordingly, low-level performance was modeled at the 16th percentile of the distribution of reading and arithmetic fluency skills. This percentile corresponds to one standard deviation from the mean and is particularly relevant, as it is defined as the clinical cutoff for identifying reading, spelling and math difficulties in evidence-based guidelines for diagnosing learning disorders in German-speaking countries (Galuschka and Schulte-Körne, 2016; Haberstroh and Schulte-Körne, 2019).

We focused on fluency measures with the aim of having consistent operationalizations in the reading and arithmetic domains, so that regression models would predict different academic competences measured in a similar format.

In the following, we present the cognitive predictors examined in this study and reason on potential differences in their predictive role depending on the performance level. We investigated the role of so-called "domain-general" predictors, dimensions that play a role in a broad spectrum of abilities, including reading and arithmetic; and "domain-specific" predictors, dimensions that are particularly relevant for only one skill domain but not for the other.

Predictors

We included nonverbal intelligence and working memory as domain-general predictors. These dimensions underpin a wide range of cognitive skills (reasoning, problem-solving, visuo-spatial processing, storage and manipulation of visuo-spatial and verbal information) that are related to the outcome variables reading and arithmetic (Geary, 2011; Fuchs et al., 2016; Korpipää et al., 2017) as well as to the predictor variables (Clayton et al., 2020; Koponen et al., 2020). These dimensions were included for two reasons: (1) as nonspecific cognitive predictors of reading and arithmetic; (2) to control for general cognitive functioning related to both outcome and predictor variables. This way, we ensured that relations observed between predictors and outcomes are specific and not due to co-variance with general cognitive abilities.

Among domain-specific predictors, phonological awareness and RAN were expected to be preferentially associated with reading measures. Phonological awareness supports grapheme-phoneme mapping during decoding and is one of the strongest predictors of reading in samples with typical development (Caravolas et al., 2012) and with reading deficits (Landerl et al., 2013b). Studies reporting an association of phonological awareness with arithmetic (Vanbinst et al., 2020; Amland et al., 2021) or with the covariance between reading and arithmetic (Korpipää et al., 2017; Cirino et al., 2018; Koponen et al., 2020) at the beginning of formal schooling suggest that phonological awareness might share some verbal processing features with arithmetic early on. These verbal processing features are related for instance to the manipulation of verbal number words in number reading and writing and to the build-up of arithmetic facts. Phonological awareness can be considered a proxy for language skills in its association with arithmetic. As different language skills make different contributions to arithmetic over the course of development, the contribution of phonological awareness is likely to change over time. Overall, we expected to find a consistent pattern as in our previous analysis (Jöbstl et al., 2024), revealing a large contribution of phonological processing to reading and a minor contribution to arithmetic fluency. With reference to reading, we anticipated a stronger contribution of this predictor in the low- vs. high-performance range due to a higher relevance of phonological processing skills in less automatized and thus more decoding-prone reading styles. Note that this hypothesis has already been confirmed in a previous study based on a subtyping approach (de Groot et al., 2015) but the finding was not replicated in a more recent study that adopted a dimensional approach (McIlraith, 2018).

RAN is a prominent predictor of reading fluency (Landerl et al., 2022). While there is no clear consensus on the theoretical basis of RAN (Kirby et al., 2010), researchers have argued that rapid naming shares many features with reading, as for instance serial processing and access from a visual input to a verbal output (Moll et al., 2009; Georgiou et al., 2013). RAN is considered a proxy for the efficiency of serial retrieval of visual-verbal associations across skill domains. Indeed, a number of studies consistently showed that RAN is not unique to reading but predicts also arithmetic as well as reading-arithmetic co-variation (Balhinez and Shaul, 2019; Georgiou et al., 2020; Koponen et al., 2020). Notwithstanding a clear cross-domain contribution, RAN can be expected to be more strongly related to reading than to arithmetic, as Jöbstl et al. (2023) previously pointed out, because there are more subcomponents of the task that are in common with reading (such as serial processing and naming) than with arithmetic. In line with this reasoning, our former analysis (Jöbstl et al., 2024) revealed a specific contribution of RAN to reading and, to a lesser extent, to arithmetic. In the current study, we anticipated that RAN would predict more strongly the high- vs. low-performance range, because retrieval-related mechanisms as measured by RAN are engaged at highly automated levels of performance. This hypothesis has been confirmed in previous studies (McIlraith, 2018; Ozernov-Palchik et al., 2022).

Counting, magnitude processing and multi-digit number transcoding were included as domain-specific predictors with an anticipated preferential importance for arithmetic. These dimensions have a straightforward relevance for number processing and are established predictors of arithmetic (Starr et al., 2013; Göbel et al., 2014; Banfi et al., 2022; Träff et al., 2023). Evidence about their contribution to reading and reading-arithmetic covariance is mixed.

Knowledge of the counting chain is an important prerequisite to correctly enumerate quantities, a numerical ability that enables to grasp the cardinality principle (the last number word of the counting chain reflects the counted quantity), which in turn boosts the understanding and the connection of different number codes (Geary and VanMarle, 2018). Reciting the counting chain also supports arithmetic early on, as it allows to solve simple additions and subtractions by counting up- or downward from a given number (Peters and De Smedt, 2018). Counting fluency was found to be highly correlated with RAN and to account for variance in reading (Koponen et al., 2013; Bernabini et al., 2021) and covariance in reading and arithmetic (Korpipää et al., 2017; Koponen et al., 2020). This evidence was taken as an indication that counting (together with RAN) measures the efficiency of serial retrieval skills, that is, the speed with which verbal traces are retrieved from memory. Another study (Cirino et al., 2018), however, showed that counting was more strongly predictive of arithmetic than reading. Cirino et al. (2018) did not control for reading while predicting arithmetic and it is therefore unclear whether the contribution of counting was specific to arithmetic or not. Note that in our previous analysis (Jöbstl et al., 2024), counting made a very marginal, specific contribution to arithmetic fluency (a not significant trend), and it did not reliably predict reading. Against this background and given its foundational role in supporting number processing and basic calculation strategies, we anticipated a prominent prediction of counting to arithmetic at the beginning of formal schooling and especially among children with low arithmetic fluency. Its contribution to reading fluency is expected to be minor.

Magnitude processing refers to the ability to compare or judge quantities such as dots or Arabic numbers. According to the triple code model, it requires understanding of quantities as analog representations and (for symbolic magnitude processing) to connect these to their symbolic counterpart in the visual Arabic code. There is evidence indicating unique prediction of symbolic and of non-symbolic magnitude comparison to arithmetic but not to reading (Durand et al., 2005; Jöbstl et al., 2023). Other studies, however, found that symbolic magnitude processing explained variance in reading comprehension, but not fluency (Cirino et al., 2018) and it also explained reading and arithmetic covariance (Koponen et al., 2020), although to a small extent. The reason for the involvement of magnitude processing in reading is not straightforward. It can be assumed that, similar to word reading, magnitude processing requires cross-format integration of the visual and semantic codes. In this sense, magnitude processing and reading might share similar symbolic processing strategies. Note that our former analysis (Jöbstl et al., 2024) revealed a reliable, specific contribution of magnitude processing to arithmetic fluency but not to reading. In the current study we tested the hypothesis that magnitude processing skills are more relevant in the low vs. high arithmetic fluency range. The ability to compare dots and single digits reflects basic numerical processing skills related to understanding quantities and being able to map Arabic digits onto number semantic. As arithmetic skills develop, access to and use of these basic number processing skills get more and more automatized. At the same time, symbolic number processing and especially knowledge of arithmetic facts become essential to efficiently solve calculations and thus retrieval of verbal number traces from memory supersedes the understanding of the meaning of numbers. Note that a previous longitudinal study with 209 Dutch-speaking children followed from kindergarten to Grade 1 already tested a similar hypothesis and did not confirm it (Bartelet et al., 2014). Bartelet et al. found a comparable contribution of symbolic magnitude processing to arithmetic across performance levels, thus suggesting a linear relation between these variables. It will be highly relevant to consider whether the results by Bartelet et al. (2014) can be confirmed in the current study, which investigates an older cohort.

Multi-digit number transcoding refers to the ability to shift between spoken number words (e.g., "thirty-two") and Arabic digits (e.g., 32) and requires mastery of two number codes. On the one hand, understanding of morpho-syllabic structures enables to build-up and understand complex spoken number words in the verbal code (e.g., "sixty" is made up of two morphemes: "six," indicating the magnitude, and "ty" indicating the place-value class). On the other hand, Arabic digit knowledge and place-value understanding are fundamental to grasp the Arabic number code. Accordingly, previous evidence suggests that multi-digit transcoding explains variance in tasks requiring place-value understanding (Cheung and Ansari, 2021). Multi-digit transcoding was reported to predict reading and reading-arithmetic co-variation (Cirino et al., 2018; Koponen et al., 2020; Amland et al., 2021). The underlying sources of shared variance between multi-digit transcoding and reading may be twofold: (1) Matching of verbal number words and Arabic digits in number transcoding parallels letter-sound binding during reading. These two dimensions may thus share variance related to cross-format mapping; (2) Reading words and transcoding numbers involves also higher-level language resources like morpho-syllabic processing of verbal codes, which is necessary for word construction and analysis. Against this background, we developed two competing hypotheses. The first hypothesis assumes that multi-digit transcoding and reading fluency co-vary due to shared cross-format mapping. Being grapheme-phoneme conversion a foundational skill of reading, we anticipated a contribution of multi-digit transcoding in children with low reading skills at the start of formal schooling, a time window in which some children are still struggling with decoding procedures. Our alternative hypothesis posits that multi-digit transcoding predicts high-level reading fluency, due to shared resources related to morpho-syllabic processing. The use of morpho-syllabic features of words during reading emerges after grapheme-phoneme conversion routines have been consolidated (Ehri, 2005, 2014). A recent review highlighted that reading strategies based on morpho-syllabic processing can be observed as early as in Grade 1 (Levesque et al., 2021). We therefore anticipated to observe a contribution of multi-digit transcoding to high-level reading fluency already in Grade 1.

Method

Participants

This study reports the analysis of data from a cross-linguistic project that examined the development of numerical and reading skills in English- and German-speaking primary school children followed from Grade 1 to Grade 3 (Göbel et al., 2020). Longitudinal data relevant for the current study variables was available from a total of 357 children: 191 English-speaking children (t1: Mage = 6 years, 2 months; SDage = 4 months; 50% female; 97% monolingual) and 166 German-speaking children (t1: Mage = 7 years, 2 months; SDage = 3 months; 47% female; 87% monolingual). German-speaking children in Graz (Austria) came from a middle-income urban school district. English-speaking children in Yorkshire (United Kingdom) came from four urban, three town, and four rural schools, with a mean deprivation index decile score of 8 (indicating the 30% of least deprived neighborhoods) (Department for Communities and Local Government, 2015) and an average of 11% of free school meals.

The use of a cross-linguistic design resulted in differences between the samples: (1) English-speaking children were on average one year younger than German-speaking children because compulsory education starts earlier in the UK as compared to Austria; (2) The decade-unit inversion in number transcoding tasks is an additional challenge for German- as compared to English-speaking children. We accounted for language-related differences in the statistical analysis (see data preparation and analysis section for more details). Note that the cognitive dimensions were correlated to the outcome variables in both samples, language group had no substantial impact on these associations. The correlation table is reported in Jöbstl et al. (2024). As language did not impact associations, we merged the two language samples to increase the statistical power of the multivariate quantile regression models.

The study adhered to the principles outlined in the Declaration of Helsinki, and consent was obtained from both the children and their guardians. The ethics committees of the Universities of York and Graz approved the study (Reference number, University of York: 559; University of Graz: 39/23/63 ex 2016/17).

Measures

Reading and arithmetic fluency tasks were administered at each time point with similar tasks (t1: Fall of Grade 1, April–July 2017; t2: Fall of Grade 2, April–July 2018; t3: Fall of Grade 3, April–July 2019). Reading fluency was assessed by means of standardized timed tests with words and pseudowords [English: TOWRE-2, Torgesen et al. (2012); German: SLRT-II; Moll and Landerl (2010)]. Arithmetic fluency was measured with one-minute timed tasks including additions and subtractions at t1 and t2, multiplication and division subtasks were added at t3. The assessment of predictors of reading and arithmetic skills took place at t1. These included: (1) Nonverbal IQ, measured with the Raven's Standard Progressive Matrices Plus (Raven et al., 1998) adapted for group use; (2) Working memory, assessed with Digit Recall Forward, Backward and Block Recall Forward from the Working Memory Test Battery for Children (Pickering and Gathercole, 2001) and a non-standardized Block Recall Backward task; (3) Phonological awareness, collected with phoneme deletion tasks. The English-speaking sample was administered the York Assessment of Reading Comprehension (Snowling et al., 2009), the German-speaking sample was administered a comparable task developed within our lab; (4) RAN tasks with letters and digits from the Comprehensive Test of Phonological Processing [CTOPP; Wagner et al. (2013)]; (5) Counting fluency, administered by means of a timed forward counting task; (6) Magnitude processing, assessed with dots and digits comparison tasks; (7) Multi-digit transcoding, measured with three indicators: number identification, number reading, and number writing. A full description of the tasks is available in the Supplementary Appendix and can also be found in Jöbstl et al. (2024).

Data preparation and analysis

Predictors consisting of only one dimension were z-standardized separately in each language group. This was done for two reasons: (1) some measures as for example phonological awareness were assessed with slightly different tasks in the two language groups and therefore raw scores were not on the same scale; (2) The English- and German-speaking groups were matched on duration of formal education. English children, however, start school 1 year earlier than Austrian children, resulting in about 1 year age difference between language groups. Z-standardizing the data in each language group prevented a bias due to this age gap.

For tasks with multiple conditions, principal component analysis (PCA) was employed to derive a composite score. PCAs were conducted in each language group, separately (see Supplementary Appendix for further details). A comprehensive summary of factor loadings can be found in Jöbstl et al. (2024).

We employed quantile regression models to investigate whether prediction patterns depend on the skill level of reading fluency and arithmetic fluency. Quantile regression models enable the investigation of prediction patterns at different performance levels, thus testing for non-linear relationships between variables. Like Bartelet et al. (2014), we performed quantile regressions at the 16th, 50th, and 84th quantiles, which correspond to criterion scores below one standard deviation from the sample mean, the median, and above one standard deviation from the mean, respectively.

We computed two sets of quantile regression models: (1) To address the cognitive basis of the association between reading and arithmetic fluency, we predicted shared variance in reading and arithmetic fluency at each time point. The dependent variables in this set of models were computed by means of PCA on reading and arithmetic fluency variables in each grade (see Supplementary Table S1); (2) To investigate prediction patterns specific to either reading or arithmetic, we included the respective other domain in each regression model (e.g., arithmetic in models predicting reading and vice versa). This enabled us to specifically target one skill domain by controlling for shared variance between the two skill domains. The anova() function (Chambers and Hastie, 1992) was used to compare significant slopes between different quantile models. Specifically, it performs an analysis of variance (ANOVA) to test the equality of distinct slopes across the different quantile models. Goodness of fit was reported as pseudo-R^2 for quantile regression (*R1*, Koenker and Machado, 1999).

Note that the grouping variable referring to the language spoken by participants was not included in any of the regression models. Data were z-standardized separately in each sample and therefore language-related differences in intercepts were controlled for. We also performed OLS regression models, in which we tested whether language group predicted shared and specific variance in reading and arithmetic. Results indicated no significant contribution of this variable ($\beta \leq 0.15$, $p \geq 0.285$), which also yielded moderate collinearity with age (VIF = 3.6) and was thus redundant in the models.

We additionally conducted sensitivity analyses to ensure that the results of the quantile regression models conducted in the whole sample were robust against differences in the distribution of variables in the two language groups. We computed quantile regression models in the two language samples separately and then pooled regression coefficients using meta-analysis methods in SPSS (fixed effects). Results are reported in the supplement.

The analysis was run in IBM SPSS Statistics (Version 29) and in R (R Core Team, 2024), *p*-values <0.05 were considered significant.

Results

Descriptive statistics, including raw scores of the administered tasks, and the correlation matrix can be found in Jöbstl et al. (2024). Quantile regression models predicting shared variance in reading and arithmetic fluency are reported in Table 1. Table 2 displays results of the regression model predicting specific variance in reading fluency, the prediction of arithmetic fluency is in Table 3.

Predictors of shared variance in reading and arithmetic fluency

Nonverbal IQ predicted common variance in reading and arithmetic fluency in Grade 1 at percentile 50, in Grade 2 at percentiles 50 and 84, in Grade 3 again only at percentile 50. Slopes did not differ significantly. Working memory made no significant contribution.

Phonological awareness predicted common variance in reading and arithmetic at Grade 1 across performance levels. In Grade 2, this predictor showed only a trend for a significant contribution to quantile 50 (p =0.054). In Grade 3, phonological awareness predicted performance at percentile 16 and 50. There was no significant difference among slopes across performance levels at any grade. RAN was a robust predictor of common variance in fluency tasks across grades and performance levels, with no difference in slopes.

Counting made no significant contribution. Magnitude processing showed a trend for a significant prediction in Grade 1 at percentile 50 (p =0.052), and was a significant predictor in Grade 2 at percentile 84. Finally, multi-digit transcoding predicted reading and arithmetic fluency in all grades and across all performance levels. The slope in Grade 1 at percentile 84 was significantly higher than percentile 50 and percentile 16, though as a trend in the latter (p =0.054).

Specific predictors of reading fluency

Age did not uniquely predict reading fluency in any grade or at any quantile level.

Nonverbal IQ predicted 1st grade reading fluency at quantile 84, but not in subsequent grades or at quantile levels 16 and 50. Working memory predicted 2nd grade reading at quantile 16 and 3rd grade reading at quantile level 50. However, the effect of working memory on reading was negative, indicating a suppressor effect. This negative effect is due to the overlap of working memory with other predictors, such as RAN, multi-digit transcoding and arithmetic fluency, which accounted for shared variance while working memory explained criterion-irrelevant variance.

Phonological awareness and RAN were consistent predictors of reading fluency across different performance levels and grades. Counting and magnitude processing, generally considered to be domain-specific to arithmetic, did not explain variance in reading at any quantile or in any grade. Multi-digit transcoding explained variance above arithmetic at quantile levels 50 and 84, but only in 1st grade.

Arithmetic fluency was a significant predictor of reading fluency in each grade. In 1st grade, there was a significant effect at quantiles 16 and 84. In 2nd and 3rd grade, arithmetic fluency consistently explained variance across all quantiles.

It is important to note that while different predictors showed varying effects depending on quantile models, there were no significant differences between slopes.

Specific predictors of arithmetic fluency

There was a significant effect of age in 2nd grade arithmetic fluency at quantile level 84, which differed significantly from quantile level 16.

Nonverbal IQ predicted arithmetic fluency in 2nd grade at quantile levels 16 and 50, and in 3rd grade at quantile level 16. While there was no significant difference in slopes in 1st or 2nd grade, there was a significant difference in 3rd grade. The effect at quantile16 was significantly higher than the effect at quantile 84. Working memory was a significant predictor in 1st grade and marginally 3rd grade (quantile 16; p =0.052). In 1st grade, working memory predicted arithmetic fluency at quantiles 16 and 50, but there was no significant difference among slopes.

Phonological awareness, an established predictor of reading, did not predict arithmetic fluency significantly. Phonological awareness acted as a suppressor variable, likely due to shared variance with reading fluency. RAN accounted for variance in 1st grade arithmetic performance at quantile 16, and marginally quantile 50 (p =0.055), with no significant differences between slopes.

Counting, a domain-specific predictor of arithmetic, did not predict 1st or 2nd grade arithmetic fluency. However, it predicted 3rd grade arithmetic fluency at quantile 16, with a significant difference compared to quantiles 50 and 84. Magnitude comparison was generally a stable predictor of arithmetic performance across various quantiles, with no significant differences between quantile levels. Multi-digit transcoding was a significant predictor of arithmetic fluency across grades. In 1st grade, multi-digit transcoding predicted arithmetic performance across skill levels, with the largest effects at quantile 84 compared to quantiles 16 and 50. In 2nd grade, multi-digit transcoding was again significant across performance levels, with no differences between slopes. In 3rd grade, multi-digit transcoding predicted significant variance at quantiles 50 and 84, with a significant difference between slopes in favor of quantile 84.

Reading fluency predicted arithmetic fluency across grades and quantiles, except for 1st grade. In 1st grade, only the effect at quantile 50 was significant with no significant differences between slopes in any grade.

Sensitivity analyses with pooled regression coefficients are reported in the Supplementary Tables S2–S4. Results are very consistent with the quantile regression models conducted in the whole sample.

Discussion

The current study investigated cognitive dimensions that contribute to reading and arithmetic fluency through the first 3 years of primary school, examining whether these contributions differed at different levels of performance. We ran quantile regression models that fitted low-, median- and high-performance

TABLE 1 Quantile regression models predicting common variance in reading and arithmetic fluency in Grade 1, 2, and 3.

	D. Slopes			Qu 16		Qu 50		Qu 84	
T1 reading and arithmetic	1–2	1–3	2–3	*β*	*p*	*β*	*p*	*β*	*p*
Age				−0.07	0.111	−0.04	0.396	−0.001	0.984
Nonverbal IQ				0.07	0.180	0.13	0.008	0.09	0.135
Working memory				0.06	0.296	0.06	0.258	0.07	0.316
Phonological awareness				0.10	0.07	0.17	0.002	0.16	0.002
RAN				0.27	< 0.001	0.34	< 0.001	0.37	< 0.001
Counting				0.11	0.052	0.06	0.276	−0.01	0.902
Magnitude processing				0.08	0.119	0.10	0.052	0.11	0.102
Multi-digit transcoding		0.054	0.016	0.26	< 0.001	0.25	< 0.001	0.42	< 0.001
R1				0.38		0.42		0.42	
T2 reading and arithmetic									
Age				0.01	0.920	−0.02	0.765	−0.03	0.56
Nonverbal IQ				0.06	0.366	0.12	0.039	0.15	0.024
Working memory				0.04	0.547	0.06	0.398	0.03	0.734
Phonological awareness				0.11	0.097	0.13	0.054	0.06	0.342
RAN				0.33	< 0.001	0.40	< 0.001	0.38	< 0.001
Counting	0.018			0.10	0.112	−0.07	0.32	0.05	0.436
Magnitude processing				0.06	0.317	0.11	0.111	0.13	0.036
Multi-digit transcoding				0.22	0.003	0.32	< 0.001	0.28	< 0.001
R1				0.33		0.36		0.34	
T3 reading and arithmetic									
Age		0.028		−0.09	0.162	0.003	0.957	0.08	0.153
Nonverbal IQ				0.12	0.069	0.15	0.012	0.06	0.326
Working memory				−0.06	0.386	−0.04	0.564	0.05	0.515
Phonological awareness				0.14	0.045	0.16	0.011	0.08	0.180
RAN				0.40	< 0.001	0.40	< 0.001	0.49	< 0.001
Counting				0.02	0.798	0.06	0.376	0.03	0.691
Magnitude processing				−0.03	0.707	0.06	0.330	0.07	0.276
Multi-digit transcoding				0.30	< 0.001	0.26	0.001	0.31	< 0.001
R1				0.30		0.34		0.34	

The columns "D-Slopes" report significant *p*-values for differences between significant regression coefficients in different quantile models: 1–2 reflects the comparison of quantiles 16 and 50, 1–3 of quantiles 16 and 84, 2–3 of quantiles 50 and 84. *R1* is the pseudo-R^2 for quantile regression models.

in the outcome variables, referring to the 16th, 50th, and 84th percentile, respectively. Note that the 16th percentile is recognized as the clinical cutoff for identifying reading, spelling and arithmetic difficulties in evidence-based guidelines for diagnosing dyslexia and dyscalculia in German-speaking countries (Galuschka and Schulte-Körne, 2016; Haberstroh and Schulte-Körne, 2019). Our analyses aimed to unravel whether there is a substantial difference in the predictive pattern of cognitive dimensions in low-achieving individuals as compared to individuals with typical development and thus represent an important first step in reconciling seemingly inconsistent evidence from two research streams, namely the subtyping and the dimensional approaches.

We addressed the cognitive basis of associations between reading and arithmetic fluency by predicting their shared variance. In a second set of models, we aimed at unveiling the cognitive basis of dissociations between academic skills by introducing arithmetic as a covariate while modeling reading and vice versa, thus addressing the specificity of cognitive predictors for one skill domain.

In a previous analysis of the same data based on OLS multiple regression models Jöbstl et al. (2024), phonological awareness was identified as a reading-specific predictor and magnitude processing as arithmetic-specific predictor. Rapid naming and multi-digit transcoding explained shared variance in reading and arithmetic fluency. Our findings are broadly consistent with this analysis. Importantly, the current study shows that most cognitive dimensions contribute similarly to reading and arithmetic at different levels of attainment, suggesting the existence of linear relations, as the multifactorial perspective would predict. There were few exceptions to this general pattern. We discuss these results in detail in the next sections.

TABLE 2 Quantile regression models predicting reading fluency in Grade 1, 2, and 3.

	D. Slopes			Qu 16		Qu 50		Qu 84	
T1 reading	1–2	1–3	2–3	*β*	*p*	*β*	*p*	*β*	*p*
Age				−0.03	0.576	−0.01	0.889	−0.01	0.896
Nonverbal IQ				0.07	0.248	0.08	0.212	0.16	0.008
Working memory				−0.06	0.293	−0.11	0.086	−0.10	0.187
Phonological awareness				0.23	< 0.001	0.29	< 0.001	0.28	< 0.001
RAN				0.31	< 0.001	0.40	< 0.001	0.47	< 0.001
Counting				0.02	0.704	0.01	0.844	0.003	0.967
Magnitude processing				−0.03	0.648	−0.02	0.700	−0.09	0.195
Multi-digit transcoding				0.02	0.814	0.16	0.023	0.15	0.040
Arithmetic fluency (t1)				0.17	0.031	0.08	0.262	0.25	0.010
R1				0.29		0.34		0.37	
T2 reading									
Age				−0.07	0.238	−0.04	0.497	< 0.001	0.997
Nonverbal IQ				0.04	0.541	0.08	0.268	0.11	0.053
Working memory				−0.14	0.027	−0.09	0.230	−0.02	0.769
Phonological awareness				0.28	< 0.001	0.24	0.001	0.18	0.006
RAN				0.48	< 0.001	0.48	< 0.001	0.47	< 0.001
Counting				0.05	0.427	−0.08	0.291	−0.01	0.911
Magnitude processing				−0.09	0.204	−0.02	0.743	−0.04	0.520
Multi-digit transcoding				0.06	0.473	0.12	0.133	0.07	0.385
Arithmetic fluency (t2)				0.15	0.021	0.21	0.017	0.17	0.020
R1				0.32		0.29		0.30	
T3 reading									
Age				−0.05	0.452	−0.09	0.112	−0.01	0.875
Nonverbal IQ				0.07	0.346	0.13	0.053	−0.02	0.803
Working memory				−0.05	0.396	−0.15	0.043	−0.04	0.659
Phonological awareness				0.27	< 0.001	0.22	0.001	0.22	0.001
RAN				0.39	< 0.001	0.47	< 0.001	0.40	< 0.001
Counting				0.06	0.441	−0.03	0.682	−0.13	0.115
Magnitude processing				−0.05	0.546	−0.04	0.610	−0.13	0.062
Multi-digit transcoding				0.001	0.988	0.07	0.387	0.16	0.069
Arithmetic fluency (t3)				0.31	< 0.001	0.26	0.001	0.25	0.003
R1				0.32		0.33		0.27	

The columns "D-Slopes" report significant *p*-values for differences between significant regression coefficients in different quantile models: 1–2 reflects the comparison of quantiles 16 and 50, 1–3 of quantiles 16 and 84, 2–3 of quantiles 50 and 84. *R1* is the pseudo-R^2 for quantile regression models.

Predictors of shared variance in reading and arithmetic fluency

Regression models investigating the cognitive basis of shared variance in reading and arithmetic fluency are consistent with the previous partial correlation analysis by Jöbstl et al. (2024), which indicated a minor contribution of domain-general predictors and a major contribution of domain-specific predictors to covariance in reading and arithmetic fluency, with an important portion of it explained by RAN and multi-digit transcoding.

In the current study, five dimensions predicted shared variance in reading and arithmetic: Nonverbal IQ, phonological awareness, RAN, magnitude processing and multi-digit transcoding. Nonverbal IQ is a domain-general dimension, its involvement in reading-arithmetic covariance indicates that visuo-spatial and reasoning skills are relevant for both literacy and math skills.

The contribution of phonological awareness to shared variance in reading and arithmetic fluency highlights that language processing underpins the covariation between these domains. The fact that this dimension predicted shared variance in Grade 1, was not significant in Grade 2, and again made a reliable contribution in Grade 3 suggests

TABLE 3 Quantile regression models predicting arithmetic fluency in Grade 1, 2, and 3.

	D. Slopes			Qu 16		Qu 50		Qu 84	
T1 arithmetic fluency	1–2	1–3	2–3	β	p	β	p	β	p
Age				−0.02	0.739	−0.01	0.778	0.04	0.503
Nonverbal IQ				0.09	0.112	−0.01	0.898	−0.01	0.914
Working memory				0.15	0.011	0.20	0.001	0.13	0.084
Phonological awareness				0.02	0.739	−0.12	0.062	−0.08	0.269
RAN				0.15	0.006	0.12	0.055	0.11	0.227
Counting				0.05	0.389	0.08	0.213	0.09	0.144
Magnitude processing				0.12	0.026	0.22	< 0.001	0.18	0.008
Multi-digit transcoding		0.009	0.022	0.23	0.004	0.31	< 0.001	0.50	< 0.001
Reading (t1)				0.07	0.298	0.17	0.021	0.09	0.345
R1				0.32		0.33		0.36	
T2 arithmetic fluency									
Age		0.027		−0.06	0.371	0.02	0.763	0.12	0.050
Nonverbal IQ				0.13	0.037	0.14	0.036	0.05	0.511
Working memory				0.13	0.099	0.11	0.135	0.13	0.124
Phonological awareness				−0.08	0.300	−0.10	0.164	−0.12	0.094
RAN				0.03	0.707	0.07	0.311	0.12	0.210
Counting				0.09	0.229	0.08	0.235	−0.05	0.539
Magnitude processing				0.09	0.273	0.18	0.008	0.22	0.003
Multi-digit transcoding				0.24	0.006	0.24	0.002	0.36	< 0.001
Reading (t2)				0.21	0.003	0.20	0.008	0.27	0.005
R1				0.22		0.29		0.29	
T3 arithmetic fluency									
Age				−0.07	0.137	0.01	0.830	0.09	0.146
Nonverbal IQ		0.046		0.15	0.012	0.11	0.092	−0.02	0.755
Working memory				0.13	0.052	0.09	0.171	0.11	0.115
Phonological awareness				−0.05	0.553	−0.03	0.662	−0.06	0.305
RAN				0.03	0.680	0.11	0.160	0.10	0.202
Counting	0.040	0.036		0.24	0.002	0.07	0.301	0.04	0.587
Magnitude processing				0.15	0.014	0.13	0.044	0.13	0.057
Multi-digit transcoding		0.041		0.12	0.092	0.18	0.024	0.33	< 0.001
Reading (t3)				0.27	< 0.001	0.24	0.002	0.38	< 0.001
R1				0.26		0.29		0.32	

The columns "D-Slopes" report significant p-values for differences between significant regression coefficients in different quantile models: 1–2 reflects the comparison of quantiles 16 and 50, 1–3 of quantiles 16 and 84, 2–3 of quantiles 50 and 84. *R1* is the pseudo-R^2 for quantile regression models.

that the language processing dimensions subtended by phonological awareness differ at different levels of instruction. Cross-format mapping as in grapheme-phoneme conversion (reading) and number transcoding (arithmetic) are probably the most likely candidates to explain the contribution of this predictor to shared variance in Grade 1. Note that phonological awareness did not predict low-level shared performance in Grade 1. This may indicate a non-linear relation, although this interpretation is highly unlikely given the robust linear pattern observed in the prediction of specific variance in reading fluency (see further for a discussion of this result). Alternatively, it is possible that the slope in the low-performance range just missed the threshold for significant results ($p = 0.07$). The role of phonological awareness in Grade 3 may be related to the manipulation of verbal material as for example during memory retrieval of orthographic representations and of arithmetic facts. The prediction did not hold in the high-performance range, indicating that this kind of language processing resources are no longer required once reading and arithmetic competences are highly automatized. Note, however, that our interpretations are very speculative. The observational design of this study and the assessed variables do not allow us to address more precisely the nature of this association. Furthermore, phonological awareness is an intrinsically multifactorial dimension with a complex and bidirectional relation with reading (Landerl et al., 2019). Future studies should investigate in more detail the linguistic features that underlie the covariation between reading and arithmetic, perhaps moving beyond the concept of phonological awareness.

RAN contributed to shared variance in reading and arithmetic fluency across grades and performance levels. Our findings are consistent with a number of previous studies in Finnish- and German-speaking children (Koponen et al., 2007, 2020; Korpipää et al., 2017; Jöbstl et al., 2023). According to the model proposed by Jöbstl et al. (2023), associations between reading and arithmetic can be explained by shared integration mechanisms of visual, verbal and semantic information. Rapid naming is supposed to measure serial retrieval fluency and is therefore a good candidate for measuring these integration processes especially between visual and verbal codes.

Magnitude comparison made a small and isolated contribution to high-level shared performance in Grade 2. This might be an indication of the role of symbolic processing resources in reading-arithmetic covariation. Note, however, that this scattered significant finding may as well be driven by the high number of regression models performed (see limitation section for a more in-depth consideration of this problem). It should thus be interpreted with caution until replicated.

Multi-digit transcoding was a reliable predictor of shared variance at all grades and performance levels, with the strongest prediction in Grade 1 at high-level performance. There are two likely explanations for the involvement of this dimension in reading and arithmetic covariance. First, multi-digit transcoding requires cross-format mapping, that is, mapping of Arabic digits onto number words and vice versa, a feature that is necessary to do arithmetic and that parallels letter-sound binding in reading. Second, multi-digit transcoding requires morpho-syllabic processing, which enables to convert spoken number words into Arabic numbers (and vice versa) using the so-called transcoding routines (Barrouillet et al., 2004). This form of higher-level language processing is relevant for reading too and therefore it may represent a second important source of shared variance between reading and arithmetic fluency.

Specific predictors of reading

Nonverbal IQ predicted high-level performance in Grade 1. Its minor contribution suggests a marginal involvement of general cognitive resources in reading fluency. Working memory predicted reading fluency in Grade 2 at percentile 16 and in Grade 3 at percentile 50. Note that all significant and non-significant beta coefficients were negative for this regressor, meaning that higher working memory resources were related to lower reading fluency. This is unexpected given that working memory was positively correlated with reading fluency (rs between 0.22 and 0.29). The pattern reflects a suppression effect due to the association of working memory with RAN, multi-digit transcoding and arithmetic fluency. This suppressor effect indicates that the four dimensions share a common component, most likely related to the central executive resources of working memory (Baddeley, 2012). The central executive includes sustained, selective attention and monitoring skills that are transversal to many tasks and situations of cognitive effort. These resources are not specific to the prediction of reading fluency, and probably enhanced the effect of other variables.

Our hypothesis that phonological awareness and RAN would preferentially contribute to reading fluency was confirmed. The two dimensions made specific contributions throughout performance levels and time points. This finding is in line with our previous analysis Jöbstl et al. (2024) and with a number of previous studies [see Landerl et al., 2022 for a recent review] and highlights the importance of these two components for reading fluency and its development through the first 3 years of school in a less (English) and a more transparent orthography (German). Our results indicate the existence of a linear relation between the two constructs and reading fluency, thus suggesting that phonological awareness and RAN predict low and high reading fluency skills to the same extent. Our findings are partly in line with a previous longitudinal study based on a dimensional approach (McIlraith, 2018). This study used quantile regression to predict Grade 1 reading performance from preschool and kindergarten predictors in 293 English-speaking children. They found reliable contributions of RAN and PA measured in kindergarten to reading at median and high-performance levels but not at low-performance level. There may be multiple reasons for the existing inconsistencies between our study and previous evidence, as for instance differences in the statistical analysis, the type of tasks, the assessment time point and the level of instruction. Future studies should aim to disentangle these confounding effects in the attempt to understand these inconsistencies.

As to cross-domain contributions, multi-digit transcoding predicted reading fluency in Grade 1 at median and high-level performance. This pattern of specific contribution does not align well with the idea that cross-format mapping is the primary source of co-variation between the two skills. If that was the case, we should have found a reliable contribution of multi-digit transcoding to reading at low-performance level. Our findings are more consistent with the hypothesis that multi-digit transcoding and reading share morpho-syllabic processing resources that are particularly relevant for high-level reading fluency. Note, however, that this hypothesis would have predicted a reliable contribution of transcoding to reading throughout grades, which was not found. The lack of contribution at higher grades is difficult to explain. We might speculate that the morpho-syllabic processing skills shared between the two dimensions are simple, mostly consisting in derivational rules for word assembly (as in "friend-ship" and "six-ty"), and these may be well consolidated already at the end of Grade 1, thus being not relevant for reading fluency in higher grades.

Counting and magnitude processing made no significant prediction to reading fluency, thus suggesting that these dimensions are not relevant for reading fluency once variance in arithmetic is controlled for.

Specific predictors of arithmetic

Nonverbal IQ predicted low- and median-level arithmetic fluency in Grade 2 and low-level arithmetic fluency in Grade 3. The nonverbal IQ task involved reasoning and visuospatial processing, characteristics that are intrinsic to number processing (Cipora et al., 2020). This pattern of prediction suggests that the involvement of these skills in arithmetic becomes relevant above Grade 1, probably coinciding with more mature processing of multi-digit numbers and understanding of the positional system. The fact that nonverbal IQ contributed to low/median performance indicates that resources like reasoning and visuospatial processing are engaged by children who do not yet rely upon efficient verbal retrieval skills for arithmetic facts and need these

10.3389/feduc.2024.1335957

resources as a support for effortful calculation strategies based on algorithms.

Working memory predicted arithmetic fluency at low- and median-level performance in Grade 1 (there was a similar trend also in Grade 3). These findings suggest that working memory resources play a more prominent role in young children with rather low arithmetic fluency skills. This is in line with previous evidence reporting a working memory deficit in dyscalculia (Szucs et al., 2013). Low performance on speeded addition and subtraction tasks can either derive from erroneous fact retrieval due to interfering memory traces (De Visscher and Noël, 2013) or from limited availability of arithmetic facts, which results in the use of simple counting strategies as an alternative method to achieve a correct solution. In both cases, children require an important engagement of storage and manipulation of numerical information, thus taxing working memory resources. In contrast, children with high performance in speeded addition and subtraction tasks most likely rely on robust memory traces that can be effortlessly accessed with no need for any further manipulation.

Among numerical predictors, counting predicted arithmetic fluency in Grade 3 in the low-level performance range. This result is only partly in line with our hypotheses. We expected counting to be preferentially related to arithmetic, with a more important contribution at the beginning of formal schooling. The fact that counting predicted arithmetic fluency in Grade 3 but not in Grade 1 is difficult to reconcile with our hypothesis. Our results are also against the idea that counting is related to serial retrieval mechanisms, as it should predict high-level arithmetic fluency performance, which is supposed to rely heavily on retrieving facts from memory. In contrast, counting only contributed to low-level performance in Grade 3. Our findings rather suggest that counting is used as a compensatory strategy to solve arithmetic problems in children that cannot fluently retrieve arithmetic facts. This interpretation is in line with clinical descriptions of children with dyscalculia that mention preponderant use of counting instead of fact retrieval strategies in solving math problems (Kucian and von Aster, 2015; Haberstroh and Schulte-Körne, 2019).

With sparse exceptions, magnitude processing was a reliable predictor of arithmetic fluency throughout grades and performance levels. This is consistent with previous literature reporting a specific contribution of magnitude processing to arithmetic (Jöbstl et al., 2023). In line with Bartelet et al. (2014), we found no difference in the predictive pattern depending on performance level. Our results thus highlight that magnitude processing makes a similar contribution at any level of arithmetic achievement during the first 3 years of formal schooling. This is in line with the assumption of the multiple deficit model and highlights that the ability to understand non-symbolic and symbolic number semantic contributes to arithmetic fluency at any level of proficiency.

Multi-digit number transcoding contributed robustly to arithmetic through grades and performance levels (the only exception being its non-significant contribution to low-level performance in Grade 3). In line with our previous analysis Jöbstl et al. (2024), the current findings indicate that the contribution of multi-digit number transcoding to arithmetic fluency is specific, as it holds after controlling for variance in reading fluency. This finding adds up to previous evidence indicating multi-digit number transcoding as one of the most important predictors of arithmetic over and above other numerical predictors (Göbel et al., 2014; Habermann et al., 2020; Banfi et al., 2022). Across grades, the increase in the slopes of the transcoding-predictors with increasing levels of arithmetic fluency likely indicates that these two skills co-vary more strongly when both are highly efficient and automated.

As to cross-domain predictors, we found that RAN predicted Grade 1 arithmetic fluency at the low-performance level and, as a trend, at median-performance level. The significant contribution of RAN to Grade 1 arithmetic fluency was already reported in our previous analysis with OLS regression. Our results partly support previous literature reporting rapid naming as a predictor of arithmetic fluency (Koponen et al., 2013, 2020; Korpipää et al., 2017), and extends it by showing that the association between RAN and arithmetic fluency holds after controlling for reading. However, the reason why in the present study this variable contributed to low-level reading performance is unclear. We expected RAN to predict high-performance level, as it is supposed to measure serial retrieval efficiency relevant for automatized and retrieval-based skills. One possible explanation for the current results is that RAN captured variance in arithmetic fluency that was not already explained by reading fluency. Beginning readers with low reading fluency are unlikely to engage in serial retrieval skills, as they are still committed in training grapheme-phoneme conversion rules. It is thus possible that, at low-level reading performance, arithmetic shares more similarity with RAN than with reading.

Limitations

Our analysis consisted of 27 multivariate regression models (three main dependent variables investigated over three time points and three levels of performance). Given that the likelihood of committing Type I errors increases with the number of outcome variables investigated, we cannot rule out the possibility that some of our significant results were actually driven by chance. All regression coefficients and the corresponding *p*-values are reported in the tables of regression results. We invite the reader to interpret the most scattered findings with caution, because these may not be robust. Beside p-values, it is important to consider the magnitude of the effects and whether these will be replicated by future studies with a comparable design.

Due to power constraints related to the sample size, only a limited number of domain-general dimensions (nonverbal intelligence and working memory) were included in the current study. Research evidence suggests that other domain-general dimensions as for instance different executive functions explain variance in reading, arithmetic and their overlap (Balhinez and Shaul, 2019; Koponen et al., 2020; Zoccolotti et al., 2020). Note, however, that these domain-general predictors were shown to be indirectly related to the outcome variables (Koponen et al., 2020) and their contribution was smaller than the one of reading- and arithmetic-specific predictors (Moll et al., 2014; Peterson et al., 2017; Cirino et al., 2018; Koponen et al., 2020; Malone et al., 2020). Nevertheless, literature on the role and importance of these dimensions to reading, arithmetic and their overlap is scattered. Future studies with large sample sizes should systematically test the contribution of these dimensions.

Conclusion

The current study based on an analysis of longitudinal data highlights that dimensions contributing to reading and arithmetic fluency have a similar predictive role across time points and performance levels. This indicates that the prediction in the low-performance range does not differ substantially from the prediction in the typical performance range. One relevant exception to this trend is counting fluency, which may serve compensatory calculation strategies for children with low arithmetic fluency in Grade 3.

Dimensions that explain shared variance in reading and arithmetic fluency or that make specific cross-domain contributions relate to retrieval efficiency, language/symbolic processing and cross-format mapping, skills that are necessary for successful integration of visual, verbal and semantic information relevant for both reading and arithmetic fluency.

Practical implications

The existence of a general linear prediction pattern suggests that the investigated cognitive dimensions are relevant for children with very different attainment in reading and arithmetic fluency. The results of the present study therefore highlight the benefits of early assessment of literacy and numeracy skills, as in currently available screening tools [see for example Jöbstl et al., 2022]. Having shown that cognitive dimensions in Grade 1 predict both high and low levels of performance later on, risk factors for reading and arithmetic difficulties can be identified at an early stage, which in turn makes it possible to design support programs that focus on strengthening below-average skills. Similarly, the early identification of skilled cognitive resources enables to target children with exceptionally high literacy and arithmetic skills and support them appropriately throughout formal schooling.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: https://reshare.ukdataservice.ac.uk/854335/.

Ethics statement

The studies involving humans were approved by the ethics committees of the Universities of Graz and York. The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation in this study was provided by the participants' legal guardians/next of kin.

Author contributions

CB: Data curation, Formal analysis, Investigation, Methodology, Visualization, Writing – original draft. VJ: Data curation, Formal analysis, Investigation, Methodology, Visualization, Writing – original draft. SG: Conceptualization, Funding acquisition, Investigation, Project administration, Resources, Supervision, Writing – review & editing. KL: Conceptualization, Funding acquisition, Investigation, Project administration, Resources, Supervision, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This research was funded in part by the Austrian Science Fund (FWF; I 2778-G16) and by the Economic and Social Research Council (ESRC; ES/N014677/1, ES/W002914/1). For the purpose of open access, the authors have applied a CC BY public copyright license to any author accepted manuscript version arising from this submission.

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/feduc.2024.1335957/full#supplementary-material

References

Amland, T., Lervåg, A., and Melby-Lervåg, M. (2021). Comorbidity between math and reading problems: is phonological processing a mutual factor? *Front. Hum. Neurosci.* 14:e577304. doi: 10.3389/fnhum.2020.577304

Astle, D. E., and Fletcher-Watson, S. (2020). Beyond the core-deficit hypothesis in developmental disorders. *Curr. Dir. Psychol. Sci.* 29, 431–437. doi: 10.1177/0963721420925518

Baddeley, A. D. (2012). Working memory: theories, models, and controversies. *Annu. Rev. Psychol.* 63, 1–29. doi: 10.1146/annurev-psych-120710-100422

Balhinez, R., and Shaul, S. (2019). The relationship between reading fluency and arithmetic fact fluency and their shared cognitive skills: a developmental perspective. *Front. Psychol.* 10:1281. doi: 10.3389/fpsyg.2019.01281

Banfi, C., Clayton, F. J., Steiner, A. F., Finke, S., Kemény, F., Landerl, K., et al. (2022). Transcoding counts: longitudinal contribution of number writing to arithmetic in different languages. *J. Exp. Child Psychol.* 223:105482. doi: 10.1016/j.jecp.2022.105482

Barrouillet, P., Camos, V., Perruchet, P., and Seron, X. (2004). ADAPT: a developmental, asemantic, and procedural model for transcoding from verbal to Arabic numerals. *Psychol. Rev.* 111, 368–394. doi: 10.1037/0033-295X.111.2.368

Bartelet, D., Vaessen, A., Blomert, L., and Ansari, D. (2014). What basic number processing measures in kindergarten explain unique variability in first-grade arithmetic proficiency? *J. Exp. Child Psychol.* 117, 12–28. doi: 10.1016/j.jecp.2013.08.010

Bernabini, L., Bonifacci, P., and de Jong, P. F. (2021). The relationship of reading abilities with the underlying cognitive skills of math: a dimensional approach. *Front. Psychol.* 12:577488. doi: 10.3389/fpsyg.2021.577488

Caravolas, M., Lervåg, A., Mousikou, P., Efrim, C., Litavský, M., Onochie-Quintanilla, E., et al. (2012). Common patterns of prediction of literacy development in different alphabetic orthographies. *Psychol. Sci.* 23, 678–686. doi: 10.1177/0956797611434536

Chambers, J. M., and Hastie, T. J. (1992). *Statistical models in S*. New York: Chapman & Hall.

Cheung, P., and Ansari, D. (2021). Cracking the code of place value: the relationship between place and value takes years to master. *Dev. Psychol.* 57, 227–240. doi: 10.1037/dev0001145

Child, A. E., Cirino, P. T., Fletcher, J. M., Willcutt, E. G., and Fuchs, L. S. (2019). A cognitive dimensional approach to understanding shared and unique contributions to Reading, math, and attention skills. *J. Learn. Disabil.* 52, 15–30. doi: 10.1177/0022219418775115

Cipora, K., He, Y., and Nuerk, H. C. (2020). The spatial–numerical association of response codes effect and math skills: why related? *Ann. N. Y. Acad. Sci.* 1477, 5–19. doi: 10.1111/nyas.14355

Cirino, P. T., Child, A. E., and Macdonald, K. T. (2018). Longitudinal predictors of the overlap between reading and math skills. *Contemp. Educ. Psychol.* 54, 99–111. doi: 10.1016/j.cedpsych.2018.06.002

Cirino, P. T., Fuchs, L. S., Elias, J. T., Powell, S. R., and Schumacher, R. F. (2015). Cognitive and mathematical profiles for different forms of learning difficulties. *J. Learn. Disabil.* 48, 156–175. doi: 10.1177/0022219413494239

Clayton, F. J., Copper, C., Steiner, A. F., Banfi, C., Finke, S., Landerl, K., et al. (2020). Two-digit number writing and arithmetic in year 1 children: does number word inversion matter? *Cogn. Dev.* 56, 1–14. doi: 10.1016/j.cogdev.2020.100967

Coltheart, M., Rastle, K., Perry, C., Langdon, R., and Ziegler, J. (2001). DRC: a dual route cascaded model of visual word recognition and reading aloud. *Psychol. Rev.* 108:204. doi: 10.1037//0033-295X.108.1.204

de Groot, B. J. A., van den Bos, K. P., Minnaert, A. E. M. G., and van der Meulen, B. F. (2015). Phonological processing and word reading in typically developing and reading disabled children: severity matters. *Sci. Stud. Read.* 19, 166–181. doi: 10.1080/10888438.2014.973028

De Visscher, A., and Noël, M. P. (2013). A case study of arithmetic facts dyscalculia caused by a hypersensitivity-to-interference in memory. *Cortex* 49, 50–70. doi: 10.1016/j.cortex.2012.01.003

Dehaene, S. (1992). Varieties of numerical abilities. *Cognition* 44, 1–42,

Department for Communities and Local Government (2015). The English indices of deprivation 2015 [statistical release]. Available at: https://www.gov.uk/government/statistics/english-indices-of-deprivation-2015.

Devlin, B. L., Jordan, N. C., and Klein, A. (2022). Predicting mathematics achievement from subdomains of early number competence: differences by grade and achievement level. *J. Exp. Child Psychol.* 217:105354. doi: 10.1016/j.jecp.2021.105354

Dirks, E., Spyer, G., Van Lieshout, E. C. D. M., and De Sonneville, L. (2008). Prevalence of combined reading and arithmetic disabilities. *J. Learn. Disabil.* 41, 460–473. doi: 10.1177/0022219408321128

Durand, M., Hulme, C., Larkin, R., and Snowling, M. (2005). The cognitive foundations of reading and arithmetic skills in 7- to 10-year-olds. *J. Exp. Child Psychol.* 91, 113–136. doi: 10.1016/j.jecp.2005.01.003

Ehri, L. C. (2005). Learning to read words: theory, findings, and issues. *Sci. Stud. Read.* 9, 167–188. doi: 10.1207/s1532799xssr0902_4

Ehri, L. C. (2014). Orthographic mapping in the acquisition of sight word reading, spelling memory, and vocabulary learning. *Sci. Stud. Read.* 18, 5–21. doi: 10.1080/10888438.2013.819356

Frith, U. (1985). "Beneath the surface of developmental dyslexia" in *Surface dyslexia: Neuropsychological and cognitive studies of phonological reading*. eds. K. Patterson, J. C. Marshall and M. Coltheart (Erlbaum), 301–330.

Fuchs, L. S., Geary, D. C., Fuchs, D., Compton, D. L., and Hamlett, C. L. (2016). Pathways to third-grade calculation versus word-reading competence: are they more alike or different? *Child Dev.* 87, 558–567. doi: 10.1111/cdev.12474

Galuschka, K., and Schulte-Körne, G. (2016). The diagnosis and treatment of reading and / or spelling disorders in children and adolescents. *Dtsch. Arztebl. Int.* 113, 279–286. doi: 10.3238/arztebl.2016.0279

Geary, D. C. (2011). Cognitive predictors of achievement growth in mathematics. *Dev. Psychol.* 47, 1539–1552. doi: 10.1037/a0025510.Cognitive

Geary, D. C., and VanMarle, K. (2018). Growth of symbolic number knowledge accelerates after children understand cardinality. *Cognition* 177, 69–78. doi: 10.1016/j.cognition.2018.04.002

Georgiou, G. K., Parrila, R., Cui, Y., and Papadopoulos, T. C. (2013). Why is rapid automatized naming related to reading? *J. Exp. Child Psychol.* 115, 218–225. doi: 10.1016/j.jecp.2012.10.015

Georgiou, G. K., Wei, W., Inoue, T., and Deng, C. (2020). Are the relations of rapid automatized naming with reading and mathematics accuracy and fluency bidirectional? Evidence from a 5-year longitudinal study with Chinese children. *J. Educ. Psychol.* 112, 1506–1520. doi: 10.1037/edu0000452

Göbel, S. M., Watson, S. E., Lervåg, A., and Hulme, C. (2014). Children's arithmetic development: it is number knowledge, not the approximate number sense, that counts. *Psychol. Sci.* 25, 789–798. doi: 10.1177/0956797613516471

Göbel, S., Wesierska, M., and Landerl, K. (2020). *Three-hundred-and-twenty-eight and 328: Cross-format number integration and its relationship to mathematics performance 2017–2020 [data set]*.

Habermann, S., Donlan, C., Göbel, S. M., and Hulme, C. (2020). The critical role of Arabic numeral knowledge as a longitudinal predictor of arithmetic development. *J. Exp. Child Psychol.* 193:104794. doi: 10.1016/j.jecp.2019.104794

Haberstroh, S., and Schulte-Körne, G. (2019). The diagnosis and treatment of dyscalculia. *Deutsches Arzteblatt Int.* 116, 107–114. doi: 10.3238/arztebl.2019.0107

Jöbstl, V., Banfi, C., Göbel, S. M., and Landerl, K. (2024). Cognitive predictors of reading-arithmetic associations and dissociations: A longitudinal study from Grade 1 to Grade 3. [Manuscript submitted for publication].

Jöbstl, V., Steiner, A. F., Deimann, P., Kastner-Koller, U., and Landerl, K. (2023). A-B-3-associations and dissociations of reading and arithmetic: is domain-specific prediction outdated? *PLoS One* 18:e0285437. doi: 10.1371/journal.pone.0285437

Jöbstl, V., Steiner, A. F., Kastner-koller, U., Deimann, P., Kaltenberger, A., Wagner, V., et al. (2022). Entwicklung eines förderorientierten Schuleingangsscreenings. Erste Befunde zur prognostischen Validität. [development of support-oriented school entrance screening. First results about prognostic validity]. *Frühe Bildung* 11, 194–200. doi: 10.1026/2191-9186/a000585

Joyner, R. E., and Wagner, R. K. (2020). Co-occurrence of reading disabilities and math disabilities: a meta-analysis. *Sci. Stud. Read.* 24, 14–22. doi: 10.1080/10888438.2019.1593420

Kirby, J. R., Georgiou, G. K., Martinussen, R., and Parrila, R. (2010). Naming speed and reading: from prediction to instruction. *Read. Res. Q.* 45, 341–362. doi: 10.1598/RRQ.45.3.4

Koenker, R., and Machado, J. A. F. (1999). Goodness of fit and related inference processes for quantile regression. *J. Am. Stat. Assoc.* 94, 1296–1310,

Koponen, T., Aro, M., Poikkeus, A. M., Niemi, P., Lerkkanen, M. K., Ahonen, T., et al. (2018). Comorbid fluency difficulties in reading and math: longitudinal stability across early grades. *Except. Child.* 84, 298–311. doi: 10.1177/0014402918756269

Koponen, T., Aunola, K., Ahonen, T., and Nurmi, J. E. (2007). Cognitive predictors of single-digit and procedural calculation skills and their covariation with reading skill. *J. Exp. Child Psychol.* 97, 220–241. doi: 10.1016/j.jecp.2007.03.001

Koponen, T., Eklund, K., Heikkilä, R., Salminen, J., Fuchs, L., Fuchs, D., et al. (2020). Cognitive correlates of the covariance in reading and arithmetic fluency: importance of serial retrieval fluency. *Child Dev.* 91, 1063–1080. doi: 10.1111/cdev.13287

Koponen, T., Salmi, P., Eklund, K., and Aro, T. (2013). Counting and RAN: predictors of arithmetic calculation and reading fluency. *J. Educ. Psychol.* 105, 162–175. doi: 10.1037/a0029285

Korpipää, H., Koponen, T., Aro, M., Tolvanen, A., Aunola, K., Poikkeus, A. M., et al. (2017). Covariation between reading and arithmetic skills from grade 1 to grade 7. *Contemp. Educ. Psychol.* 51, 131–140. doi: 10.1016/j.cedpsych.2017.06.005

Kucian, K., and von Aster, M. (2015). Developmental dyscalculia. *Eur. J. Pediatr.* 174, 1–13. doi: 10.1007/s00431-014-2455-7

Landerl, K., Castles, A., and Parrila, R. (2022). Cognitive precursors of reading: a cross-linguistic perspective. *Sci. Stud. Read.* 26, 111–124. doi: 10.1080/10888438.2021.1983820

Landerl, K., Freudenthaler, H. H., Heene, M., De Jong, P. F., Desrochers, A., Manolitsis, G., et al. (2019). Phonological awareness and rapid automatized naming as longitudinal predictors of reading in five alphabetic orthographies with varying degrees of consistency. *Sci. Stud. Read.* 23, 220–234. doi: 10.1080/10888438.2018.1510936

Landerl, K., Fussenegger, B., Moll, K., and Willburger, E. (2009). Dyslexia and dyscalculia: two learning disorders with different cognitive profiles. *J. Exp. Child Psychol.* 103, 309–324. doi: 10.1016/j.jecp.2009.03.006

Landerl, K., Göbel, S. M., and Moll, K. (2013a). Core deficit and individual manifestations of developmental dyscalculia (DD): the role of comorbidity. *Trends Neurosci. Educ.* 2, 38–42. doi: 10.1016/j.tine.2013.06.002

Landerl, K., and Moll, K. (2010). Comorbidity of learning disorders: prevalence and familial transmission. *J. Child Psychol. Psychiatry Allied Discip.* 51, 287–294. doi: 10.1111/j.1469-7610.2009.02164.x

Landerl, K., Ramus, F., Moll, K., Lyytinen, H., Leppänen, P. H. T., Lohvansuu, K., et al. (2013b). Predictors of developmental dyslexia in European orthographies with varying complexity. *J. Child Psychol. Psychiatry Allied Discip.* 54, 686–694. doi: 10.1111/jcpp.12029

Levesque, K. C., Breadmore, H. L., and Deacon, S. H. (2021). How morphology impacts reading and spelling: advancing the role of morphology in models of literacy development. *J. Res. Read.* 44, 10–26. doi: 10.1111/1467-9817.12313

Malone, S. A., Burgoyne, K., and Hulme, C. (2020). Number knowledge and the approximate number system are two critical foundations for early arithmetic development. *J. Educ. Psychol.* 112, 1167–1182. doi: 10.1037/edu0000426

McGrath, L. M., Peterson, R. L., and Pennington, B. F. (2020). The multiple deficit model: Progress, problems, and prospects. *Sci. Stud. Read.* 24, 7–13. doi: 10.1080/10888438.2019.1706180

McIlraith, A. L. (2018). Predicting word reading ability: a quantile regression study. *J. Res. Read.* 41, 79–96. doi: 10.1111/1467-9817.12089

Moll, K., Fussenegger, B., Willburger, E., and Landerl, K. (2009). RAN is not a measure of orthographic processing. Evidence from the asymmetric German orthography. *Sci. Stud. Read.* 13, 1–25. doi: 10.1080/10888430802631684

Moll, K., Göbel, S. M., and Snowling, M. J. (2015). Basic number processing in children with specific learning disorders: comorbidity of reading and mathematics disorders. *Child Neuropsychol.* 21, 399–417. doi: 10.1080/09297049.2014.899570

Moll, K., Kunze, S., Neuhoff, N., Bruder, J., and Schulte-Körne, G. (2014). Specific learning disorder: prevalence and gender differences. *PLoS One* 9:e103537. doi: 10.1371/journal.pone.0103537

Moll, K., and Landerl, K. (2010). *SLRT-II: Lese-und Rechtschreibtest [SLRT-II: Reading and spelling test]*: Huber.

Moll, K., Ramus, F., Bartling, J., Bruder, J., Kunze, S., Neuhoff, N., et al. (2014). Cognitive mechanisms underlying reading and spelling development in five European orthographies. *Learn. Instr.* 29, 65–77. doi: 10.1016/j.learninstruc.2013.09.003

Ozernov-Palchik, O., Sideridis, G. D., Norton, E. S., Beach, S. D., Wolf, M., Gabrieli, J. D. E., et al. (2022). On the cusp of predictability: disruption in the typical association between letter and word identification at critical thresholds of RAN and phonological skills. *Learn. Individ. Differ.* 97:102166. doi: 10.1016/j.lindif.2022.102166

Pennington, B. F. (2006). From single to multiple models of developmental disorders. *Cognition* 101, 385–413. doi: 10.1016/j.cognition.2006.04.008

Perfetti, C. A., and Hart, L. (2001). "The lexical quality hypothesis" in *Precursors of functional literacy2*. eds. L. Verhoeven, C. Elbro and P. Reitsma (Amsterdam: John Benjamins Publishing Company), 189–213.

Peters, L., and Ansari, D. (2019). Are specific learning disorders truly specific, and are they disorders? *Trends in Neuroscience and Education* 17:100115. doi: 10.1016/j.tine.2019.100115

Peters, L., and De Smedt, B. (2018). Arithmetic in the developing brain: a review of brain imaging studies. *Dev. Cogn. Neurosci.* 30, 265–279. doi: 10.1016/j.dcn.2017.05.002

Peterson, R. L., Boada, R., Mcgrath, L. M., Willcutt, E. G., Olson, R. K., and Pennington, B. F. (2017). Cognitive prediction of reading, math, and attention: shared and unique influences. *J. Learn. Disabil.* 50, 408–421. doi: 10.1126/science.1249098.Sleep

Pickering, S., and Gathercole, S. E. (2001). *Working memory test battery for children (WMTB-C)*: Psychological Corporation.

Protopapas, A., and Parrila, R. (2018). Is dyslexia a brain disorder? *Brain Sci.* 8, 1–18. doi: 10.3390/brainsci8040061

Raddatz, J., Kuhn, J. T., Holling, H., Moll, K., and Dobel, C. (2017). Comorbidity of arithmetic and Reading disorder: basic number processing and calculation in children with learning impairments. *J. Learn. Disabil.* 50, 298–308. doi: 10.1177/0022219415620899

Raven, J., Raven, J. C., and Court, J. H. (1998). *Raven's standard progressive matrices and vocabulary scales.* Oxford (UK): Oxford Psychologists Press.

R Core Team (2024). *R: A language and environment for statistical computing_*. Vienna, Austria: R Foundation for Statistical Computing. Available at: https://www.R-project.org/.

Siegel, L. S., and Heaven, R. K. (1986). "Categorization of learning disabilities" in *Handbook of cognitive, social and neuropsychological aspects of learning disabilities* (Routledge), 95–122.

Snowling, M. J., Stothard, S. E., Clarke, P., Bowyer-Crane, C., Harrington, A., Truelove, E., et al., (2009). *York assessment of Reading for comprehension (YARC).* GL Assessment.

Starr, A., Libertus, M. E., and Brannon, E. M. (2013). Number sense in infancy predicts mathematical abilities in childhood. *Proc. Natl. Acad. Sci. USA* 110, 18116–18120. doi: 10.1073/pnas.1302751110

Szucs, D., Devine, A., Soltesz, F., Nobes, A., and Gabriel, F. (2013). Developmental dyscalculia is related to visuo-spatial memory and inhibition impairment. *Cortex* 49, 2674–2688. doi: 10.1016/j.cortex.2013.06.007

Torgesen, J. K., Wagner, R. K., and Rashotte, C. A. (2012). *Test of word reading efficiency – Second edition (TOWRE-2)*. Pro-Ed.

Träff, U., Skagerlund, K., Östergren, R., and Skagenholt, M. (2023). The importance of domain-specific number abilities and domain-general cognitive abilities for early arithmetic achievement and development. *Br. J. Educ. Psychol.* 93, 825–841, e12599. doi: 10.1111/bjep.12599

Vanbinst, K., van Bergen, E., Ghesquière, P., and De Smedt, B. (2020). Cross-domain associations of key cognitive correlates of early reading and early arithmetic in 5-year-olds. *Early Child. Res. Q.* 51, 144–152. doi: 10.1016/j.ecresq.2019.10.009

Wagner, R. K., Torgesen, J. K., Rashotte, C. A., and Pearson, N. A. (2013). *Comprehensive test of phonological processing 2nd ed. (CTOPP-2). Pro-Ed.*

Willburger, E., Fussenegger, B., Moll, K., Wood, G., and Landerl, K. (2008). Naming speed in dyslexia and dyscalculia. *Learn. Individ. Differ.* 18, 224–236. doi: 10.1016/j.lindif.2008.01.003

Willcutt, E. G., McGrath, L. M., Pennington, B. F., Keenan, J. M., DeFries, J. C., Olson, R. K., et al. (2019). "Understanding comorbidity between specific learning disabilities" in *Models for innovation: Advancing approaches to Higher-Rick and Higher-impact Learnig disabilities science. New directions for Child and adolescent development*. eds. L. S. Fuchs and D. L. Compton, vol. 165, 91–109.

Willcutt, E. G., Petrill, S. A., Wu, S., Boada, R., DeFries, J. C., Olson, R. K., et al. (2013). Comorbidity between reading disability and math disability: concurrent psychopathology, functional impairment, and neuropsychological functioning. *J. Learn. Disabil.* 46, 500–516. doi: 10.1038/jid.2014.371

Zoccolotti, P., Angelelli, P., Marinelli, C. V., and Romano, D. L. (2021). A network analysis of the relationship among reading, spelling and maths skills. *Brain Sci.* 11:656. doi: 10.3390/brainsci11050656

Zoccolotti, P., De Luca, M., Marinelli, C. V., and Spinelli, D. (2020). Predicting individual differences in reading, spelling and maths in a sample of typically developing children: a study in the perspective of comorbidity. *PLoS One* 15:e0231937. doi: 10.1371/journal.pone.0231937

Frontiers in Education

TYPE Original Research
PUBLISHED 15 May 2024
DOI 10.3389/feduc.2024.1340322

Check for updates

OPEN ACCESS

EDITED BY
Natalie Förster,
University of Wuppertal, Germany

REVIEWED BY
Boris Forthmann,
University of Münster, Germany
Lukas Baumanns,
Technical University Dortmund, Germany

*CORRESPONDENCE
Stefan Ufer
ufer@math.lmu.de

RECEIVED 17 November 2023
ACCEPTED 09 April 2024
PUBLISHED 15 May 2024

CITATION
Ufer S, Kaiser A, Niklas F and Gabler L (2024) I have three more than you, you have three less than me? Levels of flexibility in dealing with additive situations.
Front. Educ. 9:1340322.
doi: 10.3389/feduc.2024.1340322

I have three more than you, you have three less than me? Levels of flexibility in dealing with additive situations

Stefan Ufer[1]*, Anna Kaiser[1], Frank Niklas[2] and Laura Gabler[1]

[1]Department of Mathematics, LMU Munich, Munich, Germany, [2]Department of Psychology, LMU Munich, Munich, Germany

Assessment and intervention in the early years should ideally be based on evidence-based models describing the structure and development of students' skills. Mathematical word problems have been identified as a challenge for mathematics learners for a long time and in many countries. We investigate flexibility in dealing with additive situations as a construct that develops during grades 1 through 3 and contributes to the development of students' word problem solving skills. We introduce the construct based on prior research on the difficulty of different situation structures entailed in word problems. We use data from three prior empirical studies with $N = 383$ German grade 2 and 3 students to develop a model of discrete levels of students' flexibility in dealing with additive situations. We use this model to investigate how the learners in our sample distribute across the different levels. Moreover, we apply it to describe students' development over several weeks in one study comprising three measurements. We derive conclusions about the construct in terms of determinants of task complexity, and about students' development and then provide an outlook on potential uses of the model in research and practice.

KEYWORDS

flexibility in dealing with additive situations, level model, mathematics, word problem solving, primary school, assessment, assessment-based intervention

1 Introduction

Mathematics instruction at school not only aims at conveying mathematical concepts and procedures, but also at students' skills in mathematical modelling, which means to apply these concepts and procedures in more or less realistic real-life situations (Cevikbas et al., 2022). Word problems are mathematical tasks, which are presented in verbal form and embedded in a short narrative, e.g., "Chris has 4 marbles. Chris has 3 marbles less than Alex. How many marbles does Alex have? "They constitute a standard part of school curricula world-wide (e.g., Verschaffel et al., 2020; Krawitz et al., 2022), where they serve two main purposes: They can be seen as very basic exercises in mathematical modelling, but more importantly they allow to engage students with relationships between mathematical concepts and real-world phenomena they can describe (Freudenthal, 1983; Verschaffel et al., 2020). In grade 1 and early grade 2, most of the encountered word problems require addition or subtraction since other operations are often not introduced before mid-grade 2. We focus on additive one-step word problems in this contribution, since they are frequent in mathematics text books (Gabler et al., 2023) and research (Verschaffel et al., 2020). These word problems can be solved with a single additive arithmetic operation (addition or subtraction) and contain no irrelevant information. They are part of primary school curricula from grade 1 onwards.

It is a long-standing result that solving word problems poses a substantial challenge for mathematics learners in primary school and beyond (Greeno, 1980; Verschaffel et al., 1992; Daroczy et al., 2015, 2020). Further, it is also well-established that these problems are modulated by a range of individual characteristics such as general cognitive abilities, language skills, and arithmetic (symbolic) calculations skills as well as by linguistic or mathematical task characteristics (Daroczy et al., 2015). In addition, strategies to solve word problems have received substantial attention in prior research. Based on traditional models of word problem solving (e.g., Kintsch and Greeno, 1985; Blum and Leiß, 2007), some authors have proposed that, after initially reading a difficult word problem, learners may be able to reinterpret the presented situation in a way, that results in an easier word problem type (Greeno, 1980; Stern, 1993). Recent research has found that primary school students differ systematically in their ability to reinterpret situations presented in additive one-step word problems in the proposed way (Gabler and Ufer, 2021, 2022). The skill to engage in such re-interpretations has been studied as *flexibility in dealing with additive situations* (Gabler and Ufer, 2021, 2022). To provide a conceptual basis for assessments and interventions in the early school years, we propose a level model of this skill construct, that is based on a re-analysis involving an IRT-scaling of data from three empirical studies. Such level models structure a one-dimensional skill construct into a sequence of discrete levels, that describe increasing demands and item complexity associated with the skill construct. They may also allow a criterial interpretation of students' individual skills in terms of the demands they can or cannot master (Koeppen et al., 2008; Ufer and Neumann, 2018). Further, they may provide an important link from assessing the skill towards interventions building on the students' current skill levels and targeting more complex, but still within-reach demands.

To introduce the flexibility skill construct and our study, we will first discuss models of the word problem solving process and task characteristics influencing students' performance on additive one-step word problems. Based on this, we will introduce the flexibility skill construct and present methodological backgrounds on level models of mathematical skill constructs based on item response theory.

1.1 Word problem solving

Many established models of the word problem solving process are transformative in nature (Czocher, 2018): They assume that learners transform an initial *situation model* of the presented real-world situation into a *mathematical model*, and after solving the problem entailed in this mathematical model, they transform their solution back to their initial situation model (*cf.* Kintsch and Greeno, 1985; Verschaffel et al., 2020). The situation model comprises the learner's mental representation of the textual presentation of the situation in the word problem, the *text base* (Figure 1). For additive one-step word problems, a situation model should at least contain the two quantities given in the word problem (number of Chris' marbles, and the difference between the number of Chris' and Alex's marbles), the related numbers (4 and 3), an unknown quantity (number of Alex's marbles), and a relation between the three quantities as described in the problem text (direction of the difference: Chris has 3 less than Alex). According to models of reading comprehension (e.g., Kintsch, 1998), (re-)constructing this situation model from the text base is not limited to decoding and representing the presented verbal information, but may also comprise making further inferences about the situation based on students' knowledge about the context presented in the situation itself (e.g., exchanging marbles, or shopping), or about typical situation structures that can be described mathematically by addition or subtraction. The reinterpretation strategies mentioned above (Greeno, 1980; Stern, 1993) are examples of such further inferences.

The mathematical model contains the numbers from the situation model, and the relation between them in terms of a mathematical operation (addition or subtraction). It may have the form of an operation that directly provides the number related to the unknown quantity ($7-3=$___), or of an implicit characterization of this number (e.g., ___$+3=7$) like a mathematical equation. Based on knowledge about arithmetic operations, learners may transform a mathematical model into an equivalent one in the further solution process, e.g., by solving the subtraction problem $7-3=$___ by the indirect addition $3+$___$=7$ (Torbeyns et al., 2009).

Even though more extensive models, for example the modelling cycle (Blum and Leiß, 2007), comprise further phases such as interpreting and validating the results, these phases may not be necessary and can be omitted in simple situations such as one-step word problems (Kaiser, 2017). It is well-known that learners do not necessarily follow the model of the word problem solving process. Hegarty et al. (1992, 1995) provide evidence of a *direct translation strategy* (also *keyword strategy*), in which students focus primarily on the numbers and specific terms in the text base, e.g., relational terms such as "more" or "less," or actions such as "getting" or "losing" to directly infer the necessary mathematical operation. They offer convergent evidence from behavioral, eye-tracking and problem-recall studies, that less successful word problem solvers frequently apply this strategy, resulting in difficulties extracting and representing relational information from the text. In the literature, the importance of rich situation models, "in which all key elements and relations in the problem situation that are relevant to the solution of the mathematical problem posed are represented" is highlighted repeatedly (e. g., Mellone et al., 2017, p. 3). Accordingly, many authors call to "provide instruction in a method that emphasizes understanding the situation described in the problem" (Hegarty et al., 1995, p. 29), for example by conveying strategies to build up and make use of rich situation models. They also point out, that applying this strategy is not a stable person characteristic but may depend on person and task characteristics.

1.2 Task characteristics

Several task characteristics have been found to influence word problem solution rates. In their review, Daroczy et al. (2015) distinguish between mathematical factors such as the complexity of the entailed numerical calculation, and linguistic factors, such as the structure of the presented situation or the way it is presented verbally. In a sample of grade 2 students, and in a restricted number range up to 20, Gabler and Ufer (2020, p. 77, 79) could not find systematic differences in word problem difficulty related to the specific numbers (or contexts) used. Consequently, we will focus on three linguistic

10.3389/feduc.2024.1340322

	Static situation	Dynamic situation
Part-whole situation	*Combine* Chris has 4 marbles. Alex has 3 marbles. How many marbles do Alex and Chris have together?	*Change* Chris had 4 marbles. Then, Chris got 3 marbles from Alex. How many does Chris have now?
Situation with disjoint sets	*Compare* Chris has 4 marbles. Chris has 3 marbles less than Alex. How many marbles does Alex have?	*Equalize* Chris has 4 marbles. If Chris gets 3 more marbles from his parents, Chris and Alex have the same number of marbles. How many marbles does Alex have?

4 + 3 = 7

FIGURE 1
Different semantic structures relating to the same mathematical structure.

factors that describe the *situation structure* presented in the word problem: the *semantic structure*, which describes the structure of the presented situation, the *unknown set*, that needs to be determined in the word problem, and the *additive* vs. *subtractive wording* of the word problem, which relates to additive ("more," "getting") vs. subtractive ("less," "loosing") terms used to describe the situation.

1.2.1 Semantic structure

A range of different real-world phenomena can be described by the same mathematical model (e.g., an additive operation such as $4+3=7$, Figure 1). Up to four different types of so-called semantic structures in additive one-step word problems have been differentiated (Riley et al., 1983; Figure 1). While word problems can also refer to other quantities, these problem types are usually exemplified with word problems relating to the numbers of objects in different sets. *Change structures* refer to an increase or decrease of a set of objects. *Combine structures* relate to part-whole structures between a set, and two subsets that together make up the whole set. *Compare structures* comprise two disjoint quantities, and a relational statement about their difference. *Equalize structures* have been studied less frequently. Like compare structures, they contain two disjoint sets (Chris's and Alex's marbles), but the relation between the two sets is presented by an action—as in change structures—that would (hypothetically) equalize the two sets (someone giving Chris 3 more marbles). Change and equalize structures are called *dynamic*, because they entail an action, while compare and combine structures are called *static* (Riley et al., 1983).

Word problems with different semantic structures have been found to be of systematically different difficulty. Numerous studies (e.g., Riley and Greeno, 1988) show that compare problems are more difficult than change and combine problems. Few studies have investigated equalize problems. However, Stern (1994) reported high solution rates (96%) for equalize problems, similar to those for change and combination problems, in a sample of first graders. Several reasons for this specific difficulty of compare problems have been discussed. Stern (1993) points out that in combine and change problems, all sets exist as *concrete* sets, that are observable separate quantities or in terms of an observable action. The difference in compare problems, in contrast does not describe a concrete set, but a relation between two concrete sets (a *difference set*), that is only observable when considering both concrete sets. This can be done by setting up a one-to-one correspondence between one concrete set and a subset of the other concrete set, and then counting the excess objects (Stern, 1998). Representing relational statements as difference sets may pose a substantial problem for learners. Here, it makes a substantial difference, if quantitative (i.e., described by numbers, e.g., three more than) differences are describe, or if only qualitative differences (more than, less than) are considered. For example, some learners understand quantitative relational statements such as "Chris has 3 marbles more than Alex" as being equivalent to "Chris has 3 marbles, and Chris has more marbles than Alex" – interpreting the quantitative relation as a qualitative relation (more than) and a statement about a concrete set (Mekhmandarov et al., 1996; Gabler and Ufer, 2021). In line with these difficulties, understanding numbers as quantitative comparisons between sets is allocated to later phases in models of the development of the number concept (under the term "relationality"; for an overview see Hartmann and Fritz, 2021).

1.2.2 Unknown set

Three different sets are involved in one-step word problems. In change, compare and equalize problems, one set provides a *reference*: This can be the start of a change of equalize action, or the set to which another set is compared (e.g., "Chris has 4 marbles. Chris has 3 marbles less than Alex. *How many marbles does Alex have*?"). One set serves as a *result*, which can be the result of the change or equalize action, or the set that is compared to another one (e.g., "Alex has 7 marbles. Chris has 3 marbles less than Alex. *How many marbles does Chris have*?"). The third set describes the *relation* between the other two sets as a concrete set in the change or equalize actions, or as a difference set in compare statements (e.g., "Alex has 7 marbles. Chris has 4 marbles. *How many marbles does Chris have less than Alex?*"). The situation is different for combine problems, where two parts, which play similar but complementary roles in the situation, and the whole set need to be distinguished.

There is evidence that, at least for young learners, word problems with unknown reference or relation set are more difficult than those with unknown result set (Gabler and Ufer, 2020; Van Lieshout and Xenidou-Dervou, 2020). One reason may be that the learners' standard mathematical model related to the situation structures is of the form *<reference><operation (+/−)><relation>=<result>*. This would imply, that the unknown result set situations directly provide all numerical information to perform the arithmetic operation entailed in the model. For unknown reference and relation sets, this standard model would result in an implicit characterization of the required numerical solution (e.g., *<unkown reference>+<known relation>=<known result>*). In the second case, either the implicit problem needs to be solved directly (e.g., by fact retrieval or trial and error) or it needs to be transferred into an equivalent, directly solvable mathematical model (e.g., *<unkown reference>=<known result>−<known relation>*). Yet, the exact reasons for the observed difficulty pattern are still to be clarified.

1.2.3 Additive vs. subtractive wording and consistency

In change, equalize, and compare word problems, the change and relations can be expressed additively ("more," "getting": additive wording) or subtractively ("less," "loosing": subtractive wording) in the problem text. If students apply the direct translation strategy (Hegarty et al., 1995), they will use addition of the two given numbers as mathematical model in case of additive wording and subtraction of the two numbers in case of subtractive wording. For some word problems (e.g., those with unknown result set) this results in a correct mathematical model, for others (e.g., those with unknown reference set) it leads to a wrong model. For problems with unknown relation set, a correct mathematical model results only for subtractive wording. Word problems are called *consistent*, if the direct translation strategy results in a (addition/subtraction), that correct mathematical model, i.e., if the wording of the word problem (additive/subtractive) reflects the operation (addition/subtraction) that can be applied to the given numbers directly, to obtain a valid mathematical model. Consistent word problems have been found to be easier than inconsistent ones (Lewis and Mayer, 1987, "consistency hypothesis") for primary school students (Verschaffel, 1994; Gabler and Ufer, 2020) and adults (Daroczy et al., 2020). This means, for example, that the inconsistent word problem "Chris has 4 marbles. *Chris has 3 marbles less than Alex*. How many marbles does Alex have?" can be expected to be more difficult than the very similar, but consistent word problem "Chris has 4 marbles. *Alex has 3 marbles more than Chris*. How many marbles does Alex have?." One explanation of this effect could be that, while consistent word problems can be directly solved using the direct translation strategy, inconsistent word problems require a deeper conceptual analysis of either the situation model or the mathematical model, to arrive at a correct solution (Scheibling-Sève et al., 2020).

1.3 Flexibility in dealing with additive situations

Already slight changes to the way a word problem is presented can substantially affect their difficulty. Several researchers have argued that this could provide a starting point to help students solve more difficult word problems.

For example, Stern (1993) and other researchers (Verschaffel, 1994; Fuson et al., 1996) stress the importance of understanding the meaning of relational statements and being able to deal with them. The symmetry of relational statements poses substantial challenges to students. Only 30% of the first-graders in Stern (1993) interview study could identify statements such as "*Chris has 3 marbles less than Alex.*" as equivalent to the symmetric statement *"Alex has 3 marbles more than Chris."* Being able to do so, however, could allow students to convert the inconsistent word problem "Chris has 4 marbles. *Chris has 3 marbles less than Alex*. How many marbles does Alex have?" with unknown reference set into the easier, consistent problem "Chris has 4 marbles. *Alex has 3 marbles more than Chris*. How many marbles does Alex have?" with unknown result set.

Similarly, Greeno (1980) argued that learners might find it easier to solve change problems with unknown relation set (change) such as "Jill had 3 apples. Betty gave her some more apples. Now Jill has 8 apples. How many did Betty give her?," if they reinterpret the situation as a combine situation with the result set as a whole (8 apples afterwards), the reference set (3 apples initially) as one part, and the relation set (change) as the other part. Considering that compare problems have been found to be harder than equalize problems speaks for a similar idea for solving compare problems (Nesher et al., 1982; Fuson et al., 1996). Both structures contain two disjoint sets offering the opportunity to reinterpret static compare statements such as "Alex has 3 marbles more than Chris" in terms of a (dynamic) equalization action "If Chris gets 3 more marbles (from someone else), Chris has as many as Alex."

One idea is common to both arguments: Re-interpreting word problems in terms of a different situation structure may allow students to turn difficult word problems into easier ones. Providing students with a range of different perspectives—connected to different situation structures—on the same situation may support word problem solving. This resonates with works stressing the importance of deep processing of situation models in word problem solving or mathematical modelling in general (e.g., Stern and Lehrndorfer, 1992; Thevenot et al., 2007; Leiss et al., 2010). Having perspectives available that correspond to easier word problem types may increase the chance to find a mathematical model and to solve the word problem. This re-interpretation can be seen as a part of the reading process in terms of Kintsch (1998) model of reading comprehension: Based on a mental representation of the text base in an initial situation model, the learner adds new perspectives on the situation by making further inferences based on his or her knowledge about connections between different situation structures. This idea, however, strongly depends on students' knowledge about these connections, their ability to identify similarities and differences between different perspectives on the same situation, and their ability to infer new perspectives that are fruitful for solving the problem.

Based on these considerations, Gabler and Ufer (2020, 2021) hypothesize that being able to perform these re-interpretations for additive one-step word problems might be a person characteristic, that shows systematic inter-individual variation between primary school students. *Flexibility in dealing with additive situations* can be defined as the skill to compare or restructure situation models of additive one-step word problems by inferring alternative perspectives on the situation that relate to different situation structures (e.g., different wording, semantic structures, or unknown sets). Flexibility is understood here similar as in cognitive flexibility theory, which "includes the ability to represent knowledge from different conceptual and case perspectives and then, when the knowledge must later be used, the ability to construct from those different conceptual and case representations a knowledge ensemble tailored to the needs of the understanding or problem-solving situation at hand" (Spiro et al., 1991, p. 24). It can be seen as a special case of conceptual knowledge of addition and subtraction in the sense of "implicit or explicit understanding of the principles that govern a domain and of the interrelations between units of knowledge in a domain" (Rittle-Johnson et al., 2001, p. 346). We note that the conceptualization of flexibility applied in this manuscript differs from other views that relate flexibility to humans' ability to shift between different tasks (Ionescu, 2012). Instead, flexibility in this manuscript refers to students' knowledge how to solve a specific (word) problem in different ways (Heinze et al., 2009; Ionescu, 2012).

Starting from this idea, a test instrument for flexibility in dealing with additive situations was investigated in Gabler and Ufer (2022). Here, students are asked to compare two verbal descriptions of the same situation and decide whether they believe these are description about the same situation, or not. Note that this is very different from the work by Zorrilla et al. (2024, p. 6, Figure 3) where students were provided with a number of related *complete* word problems, among which one problem contained the solution to another one—potentially leading to the high frequency of superficial strategies found there. The analyses of our skill construct assumed a one-dimensional scale, and results showed good reliability ($\alpha = 0.80$) with second graders rating 20 dichotomous items. This finding supports the assumption of a flexibility in dealing with additive situations as a one-dimensional personal characteristic. Moreover, Gabler and Ufer (2022) report that the construct predicts students' word problem solving skills above general cognitive abilities, language skills, and symbolic arithmetic calculation skills. They conclude, that language skills and symbolic arithmetic calculation skills, but not general cognitive abilities significantly explain inter-individual differences in students' flexibility. Gabler and Ufer (2024) provide evidence from an experimental intervention study, in which training second graders flexibility in dealing with additive situations increased not only their flexibility, but also their word problem solving performance. The qualitative analysis in Gabler and Ufer (2021) showed that students varied substantially in their progress during the intervention, depending on their flexibility at the start of the intervention. To conceptualize adaptive support in future studies, models are needed that allow a criterial interpretation of students' current performance in terms of concrete demands the students can (and cannot yet) master systematically. Moreover, little is known about how students' flexibility develops over time, for example, if certain transitions between levels are less frequent than others or take more time or if certain demands take longer to cope with than others.

1.4 Level models for assessment and intervention

Prior studies have unveiled substantial inter-individual differences in students' number-related knowledge and skills at the start of primary school (Schmidt and Weiser, 1982; Fuson, 1988), which persist during the first years of schooling and beyond (Krajewski and Schneider, 2009; Niklas and Schneider, 2017; Balt et al., 2020). Current models of numerical development often propose a sequence of levels, which describe demands of increasing complexity and are assumed to also reflect students' temporal development (Hartmann and Fritz, 2021). Each level in these models is characterized by specific content-related insights, e.g., grasping the idea of cardinality of a set. Often knowledge and skills related to previous levels are considered necessary to acquire the next one. This underpins the necessity to align instruction to students' current knowledge and skill level, optimally targeting the next level that is within students' reach. Indeed, Wildgans-Lang et al. (2020) argue that level models provide useful information that support teachers' diagnosis of students' current level of understanding.

Level models are usually generated based on analytic approaches towards test performance based on item response theory (IRT) models. They are one special case of models for cognitive knowledge or skill constructs, which differentiate skills along one coherent dimension of individual scores and define discrete levels of observable performance by defining the demands, that usually can be mastered on each level (Koeppen et al., 2008; Ufer and Neumann, 2018). Level models can be created for one-dimensional skill constructs, or for single dimensions of multi-dimensional skill constructs. This has a long tradition in large scale assessments (Heine et al., 2013), but level models also exist for different mathematical knowledge areas such as broad arithmetic skills in primary school (Reiss and Obersteiner, 2019), proportional reasoning skills or fraction knowledge (Schadl and Ufer, 2023), and mathematical knowledge required for undergraduate mathematics learning at university (Rach and Ufer, 2020; Pustelnik et al., 2023). Often level models are seen as a preliminary step towards models that describe development such as learning trajectories (Simon, 1995) or learning progressions (Jin et al., 2019). In this sense, level models may not only support diagnosis, but also provide useful (though often in the first place heuristic) hints towards reasonable learning goals and possible learning opportunities that would be most promising for a student on a specific level. If a student can be assigned to a level in a level model, providing learning opportunities and—if necessary—scaffolding on the next more complex level might be a plausible heuristic to support this student's progress. In this sense, level models provide a heuristic to identify what Lave and Wenger (1991) call a "scaffolding interpretation" of the concept of "zone of proximal development" (Vygotsky et al., 1978). It can be assumed that level models are the more useful to this end, the more they focus on a coherent, well-defined, and well-delineated skill or conceptual knowledge construct.

Level models are usually generated by analyzing the difficulties of a set of test items based on a scaling study. Two main approaches can be distinguished: Some researchers use statistical clustering methods (e.g., Marcoulides and Drezner, 2000) based on the empirical item difficulties, to obtain item subsets with coherent difficulty within each set and large difficulty gaps between the item subsets (e.g., Jiang et al., 2021). The main advantage of this approach is, that it results in clearly distinguishable item subsets. Its disadvantages are that it does not consider the sampling error of item difficulties and that it may result in item subsets that are hard to interpret in terms of common item demands. Variants of the so-called bookmark method are an alternative (Mitzel et al., 2013; Dimitrov, 2022). Items are sorted in an item booklet by increasing difficulty. This booklet is then analyzed for subsets of consecutive items, that share common item demands. This is an interpretative process, that results in subsets of items that cluster along the difficulty scale alongside with verbal descriptions of the common item demands. Depending on the implementation of the method, one expert or a group of experts are involved in this interpretation (e.g., Reiss and Obersteiner, 2019; Rach and Ufer, 2020;

Pustelnik et al., 2023; Schadl and Ufer, 2023). The item subsets from these analyses are interpreted as levels of demands regarding the skill construct. If the item difficulties stem from a one-dimensional item response theory (IRT) model, that can align test participants' performance scores and item difficulties on the same scale, test participants can be allocated to one of the levels, indicating which levels of demand the participant usually can already master, and which levels the participant probably will struggle with. In this sense, the levels can also be interpreted as levels of participants' knowledge or skill.

2 The current study

Assuming a relevance of flexibility in dealing with additive situations, central steps for targeted assessment and intervention in early primary school have been achieved in prior research. An applicable test instrument is available, that is based on a clear skill construct definition, and there is evidence speaking for its validity in terms of the importance for learning word problem solving. However, a level model that can guide teachers' diagnosis and adaptive instruction is still missing. The main goal of this study was to develop a level model for flexibility in dealing with additive situations, and to describe students' current flexibility and its development using this model. We analyzed data of students' flexibility in dealing with additive situations from prior studies with an IRT approach aiming at the construction of a level model for this skill construct.

To clarify whether constructing a level model is reasonable, at all, we first investigated if a one-dimensional model is suited to describe participants' flexibility in dealing with additive situations in our data (Q1).

Since models allowing to describe students' current performance in terms of concrete demands, they can master are rare, we the aimed to construct a level model for the flexibility construct, focusing on question Q2: Can we distinguish levels of coherent difficulty and item demands in primary school students' flexibility in dealing with additive situations? Which task features differentiate consecutive levels in terms of item demands?

Based on the generated level model, we aimed to characterize students' performance on the flexibility construct in more detail and investigated how students from our sample distributed across these levels (Q3).

Finally, we were interested whether such a model would be of added value to describe how students' flexibility develops over the span of few weeks. We investigated to which extent the model can be used to descriptively characterize students' progress in flexibility across a span of several weeks (Q4).

3 Materials and methods

3.1 Design and sample

We reanalyzed data from three studies which measured students' flexibility in dealing with additive situations. Study 1 was an unpublished, cross-sectional scaling study and comprised $N = 130$ grade 3 students (62 female, 68 male, $M_{age} = 8.5$ years) who worked only on the flexibility test instrument and provided some demographic data. In study 2, $N = 119$ grade 2 students (56 female, 63 male, $M_{age} = 7.6$ years) worked on this and other instruments in a cross-sectional experimental study. Some data from this study were reported in Gabler and Ufer (2021), but not including data on flexibility. Study 3 was an intervention study with $N = 134$ grade 2 students (66 female, 73 male, $M_{age} = 7.6$ years), in which the instrument was applied together with other measures in a pre-test, five weeks later in a post-test and four more weeks later in a follow-up test. Here, data from all three measurements were included in the analysis, including students who participated in only one ($N = 5$) or two ($N = 17$) of the measurements. Results of the intervention study are reported in (Gabler and Ufer, 2024). Overall, data from $N = 624$ test participations by $N = 383$ students were included in our analyses.

3.2 Flexibility instrument

The flexibility instrument was slightly adapted between the three studies. In study 1, an initial instrument with 18 items was used. Two items were removed due to their psychometric properties after initial scaling and replaced by four new items in study 2. After study 2, again five items were removed, because they turned out to be of very similar difficulty as other items, and five additional items were introduced to better cover areas of higher difficulty. Consequently, 27 different items in total were included in our initial analyses. The Marginal Maximum Likelihood Approach was used, so that the missing data for some items in small groups of participants could be accommodated when estimating the IRT models, by only including those items into the Likelihood Function for a specific person, which the person had worked on.

The test was framed as a story about a birthday party of two twins, Alma and Ben. In each item, two statements from participants of the party were given, and the students were asked, whether the two participants tell the same thing about the party. All presented statements were similar to the sentences that present changes or relations in usual word problems, e.g., "There were four chairs less than children on the party." and "There were four more children than chairs on the party." Most statements comprised change actions, equalize actions, and quantitative or qualitative comparisons. Integrating combine situations into the item format proved difficult, thus only few items contained two combine statements. Since combine situations are considered quite easy anyway, we did not see this as an issue. All statements contained exactly one numerical information, except for qualitative comparisons, which contained no numerical information. If an item contained a numerical information, the same number was provided in both statements. The two statements within each item differed in terms of different additive vs. subtractive wording, different semantic structures, or both. Note that calculations were neither possible nor useful, as at most one numeric information was given in each item. The answer options were "yes" (the two participants tell the same thing), "no" (they do not tell the same thing) or "I do not know." The last answer option was used very rarely, so that a probability of 50% for choosing the correct solution just by guessing must be assumed for all items.

3.3 Analyses

The generation of a level model (Q2) usually requires modelling the response data with an appropriate IRT model (Q1). Different IRT models were estimated using the R package tam (Robitzsch et al., 2020) and compared using Chi-Square-Likelihood-Ratio tests (investigating differences in model-data-fit), and the Akaike (AIC) and Bayes Information Criterion (BIC, lower values reflect better model fit) as information indices. First, we explored different one-dimensional models, that reflect different assumptions about how to model the item answer process: A three-parameter model, that assumes varying difficulty, guessing, and discrimination parameters over all items (model 1), a model that assumes constant item discriminations, but varying difficulty and guessing parameters (model 2), a model that assumed constant item discriminations, constant guessing parameters of 50%, and varying item difficulty parameters (model 3), and a one-parameter Rasch model with constant item discriminations and no (zero) guessing parameter (model 4). When testing these models, we balanced model parsimony with fit to the data and interpretability. For example, model 4 with varying item discriminations has the disadvantage of crossing item characteristic curves, which implies "substantive illogic in attempting to define a construct with item characteristic curves (ICC) that cross, because their slopes differ due to differing discriminations, or their asymptotes differ due to differing guessing parameters. Crossing curves cause the hierarchy of relative item difficulty to change at every ability level. This destroys the variable's criterion definition" (Wright, 1999, p. 74). We primarily used Warm's unbiased maximum likelihood estimator (WLE) for person parameters (Warm, 1989). Expected-A-Posteriori (EAP) (Bock and Aitkin, 1981) estimators were used to calculate a second measure of person reliability, but not for person parameter estimation, since these typically have a "larger inward bias toward the prior mean but smaller variance" (Wang, 2015, p. 445) than WLE estimates.

For example, the one-dimensional, restricted three-parameter model 3 would mean, that a difficulty parameter for each item and a single ability estimate for each person are estimated, assuming equal discrimination indices for all items, as well as a constant guessing parameter of 50%. The probability that a student *i* with performance parameter θ_i solves an item *j* with difficulty parameter δ_j in this model can be calculated as

$$P\left(X_{ij}=1\right)=\alpha_j+\left(1-\alpha_j\right)\frac{e^{\beta_j\left(\theta_i-\delta_j\right)}}{1+e^{\beta_j\left(\theta_i-\delta_j\right)}}$$

with guessing parameter $\alpha_j = 0.5$ and discrimination parameter $\beta_j = 1$ for all items *j*.

Given the decision for specific assumptions on the item answering process, we furthermore explored whether a two-dimensional model (model 5) would be superior for our purpose compared to the corresponding one-dimensional model. To this end, we investigated the two-dimensional model that was most plausible from our perspective, separating the ability to identify equivalent statements as equivalent from the ability to identify non-equivalent statements as non-equivalent into separate dimensions.

When deciding for a final model we used information criteria (AIC, BIC) and Chi-Square-Difference tests to compare models. Infit (weighted) and outfit (unweighted) Root Mean Square Deviation (RMSD) item fit measures (Adams and Wu, 2007) were used to evaluate model fit, with values above 1.5 considered unproductive for measurement (Linacre, 2002). Additionally, we inspected the Q3 statistics, interpreting values above 0.25 as indication of a violation of the local independence assumption in IRT models (Christensen et al., 2017). Consideration of these model comparison and model fit measures was balanced against model parsimony (preferring models with fewer parameters, that were equally plausible) and usefulness of the model for measurement (e.g., acceptable reliabilities).

To generate a level model (Q2), items were first ordered by their difficulty in an item booklet as in the bookmark method (*cf.* Mitzel et al., 2013 for a similar approach). Then, each item was characterized in its specific demands (e.g., involved statement types, necessity to deal with qualitative or quantitative comparisons, direction of the wording, equivalence of the two statements), to identify similarities between consecutive items as well as differences between neighboring groups of items with similar demands (*cf.* Rach and Ufer, 2020). In this way, groups of items with similar difficulty and similar item demands were identified. This mostly focused on the similarities and differences between the two statements in each item, but also considered reasonable comparison strategies and possible errors from the literature. This resulted in verbal descriptions of the common demands of the items in each level. Later, the differences in the demands between adjacent levels were analyzed based on these texts to make the model more accessible. Moreover, thresholds between the levels were established by calculating the average difficulty of the easiest item of one level and the most difficult item of the next lower level. Furthermore, the difficulty of the easiest and the most difficult items in the test were used as lower threshold of the lowest and upper threshold of the highest level.

To allocate student performance on the item difficulty scale for Q3, we assigned a student to that point on the scale, where she or he would have a 75% probability of solving the item, which is exactly the position of the students' performance parameter θ. In large scale studies that do not apply guessing parameters, values from 62.5 to 70% have been used. The exact cut-off value has been found to make little difference in some analyses (Rolfes and Heinze, 2022). We decided to use a substantially higher cut-off due to the high assumed guessing probability of 50%.

Assigning students to levels is usually done by assigning a student to the level, in which his or her performance parameter θ is located (i.e., between the lower and the upper threshold of this level). Consequently, *being on a level* means that a student can solve all items of the easier levels with more than 75% probability and all items of the more difficult levels with less than 75% probability. The assigned level comprises items, which the students has not (yet) mastered, but which are closest to being mastered. A far-reaching interpretation would be, that this level is the students' "zone of proximal development" (Vygotsky et al., 1978). Since person parameters often carry substantial measurement error, this assignment procedure with strict cut-offs can be debated: A student that has an estimated performance parameter close to a level threshold has a high probability of being assigned to the wrong level, since the (unknown) real performance parameter might be on the other side of the threshold. Therefore, we applied an alternative approach: For each student with estimated person parameter θ, corresponding standard error se_θ and each level *i*, we calculated a statistics $S(\theta, i)$ as the percentage of a normal distribution around θ with variance se_θ, that falls within the respective level. This results in one value per level and student, and the sum of the values across all levels for a single student is one. $S(\theta, i)$ can be interpreted as the probability density of obtaining a measured person parameter around θ, if one repeatedly would test students, whose real person parameter was surely at the respective level. Thus, summing $S(\theta, i)$ over all levels i yields a sum of one (since each real person parameter is one of the summed levels) for each (measured) person parameter value. We interpret the sum of $S(\theta, i)$ over all participants as a measure of the number of participants that can be allocated to level i, weighting each student by its distance to the respective level. We will refer to individual values of S as *level sampling probability* and to sample sums of S as *number of persons allocated to this level*.

Students' average development (Q4) was investigated using analysis of variance. A deeper, more qualitative description using the generated level model was based on bivariate density plots and local regression models.

4 Results

4.1 Item selection and model comparison (Q1)

Initial scaling analyses revealed that one item asking for a comparison of two combine statements showed a very high solution rate (90%) and problematic item fit values. Consequently, this item was excluded from further analyses. During the further analysis, another item that required to match equivalent compare and equalize statements was removed from the analysis due to substantial residual correlations with two other items ($Q3 = 0.35$ and $Q3 = 0.36$).

Table 1 shows the fit information and Table 2 shows Chi-Square-Likelihood-Ratio tests for the models reflecting different assumptions about the item answer process. Model 3 shows the lowest AIC and BIC indices, making it preferable over the other three models. Model 2 with freely estimated guessing parameters did not fit significantly better than model 3 with constant guessing parameters of 50%. It did, however, fit better than model 4 without guessing parameters. Guessing parameters in model 2 ranged between 34 and 50%. Model 1 showed significantly better fit than model 3. However, we decided to use the more restrictive model 3 for several reasons: (i) It contained less parameters but was preferable in terms of information indices. (ii) We considered the number of parameters in model 1 as high in relation to our restricted sample size. (iii) Varying item discriminations make it hard to describe students' performance in terms of item demands as item characteristic curves may intersect ("This destroys the variable's criterion definition," Wright, 1999, p. 74).

For model 3 the person reliabilities were acceptable (WLE: 0.67, EAP: 0.75). Also, item RMSD infit (0.98–1.03) and outfit (0.64–1.68) indices were mostly within the acceptable range (Linacre, 2002), except for one item with a high outfit (1.68) indicating some underfit. An analysis of the item showed that it contained statements about the height of block towers, while all other items dealt with numbers of objects. Since it was one of the few items containing qualitative comparisons, and since it was not of extremely low or high difficulty, we decided to keep the item in the analysis. The other items showed outfit values of 1.30 and below. Residual correlations between items (Q3) provided no indications of substantially violating the local independence assumption of the applied IRT model. The average person performance parameter was $M = 1.49$ ($SD = 1.54$).

Moreover, we investigated a two-dimensional model separating items with pairs of equivalent statements and items with pairs of non-equivalent statements into different dimensions (model 5). Indeed, this two-dimensional model fit the data significantly better than model 3 (Table 2). However, the WLE reliabilities of the two dimensions were unacceptably low (equivalent statements: WLE: 0.37, EAP: 0.67; non-equivalent statements: WLE: 0.34, EAP: 0.73; correlation between the two dimensions latent: 0.86, manifest WLE: 0.39). Since this indicates that the model is not well suited to derive individual person parameter estimates, the two-dimensional model was disregarded.

4.2 Level model (Q2)

Analyzing the item booklet sorted by empirical difficulty revealed, that investigating two qualitative compare statements or two combine statements constituted the easiest demands (level 1, $\delta = -2.65$ to -1.81). For example, one item required to match the (equivalent) qualitative comparison statements "Alma's tower is smaller than Ben's tower." and "Ben's tower is higher than Alma's tower."

All of the remaining items involving two equivalent statements (two statements referring to the same situation) turned out to be harder than items requiring to find differences between the descriptions in the two statements. Within those items asking the students to identify two statements as describing different situations, one set of eight items contained either two change statements in different directions (additive vs. subtractive), or a statement describing a temporal change of one of two initially equal sets and a compare statement about the two sets after the change (e.g., item 10: "Before the party, Alma and Ben had equally many wristbands. Ben got 3 bands more on the party" vs. "Alma has 3 bands more than Ben now").

TABLE 1 IRT model fit indices.

Model #	Parameter restrictions	*LogLikelihood*	N_{par}	*AIC*	*BIC*
Model 1	Free guessing and discrimination param.	−5147.6	76	10446.1	10783.2
Model 2	Free guessing param., discrimination = 1	−5177.4	51	10456.7	10683.0
Model 3	Guessing = 0.5, discrimination = 1	−5191.5	26	10435.1	10550.4
Model 4	Guessing = 0, discrimination = 1	−5260.5	26	10572.9	10688.3

LogLikelihood, Logarithmized likelihood of the model; N_{par}, number of model parameters; *AIC*, Akaike Information Criterion; *BIC*, Bayesian Information Criterion.

TABLE 2 IRT model comparisons.

Model comparison	χ^2	*df*	*p*
Model 1 vs. model 2	60.6	25	<0.001
Model 2 vs. model 3	28.4	25	0.29
Model 2 vs. model 4	166.2	25	<0.001
Model 3 vs. model 5	14.6	2	<0.001

χ^2, Chi-Square statistics; *df*, degrees of freedom.

Also, two items containing two compare statements each can be found on this level (level 2, $\delta = -1.81 \text{ to} - 0.78$, e.g., item 11: "Ben got 2 presents more than Alma" vs. "Alma got 2 presents more than Ben"). Different from the items on level 1, these items contained a quantitative relation information, but the two situations could be identified as different by a qualitative comparison – even if the quantitative relation was misunderstood as a concrete set. Similarly, one item with an additively worded equalize statement and a non-equivalent subtractively worded compare statement was in this interval, as well. Level 2 primarily requires establishing a change statement as being different from another change or a compare statement.

The next set of six items (level 3, $\delta = -0.78 \text{ to} - 0.18$) comprised items that involved either two equalize statements (e.g., item 18. "If Ben takes 3 cards more, he has as many cards as Alma." vs. "If Ben puts 3 cards away, he has as many cards as Alma.") or an equalize statement and a compare statement with the same (additive vs. subtractive) wording (e.g., item 21: "Currently, Alma has 4 postcards more than Ben." vs. "Alma needs to get 4 more postcards, so that she has as many postcards as Ben"). Regarding the matching of equalize and compare statements, statement pairs with different wording (level 2) were assigned to lower levels than statement pairs with the same wording (both additive or both subtractive). Level 3 requires establishing an equalize statement as being different from another equalize statement or a compare statement with the same wording.

Within those items asking to identify two statements as being equivalent (describing the same situation), three items contained either two change statements or two equalize statements (level 4, $\delta = 0.18 \text{ to } 0.79$, e.g., item 23: "Alma sold 3 lemonades to Ben" vs. "Ben bought 3 lemonades from Alma."). Thus, level 4 requires establishing two dynamic (change or equalize) statements as equivalent.

The next set of four items contained either a change and a compare statement or an equalize and a compare statement (similar to level 2 resp. 3, but with two statements matching the same situation; level 5, $\delta = 0.79 \text{ to } 1.79$, e.g., item 19: "If Ben gives 3 cookies to another child, he has as many cookies as Alma." vs. "Ben currently has 3 cookies more than Alma."). Level 5 requires establishing quantitative comparisons as being equivalent to change or equalize statements.

Each of the two items from the last set (level 6, $\delta = 1.79 \text{ to } 3.90$) contained two symmetric compare statements, one with additive and one with subtractive wording (e.g., "On the party, there were 4 chairs less than children." vs. "On the party, there were 4 children more than chairs."). Thus, level 6 is characterized by identifying symmetric quantitative compare statements as equivalent.

Figure 2 shows the level descriptions and the differences between consecutive levels.

4.3 Student distribution across the levels (Q3)

Figure 3 displays the smoothed level sampling probability of a participant being allocated to each level, depending on the participant's performance parameter θ. As expected, there is substantial overlap around the level thresholds. Note that participants have the highest level sampling probability for the narrow level 4 only in a very small interval of performance parameters. Moreover, for participants with very low performance parameters, level 1 assignment is hard to differentiate from assignment below level 1. On the other hand, for example, a high-level sampling probability of belonging to level 3 can be determined for performance parameters around zero.

To investigate students' distribution over the levels, we considered data from study 1, study 2, and the first measurement of study 3 ($N = 374$). Figure 4 and Table 3 display participants' distribution across the levels. Note that Figure 4 and the second column of Table 3 show the expected number of persons per level. In this sense, every participant counts into all levels, with higher shares on levels near to the participant's performance parameter and lower shares on more distal levels, as described by their level sampling probabilities. The findings indicate that while few participants can be found at level 1 or below, about one quarter of all participants was allocated to level 2 or 3. This indicates that these students still struggle with establishing two situation descriptions as being different. On the other hand, 26.0% of the participants can be assumed to be at level 6, indicating that they are able to identify change or equalize statements and quantitative compare statements as equivalent, but still struggle with symmetric quantitative comparisons. The finding, that only 7.6% of the participants seem to be able to identify symmetric quantitative compare statements as equivalent must be interpreted with care, since only two items of very different difficulty constitute this highest level.

4.4 Describing development using the level model (Q4)

To analyze the development in students' levels, we considered data from $N = 112$ grade 2 students only, who participated in all measurements of study 3. A repeated measures ANOVA with flexibility as dependent variable showed a significant effect for measurement $[F(222,2) = 33.63, p < 0.001, \eta^2_{\text{part}} = 0.23]$ and pairwise contrasts between each of the two measurements were significant (p's < 0.05, Tukey correction). Figure 5 shows the distribution of participants (counted by their level sampling probability) across the levels by measurement. A decrease can be observed for the levels 4 and below, primarily between T1 and T2. Level 5 assignment remained stable, most likely due to students from lower levels progressing to level 5 and students from level 5 progressing to higher levels. An increase can be observed for level 6 between T1 and T2, and above level 6 between each pair of measurements.

Even though Figure 5 may indicate that level population decreases for lower levels and increases for higher levels, it allows only restricted insight into students' development. To obtain a clearer picture, we plotted students' performance parameter in T2

level	l.thr.	...	u.thr.	Items demand mostly to identify...
above 6	+ 3.90	...		
6	+ 1.79	...	+ 3.90	...pairs of two comparison statements with different wordings as equivalent
5	+ 0.79	...	+ 1.79	...pairs of a change or equalize and a comparison statements as equivalent
4	+ 0.18	...	+ 0.79	...pairs of two change or two equalize statements with different wording (direction) as equivalent.
3	− 0.78	...	+ 0.18	...pairs of two equalize statements or a comparison and an equalize statement as different.
2	− 1.81	...	− 0.78	...pairs of two change statements or a comparison and a change statement as different.
1	− 2.65	...	− 1.81	...pairs of combination or qualitative comparison statements as equivalent.
below 1		...	− 2.65	

+ Establishing symmetric quantitative comparison statements as equivalent.
+ Establishing quantitative comparisons as equivalent to change or equalize statements.
+ Establishing two equalize or change statements as equivalent.
+ Seeing differences between equalize statements and quantitative comparisons.
+ Seeing differences between change statements and quantitative comparisons.
+ Matching combinations and qualitative comparisons.

FIGURE 2
Level thresholds, level descriptions, and differences between consecutive levels.

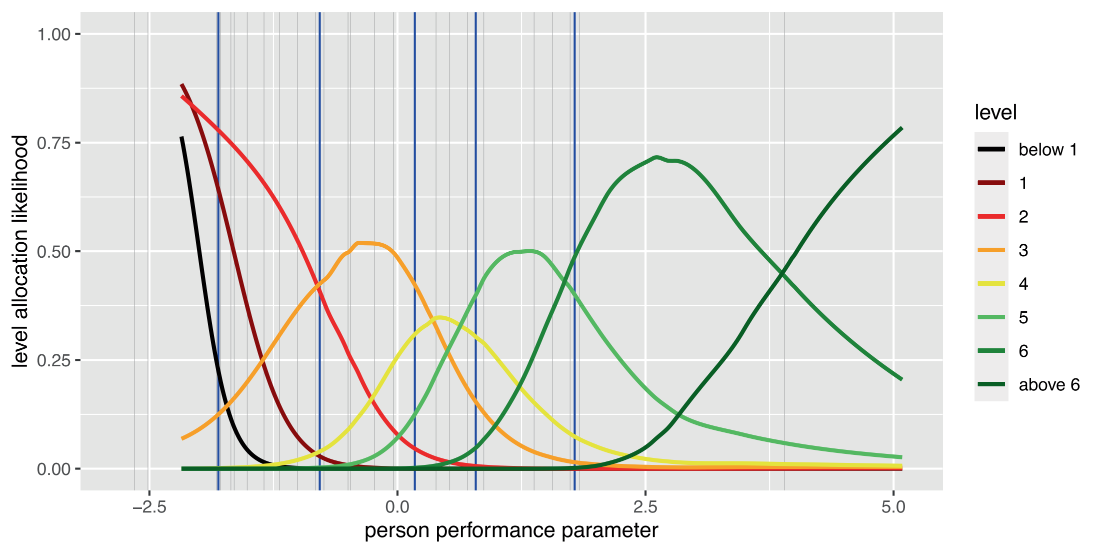

FIGURE 3
Level sampling probability by level and person performance parameter. Blue lines indicate level thresholds, light grey lines indicate item difficulties.

(resp. T3) against their performance in T1 (resp. T2) using two-dimensional density plots (Figures 6A,B). Lighter colors indicate more students in the respective areas. The number of students allocated to each combination of levels from pre-and post-test (resp. pre-and follow-up-test), calculated from the sum of level sampling percentages, is provided in Supplementary Tables S1, S2. Moreover, we used bivariate non-linear regression methods to estimate the average local T3 performance score for each T1 performance score.

Figure 6A indicates that, on average, participants throughout the whole range of initial (T1) performance, except those with very high values (above 2.5), show significant progress until T2 (the non-linear regression line and its confidence interval lie below the diagonal line indicating no progress from T1 to T2 for these students). This may be due to a ceiling effect, meaning that the flexibility test was very easy at T2 for participants with initially high scores. The regression line is a bit steeper at the upper end of level 4, indicating that being able to identify change (or equalize) and compare statements as equivalent already initially may substantially benefit students' progress from T1 to T2.

Analyzing the two-dimensional distribution in Figure 6A indicates that few students outperform level 6 even at T2 (dark colors at the upper end of the figure). Out of those students who were on level 5 at T1, quite some—but not all—were allocated to level 6 at T2,

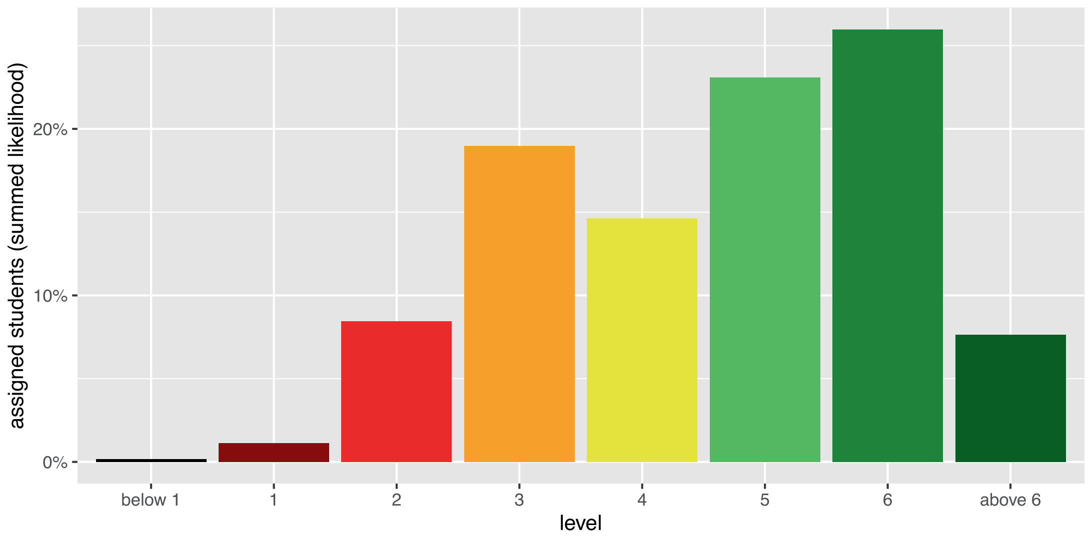

FIGURE 4
Allocated students per level for the whole dataset.

TABLE 3 Absolute and relative number of persons allocated per level.

Level	Number of persons allocated	Proportion
Above 6	28.5	7.6%
6	97.2	26.0%
5	86.4	23.1%
4	54.7	14.6%
3	70.9	19.0%
2	31.6	8.4%
1	4.2	1.1%
Below 1	0.6	<0.1%
Sum	374	100.0%

showing that these students could benefit and learn how to match equalize and compare statements. Students who were on the (narrow) level 4 at T1 spread out across levels 3–5 and—less frequently—level 6 at T2. Some students remain on level 4 or even show lower scores at T2—possibly due to measurement error variation. Others show scores on level 5 and the lower end of level 6. This indicates that students at the transition to being able to identify statements as equivalent benefited very differently from learning opportunities between T1 and T2. Moreover, many of those students who were on level 3 at T1, who were still struggling to identify compare and equalize statements as different, remained at this level and showed little progress.

Figure 6B indicates significant progress also from T2 to T3 over almost the whole range of T2 performance, again except for T2 scores above 2.5. Now, some of the students who were at level 3 at T2 progress to level 4 at T3. The pattern is less clear for level 4 at T2, which is less populated than at T1 (*cf.* Figure 5). Similarly, there is some progress for students from level 5 at T2 to level 6 at T3.

5 Discussion

In this study, we report the development of a level model of primary school students' flexibility in dealing with additive situations. We were able to distinguish six different levels of coherent difficulty and coherent item demands. Further, students were assigned to the most likely levels according to their person performance parameters which were estimated from their ability to rate whether two mathematical statements referred to the same or a different situation. The data from three different studies with a total of 383 primary school children in Grades 2 and 3 were used in the development of the model and children were assigned mostly to levels 5 and above with only few children in level 2 and below. However, a substantial number of students remained below level 6, which describes beginning mastery of symmetric comparison statements. In addition, we were able to analyze children's progression across about 2 months. Children on level 4 and 5 at the beginning of the assessments often progressed to the next level, whereas children on level 3 often did not progress to higher levels. This may indicate a qualitative step between levels 3 (and below) and levels 4 (and above). This kind of information will be valuable for teachers in their support of children's flexibility and their mathematical word problem solving skills.

5.1 A level model for flexibility in dealing with additive situations

To generate a level model, we applied an exploratory approach using data from an existing test instrument. With our IRT approach, we were able to identify a fairly clear level structure for children's flexibility in dealing with additive situations. The most pronounced differentiation was visible for identifying statements as non-equivalent (on lower levels) and as equivalent (on higher levels). It remains to be explored, if this reflects a general tendency of students to mark statements as different in the case of doubt. Taking a closer look, it is

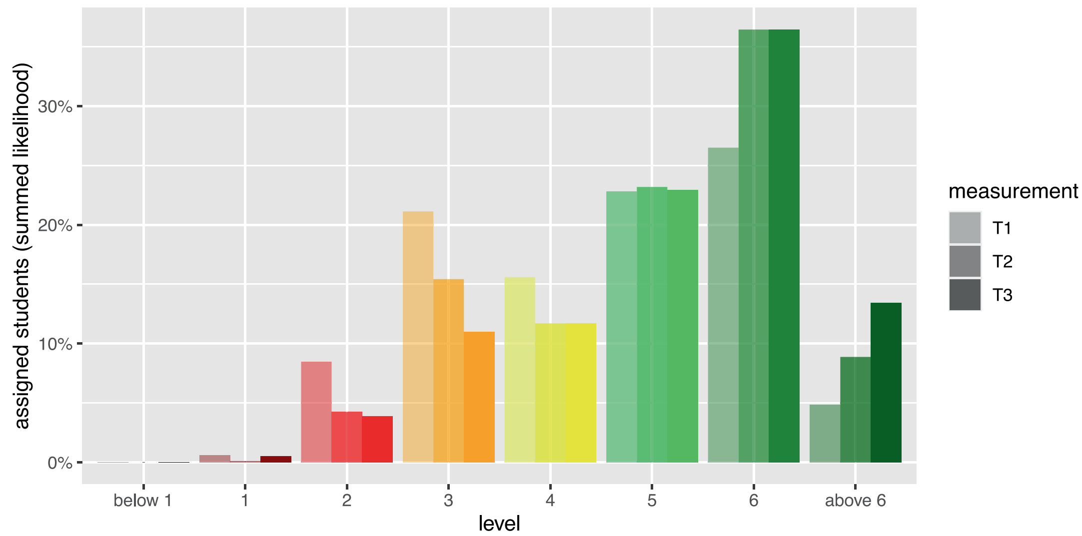

FIGURE 5
Assigned students per level and measurement for study 3.

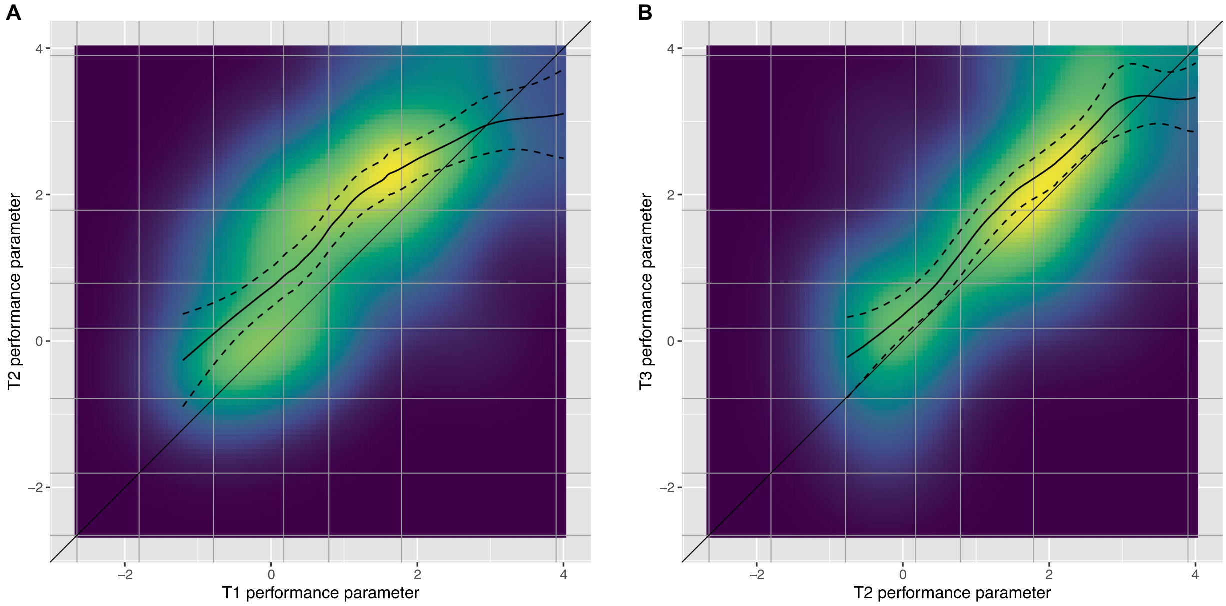

FIGURE 6
Density plots for person performance parameter at T2 **(A)** resp. T3 **(B)** by person performance parameter at T1 **(A)** resp. T2 **(B)** with bivariate non-linear regression line and 95% confidence interval. Grey lines indicate level thresholds.

sufficient to identify even a small difference in the described situations to establish two statements as non-equivalent. Contrary, to establish two statements as equivalent, the existence of such two differences must be ruled out, for example by explicitly transforming perspective described by one statement into the one described by the other (e.g., by seeing that an equalization action described in one statement would balance the quantitative relation described in the other one). Apart from these two broad areas, the easiest level 1 comprised dealing with combine statements and qualitative comparisons.

A bit different patterns occurred within each of the two areas (identifying statements as non-equivalent resp. equivalent). For identifying statements as equivalent, matching with two dynamic statements was easiest (level 4) followed by matching a dynamic statement and a compare statement (level 5) and two compare statements (level 6). This speaks for a learning trajectory from dynamic situations to static situations, as it was also conjectured in (Gabler and Ufer, 2021). For identifying two statements as different, matching two quantitative compare statements was part of the lower level 2, together with dealing with two change statement or one change and one compare statement. This low difficulty of identifying two quantitative compare statements as different should not be overestimated, as it can be solved by just considering the entailed qualitative comparisons ("Chris has 3 marbles more than Alex" vs. "Alex has 3 marbles more than Chris"), and it can be even solved correctly if statements such as "Chris has 3 marbles more than Alex" are interpreted as "Chris has 3 marbles, and Chris has more marbles than Alex" (Mekhmandarov et al., 1996). Apart from this, dealing with equalize statements (possibly along with one compare statement, level 3) turned out more difficult than dealing with change statements (possibly along with one compare statement, level 2). Since equalize statements comprise a (imagined) change action together with a statement about the effect of this action, this can be explained by the different complexity the two dynamic statement types.

It should be noted that these differences in item difficulty can not necessarily be interpreted as a naturally occurring, individual learning trajectory in the psychological sense. As a didactic model, progressing from identifying statements as non-equivalent to identifying statements as equivalent would not make sense. Our results still may serve to generate hypothetical learning trajectories, which are not purely descriptive psychological models, but are strongly engendered within a specific instructional context (Simon, 1995; Clements and Sarama, 2012). They also carry a normative component, making proposals on how to (more effectively) treat a certain content in instruction. Progressing from the analysis of qualitative comparison and dynamic statements to matching dynamic statements with quantitative compare statements and finally pairs of compare statements would be reasonable based on our findings. When informally reviewing first grade mathematics textbooks in Germany, we did not find activities related to the interpretation of quantitative comparison statements and Gabler et al. (2023) found few compare word problems in German first and second grade textbooks. Even though it does not follow from our findings, discussing the meaning of quantitative compare statements, for example by considering two concrete sets, establishing a one-to-one-correspondence between the smaller set and a subset of the larger one, and highlighting the remaining items as the difference set, seems to be a reasonable learning opportunity before matching compare statements to other statements.

The most striking difference in difficulty was between qualitative comparison statements, which belonged to the lowest levels, and quantitative comparison statements, which made up the upper end of the difficulty continuum. This reflects the specific difficulty of representing (Mekhmandarov et al., 1996; Gabler and Ufer, 2021) and flexibly interpreting (Stern, 1993) quantitative comparison statements.

5.2 Students' flexibility in dealing with additive situations

Even though the average student performance parameter was in the range of level 5, the substantial standard deviation raises the question about a qualitative interpretation of inter-individual differences in students' performance. Our model allows to allocate students to the levels, offering ways to provide teachers with more specific information than a rough overall performance measure (in terms of, for example, percent tasks solved correctly). Each level comes with a criterial interpretation (Ufer and Neumann, 2018) of the demands students can systematically master (below the assigned level), which they master partially (assigned level) and which demands they will typically fail currently (above the assigned level). This may allow teachers to identify individual learning needs of each student. However, we also note that level assignment underlies measurement error. Determining the level sampling probability of a student "being" on this level allows to make this natural uncertainty transparent to teachers by reporting the two or three most likely levels for a student—at the cost of providing a quite conservative (i.e., too much spread out) picture of the assignment uncertainty. It might be of interest how teachers deal with this kind of uncertainty when evaluating this information as diagnostic evidence and deriving didactical conclusions (Heitzmann et al., 2019). This uncertainty is part of teachers' everyday practice and should also be acknowledged in teacher education—instead of, for example, asking (active or pre-service) teachers to make definite decisions between levels in simulations (Wildgans-Lang et al., 2020).

We found that around 60% of the students are systematically able to identify non-equivalent statements as non-equivalent. However, this also means that about 40% of the students in our sample are not able to do so. Thus, comparing and contrasting (Hattikudur and Alibali, 2011) different verbal statements about mathematical structures in real-world situations may be promising for all students—either learning about differences between non-equivalent pairs of statements, or learning to correctly match equivalent statements.

It must be noted that almost all students in our study mastered level 1, comprising combine and qualitative compare statements, and few still struggled with differentiating descriptions of different change situations (level 2). This indicates, that change situations might be a good didactical starting point for the learning trajectory proposed above. Dealing with equalize statements posed problems to at least 40% of the students, indicating that understanding them may not be taken for granted and should also be addressed in instruction. Since equalize statements may provide a didactical bridge to reinterpreting and understanding compare statements, more than the currently still restricted evidence (one example is Stern, 1994) on students' understanding of and performance on equalize word problems might be desirable. Finally, less than 10% or our sample systematically mastered to identify symmetric quantitative compare statements as equivalent. This is in line with findings by Stern (1993) and underpins the importance of dealing with compare statements in more depth in first years mathematics instruction.

5.3 Development of flexibility in dealing with additive situations

The longitudinal analyses provide insights into students' development of flexibility in dealing with additive situations. As may be expected when a skill develops positively over time in a population, lower levels decrease in frequency, while higher levels increase, and middle levels remain rather stable over the short period of several weeks in our study. Thus, in terms of an interpretation of students' development, level frequencies contain little more information than analyses of average development. At least we can say, that an increasing number of students master symmetric quantitative compare statements after several weeks.

Deeper, qualitative analyses of the bivariate distributions based on performance from two consecutive measurements allows to derive more elaborate conjectures. Firstly, some progress seems to be observable for students all over the whole range of prior performance range. Students with high initial flexibility form an exception, most likely due to a ceiling effect of the instrument. This indicates that flexibility is a skill that can be learned and developed.

However, there are first indications that this progress is not equally strong for all students. Until T2, there was little progress of students who were initially on level 3, and from T2 to T3 only few students progress towards level 4. The levels 3 and 4 differentiate between finding differences between statements (level 2 and 3) and identifying statements as equivalent (levels 4 and above). It seems that students, who were on level 3 initially, cannot really benefit from eventual opportunities in the investigated time span between T1 and T3. It seems that the challenge to identify different descriptions of the same situation as equivalent requires specific instructional support, for example by establishing strategies such as the dynamization or the inversion strategies described above. On the other hand, the acquisition of these strategies might still require a basic understanding of the corresponding situation structures. It remains an open question if what constitutes level 3 in our model provides sufficient prior knowledge and skills for this purpose.

Indeed, students who were on level 4 initially, showed varying progress from T1 to T2. This may indicate that other factors, not explicitly investigated in our analyses, influence students' progress here. Given that flexibility comprises verbal descriptions of mathematical structures in everyday situations, students language skills (Purpura and Reid, 2016; Peng et al., 2020; Ufer and Bochnik, 2020) may be an important factor here, but also their mathematical prior knowledge. Future research should clarify not only which student characteristics explain students' performance in word problem solving or flexibility, but also the development of this performance (e.g., Paetsch et al., 2016).

Moreover, almost exclusively those students progressed to level 6, who were on level 5 at the preceding measurement. This indicates that being able to match compare statements to equivalent dynamic statements might be an important prerequisite for (at least partially) mastering symmetric compare statements. Future research considering students' development with a better temporal resolution or applying deeper analysis of their learning and reasoning processes when dealing with symmetric quantitative compare statements should investigate this hypothesis in more detail.

One option to put the developmental hypotheses from this study to a test would be to provide students adaptively with instruction targeting the assigned next level and compare this to offering instruction targeting different higher levels. This may provide useful insights into whether students' prior flexibility plays a role, at all, for further learning, and if yes, which level of instruction requires which prior skills. Furthermore, such evidence may support the conceptualization of adaptive instruction (Plass and Pawar, 2020) that teachers (or other actors in the educational system) may provide based on a diagnosis of students' flexibility. Finally, the results are also of practical significance, as they may be used in the development of textbooks which may even contain activities to train flexibility, and for specific diagnostic processes, in which the current student's approach towards additive situations and their current level of flexibility are assessed and considered in a subsequent training.

5.4 Limitations

This study is marked by several limitations. We used an interpretational method based on item contents and item difficulties to arrive at our levels and their characterizations. It would be helpful to replicate the model with an independent sample if persons and, optimally, with a new set of items that was constructed based on the model. This would also allow to extend the test beyond the few items, that were used to build the level model. It would also be helpful to improve the level model and to contribute to a better differentiation between different learners. Future extensions of the instrument should cover typical errors, such as the misinterpretation of quantitative compare statements as statements about a concrete set and a qualitative comparison. Moreover, demands on higher levels of flexibility still need to be conceptualized. One idea would be to have students actively produce alternative, equivalent descriptions of the same situation (using, for example, provided words) instead of judging given statements. This would, however, also come with stronger demands in terms of language skills (Gabler and Ufer, 2022). Finally, other kinds of situations could be included, covering not only statements about the cardinality of sets, but also about other measures such as lengths, weights, etc.

When doing so, dimensionality of the construct should be considered further. The observed reliability of the whole scale was satisfactory, but not optimal. Including more items per dimension by varying the existing situation description statements could potentially remedy the very low reliabilities of the whole scale or the two assumed subscales. However, an eventual two-dimensional model would differentiate matching vs. unmatching statement pairs, which are also separated into the upper and the lower end of the one-dimensional current difficulty scale. For many participants one the two dimensions would end up being either very easy or very hard. Furthermore, the two dimensions showed a very high latent correlation in our analysis, also pointing to little added explanatory value above a one-dimensional model. If multi-dimensional models are considered, level models would need to be established for each dimension separately. For the two-dimensional model considered in this manuscript, it would be plausible that levels 1 to 3 would arise for one scale and levels 4 to 6 for the other one. Alternative multi-dimensional could be explored, e.g., differentiating the type of statements entailed each item.

Our study applied mostly descriptive methods to either characterize observed student performance or to derive hypotheses about developmental questions. Formal statistical inference methods could be applied to investigate the average development in the study 3 sample—which is not the main contribution of the study. Other methods such as think aloud methods, more intensive tracking of student performance over a longer time span, or intervention studies are necessary to further investigate our exploratory findings. Developmental conclusions are also limited by the ceiling effect of the test instrument, speaking again for an extension of the test with more complex demands. Relatedly, we assumed sufficient longitudinal measurement invariance in this manuscript, but this assumption needs to be tested. Such analyses might also provide deeper insights into the development of students' flexibility. Finally, note that our approach to allocate participants to the levels is still quite pragmatic and not based on a systematic estimation of participants' probability to belong to the profiles. Future works might explore more advanced allocation techniques, such as calculating this probability by using the overall score distribution, or even *a-priori*-distributions for each participant based on prior measurements as a reference.

Finally, our results are limited by the fact that they are solely based on an observation of students' final performance on the flexibility test items. A closer consideration of students' solution strategies, in particular for items containing equivalent vs. non-equivalent findings could extend prior results that primarily address the interpretation of quantitative compare statements (Mekhmandarov et al., 1996).

6 Conclusion and outlook

The contribution proposes a model of different levels of complexity for students' flexibility in dealing with additive situations. Even though longitudinal interpretations should be taken with care, the model itself may support informal observation and formative assessment of students' flexibility in the first 2 years of primary school. To this end, it should be investigated to which extent teachers can make use of this model to either observe students or interpret students' solutions on statement matching tasks or word problems, or to make use of findings from an externally applied assessment of students' skills with a flexibility test like the one analyzed here. Regarding the latter use of the instrument, the uncertainty entailed in the level assignment from such tests is a matter that should be addressed explicitly when informing teachers about their students' performance, but also when investigating teachers' use of such information.

A further step would be to conceptualize and investigate adaptive instruction based on the model. This could provide useful and more reliable insights into the contingencies between different subskills in the development of students' flexibility, but also evidence about effective strategies to support this development. Specific focus could be put on the difference between levels 1–3 and levels 4–6, which turned out to be particularly challenging in our longitudinal analysis.

Summarizing, we highlight a skill in this manuscript that plays an important role in students' development of word problem solving skills. The proposed level model requires further validation, extension, and investigation, but may develop into a useful tool for assessment and intervention in mathematics instruction in the early school years.

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

Ethical approval was not required for the study involving humans in accordance with the local legislation and institutional requirements. The study was approved by the local school authorities under the Bavarian Ministry of Education. Written informed consent for participation in this study was provided by the participants' legal guardians/next of kin.

Author contributions

SU: Conceptualization, Formal analysis, Funding acquisition, Methodology, Resources, Supervision, Visualization, Writing – original draft, Writing – review & editing. AK: Validation, Writing – review & editing. LG: Conceptualization, Data curation, Investigation, Methodology, Project administration, Writing – review & editing. FN: Validation, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

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.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/feduc.2024.1340322/full#supplementary-material

References

Adams, R. J., and Wu, M. L. (2007). The mixed-coefficients multinomial logit model: a generalized form of the Rasch model. In: *Multivariate and mixture distribution Rasch models: extensions and applications*, Eds. M. van Davier and C. H. Carstensen (New York, US: Springer). 57–75.

Balt, M., Fritz, A., and Ehlert, A. (2020). Insights into first grade Students' development of conceptual numerical understanding as drawn from progression-based assessments. *Front. Educ.* 5:80. doi: 10.3389/feduc.2020.00080

Blum, W., and Leiß, D. (2007). "How students and teachers deal with modelling problems." In: *Mathematical Modelling: Education, Engineering and Economics - ICTMA12.* Eds. C. Haines, P. Galbraith, W. Blum and S. Khan (Chichester, UK: Horwood Publishing). 222–231.

Bock, R. D., and Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. *Psychometrika* 46, 443–459. doi: 10.1007/BF02293801

Cevikbas, M., Kaiser, G., and Schukajlow, S. (2022). A systematic literature review of the current discussion on mathematical modelling competencies: state-of-the-art developments in conceptualizing, measuring, and fostering. *Educ. Stud. Math.* 109, 205–236. doi: 10.1007/s10649-021-10104-6

Christensen, K. B., Makransky, G., and Horton, M. (2017). Critical values for Yen's Q 3: identification of local dependence in the Rasch model using residual correlations. *Appl. Psychol. Meas.* 41, 178–194. doi: 10.1177/0146621616677520

Clements, D. H., and Sarama, J. (2012). "Learning trajectories in mathematics education." In: *Hypothetical learning trajectories*. Eds. D. H. Clements and J. Sarama (New York, US: Springer), 81–90.

Czocher, J. A. (2018). How does validating activity contribute to the modeling process? *Educ. Stud. Math.* 99, 137–159. doi: 10.1007/s10649-018-9833-4

Daroczy, G., Meurers, D., Heller, J., Wolska, M., and Nürk, H.-C. (2020). The interaction of linguistic and arithmetic factors affects adult performance on arithmetic word problems. *Cogn. Process.* 21, 105–125. doi: 10.1007/s10339-019-00948-5

Daroczy, G., Wolska, M., Meurers, W. D., and Nuerk, H.-C. (2015). Word problems: a review of linguistic and numerical factors contributing to their difficulty. *Front. Psychol.* 6:348. doi: 10.3389/fpsyg.2015.00348

Dimitrov, D. M. (2022). The response vector for mastery method of standard setting. *Educ. Psychol. Meas.* 82, 719–746. doi: 10.1177/00131644211032388

Freudenthal, H. (1983). *Didactical phenomenology of mathematical structures*. Dordrecht, NL: D. Reidel.

Fuson, K. C. (1988). *Children's counting and concepts of number* New York, US: Springer.

Fuson, K. C., Carroll, W. M., and Landis, J. (1996). Levels in conceptualizing and solving addition and subtraction compare word problems. *Cogn. Instr.* 14, 345–371. doi: 10.1207/s1532690xci1403_3

Gabler, L., and Ufer, S. (2020). Flexibilität im Umgang mit mathematischen Situationsstrukturen: Eine Vorstudie zu einem Förderkonzept zum Lösen von Textaufgaben zu Addition und Subtraktion [Flexibility when dealing with situational structures in mathematical contexts—a preliminary study investigating a learning framework on solving additive word problems]. *J. Math.-Didakt.* 42, 61–96. doi: 10.1007/s13138-020-00170-3

Gabler, L., and Ufer, S. (2021). Gaining flexibility in dealing with arithmetic situations: a qualitative analysis of second graders' development during an intervention. *ZDM* 53, 375–392. doi: 10.1007/s11858-021-01257-y

Gabler, L., and Ufer, S. (2022). Contribution of flexibility in dealing with mathematical situations to word-problem solving beyond established predictors. In: *Proceedings of the 45th Conference of the International Group for the Psychology of Mathematics Education*, Eds. C. Fernández, S. Llinares, A. Gutiérrez, and N. Planas (Alicante, Spain: PME). 2:267–274.

Gabler, L., and Ufer, S. (2024). Training flexibility in dealing with additive word problems. *Learn. Instr.* 92:101902. doi: 10.1016/j.learninstruc.2024.101902

Gabler, L., von Damnitz, F., and Ufer, S. (2023). "Additive word problems in German 1st and 2nd grade textbooks." In: *Proceedings of the 46th Conference of the International Group for the Psychology of Mathematics Education*. Eds. M. Ayalon, B. Koichu, R. Leikin, M. Rubel and M. Tabach (Haifa, Israel: PME), 2, 355–362.

Greeno, J. G. (1980). "Some examples of cognitive task analysis with instructional implications." In: *Aptitude, learning, and instruction: volume 2: cognitive process analysis of learning and problem solving*. Eds. E. Snow, P.-A. Frederico and W. E. Montague (Hillsdale, US: Lawrence Erlbaum Associates), 1–21.

Hartmann, J., and Fritz, A. (2021). "Language and mathematics: how children learn arithmetic through specifying their lexical concepts of natural numbers." In: *Diversity dimensions in mathematics and language learning*. Eds. A. Fritz, E. Gürsoy and M. Herzog (Boston, US: De Gruyter), 21–39.

Hattikudur, S., and Alibali, M. (2011). The role of comparison in mathematics learning. *Proc. Annu. Meet. the Cogn. Sci. Soc.* 33, 306–311.

Hegarty, M., Mayer, R. E., and Green, C. E. (1992). Comprehension of arithmetic word problems: evidence from students' eye fixations. *J. Educ. Psychol.* 84, 76–84. doi: 10.1037/0022-0663.84.1.76

Hegarty, M., Mayer, R. E., and Monk, C. A. (1995). Comprehension of arithmetic word problems: a comparison of successful and unsuccessful problem solvers. *J. Educ. Psychol.* 87, 18–32. doi: 10.1037/0022-0663.87.1.18

Heine, J.-H., Sälzer, C., Borchert, L., Sibberns, H., and Mang, J. (2013). "Technische Grundlagen des fünften internationalen Vergleichs." In: *PISA 2012- Fortschritte und Herausforderungen in Deutschland*. Eds. M. Prenzel, C. Sälzer, E. Klieme and O. Köller (Münster, DE: Waxmann), 309–346.

Heinze, A., Star, J. R., and Verschaffel, L. (2009). Flexible and adaptive use of strategies and representations in mathematics education. *ZDM* Münster, DE 41, 535–540. doi: 10.1007/s11858-009-0214-4

Heitzmann, N., Seidel, T., Opitz, A., Hetmanek, A., Wecker, C., Fischer, M., et al. (2019). Facilitating diagnostic competences in simulations: a conceptual framework and a research agenda for medical and teacher education. *Frontline Learn. Res.* 7, 1–24. doi: 10.14786/flr.v7i4.384

Ionescu, T. (2012). Exploring the nature of cognitive flexibility. *New Ideas Psychol.* 30, 190–200. doi: 10.1016/j.newideapsych.2011.11.001

Jiang, Z., Mok, I. A. C., and Li, J. (2021). Chinese students' hierarchical understanding of part-whole and measure subconstructs. *Int. J. Sci. Math. Educ.* 19, 1441–1461. doi: 10.1007/s10763-020-10118-1

Jin, H., Mikeska, J. N., Hokayem, H., and Mavronikolas, E. (2019). Toward coherence in curriculum, instruction, and assessment: a review of learning progression literature. *Sci. Educ.* 103, 1206–1234. doi: 10.1002/sce.21525

Kaiser, G. (2017). The teaching and learning of mathematical modeling. In: *Compendium for research in mathematics education*. Ed. J. Cai (Reston, US: NCTM), 267–291.

Kintsch, W. (1998). *Comprehension: a paradigm for cognition* Cambridge, UK: Cambridge University Press.

Kintsch, W., and Greeno, J. G. (1985). Understanding and solving word arithmetic problems. *Psychol. Rev.* 92, 109–129. doi: 10.1037/0033-295X.92.1.109

Koeppen, K., Hartig, J., Klieme, E., and Leutner, D. (2008). Current issues in competence modeling and assessment. *J. Psychol.* 216, 61–73. doi: 10.1027/0044-3409.216.2.61

Krajewski, K., and Schneider, W. (2009). Early development of quantity to number-word linkage as a precursor of mathematical school achievement and mathematical difficulties: findings from a four-year longitudinal study. *Learn. Instr.* 19, 513–526. doi: 10.1016/j.learninstruc.2008.10.002

Krawitz, J., Chang, Y.-P., Yang, K.-L., and Schukajlow, S. (2022). The role of reading comprehension in mathematical modelling: improving the construction of a real-world model and interest in Germany and Taiwan. *Educ. Stud. Math.* 109, 337–359. doi: 10.1007/s10649-021-10058-9

Lave, J., and Wenger, E. (1991). *Situated learning: legitimate peripheral participation*. Cambridge, UK: Cambridge University Press.

Leiss, D., Schukajlow, S., Blum, W., Messner, R., and Pekrun, R. (2010). The role of the situation model in mathematical modelling: task analyses, student competencies, and teacher interventions. *J. Math.-Didakt.* 31, 119–141. doi: 10.1007/s13138-010-0006-y

Lewis, A. B., and Mayer, R. E. (1987). Students' miscomprehension of relational statements in arithmetic word problems. *J. Educ. Psychol.* 79, 363–371. doi: 10.1037/0022-0663.79.4.363

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean. *Rasch Meas. Trans.* 16:878.

Marcoulides, G. A., and Drezner, Z. (2000). "A procedure for detecting pattern clustering in measurement designs." In: *Objective measurement: theory into practice*. Eds. M. R. Wilson and G. Engehlhard (Norwood, US: Ablex Publishing Corporation), 287–302.

Mekhmandarov, I., Meron, R., and Peled, I. (1996). Performance and understanding: a closer look at comparison word problems. In: *Proceedings of the 20th Conference of the International Group for the Psychology of Mathematics Education*, Eds. L. Puig and A. Gutiérrez, Lisbon, Portugal: PME, 3, 385–390.

Mellone, M., Verschaffel, L., and Van Dooren, W. (2017). The effect of rewording and dyadic interaction on realistic reasoning in solving word problems. *J. Math. Behav.* 46, 1–12. doi: 10.1016/j.jmathb.2017.02.002

Mitzel, H. C., Lewis, D. M., Patz, R. J., and Green, D. R. (2013). "The bookmark procedure: psychological perspectives." In: *Setting performance standards*. Ed. G. J. Cizek (New York, US: Routledge), 263–296.

Nesher, P., Greeno, J. G., and Riley, M. S. (1982). The development of semantic categories for addition and subtraction. *Educ. Stud. Math.* 13, 373–394. doi: 10.1007/BF00366618

Niklas, F., and Schneider, W. (2017). Home learning environment and development of child competencies from kindergarten until the end of elementary school. *Contemp. Educ. Psychol.* 49, 263–274. doi: 10.1016/j.cedpsych.2017.03.006

Paetsch, J., Radmann, S., Felbrich, A., Lehmann, R., and Stanat, P. (2016). Sprachkompetenz als Prädiktor mathematischer Kompetenzentwicklung von Kindern deutscher und nicht-deutscher Familiensprache. *Z. Entwicklungspsychol. Pädagog. Psychol.* 48, 27–41. doi: 10.1026/0049-8637/a000142

Peng, P., Lin, X., Ünal, Z. E., Lee, K., Namkung, J., Chow, J., et al. (2020). Examining the mutual relations between language and mathematics: a meta-analysis. *Psychol. Bull.* 146, 595–634. doi: 10.1037/bul0000231

Plass, J. L., and Pawar, S. (2020). Toward a taxonomy of adaptivity for learning. *J. Res. Technol. Educ.* 52, 275–300. doi: 10.1080/15391523.2020.1719943

Purpura, D. J., and Reid, E. E. (2016). Mathematics and language: individual and group differences in mathematical language skills in young children. *Early Child. Res. Q.* 36, 259–268. doi: 10.1016/j.ecresq.2015.12.020

Pustelnik, K., Rach, S., Ufer, S., and Sommerhoff, D. (2023). Levels of mathematical knowledge in linear algebra for entering university. In: *Proceedings of the 46th Conference of the International Group for the Psychology of Mathematics Education*, Eds. M. Ayalon, B. Koichu, R. Leikin, M. Rubel and M. Tabach Haifa, Israel: PME, 4, 75–82.

Rach, S., and Ufer, S. (2020). Which prior mathematical knowledge is necessary for study success in the university study entrance phase? Results on a new model of knowledge levels based on a reanalysis of data from existing studies. *Int. J. Res. Undergrad. Math. Educ.* 6, 375–403. doi: 10.1007/s40753-020-00112-x

Reiss, K., and Obersteiner, A. (2019). "Competence models as a basis for defining, understanding, and diagnosing students' mathematical competences." In: *International handbook of mathematical learning difficulties: from the laboratory to the classroom*. Eds. A. Fritz, V. G. Haase and P. Räsänen, (Cham, DE: Springer), 43–56.

Riley, M. S., and Greeno, J. G. (1988). Developmental analysis of understanding language about quantities and of solving problems. *Cogn. Instr.* 5, 49–101. doi: 10.1207/s1532690xci0501_2

Riley, M. S., Greeno, J. G., and Heller, J. I. (1983). "Development of children's problem-solving ability in arithmetic." In: *The development of mathematical thinking*. Ed. H. P. Ginsburg (New York, US: Academic Press), 153–196.

Rittle-Johnson, B., Siegler, R. S., and Alibali, M. W. (2001). Developing conceptual understanding and procedural skill in mathematics: an iterative process. *J. Educ. Psychol.* 93, 346–362. doi: 10.1037/0022-0663.93.2.346

Robitzsch, A., Kiefer, T., and Wu, M. (2020). *TAM: Test analysis modules. R package version, 3*. Online source: https://cran.r-project.org/web/packages/TAM/index.htm

Rolfes, T., and Heinze, A. (2022). "Nur 30 Prozent der Abiturientinnen und Abiturienten erreichen Mindeststandards in voruniversitärer Mathematik!?." In: *Das Fach Mathematik in der gymnasialen Oberstufe*. Eds. T. Rolfes, S. Rach, S. Ufer and A. Heinze (Münster, DE: Waxmann), 237–260.

Schadl, C., and Ufer, S. (2023). Beyond linearity: using irt-scaled level models to describe the relation between prior proportional reasoning skills and fraction learning outcomes. *Child Dev.* 94, 1642–1658. doi: 10.1111/cdev.13954

Scheibling-Sève, C., Pasquinelli, E., and Sander, E. (2020). Assessing conceptual knowledge through solving arithmetic word problems. *Educ. Stud. Math.* 103, 293–311. doi: 10.1007/s10649-020-09938-3

Schmidt, S., and Weiser, W. (1982). Zählen und Zahlverständnis von Schulanfängern. *J. Math.-Didakt.* 3, 227–263. doi: 10.1007/BF03338666

Simon, M. A. (1995). Reconstructing mathematics pedagogy from a constructivist perspective. *J. Res. Math. Educ.* 26, 114–145. doi: 10.2307/749205

Spiro, R. J., Feltovich, P. J., Jacobson, M. J., and Coulson, R. L. (1991). Cognitive flexibility, constructivism, and hypertext: random access instruction for advanced knowledge in ill-structured domains. *Educ. Technol.* 31, 24–33.

Stern, E. (1993). What makes certain arithmetic word problems involving the comparison of sets so difficult for children? *J. Educ. Psychol.* 85, 7–23. doi: 10.1037/0022-0663.85.1.7

Stern, E. (1994). Die Erweiterung des mathematischen Verständnisses mit Hilfe von Textaufgaben. *Grundschule* 26, 23–25.

Stern, E. (1998). *Die Entwicklung des mathematischen Verständnisses im Kindesalter*. Lengerich, DE: Pabst Science Publishers.

Stern, E., and Lehrndorfer, A. (1992). The role of situational context in solving word problems. *Cogn. Dev.* 7, 259–268. doi: 10.1016/0885-2014(92)90014-I

Thevenot, C., Devidal, M., Barrouillet, P., and Fayol, M. (2007). Why does placing the question before an arithmetic word problem improve performance? A situation model account. *Q. J. Exp. Psychol.* 60, 43–56. doi: 10.1080/17470210600587927

Torbeyns, J., De Smedt, B., Stassens, N., Ghesquière, P., and Verschaffel, L. (2009). Solving subtraction problems by means of indirect addition. *Math. Think. Learn.* 11, 79–91. doi: 10.1080/10986060802583998

Ufer, S., and Bochnik, K. (2020). The role of general and subject-specific language skills when learning mathematics in elementary school. *J. Math.-Didakt.* 41, 81–117. doi: 10.1007/s13138-020-00160-5

Ufer, S., and Neumann, K. (2018). "Measuring competencies." In: *International handbook of the learning sciences*. Eds. F. Fischer, C. E. Hmelo-Silver, S. R. Goldman and P. Reimann (New York, US: Routledge), 433–443.

Van Lieshout, E. C., and Xenidou-Dervou, I. (2020). Simple pictorial mathematics problems for children: locating sources of cognitive load and how to reduce it. *ZDM* 52, 73–85. doi: 10.1007/s11858-019-01091-3

Verschaffel, L. (1994). Using retelling data to study elementary school children's representations and solutions of compare problems. *J. Res. Math. Educ.* 25, 141–165. doi: 10.5951/jresematheduc.25.2.0141

Verschaffel, L., De Corte, E., and Pauwels, A. (1992). Solving compare problems: an eye movement test of Lewis and Mayer's consistency hypothesis. *J. Educ. Psychol.* 84, 85–94. doi: 10.1037/0022-0663.84.1.85

Verschaffel, L., Schukajlow, S., Star, J., and Van Dooren, W. (2020). Word problems in mathematics education: a survey. *ZDM* 52, 1–16. doi: 10.1007/s11858-020-01130-4

Vygotsky, L. S., Cole, M., John-Steiner, V., Scribner, S., and Souberman, E. (1978). *Mind in society: the development of higher psychological processes*. Havard, US: Harvard University Press.

Wang, C. (2015). On latent trait estimation in multidimensional compensatory item response models. *Psychometrika* 80, 428–449. doi: 10.1007/s11336-013-9399-0

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika* 54, 427–450. doi: 10.1007/BF02294627

Wildgans-Lang, A., Scheuerer, S., Obersteiner, A., Fischer, F., and Reiss, K. (2020). Analyzing prospective mathematics teachers' diagnostic processes in a simulated environment. *ZDM* 52, 241–254. doi: 10.1007/s11858-020-01139-9

Wright, B. D. (1999). "Fundamental measurement for psychology" in *The new rules of measurement: what every psychologist and educator should know*. eds. S. E. Embretson and S. L. Hershberger (New York, US: Erlbaum), 65–104.

Zorrilla, C., Roos, A.-K., Fernández, C., Llinares, S., and Prediger, S. (2024). Connecting operation-choice problems by the variation principle: sixth graders' operational or deeper relational pathways. *J. Math. Behav.* 73:101104. doi: 10.1016/j.jmathb.2023.101104

frontiers | Frontiers in Education

TYPE Original Research
PUBLISHED 30 May 2024
DOI 10.3389/feduc.2024.1379491

OPEN ACCESS

EDITED BY
Martin Schöfl,
University of Education Upper Austria, Austria

REVIEWED BY
Michael Schurig,
Technical University Dortmund, Germany
Verena Dresen,
University of Innsbruck, Austria

*CORRESPONDENCE
Sarah R. Powell
srpowell@utexas.edu

RECEIVED 31 January 2024
ACCEPTED 09 May 2024
PUBLISHED 30 May 2024

CITATION
Powell SR, Moore CE, Vander Tuin M, Fall A-M and Roberts G (2024) Investigation of the initial feasibility of extended mathematics read-alouds used by kindergarten teachers. *Front. Educ.* 9:1379491.
doi: 10.3389/feduc.2024.1379491

Investigation of the initial feasibility of extended mathematics read-alouds used by kindergarten teachers

Sarah R. Powell*, Clinton E. Moore, Mackenna Vander Tuin, Anna-Mari Fall and Greg Roberts

Department of Special Education, The University of Texas at Austin, Austin, TX, United States

Read-alouds provide teachers with an efficacious method for introducing vocabulary and content to young children. In this study, kindergarten teachers were randomly assigned to implement a mathematics-focused extended read-aloud program called TEMPLE (Teaching Early Math by Providing Language Exploration) with kindergarten children ages 5–6 (19 teachers with 169 children) or to a business-as-usual (BaU) condition (17 teachers and 146 children). With TEMPLE, teachers used read-alouds with storybooks to introduce mathematics vocabulary and content with each read-aloud accompanied by a brief mathematics activity for a period of up to 18 weeks. At posttest, TEMPLE children demonstrated an advantage on two measures of mathematics, but the difference between TEMPLE and BaU children was not significant. In an analysis related to implementation fidelity, we identified a positive and significant association between the number of book readings that TEMPLE teachers reported and scores on one of the mathematics measures. Limited implementation of the program may have led to these results.

KEYWORDS

mathematics, read-alouds, kindergarten, mathematics vocabulary, numeracy

Introduction

Children begin kindergarten with many things in hand: a backpack, crayons, and early mathematics skills. However, upon kindergarten entry, children may already differ in their mathematics knowledge (von Hippel et al., 2018; Throndsen et al., 2020). These differences in mathematics knowledge may impact the later academic success of children. Evidence from the longitudinal data supports that school-entry mathematics skills are predictive of fifth-grade achievement (Claessens et al., 2009). Similarly, Duncan et al. (2007) findings support that mathematics skills at school entry are associated with higher academic performance in later grades. Success in early mathematics requires knowledge of numbers, operations, and vocabulary (Krajewski and Schneider, 2009; Toll and Van Luit, 2014). Due to initial skill differences at the start of kindergarten (von Hippel et al., 2018), children may not have the same chance to develop the interdisciplinary skills they need to succeed in mathematics. To provide children with access to early mathematical language and content, we developed an extended read-aloud routine to be used with mathematics-focused trade books (i.e., story books that can be purchased online or in a bookstore and found in a school or community library, a teacher's classroom, or in the home).

In this introduction, we provide an overview of early mathematics language and review read-aloud efforts in early education settings. Then we discuss the implementation of read-aloud studies. Finally, we provide the purpose and research questions guiding this study.

Early mathematics language

Children learn and communication through language. As children start to participate in school, there is an academic language that children learn. This academic language may overlap with everyday spoken language that children hear in the home and in their community. However, academic language may differ from the language that children use in their home environments. In either case, some children need support with the learning of academic language (Schleppegrell, 2012). Within the language of mathematics, children learn about mathematics through language, whether that language is words (e.g., *three*, *more*, *triangle*), symbols (e.g., 4, =), drawings, or gestures. In this study, we focus on mathematics language presented primarily with words.

Children hear and use mathematics language very early in their lives, and this language is related to an understanding of mathematics (Turan and De Smedt, 2022). For example, Susperreguy and Davis-Kean (2016) determined that high-quality mathematics language used by caregivers at mealtimes was predictive of children' mathematics performance 1 year later. In other research, Levine et al. (2010) noted the caregiver talk about numeracy in the home related to child-level understanding of cardinality.

With preschoolers, Purpura et al. (2017a) identified mathematics language as a mediator between mathematics performance and literacy scores. Hornburg et al. (2018) noted mathematics language shared significant correlations with a variety of preschool mathematics tasks, including counting, numeral identification, cardinality, comparison, ordering numbers, and story problems. Importantly, mathematics language appears to predict later success with mathematics. As such, Toll and Van Luit (2014) identified mathematics language as a predictor of early numeracy growth for a wide distribution of children (i.e., children who demonstrated typical or weak mathematics performance). Purpura and Logan (2015) determined mathematics language as a predictor of later mathematics scores, and King and Purpura (2021) suggested mathematics language explained some of the variance in later mathematics performance in preschool. In fact, mathematics language has been identified as one of the strongest predictors of early childhood mathematical success (Purpura et al., 2017a). Therefore, extra efforts supporting early mathematics language could prove useful for acquiring early mathematical skills. One approach for focusing on early mathematics language is through read-alouds, which we explore in the next section.

Mathematics-focused read-alouds

The combination of mathematics and literacy can be a useful strategy for enhancing mathematical language skill for children. Specifically, mathematics-focused read-alouds can be used to develop mathematical knowledge and build mathematical language (Casey et al., 2008; Van den Heuvel-Panhuizen et al., 2016; Purpura et al., 2017b). In this study, we use the term "read-aloud" to refer to this activity even as it is sometimes referred to as "dialogic reading" or "shared book reading" (Anders et al., 2012; Hassinger-Das et al., 2015). A read-aloud is defined as a book or book passage read aloud to a child or group of children, often by an adult (Alatalo and Westlund, 2021).

An *interactive* read-aloud is defined by Johnston (2016, p. 40) as a practice where teachers or caregivers read text to children while incorporating "pitch, tone, pace, volume changes, questions, and comments to produce a fluent and engaging delivery." These interactive components engage children and promote oral language by eliciting and reinforcing verbal responses (Mol et al., 2009). Additionally, literacy-focused read-alouds help develop content knowledge, grammar, syntax, and vocabulary knowledge (Biemiller, 2001; Santoro et al., 2008; Heisey and Kucan, 2010; Stevens et al., 2010; Strachan, 2015). Similar tenets of engagement and content-focused talk lend themselves to mathematic-specific read-alouds. Within mathematics, read-alouds can enhance mathematics content knowledge as well as mathematics language (Purpura et al., 2021).

While much of the literature about read-alouds has focused on literacy topics, multiple research teams have investigated mathematics-focused efforts (Casey et al., 2008; Green et al., 2018; McGuire et al., 2021). For example, Casey et al. (2008) developed an intervention for kindergarten children using read-alouds to improve geometry concepts like part-whole relations. Teachers implemented the intervention for 63 kindergarten children over 8 sessions. The read-aloud plus geometry intervention group's progress ($n = 35$) was compared to children who received geometry instruction without the read-alouds ($n = 28$). Teachers using the read-aloud program worked from the book, *Tan and the Shape Changer* (Schiro et al., 2002), and systematically introduced geometric concepts to children. Teachers received a brief training session from the researchers where they watched video clips of how to best use the materials and activities in their classes. The read-aloud plus geometry incorporated poems, chants, puppets, an interactive storyline, and activities (i.e., combining isosceles triangles to make the body of a dragon) for children to engage with geometry content. At posttest, children who participated in the read-aloud plus geometry condition significantly outperformed those in the geometry-alone condition on near (ES [effect size] = 0.09) and far transfer (ES = 0.12) tasks about triangles. These findings provided support for the use of interactive storytelling as a method for teaching mathematics content related to geometry.

Also at kindergarten, Hassinger-Das et al. (2015) implemented mathematics-focused read-alouds and compared the read-alouds to mathematics instruction without read-alouds. Their primary goal was to increase kindergarten mathematics vocabulary knowledge of children who exhibited low number sense scores on a screening measure. Children ($n = 124$) from 17 kindergarten classrooms were randomly assigned to participate in small-group instruction using read-alouds, small-group number sense instruction, or a business-as-usual (BaU). In the two conditions with small-group instruction, children participated in 24 sessions (30 min each) with researchers. With the read-alouds condition, researchers used each of seven storybooks over three sessions; researchers read and re-read the books, provided explicit instruction on vocabulary within the books, and engaged children in activities and games related to the mathematics content of the books. With the number sense instruction condition, children engaged in learning about counting, number relations, and operations without read-alouds. At posttest, children who participated in instruction with the read-alouds demonstrated stronger mathematics vocabulary knowledge than children in the other two conditions (ESs = 0.51–0.57). However, the authors identified no significant differences between conditions on a posttest measure of applied problem solving. Children who received number sense instruction demonstrated higher scores on a posttest calculation measure than children in the read-alouds condition or BaU (ESs = 0.58–0.59). Results demonstrated the promise of read-alouds for increasing mathematics-vocabulary knowledge but without transfer to overall mathematics performance.

At preschool, McGuire et al. (2021), investigated the efficacy of a 10-week program focused on early mathematics. Their program, *Booked on Math*, utilized mathematics-focused book readings and inquiry-based lesson plans. The researchers chose 10 books that aligned with the targeted mathematics domains of number concepts and operations, spatial relationships and shapes, comparison and measurement, and patterns. The program was implemented at a university-based family development center with 6 teachers and 74 children. Two of the teachers voluntarily agreed to be in the BaU condition. Teachers who implemented *Booked on Math* did not receive specific training on implementing the program before implementing it with their classrooms. At posttest, across seven mathematics constructs, McGuire et al. noted a significant impact on two of the seven constructs (i.e., quantifies and knowledge of patterns) for the children who participated in *Booked on Math*.

With a focus on children with disabilities, Green et al. (2018) examined the effects of an integrated mathematics and literacy intervention on the early numeracy skills of 50 preschool children (aged 3 to 5) with disabilities (Green et al., 2018). The intervention was implemented by trained researchers and took place 3 times a week for 6 weeks. This quasi-experimental study consisted of two conditions, treatment and BaU. The intervention condition utilized read-alouds and integrated mathematical content through scripted questions. Children engaged in mathematics-related activities after each story was read aloud to them. Children in the BaU were read the same books without the extra mathematical content or activities. At posttest, children who received the intervention significantly outperformed the control group in overall mathematics ability (ES = 0.22), comparison (ES = 0.12), oral counting (ES = 0.15), and one-to-one correspondence (ES = 0 0.13).

While there are more studies about mathematics-focused read-alouds for young children, these studies represent a sample of studies and their approaches to combining mathematics and literacy to promote early numeracy skills. This research establishes how read-aloud interventions can act as a vessel for content such as geometric reasoning (Casey et al., 2008), counting (Green et al., 2018), and shape knowledge (McGuire et al., 2021). As mentioned, many other efforts within this area, particularly with parents and their children, have produced positive results for early childhood mathematics development (Hojnoski et al., 2014; Purpura et al., 2017b; Hendrix et al., 2019; Uscianowski et al., 2020; Purpura et al., 2021).

Purpose and research questions

Because mathematics language is important to concurrent and later mathematics performance (Turan and De Smedt, 2022), we wanted to investigate the initial feasibility of an extended read-alouds program designed to help teachers focus on both mathematics language and mathematics content. Toward this end, we designed the TEMPLE (Teaching Early Mathematics by Providing Language Exploration) project which relied on mathematics-focused read-alouds. We describe this as an *extended* read-alouds program because each TEMPLE teacher used a read-aloud routine paired with a mathematics activity.

With TEMPLE read-alouds, we relied on previous read-aloud efforts but with slight differences in execution and design. Similar to *Booked on Math*, developed by McGuire et al. (2021), we designed TEMPLE for implementation at the whole-class level. While other read-aloud studies employed a quasi-experimental design (Casey et al., 2008; Green et al., 2018), we used a randomized controlled trial in which we assigned teachers to participate in TEMPLE read-alouds or a BaU. Past research has investigated similar mathematics read-alouds with smaller sample sizes (i.e., 50 to 84 children; Casey et al., 2008; Green et al., 2018; McGuire et al., 2021; Purpura et al., 2021). Therefore, we aimed to increase the sample size of participating teachers and children. Furthermore, we developed the TEMPLE read-aloud routine for use with any mathematics-focused trade book by adapting a routine previously developed to increase pre-reading skills (i.e., *Read-Aloud Routine for Building Vocabulary and Comprehension Skills in Prekindergarten*; The Meadows Center for Preventing Educational Risk, 2014). In contrast to McGuire et al. (2021), we trained teachers to implement and use the read-aloud routine with the children in their classrooms. But similar to McGuire et al. (2021), we selected 20 mathematics-focused trade books and considered their match to the Texas mathematics standards for kindergarten as well as each book's mathematics-related vocabulary, quality of mathematics concepts, and appeal to kindergarten children. We planned for each teacher to select one book each week for at least 18 weeks during the school year.

We explored the following research questions. (1) Does participation in the TEMPLE read-alouds lead to improved child outcomes at posttest on two measures (i.e., one researcher created measure and one commercially available measure) of early mathematics? (2) Do TEMPLE children perform better on mathematics measures when they participated in more book readings with their teacher?

Method

Context and setting

Five school districts in a state in the south central of the U.S. participated in this study. A school district is an organization bound by a geographical area that provides educational services to students across preschool through high school. Our 5 school districts oversaw 15 elementary schools (23,000 students in district), 7 elementary schools (9,000 students), 6 elementary schools (8,000 students), 6 elementary schools (7,000 students), and 5 elementary schools (6,000 students).

Across the 5 districts, we recruited 36 kindergarten teachers from 8 elementary schools. Table 1 presents the demographics of the participating school districts. One district with 15 participating teachers was in a mid-sized city on the eastern side of the state. Three districts were in a rural area of the central portion of the state and had

TABLE 1 School district demographics.

	District A	District B	District C	District D	District E
	(*n* = 46)	(*n* = 81)	(*n* = 25)	(*n* = 69)	(*n* = 126)
Gender (%)					
Female	49.0	49.0	49.0	50.0	48.0
Male	51.0	51.0	51.0	50.0	52.0
Race/ethnicity (%)					
Asian	13.3	1.1	1.9	0.3	1.1
Black	0.8	3.0	13.4	2.5	28.7
Hispanic	15.2	65.5	46.1	79.1	43.1
White	64.6	27.4	32.1	16.4	24.0
Two or More	6.1	2.8	6.3	1.4	3.0
Other	0.3	0.2	0.3	0.3	0.2
Free/reduced lunch (%)					
Yes	3.6	46.1	39.2	77.5	85.8
With disability (%)					
Yes	8.4	13.5	15.9	11.4	11.9

a combined total of 17 teachers. All three of these districts met the federal definition for rural school districts (i.e., census-defined rural territory more than 25 miles from an urbanized area). One district, with four teachers, was in a suburban area adjacent to a large metropolitan city in the central portion of the state.

Participants

Before beginning the study, we received Institutional Review Board approval from our university. All 36 teachers signed written consent forms for participation. Teachers sent home consent forms to the caregivers of all children in each of their classrooms. We collected caregiver consent from 347 children ages 5 to 6 years old. Blocking by school, we randomly assigned 19 teachers to implement the TEMPLE Read-Aloud routine ($n = 179$ children) or 17 teachers to participate in kindergarten BaU ($n = 154$ children). Of the classrooms randomized, none dropped out of the study.

Teachers as implementers

Before the program began, we provided one 2 h professional development session to TEMPLE teachers in a synchronous virtual workshop. The virtual workshop allowed teachers to ask questions and clarify important aspects of the project in a collaborative environment. The session was recorded and provided to teachers so they could reference it later. The workshop had the dual purpose of introducing teachers to the idea of mathematics as a language-based activity and teaching them how to use the TEMPLE Read-Aloud routine. A key feature of the TEMPLE Read-Aloud routine was that teachers used mathematics-focused children's books to increase children's familiarity with mathematics vocabulary. We provided each teacher with a binder that contained TEMPLE Read-Aloud routine materials we developed for use with any mathematics-focused children's trade book and a set of 20 books we selected for use during the study period. Table 2 provides an overview of the 20 books. We asked teachers to try to use the TEMPLE Read-Aloud routine for 18 weeks (with 18 books).

During this initial workshop, we addressed the following. (a) We discussed the predictive nature of kindergarten mathematics skills on future academic success in both mathematics and reading (Pagani et al., 2010). (b) We presented research indicating that proficiency with mathematical language is an underlying factor that facilitates both numeracy and reading skills (Purpura et al., 2017a). (c) We highlighted categories of mathematics-related terminology [e.g., terms related to precise quantities (*one*, *two*, *three*), imprecise quantities (*a little bit*, *many*, *all*), spatial location (*in*, *out*, *above*, *below*), movement or direction (*come*, *go*, *into*, *left*, *right*), comparisons (*more*, *less*, *equal*, *bigger*), and measurement (*minute*, *inch*, *pound*), etc.]. (d) We provided examples of common mathematics-related vocabulary terms that may cause confusion. (e) We discussed various groups of mathematics-related terminology represented on a list of early sight words. (f) We demonstrated how to use the TEMPLE Read-Aloud routine to introduce and explore mathematics content and mathematics vocabulary terms in the kindergarten classroom.

Components of the TEMPLE read-aloud routine

When using the weekly TEMPLE Read-Aloud routine, children participated in interactive read-alouds and activities with one mathematics-focused book over 3 days each week. It was intended for the classroom teacher to use the extended read-aloud routine for approximately 45 min per week of classroom time over those 3 days (i.e., about 20 min, 15 min, and 10 min for the first, second, and third days, respectively).

We selected books that covered as many of the objectives in the Mathematics Texas Essential Knowledge and Skills at Kindergarten as possible. Objectives in this mathematics standards document

TABLE 2 Mathematics-focused read-alouds used by TEMPLE teachers.

Book	Kindergarten content
20 Big Trucks in the Middle of the Street (Lee, 2013)	Counting (counting forward 1–20)
100 Bugs! A Counting Book (Narita and Kaufman, 2018)	Counting to 100
A Dollar, a Penny, How Much and How Many? (Cleary and Gable, 2012)	Identifies different U.S. coins and bills; some addition
Arithmechicks Add Up: A Math Story (Stephens and Liu, 2019)	Adding numbers between 1 and 10
Arithmechecks Take Away (Stephens and Liu, 2020)	Taking away using 0–10
Captain Invincible and the Space Shapes (Murphy and Simard, 2001)	Geometry, 2- and 3-dimensional shapes
Domino Addition (Long, 1996)	Adding to and taking away
Elevator Magic (Murphy and Karas, 1997)	Taking away between 0 and 10
How Long or How Wide? A Measuring Guide (Cleary, 2007)	Measurable attributes of different objects
How Many Snails? (Giganti and Crews, 1988)	Counting with different objects
How Tall? Wacky Ways to Compare Height (Weakland and Sinkovec, 2014)	Measures and compares objects
Money Plan (Eaton and Symone, 2021)	Money management
Round is a Tortilla: A Book of Shapes (Thong and Parra, 2013)	Identifies 2-dimensional shapes
Shaping Up Summer (Flatt and Barron, 2014)	Geometry, identifies 2- and 3-dimensional shapes
Sheep Won't Sleep: Counting by 2s, 5s, and 10s (Cox and Cuneo, 2017)	Counting to 20; counting by set to 100
Super Sand Castle Saturday (Murphy and Gorton, 1999)	Uses different objects to measure and compare
The Animals Would Not Sleep! (Levine, 2020)	Identifying and sorting objects into 3 main categories
The Four Money Bears (Gardner, 2015)	Emphasizes financial literacy, investment, and saving money
The Great Graph Contest (Leedy, 2006)	Identifies and categorizes different graphs
The Penny Pot (Murphy, 1998)	Introduces coins with addition

included: represent and compare whole numbers; describe the relative positive and magnitude of whole numbers; explore relationships within the number system; develop an understanding of addition and subtraction; identify coins; recite numbers to 100 by ones and tens; analyze attributes of two-dimensional shapes and three-dimensional figures; compare measurable attributes; and collect and organize data to interpret information.

For each book, we also considered the mathematics-related vocabulary terms presented, the probable appeal to kindergarten children, and the quality of the presentation of mathematics concepts (e.g., illustrations supported mathematical understanding or the book contained multiple exposures to a concept). Figure 1 provides an overview of the three-day TEMPLE framework used each week.

Day one of TEMPLE read-aloud routine

For the first session each week, the teacher introduced the book and three to five target vocabulary terms from the story. Each week, the classroom teacher selected the target vocabulary terms from the book based on a list of suggested early mathematics terminology. For example, for *20 Big Trucks in the Middle of the Street,* vocabulary terms and child-friendly definitions may have included: (a) *middle*: in between two things, (b) *count*: find the number, (c) *big*: not small but large, and (d) *small*: not big but little. Teachers received a set of vocabulary cards (see Figure 2) they could use with the book. The teacher provided both a kindergarten-friendly definition and a non-verbal representation (e.g., a hand gesture, picture, or object) for each term, and invited children to listen for these words in the story and to give a thumbs-up (or other gesture) when the words were read by their teacher.

Based on best practice recommendations for read-alouds (Giroir et al., 2015), teachers read the selected book straight through the first time, without stopping for discussion, to allow children to focus on understanding the storyline. After completing this first reading, children talked to one another about the story and what they learned with prompting, if needed, from the teacher. The same day, either immediately or later in the day, the teacher read the book a second time, pausing to interact with children about the story, the selected vocabulary terms, and the mathematics content addressed. During this second reading, the teacher also showed children how the pictures supported the story and the mathematics content. After the second reading, the teacher and children engaged in a conversation about how the story and the mathematics concepts addressed related to the children's everyday lives. Within the routine materials, we provided suggestions on common ways kindergartners use mathematics for each strand in the preschool guidelines (see Figure 2). We asked teachers to elicit ideas from the children in their classroom and to add their own ideas about how everyday connections can be made to the book they read each week.

Finally, children participated in a hands-on activity targeting the mathematics-related terms and concepts in the story. This activity gave children opportunities for further exploration of the concepts and vocabulary terms encountered during the read-aloud time (Massey, 2013). We developed a variety of activity ideas and included these suggestions in the TEMPLE Read-Aloud materials (see Figure 3). We encouraged teachers to use our suggestions as a starting point but to be creative when planning activities related to each book (i.e., a different activity for each day of the routine week, for a total of three extension activities across the week). We emphasized incorporating playfulness, concrete items, and body

Framework for Delivery of the Read-Alouds

After planning for the week, use the following general framework for each book.

Day 1

- Introduce the book.
- Introduce the target math concept.
- Introduce three to five new vocabulary words from the book.
- **Read the book**.
- Review the new vocabulary words.
- **Reread the book**, stopping at each target vocabulary word and illustration to clarify concepts and check for understanding.
- Explore new concepts and build skill with a hands-on activity.
- Use the skill as you go through the day.

Day 2

- Review concepts and vocabulary from the previous day's reading.
- **Reread the book**, stopping at each target vocabulary word and illustration to clarify concepts and check for understanding.
- Reinforce the new skill with another hands-on activity.
- Use the skill as you go through the day.

Day 3

- Review the target vocabulary words.
- Reinforce the new skill with another related math activity.
- Ask students how they used their new math skills throughout the week.

FIGURE 1
Overview of TEMPLE read-aloud routine.

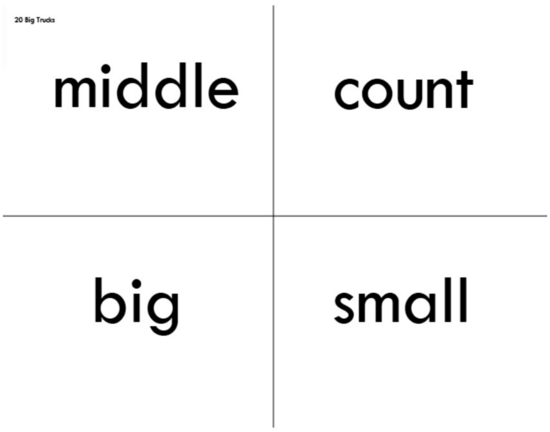

FIGURE 2
Vocabulary cards.

movements into the activities. We also encouraged teachers to use items already in the classroom to help children connect the mathematics-related content and vocabulary terms to children's everyday lives. We did not plan specific activities for each book because we designed the TEMPLE Read-Aloud routine to be a framework to be used with any mathematics-focused children's book. We intended to support teachers in developing their own skills at helping kindergartners connect to the mathematics-related content and vocabulary terms in any mathematics-focused children's book.

Day two of TEMPLE read-aloud routine

On the second day of the extended read-aloud routine, teachers read the story for a third time, again pausing to interact with the children about the story, pictures, mathematics-related vocabulary terms, and mathematical content. Children then participated in a second hands-on activity connecting to the story's mathematics content and target vocabulary terms.

Day three of TEMPLE read-aloud routine

For the third day in the weekly routine, teachers provided a third hands-on activity and had the option of rereading the book depending upon child interest. To strengthen the kindergartners' connections to target vocabulary terms and mathematics concepts, we asked teachers to leave related materials in their mathematics-play center over the week and to make connections to the target terminology and mathematics content as the children went through their everyday routines and activities.

Business-as-usual

Teachers in the BaU continued their kindergarten mathematics instruction with the children in their classroom as originally planned. We did not provide any training to the BaU

Sample Lesson Outline for Kindergarten

Super Sandcastle Saturday by Stuart J. Murphy

Target Math Concept

Use the standards to select a target math concept that can be taught using this book.

Vocabulary Suggestions

Select three to five of the following (use gestures as appropriate):

- tallest: Having the greatest height, the highest
- measure: To find out the size
- longest: The greatest distance from end to end
- deepest: Farthest down from the top
- up: The direction toward the sky
- inch: A standard unit of measure

Suggested Questions for Discussion

- Who had the deepest moat?
- Who had the longest wall?
- What is more correct, measuring the wall of the tower with inches or using your feet?
- Why couldn't Sarah, Juan, and Laura use shovels to measure how deep their moat was?
- Which tower measured the tallest in inches on the ruler, Sarah's or Juan's?
- Why is it important to use a ruler in measuring how tall, long, and deep something is?

Activity Ideas

How Tall Am I?

- Pair each student with a partner.
- Take 2 to 5 minutes for each group to select a small object in the classroom to use as a measuring tool.
- Take 10 to 15 minutes for partners to measure one another using their object and record their results.
- Extension activity: Ask the groups to go around the classroom measuring the height of objects. For example, ask students how many pencils tall the door is.

Tower Contest

- Gather materials to stage a tower-building game like the one in *Super Sandcastle Saturday*: wooden or plastic blocks and a timer.
- In the first round, see who can make the tallest tower.
- In the second round, see who can make the longest tower.

FIGURE 3
Teacher activity guide.

teachers, and they did not receive any TEMPLE materials from our research team.

Fidelity of implementation

We collected data on fidelity of TEMPLE implementation via survey and direct observations. Each week, we asked each TEMPLE teacher to complete a self-report survey via Qualtrics. Note, every week we sent each teacher an email about filling out the survey but did not follow-up when teachers did not complete the survey. We made this decision because of restrictions related to visitors in school buildings and to see how well teachers implemented the program without much interference from the research team. On the survey, TEMPLE teachers (a) named the book they used during the week; (b) described the number of times they read the book with their class; (c) provided a rating of the book on a scale of 1 (would not recommend this book) to 10 (would highly recommend the book); (d) explained the strengths of the book and activities; and (e) described the weaknesses of the book and activities.

We also conducted direct fidelity observations of those TEMPLE teachers who worked at participating schools within an hour's proximity to our university ($n = 11$) in Districts A, B, C, and D. Research staff developed an observation form that could be used to gather observational data specific to each day's expectations of the TEMPLE read-aloud routine (i.e., Day 1, Day 2, Day 3). For instance, if a teacher was doing the read-aloud for the first time that week (i.e., Day 1), the observer noted whether the teachers activated prior knowledge and introduced the vocabulary terms according to the procedure in the routine. If a teacher was observed reading the book on the second or third day that week, the observer noted how vocabulary was reviewed and how the connecting questions and enrichment activities were implemented.

Measures

We tested consented children in both conditions at pretest and posttest. We used one brief measure with all children: the *Kindergarten Texas Early Assessment of Mathematics* (KTEAM; Powell and Moore, 2021). We also randomly selected and tested approximately four children in each classroom using the *Test of Early Mathematics Ability*, 3rd edition (TEMA-3; Ginsberg and Baroody, 2003). We tested these same randomly selected children with the TEMA-3 at pre- and posttest. Children were not administered the KTEAM and the TEMA-3 on the same day.

KTEAM

We developed the KTEAM (Powell and Moore, 2021) to measure child knowledge of the mathematics objectives listed in the Texas Essential Knowledge and Skills (TEKS). These objectives fell into the categories of number and operations, algebraic reasoning, geometry and measurement, data analysis, and personal financial literacy. The KTEAM had two parts. In Part A, examiners tested children in a whole class setting. Children received a paper booklet and drew or circled responses to 37 items in the booklet. The questions focused on an understanding of mathematics vocabulary with questions that asked about *more* (2 questions), *most, greatest, less, least* (3 questions), *taller, longest, shortest, one more, one less, equal, add, subtract, take away, group of 10, penny, nickel, dime, quarter, pattern* (2 questions), *rectangle* (2 questions), *triangle* (2 questions), *square, circle, sphere, cone, cylinder, cube,* and three questions about number writing of *three, sixteen,* and *forty-five.* Some questions asked children to circle an option (e.g., "Circle the jug that holds the least water."). Other questions asked children to draw (e.g., "Draw a triangle.") or write ("What's next in this pattern? 1 3 5 1 3 5 1 3 5 ___). In Part B, examiners tested children individually on an additional 10 items. For four items, children looked at pictures and named the number of dots, apples, or bears. On two items, children counted forward from 1 to 20 or backwards from 20 to 1. The final four questions were orally presented word problems (e.g., "You have 3 pennies. I have 8 pennies. How many pennies do we have together?"). Children earned one point for each correct response for a total possible score of 47. At pretest, the range of scores was from 10 to 45. At posttest, the range was from 16 to 47, with only one student scoring a perfect 47. Cronbach's alpha for this sample was 0.82.

TEMA-3

We used the commercially available TEMA-3 (Ginsberg and Baroody, 2003) to measure mathematics performance across a broad range of early numerical skills including: subitizing, counting, discerning more and less, addition and subtraction, recognizing and writing numerals, and solving word problems with manipulatives. Examiners read from the TEMA-3 administration booklet and administered this measure individually to children. On the TEMA-3, children demonstrated their mathematics understanding using fingers, pictures, manipulatives, pencil and paper, and verbal responses. Testing began at a point determined by the child's age and stopped when a child scored five consecutive incorrect responses. Because of time constraints, we randomly selected four children in each classroom to take the TEMA-3. In classrooms with four or fewer consented children, we tested all children with consent. Reliability was reported as 0.98. Test–retest reliability, as reported by the manual, was 0.82 to 0.93.

Data collection and scoring

The second author and three additional examiners administered the two measures. We trained the examiners on administration of the KTEAM in one 2 h initial training session and, later, one 1 h review session. The four examiners responsible for testing also administered the more complex TEMA-3, and we provided one 2 h training and practice of the administration of this measure. We required each examiner to demonstrate accuracy in scoring before administering the measures to children in school settings.

We double coded all scoring for the KTEAM and TEMA-3. For these measures, two examiners, working independently, scored measures, and entered data into two separate databases. Upon the completion of data entry, we ran a comparison of the two databases to identify the discrepancies between the two sets of scores. We used the discrepancies to calculate the inter-rater reliability of the scoring as 96.6%. We resolved all discrepancies to ensure almost 100% accuracy in the database.

Procedure

We recruited districts and trained teachers in the early fall of the school year. Pretesting of children occurred in November and December. We asked TEMPLE teachers to use the TEMPLE read-aloud routine in their classrooms for up to 18 weeks, selecting one of the 20 mathematics-focused children's books each week. Posttesting of all children occurred in April and May.

Data analysis

Children were nested within teachers and teachers were nested within schools. Given the clustered nature of the data, we used multilevel models to examine the effect of TEMPLE read-alouds on child-level outcomes. Multilevel models account for dependencies in nested data structures by estimating residual components at each level and by partitioning total variance into its level-specific component parts, yielding unbiased parameter estimates (Singer and Willett, 2003). We estimated treatment effects at the teacher-level using dummy codes for experimental conditions (1 = TEMPLE, 0 = BaU). We included pretest scores as covariates to optimize statistical power and to aid the interpretation of model parameters (Bloom et al., 2007). Specifically, child-level scores were centered on teacher means (Enders and Tofighi, 2007) in level 1 of the model and class-level means were centered on the grand mean and included as a teacher-level covariate to optimize precision (Hoffman, 2015). The reduced-form equation is:

$$Outcome_{ijk} = \gamma_{000} + \gamma_{100}\left(Pretest\right)_{ijk} + \gamma_{010}\left(Pretest\right)_{jk} + \gamma_{020}Temple_{jk} + e_{ijk} + r_{jk} + u_{k}.$$

Here, i represents children, j represents classrooms, and k represents schools. Parameter γ_{000} refers to the average child-level outcome across all classrooms and schools; γ_{100}represents child-level pretest scores centered around the classroom mean; γ_{010} is the classroom-level pretest aggregate centered around the grand mean; $Temple_{jk}$ is classroom-level dummy variable representing TEMPLE program; residuals e_{ijk}, r_{jk}, and u_k are Level 1, 2, and 3 random effects, respectively. We estimated effect sizes (Hedges' g) for the main effect of TEMPLE as the ratio of the model-derived treatment coefficient and the unadjusted pooled within-group standard deviation across conditions at posttest (What Works Clearinghouse, 2020).

Our second research question focused on the association between the number of book readings by teachers and children's math achievement. We evaluated this research question using data from treatment children only because this information was not collected from the teachers in the BaU classes. The three-level model failed to converge, and we removed the random intercept at the school level and estimated a two-level model with the children at Level 1 and teachers at Level 2. The clustering of teachers within schools was accounted for using the design-based correction of standard errors implemented in Mplus (type = complex; McNeish et al., 2017). We estimated the effect of number of book readings at the teacher-level with child-level pretest data centered on classroom means as a level 1 covariate and with classroom-level aggregates centered on the grand means at level 2. We fit models using Mplus 8.7 (Muthén and Muthén, 2023). We handled missing data using the full information maximum likelihood (FIML) estimation, which uses all available information in the data and provides unbiased results with either missing completely at random or missing at random data (Enders, 2010).

Results

Fidelity of TEMPLE implementation and social validity

We had complete TEMPLE survey data in 144 instances (with 17 of the 19 TEMPLE teachers responding at least once; $M = 8.0$ surveys, $SD = 4.3$); range of completed surveys by teacher was 1 to 15. During implementation of TEMPLE, we encouraged teachers to use 18 of the 20 books in any order. Teachers reported reading each book an average of 2.5 times during the week ($SD = 0.9$). Over the course of TEMPLE, teachers reported that children participated in 21.2 read-alouds ($SD = 9.0$), with a range of 4 to 39 reported total read-alouds.

We collected in-person fidelity of TEMPLE implementation from 11 (of 19) TEMPLE teachers on 32 occurrences, with 11 observations for Day 1, seven for Day 2, and 14 for Day 3. The observation checklist for Day 1 contained 10 items, the Day 2 checklist contained four items, and the checklist for Day 3 contained five items. For Day 1 TEMPLE implementation, we noted 78.9% fidelity. Day 2 adherence was 89.3%. The fidelity for teachers we observed implementing Day 3 was 88.8%.

On the weekly survey, we asked teachers to rate each book on a scale of 1 (terrible) to 10 (amazing), with an average score of 6.6 ($SD = 2.8$) across the books. When teachers provided strengths of the books and activities, they mentioned "They liked the different shoes and they LOVED hearing the words they had not heard before," "They like anything with cars and trucks," "It was engaging and easy for students to understand," "The students loved the illustrations," "The loved how they used silly objects to measure the height of the objects and they liked the activities that went with the lesson," and "This book was super engaging and made it fun and easy to teach multiple concepts!" When teachers gave books a lower rating, they provided comments such as "the book was too long for this age group," "This book did not seem to keep the students' attention as well," or "It included other math/science terms that our students were not familiar with."

Attrition, baseline equivalence, and variance

Overall attrition ranged from 1.2 to 5.4% at posttest. Rates of attrition were comparable across treatment conditions. At posttest, the difference in attrition rates ranged from 0.4 to 2.4% across outcomes. After the posttest, the potential for bias due to attrition is considered acceptable when following the cautious guidelines provided by the What Works Clearinghouse (2020). The two groups were very similar on pretest scores (Hedges' g ranged from 0.03 to 0.07; p-values ranged from 0.76 to 0.77), suggesting equivalence prior to onset of TEMPLE.

We estimated an unconditional three-level model with random intercepts to estimate the intraclass correlation (ICC) and to examine the proportion of total variance due to differences between teachers and schools. The estimate of ICC indicated most of the total variance was at the child-level (0.83 for KTEAM and 0.77 for TEMA-3) with smaller amounts at the teacher-level (0.11 for KTEAM and 0.06 for TEMA-3) and the school-level (0.06 for KTEAM and 0.17 for TEMA-3). We fit a three-level conditional model to estimate the main effect of TEMPLE.

Main effect of TEMPLE

Table 3 displays observed pretest and posttest means and standard deviations for each measure by treatment conditions. The estimated effects of the TEMPLE program on mathematics outcomes are reported in Table 4. Children in classes assigned to the TEMPLE program outperformed children in the control classrooms; however, the difference was not significant on the TEMA-3 (ES = 0.11, 95% CI [−0.19, 0.42]) or on the KTEAM outcome measures (ES = 0.06, 95% CI [−0.07, 0.20]).

Effect of number of book readings on child outcomes

Next, we examined the number of book readings completed by TEMPLE teachers (i.e., the number of times a TEMPLE teacher reported reading any books during the TEMPLE program). We calculated this based on the weekly surveys that teachers completed with the survey question of "how many times did you read the book this week?" We investigated whether the number of total book readings by a TEMPLE teacher was associated with child outcomes, and we estimated a multilevel model with total number of book readings as the between-teacher predictor. Results presented in Table 5 indicate a positive and significant association between number of book readings and children's scores on the TEMA-3 (γ_{01} = 0.05, SE = 0.02, p = 0.01). It is important to note that although the regression coefficient was significant, the ES was not (ES = 0.34, 95% CI [−0.30, 0.98]).

Our analysis suggests that there may be an effect of the total number of book readings on TEMA-3 scores, as indicated by an ES of 0.34. However, it's important to interpret these results with caution due to the wide confidence interval, which ranges from −0.30 to 0.98. This wide range suggests considerable uncertainty surrounding the estimated effect. Given our small sample size (n = 38), it is plausible that our analysis lacks the statistical power to detect effects. Therefore, while the observed effect provides some evidence for the impact of number of book readings, further research with larger sample sizes is needed to provide a more conclusive finding. The association between number of book readings and KTEAM scores was positive but not significant (γ_{01} = 0.01, SE = 0.01, p = 0.36, ES = 0.14, 95% CI [−0.38, 0.66]).

Discussion

In this study, we investigated the initial feasibility of mathematics-focused extended read-alouds (i.e., TEMPLE) implemented by kindergarten teachers with their entire classrooms of kindergarten children. First, we examined the impact of the TEMPLE Read-Alouds on two measures of mathematics, and we compared children who participated in TEMPLE versus children in a BaU. Based on those results, we investigated whether teachers who reported book use of TEMPLE materials had differential results within the TEMPLE condition.

TABLE 3 Means and standard deviations at pretest and posttest.

	Pretest			Posttest		
	n	*M*	SD	*n*	*M*	SD
KTEAM						
BaU	154	30.25	6.37	146	36.10	6.17
TEMPLE	179	29.93	6.44	169	36.19	5.63
TEMA-3						
BaU	40	30.18	10.44	40	36.53	9.50
TEMPLE	42	31.02	9.95	41	38.71	9.05

KTEAM, Kindergarten Texas Early Assessment of Mathematics; BaU, Business as usual; TEMPLE, Teaching Early Math by Providing Language Exploration; TEMA-3, Test of Early Mathematics Ability (3rd Edition).

Effect of TEMPLE read-alouds

With our first research question, we asked whether teacher and child participation in TEMPLE led to improved child outcomes at posttest on two different mathematics measures. We determined baseline equivalence between the two conditions (i.e., TEMPLE and BaU); therefore, we could attribute any posttest differences to participation in TEMPLE. Our analysis, however, suggested no significant difference on either the KTEAM (for which we had scores for all children) or the TEMA-3 (for which we had scores for approximately 4 children in each classroom). We noted that the TEMPLE children had greater posttest scores than BaU children on both measures, but these scores, while positive, were not significant.

Our effect size on the TEMA-3 of 0.11 was fairly similar to that in other read-aloud research studies, such as Casey et al. (2008, ESs = 0.09, 0.12) and Green et al. (2018, ESs = 0.12, 0.13, 0.15, 0.22). But this was not similar to the effects (0.51, 0.57) of Hassinger-Das et al. (2015), perhaps because the outcome in Hassinger-Das was a mathematics language measure whereas our KTEAM emphasized both mathematics vocabulary and mathematics content. In a preschool evaluation of TEMPLE (Powell et al., under review), we determined the TEMPLE read-aloud routine as efficacious for improving child-level mathematics outcomes. That we did not replicate the TEMPLE results from preschool (i.e., children ages 3 and 4) in kindergarten (i.e., children ages 5 and 6) could indicate that read-alouds are not as appropriate in kindergarten as preschool; but prior research shows that is not the case. For that reason, we turned our attention – with our second research question – toward teacher implementation of TEMPLE as an explanation for this difference in results.

TABLE 4 Estimated effects of the intervention.

	KTEAM				TEMA-3			
Fixed effects	Est.	*SE*	*p*	ES [95% CI]	Est.	*SE*	*p*	ES [95% CI]
Intercept	35.82	0.53	0.00		36.93	0.86	0.00	
Pretest L1	0.72	0.03	0.00		0.67	0.06	0.00	
Pretest L2	0.68	0.11	0.00		0.91	0.11	0.00	
TEMPLE	0.36	0.40	0.37	0.06 [−0.07, 0.20]	1.33	1.31	0.31	0.11 [−0.19, 0.42]
Random Effect	Est.	ICC			Est.	ICC		
L1 variance	13.68	0.93			24.45	0.93		
L2 variance	0.08	0.01			1.64	0.06		
L3 variance	1.02	0.07			0.19	0.01		

KTEAM, Kindergarten Texas Early Assessment of Mathematics; TEMA-3, Test of Early Mathematics Ability (3rd Edition).

TABLE 5 Estimated effects on the number of book readings by teachers.

	KTEAM				TEMA-3			
Predictors	Est.	*SE*	*p*	ES [95% CI]	Est.	*SE*	*p*	ES [95% CI]
Intercept	36.45	0.73	0.00		38.76	0.69	0.00	
Pretest L1	0.66	0.05	0.00		0.72	0.11	0.00	
Pretest L2	0.72	0.13	0.00		0.91	0.14	0.00	
Book readings	0.01	0.01	0.36	0.14 [−0.38, 0.66]	0.05	0.02	0.01	0.34 [−0.30, 0.98]
Random Effect	Variance	ICC			Variance	ICC		
L1 variance	13.55	0.92			24.97	1.00		
L2 variance	0.07	0.00			0.06	0.00		
L3 variance	1.13	0.08						

KTEAM, Kindergarten Texas Early Assessment of Mathematics; TEMA-3, Test of Early Mathematics Ability (3rd Edition).

Effect of TEMPLE read-alouds with fidelity

With our second research question, we pursued the idea that teacher implementation of TEMPLE may have had a differential impact on child-level mathematics outcomes. Our analysis suggested that fidelity of TEMPLE implementation had some effect of the results from our study. When teachers reported doing more book readings, the association between the number of readings and scores on the TEMA-3 was both positive and significant. We saw the same positive trend for scores on the KTEAM yet this was not significant.

It was important to investigate this research question for several reasons. First, we relied on teachers to fill in a brief weekly survey to report which book they read during the week. Even with a few gentle reminders from the research team, teachers did not submit surveys regularly. In fact, two teachers never submitted a single survey, but 17 teachers submitted a total of 144 surveys. The average number of returned surveys from each teacher was 8, which was less than half of the expected number of weeks of TEMPLE implementation (i.e., 18 weeks).

In prior read-aloud research, researchers often implemented the read-alouds in classrooms (e.g., Hassinger-Das et al., 2015; Purpura et al., 2017b; Green et al., 2018). Given that researchers would drive to schools for the project, all read-aloud sessions would have been implemented as planned. As such, both Green et al. (2018) and Hassinger-Das et al. (2015) measured fidelity of implementation of the researchers (not teachers) who implemented their read-aloud intervention. Fidelity for such studies is important to learn about whether the program was implemented and the quality of that implementation (O'Donnell, 2008). In Casey et al. (2008), kindergarten teachers implemented the read-alouds with one book for eight sessions. However, the authors provided no information about fidelity of intervention implementation. This is similar to no information about fidelity from the classroom teachers who implemented read-alouds in McGuire et al. (2021). While our return rate from the weekly teacher surveys was underwhelming, it does provide a snapshot as to how many teachers implemented a read-alouds program when provided with initial training but then, for the most part, left to implement the program on their own.

Second, in the small number of in-person fidelity observations we conducted, we noted fidelity for Day 1 of each week's TEMPLE implementation to average 78.9%. This percentage is almost identical to Peltier et al. (2021) study in which a classroom teacher implemented a word-problem intervention with 78.6% fidelity. We would expect this percentage to be higher, particularly for the day in which each book and its vocabulary would be introduced. Our fidelity rate might be acceptable, but there is room for improvement. As fidelity of implementation for mathematics interventions has been identified as an important predictor of child-level outcomes (Nelson et al., 2019), not only should future research investigate whether teachers are doing the TEMPLE read-alouds but whether they are using the materials effectively.

Third, we provided all TEMPLE teachers with a brief (i.e., 2 h) professional development workshop about the TEMPLE read-alouds and routine. Teachers attended this session before implementation. Due to constraints on school visits, we did not provide any follow-up workshops or embedded coaching. As described in the literature, coaching is often an integral companion to professional development (Rudd et al., 2009; Wood et al., 2016) which may or may not be important to embed in future iterations of this research (Kennedy, 2016). Beyond including follow-up professional development and coaching, however, teachers, schools, and districts may need to be open to a change in routines and practices for teachers to incorporate a new practice into their regular weekly teaching routine (Thoonen et al., 2011).

Limitations

Before concluding, we note several limitations to this research study. First and foremost, we did not have regular implementation of TEMPLE from all teachers. Teachers did not fill in surveys regularly, even with a weekly email reminder from our research team. We also did not conduct regular in-person checks of implementation fidelity. Much of our hands-off approach was due to COVID-related protocols still in place in schools with schools excluding or limiting visitors in the school building. But without this information, we do not know if TEMPLE teachers were implementing TEMPLE and not filling in the weekly survey or if they were not implementing TEMPLE at all. We also have a limited understanding of the mathematics activities that teachers used alongside each book. Future research should focus on the initial efficacy of the extended TEMPLE Read-Alouds when implemented with high levels of fidelity. Furthermore, future research should investigate TEMPLE Read-Alouds with a prescribed mathematics activity for each session or a teacher-selected activity. And if teachers select the activity to accompany each book reading, more information should be collected on the type of activity, duration of activity, and emphasis on vocabulary within each activity. In the future, researchers may also want to collect observational data on how children used the target vocabulary terms within the TEMPLE Read-Alouds program and across the school day.

Second, we did not administer the TEMA-3 to all participating children. We randomly selected four children in each classroom to participate in the TEMA-3 to limit the in-person interaction with our research staff in school buildings. In the future, we would administer this assessment to all children as it is a standardized assessment of early mathematics knowledge. Third, our KTEAM measure produced higher average scores at pretest, which may have contributed to ceiling effects at posttest. In the future, we would revise this assessment and include more mathematics vocabulary terms appropriate for the early elementary grades.

Conclusion

We learned of no differences between TEMPLE children and BaU children after participation in TEMPLE read-alouds; however, we did see a significant different between conditions when we considered the number of book readings conducted by TEMPLE teachers. If we were to describe this study, we would suggest that we conducted an initial feasibility study in which we learned about implementation fidelity, or lack thereof. To determine the efficacy of the extended TEMPLE read-alouds for kindergarten children, we would need to provide better training to teachers, and we would need to better monitor fidelity related to dosage and quality of the instruction.

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 humans were approved by The University of Texas at Austin Institutional Review Board. The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation in this study was provided by the participants' legal guardians/next of kin. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

Author contributions

SP: Writing – original draft, Writing – review & editing. CM: Project administration, Writing – original draft. MVT: Writing – original draft. A-MF: Formal analysis, Writing – original draft. GR: Formal analysis, Methodology, Writing – original draft.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This research was supported in part by a grant from the T.L.L. Temple Foundation.

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

Alatalo, T., and Westlund, B. (2021). Preschool teachers' perceptions about read-alouds as a means to support children's early literacy and language development. *J. Early Child. Lit.* 21, 413–435. doi: 10.1177/1468798419852136

Anders, Y., Rossbach, H.-G., Weinert, S., Ebert, S., Kuger, S., Lehrl, S., et al. (2012). Home and preschool learning environments and their relations to the development of early numeracy skills. *Early Child. Res. Q.* 27, 231–244. doi: 10.1016/j.ecresq.2011.08.003

Biemiller, A. (2001). Teaching vocabulary: early, direct, and sequential. *Am. Educ.* 25, 24–28, 47. Available at: https://www.aft.org/ae/spring2001/biemiller

Bloom, H. S., Richburg-Hayes, L., and Black, A. R. (2007). Using covariates to improve precision for studies that randomize schools to evaluate educational interventions. *Educ. Eval. Policy Anal.* 29, 30–59. doi: 10.3102/0162373707299550

Casey, B., Erkut, S., Ceder, I., and Young, J. M. (2008). Use of a storytelling context to improve girls' and boys' geometry skills in kindergarten. *J. Appl. Dev. Psychol.* 29, 29–48. doi: 10.1016/j.appdev.2007.10.005

Claessens, A., Duncan, G., and Engel, M. (2009). Kindergarten skills and fifth-grade achievement: evidence from the ECLS-K. *Econ. Educ. Rev.* 28, 415–427. doi: 10.1016/j.econedurev.2008.09.003

Cleary, B. P. (2007). *How long or how wide?: a measuring guide*. Kildare, Ireland: Millbrook Press.

Cleary, B. P., and Gable, B. (2012). *A dollar, a penny, how much and how many?*. Kildare, Ireland: Millbrook Press.

Cox, J., and Cuneo, N. (2017). *Sheep won't sleep: Counting by 2s, 5s, and 10s*. New York, NY: Holiday House.

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., et al. (2007). School readiness and later achievement. *Dev. Psychol.* 43, 1428–1446. doi: 10.1037/0012-1649.43.6.1428

Eaton, M., and Symone, A. (2021). *Money plan*. Solon, OH: Findaway Voices.

Enders, C. K. (2010). *Applied missing data analysis*. New York, NY: Guilford Press.

Enders, C. K., and Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: a new look at an old issue. *Psychol. Methods* 12, 121–138. doi: 10.1037/1082-989x.12.2.121

Flatt, L., and Barron, A. (2014). *Shaping up summer*. Montreal, Quebec, Canada: Owlkids.

Gardner, M. (2015). *The four money bears*. Tampa, FL: Motivate Your Money.

Giganti, P., and Crews, D. (1988). *How many snails?: A counting book*. New York, NY: Greenwillow Books.

Ginsberg, H. P., and Baroody, A. J. (2003). *Test of early mathematics ability (3rd Ed)*. Austin, TX: Pro-ed.

Giroir, S., Grimaldo, L. R., Vaughn, S., and Roberts, G. (2015). Interactive read-alouds for English learners in the elementary grades. *Read. Teach.* 68, 639–648. doi: 10.1002/trtr.1354

Green, K. B., Gallagher, P. A., and Hart, L. (2018). Integrating mathematics and children's literature for young children with disabilities. *J. Early Interv.* 40, 3–19. doi: 10.1177/1053815117737339

Hassinger-Das, B., Jordan, N. C., and Dyson, N. (2015). Reading stories to learn math: mathematics vocabulary instruction for children with early numeracy difficulties. *Elem. Sch. J.* 116, 242–264. doi: 10.1086/683986

Heisey, N., and Kucan, L. (2010). Introducing science concepts to primary students through read-alouds: interactions and multiple texts make the difference. *Read. Teach.* 63, 666–676. doi: 10.1598/RT.63.8.5

Hendrix, N. M., Hojnoski, R. L., and Missall, K. N. (2019). Shared book reading to promote math talk in parent-child dyads in low-income families. *Top. Early Child. Spec. Educ.* 39, 45–55. doi: 10.1177/0271121419831762

Hoffman, L. (2015). *Longitudinal analysis: modeling within-person fluctuation and change*. London, England: Routledge.

Hojnoski, R. L., Columba, H. L., and Polignano, J. (2014). Embedding mathematical dialogue in parent-child shared book reading: a preliminary investigation. *Early Educ. Dev.* 25, 469–492. doi: 10.1080/10409289.2013.810481

Hornburg, C. B., Schmitt, S. A., and Purpura, D. J. (2018). Relations between preschoolers' mathematical language understanding and specific numeracy skills. *J. Exp. Child Psychol.* 176, 84–100. doi: 10.1016/j.jecp.2018.07.005

Johnston, V. (2016). Successful read-alouds in today's classroom. *Kappa Delta Pi Rec.* 52, 39–42. doi: 10.1080/00228958.2016.1123051

Kennedy, M. M. (2016). How does professional development improve teaching? *Rev. Educ. Res.* 86, 945–980. doi: 10.3102/0034654315626800

King, Y. A., and Purpura, D. J. (2021). Direct numeracy activities and early math skills: math language as a mediator. *Early Child. Res. Q.* 54, 252–259. doi: 10.1016/j.ecresq.2020.09.012

Krajewski, K., and Schneider, W. (2009). Exploring the impact of phonological awareness, visual-spatial working memory, and preschool quantity-number competencies on mathematics achievement in elementary school: findings from a longitudinal study. *J. Exp. Child Psychol.* 103, 516–531. doi: 10.1016/j.jecp.2009.03.009

Lee, M. (2013). *20 big trucks in the middle of the street*. Somerville, MA: Candlewick.

Leedy, L. (2006). *The great graph contest*. New York, NY: Holiday House.

Levine, S. (2020). *The animals would not sleep!*. Watertown, MA: Charlesbridge.

Levine, S. C., Suriyakham, L. W., Rowe, M. L., Huttenlocher, J., and Gunderson, E. A. (2010). What counts in the development of young children's number knowledge? *Dev. Psychol.* 46, 1309–1319. doi: 10.1037/a0019671

Long, L. (1996). *Domino addition*. Watertown, MA: Charlesbridge.

Massey, S. L. (2013). From the reading rug to the play center: enhancing vocabulary and comprehensive language skills by connecting storybook reading and guided play. *Early Childhood Educ. J.* 41, 125–131. doi: 10.1007/s10643-012-0524-y

McGuire, H. B., Clayton, G., Yoo, M., and Logue, M. E. (2021). Booked on math: developing math concepts in pre-k classrooms using interactive read-alouds. *Early Childhood Educ. J.* 49, 313–323. doi: 10.1007/s10643-020-01073-1

McNeish, D., Stapleton, L. M., and Silverman, R. D. (2017). On the unnecessary ubiquity of hierarchical linear modeling. *Psychol. Methods* 22, 114–140. doi: 10.1037/met0000078

Mol, S. E., Bus, A. G., and de Jong, M. T. (2009). Interactive book reading in early education: a tool to stimulate print knowledge as well as oral language. *Rev. Educ. Res.* 79, 979–1007. doi: 10.3102/0034654309332561

Murphy, S. (1998). *The penny pot*. New York, NY: HarperCollins.

Murphy, S. J., and Gorton, J. (1999). *Super sand castle Saturday*. New York, NY: HarperCollins.

Murphy, S. J., and Karas, G. B. (1997). *Elevator magic*. New York, NY: HarperCollins.

Murphy, S. J., and Simard, R. (2001). *Captain invincible and the space shapes*. New York, NY: HarperCollins.

Muthén, L., and Muthén, M. (2023). *Mplus*. Available at: https://www.statmodel.com/index.shtml.

Narita, K., and Kaufman, S. (2018). *100 bugs!: a counting book*. New York, NY: Farrar Straus Giroux.

Nelson, P. M., Van Norman, E. R., Parker, D. C., and Cormier, D. C. (2019). An examination of interventionist implementation fidelity and content knowledge as predictors of math intervention effectiveness. *J. Appl. Sch. Psychol.* 35, 234–256. doi: 10.1080/15377903.2019.1568334

O'Donnell, C. L. (2008). Defining, conceptualizing, and measuring fidelity of implementation and its relationship to outcomes in K–12 curriculum intervention research. *Rev. Educ. Res.* 78, 33–84. doi: 10.3102/0034654307313793

Pagani, L. S., Fitzpatrick, C., Archambault, I., and Janosz, M. (2010). School readiness and later achievement: a French-Canadian replication and extension. *Dev. Psychol.* 46, 984–994. doi: 10.1037/a0018881

Peltier, C., Lingo, M. E., Autry, F., Deardorff, M. E., and Palacios, M. (2021). Schema-based instruction implemented under routine conditions. *J. Appl. Sch. Psychol.* 37, 246–267. doi: 10.1080/15377903.2020.1821273

Powell, S. R., and Moore, C. (2021). *Kindergarten Texas early assessment of mathematics*. Austin, TX: S. R. Powell, 1912 Speedway, 78712.

Powell, S. R., Moore, C. E., Forsyth, S. R., Fall, A.-M., and Roberts, G. (under review). The impact of mathematics-focused read-alouds on mathematics outcomes for preschool children.

Purpura, D. J., and Logan, J. A. R. (2015). The nonlinear relations of the approximate number system and mathematical language to early mathematics development. *Dev. Psychol.* 51, 1717–1724. doi: 10.1037/dev0000055

Purpura, D. J., Logan, J. A. R., Hassinger-Das, B., and Napoli, A. R. (2017a). Why do early mathematics skills predict later reading? The role of mathematical language. *Dev. Psychol.* 53, 1633–1642. doi: 10.1037/dev0000375

Purpura, D. J., Napoli, A. R., Wehrspann, E. A., and Gold, Z. S. (2017b). Causal connections between mathematical language and mathematical knowledge: a dialogic reading intervention. *J. Res. Educ. Effect.* 10, 116–137. doi: 10.1080/19345747.2016.1204639

Purpura, D. J., Schmitt, S. A., Napoli, A. R., Dobbs-Oates, J., King, Y. A., Hornburg, C. B., et al. (2021). Engaging caregivers and children in picture books: a family-implemented mathematical language intervention. *J. Educ. Psychol.* 113, 1338–1353. doi: 10.1037/edu0000662

Rudd, L. C., Lambert, M. C., Satterwhite, M., and Smith, C. H. (2009). Professional development + coaching = enhanced teaching: increasing usage of math mediated language in preschool classrooms. *Early Childhood Educ. J.* 37, 63–69. doi: 10.1007/s10643-009-0320-5

Santoro, L. E., Chard, D. J., Howard, L., and Baker, S. K. (2008). Making the "very" most of classroom read-alouds to promote comprehension and vocabulary. *Read. Teach.* 61, 396–408. doi: 10.1598/RT.61.5.4

Schiro, M., Casey, B., and Anderson, K. (2002). *Tan and the shape changer*. St. Petersburg, FL: The Wright.

Schleppegrell, M. J. (2012). Academic language in teaching and learning: introduction to the special issue. *Elem. Sch. J.* 112, 409–418. doi: 10.1086/663297

Singer, J. D., and Willett, J. B. (2003). *Applied longitudinal data analysis: modeling change and event occurrence*. Oxford, England: Oxford University Press.

Stephens, A. M., and Liu, J. (2019). *Arithmechicks add up: a math story*. Honesdale, PA: Boyds Mills Press.

Stephens, A. M., and Liu, J. (2020). *Arithmechicks take away*. Honesdale, PA: Boyds Mills Press.

Stevens, R. J., Van Meter, P., and Warcholak, N. D. (2010). The effects of explicitly teaching story structure to primary grade children. *J. Lit. Res.* 42, 159–198. doi: 10.1080/10862961003796173

Strachan, S. L. (2015). Kindergarten students' social studies and literacy learning from interactive read-alouds. *J. Soc. Stud. Res.* 39, 207–223. doi: 10.1016/j.jssr.2015.08.003

Susperreguy, M. I., and Davis-Kean, P. E. (2016). Maternal math talk in the home and math skills in preschool children. *Early Educ. Dev.* 27, 841–857. doi: 10.1080/10409289.2016.1148480

The Meadows Center for Preventing Educational Risk (2014). *Read-aloud routine for building vocabulary and comprehension skills in prekindergarten*: Author.

Thong, R., and Parra, J. (2013). *Round is a tortilla: A book of shapes*: Chronicle Books.

Thoonen, E. E. J., Sleegers, P. J. C., Oort, F. J., Peetsma, T. T. D., and Geijsel, F. P. (2011). How to improve teaching practices: the role of teacher motivation, organizational factors, and leadership practices. *Educ. Adm. Q.* 47, 496–536. doi: 10.1177/0013161X11400185

Throndsen, J. E., Shumway, J. F., and Moyer-Packenham, P. S. (2020). The relationship between mathematical literacy at kindergarten entry and public preschool attendance, type, and quality. *Early Childhood Educ. J.* 48, 473–483. doi: 10.1007/s10643-019-01014-7

Toll, S. W. M., and Van Luit, J. E. H. (2014). Explaining numeracy development in weak performing kindergartners. *J. Exp. Child Psychol.* 124, 97–111. doi: 10.1016/j.jecp.2014.02.001

Turan, E., and De Smedt, B. (2022). Mathematical language and mathematical abilities in preschool: a systematic literature review. *Educ. Res. Rev.* 36:100457. doi: 10.1016/j.edurev.2022.100457

Uscianowski, C., Almeda, M. V., and Ginsberg, H. P. (2020). Differences in the complexity of math and literacy questions parents pose during storybook reading. *Early Child. Res. Q.* 50, 40–50. doi: 10.1016/j.ecresq.2018.07.003

Van den Heuvel-Panhuizen, M., Elia, I., and Robitzsch, A. (2016). Effects of reading picture books on kindergartners' mathematics performance. *Educ. Psychol.* 36, 323–346. doi: 10.1080/01443410.2014.963029

von Hippel, P. T., Workman, J., and Downey, D. B. (2018). Inequality in reading and math skills forms mainly before kindergarten: a replication, and partial correct, or "are schools to great equalizer?". *Sociol. Educ.* 91, 323–357. doi: 10.1177/0038040718801760

Weakland, M., and Sinkovec, I. (2014). *How tall?: wacky ways to compare height*. Bloomington, MN: Picture Window Books.

What Works Clearinghouse (2020). *What works clearinghouse procedures handbook, version 4.1*: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. Available at: https://ies.ed.gov/ncee/wwc/handbooks.

Wood, C. L., Goodnight, C. I., Bethune, K. S., Preston, A. I., and Cleaver, S. I. (2016). Role of professional development and multi-level coaching in promoting evidence-based practice in education. *Learn. Disabil. Contemp. J.* 14, 159–170.

frontiers | Frontiers in Education

TYPE Original Research
PUBLISHED 24 July 2024
DOI 10.3389/feduc.2024.1376805

OPEN ACCESS

EDITED BY
Rüdiger Christoph Pryss,
Julius Maximilian University of Würzburg, Germany

REVIEWED BY
Young S. Seo,
University at Buffalo, United States
Michael Winter,
Julius Maximilian University of Würzburg, Germany

*CORRESPONDENCE
Susanne Seifert
✉ susanne.seifert@uni-graz.at

RECEIVED 26 January 2024
ACCEPTED 10 July 2024
PUBLISHED 24 July 2024

CITATION
Seifert S, Paleczek L, Schöfl M and Weber C (2024) Unveiling mode effects in grade 1 vocabulary assessment: the intriguing influence of test mode.
Front. Educ. 9:1376805.
doi: 10.3389/feduc.2024.1376805

Unveiling mode effects in grade 1 vocabulary assessment: the intriguing influence of test mode

Susanne Seifert[1,2]*, Lisa Paleczek[1,2], Martin Schöfl[3,4] and Christoph Weber[3,4]

[1]Department of Education Research and Teacher Education, University of Graz, Graz, Austria, [2]Research Center for Inclusive Education, Graz, Austria, [3]Department of Educational Sciences, University of Education Upper Austria, Linz, Upper Austria, Austria, [4]Research Institute for Developmental Medicine, Johannes Kepler University Linz, Linz, Upper Austria, Austria

Background: Vocabulary knowledge plays a pivotal role in academic development, particularly among Grade 1 students. To support students in their academic development, effective assessment instruments in educational settings are crucial. The GraWo (Graz Vocabulary Test) is introduced as a tool designed to evaluate receptive vocabulary in German-speaking countries in print and in digital mode.

Objectives: This study aims to investigate mode effects in the GraWo among Grade 1 students, comparing vocabulary gains in digital and print versions. Additionally, it explores the influence of student characteristics, such as gender and language status, and examines item-level differences between the two modes in order to gain a more comprehensive understanding of test performance.

Design: The research design entails a longitudinal approach, following children ($n = 421$) from the beginning to the end of Grade 1, varying the test modes (digital or print) only at second measurement (40% receiving the print version), while at first measurement all children worked with the digital version.

Results: Baseline comparisons of test mode groups indicated almost no significant differences. In terms of growth in vocabulary during Grade 1, an ANOVA with repeated measures revealed a main effect for time, indicating increased performance in both groups at second measurement. Moreover, an interaction effect between time and test mode group showed that the print group exhibited higher gains in the vocabulary test compared to the digital group. Further analysis using MNLFA confirmed that the print mode group outperformed the digital group overall and that four items were also individually affected by differences between the digital and print versions.

Conclusion: The study emphasizes the need for nuanced investigations into the impact of test mode on student performance and suggests incorporating observational methods to comprehensively understand student interactions with digital and print modes. In acknowledging potential variations in performance, educators and policymakers need to tailor practices to accommodate the demands of hybrid test procedures and to consider the role of digital competence in shaping testing experiences.

KEYWORDS

vocabulary, German, assessment, mode effect, test mode, grade 1

1 Introduction

Vocabulary plays a pivotal role in nurturing various linguistic abilities and is a crucial requirement for many skills essential in the academic context. Specifically, a child's proficiency in reading, including word recognition and comprehension, can be predicted by their vocabulary knowledge during their early years (Muter et al., 2004; Ennemoser et al., 2012; Juska-Bacher et al., 2021). The association between vocabulary and reading skills becomes even more pronounced among second language (L2) learners. Research consistently demonstrates that L2 learners tend to possess a lesser command of vocabulary in the language of instruction when compared to their first language (L1) counterparts (for example, Cremer and Schoonen, 2013; research specifically focusing on the German language: Klassert, 2011; Seifert et al., 2019). Consequently, L2 learners often exhibit poorer performance on reading assessments conducted in the language of instruction (Melby-Lervåg and Lervåg, 2014; Wendt and Schwippert, 2017). Hence, vocabulary knowledge is a relevant precursor for reading development, encompassing both receptive and productive vocabulary. Receptive vocabulary refers to the words that students can recognize and understand when they encounter them in listening or reading, while productive vocabulary includes the words they can use correctly in speaking and writing. Developing both types of vocabulary is crucial for Grade 1 students.

1.1 Vocabulary assessment

To foster vocabulary skills, reliable und valid assessment instruments for use in the school context are needed, as is also true regarding other precursor skills for written language (Ennemoser et al., 2012). Productive vocabulary, indicating a child's ability to express words, is usually assessed individually, especially when concerns about potential deficits arise. Conversely, receptive vocabulary, revealing a child's understanding of words, is intricately linked to reading proficiency since reading hinges on the comprehension of written words. Although it is possible to assess receptive vocabulary in a group setting with an entire class, there are only relatively few assessment tools available in German-speaking countries that support this setting.

The majority of German-language group assessments for receptive vocabulary skills primarily involve written language, such as requiring students to match one of four written words to a corresponding picture (e.g., word comprehension test as a subtest of the ELFE II reading comprehension test: Lenhard et al., 2020) or to decide whether a written word is a word of the German language or not (e.g., WOR-TE vocabulary test: Trautwein and Schroeder, 2019). These assessments tend to emphasize the orthographic component of vocabulary. However, according to Perfetti and Hart (2002), vocabulary encompasses two additional facets: the phonological component, encompassing knowledge about word pronunciation, and the semantic component, related to understanding word meanings. At the beginning of schooling, when reading skills are about to be acquired, methods that primarily focus on the orthographic component of vocabulary are not a suitable measurement technique. In contrast, the Graz Vocabulary Test (GraWo, Seifert et al., 2017) places particular emphasis on the semantic component of vocabulary by avoiding written language and therefore not relying on children's reading skills. In this receptive vocabulary assessment, designed for screening purposes and comprising only 30 items, children are tasked with matching a word presented orally to one of four pictures (see Figure 1). For the present study, we focus on Grade 1 students and assess receptive vocabulary knowledge.

1.2 Digital and print assessment

When using assessments in the school context, efficiency and user-friendliness are crucial. Group screenings like GraWo (Seifert et al., 2017 are preferred over complex individual tests. Additionally, digital assessment tools provide time-saving advantages. This increases their relevance and also furthers their acceptance (e.g., Neumann et al., 2019). Digital tools also enhance standardization of test instructions, item presentation, and response encoding, leading to more accurate and fairer assessments (Wang et al., 2021).

While the advantages of digital assessments are numerous, and their acceptance is widespread (e.g., Dawidowsky et al., 2021), the transition to digital assessments can only succeed if all schools are adequately equipped with digital devices, and if teachers as well as children possess the necessary expertise. Therefore, at least for a transitional period, both digital and traditional paper-based (print) assessments will coexist and remain relevant (Paleczek et al., 2021). Hybrid instruments that can be applied in both digital and print versions will continue to serve a purpose, allowing test users to flexibly utilize an instrument according to their individual needs. However, it is essential that the results from both versions are comparable since the goal is to assess the same skill with the instrument, and not to assess additional competencies such as digital skills (Puhan et al., 2007; Paleczek et al., 2021; Seifert and Paleczek, 2022).

1.3 Test mode effects

Due to the fundamental differences in the answering processes required for digital and print assessments, the research on mode effects continues to be a subject of interest among researchers (e.g., Wang et al., 2021). Test performance may also be influenced by the individual digital experience of examinees with the specific digital device that is used, which in turn raises concerns about potential variations in validity between digital and paper-based assessments. However, there is a lack of consistent research evidence regarding the equivalence and interchangeability of scores with respect to digital and print assessments. Across various instructional domains (reading, math, sciences), some studies (Johnson and Green, 2006; Puhan et al., 2007; Hamhuis et al., 2020) and meta-analyses (Wang et al., 2007, 2008) indicate comparability, while others suggest the opposite. In case of non-comparability, print assessments sometimes yield higher results compared to digital ones (Taherbhai et al., 2012; Lenhard et al., 2017; Backes and Cowan, 2019; Seifert and Paleczek, 2022; Wagner et al., 2022), and in other cases, digital assessments lead to higher scores (Lee et al., 2010; Wang et al., 2021).

The reasons for mode effects are still being debated. Some argue that mode effects are item-specific, tied to factors like response formats or item order (Buerger et al., 2019). Others suggest that differences in test-taking behavior and answering strategies explain such effects. For instance, computer-based test-takers tend to complete tests more quickly,

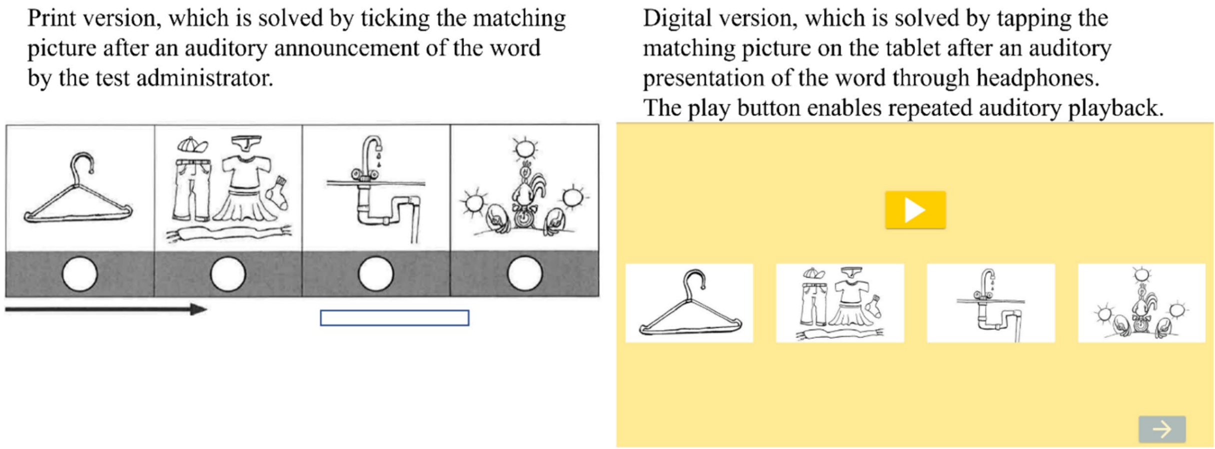

FIGURE 1
Print and digital versions of item "Kleidung" (clothes) in test GraWo.

guessing more often or responding more shallowly (Bodmann and Robinson, 2004; Karay et al., 2015; Lenhard et al., 2017; Singer-Trakham et al., 2019; Støle et al., 2020), although, here too, contradictory findings also exist (Steedle et al., 2022). Lenhard et al. (2017) noted that such behaviour is more prevalent in younger children (e.g., Grade 1 students) and when dealing with simpler tasks. Moreover, it is also prominent when a single test item is displayed on the screen (instead of more items or representation formats where scrolling is needed), as seen by Leeson (2006) as well as Bodmann and Robinson (2004). In contrast, paper-based assessments lead to more frequent answer reviews and amendments (Johnson and Green, 2006; Wang et al., 2021).

Mode effects have also been examined concerning specific student characteristics, with gender being a frequently explored variable. For instance, Jeong (2012) observed a difference between girls and boys, with girls performing significantly worse in digital assessments compared to paper-based ones in three out of four academic domains, while for boys this difference was only found in one domain. This was attributed to a gender gap in computer usage. However, most studies found no evidence of gender influencing mode effects (Clariana and Wallace, 2002; Poggio et al., 2005; Lee et al., 2010). Another student variable frequently discussed in relation to the mode effect is language status (L1 or L2). While some studies focused on English language learners (who can be called L2 learners), revealed mode effects in this specific sample (Hosseini et al., 2014), few have examined differences between L1 and L2 learners. Existing studies present a mixed picture. For example, Backes and Cowan (2019) identified stronger mode effects in language assessments for L2 learners compared to their L1 peers. In contrast, our own previous study (Seifert and Paleczek, 2022) found no moderating effect of language status on the mode effect found with a reading assessment tool. Apart from gender and language status, digital competence or familiarity with the corresponding digital devices seem to play a more significant role in moderating mode effects (Clariana and Wallace, 2002).

1.4 Scientific aim

The present study aims to contribute to a better understanding of mode effects, focusing on a vocabulary test previously employed to examine mode effects in the last year of kindergarten (Paleczek et al., 2021). While no significant differences in mean scores between the print and digital versions of this test were observed in the last year of kindergarten, a preference for the digital format was identified (as also revealed, e.g., by Gnambs and Lenhard, 2023). Furthermore, examiners were able to identify a distinct form of working behavior, indicating that children in the digital version worked more rapidly and with less self-monitoring.

The present study seeks to extend the investigation of mode effects with regard to the same instrument to the early stages of primary school, specifically, to explore whether the digital version truly aligns with the print version. Additionally, the study aims to scrutinize the moderating influence of the variables "language status" and "gender" concerning mode effects.

In detail, the present study examines mode effects in the use of the vocabulary test GraWo in Grade 1. Experimental design ensured that familiarity effects with the instrument and the vocabulary task itself were avoided. All children were given the digital version at the first measurement at the beginning of Grade 1 and the mode was only varied at the second measurement at the end of the school year.

The digital version was provided via tablets. We chose tablets over computers due to several factors. Tablets offer greater portability and ease of use, which are particularly important for young learners (Merchant, 2015). The touch-based interaction of tablets can be more intuitive and engaging for Grade 1 students, enhancing their learning experience (Ricoy and Sánchez-Martínez, 2020). Additionally, tablets are more commonly used in modern classrooms (Ricoy and Sánchez-Martínez, 2020) and households (Chaudron et al., 2017), making the study conditions more ecologically valid. Had we used computers, we might have encountered different challenges, such as the need for more advanced motor skills for mouse and keyboard use, potentially impacting the students' performance and the study's outcomes.

Based on our previous results with the same instrument in the last kindergarten year (Paleczek et al., 2021), we assume that there will be no significant differences between the two groups using different test versions at second measurement (digital vs. print). However, if there is a difference between the gains of the two groups, and both groups are comparable in terms of key baseline characteristics and composition - then it may be assumed that the test version used at the second measurement does indeed have an impact on the results. We would then hypothesize that the group assessed with the paper version at second measurement would show higher gains (Taherbhai et al., 2012; Lenhard et al., 2017; Backes and Cowan, 2019; Seifert and Paleczek, 2022; Wagner et al., 2022). We also look at the influence of student gender and language status on mode effect. Based on our earlier work on mode effects in reading comprehension (Seifert and Paleczek, 2022), we suspect these two student characteristics will have no influence. However, given the mixed results in existing literature, it is crucial to empirically investigate these characteristics to ensure a comprehensive understanding of their potential impact in our specific context.

Furthermore, in addition to examining mode effects with respect to overall test scores, we also investigate mode differences in single items. In detail, we examine whether the probability of correctly solving an item differs between the two modes beyond what would be expected from mode differences on the overall test (i.e., differential item functioning; e.g., Bauer, 2023).

2 Materials and methods

This study is part of a longitudinal study concerning assessment of, and intervention in, reading and writing from the very beginning of schooling (see Schöfl et al., 2022).

In the present study, we followed children from the beginning (autumn 2021, Measurement 1) to the end of Grade 1 (summer 2022, Measurement 2). Vocabulary was assessed as part of a broader screening procedure which is described in Schöfl et al. (2022). At the beginning of Grade 1, vocabulary skills were assessed using the digital version of a vocabulary test provided on tablets. At the end of Grade 1, vocabulary was assessed using either the digital or the print version of the test (groups: digital vs. print). Only those students that performed the vocabulary test at both times were included in the present study.

The different measures used at first measurement in the screening procedure (assessing the domains sentence repetition and phonological information processing) are described below in section 2.2 (Instruments).

2.1 Participant recruitment and characteristics

The majority of the participants enrolled in this study were from a district in Upper Austria, encompassing four prominent community-based schools. Initial contact and invitations for study participation were extended to eligible schools through telephone communication, followed by subsequent in-person visits. All school headmasters provided their consent for participation, and an additional four schools expressed interest and subsequently joined the project. Ultimately, parents of 459 students agreed (providing written permission) at the beginning of Grade 1 that their child(ren) could participate in the study. Data on the GraWo vocabulary test is available solely at Measurement 1 for 38 students [8.28% of the consented sample; 20 of them male, 30 monolinguals in German (language status L1)]. For the remaining 421 children (91.72% of the consented sample), data existed for both measurements (beginning and end of Grade 1). Consequently, these students were included in the analyses for this study.

The resulting study sample exhibited a diverse mix of children, and mirrored the characteristics of the Austrian primary school population in terms of gender, the proportion of L2 learners, and parental educational levels.

The characteristics of participants are representative for Austrian elementary school children. Of the 421 students, 48.9% ($n = 206$) are female, 73.6% ($n = 310$) are monolinguals in German (language status L1). These proportions correspond to those found in Austrian school statistics (Statistik Austria, 2022). Parents' highest educational attainment was used as a proxy for the children's socioeconomic background. The sample consisted of students with parents from all educational backgrounds: among the mothers, 7.6% had a maximum educational attainment of a secondary school diploma, 28.3% completed an apprenticeship or a VET school, 19.2% had a high school diploma, and 38.2% had a university diploma. The fathers' educational levels were comparable (7.8% secondary school diploma, 34% apprenticeship or VET school,14.3% high school diploma, and 30.9% with a university degree).

The individualized screening process commenced in the autumn of 2020, and was initiated within 2 weeks of school onset in 27 classes. Within 3 weeks, assessments had been completed for 85% of the sample. Subsequently, over the following 2 weeks, children who were either unwell or unable to attend during the initial survey period were also assessed.

2.2 Instruments

The Graz Vocabulary Test (GraWo; Seifert et al., 2017) was used to assess vocabulary skills. The GraWo is a standardized screening instrument that assesses receptive vocabulary in first to third graders. It consists of two sample items and 30 test items (5 verbs, 5 adjectives, 5 prepositions, and 15 nouns, with the latter being split into 5 mono-morphematic nouns, 5 composite nouns and 5 nouns referring to categories; for a list of all items, see Supplementary Table S1). The children are required to select one out of four pictures matching a word that is presented audibly. This test can be administered in various formats and settings, broadly classified as group or single settings, and either as a print version or a digital version (conducted on tablets). For the scope of our study, we provided the print version in a single setting, where the word was pronounced out loud by the instructor. The digital version was delivered in a single setting as well, in which children used headphones to listen to prerecorded pronunciations by professional speakers (as illustrated in Figure 1). This approach was chosen to ensure uniformity in the conditions of test administration across both modes, allowing for a fair comparison of performance outcomes.

This vocabulary test was either provided in the digital version [as a part of a wider screening procedure, for all of the students at the beginning of Grade 1 ($N = 421$) and for nearly 60% of the sample at the end of Grade 1 ($n = 249$)], or as a print version [for nearly 40% of the sample at the end of the school year ($n = 172$)]. Reliability data are given for the print version of the GraWo (Seifert et al., 2017): Cronbach's Alpha ranged from 0.89 (end of Grade 1) to 0.82 (end of Grade 2). Retest reliability was $r_{tt} = 0.93$ (Grade 1).

A parent questionnaire was used to obtain information about children's L1 and socioeconomic background. When this indicated that the L1 was only German, or that contact with German occurred from birth up to and including the age of 2, then children were classified as having language status L1. Children whose contact with German occurred only after the age of 2 were classified as having language status L2.

Additionally, a screening of potential precursors of reading skills (covering sentence repetition: adapted from Ibrahim et al., 2018; phonological awareness (PA): Schöfl et al., 2022; rapid automatized naming (RAN) objects: Schöfl et al., 2022; RAN digits: Denckla and Rudel, 1974; letter knowledge: Schöfl et al., 2022; phonological working memory: Grob et al., 2009) was used at the beginning of Grade 1 (for more information, see also Schöfl et al., 2022). The results of these tests were used to analyze whether the two test mode groups differed in terms of baseline results.

2.3 Procedure

Before assessment, school principals received information and a letter for parents explaining the testing procedure, along with consent forms, a data protection declaration, and questions about children's language and parents' educational background. Teachers entered student names into an online database, converting them to IDs for tablet use. Pre-service teachers enrolled in a university seminar on testing procedures carried out the testing and received course credit for it (amounting to 4 h). The materials for testing were brought to the schools by a research coordinator. On test mornings, the test team (pre-service teachers and project staff) selected students alphabetically, and assessed them individually. Instructor and child were seated at a table across from each other. Each child was given a tablet and guided through the screening process, starting with an introduction to the friendly dragon SCHWUPP. App navigation allowed independent use, with instructors intervening if needed. Instructions were recorded as audio files, opened automatically, and could be listened to repeatedly if necessary. The assessment, including all subtests, averaged 38.4 min per child (SD = 9.3). Vocabulary assessment with the GraWo was just one part of the screening.

At the end of the school year, the procedure remained similar, except for one key modification: to address the present research questions, classes were randomly assigned to either receive the digital mode again or switch to the print mode of the vocabulary test. Half of the classes, comprising 40% of the students, received the print version of the vocabulary test instead of the digital one. Due to logistical practicalities, such as the allocation of tablets and paper materials, once a class was randomly selected for a test mode, all students within that class were tested in the same mode to maintain consistency and manage resources efficiently.

The total number of items solved was either recorded directly by the app in the case of digital testing, or was counted manually by the research team. Data on individual items are available for all children in the digital version for both measurements (measurement 1: $n = 421$ of 421 students, measurement 2: $n = 249$ of 249 students, 100%), but only for some of the children in the print version ($n = 69$ of 197 students, 35%). Our analysis revealed that the two print version groups, one with available data on individual items ($n = 69$) and one without ($n = 128$), did not differ significantly in their performance on the GraWo print version of the test [$t(195) = 1.37$, $p = 0.17$] and were comparable in gender distribution [$\chi^2(1) = 0.64$, $p = 0.424$, $\varphi = -0.06$] and distribution of L1 and L2 students [$\chi^2(1) = 0.98$, $p = 0.366$, $\varphi = 0.07$]. However, disparities were noted in terms of socioeconomic background: the group with available data had a lower socioeconomic status, as suggested by mothers' and fathers' highest educational attainment (mothers': $U = 2894.00$, $Z = -3.40$, $p < 0.001$; fathers': $U = 3032.50$, $Z = -2.69$, $p = 0.007$), and had fewer years of kindergarten attendance ($U = 3561.50$, $Z = -2.05$, $p = 0.040$), than the group lacking individual item data.2.5 Methods.

First, the two test mode groups (digital vs. print) were analyzed in terms of sample composition (using Chi-square tests) and baseline differences (using t-tests).

Second, to analyze for mode effects, an ANOVA with repeated measures with the inner-subject variables "GraWo scores beginning of Grade 1" and "GraWo scores end of Grade 1" and the inter-subject factor test mode group (digital at the end of Grade 1 vs. print at the end of Grade 1) was conducted to find out whether it makes a difference whether the GraWo was used digitally or in print at the end of Grade 1. A significant interaction of time and mode would indicate a mode effect. Additionally, the inter-subject factors gender (girls vs. boys) and language status (L1 vs. L2) were included to look for interaction effects with the test mode group.[1]

Third, to deepen the analyses on mode effects, we applied moderated non-linear factor analysis (MNLFA; Bauer, 2023) – a relatively new and flexible approach – to test for differential item functioning (DIF). We used the aMNLFA R-package (Gottfredson et al., 2019) together with Mplus 8 (Muthén and Muthén, 1998-2017). The DIF analysis consisted of the following steps (see Gottfredson et al., 2019, p. 68): (1) We tested whether a unidimensional model for the GraWo fit the data well, using the following guidelines for the evaluation of model fit (Schermelleh-Engel et al., 2003): $\chi^2/df \leq 2$, comparative fit index (CFI) ≥ 0.975, and root mean square error of approximation (RMSEA) ≤ 0.05 for a good fit and $\chi^2/df \leq 2$, CFI ≥ 0.975, RMSEA ≤ 0.05 for an acceptable fit. As the standardizes root mean square residual (SRMR) has been shown to over-reject models with binary indicators, we do not report this index (Yu, 2002). At this stage we used a weighted least square parameter estimation (WLSMV) with a probit link, as this estimation provides more model fit indices than maximum likelihood (ML) estimation. For all subsequent analyses, a ML estimation with Monte Carlo integration was required. (2) We assessed whether there are mode effects on the GraWo latent mean and the GraWo latent variance. At this stage, an α-level of 0.10 was used. (3) We applied an item-by-item approach to test for DIF, i.e., we tested mode effects on a single item (threshold and loading) while holding thresholds and loadings of all other items constant. At this stage, due to multiple testing issues an α-level of 0.05 was used. (4) Finally, we tested all significant mode effects of step 2 and 3 in a simultaneous model. The Benjamini-Hochberg correction was applied to correct for the effects of multiple testing. Mode effects that remained statistically significant in this final simultaneous model

1 Taking a different approach on the analysis of mode effects, we also conducted a regression analysis with GraWo at the end of Grade 1 as dependent variable and mode as independent variable. We additionally controlled for GraWo scores assessed at the beginning of Grade 1, gender and language status.

are reported below. For more details on the DIF testing procedure see Gottfredson et al. (2019).

3 Results

3.1 Baseline comparison of the test mode groups

With respect to sample composition, we investigated whether the proportion of girls and the proportion of students with language status L1 was comparable in the test mode groups. Results show neither a significant relationship between group and gender [$\chi^2(1) = 3.81$, $p = 0.051$, $\varphi = 0.09$], nor between group and language status [$\chi^2(1) = 0.36$, $p = 0.548$, $\varphi = 0.03$].

As shown by the Mann–Whitney-*U*-Test, socioeconomic background was not comparable in the test mode groups. The distributions of parents' highest educational attainment differed between both groups. There was a statistically significant difference between test mode group and mothers' highest educational attainment ($U = 17382.50$, $Z = -2.30$, $p = 0.022$) and fathers' highest educational attainment ($U = 16763.00$, $Z = -2.58$, $p = 0.010$), respectively. The digital test mode group showed higher educational attainments of both mothers and fathers than the print test mode group.

On considering the number of years of kindergarten attendance, no significant difference was found between the two groups ($U = 19920.50$, $Z = -0.56$, $p = 0.577$).

In terms of baseline differences, we first investigated the test mode groups with regard to differences in the mean values of the vocabulary test at the beginning of Grade 1. Results show no significant differences at first measurement in GraWo [$t(419) = 1.16$, $p = 0.249$; for mean values, see also Table 1]. Moreover, no other baseline assessments revealed differences between the two groups either, with one exception: the digital group achieved significantly higher sentence repetition scores (see Table 1). To facilitate a clear understanding of this particular outcome, it is pertinent to expound upon the sentence repetition task. This assessment comprised a block of 15 items representing morphosyntactic constructions with varying degrees of complexity. Each item was scored according to whether or not the sentence was reproduced correctly. Correct reproduction required that the sentence structure be mirrored precisely as presented, irrespective of any articulatory deficits that did not interfere with structural accuracy (for more information, see also Schöfl et al., 2022).

To summarize, there are hardly any differences between the two test mode groups. The existing differences (highest educational attainments of parents and sentence repetition) favor the group that was tested digitally at the second measurement compared to the group tested in print.

3.2 Comparison of vocabulary growth during the first grade across the test mode groups

Table 2 presents the means of the scores in the vocabulary test GraWo for both test mode groups.

An ANOVA with repeated measures showed a main effect for time [$F(1, 419) = 377.93$, $p < 0.001$, $\eta^2 = 0.47$], showing that both groups solved more items at the end of Grade 1 compared to the beginning of Grade 1. Additionally, an interaction effect of time and test mode group was uncovered [$F(1, 419) = 27.34$, $p < 0.001$, $\eta^2 = 0.06$]. The test mode group "print" showed higher gains in the vocabulary test (see Figure 1), suggesting a positive print mode effect[2] (Figure 2).

No main effects for gender [$F(1, 413) = 0.00$, $p = 0.996$] and language [$F(1, 413) = 0.00$, $p = 0.959$] were detected, nor were any further interaction effects with the test mode group revealed.

3.3 Comparison of test modes through DIF analyses

Before performing DIF analyses, we fitted a one-factor model to the data to evaluate whether the GraWo is unidimensional. In detail, we fit a one-factor model with loadings constrained to be equal (corresponding to a one-parameter (1pl) item response theory (IRT) model) and a one-factor model with loadings freely estimated [corresponding to a two-parameter (2pl) IRT model].

The 2pl model resulted in a good fit ($\chi^2(405) = 454.8$, $p < 0.05$; CFI = 0.980; RMSEA = 0.018, 90%-CI [0.023, 0.026]), whereas, the more restrictive 1pl model did not fit the data well ($\chi^2(434) = 802.5$, $p < 0.001$; CFI = 0.855; RMSEA = 0.047, 90%-CI [0.042, 0.052]). The 2pl model was also supported by the results of a χ^2-difference test ($\chi^2(29) = 171.0$; $p < 0.001$). Taking the 2pl as a starting point, we subsequently tested DIF using MNLFA. Although the digital mode group demonstrated more consistent performance, as evidenced by a significantly ($p < 0.001$) smaller standard deviation ($SD = 0.87$ while the SD in the print mode group was fixed at 1), it was the print mode group that, on average, outperformed the digital mode group in the vocabulary test scores (latent mean difference $d = 0.400$, $p < 0.001$). Finally, four items showed statistically significant DIF. Items 8, 15, and 21 were less often correctly solved in the digital mode group than in the print group compared to what would be expected from the overall mode effect. In contrast, the opposite was true for item 14, i.e., children were more likely to solve this item when it was offered in digital mode.

4 Discussion

4.1 Test mode effect in vocabulary assessment

The findings of this study stress the significant impact of test mode on student scores in assessments in Grade 1 students, particularly in the realm of vocabulary testing. Despite both test mode groups displaying comparable vocabulary results at the initial measurement, where both were assessed digitally, the subsequent gains exhibited noteworthy differences depending on test mode. Notably, students

2 This conclusion was also supported by the results of a supplementary regression analysis. Controlling for GraWo at the beginning of Grade 1, gender, and language status, students from the print mode group scored 1.65 points higher ($b = 1.65$, SE = 0.31, $p < 0.001$, Cohen's $d = 0.36$) on the GraWo at the end of Grade 1.

TABLE 1 Domains assessed at the beginning of the school year, information about the used assessments and results of the two test mode groups and tests of mean differences.

Domain[a]	Subtests	Number of test items	Test medium	Measured value	Results of both groups M (SD)		Differences between groups[a], $df = 419$
					Digital	Print	
Vocabulary		30	Digital	Score	21.17 (4.82)	20.58 (5.56)	$T = 1.16, p = 0.249$
Sentence repetition		15	Digital	Score	10.32 (3.77)	9.24 (4.25)	$T = 2.74, p = 0.006$
Phonological information processing	PA: Rhyme detection	10	Digital	Score	8.91 (1.25)	8.67 (1.42)	$F = 3.32, p = 0.07$
	PA: Syllable count	10	Digital	Score	8.78 (1.87)	8.71 (1.75)	$F = 0.15, p = 0.70$
	PA: Initial phoneme detection	10	Digital	Score	8.53 (1.67)	8.44 (1.97)	$F = 0.30, p = 0.58$
	RAN objects	30	Print	Time in seconds	34.03 (8.75)	34.80 (9.59)	$F = 0.72, p = 0.40$
	RAN digits	30	Print	Time in seconds	33.92 (11.97)	34.07 (16.53)	$F = 0.01, p = 0.91$
	Letter knowledge	26	Print	Score	14.76 (8.16)	15.31 (8.26)	$F = 0.46, p = 0.50$
	Phonological working memory – letter number span forward	adaptive	Verbal (instructor)	Score	7.32 (1.84)	7.32 (2.44)	$F = 0.00, p = 0.97$
	Phonological working memory – letter number span backward	adaptive	Verbal (instructor)	Score	3.90 (1.96)	4.10 (2.18)	$F = 0.96, p = 0.33$

[a]Domains without subtests: *t*-tests were conducted, Domains with at least two subtests: multivariate variance analyses were conducted.

TABLE 2 GraWo scores for both test mode groups.

Test mode group	Beginning of Grade 1 (measurement 1) – all students digital		End of Grade 1 (measurement 2)	
	M	*SD*	*M*	*SD*
Digital	21.17	4.82	23.71	4.62
Print	20.58	5.56	24.99	4.40

assessed with the print version at the second measurement demonstrated significantly higher gains compared to their counterparts who were tested using the digital version. Assessment of DIF showed that the mode effect occurs overall, but that four items are also individually affected by these differences between the digital and print versions beyond what we would expect from the overall mode effect alone. For item 8, one possible explanation for the advantage of the print version could be that it is easier to differentiate from an auditory distractor (target item “hell” [bright], distractor “Fell” [fur], see also Supplementary Table S1) due to the mouth of the instructor being visible in the print version. Such cues are absent in the digital version. For the other three items, we cannot explain the differences between the print and digital modes. Future studies are needed to address this topic more systematically.

The revealed test mode differences were contrary to our original assumptions (i.e., based on our previous results with the same instrument in the last kindergarten year no differences were found) (Paleczek et al., 2021). However, the results are in line with the existing literature, where higher student scores are found for print version than for the digital version assessment (e.g., Taherbhai et al., 2012; Lenhard et al., 2017; Backes and Cowan, 2019; Seifert and Paleczek, 2022; Wagner et al., 2022). The present study even adds further weight to the argument, since at second measurement, all students were familiar not only with the assessment instrument, but also with the digital test mode as all of them had been assessed with the digital version at the first measurement (at the beginning of Grade 1). They knew what to do in the digital mode. Hence, inexperience with the digital instrument cannot be used as the cause of the mode effect. Moreover, the rare baseline differences we detected between the two groups actually favored the digital group (highest parental educational attainment, sentence repetition). Yet, apparently, these were not sufficiently strong to counteract the influence of test mode.

As evidenced by prior studies (e.g., Lenhard et al., 2017), it is likely that the unique demands present in digital assessment instruments contribute to the observed differences. In the GraWo, both the setting and the response format differ for the digital and print versions. The digital setting, with instructions received via headphones, is rather individualized and self-directed, allowing students to decide for themselves when to proceed. In contrast, the print setting is a guided process, where students only proceed when the instructor decides to do so. The specific response format also differs between the digital and the print mode. In the digital version, students tap the correct image and the “proceed” button on a tablet, which makes items visible individually, one after the other. Although listening to the audio

10.3389/feduc.2024.1376805

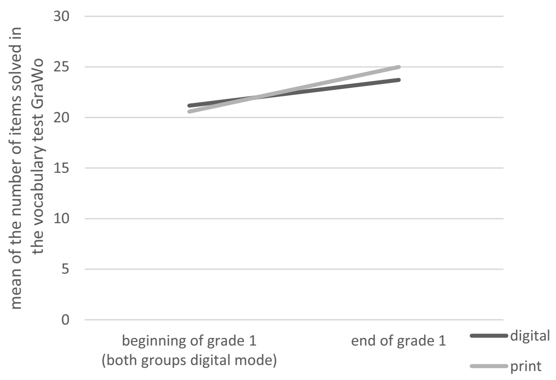

FIGURE 2
GraWo scores at the beginning and end of Grade 1, using either the digital or print version of the GraWo at the end of Grade 1.

announcement again and returning to a previous item is possible, students may not make sufficient use of these functions. In the print mode, students make a cross in the test booklet following the instructor's command. The instructor can easily be asked to repeat the announcement. Additionally, each page in the print version presents five items, which leaves previous items visible and might facilitate access to rethinking and, if necessary, correction of earlier answers. Observations on students' test behaviors with the kindergarten version of this test (Paleczek et al., 2021) suggest that students using the digital version in the present study provided faster and potentially less-considered responses, as has been found in previous studies (e.g., Bodmann and Robinson, 2004; Lenhard et al., 2017; Singer-Trakham et al., 2019). Previous studies have also shown this effect to be especially prominent when using digital assessments with single items (and not multiple ones) displayed on the screen (Bodmann and Robinson, 2004; Leeson, 2006), as was the case in the present study. This highlights the potential impact of the self-directed nature of the digital mode of the GraWo, where students may aim to complete tasks more swiftly and perhaps be less accurate in their responses.

Despite these insights, there remains a dearth of observational studies comparing the actual behaviors of students in digital and print modes, especially in the investigation of mode effects. Future research should prioritize such observations in order to gain a more nuanced understanding of how students interact with different test modes and also to shed more light on the reasons for different student interactions.

In the current study, we also delved into student characteristics concerning the mode effect. The findings revealed that neither gender nor language status exerted an influence on the mode effect, which is also in line with previous research (for gender: Clariana and Wallace, 2002; Poggio et al., 2005; Lee et al., 2010; for language status: Seifert and Paleczek, 2022). However, a significant factor remained largely unexplored: digital competence. This omission is particularly noteworthy given the evolving landscape of technology in education. Existing research (e.g., Clariana and Wallace, 2002) emphasizes the importance of considering digital competence in understanding students' interactions with digital and print modes. However, while frequency of device use may not be the primary determinant, the actual application skills, such as tablet orientation and proficiency in app-specific functions (e.g., tapping), may moderate effects. Future studies investigating mode effects need to scrutinize digital competence and associated application components in order to further elucidate the nuanced dynamics at play.

In conclusion, this study contributes valuable insights into the multifaceted impact of test mode on student scores, emphasizing the need for a comprehensive understanding of setting, task requirements, and student digital competence. The findings prompt a re-evaluation of assumptions about digital testing behaviors and underscore the imperative for more extensive observational research in this domain.

4.2 Limitations

This study shares common limitations with previous research (e.g., Gnambs and Lenhard, 2023) as we did not employ a true experimental design that involves randomly assigning children to test mode groups. Instead, for administrative convenience, entire classrooms were assigned to either the digital or print condition at the second measurement. While this approach was logistically more feasible, it introduced a limitation in terms of establishing a more robust causal relationship due to the potential for selection bias. Specifically, differences in classroom environments, teachers' instructional styles, or other classroom-level characteristics could systematically differ between conditions and influence student performance. To address this, we have controlled for baseline performance and other individual-level characteristics, such as gender and language status. This strategy aimed to reduce the effects of measured confounders at the student level. We acknowledge, however, that given the low number of classrooms that were randomly assigned to the test mode groups, unmeasured confounders (at student and classroom level) that differ by chance between print and digital mode classrooms may be an issue. The lack of randomization at individual level limits the strength of causal inferences that can be drawn from our findings. Therefore, while our analysis provides valuable insights into the test mode effect, the results should be interpreted with an understanding of these methodological constraints. Going forward, future studies may benefit from employing design improvements such as randomization at individual level to enhance causal inference. Replication in diverse educational settings with random assignment would further validate our findings and contribute to a more comprehensive understanding of the impact of assessment modes.

The fact that the processing of the individual items in the print version was not available for all children at second measurement, but only for around 35% of the children, represents a further limitation. It was shown that the GraWo scores did not differ between those children for whom the data for the processing of the individual items was available and those for whom this was not the case. However, there were observed differences in socioeconomic background and years of kindergarten attendance between the group of children in the print version with individual item data and the group without such data. These factors are known to influence educational outcomes, and as such, they may have impacted our results. The group with complete data had less favorable socioeconomic conditions and fewer years of kindergarten attendance, which could potentially confound the comparisons of performance between the print and digital groups as well. Future research should aim to ensure a more homogeneous sample or control for these variables to accurately isolate the effects of the mode of test administration.

While our research sheds light on the comparative efficacy of print versus digital test modes, we must note the omission of participants' digital competence data and the absence of observational data as two notable limitations. The potential variation in digital competence among students could be a confounding factor that influences their ability to perform optimally in a digital environment. Simultaneously, without observational data, the behavioral and interactive facets of testing that might differentially impact performance in digital versus print formats remain unexplored. Future studies should seek to incorporate both assessments of digital competence and structured observations during test-taking to derive a more granular understanding of how these variables intersect to influence students' assessment outcomes. Such an approach would contribute to a more rigorous and nuanced understanding of the dynamics involved in mode effects.

While our findings contribute valuable insights into the efficacy of print versus digital test modes, we acknowledge that the generalizability of these results may be limited by contextual factors unique to our study's setting. The implications of this research could vary significantly in different educational environments, particularly across regions with disparate levels of technological infrastructure and varying pedagogical traditions. For instance, students in areas with limited access to digital devices or those not routinely integrated into classroom learning may respond differently to test modes than those in our study cohort. Additionally, the familiarity with specific test formats due to prevailing instructional strategies could impact the outcomes observed in other contexts. Therefore, we emphasize the necessity for caution when extrapolating our study's conclusions to dissimilar settings. To ascertain the broader applicability of these findings, we endorse the undertaking of replication studies encompassing a diverse array of educational contexts. This would enhance the understanding of the conditions under which the observed advantages of either test mode are most pronounced.

4.3 Implications for research and practice

The findings emphasize the need for further investigations into the impact of test mode on student scores in hybrid assessment tools, considering not only test outcomes but also the cognitive processes and behaviors associated with different modes. In practice, educators and policymakers should acknowledge the potential differences in student scoring arising from test mode variations. The study highlights the importance of considering the specific demands of hybrid test procedures and tailoring educational practices accordingly. Moreover, educators need to be aware of the potential influence of digital competence on students' assessment experiences. As technology continues to play a central role in education, understanding and addressing the nuances brought forth by different test modes are imperative for informed decision-making in both research and educational settings.

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 at: https://osf.io/y29ps/?view_only=a8b09d29c9bc4509aa4a84a60f2e91e9.

Ethics statement

The studies involving humans were approved by Regional School Board for Upper Austria (Bildungsdirektion Upper Austria). The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation in this study was provided by the participants' legal guardians/next of kin.

Author contributions

SS: Conceptualization, Formal analysis, Methodology, Writing – original draft, Writing – review & editing. LP: Writing – review & editing. MS: Conceptualization, Data curation, Funding acquisition, Investigation, Project administration, Supervision, Writing – review & editing. CW: Formal analysis, Methodology, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. All costs related to this research project were borne by the University of Education Upper Austria and the Research Institute for Developmental Medicine (RID), Johannes Kepler University Linz. The tablets were provided by the project team with funds from the University of Education Upper Austria. The article processing charge was covered by the University of Graz Open Access Publishing Fund.

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/feduc.2024.1376805/full#supplementary-material

References

Backes, B., and Cowan, J. (2019). Is the pen mightier than the keyboard? The effect of online testing on measured student achievement. *Econ. Educ. Rev.* 68, 89–103. doi: 10.1016/j.econedurev.2018.12.007

Bauer, D. J. (2023). Enhancing measurement validity in diverse populations: modern approaches to evaluating differential item functioning. *Br. J. Math. Stat. Psychol.* 76, 435–461. doi: 10.1111/bmsp.12316

Bodmann, S. M., and Robinson, D. H. (2004). Speed and performance differences among computer-based and pencil-paper tests. *J. Educ. Comput. Res.* 31, 51–60. doi: 10.2190/GRQQ-YT0F-7LKB-F033

Buerger, S., Kroehne, U., Koehler, C., and Goldhammer, F. (2019). What makes the difference? The impact of item properties on mode effects in reading assessments. *Stud. Educ. Eval.* 62, 1–9. doi: 10.1016/j.stueduc.2019.04.005

Chaudron, S., Di Gioia, R., and Gemo, M. (2017). Young children (0-8) and digital technology - a qualitative study across Europe. Luxembourg: Publications Office of the European Union.

Clariana, R., and Wallace, P. (2002). Paper–based versus computer–based assessment: key factors associated with the test mode effect. *Br. J. Educ. Technol.* 33, 593–602. doi: 10.1111/1467-8535.00294

Cremer, M., and Schoonen, R. (2013). The role of accessibility of semantic word knowledge in monolingual and bilingual fifth-grade reading. *Appl. Psycholinguist.* 34, 1195–1217. doi: 10.1017/S0142716412000203

Dawidowsky, K., Holz, H., Schwerter, J., Pieronczyk, I., and Meurers, D. (2021). Development and evaluation of a tablet-based Reading fluency test for primary school children, 1–17. doi: 10.1145/3447526.3472033

Denckla, M. B., and Rudel, R. (1974). Rapid "automatized" naming of pictured objects, colors, letters and numbers by normal children. *Cortex* 10, 186–202. doi: 10.1016/S0010-9452(74)80009-2

Ennemoser, M., Marx, P., Weber, J., and Schneider, W. (2012). Spezifische Vorläuferfertigkeiten der Lesegeschwindigkeit, des Leseverständnisses und des Rechtschreibens. *Zeitschrift Entwicklungspsychologie Pädagogische Psychologie* 44, 53–67. doi: 10.1026/0049-8637/a000057

Gnambs, T., and Lenhard, W. (2023). Remote testing of Reading comprehension in 8-year-old children: mode and setting effects. *Assessment* 31, 248–262. doi: 10.1177/10731911231159369

Gottfredson, N. C., Cole, V. T., Giordano, M. L., Bauer, D. J., Hussong, A. M., and Ennett, S. T. (2019). Simplifying the implementation of modern scale scoring methods with an automated R package: automated moderated nonlinear factor analysis (aMNLFA). *Addict. Behav.* 94, 65–73. doi: 10.1016/j.addbeh.2018.10.031

Grob, A., Meyer, C. S., and Hagmann-von Arx, P. (2009). Intelligence and Development Scales (IDS).: Intelligenz-und Entwicklungsskalen für Kinder von 5–10 Jahren. Bern: Verlag Hans Huber.

Hamhuis, E., Glas, C., and Meelissen, M. (2020). Tablet assessment in primary education: are there performance differences between TIMSS' paper-and-pencil test and tablet test among Dutch grade-four students? *Br. J. Educ. Technol.* 51, 2340–2358. doi: 10.1111/bjet.12914

Hosseini, M., Abidin, M. J. Z., and Baghdarnia, M. (2014). Comparability of test results of computer based tests (CBT) and paper and pencil tests (PPT) among English language learners in Iran. *Procedia Soc. Behav. Sci.* 98, 659–667. doi: 10.1016/j.sbspro.2014.03.465

Ibrahim, L., Hamann, C., and Öwerdick, D. (2018). Identifying specific language impairment (SLI) across different bilingual populations: German sentence repetition task (SRT), vol. 2. Proceedings of the 42nd annual Boston University Conference on Language Development, 1–14. Somerville, MA: Cascadilla Press.

Jeong, H. (2012). A comparison of the influence of electronic books and paper books on reading comprehension, eye fatigue, and perception. *Electron. Libr.* 30, 390–408. doi: 10.1108/02640471211241663

Johnson, M., and Green, S. (2006). Online mathematics assessment: the impact of mode on performance and question answering strategies. *J. Technol. Learn. Assesment* 4, 1–33.

Juska-Bacher, B., Röthlisberger, M., Brugger, L., and Zangger, C. (2021). "Lesen im 1. Schuljahr: Die Bedeutung von phonologischer Bewusstheit, Benennungsgeschwindigkeit und Wortschatz" in Weiterführende Grundlagenforschung in Lesedidaktik und Leseförderung: Theorie, Empirie, Anwendung. eds. S. Gailberger and C. Sappok, (Bochum: Universitätsbiliothek der Ruhr Universität Bochum) 11–26.

Karay, Y., Schauber, S. K., Stosch, C., and Schüttpelz-Brauns, K. (2015). Computer versus paper--does it make any difference in test performance? *Teach. Learn. Med.* 27, 57–62. doi: 10.1080/10401334.2014.979175

Klassert, A. (2011). Lexikalische Fähigkeiten bilingualer Kinder mit Migrationshintergrund-Eine Studie zum Benennen von Nomen und Verben im Russischen und Deutschen [Lexical Abilities of Bilingual Children with an Immigrant Background – A Study of Naming of Nouns and Verbs in Russian and German]. Dissertation: University of Marburg.

Lee, K. S., Osborne, R. E., and Carpenter, D. N. (2010). Testing accommodations for university students with AD/HD: computerized vs. paper-pencil/regular vs. extended time. *J. Educ. Comput. Res.* 42, 443–458. doi: 10.2190/EC.42.4.e

Leeson, H. V. (2006). The mode effect: a literature review of human and technological issues in computerized testing. *Int. J. Test.* 6, 1–24. doi: 10.1207/s15327574ijt0601_1

Lenhard, A., Lenhard, W., and Schneider, W. (2020). ELFE II. Ein Leseverständnistest für Erst-bis Siebtklässler [A reading comprehension test for Grade 1 to 7 students]: Version II. Göttingen: Hogrefe, 4.

Lenhard, W., Schroeders, U., and Lenhard, A. (2017). Equivalence of screen versus print Reading comprehension depends on task complexity and proficiency. *Discourse Process.* 54, 427–445. doi: 10.1080/0163853X.2017.1319653

Melby-Lervåg, M., and Lervåg, A. (2014). Reading comprehension and its underlying components in second-language learners: a meta-analysis of studies comparing first-and second-language learners. *Psychol. Bull.* 140, 409–433. doi: 10.1037/a0033890

Merchant, G. (2015). Keep taking the tablets: iPads, story apps and early literacy. *AJLL* 38, 3–11. doi: 10.1007/BF03651950

Muter, V., Hulme, C., Snowling, M. J., and Stevenson, J. (2004). Phonemes, rimes, vocabulary, and grammatical skills as foundations of early reading development: evidence from a longitudinal study. *Dev. Psychol.* 40, 665–681. doi: 10.1037/0012-1649.40.5.665

Muthén, L. K., and Muthén, B. O. (1998-2012). Mplus User's Guide. *7th* Edn. Los Angeles, CA: Muthén & Muthén.

Muthén, L. K., and Muthén, B. (1998-2017). Los Angeles, CA: Mplus User's Guide.

Neumann, M. M., Anthony, J. L., Erazo, N. A., and Neumann, D. L. (2019). Assessment and technology: mapping future directions in the early childhood classroom. *Front. Educ.* 4. doi: 10.3389/feduc.2019.00116

Paleczek, L., Seifert, S., and Schöffl, M. (2021). Comparing digital to print assessment of receptive vocabulary with GraWo-KiGa in Austrian kindergarten. *Br. J. Educ. Technol.* 52, 2145–2161. doi: 10.1111/bjet.13163

Perfetti, C. A., and Hart, L. (2002). The lexical quality hypothesis. *Precurs. Funct. Liter.* 11, 67–86. doi: 10.1075/swll.11.14per

Poggio, J., Glasnapp, D. R., Yang, X., and Poggio, A. J. (2005). A comparative evaluation of score results from computerized and paper and pencil mathematics testing in a large scale state assessment program. *J. Technol. Learn. Assesment* 3.

Puhan, P., Boughton, K., and Kim, S. (2007). Examining differences in examinee perfomance in paper and pencil and computerized testing. *J. Technol.* 6, 1–20.

Ricoy, M.-C., and Sánchez-Martínez, C. (2020). Revisión sistemática sobre el uso de la tableta en la etapa de educación primaria [A systematic review of tablet use in primary education]. *Revista Española Pedag.* 78, 273–290. doi: 10.22550/REP78-2-2020-04

Schermelleh-Engel, K., Moosbrugger, H., and Müller, H. (2003). Evaluating the fit of structural equation models: tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research*. 8, 23–74. doi: 10.23668/psycharchives.12784

Schöfl, M., Steinmair, G., Holzinger, D., and Weber, C. (2022). Predicting word reading deficits using an app-based screening tool at school entry. *Front. pediatr.* 10:863477. doi: 10.3389/fped.2022.863477

Seifert, S., and Paleczek, L. (2022). Comparing tablet and print mode of a German reading comprehension test in grade 3: Influence of test order, gender and language. *Int. J. Educ. Res.* 113. doi: 10.1016/j.ijer.2022.101948

Seifert, S., Paleczek, L., and Gasteiger-Klicpera, B. (2019). Rezeptive Wortschatzleistungen von Grundschulkindern mit Deutsch als Erst- und Zweitsprache und der Zusammenhang zu den Lesefähigkeiten: Implikationen für einen inklusiven Unterricht. *Empirische Sonderpädagogik*, 4, 259–278. doi: 10.25656/01:18334

Seifert, S., Paleczek, L., Schwab, S., and Gasteiger-Klicpera, B. (2017). *Grazer Wortschatztest – GraWo [Graz vocabulary test)*. Göttingen: Hogrefe.

Singer-Trakham, L. M., Alexander, P. A., and Berkowitz, L. E. (2019). Effects of processing time on comprehension and calibration in print and digital mediums. *J. Exp. Educ.* 87, 101–115. doi: 10.1080/00220973.2017.1411877

Statistik Austria. (2022). Bildung in Zahlen 2020/21- Schlüsselindikatoren und Analysen. Wien: Statistik Austria.

Steedle, J. T., Cho, Y. W., Wang, S., Arthur, A. M., and Li, D. (2022). Mode effects in college admissions testing and differential Speededness as a possible explanation. *Educ. Meas. Issues Pract.* 41, 14–25. doi: 10.1111/emip.12484

Støle, H., Mangen, A., and Schwippert, K. (2020). Assessing children's reading comprehension on paper and screen: a mode-effect study. *Comput. Educ.* 151:103861. doi: 10.1016/j.compedu.2020.103861

Taherbhai, H., Seo, D., and Bowman, T. (2012). Comparison of paper-pencil and online performances of students with learning disabilities. *Br. Educ. Res. J.* 38, 61–74. doi: 10.1080/01411926.2010.526193

Trautwein, J., and Schroeder, S. (2019). WOR-TE: Ein Ja/Nein-Wortschatztest für Kinder verschiedener Altersgruppen. *Diagnostica* 65, 37–48. doi: 10.1026/0012-1924/a000212

Wagner, I., Loesche, P., and Bißantz, S. (2022). Low-stakes performance testing in Germany by the VERA assessment: analysis of the mode effects between computer-based testing and paper-pencil testing. *Eur. J. Psychol. Educ.* 37, 531–549. doi: 10.1007/s10212-021-00532-6

Wang, S., Jiao, H., Young, M. J., Brooks, T., and Olson, J. (2007). A Meta-analysis of testing mode effects in grade K-12 mathematics tests. *Educ. Psychol. Meas.* 67, 219–238. doi: 10.1177/0013164406288166

Wang, S., Jiao, H., Young, M. J., Brooks, T., and Olson, J. (2008). Comparability of computer-based and paper-and-pencil testing in K–12 Reading assessments. *Educ. Psychol. Meas.* 68, 5–24. doi: 10.1177/0013164407305592

Wang, T.-H., Kao, C.-H., and Chen, H.-C. (2021). Factors associated with the equivalence of the scores of computer-based test and paper-and-pencil test: presentation type, item difficulty and administration order. *Sustain. For.* 13:9548. doi: 10.3390/su13179548

Wendt, H., and Schwippert, K. (2017). "Lesekompetenzen von Schülerinnen und Schülern mit und ohne Migrationshintergrund" in IGLU 2016: Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich. eds. A. Hußmann, H. Wendt, W. Bos, A. Bremerich-Vos, D. Kasper and E.-M. Lankeset al. (Münster, New York: Waxmann), 219–234.

Yu, C.-Y. (2002). Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes. Dissertation. University of California, Los Angeles, CA.

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 20 August 2024
DOI 10.3389/fpsyg.2024.1380036

OPEN ACCESS

EDITED BY
Natalie Förster,
University of Wuppertal, Germany

REVIEWED BY
Peter Adriaan Edelsbrunner,
ETH Zürich, Switzerland
Sarah Powell,
The University of Texas at Austin, United States

*CORRESPONDENCE
Marco Ennemoser
✉ ennemoser@ph-ludwigsburg.de

†These authors share senior authorship

RECEIVED 31 January 2024
ACCEPTED 29 July 2024
PUBLISHED 20 August 2024

CITATION
Ennemoser M, Sinner D, Nguyen L and Krajewski K (2024) From developmental theory to effective training: long-term and transfer effects of promoting the quantity–to–number word linkage in first-graders at risk for mathematical difficulties. *Front. Psychol.* 15:1380036. doi: 10.3389/fpsyg.2024.1380036

From developmental theory to effective training: long-term and transfer effects of promoting the quantity–to–number word linkage in first-graders at risk for mathematical difficulties

Marco Ennemoser[1*†], Daniel Sinner[2], Linda Nguyen[1] and Kristin Krajewski[3†]

[1]Faculty of Participation Science, University of Education Ludwigsburg, Ludwigsburg, Germany, [2]Institute for Psychology, Justus Liebig University Gießen, Giessen, Germany, [3]Institute for Psychology, University of Education Ludwigsburg, Ludwigsburg, Germany

Introduction: The model of quantity–to–number word linkage (*QNL model*) identifies relevant milestones in the process of early numerical acquisition and describes a developmental sequence that can guide the fostering of foundational mathematical abilities in at-risk children. While there is substantial evidence for the predictive value of the quantity-number competencies (QNC) described by the model, evidence supporting the preventive potential of interventions targeting these QNC is so far largely restricted to short-term effects. Findings regarding their long-term preventive impact, especially in terms of transfer to mathematical school achievement, are still limited. This quasi-experimental study aimed to address this gap by evaluating the long-term transfer effects of an intervention program that is strictly derived from the QNL model of mathematical development [QNL training; in German "Mengen, zählen, Zahlen" (MZZ)].

Methods: We assessed the quantity-number competencies of 575 first-graders and identified 119 of them as being at risk for mathematical learning difficulties, who were then assigned to three experimental conditions. Sixty one children received 12 sessions of the QNL training, while 30 underwent training in inductive reasoning. Another 28 children served as a control group, receiving no specific intervention.

Results and Discussion: Multi-level analyses confirmed both significant short- and long-term effects in the specifically trained quantity–number competencies as well as transfer effects on subsequent mathematical school achievement. In accordance with previous findings, transfer effects of the QNL training on mathematical school achievement were not yet evident immediately after the intervention but turned out to be significant after a delay of 6 months and remained stable even 15 months after training. Effect sizes ranged from $d = 0.32$ to $d = 1.12$. These findings both underscore the preventive potential of interventions that are strictly driven by developmental theory and, conversely, support the theoretical assumptions of the QNL model.

KEYWORDS

number sense, prevention, developmental dyscalculia, approximate number system, math intervention, math development, QNL model, numerical vocabulary development

1 Introduction

1.1 Perspectives on the prevention of mathematical learning difficulties

One of the most important challenges in research on developmental dyscalculia is to find effective means of prevention. However, as noted by Szücs and Goswami (2013), many theoretical approaches in the field provide a rather pessimistic view of the prospects for achieving this goal, in that they suggest that children with developmental dyscalculia are lacking a cognitive core module critical to understanding mathematics, i.e., they display a neurological dysfunction (Shalev, 2004; Kucian et al., 2006; Mussolin et al., 2010). Those authors point out that this notion "may suggest to educators an irreversible condition" (p. 36). In the current paper we take a more optimistic position. Based on the *developmental model of quantity–to–number word linkage* (*QNL model*) (Krajewski, 2008; see Krajewski and Schneider, 2009a,b), we suggest that what is often labeled as developmental dyscalculia might more accurately be described as a delay in numerical development rather than a deficient cognitive module. Consequently, while we also use the term "developmental dyscalculia" to stay aligned with established clinical and educational terminology, we prefer to describe persistent challenges using the term "mathematical difficulties." Unlike dyscalculia, this term does not imply assumptions on an immutable neurological deficit but rather emphasizes developmental gaps that can potentially be closed through specific training.

In the following section, we will give a brief description of the underlying model of early mathematical development presumed here and will suggest that effective prevention programs should address the particular levels of competence described in the model. Furthermore, we will report findings on the predictive value of the corresponding *quantity–number competencies* (QNC) and provide a discussion of existing studies on the effectiveness of early interventions for children with low mathematical achievement or at-risk children, respectively.

A particular focus will be on methodological limitations of the available studies, limitations that lead to the notion that the most important questions regarding the effectiveness of dyscalculia prevention programs remain unanswered, although most available studies report some kind of significant effect on numerical abilities. To overcome this unsatisfactory situation, we will make two suggestions for future research on the prevention and early intervention in math difficulties or dyscalculia:

(1) The content of a training program should be derived strictly based on a developmental theory able to explain why and how a particular training approach should lead to (long-term) improvement in subsequent school achievement in mathematics; and

(2) Research on the prevention of math difficulties should more seriously address some (well-known) methodological challenges that have yet to be met by evaluation studies, in order to provide more reliable evidence about what works (and does not), why, and how.

On the basis of these two suggestions, we will introduce a training program in quantity-to-number-word linkage (*QNL*) "Mengen, zählen, Zahlen" (Quantities, counting, numbers) (Krajewski et al., 2007) based on the developmental QNL model. Finally, we will report the results of a longitudinal intervention study on the effectiveness of this training program, through which we tried to address the methodological challenges that we will have discussed.

1.2 Developmental model of quantity–to–number word linkage (QNL model)

1.2.1 Key principle and core assumptions of the QNL model

Principle of conservative competence attribution. The model of quantity–to–number word linkage (Krajewski, 2008; for English-language discussion, see also Krajewski and Schneider, 2009a,b, and the meta-analysis by Lin and Powell, 2023) describes numerical development (that is, the development of quantity–number competencies) from birth through to primary school. A key distinction of the QNL model from other approaches is the *principle of conservative competence attribution*. This principle posits that any observable mathematical "output" in a child's behavior should be interpreted conservatively to avoid overestimating the child's numerical abilities. For example, unlike other researchers in the field (e.g., Fuson, 1988; von Aster and Shalev, 2007), the QNL model suggests that merely saying a single number word does not necessarily imply that the child fully understands its meaning (i.e., that the word precisely represents a specific quantity). The QNL model attributes observed performances, such as reciting number words backwards, only to competencies deemed necessary, essential, and sufficient to produce that performance.

This conservative attribution allows for a more sensitive detection of potential hurdles that make some children struggle with mathematics. It provides a finer grain size in the analysis of developmental steps that need to be taken and their significance in the subsequent development of mathematical competency. This precise identification of potential hurdles in the developmental process enables the design of more tailored and potentially more effective interventions.

Innate versus emerging numerical abilities. While other theoretical frameworks (e.g., Fuson, 1988; Dehaene, 1992; Butterworth, 2005; von Aster and Shalev, 2007) suggest that children are born with a notion of what a number is and understand that number words—once separated from the "numerical string" (Fuson, 1988)—are associated with quantities, the QNL model takes a different position. It posits that the connection between quantities and number words (as well as other symbolic representations) is something that develops gradually over a longer period of time. This is the reason why the framework is referred to as the model of quantity-to-number-word linkage. It makes the following three claims about the linkage between quantities and number words.

(1) Conceptual understanding of the linkage between quantities and numbers is the most critical prerequisite for subsequent acquisition of school mathematics—that is, here lies the *core deficit* of mathematical difficulties and developmental dyscalculia.

(2) All children are naturally born with a "deficit" in grasping discrete numerical quantities and, consequently, have to learn that number words are linked to particular quantities rather than to other characteristics of an object like color or usage. This is also true for small numbers, which are usually supposed to be apprehended with a single glance as children grow older (i.e., that lie within the subitizing range). The *quantity–to–number word linkage has to be acquired* via a developmental process; that is, this so-called deficit is not specific to children with developmental dyscalculia.

(3) While many children succeed effortlessly in acquiring this linkage, others need—and can effectively receive—explicit instruction.

These claims imply that early instruction in quantity–to–number word linkage should lead to long-term improvement in subsequent acquisition of mathematics during regular school instruction.

1.2.2 Developmental steps

In her model, Krajewski identifies three key milestones. Achieving these milestones marks and facilitates children's progression toward a deeper understanding of the quantity–to–number word linkage (see Figure 1).

Level 1: unlinked numerical skills (number words and numerals isolated from quantities). Newborns, starting at level 1 in the model, can assess whether objects or continuous quantities differ (*quantity discrimination*), effectively distinguishing between *indiscrete* amounts, i.e., between amounts that are not precisely countable but differ in scope and extent (e.g., Feigenson et al., 2002; Rousselle et al., 2004; e.g., Turati et al., 2013). Researchers have different views on whether this also implies that newborns and infants can differentiate between discrete (i.e., potentially countable) quantities at this point; however, according to the QNL model this capability is not considered necessary for normal developmental progression.

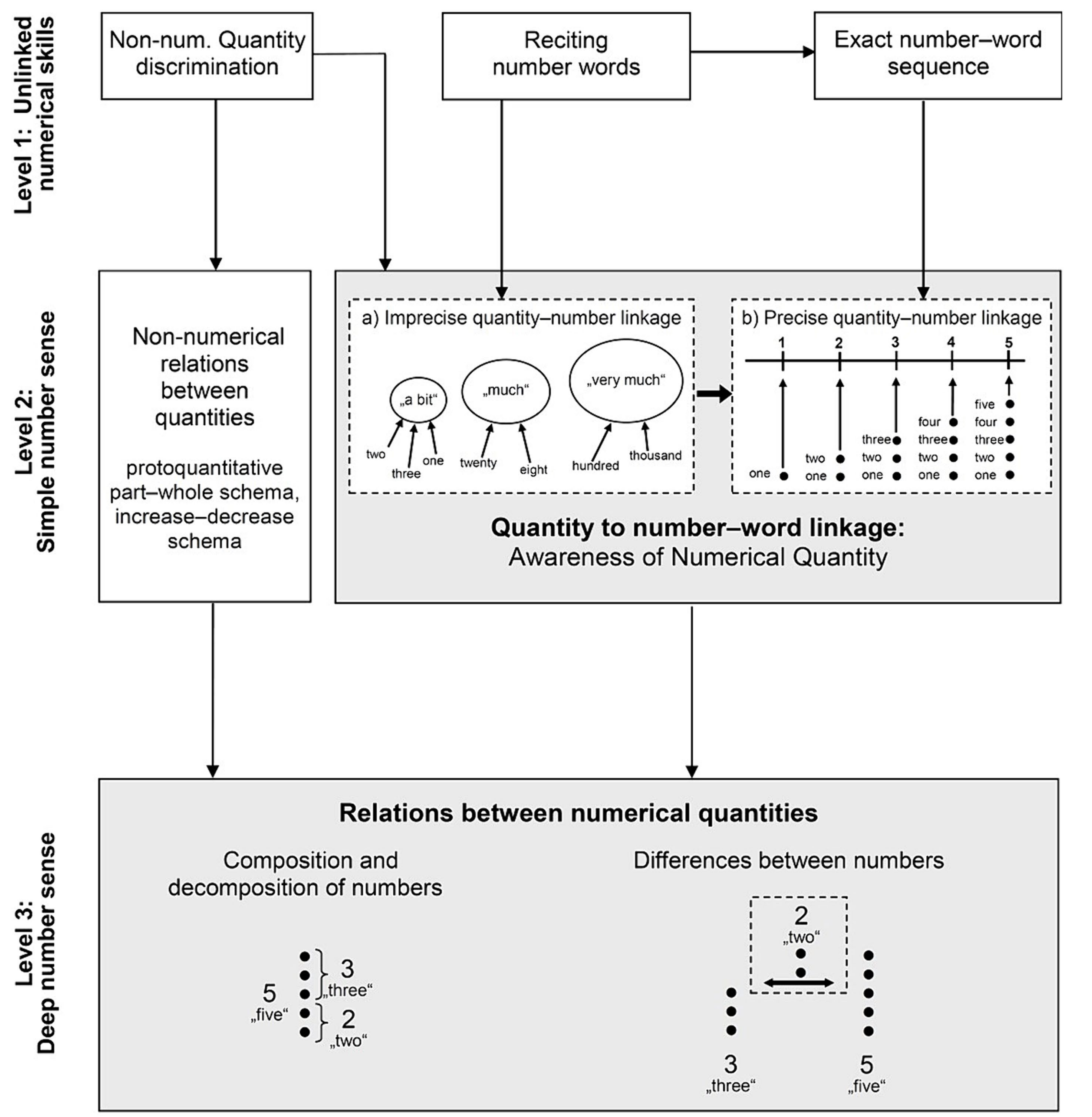

FIGURE 1
Developmental model of quantity–to–number word linkage (QNL model) (Krajewski, 2008).

A second skill that develops in the following years but is still located at level 1 is the knowledge of *number words* and the number-word sequence. According to the QNL model, these two skills develop—at least initially—independently, that is, when children start to recite number words (e.g., "one, two, three, four, five"), they do not necessarily know that a given number word (e.g., *three*) refers to a quantity or magnitude (e.g., to all three fingers that were just counted). Rather, number words might be perceived as just reflecting a kind of labeling used when things are pointed to in sequence (e.g., *three* may be a roughly synonymous label for "middle finger"). In other words, at this early level, the use of number words by a child does not necessarily imply an understanding of their quantitative meaning.

Level 2: simple number sense (quantity–to–number word linkage). Krajewski argues that children gradually learn to understand the linkage between numbers and quantities. Most children already acquire this knowledge before the beginning of formal mathematics instruction in school. In the beginning, they only have a very vague and imprecise understanding of the quantitative meaning of a number word (*imprecise quantity–to–number word linkage*, level 2a). For instance, they may understand that some number words correspond with rough verbal categories like "a bit," that is, represent only small quantities, while other number words are associated with terms like "much" or "very much," corresponding to large or very large quantities. For example, the number word *three* may correspond to the verbal category "a bit," while *a hundred* might indicate "very much." Of course, the verbal terms for magnitudes ("a bit," "much," "very much," etc.) are just very rough categories that do not sharply distinguish between clearly defined number spaces. Consequently, children at this level often fail when they have to differentiate between number words that belong to the same category of verbal magnitude. Thus, if a child assigns the words *twenty-three* and *twenty-five* to the same category (e.g., "much") he or she will not be able to tell which of these two number words represents "more" than the other. This would require a more precise categorical system than one of rough verbal categories—that is, a *precise quantity–to–number word linkage*, which characterizes the next phase of level 2 (level 2b). At this level, children are able to assign exactly countable quantities to an exact position in the number-word sequence, and therefore to differentiate even near numbers on the basis of their quantitative meaning (e.g., *twenty-four* is more than *twenty-three* and both are less than *twenty-five*), because now each number word is associated with an exactly defined quantity.

The developmental process described from level 2a to level 2b (i.e., from an initially imprecise to a precise concept of quantity) aligns with theoretical assumptions regarding the development of an approximate number sense (*ANS*) (Halberda and Feigenson, 2008). Specifically, this process involves a shift from a very imprecise representation of numbers, where the degree of imprecision increases with larger numerosities (corresponding to QNL model Level 2a), to a more precise numerical representation (corresponding to QNL model Level 2b). This transition reflects the enhancement of children's ability to differentiate between quantities more accurately, thereby improving their overall numerical understanding and competence.

At the same time as children become able to elaborate these relations between the quantities indicated by number words, children also gather experiences with *relations between quantities (without reference to number words)*. They become aware that the size of a non-numerical quantity only changes if something is added or taken away, and that quantities can be divided into pieces that, when put together again, equal the original quantity (see also Resnick, 1989, on the *protoquantitative decrease/increase* and *part–whole schema*).

Level 3: deep number sense (concept of number relationships). At the third level, the awareness of relations between non-numerical quantities (that has emerged at level 2, see above) is transferred to *numerical relations*. Children now grasp that when a discrete number of items are divided into smaller pieces, both the resulting subsets and the *relation* between these subsets can themselves be represented (exactly quantified) by means of number words. In practice this includes the ability to recognize that the number *eight* can be divided into the numbers *five* and *three (de/composition of numbers)* and that the subset *five* includes exactly two more than the subset *three (differences between numbers)*. The assumption upon which Krajewski's model rests is that these insights basically have to be understood on a verbal level, that is, by the use of number words. The knowledge of Arabic numeral symbols (e.g., 8, 5, 3) may be helpful, but it is not a necessary precondition for this level of development. Nor is calculating on a symbolic level (e.g., $8-5=3$) required.

1.3 Learning number words as a case of vocabulary acquisition

As outlined above, based on the *principle of conservative competence attribution* the QNL model suggests that merely saying a number word does not necessarily imply that a child precisely understands the meaning of this word. In our view, this phenomenon perfectly parallels the acquisition of any other word in our vocabulary. When a child uses a new word like *dog* for the first time (maybe just by parroting what it has heard another person say), it usually does not have a clear notion of what this word means. Rather, the exact meaning of the word *dog* has to be learned via an evolving process of comparing and categorizing, of testing and rejecting hypotheses, until the child has developed a sophisticated network of more or less distinctive attributes that constitute the mental representation (i.e., the meaning) of the word *dog* (e.g., Bloom, 2000). This representation might include: has legs, is barking, is not a cat, is a living being, or—somewhat later and more abstractly—is an animal, includes various different breeds. A very similar developmental process is necessary to acquire an appropriate mental representation of a number word. For example, in order to attain a deep conceptual understanding of the number word *six*, there are a lot of relevant attributes that have to be linked to this word. Examples of attributes that have to be acquired progressively are that *six*:

- is usually said between the words *five* and *seven* (level 1),
- is a kind of label that is often used when objects are pointed to in sequence (level 1),
- has something to do with quantities (level 2a),
- is not very much (level 2a),
- is less than *one hundred* (level 2a),
- is more than *five* and less than *seven* (level 2b),
- is exactly one more than *five* and one less than *seven* (level 3) – which is not the same as simply knowing that six is usually said *between* five and seven (see above)
- can be decomposed into *four* and *two* (level 3), and
- is the difference between *ten* and *four* (level 3).

Our notion of learning number words as a case of vocabulary acquisition seems particularly suited to underpin and to clarify the core assumption of the QNL model, that the linkage between number words and quantities has to be established in a developmental process.

1.3.1 Practical implications

Both the principle of conservative competence attribution and the notion of number-word learning as a case of vocabulary acquisition imply the need for a much closer look than has been taken by other theoretical frameworks at how children acquire an understanding of number words as the representatives of numerical quantities. The linkage between quantities and numbers is not simply "on" or—in a few children, those with developmental dyscalculia—"off." Rather, it has to be acquired passing through a developmental sequence. The QNL model gives a detailed theoretical description of this developmental sequence, including particular milestones that have to be mastered gradually and, thus, provides promising starting points for targeted interventions, starting points that might otherwise be neglected. According to the QNL model, dyscalculia prevention programs should strictly address this developmental sequence from level 1 to level 3, with a particular focus on establishing the linkage between number words and quantities (level 2).

The practical value of such a developmental model can be evaluated, first, by investigating the predictive power of the proposed quantity–number competencies for subsequent mathematical school achievement in typically-performing children. Second, it can be tested and determined whether children with math difficulties show deficits in the proposed developmental milestones. Third, a training program based on this model and designed to foster children's mathematical understanding along the path laid out by these milestones is likely to effectively facilitate subsequent development and—in the long run—lead to better mathematical school achievement. In what follows, we will first conduct a theoretical discussion to shed light on points one and two above. Afterwards, we will take a look at existing studies on fostering numerical abilities in children, considering the third requirement in their light, before finally leading into our empirical study.

1.4 Validity of the described quantity–number competencies for numerical learning in children with and without dyscalculia

Longitudinal studies from many countries provide consistent evidence for the importance of the quantity–number competencies described above[1] for primary mathematical school achievement when assessed before or around school entry (e.g., Passolunghi et al., 2007; Jordan et al., 2012; Watts et al., 2014). For example, Aunola et al. (2004), in a Finnish longitudinal study, found that level-1 competencies (reciting the number-word sequence forwards and backwards from a given starting point) predicted 38% of the level and 5% of the growth rate of central domains of first-and second-graders' elementary mathematical school curriculum. Moreover, in a German longitudinal study by Krajewski and Schneider (2009a), 24% of the variance in mathematical school achievement at the end of fourth grade was predicted by level-2 competencies (quantity–to–number word linkage) that had been assessed in kindergarten 4 years before, while in turn 58% of level-2 competencies were predicted by competencies located on level 1. Similarly, Ennemoser et al. (2017) found that quantity-number competencies assessed at school entry explained 42% of the variance in math achievement by the end of 4th grade, even after controlling for nonverbal intelligence and arithmetic fact retrieval. In sum, quantity–number competencies are the strongest known predictor of later mathematical school achievement in normal-performing children. Moreover, these predictions reliably distinguish between primary school children with and without mathematical learning difficulties. Children with mathematical difficulties consistently show deficits on all three levels of quantity–number competency, in various domains: naming, reading, and writing numbers, arranging numbers on a scaled number line as well as reciting number words (all situated at level 1); matching Arabic numerals with representations of their magnitudes and comparing numbers according to their cardinal value (level 2); deficits in part-whole relations of numbers (level 3) (e.g., Gaupp et al., 2004; Geary et al., 2004; Landerl et al., 2004; Krajewski and Schneider, 2009a; Sinnakaudan and Ghazali, 2015).

Cross-sectional and longitudinal data provide important evidence for the assumptions of the QNL model. Additionally, intervention studies can make a particularly significant contribution by testing whether fostering QNL competencies, as defined by the model, indeed enhances future mathematical competence development. The evidence available to date and the research needs in this area are discussed in the following sections.

1.5 Evidence from early interventions to foster mathematical achievement

1.5.1 Essential criteria for the evaluation of QNC-based mathematics interventions

Merely demonstrating that quantity-number competencies (QNC) substantially predict subsequent mathematical skill acquisition is not enough to fully test the theoretical assumptions of the QNL model. Given the perspective that mathematical difficulties are rooted not in a deficient number module but in a developmental delay in QNC—a delay that can be addressed through targeted intervention—it is critical to demonstrate the efficacy of such interventions. Three aspects are essential in this regard. First, it should be shown that QNC can be effectively fostered—particularly in children with weak initial skills (i.e., those with a developmental delay). Second, these effects should be sustainable, persisting beyond the intervention period. Third, it is crucial that the benefits of interventions extend beyond the trained QNC, transferring to mathematical skills that are subsequently taught during the primary school curriculum, as measured by a standardized, curriculum-based mathematics achievement test.

1 For instance, level 1: reciting the number-word sequence forwards and backwards, naming preceding and subsequent numbers; level 2: linking number words to corresponding quantities, comparing numbers in terms of magnitude; level 3: determining numerical differences between numbers.

1.5.2 The challenge of heterogeneity in mathematics intervention research

Given the abundance of research on math interventions published in the last two decades, it is surprising that the evidence specifically addressing the three aforementioned criteria remains notably incomplete. One main reason is the heterogeneity of research in the field of mathematics interventions, a diversity that is also reflected in the available reviews and meta-analyses (e.g., Mononen et al., 2014; Wang et al., 2016; Jitendra et al., 2021; Myers et al., 2022; Schnepel and Aunio, 2022; for a detailed overview, see Svane et al., 2023). Studies exhibit considerable variability, covering a range of age groups (from 2 years old to adulthood), targeting diverse populations (including unselected samples, various "at-risk" groups, and children with learning difficulties), and encompassing different educational settings (such as kindergarten and special education programs). There is also a considerable variability in the primary focus of the interventions, with some focusing narrowly on specific mathematical skills such as counting skills or arithmetic. Others take a more comprehensive approach, integrating a broader spectrum of mathematical abilities, and sometimes they even incorporate activities from other domains, such as reading (Svane et al., 2023). In the following sections, we provide an overview of studies that specifically focus on fostering QNC.

1.5.3 Evidence from QNC-based interventions: short-term specific effects

Studies with a specific focus on quantity-number competencies (QNC), in line with the QNL model, are predominantly conducted in early childhood education settings and in the early years of formal schooling. Encouragingly, these studies consistently indicate that targeted interventions can effectively enhance QNC across a diverse range of groups, including children with initially weak QNC, i.e., those at a high risk of developing math difficulties (Mononen et al., 2014). However, this addresses only the first of the three criteria previously mentioned. Findings related to the two remaining aspects—the sustainability of effects and their transfer to later (curriculum-based) mathematics performance—are largely lacking. Intervention studies in the field rarely include a follow-up assessment to verify the longevity of effects over time, and even fewer studies explicitly examine whether training-induced gains in QNC actually translate into improved school mathematics performance (Mononen et al., 2014). Table 1 provides an overview of 20 intervention studies that focus on promoting foundational mathematical abilities. It details the setting (kindergarten vs. school), the characteristics of the sample (unselected children, children at risk for mathematical difficulties, children with low performance in mathematics, or those diagnosed with developmental dyscalculia), and whether a control group was included, specifying if it was an active control group (i.e., one that received an alternative intervention to control for attention effects). Additionally, the table reports the types of effects examined in each study, distinguishing whether the research focused solely on training-specific effects, on targeted skills or also explored transfer effects on non-targeted aspects of mathematical achievement. It also differentiates whether these effects were examined in the short-term or long-term. The overview, while potentially not exhaustive, effectively highlights the aforementioned challenges. For a detailed description of these studies, including what was investigated, how it relates to the QNL model, and which methodological weaknesses might have been present, please refer to Supplementary material.

1.5.4 Limited evidence for long-term and transfer effects on later mathematics performance

As illustrated in Table 1, nine of the reported studies examined the effectiveness of different mathematics intervention programs (e.g., Number Worlds, Numerical Board Game, The Number Race) in unselected samples of kindergarten children. While all of them reported short-term training effects on the targeted numerical skills, only two also investigated the sustainability of these effects, with only one providing evidence for significant long-term improvements in the trained numerical abilities.

Three additional studies also examined the effectiveness of various intervention programs in a kindergarten setting (e.g., Additional Early Mathematics), specifically focusing on children identified as being at-risk based on their low mathematical skill levels. Again, all three studies reported significant short-term training effects on the targeted numerical skills. Furthermore, two of these studies examined long-term effects and found supportive evidence for the sustainability of the interventions on numerical abilities that were specifically trained.

Eight studies investigated the effectiveness of early mathematics interventions for school children, with four using unselected samples (e.g., Fuchs et al., 2014) and four focusing on school children with mathematical difficulties (Rescue Calcularis, Galaxy Math Program). Except for the study by Hellstrand et al. (2020, an evaluation of The Number Race) all reported significant short-term effects on the trained numerical abilities. However, none of the eight school studies examined long-term effects on mathematical abilities. Moreover, none of the 20 intervention studies examined long-term transfer effects on mathematical abilities that were not targeted by the training or even included a standardized curriculum-based mathematics test, which would be particularly noteworthy.

In sum, there is a clear lack of evidence regarding long-term effects, particularly long-term transfer effects related to subsequent school achievement in mathematics. It is important to note that this conclusion is not based on a large proportion of insignificant results, which would not prove the absence of an effect (Edelsbrunner and Thurn, 2024), but on the gaping lack of studies that even explore these aspects.

1.6 The training of quantity–to–number-word linkage "Mengen, zählen, Zahlen" (QNL training)

1.6.1 Theoretical background

Most of the training approaches mentioned above are based on theories positing the existence of an innate core deficit in number sense among children with developmental dyscalculia. The program Mengen, zählen, Zahlen (Krajewski et al., 2007) is founded on Krajewski's model of quantity–to–number word linkage that proposes that the basis of dyscalculia and mathematical difficulties lies in a delay in children's numerical development rather than an inherent core deficit. Consequently, the aim of the program is to close developmental gaps by tracing the developmental sequence as described in the QNL model in order to prevent mathematical difficulties or developmental dyscalculia, respectively. It is

TABLE 1 Interventions targeting foundational mathematical skills.

Authors/ Intervention	Setting	Control group(s)	Sample	Short-term effects		Long-term effects	
				Specific[a]	Transfer[b]	Specific[a]	Transfer[b]
Griffin et al. (1995) Number Worlds	Kindergarten	Yes (active)	Unselected	Sig.	Sig.[c]	–	–
Sterner et al. (2019) Number Worlds	Kindergarten	Yes (active)	Unselected	Sig.	–	n.s.	–
Additional Early Mathematics	Kindergarten	Yes	At-risk	Sig.	–	Sig.	–
Van Luit and Schopman (2000) Additional Early Mathematics	Kindergarten	Yes	At-risk	Sig.	n.s.	Sig.	–
Wilson et al. (2006) The Number Race	Kindergarten	No	Unselected	Sig.	–	–	–
Wilson et al. (2009) The Number Race	Kindergarten	Yes	Unselected	Sig.	–	–	–
Räsänen et al. (2009) The Number Race	Kindergarten	Yes (active)	At-risk	Sig.	n.s.	–	–
Hellstrand et al. (2020) The Number Race	School	Yes	At-risk	n.s.	n.s.	–	–
Obersteiner et al. (2013) The Number Race	School	Yes (active)	Unselected	Sig.	–	–	–
Siegler and Ramani (2008) Numerical Board Game	Kindergarten	Yes (active)	Unselected	Sig.	–	–	–
Whyte and Bull (2008) Numerical Board Game	Kindergarten	Yes (active)	Unselected	Sig.	–	–	–
Hawes et al. (2019) Numerical Board Game	Kindergarten	Yes (active)	Unselected	Sig.	–	–	–
Kucian et al. (2011) Rescue Calcularis	School	No	Developmental Dyscalculia	Sig.	n.s.	–	–
Käser et al. (2013) Rescue Calcularis	School	Yes (waiting list)	Low achievers[d]	Sig.	Sig.	–	–
Fischer et al. (2011)	Kindergarten	Yes (active)	Unselected	Sig.	–	–	–
Link et al. (2013)	School	Yes (active)	Unselected	Sig.	–	–	–
Hyde et al. (2014)	School	Yes (active)	Unselected	Sig.	–	–	–
Fuchs et al. (2013) Galaxy Math Program	School	Yes	At-risk	Sig.	–	–	–
Fuchs et al. (2014)	School	Yes (active)	Unselected	Sig.	Sig.	–	–
Praet and Desoete (2014)	Kindergarten	Yes (active)	Unselected	Sig.	–	Sig.	–

–, not assessed; Sig., significant; n.s., not significant.
[a]Effects on training specific tasks.
[b]Transfer effects on non-trained numerical or arithmetic tasks.
[c]Transfer effects were assessed 4 weeks after training, which we did not consider indicative of long-term effects.
[d]Selection based on teacher judgment.

important to note that the QNL training does—of course—comprise the use of number words and corresponding numerals, but it does not contain any formal arithmetic tasks at the symbolic level.

1.6.2 Structure and succession of the training

The QNL training was originally developed for five-to six-year-old kindergarten children. In its original kindergarten version, it is conducted three times per week in small groups with four to six children over the course of 8 weeks (in all, 24 sessions at 30 min each). In several studies with at-risk first graders, which are described in the following section, the program was condensed to twelve 45-min sessions conducted twice a week. The initial focus of the training is on exercising foundational numerical skills, for example, naming verbal and Arabic numerals, or comparing non-numerical quantities. During the corresponding sessions, number-word sequences and Arabic numerals up to 10 are introduced (e.g., which number belongs between two other numbers; level 1) with a focus on the awareness that individual verbal and Arabic numerals are associated with particular quantities (*quantity–to–number word linkage*, level 2; e.g., which number word belongs to a given discrete quantity). Afterwards, children are trained to understand that numbers placed later in the number-word sequence are associated with larger quantities (*number seriation*, level 2). Here children learn that ascending numbers represent an increase in discrete pieces/items or in amounts/quantities of space, while descending numbers represent fewer items or lesser amounts. The aim of the final section is to facilitate the awareness that numbers are related to each other in a particular way (*number relations*). Corresponding exercises support the understanding that numbers can be decomposed into smaller numbers and that pooling these subsets in turn leads to the original number again (de/composition of numbers, level 3; e.g., learning which number results when we put *two* and *three* together and what numbers (can) result when we break *five* apart). A further set of exercises aims at promoting children's understanding that the difference between two numbers can be exactly described by a third number (the *difference between numbers*, level 3; e.g., finding out what number includes two more than *four*).

1.6.3 Additional principles of the program: accounting for language, working memory, executive functioning

Additional principles of the program proactively consider relevant third variables that might impact the training's effectiveness. Numerous studies have shown that deficits in mathematics are associated with deficits in other areas. These include, in particular, language competence (Purpura and Ganley, 2014; Lin et al., 2021), working memory (Friso-van den Bos et al., 2013; Zhang et al., 2022), meta cognition (Schneider and Artelt, 2010; see also Baten et al., 2017) and executive functions (Simanowski and Krajewski, 2019; Ribner et al., 2023). Although a focused mathematics training program cannot directly enhance general language competence, it can notably emphasize the use of mathematical language. To achieve this, the training extends beyond linking numbers to number words by also incorporating how to verbally express quantity-related situations. Appropriate verbal descriptions of such situations are introduced and practiced. Thus, children are guided to understand relevant principles not only visually (e.g., "Larger numbers include more space") but also to describe and justify them using language (e.g., "Four is bigger than three because four includes more pieces than three").

The program also takes into account low working memory and executive functions by utilizing materials that are particularly suited to facilitate the development of accurate mental representations of numbers. Initially lacking a precise internal representation of quantities, children benefit from external representations provided by the program that vividly illustrate numbers and their relationships. Thus, the QNL training puts a special emphasis on ensuring that external representations make the numerical aspect visible in isolation, i.e., a clear focus is set on quantity as a particular feature of objects (for an example, see Figure 2). Any "seductive details" or narrative contexts that might distract children's attention from the numerical focus are avoided (e.g., Mayer et al., 2001). This approach not only alleviates the burden on working memory but also diminishes the demands on executive functions, such as discriminating and inhibiting irrelevant information.

1.7 Empirical evidence for the effectiveness of the QNL training

1.7.1 Evidence from studies with unselected kindergarten children

In a pilot study (Krajewski et al., 2008) the first version of the QNL program by Krajewski et al. (2007), was evaluated with a sample of 260 five-to six-year-olds, assigned to four groups. One year before school entry, a first group of children was trained by kindergarten teachers using the *QNL program*, while a second group received training focused on *inductive reasoning* (Klauer, 1989) and the third group participated in another mathematical training program that was conducted as standard in their kindergarten (*Zahlenland* [Numberland]; Friedrich and de Galgóczy, 2004). In contrast to the QNL training, this program is based on the concept of introducing numbers through stories, personifying them as characters (i.e., introducing the number 2 as a swan). A fourth group served as a control group that took part in the regular kindergarten program and did not receive any specific training. An immediate posttest and a seven-month follow-up measure in kindergarten, respectively, displayed significant short-and long-term effects in quantity–number competencies for the QNL group as compared to all other groups. Further, these effects were limited to numerical performance; the four groups did not differ in their developmental growth of phonological awareness, which reflects that the QNL training specifically facilitated mathematical development. Next, at the end of the first grade (19 months after training), a second follow-up was conducted to evaluate long-term effects on mathematical school achievement: the QNL group performed better in basic arithmetic than the students who had experienced the alternative mathematical training, where numbers were introduced embedded in a narrative context. That is, there were significant long-term transfer effects of QNL training as compared to an alternative math training method. Unexpectedly, however, these effects were not confirmed in comparison to the control group.

A study conducted by Hauser et al. (2014) involving 329 kindergarten children also incorporated the QNL training. Unfortunately, the training was not carried out as outlined in the manual; for example, the recommended group size of a maximum of

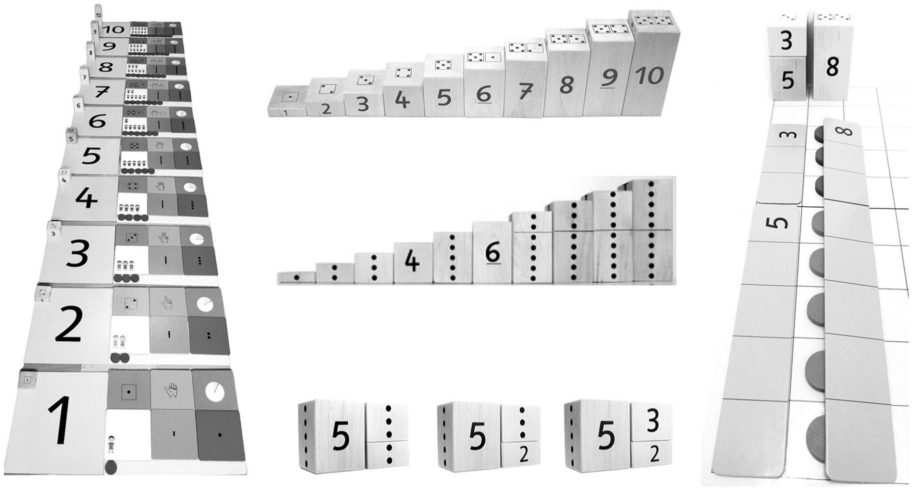

FIGURE 2
Examples of training materials from the "Mengen, zählen, Zahlen" program for teaching quantity-to-number word linkage (QNL training); "Numerical Street" on the left and "Number Staircase" on the right (Krajewski et al., 2007).

six kindergarten children was substantially exceeded, with up to 24 children per group. Deviations of this magnitude are very likely to compromise the intended quantity and quality of teacher-child interaction. Consequently, the study could not replicate the short-term effects of QNL training compared to another training. However, given the inadequate training fidelity, it does not allow for valid conclusions about the program's effectiveness.

1.7.2 Evidence from studies with at-risk kindergarten children

Renner et al. (2024) conducted a study to specifically evaluate the QNL program as a secondary preventive measure using a cohort of 567 preschoolers. Among them, 190 were identified as at-risk for developmental dyscalculia based on their performance in a test of quantity number competencies. The participants were divided into three distinct groups: the first group received QNL training, the second participated in the previously mentioned "Zahlenland" program (Numberland, NL) (Friedrich and de Galgóczy, 2004), and the third group followed the standard preschool curriculum. The study found that, both short-term and long-term, children in the QNL group showed significantly more improvement in quantity-number skills compared to those in the NL program and the control group (CG). Notably, 1 year after the training, at the end of 1st grade, the QNL group demonstrated enduring long-term transfer effects on arithmetic comprehension (compared to NL, $d = 0.90$ and to CG, $d = 0.58$). However, these effects did not extend to a standardized test that was aligned with the first-grade curriculum.

1.7.3 Evidence from studies with children deferred from school entry

Ennemoser (2010) evaluated the effectiveness of the QNL program in a classroom setting with children at risk who had been deferred from regular school entrance. While the control group received the regular math curriculum, the training group's first lessons of the school year were replaced by the QNL training. Results showed short-term effects for the QNL program in quantity–number competencies (QNC) compared to the control group. There were large effects on QNC level 2 ($d = 1.29$) and QNC level 3 ($d = 0.94$); these effects were already observable after only half of the program had been conducted. Transfer and long-term effects were not investigated in this study.

Hasselhorn and Linke-Hasselhorn (2013) also applied the QNL program to at-risk children who had been deferred from regular school attendance, but in this study, children were instructed in a small group setting. One group of children was trained with QNL, while the other served as a waiting control group. At posttest, the QNL group was superior to the control group in mathematical thinking, with a large effect size ($d = 4.6$). Although the subsequent training of the waiting control group constituted only a shortened version of the QNL training, large effect sizes were observed here as well ($d = 2.5$). However, this study had a very small sample size ($n = 8$), and neither transfer effects nor long-term effects were investigated. Furthermore, because no additional treatment was provided in the control condition, effects might not conclusively be attributed to the particular contents of the training.

1.7.4 Evidence from studies on children with special needs or intellectual disabilities

A further school study by Sinner and Kuhl (2010) was conducted in schools for children with special needs. Forty children were trained with 12 selected sessions of the QNL program, while another group received a training of inductive reasoning (*IR*) (Klauer, 1989). Compared to the control training the QNL training had a medium effect on quantity–number competencies ($d = 0.56$). In this study, both long-term effects and transfer effects on basic arithmetic were investigated. However, both analyses yielded insignificant results, indicating that training-induced advantages were not stable over time and did not transfer to general mathematics achievement.

Similar results were found in a study by Kuhl et al. (2012) with two groups of students with intellectual disabilities. While one group received a shortened version of the QNL training (in which level-3 exercises were omitted), the other group participated in language training based on the principles of the dialogic reading approach (Whitehurst et al., 1988). Compared to the language training, the QNL training led to higher gains in the student's quantity–number competencies (d = 0.44). However, this effect was only observable immediately after training, and did not persist until follow-up. Transfer effects on arithmetic were not investigated in this study.

1.7.5 Evidence from studies with first grade at-risk children

The first QNL intervention study in a regular school setting was conducted by Ennemoser and Krajewski (2007). They compared two groups of first-graders with mathematical difficulties, who received either the QNL program or a reading training. The QNL training was applied in the last quarter of the first school year, which is substantially later than in the other studies reported above. A curriculum-based mathematics test was administered at pre-and posttest in order to investigate training effects. The results confirmed medium-to-large effects of the QNL training compared to the reading training. Superior achievement gains were evident in the total math curriculum test score (d = 0.58) and in subtests on part–whole relationships (d = 0.79) and word problems (d =0.85). These results indicate that QNL training with low-achieving first-graders at the end of first grade has short-term transfer effects on mathematical school achievement. Again, long-term effects were not investigated.

In another study, Ennemoser et al. (2015) conducted a study with 238 first-graders, of whom they identified 64 as being at risk of developing mathematical difficulties. Of these children, 32 were assigned to the training group and participated in 10 selected sessions of the QNL training. A second group of 32 children served as a control group, receiving remedial mathematical instruction provided by the school. The QNL training group outperformed the control group in the quantity–number competencies that were focused on. Both short-and long-term effects (3 months after training) were observed, with effect sizes in the medium range (d=0.64 and d=0.69, respectively). While the transfer effect on arithmetic was not significant immediately after training, the QNL group subsequently displayed significantly larger improvements in basic arithmetic than the control group (d=0.52). This time-lagged effect indicated that the students' progress in quantity–number competencies as a result of QNL might have continued to facilitate more efficient learning during regular mathematics instruction in school. Unfortunately, no comprehensive curriculum-based test of mathematics achievement was applied, so this interpretation of the transfer effect remains restricted to the children's performance in basic arithmetic.

1.7.6 Evidence from studies in a regular first-grade classroom setting

Finally, Olyai et al. (2014) examined the effectiveness of QNL training implemented in a regular classroom setting in school. A total of 842 first-graders were assigned to one three groups. One group received QNL training, another group underwent self-regulation (SR) training, and a third group received a combination of QNL and self-regulation training (QNL + SR). A fourth group received regular mathematics instruction and acted as the control group (CG). The QNL group displayed significantly larger short-and long-term improvements in basic arithmetic compared to the SR group and the control group. Similar gains were reported for the combined training (QNL + SR), but no evidence was found for an added value of the additional self-regulation component. The study also investigated transfer effects to a standardized, curriculum-based mathematics test and found no significant differences between the four groups on this measure.

1.7.7 Summary

All prior studies that implemented the QNL training with fidelity consistently demonstrate significantly greater improvements in quantity-number competencies than the control groups. Most of these studies included active control groups that had received alternative treatments. These results constitute a growing body of evidence that QNL training can effectively stimulate the numerical development of children and is a suitable means of prevention of math difficulties. Particularly encouraging are the follow-up assessment results of several studies with low-achieving first-graders in regular school settings. These studies demonstrate that there is hope for *long-term* training effects, both in the quantity–number competencies that are actually the focus of training and, importantly, also in basic arithmetic, which can be considered a transfer effect. Nevertheless, there is currently insufficient evidence to confirm long-term transfer effects on (curriculum-based) mathematics achievement. Such confirmation is crucial, serving both as a primary criterion for evaluating the effectiveness of a prevention program and as a key validation of the QNL model's theoretical assumptions. It would specifically substantiate the idea that mathematical difficulties are rooted in a developmental delay in QNC, which is remediable. In other words, bridging this developmental gap in QNC can significantly enhance the subsequent acquisition of more advanced mathematical skills, beyond foundational numerical competencies.

1.8 Aims of the study and research questions

As outlined above, quantity–number competencies provide a promising starting point for the prevention of developmental dyscalculia. However, although previous studies on various trainings report on positive results, conclusive empirical evidence for the preventive potential of numerical training is still lacking. First, a large proportion of the interventions investigated by these studies do not constitute comprehensive training programs that trace the developmental sequence in which mathematical abilities are acquired; instead, they focus on particular facets or task formats. Second, long-term effects and—most importantly—transfer effects on subsequent mathematical school achievement are scarcely addressed in the available studies. Third, many studies fail to meet basic methodological challenges (e.g., inclusion of appropriate control conditions) and do not allow reliable conclusions about the effectiveness of the particular intervention to be drawn.

Against this background, the goal of the present study was to evaluate the effectiveness of a small group prevention program for children at risk of developmental dyscalculia. Our main research question was, "Does a training program aligned with the developmental sequence of mathematical competencies, as outlined in the QNL model, effectively prevent math difficulties?" In line with the assumptions of the model, we expected that the training would yield both significant short-and long-term effects specific to the training focus (i.e., on QNC) as well as transfer effects on subsequent mathematical school achievement. Further, we addressed the question of whether transfer effects on school achievement, if observable, are mediated by preceding gains in quantity–number competencies. According to the prerequisite role of these competencies for subsequent success in the mathematical school curriculum, this mediation effect was to be expected. To address the methodological weaknesses of many prior studies, potential achievement gains from sources other than QNL training (e.g., those due to attention effects, repeated testing effects, or natural development) were controlled for by the inclusion of two appropriate control conditions.

2 Materials and methods

2.1 Participants and design of the study

A total of 575 first-grade students (mean age 6;11 years) from 14 primary schools participated in a standardized test of quantity–number competencies (MBK 1+; see below), 4 months after school entry. The assessments were conducted by trained student test administrators following a rigorously standardized procedure to ensure that the testing conditions were consistent for all children. Based on their performance in the MBK 1+, 119 of the children were identified as being at risk of math difficulties (below 20th percentile) and assigned to (quasi-)experimental conditions. Initially, a randomized assignment of participants was planned. However, during the randomization process, it became evident that diffusion effects were likely to occur, potentially undermining treatment validity (Cook and Campbell, 1979). Specifically, assigning different groups within the same school or neighboring schools, which sometimes cooperate closely, to different conditions could lead to teachers sharing information and adjusting their methods. To avoid these issues and maintain treatment validity, we opted for spatial assignment instead of randomization, ensuring that groups assigned to different conditions were not from the same or closely collaborating schools. Half of the at-risk sample (the QNL group, $n = 61$; 34 girls and 27 boys) received the training in quantity-to-number-word linkage, as described above. The other half was assigned to the two control groups, one of which ($n = 30$; 15 girls and 15 boys) participated in training in inductive reasoning (Klauer, 1989; see also Klauer and Phye, 2008; this was the IR group), while the remaining group did not receive any additional training (the control group, CG, $n = 28$; 16 girls and 12 boys). All trainings were conducted in a small-group setting with 3 to 7 children.

As shown in Figure 3, four measurement points were set: in addition to the pre-and posttest (T1, T2), two follow-up sessions were included in order to assess long-term effects. These two additional measurement points were, respectively, 6 months (T3) and 15 months after training (T4; end of second grade). Quantity–number competencies were measured from T1 until T3; mathematical achievement was assessed at all measurement points (T1 until T4).

Until T4 (i.e., 15 months after training) eight children had moved to another town or school, respectively, and further four children did not participate in the final follow-up session because of illness. This constitutes an attrition rate of 10%. Six of these children dropped out of the QNL group, while each of the other two groups lost three participants.

2.2 Measures

2.2.1 Quantity–number competencies at T1, T2, and T3

As a measure of quantity–number competencies (as described under the QNL model), we used a standardized test (*Test mathematischer Basiskompetenzen in der ersten Klasse; MBK 1+*; Ennemoser et al., 2017), containing the following subtests. To tap *QNC level 1*, first, 10 number words were presented orally to the children, who had to write down the corresponding Arabic numerals (*number-word transcoding*). Next, children were given pairs of numerals (e.g., 11 and 13) and had to fill in the number missing between them (12) (number gaps). Four subtests were given to assess mastery of *level 2*. In these subtests children were required to link quantities to numbers (*numerical quantities*), to compare the numerical size of numbers (*number comparison*), to place numbers on number lines (*number lines*), and to fill in gaps in several series of numerical quantities (*number seriation*). Finally, in order to assess competencies on *level 3*, children were asked to determine which number is exactly one more (or one less, respectively) than a given number (*one more, one less*), to complete part–whole tasks where either one or two numbers were decomposed into two components (*decomposition, part–whole*) and, finally, to solve word problems presented as comparison problems (see Riley et al., 1983). Each correct solution was given 1 point, except for number-word transcoding (0.5 points). The maximum score on the QNC test was 49. Reliability was 0.90 (Cronbach's α).

2.2.2 Mathematical school achievement at T1, T2, T3, and T4

Since no standardized test for mathematical school achievement was available for the initial months of the school year, we utilized the supplemental *basic arithmetic test* included in the MBK 1+ (see above) to evaluate the children's basic arithmetic at T1. In 8 s, children had to solve as many addition tasks as possible, up to 10 (maximum score = 20, Cronbach's α = 0.92). At the end of first grade (T2) and the start of second grade (T3) the standardized German mathematics test *Deutscher Mathematiktest für erste Klassen* (*DEMAT 1+*; Krajewski et al., 2002), which is based on the primary school curriculum of all German federal states, was administered to the children (maximum score = 36, Cronbach's α = 0.82). Mathematical school performance at the end of second grade (T4) was assessed using the standardized *Heidelberger Rechentest für erste bis vierte Klassen* (*HRT 1–4*; Haffner et al., 2005). As per the test-developers, reliability is 0.93.

2.2.3 Nonverbal intelligence (IQ)

In order to control for effects of intellectual ability, the short version of the *Culture Fair Intelligence Test* (*CFT 1, Scale 1*; Cattell et al., 1997) was included as a pretest measure. The short version consists of three subtests, on classifications, similarities, and matrices, respectively, (maximum score = 36; Cronbach's α = 0.90).

2.3 Interventions

2.3.1 Training of quantity-to-number-word linkage

The QNL training program *Mengen, zählen, Zahlen* (Krajewski et al., 2007) was used to foster *numerical development* in the first

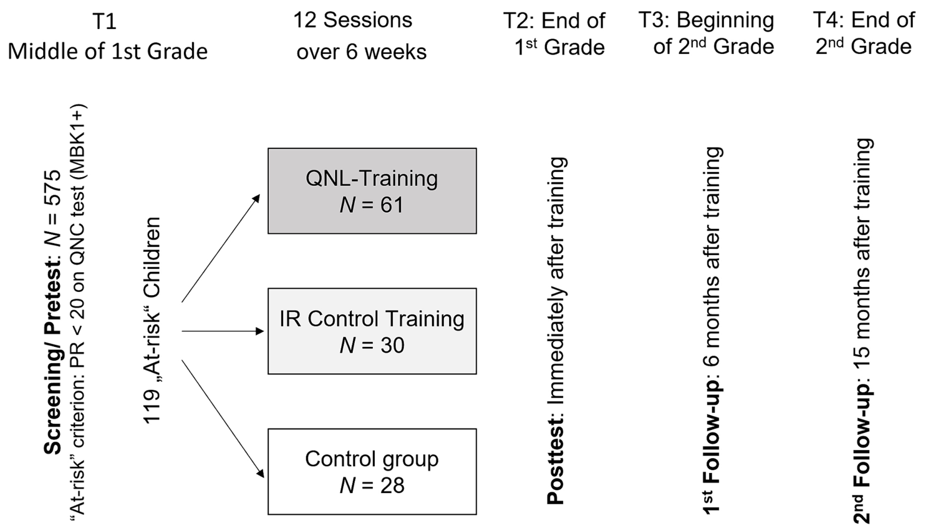

FIGURE 3
Design of the study.

training group. The training was originally designed for kindergartners and, thus, had to be tailored for a school setting. The original program's 24 sessions, each lasting 30 min, were condensed to twelve 45-min sessions. This was achieved by omitting five sessions focused on the introduction of the numbers one to ten and by reducing repetition exercises. The remaining sessions especially focused on number seriation and number relations (level-2 and level-3 competencies; for a detailed description see above). As a supplement to the kindergarten version of the training, additional worksheets were introduced in order to facilitate transfer of the trained level-2 and level-3 competencies to pictorial and symbolic task formats. The adapted version of the QNL training was administered over the course of 6 weeks and was conducted in groups of three to seven children. In order to strengthen treatment fidelity trainers received a four-hour training and were provided with a highly standardized manual that specifies every single session in detail (aims, materials, and relevant verbal instructions). All learning materials necessary for the training were provided as well. The delivery of sessions was recorded by the teachers. According to these records all 12 sessions were held as specified in the manual. For organizational reasons, however, in some of the groups more than 6 weeks (up to 11) were necessary to deliver the 12 training sessions.

2.3.2 Inductive reasoning training

A second group of children received a training in *inductive reasoning* (*IR*; *Denktraining für Kinder 1;* Klauer, 1989). Children are encouraged to apply systematic comparison in order to identify differences and similarities (with respect to attributes and relations, respectively) and consequently to induce regularities. Inductive reasoning is hypothesized to play a crucial role in problem-solving, and there is supportive evidence that IR training has significant and long-lasting effects on non-verbal IQ and academic learning (e.g., Klauer and Phye, 2008). The training used here consisted of 120 visually presented tasks conducted in groups of five to seven children over the course of 12 sessions held parallel to those in the QNL training.

2.3.3 Control group

The third group only participated in regular mathematics instruction provided by school and did not receive additional treatment during the six-week period of intervention.

2.4 Missing data

The analysis of missing data revealed that the percentages of missing values were low, with values across the first three measurement points ranging from 0.8% (IQ at T1) to 6.7% (math achievement at T3). The highest percentage of missing data occurred for math achievement at T4, with 10.1% missing due to cumulative relocations over the course of the study and student absences caused by illness at the last measurement point. To determine if the missing data were missing completely at random (MCAR), Little's MCAR test was conducted. Results indicate that it is reasonable to assume that the values are completely missing at random ($\chi^2 = 52.404$, $df = 46$, $p = 0.24$). Multiple imputation was employed to address the missing data, resulting in 5 imputed datasets. The final analyses were conducted on the pooled imputed datasets.

2.5 Statistical analysis

The analyses were conducted using SPSS Version 29.0 (IBM Corp, 2023a) and AMOS Version 29.0 (IBM Corp, 2023b). Descriptive statistics for the variables were calculated for all data, and pretest differences were examined using analyses of variance (ANOVAs).

To account for the clustered data structure, with children trained in small groups, we employed multilevel modeling (MLM). Traditional linear regression assumes that all observations are independent, but this assumption is violated in our study due to the group-based training. For example, children within the same small group might perform similarly because they share the same instructor and training

TABLE 2 Descriptive statistics for all variables and all measurement points (T1–T4).

	Condition					
	QNL training (QNL)		Inductive reasoning (IR)		Control group (CG)	
	M	*SD*	*M*	*SD*	*M*	*SD*
Quantity-number competencies (QNC)						
T1: *MBK 1+*	23.91	6.28	24.27	5.53	24.63	5.07
T2: *MBK 1+*	37.90	6.30	28.60	6.89	30.42	7.22
T3: *MBK 1+*	42.18	5.96	35.79	6.66	33.315	5.72
Mathematics achievement						
T1: *Basic arithmetic*	10.28	4.68	14.73	5.66	11.61	4.41
T2: *DEMAT 1+*	20.77	6.72	20.10	6.94	20.41	5.89
T3: *DEMAT 1+*	25.68	6.35	20.42	7.08	20.61	5.85
T4: *HRT 1–4*	45.69	4.84	44.34	6.25	42.50	4.34
Intelligence						
T1: *CFT 1*	18.42	5.43	19.60	4.74	17.82	4.14

*$p<0.05$. **$p<0.01$.

environment. Ignoring this grouping could lead to underestimated standard errors and biased effect sizes (Hox, 2010).

2.5.1 Model specification

As previously mentioned in chapter 2.1, all at-risk students were organized into small training groups of three to seven children, which were assigned to either the QNL intervention, the IR training, or the CG. A two-level nested data structure was assumed, with students as level-1 units ($n = 119$) and training groups as level-2 units ($n = 22$).

Separate multilevel analyses were conducted to examine training effects on different outcome measures: effects on the trained quantity-number competencies (QNC) immediately after training (short-term effects) and 6 months after training (long-term effects), as well as transfer effects on mathematical school achievement at three time points: immediately after training (T2), 6 months after training (T3), and 15 months after training (T4).

Pretest QNC scores and IQ were included as covariates to control for initial levels of QNC and differences in intellectual ability when analyzing the effects on the trained competencies. For transfer effects on mathematical achievement, IQ and children's early mathematics performance (basic arithmetic) at T1 served as covariates to control for differences in intellectual ability and initial achievement level.

We attempted to fit four models with progressively increasing complexity for each outcome variable using maximum likelihood estimation. However, the fourth model, which included random slopes, did not converge (see below). Therefore, we compared the fit of the first three models using log-likelihood ratio tests and report the results of the most appropriate model. The models were specified as follows:

2.5.1.1 Model 1: null model (unconditional model)

This model did not include any explanatory variables and only the random intercept to partition the variance between groups. It served as a baseline to estimate and compare subsequent models. The intraclass correlation coefficient (ICC) was calculated to determine the proportion of variance in the outcome variable attributable to differences between groups (level-2 units).

An ICC greater than 0.10 is a commonly used as a threshold indicating that a significant portion of the variance in the outcome variable is due to differences between groups (Snijders and Bosker, 2012), which suggests that MLM is appropriate.

2.5.1.2 Model 2: fixed effects model

This model included the above-mentioned explanatory variables (covariates and quasi-experimental condition) as fixed effects to assess their impact on the respective outcome variable.

2.5.1.3 Model 3: random intercept model

This model included explanatory variables as fixed effects and additionally allowed the intercept to vary across different groups, acknowledging that each group might have a different baseline regarding the outcome variable.

2.5.1.4 Model 4: random slopes model

This model allowed both the intercept and slopes (coefficients of predictor variables) to vary across groups, capturing potential differences in how predictor variables influence the outcome within different groups. However, as mentioned above, models with random slopes did not converge. Several attempts to resolve the convergence issues were made, including adjusting initial values and modifying convergence criteria, scaling predictor variables and simplifying the model, using other statistical software programs like R Version 4.40 (R Core Team, 2024) and Mplus Version 8.11 (Muthén and Muthén, 2024),[2] but the problems persisted. Therefore, the estimates of random slopes models are not reported for any outcome variable. Instead, we report the results from models with random intercepts, which adequately capture the nested structure of the data and control for potential intra-group correlation.

3 Results

3.1 Descriptive statistics and preliminary analysis

Descriptive statistics for all variables and measurement points are shown in Table 2. Preliminary analyses of variance (ANOVAs) revealed no significant differences in IQ or quantity–number competencies (QNC) between the four groups at the outset of the study ($ps > 0.05$). However, there was a significant effect of group with regard to basic arithmetic [$F(2, 113) = 7.38$; $p < 0.01$]. *Post-hoc* tests confirmed that this effect was due to the IR group, which showed superior basic arithmetic compared to the two other groups ($p < 0.05$).

The correlations between pretest, posttest, and follow-up performances in QNC and curricular mathematics performances (DEMAT, HRT) range from $r = 0.29$ to $r = 0.62$ (see Table 3). These values are within the expected range for selected at-risk samples,

2 Although different software packages typically yield comparable results, they differ substantially in terms of convergence rates (McCoach et al., 2018).

 10.3389/fpsyg.2024.1380036

likely being somewhat reduced due to the limited variance in this group's abilities and the training provided between assessments. As expected, the correlations of QNC with basic arithmetic are lower and not always significant.

3.2 Specific training effects on quantity–number competencies

The results of the MLM analyses for specific short-and long-term training effects on quantity–number competencies are summarized in Table 4.

3.2.1 Short-term effects

Likelihood-Ratio tests indicated that both the fixed effects model and the random intercept model provided a significantly better fit than the null model (*p*s < 0.001) for the short-term effects at T2. The ICC for the null model was 0.34, indicating that the use of MLM was appropriate.

However, the random intercept model did not significantly improve the fit over the fixed effects model ($\chi^2 = 1.75$, $df = 1$, $p = 0.185$). Therefore, the parameter estimates of the fixed effects model are reported.

The results of the MLM revealed significant effects for the QNC pretest score ($b = 0.43$, $SE = 0.09$, 95% CI [0.26, 0.60], $p < 0.001$) and for IQ ($b = 0.49$, $SE = 0.10$, 95% CI [0.29, 0.69], $p < 0.001$). Furthermore, the QNL training group yielded significantly higher gains in quantity–number competencies from pretest to posttest than the control group ($b = 7.49$, $SE = 1.19$, 95% CI [5.16, 9.82], $p < 0.001$) and the inductive reasoning group ($b = 10.04$, $SE = 1.17$, 95% CI [7.76, 12.33], $p < 0.001$). Large effect sizes were observed, amounting to more than one standard deviation in favor of the QNL training (QNL vs. CG: $d = 0.82$, 95% CI [0.61, 1.02]; QNL vs. IR group: $d = 1.12$, 95% CI [0.89, 1.35]).

3.2.2 Long-term effects 6 months after training

For QNC-specific long-term effects, the parameter estimates of the fixed effects model are reported, since the model fit significantly surpassed the null model ($\chi^2 = 55.08$, $df = 3$, $p < 0.001$), and the addition of a random intercept did not significantly improve the model fit ($\chi^2 = 0.59$, $df = 1$, $p = 0.443$).

The analysis indicated that the observed advantage of the QNL training group was persistent over time. Six months after training (T3), the QNL group still significantly outperformed both the inductive reasoning group ($b = 6.60$, $SE = 1.22$, 95% CI [6.07, 11.40], $p < 0.001$) and the control group ($b = 8.75$, SE = 1.26, 95% CI [6.28, 11.21], $p < 0.001$). Effect sizes were in a medium to large range, between $d = 0.69$, 95% CI [0.43, 0.95] (QNL vs. IR) and $d = 0.90$, 95% CI [0.63, 1.16] (QNL vs. CG).

3.3 Transfer effects on mathematical school achievement

The results of the MLM analyses for short-and long-term transfer effects on mathematical school achievement are summarized in Table 5.

3.3.1 Short-term transfer effects immediately after training

Since the fixed effects model again provided a better fit than the null model ($\chi^2 = 30.57$, $df = 3$, $p < 0.001$) and the addition of a random intercept did not significantly improve the fit ($\chi^2 = 0.05$, $df = 1$, $p = 0.823$), the estimated parameters of the fixed effects model are reported for short-term transfer effects at T2.

The analysis of short-term effects on mathematical school achievement, assessed by the *DEMAT 1+* immediately after training, did not parallel the findings for specific effects on quantity–number competencies. Significant effects were found only for the two covariates (basic arithmetic: $b = 0.51$, $SE = 0.12$, 95% CI [0.27, 0.74], $p < 0.001$; IQ: $b = 0.28$, $SE = 0.11$, 95% CI [0.06, 0.50], $p = 0.014$), whereas the QNL group did not

TABLE 3 Correlations among pretest, posttest, and follow-up scores for the at-risk children.

			QNC			Basic arithmetic			Math achievement		
			T1	T2	T3	T1	T2	T3	T2	T3	T4
QNC											
MBK1+	T1		–								
MBK1+	T2		0.40**	–							
MBK1+	T3		0.32**	0.56**	–						
Basic arithmetic											
	T1		0.22*	0.14	0.11	–					
	T2		0.29**	0.28**	0.34**	0.37**	–				
	T3		0.15	0.21*	0.26**	0.35**	0.42**	–			
Math achievement											
DEMAT 1+	T2		0.29**	0.40**	0.36**	0.34**	0.43**	0.37**	–		
DEMAT 1+	T3		0.49**	0.55**	0.62**	0.16	0.48**	0.36**	0.53**	–	
HRT 1–4	T4		0.34**	0.44**	0.54**	0.41**	0.49**	0.54**	0.45**	0.69**	–
IQ (CFT)	T1		0.33**	0.39**	0.38**	0.31**	0.16	0.17	0.29**	0.34**	0.37*

*$p < 0.05$. **$p < 0.01$.

TABLE 4 Multilevel analysis on specific training effects on QNC.

	QNC short-term effects			QNC long-term effects		
	Null model	Fixed effects model	Random intercept model	Null model	Fixed effects model	Random intercept model
Intercept	33.80** (1.14) [31.56, 36.04]	18.61** (2.39) [13.93, 23.28]	18.36** (2.42) [13.61, 23.11]	37.86** (1.07) [35.76, 39.96]	27.21** (2.49) [22.33, 32.08]	27.17** (2.52) [22.23, 32.11]
QNL training vs. control group		7.49** (1.19) [5.16, 9.82]	7.54** (1.43) [4.75, 10.34]		8.75** (1.26) [6.28, 11.21]	8.74** (1.36) [6.07, 11.40]
		$d = 0.82$	$d = 0.69$		$d = 0.90$	$d = 0.83$
QNL training vs. IR training		10.04** (1.17) [7.76, 12.33]	9.95** (1.44) [7.14, 12.76]		6.60** (1.22) [4.21, 8.99]	6.57** (1.34) [3.94, 9.20]
		$d = 1.12$	$d = 0.90$		$d = 0.69$	$d = 0.63$
Pretest-score		0.43** (0.09) [0.26, 0.60]	0.45** (0.09) [0.28, 0.63]		0.22* (0.10) [0.03, 0.41]	0.22* (0.10) [0.03, 0.41]
IQ (CFT)		0.49** (0.10) [0.29, 0.69]	0.48** (0.10) [0.28, 0.68]		0.51** (0.11) [0.29, 0.72]	0.51** (0.11) [0.30, 0.72]
df	3	6	7	3	6	7
Pseudo R^2 (marginal)	0	0.56 (0.00)	0.56 (0.00)	0	0.46 (0.01)	0.46 (0.01)
Pseudo R^2 (adjusted)	0.34	0.56 (0.00)	0.61 (0.00)	0.32	0.46 (0.01)	0.48 (0.02)
−2 Log-likelihood	806.02	729.88 (0.17)	728.13 (0.09)	793.57 (0.41)	738.49 (1.22)	737.90 (1.34)
BIC	820.36	758.56 (0.17)	761.58 (0.10)	807.91 (0.41)	767.88 (1.39)	771.35 (1.34)
ICC	0.34			0.32 (0.01)		

Values in parentheses are standard errors; values in brackets are 95% confidence intervals; BIC, Bayesian information criterion; ICC, Intra-class correlation coefficient; $*p < 0.05$. $**p < 0.01$.

significantly differ from the other two groups (QNL vs. CG: $b = 0.52$, $SE = 1.33$, 95% CI [2.10, 3.13], $p = 0.698$; QNL vs. IR: $b = 2.54$, $SE = 1.45$, 95% [0.30, 5.39], $p = 0.08$). This result parallels previous findings indicating that transfer effects may require some time and is consistent with the expectation that, immediately after training (T2), children in the QNL group may not yet be able to transfer their improved quantity–number competencies to typical task formats used in regular mathematics instruction (Ennemoser et al., 2015).

3.3.2 Transfer effects 6 months after training

The results of the likelihood-ratio test indicated that both the fixed effects model and the random intercept model significantly improved the fit over the null model ($ps < 0.001$). Furthermore, the random intercept model also significantly improved the fit over the fixed effects model ($\chi 2 = 4.35$, $df = 1$, $p = 0.037$). Consequently, the estimated parameters of the random intercept model are reported.

Six months after training (T3), the effect of IQ was still significant ($b = 0.42$, $SE = 0.11$, 95% CI [0.21, 0.64], $p < 0.001$), while the effect for basic arithmetic was not ($b = 0.19$, $SE = 0.12$, 95% CI [0.05, 0.42], $p = 0.101$). However, in contrast to the result of the analysis of short-term effects, the QNL group showed superior mathematics performance to both the inductive reasoning group ($b = 5.86$, $SE = 1.98$, 95% CI [1.95, 9.77], $p < 0.001$, $d = 0.39$, 95% CI [0.14, 0.65]) and the control group ($b = 4.65$, $SE = 1.80$, 95% CI [1.12, 8.17], $p < 0.001$, $d = 0.33$, 95% CI [0.08, 0.59]), which indicates a significant time-lagged transfer effect of QNL training on children's mathematical achievement as compared to the two control groups.

3.3.3 Transfer effects 15 months after training

For long-term transfer effects, again both the fixed effects model and the random intercept model provided a significantly better fit than the null model ($ps < 0.001$). However, the random intercept model did not further improve the fit over the fixed effects model ($\chi^2 = 3.18$, $df = 1$, $p = 0.074$). Therefore, the estimates of the fixed effects model are reported.

Fifteen months after training (T4), significant effects were observed for the two covariates (basic arithmetic: $b = 0.41$, $SE = 0.09$, 95% CI [0.24, 0.58], $p < 0.001$; IQ: $b = 0.31$, $SE = 0.09$, 95% CI [0.14, 0.48], $p < 0.001$). Children who participated in the QNL training still significantly outperformed the two other groups 15 months after training (QNL vs. CG: $b = 2.74$, $SE = 1.00$, 95% CI [0.79, 4.70], $p = 0.006$; QNL vs. IR: $b = 2.77$, $SE = 1.15$, 95% CI [0.50, 5.05], $p = 0.017$). Effect sizes were in a small to medium range (IR training: $d = 0.32$ [0.06, 0.58]; control group: $d = 0.36$ [0.10, 0.61]).

3.4 Mediation model

A path model was specified to test the hypothesis that transfer effects on mathematical school achievement are mediated by preceding (specific) effects on the development of quantity–number competencies (see Figure 4). The dependent variable was mathematical achievement at the end of the second grade. To control for differences in initial performance, basic arithmetic at the first measurement point (T1) was included as a covariate. Quantity–number competencies (QNC) at posttest were included as the mediator, adjusted for the initial QNC levels at T1, thus representing the change in QNC during the intervention phase.

Although we expected the mediation effect to hold for both control groups, we analyzed the model separately for each control

TABLE 5 Multilevel analysis on transfer effects on mathematics achievement in school.

	Math achievement T2 (DEMAT 1+)			Math achievement T3 (DEMAT 1+)			Math achievement T4 (HRT1-4)		
	Null model	Fixed effects model	Random intercept model	Null model	Fixed effects model	Random intercept model	Null model	Fixed effects model	Random intercept model
Intercept	20.29** (0.67) [18.97, 21.61]	9.91** (2.19) [5.62, 14.20]	9.90** (2.20) [5.59, 14.21]	23.04** (0.98) [21.13, 24.96]	15.10**(2.20) [10.78, 19.41]	15.63**(2.26) [11.20, 20.06]	44.45** (0.61) [43.26, 45.64]	35.18** (1.67) [31.89, 38.46]	35.12** (1.71) [31.77, 38.47]
QNL vs. control group		0.52 (1.33) [2.10, 3.13]	0.52 (1.37) [2.16, 3.20]		4.74** (1.39) [2.02, 7.46]	4.65* (1.80) [1.12, 8.17]		2.74** (1.0) [0.79, 4.70]	2.67* (1.18) [0.36, 4.98]
		$d=0.05$	$d=0.05$		$d=0.43$	$d=0.33$		$d=0.36$	$d=0.29$
QNL vs. IR training		2.54 (1.45) [0.30, 5.39]	2.54 (1.48) [0.37, 5.45]		6.11** (1.55) [3.05, 9.18]	5.86** (1.98) [1.95, 9.77]		2.77* (1.15) [0.50, 5.05]	2.69* (1.32) [0.08, 5.30]
		$d=0.24$	$d=0.23$		$d=0.54$	d=0.39		$d=0.32$	$d=0.27$
Basic arithmetic		0.51** (0.12) [0.27, 0.74]	0.50** (0.12) [0.26, 0.74]		0.19 (0.12) [0.04, 0.43]	0.19 (0.12) [0.05, 0.42]		0.41** (0.09) [0.24, 0.58]	0.42** (0.09) [0.25, 0.59]
IQ (CFT)		0.28* (0.11) [0.06, 0.50]	0.28* (0.11) [0.06, 0.51]		0.45** (0.11) [0.23, 0.67]	0.42** (0.11) [0.21, 0.64]		0.31** (0.09) [0.14, 0.48]	0.31** (0.09) [0.14, 0.48]
df	3	6	7	3	6	7	3	6	7
Pseudo R^2 (marginal)	0	0.23 (0.02)	0.23 (0.02)	0	0.28 (0.04)	0.26 (0.04)	0	0.33 (0.01)	0.34 (0.01)
Pseudo R^2 (adjusted)	0.05 (0.01)	0.23 (0.02)	0.24 (0.01)	0.30 (0.02)	0.28 (0.04)	0.39 (0.05)	0.11 (0.02)	0.33 (0.01)	0.40 (0.01)
−2 Log-Likelihood	785.96 (1.18)	755.39 (2.40)	755.34 (2.38)	780.70 (1.88)	757.81 (2.16)	753.46 (2.49)	728.39 (2.26)	684.67 (2.68)	681.49 (2.72)
BIC	800.30 (1.19)	784.06 (2.41)	788.79 (2.38)	794.23 (1.88)	786.01 (2.17)	786.91 (2.50)	743.53 (2.26)	712.94 (2.46)	715.35 (2.49)
ICC	0.05 (0.01)			0.30 (0.02)			0.11 (0.02)		

Values in parentheses are standard errors; values in brackets are 95% confidence intervals; BIC, Bayesian information criterion; ICC, Intra-class correlation coefficient. $*p<0.05$. $**p<0.01$.

group to ensure we did not miss potential differences: one model comparing the QNL training group with the control group, and another model comparing the QNL training group with the reasoning group. As can be seen in Figure 4, the relevant coefficients for the two models are presented one below the other for easy comparison (QNL vs. control group at the top, QNL vs. inductive reasoning group at the bottom). The results were closely parallel, indicating that the effects were consistent across both comparisons.

First of all, both analyses confirm that there are significant effects on QNC in favor of the QNL group (Model 1: $\beta=0.56$; $p<0.001$, Model 2: $\beta=0.44$; $p<0.001$). These findings are consistent with the training effects observed in the multilevel analyses above, indicating that the results are robust across different analytical approaches. Results further support the assumption of an *indirect training effect* on mathematical school achievement. Higher gains in QNC during the intervention phase are subsequently related to better mathematical school achievement 15 months after training (Model 1: $\beta=0.47$; $p<0.001$, Model 2: $\beta=0.48$; $p<0.001$). Controlling for this indirect (mediation) effect, the direct effect of QNL training on mathematical achievement was not significant in either model (Model 1: $\beta=0.06$; p, Model 2: $\beta=0.03$; $p=0.73$).[3] Both models explained similar proportions of the variance in mathematics achievement at the end of second grade (Model 1: 42%, Model 2: 43%).

4 Discussion

The model of quantity–to–number word linkage (i.e., the QNL model) (Krajewski, 2008) traces the development of early numerical abilities. It identifies important milestones that have to be met over the course of development and, thus, indicates promising starting points for prevention. A large body of evidence demonstrates the predictive value of the competencies described by the QNL model. With regard to its preventive potential, however, the evidence available to date is promising but not conclusive. Most findings are limited to short-term effects on specifically trained skills, while particularly long-term transfer effects on subsequent mathematical school achievement are scarcely investigated. Moreover, the available studies usually did not include a comprehensive training program specifically designed to establish conceptual understanding in a step-by-step process in accordance with the acquisition of foundational skills hypothesized by developmental theory. Given this situation,

3 Preliminary path analyses without the mediator indicated significant direct transfer effects of QNL training on later mathematics achievement (Model 1: $\beta=0.32$; $p<0.001$, Model 2: $\beta=0.28$; $p=0.002$).

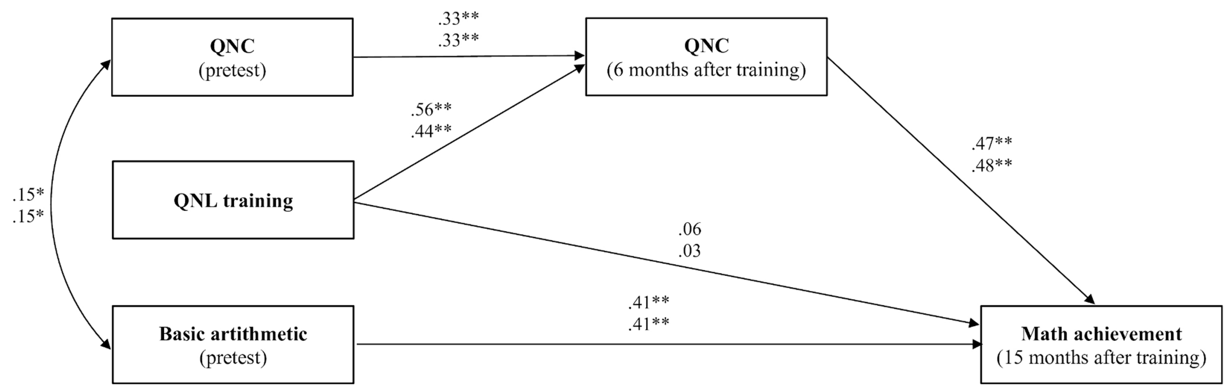

FIGURE 4
Mediation model: transfer effects of QNL training on mathematics achievement 15 months after intervention, mediated by previous (training-induced) gains in (upper values: QNL vs. control group; lower values: QNL vs. inductive reasoning group).

the goal of our study was to investigate the effectiveness of a strictly theory-based prevention program (QNL program) for first-grade children who are at risk of developing mathematical difficulties. We further aimed to investigate not only short-term but also long-term effects, 6 and 15 months after training. Both specific and transfer effects on subsequent mathematical school achievement were investigated; if both specific and transfer effects were observable, we intended also to test whether transfer effects on school achievement could be attributed to preceding (training-induced) gains in quantity–number competencies (that is, a mediation effect).

Our results align with and expand upon existing findings on the *effectiveness of QNL training*. Both short-and medium-term effects on quantity–number competencies were found up to 6 months later. These specific effects on foundational numerical insights, which were particularly addressed by the training, subsequently led to a long-term transfer effect on mathematical school achievement. Interestingly, this transfer effect was not yet evident immediately after training but grew significant 6 months later and remained stable over the course of the next 15 months. This time-lagged effect replicates the findings by Ennemoser et al. (2015) and is in accordance with the theoretical assumptions of the QNL model. It indicates that while the training successfully closed developmental gaps in quantity–number competencies, this increase did not immediately boost children's performance in a standardized mathematics test at posttest. Subsequently, however, this higher level of QNC facilitated children's understanding of what was taught during regular mathematics instruction in school.

Path analytical results supported our hypothesis that long-term transfer effects on mathematics school achievement are mediated by prior gains in quantity–number competencies, a finding that further adds to the existing evidence on the crucial role of quantity–number competencies as a prerequisite for subsequent mathematics achievement.

Our findings are particularly encouraging regarding their practical implications. The results were obtained in a regular school setting, with the teacher instruction requiring only 4 h, and the entire intervention consisting of just 12 sessions. Given this minimal investment, the long-term effects on mathematical school performance of at-risk children, observable even after 15 months, are quite remarkable. According to teacher feedback, the highly standardized approach makes the program easy to implement and follow, i.e., to maintain treatment fidelity. Deviations from the program guidelines were reported only in terms of the number of weekly sessions, resulting in some groups taking longer to complete the 12 training sessions than initially planned (up to 10 weeks instead of 6).

Considering the consistently positive findings on the effectiveness of QNL training, it may be valuable to evaluate the training in future studies not as a secondary preventive measure (targeting selected at-risk children), but as a primary preventive measure by implementing QNL instruction at the beginning of first grade, prior to the regular mathematics curriculum. This approach might be even more efficient in ensuring foundational skills in mathematics at the start of first grade, thereby reducing the number of children who lack the prerequisites to benefit from subsequent conventional mathematics instruction.

One limitation of this study is that it was not conducted as a randomized controlled trial (RCT), but rather as quasi-experimental research. The assignment of participants to the experimental conditions was partially based on non-random factors, particularly to avoid diffusion effects, which introduces potential selection bias. This approach, while valuable, may not offer the same level of robustness in controlling for confounding variables as an RCT. Additionally, the relatively small sample size is a constraint, as effects observed in smaller studies often tend to diminish in larger-scale research.

Another limitation is that the models with random slopes did not achieve convergence. Thus, it cannot be entirely ruled out that the parameter estimates might be biased due to unaccounted variability in the slopes across groups. The convergence issues may be due to the relatively small number of groups ($n = 22$) and number of students per group (three to seven), since insufficient sample sizes can lead to instability in estimating parameters (Maas and Hox, 2005; Hox, 2010). Given these limitations, however, it should be noted that significant short-and long-term training effects (including both specific and transfer effects) were also confirmed in the path analyses conducted to test the mediator hypothesis, demonstrating that the fixed effects remained robust across different analytical methods. Future studies with larger sample sizes and more level-2 units could provide more reliable and nuanced insights. Finally, we considered only a limited number of control variables in our study. Future research should also examine whether the observed intervention effects are equally effective for all children or whether they are moderated by third variables such as motivational factors or cognitive influences like working memory and language proficiency.

In sum, our findings provide evidence for the long-term effectiveness of a strictly theory-based training program for the early prevention of developmental dyscalculia. Notably, this success is achieved with remarkably little effort, both in terms of teacher instruction and the total number of sessions. Additionally, these encouraging results support the QNL model's view that dyscalculia is not due to an immutable defective number module. Instead, it supports the more optimistic notion of developmental dyscalculia as a developmental delay in foundational skills, which can be effectively addressed through targeted early prevention measures.

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 humans were approved by the Ethics Committee of the Ludwigsburg University of Education. The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation in this study was provided by the participants' legal guardians/next of kin.

Author contributions

ME: Conceptualization, Funding acquisition, Methodology, Writing – original draft, Writing – review & editing, Supervision, Formal analysis. DS: Investigation, Project administration, Formal analysis, Writing – review & editing. LN: Writing – review & editing. KK: Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This study was supported by the Research Network for Empirical Educational Research (Forschungsnetzwerk Empirische Bildungsforschung) at Justus Liebig University Gießen.

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.2024.1380036/full#supplementary-material

References

Aunola, K., Leskinen, E., Lerkkanen, M.-K., and Nurmi, J.-E. (2004). Developmental dynamics of mathematical performance from preschool to grade 2. *J. Educ. Psychol.* 96, 699–713. doi: 10.1037/0022-0663.96.4.699

Baten, E., Praet, M., and Desoete, A. (2017). The relevance and efficacy of metacognition for instructional design in the domain of mathematics. *ZDM* 49, 613–623. doi: 10.1007/s11858-017-0851-y

Bloom, P. (2000). How children learn the meanings of words. Cambridge, MA: MIT Press.

Butterworth, B. (2005). "Developmental dyscalculia" in The handbook of mathematical cognition (Psychology Press), 455–467.

Cattell, R. B., Weiß, R. H., and Osterland, J. (1997). Grundintelligenztest Skala 1 (CFT 1). *5th* Edn. Braunschweig: Westermann.

Cook, T. D., and Campbell, D. T. (1979). Quasi-experimentation: design and analysis issues for field settings: Houghton Mifflin.

Dehaene, S. (1992). Varieties of numerical abilities. *Cognition* 44, 1–42. doi: 10.1016/0010-0277(92)90049-N

Edelsbrunner, P. A., and Thurn, C. (2024). Improving the utility of non-significant results for education: a review and recommendations. *Educ. Res. Rev.* 42:100590. doi: 10.1016/j.edurev.2023.100590

Ennemoser, M. (2010). Training mathematischer Basiskompetenzen als unterrichtsintegrierte Maßnahme in Vorklassen. *Empirische Pädagogik* 24, 336–352.

Ennemoser, M., and Krajewski, K. (2007). Effekte der Förderung des Teil-Ganzes-Verständnisses bei Erstklässlern mit schwachen Mathematikleistungen. *Viertel. Heilpäd. ihre Nachb.* 76, 228–240.

Ennemoser, M., Krajewski, K., and Sinner, D. (2017). Test zur Erfassung mathematischer Basiskompetenzen ab Schuleintritt (MBK 1+). Göttingen: Hogrefe.

Ennemoser, M., Sinner, D., and Krajewski, K. (2015). Kurz-und langfristige Effekte einer entwicklungsorientierten Mathematikförderung bei Erstklässlern mit drohender Rechenschwäche [Short and long term effects of quantity-number training with first grade at-risk children]. *Lernen Lernstörungen* 4, 43–59. doi: 10.1024/2235-0977/a000091

Feigenson, L., Carey, S., and Spelke, E. (2002). Infants' discrimination of number vs continuous extent. *Cogn. Psychol.* 44, 33–66. doi: 10.1006/cogp.2001.0760

Fischer, U., Moeller, K., Bientzle, M., Cress, U., and Nuerk, H.-C. (2011). Sensori-motor spatial training of number magnitude representation. *Psychon. Bull. Rev.* 18, 177–183. doi: 10.3758/s13423-010-0031-3

Friedrich, G., and de Galgóczy, V. (2004). Komm mit ins Zahlenland: Eine spielerische Entdeckungsreise in die Welt der Mathematik. Freiburg: Christophorus.

Friso-van den Bos, I., van der Ven, S. H. G., Kroesbergen, E. H., and van Luit, J. E. H. (2013). Working memory and mathematics in primary school children: a meta-analysis. *Educ. Res. Rev.* 10, 29–44. doi: 10.1016/j.edurev.2013.05.003

Fuchs, L. S., Cirino, P. T., Powell, S. R., Schumacher, R. F., Marrin, S., Hamlett, C. L., et al. (2014). Does calculation or word-problem instruction provide a stronger route to prealgebraic knowledge? *J. Educ. Psychol.* 106, 990–1006. doi: 10.1037/a0036793

Fuchs, L. S., Geary, D. C., Compton, D. L., Fuchs, D., Schatschneider, C., Hamlett, C. L., et al. (2013). Effects of first-grade number knowledge tutoring with contrasting forms of practice. *J. Educ. Psychol.* 105, 58–77. doi: 10.1037/a0030127

Fuson, K. (1988). Children's counting and concepts of number. New York: Springer.

Gaupp, N., Zoelch, C., and Schumann-Hengsteler, R. (2004). Defizite numerischer Basiskompetenzen bei rechenschwachen Kindern der 3. und 4. Klassenstufe. *Z. Pädag. Psychol.* 18, 31–42. doi: 10.1024/1010-0652.18.1.31

Geary, D. C., Hoard, M. K., Byrd-Craven, J., and DeSoto, M. C. (2004). Strategy choices in simple and complex addition: contributions of working memory and counting knowledge for children with mathematical disability. *J. Exp. Child Psychol.* 88, 121–151. doi: 10.1016/j.jecp.2004.03.002

Griffin, S. A., Case, R., and Capodilupo, S. (1995). "Teaching for understanding: the importance of central conceptual structures in the elementary school mathematics curriculum" in Educational environments. ed. A. Strauss (Norwood, NJ: Ablex).

Haffner, J., Baro, K., Parzer, P., and Resch, F. (2005). Der Heidelberger Rechentest: Erfassung mathematischer Basiskompetenzen im Grundschulalter (HRT 1–4). Göttingen: Hogrefe.

Halberda, J., and Feigenson, L. (2008). Developmental change in the acuity of the "number sense": The approximate number system in 3-, 4-, 5-, and 6-year-olds and adults. *Developmental Psychology*, 44, 1457–1465. doi: 10.1037/a0012682

Hasselhorn, M., and Linke-Hasselhorn, K. (2013). Fostering early numerical skills at school start in children at risk for mathematical achievement problems: a small sample size training study. *Int. Educ. Stud.* 6, 213–220. doi: 10.5539/ies.v6n3p213

Hauser, B., Vogt, F., Stebler, R., and Rechsteiner, K. (2014). Förderung früher mathematischer Kompetenzen: Spielintegriert oder trainingsbasiert [promoting early mathematical competencies: play-integrated or training-based]. *Frühe Bildung* 3, 139–145. doi: 10.1026/2191-9186/a000144

Hawes, Z., Cain, M., Jones, S., Thomson, N., Bailey, C., Seo, J., et al. (2019). Effects of a teacher-designed and teacher-led numerical board game intervention: a randomized controlled study with 4-to 6-year-olds. *Mind Brain Educ.* 14, 71–80. doi: 10.1111/mbe.12215

Hellstrand, H., Korhonen, J., Linnanmäki, K., and Aunio, P. (2020). The number race – computer-assisted intervention for mathematically low-performing first graders. *Eur. J. Spec. Needs Educ.* 35, 85–99. doi: 10.1080/13488678.2019.1615792

Hox, J. J. (2010). Multilevel analysis: techniques and applications. *2nd* Edn: Routledge.

Hyde, D. C., Khanum, S., and Spelke, E. S. (2014). Brief non-symbolic, approximate number practice enhances subsequent exact symbolic arithmetic in children. *Cognition* 131, 92–107. doi: 10.1016/j.cognition.2013.12.007

IBM Corp (2023a). IBM SPSS statistics for windows (version 29.0) [software]. Armonk, NY: IBM Corp.

IBM Corp (2023b). IBM SPSS Amos for windows (version 29.0) [software]. Armonk, NY: IBM Corp.

Jitendra, A. K., Alghamdi, A., Edmunds, R., McKevett, N. M., Mouanoutoua, J., and Roesslein, R. (2021). The effects of tier 2 mathematics interventions for students with mathematics difficulties: a meta-analysis. *Except. Child.* 87, 307–325. doi: 10.1177/0014402920969187

Jordan, N. C., Glutting, J., Dyson, N., Hassinger-Das, B., and Irwin, C. (2012). Building kindergartners' number sense: a randomized controlled study. *J. Educ. Psychol.* 104, 647–660. doi: 10.1037/a0029018

Käser, T., Baschera, G. M., Kohn, J., Kucian, K., Richtmann, V., Grond, U., et al. (2013). Design and evaluation of the computer-based training program calcularis for enhancing numerical cognition. *Front. Psychol.* 4:489. doi: 10.3389/fpsyg.2013.00489

Klauer, K. J. (1989). Denktraining für Kinder I. Göttingen: Hogrefe.

Klauer, K. J., and Phye, G. D. (2008). Inductive reasoning: a training approach. *Rev. Educ. Res.* 78, 85–123. doi: 10.3102/0034654307313402

Krajewski, K. (2008). "Vorschulische Förderung mathematischer Kompetenzen" in Enzyklopädie der Psychologie, Reihe Entwicklungspsychologie, Bd. Angewandte Entwicklungspsychologie. eds. F. Petermann and W. Schneider (Göttingen: Hogrefe), 275–304.

Krajewski, K., Küspert, P., and Schneider, W. (2002). Deutscher Mathematiktest für erste Klassen (DEMAT 1+). Göttingen: Hogrefe.

Krajewski, K., Nieding, G., and Schneider, W. (2007). Mengen, zählen, Zahlen: Die Welt der Mathematik verstehen (MZZ). Berlin: Cornelsen.

Krajewski, K., Nieding, G., and Schneider, W. (2008). Kurz-und langfristige Effekte mathematischer Frühförderung im Kindergarten durch das Programm "Mengen, zählen, Zahlen". *Z. Entwicklungspsychol. Pädag. Psychol.* 40, 135–146. doi: 10.1026/0049-8637.40.3.135

Krajewski, K., and Schneider, W. (2009a). Early development of quantity to number-word linkage as a precursor of mathematical school achievement and mathematical difficulties: findings from a four-year longitudinal study. *Learn. Instr.* 19, 513–526. doi: 10.1016/j.learninstruc.2008.10.002

Krajewski, K., and Schneider, W. (2009b). Exploring the impact of phonological awareness, visual-spatial working memory, and preschool quantity-number competencies on mathematics achievement in elementary school: findings from a 3-year-longitudinal study. *J. Exp. Child Psychol.* 103, 516–531. doi: 10.1016/j.jecp.2009.03.009

Kucian, K., Grond, U., Rotzer, S., Henzi, B., Schonmann, C., Plangger, F., et al. (2011). Mental number line training in children with developmental dyscalculia. *NeuroImage* 57, 782–795. doi: 10.1016/j.neuroimage.2011.01.070

Kucian, K., Loenneker, T., Dietrich, T., Martin, E., and von Aster, M. (2006). Impaired neural networks for approximate calculation in dyscalculic children: a functional MRI study. *Behav. Brain Funct.* 2:31. doi: 10.1186/1744-9081-2-31

Kuhl, J., Sinner, D., and Ennemoser, M. (2012). Training quantity-number competencies in students with intellectual disabilities. *J. Cogn. Educ. Psychol.* 11, 128–142. doi: 10.1891/1945-8959.11.2.128

Landerl, K., Bevan, A., and Butterworth, B. (2004). Developmental dyscalculia and basic numerical capacities: a study of 8-9 year old students. *Cognition* 93, 99–125. doi: 10.1016/j.cognition.2003.11.004

Lin, X., Peng, P., and Zeng, J. (2021). Understanding the relation between mathematics vocabulary and mathematics performance: a meta-analysis. *Elem. Sch. J.* 121, 504–540. doi: 10.1086/712504

Lin, X., and Powell, S. R. (2023). Exploring the impacts of linguistics, working memory, age, and gender on the quantity-number competencies model. *Early Child. Res. Q.* 63, 85–97. doi: 10.1016/j.ecresq.2022.12.003

Link, T., Moeller, K., Huber, S., Fischer, U., and Nuerk, H.-C. (2013). Walk the number line – an embodied training of numerical concepts. *Trends Neurosci. Educ.* 2, 74–84. doi: 10.1016/j.tine.2013.06.005

Maas, C. J. M., and Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology* 1, 86–92. doi: 10.1027/1614-2241.1.3.86

Mayer, R. E., Heiser, J., and Lonn, S. (2001). Cognitive constraints on multimedia learning: when presenting more material results in less understanding. *J. Educ. Psychol.* 93, 187–198. doi: 10.1037/0022-0663.93.1.187

McCoach, D. B., Dever, M. T., Rambo, K. E., and Lavooy, M. J. (2018). Does the package matter? A comparison of five common multilevel modeling software packages via Monte Carlo simulations. *J. Educ. Behav. Stat.* 43, 594–627. doi: 10.3102/1076998618776348

Mononen, R., Aunio, P., Koponen, T., and Aro, M. (2014). A review of early numeracy interventions for children at risk in mathematics. *Int. J. Early Childh. Spec. Educ.* 6:25. doi: 10.20489/intjecse.14355

Mussolin, C., Mejias, S., and Noël, M.-P. (2010). Symbolic and nonsymbolic number comparison in children with and without dyscalculia. *Cognition* 115, 10–25. doi: 10.1016/j.cognition.2009.10.006

Muthén, L. K., and Muthén, B. O. (2024). Mplus user's guide (version 8.11) [computer software]: Muthén & Muthén Available at: https://www.statmodel.com/.

Myers, J. A., Witzel, B. S., Powell, S. R., Li, H., Pigott, T. D., Xin, Y. P., et al. (2022). A meta-analysis of mathematics word-problem solving interventions for elementary students who evidence mathematics difficulties. *Rev. Educ. Res.* 92, 695–742. doi: 10.3102/00346543211070049

Obersteiner, A., Reiss, K., and Ufer, S. (2013). How training on exact or approximate mental representations of number can enhance first-grade students' basic number processing and arithmetic skills. *Learn. Instr.* 23, 125–135. doi: 10.1016/j.learninstruc.2012.08.004

Olyai, N., Otto, B., Büttner, G., and Krajewski, K. (2014). "Numbers-Wirksamkeitsevaluation einer unterrichtsintegrierten Förderung von Zahl-Größen-Kompetenz und Selbstregulation zur Sekundärprävention von Rechenschwierigkeiten" in Paper presented at the 49th congress of the German Society for Psychology (Bochum, Germany).

Passolunghi, M. C., Vercelloni, B., and Schadee, H. (2007). The precursors of mathematics learning: working memory, phonological ability and numerical competence. *Cogn. Dev.* 22, 165–184. doi: 10.1016/j.cogdev.2006.09.001

Praet, M., and Desoete, A. (2014). Enhancing young children's arithmetic skills through non-intensive, computerised kindergarten interventions: a randomised controlled study. *Teach. Teach. Educ.* 39, 56–65. doi: 10.1016/j.tate.2013.12.003

Purpura, D. J., and Ganley, C. M. (2014). Working memory and language: skill-specific or domain-general relations to mathematics? *J. Exp. Child Psychol.* 122, 104–121. doi: 10.1016/j.jecp.2013.12.009

R Core Team (2024). R: a language and environment for statistical computing (version 4.40) [computer software]: R Foundation for Statistical Computing Available at: https://www.R-project.org/.

Räsänen, P., Salminen, J., Wilson, A. J., Aunio, P., and Dehaene, S. (2009). Computer-assisted intervention for children with low numeracy skills. *Cogn. Dev.* 24, 450–472. doi: 10.1016/j.cogdev.2009.09.003

Renner, A., Krajewski, K., and Schneider, W. (2024). Evaluation of a structure-oriented preschool training of quantity-number competencies: Frühe Bildung.

Resnick, L. B. (1989). Developing mathematical knowledge. *Am. Psychol.* 44, 162–169. doi: 10.1037/0003-066X.44.2.162

Ribner, A. D., Ahmed, S. F., Miller-Cotto, D., and Ellis, A. (2023). The role of executive function in shaping the longitudinal stability of math achievement during early elementary grades. *Early Child. Res. Q.* 64, 84–93. doi: 10.1016/j.ecresq.2023.02.004

Riley, M. S., Greeno, J. G., and Heller, J. H. (1983). "Development of children's problem-solving ability in arithmetic" in The development of mathematical thinking. ed. H. P. Ginsburg (New York: Academic Press), 153–196.

Rousselle, L., Palmers, E., and Noël, M.-P. (2004). Magnitude comparison in preschoolers: what counts? Influence of perceptual variables. *J. Exp. Child Psychol.* 87, 57–84. doi: 10.1016/j.jecp.2003.10.005

Schneider, W., and Artelt, C. (2010). Metacognition and mathematics education. *ZDM* 42, 149–161. doi: 10.1007/s11858-010-0240-2

Schnepel, S., and Aunio, P. (2022). A systematic review of mathematics interventions for primary school students with intellectual disabilities. *Eur. J. Spec. Needs Educ.* 37, 663–678. doi: 10.1080/08856257.2021.1943268

Shalev, R. S. (2004). Developmental dyscalculia. *J. Child Neurol.* 19, 765–771. doi: 10.1177/08830738040190100601

Siegler, R. S., and Ramani, G. B. (2008). Playing linear numerical board games promotes low-income children's numerical development. *Dev. Sci.* 11, 655–661. doi: 10.1111/j.1467-7687.2008.00714.x

Simanowski, S., and Krajewski, K. (2019). Specific preschool executive functions predict unique aspects of mathematics development: a 3-year longitudinal study. *Child Dev.* 90, 544–561. doi: 10.1111/cdev.12909

Sinnakaudan, S., and Ghazali, M. (2015). Why year one students have difficulties in mastering the numbers parts-whole relationship? International conference on research and education in mathematics (ICREM7), Kuala Lumpur, 147–153.

Sinner, D., and Kuhl, J. (2010). Förderung mathematischer Basiskompetenzen in der Grundstufe der Schule für Lernhilfe. *Z. Entwicklungspsychol. Pädagog. Psychol.* 42, 241–251. doi: 10.1026/0049-8637/a000026

Snijders, T. A. B., and Bosker, R. J. (2012). Multilevel analysis: an introduction to basic and advanced multilevel modeling. *2nd* Edn. Sage.

Sterner, G., Wolff, U., and Helenius, O. (2019). Reasoning about representations: effects of an early math intervention. *Scand. J. Educ. Res.* 64, 782–800. doi: 10.1080/00313831.2019.1600579

Svane, R. P., Willemsen, M. M., Bleses, D., Krøjgaard, P., Verner, M., and Nielsen, H. S. (2023). A systematic literature review of math interventions across educational settings from early childhood education to high school. *Front. Educ.* 8:1229849. doi: 10.3389/feduc.2023.1229849

Szücs, D., and Goswami, U. (2013). Developmental dyscalculia: fresh perspectives. *Trends Neurosci. Educ.* 2, 33–37. doi: 10.1016/j.tine.2013.06.004

Turati, C., Gava, L., Valenza, E., and Ghirardi, V. (2013). Number versus extent in newborns spontaneous preference for collections of dots. *Cogn. Dev.* 28, 10–20. doi: 10.1016/j.cogdev.2012.06.002

Van Luit, J. E. H., and Schopman, E. A. M. (2000). Improving early numeracy of young children with special educational needs. *Remedial Spec. Educ.* 21, 27–40. doi: 10.1177/074193250002100105

von Aster, M. G., and Shalev, R. S. (2007). Number development and developmental dyscalculia. *Dev. Med. Child Neurol.* 49, 868–873. doi: 10.1111/j.1469-8749.2007.00868.x

Wang, A. H., Firmender, J. M., Power, J. R., and Byrnes, J. P. (2016). Understanding the program effectiveness of early mathematics interventions for prekindergarten and kindergarten environments: a meta-analytic review. *Early Educ. Dev.* 27, 692–713. doi: 10.1080/10409289.2016.1116343

Watts, T. W., Duncan, G. J., Siegler, R. S., and Davis-Kean, P. E. (2014). What's past is prologue: relations between early mathematics knowledge and high school achievement. *Educ. Res.* 43, 352–360. doi: 10.3102/0013189X14553660

Whitehurst, G. J., Falco, F. L., Lonigan, C., Fischel, J. E., DeBaryshe, B. D., Valdez-Menchaca, M. C., et al. (1988). Accelerating language development through picture book reading. *Dev. Psychol.* 24, 552–559. doi: 10.1037/0012-1649.24.4.552

Whyte, J. C., and Bull, R. (2008). Number games, magnitude representation, and basic number skills in preschooler. *Dev. Psychol.* 44, 588–596. doi: 10.1037/0012-1649.44.2.588

Wilson, A. J., Dehaene, S., Dubois, O., and Fayol, M. (2009). Effects of an adaptive game intervention on accessing number sense in low-socioeconomic-status kindergarten children. *Mind Brain Educ.* 3, 224–234. doi: 10.1111/j.1751-228X.2009.01075.x

Wilson, A. J., Revkin, S. K., Cohen, D., Cohen, L., and Dehaene, S. (2006). An open trial assessment of "the number race", an adaptive computer game for remediation of dyscalculia. *Behav. Brain Funct.* 2:20. doi: 10.1186/1744-9081-2-19

Zhang, Y., Tolmie, A., and Gordon, R. (2022). The relationship between working memory and arithmetic in primary school children: a meta-analysis. *Brain Sci.* 13:22. doi: 10.3390/brainsci13010022

Frontiers in Education

TYPE Original Research
PUBLISHED 30 August 2024
DOI 10.3389/feduc.2024.1383421

Check for updates

OPEN ACCESS

EDITED BY
Irene Cadime,
University of Minho, Portugal

REVIEWED BY
Lee Branum-Martin,
Georgia State University, United States
Rui Alexandre Alves,
University of Porto, Portugal

*CORRESPONDENCE
Christoph Weber
christoph.weber@ph-ooe.at

RECEIVED 07 February 2024
ACCEPTED 15 August 2024
PUBLISHED 30 August 2024

CITATION
Schöfl M, Weber C, Steinmair G, Zepnik S and Grimm A (2024) Language abilities and phonological information processing mediate the association of spelling with bilingualism and socioeconomic status.
Front. Educ. 9:1383421.
doi: 10.3389/feduc.2024.1383421

Language abilities and phonological information processing mediate the association of spelling with bilingualism and socioeconomic status

Martin Schöfl[1,2], Christoph Weber[1,3]*, Gabriele Steinmair[2], Sabine Zepnik[2] and Angela Grimm[4]

[1]Research Institute for Developmental Medicine, Johannes Kepler University of Linz, Linz, Upper Austria, Austria, [2]Department for Elementary and Primary Education, University of Education Upper Austria, Linz, Upper Austria, Austria, [3]Department for Inclusive Education, University of Education Upper Austria, Linz, Upper Austria, Austria, [4]Institute for Psycholinguistics and Didactics of the German Language, Faculty of Modern Languages, Goethe University Frankfurt, Frankfurt, Germany

Introduction: Research to date has focused on the associations between spelling and (i) its precursors (phonological information processing and language) and (ii) socioeconomic status and bilingualism. Studies have also indicated that bilingualism and parental education are associated with spelling precursors. Whereas these associations have previously been analyzed individually, this study proposes a mediation model in which the effects of socioeconomic status and bilingualism on spelling are mediated by phonological information processing and language skills.

Methods: A total of 1,012 German-speaking first graders attending primary schools in Austria were assessed at the beginning of the first grade on their phonological information processing and language abilities, and their spelling abilities were tested again at the end of the first grade. Subsequently, a structural equation modeling approach was employed to evaluate the mediation model.

Results: In line with the mediation hypothesis, the results show indirect effects of parental education (as a measure of socioeconomic status) on spelling via language and phonological information processing. In addition to mediation, we also found a direct effect of SES on spelling performances. For bilingualism, the results support full mediation as an indirect effect via language abilities. Notably, we found no effect of bilingual status on phonological information processing.

Discussion: This study highlights the ongoing need for systematic oral language training for bilingual children and children from low-SES backgrounds that starts in preschool and continues throughout primary school. Moreover, given the predictive effect of phonological awareness on spelling achievement, phonological awareness should be part of the training for preschool and school-aged children.

KEYWORDS

Schnapp spelling test, bilingualism, first graders, phonological information processing, language abilities, word spelling

 10.3389/feduc.2024.1383421

1 Introduction

The acquisition of spelling is a protracted cognitive undertaking that requires diligent effort irrespective of the specific orthographic system under consideration. Understanding the challenges associated with spelling acquisition is important to support spelling development. Since accurate spelling is a complex cognitive skill (Pan et al., 2021), numerous factors that influence spelling development must be considered, especially phonological awareness (phonemic manipulation, phonemic discrimination), working memory, rapid naming, multisyllable word naming, lexical and grammatical abilities, and nonverbal intelligence (Ennemoser et al., 2012; Strattman and Hodson, 2005). Socio-cultural factors, such as literacy in the home environment and income/socio-economic status (Niklas et al., 2023; Sirin, 2005) also affect spelling achievement. A bilingual biography is another influencing factor, but bilingualism is typically closely connected to oral language abilities and socioeconomic status (SES). Differences from age-matched monolinguals, if reported, are therefore often explained by factors other than bilingualism *per se* (see below). Research has hitherto targeted either the associations of spelling precursors and spelling or the associations of bilingualism, SES, and spelling. This study, in contrast, investigated the impact of oral language-related spelling precursors and language abilities and considered bilingualism and SES in German spelling development. We thus shed new light on the associations between phonological information processing skills, oral language skills, SES, and bilingualism in spelling attainment at the end of grade 1. German has an alphabetical writing system that is based on the Roman script (Coulmas, 2003; Neef, 2015; Ziegler and Goswami, 2006). While German is often described as a shallow orthography that maps letters and sounds relatively consistently like, for example, Dutch, Spanish and Finnish (Goswami et al., 2003; Goswami et al., 2001; Landerl et al., 2013), more recent analyses have challenged this view (e.g., Evertz and Primus, 2013; Fuhrhop, 2018; Maas, 2015). Even if the correct spelling can be deduced from phonographic principles in many German words, such as <Mist> [mist] 'dung', this does not hold for a large proportion of 'typical' German words. For example, applying phonographic principles to words such as <Tasse> 'cup' or <rennt> 'runs' results in the incorrect outcomes *<tase> and *<rent>, respectively. These words are spelled in accordance with so-called syllabic and morphological principles (Müller, 2010). Syllabic principles refer to the graphemic marking of tense vowels (i.e., <aa> in <Waage> 'scale' or <ah> in <Wahl> 'choice'), the spelling of ambisyllabic consonants as geminates (i.e., the <ss> in <Tasse> 'cup'), and the so-called silent <h> in words such as <ruhen> 'to rest' or <ziehen> 'to pull.' Correct spelling of these words does not solely rely on phonological information processing, but also requires knowledge of orthographical, grammatical principles that define prosodic patterns, which are the foundation of written word spelling in German. Language abilities, above all oral vocabulary and grammar, are important to infer such principles of German orthography and to construct analogies between words and phrasal units. Consequently, oral language abilities, bilingualism and SES (with SES operationalized as parental education in our study) are thought to impact language competencies and spelling development.

1.1 Oral language-related abilities as precursors: phonological information processing and oral language abilities

Written language development depends on phonological information processing, which relates – among other things – to the processing of acoustic signals of the spoken language (Ptok et al., 2008; Wagner and Torgesen, 1987) and is often described as a compound of four sub-domains: phonological awareness (PA), phonological working memory (PWM), rapid automatized naming (RAN) and letter knowledge (LK) (Vellutino et al., 2004; Wagner and Torgesen, 1987; Wolf and Bowers, 1999). In addition to PA, language abilities influence spelling achievement (Ennemoser et al., 2012; Fricke et al., 2016; Lervåg and Hulme, 2010).

Phonological awareness is a metalinguistic skill defined as awareness of the sound structure of language (Wagner and Torgesen, 1987) and the ability to analyze, segment, and manipulate phonological units (e.g., Schnitzler, 2008; Tunmer and Hoover, 1992). The term phonological awareness is used extensively in research and focuses on phonological units of various sizes: larger units are word, syllable, and rhyme, while the single phoneme is a small unit (Skowronek and Marx, 1989). Smaller units are acquired later than larger units and are more difficult to manipulate.

In addition to the size of the phonological unit, the operation process matters. Analysis is easier than segmentation or synthetization, and both are easier than manipulation (Schnitzler, 2008). Addressing the question of whether children with limited phonological competencies benefit in the long term from training, Schneider et al. (1998) reported significant training effects, particularly in spelling, when the training adhered precisely to instructions.

Phonological working memory allows temporary sound-based representations or linguistic information to be kept available until cognitive processing is complete (Baddeley, 1992). It is interrelated with PA and influences spelling development (Gindri et al., 2007; Steinbrink and Klatte, 2008). Working memory is described as a "capacity-limited system that serves the short-term storage of information" (Mähler and Schuchardt, 2014, p. 77). The capacity of working memory seems particularly relevant for children at the beginning of the reading acquisition process (Schneider, 2009). It is considered a stable trait-like ability that shows resistance to systematic training and has primarily predictive characteristics (Mayer, 2016, p. 77).

Rapid automatized naming (RAN) relates to the fast pronunciation of objects, colors, or digits (Georgiou et al., 2009). In their Double-Deficit Hypothesis, Wolf and Bowers (1999) described children with dyslexia not only in terms of deficits in phonological awareness but also in terms of differences in naming speed. There is strong agreement that children with reading difficulties, especially in consistent orthographies, can be characterized by deficits in naming speed (Brizzolara et al., 2006; Mayer, 2016). The most significant specific predictor for spelling, however, remains phonological awareness (Landerl and Wimmer, 2008). Korhonen (1995) presented RAN as a personality trait, which indicates that difficulties in rapid naming persist into early adulthood. The specific components of the complex construct of naming speed, which is determined by the interplay of visual processing speed and speed of access to phonological information, are subject of intense discussions (*cf.* Mayer, 2016). Given the well-established association between RAN and literacy acquisition, it is reasonable to assume that RAN-trainings have the potential to improve literacy, nevertheless, empirical evidence of the effectivity of RAN-based training methods is rare (Berglez, 2003; Mayer, 2016). Berglez (2003) posited that children exhibit slow naming speed due to insufficient practice, and suggested that their actions in early childhood have been insufficiently accompanied by speaking or naming. "This weakness is to be compensated by training children to name different categories of objects faster" (p. 123). In a study in preschool children, she observed significant improvements in naming speed, which indicated its fundamental trainability. However, these improvements could not be attributed to specific training effects and even affected the control groups without training. Importantly, to date there have been no reports on the effects of training to improve rapid naming on later reading and writing.

Letter knowledge defines the explicit knowledge of letter names (e.g., Ennemoser et al., 2012; Foulin, 2005) and is an important predictor of reading (Foorman et al., 2016; Malling et al., 2022; Schatschneider et al., 2004). Its predictive accuracy depends on the transparency of the orthography (Martínez and Goikoetxea, 2019). Further, the relative contributions of rapid automatized naming, letter sound knowledge, and phonological awareness differ before and after school entry (Schmitterer and Schroeder, 2019).

There is ample evidence that – together with working memory (Berninger et al., 2010) – phonological awareness is the strongest predictor of later spelling abilities (e.g., for German-speaking children, see Ennemoser et al., 2012; Pfost, 2015 for a review). Wolf and Bowers (1999) emphasized the close connection between RAN and reading, and Ennemoser et al. (2012) postulated for RAN a persistent but lower correlation to spelling than to reading.

The vocabulary is built and expanded by recognizing new words and storing them in the mental lexicon together with the semantic, syntactic and morphological information and the phonetic-phonological form for each lexical unit in a network (Levelt, 1998). This organization and structure ensures efficient word retrieval in comprehension and in production and in the oral and written modalities (Aitchison, 1997). Although the size of the mental lexicon increases dramatically at preschool age (Clark, 1995; Kauschke, 2000 for German), it takes several years to establish the lexical representations in an adult-like manner because storage of detailed phonological, morphosyntactic, and semantic information requires children to hear words in various situations (Oller and Eilers, 2002). Two characteristics of lexical development are important with regard to our study: First, lexical development largely takes place on an item-by-item basis; that is, learners must hear the words in order to acquire them. Vocabulary size is thus strongly influenced by socio-cultural factors (Hoff, 2003, but see, e.g., Zaretsky and Lange, 2017; see section 1.2). Second, lexical development is never complete, because new words are integrated into the mental lexicon and existing representations are modified throughout one's entire life (Engelkamp and Rummer, 1999).

In both of the above regards, vocabulary differs from grammatical knowledge, which is largely based on rules that are mastered at a certain point in development (Schulz, 2007). Receptive and productive grammatical knowledge grows dramatically at preschool age. Between 2 and 3 years of age, German-speaking children produce main clauses with correct subject-verb agreement, and embedded clause structure is acquired between 2;6 and 4 years of age (e.g., Clahsen, 1986; Rothweiler, 1993). Due to complex rules and a high numbers of exceptions, other phenomena in morphosyntactic development, such as case and gender marking, plural formation, and inflection of irregular verbs, remain difficult until primary school age and beyond (Schulz, 2007). German-speaking children can rarely infer plural formation, gender marking, and irregular verbs by applying rules; instead, many derivations must be learned item by item (Clahsen, 1986). Given that written language encodes grammatical information (Bredel, 2015; Fuhrhop and Peters, 2023; Maas, 2015), grammatical (morphosyntactical) knowledge influences reading development (Ennemoser et al., 2012) and word spelling, even in first graders (Fricke et al., 2016; Lervåg and Hulme, 2010; Author et al., under review).

1.2 Socio-economic status

The spelling achievement of children is strongly associated with the socioeconomic status (SES) of their parents. SES – referring to the relative social position of individuals – is a multifaceted construct that is frequently measured using one or more of the following three indicators: income, education and occupational status (Bradley and Corwyn, 2002; Conger and Donnellan, 2007; Sirin, 2005). It has repeatedly been found that children from low-SES families show lower verbal academic achievement than children from high-SES backgrounds (Sirin, 2005; White, 1982). With regard to spelling, there is also consistent evidence that low-SES children have poorer orthographic skills than their peers (see Breit et al., 2016; Niemietz et al., 2023 for German speaking countries). On average, children from high-SES backgrounds have a larger vocabulary, can use more varied and complex grammatical forms and show greater phonological awareness than children from low-SES backgrounds (Fernald et al., 2013; Gilkerson et al., 2017; Hoff, 2003; Huttenlocher et al., 2010; Li et al., 2023; Noble et al., 2007; Pace et al., 2017; Rowe, 2008). This gap in language-related skills develops before school enrollment, and it influences academic achievement throughout elementary school (Hindman et al., 2010; Morgan et al., 2015; Von Stumm et al., 2020).

Pace et al. (2017) described several pathways by which SES may affect language development in general and consequently also spelling. In line with general models of the influence of SES on child development (e.g., Bradley and Corwyn, 2002; Conger and Donnellan, 2007), SES is considered to be associated with the quantity and quality of generic (e.g., quality and quantity of general parenting behaviors) and language-specific (e.g., quality and quantity of language input provided by parents) parent–child interactions. It is assumed that low-SES parents show less positive parenting (e.g., less sensitivity) due to, amongst others, strain associated with a low SES (Conger and Donnellan, 2007), and provide less (qualitative) language input, which in turn affects language development (e.g., vocabulary and grammatical knowledge). Limited economic resources of low-SES parents are also associated with limited availability of learning materials (e.g., number of books) relevant for language development. Finally, SES may be associated with other skills that are important for language learning (e.g., processing efficiency, fast mapping skills). Low-SES children may therefore be characterized by their lower levels of the skills needed for language development (Pace et al., 2017).

1.3 Bilingual status

According to Grosjean (2012), children are bilingual if they use two or more languages in their daily life. Learning the language of instruction as a second language is a further factor associated with spelling development. Overall, international large-scale studies such as PISA (Programme for International Student Assessment; OECD, 2023) indicate that children, who learn the language of instruction as a second language (and speak another language at home) often have a migration background. A migration background and the use of a heritage language at home are strongly associated with lower academic achievement, particularly in German-speaking countries (e.g., OECD, 2023) – the context of the current study. There is consistent evidence that migration background and heritage language use are associated with lower reading scores in primary and secondary education. Notably, some of these differences are explained by the lower SES of migrant families (e.g., Breit et al., 2016; Henschel et al., 2023). Likewise, studies suggest that students who acquire German as their second language (Lenhart et al., 2019) and students with migration background (Henschel et al., 2023) lag behind their peers in their spelling skills. Again, lower SES accounts for some of the association of spelling achievement with migration background (Niemietz et al., 2023).

Bilinguals can differ from age-matched monolingual children not only in SES, but also in their oral language abilities. Due to dual language exposure, bilinguals can show accelerated or decelerated language development (Paradis and Genesee, 1996). Bilingualism can foster the acquisition of language and cognitive abilities, as expressed, for example, by the notion of a 'bilingual advantage' (Bialystok, 2011; Blom et al., 2014). A bilingual advantage has sometimes been reported for phonological awareness, which is a key predictor of written language achievement (Bialystok et al., 2003; but see Antoniou, 2019; Cat et al., 2018; Ross and Melinger, 2017 for recent evidence opposing a bilingual advantage). In addition, typological differences in the oral marking of word and sentence structure can increase the awareness of word boundaries (Veldhuis, 2015), which is relevant for spelling. However, at pre-school and primary-school age, bilinguals often lag behind monolinguals in their lexical (Czapka et al., 2019; Klassert et al., 2014; Seifert et al., 2019; for an international review see Bialystok and Luk, 2012) and grammatical (e.g., Schulz and Grimm, 2019) abilities. A decelerated acquisition is reported in particular for the sub-group of successively bilingual children, that is, children who start to acquire the ambient language with entry to kindergarten or with school enrolment (Paradis et al., 2021).

The oral language abilities are associated with the spelling outcome in several ways. Lexical and morphosyntactic abilities are required to conceptualize the text and to produce written morphemes and sentences (Ehri, 2017; Silverman et al., 2015). Due to dual language development, many bilingual children cannot rely on the same language resources as monolingual children when learning to spell, for example because they started to acquire the majority language successively to the heritage language (Schulz and Grimm, 2019). Bilingualism can thus indirectly – modulated by the oral language abilities – have a negative impact on spelling achievement. We emphasize that the difficulties are not caused by the factor bilingualism *per se* and that they do not represent a deficit of the child. Rather, spelling difficulties observed in bilingual children result from the education system (here: the Austrian), which takes monolingualism as the norm ('monolingual habitus', Gogolin, 2008), and from inadequate spelling instruction (Bredel et al., 2017). Taken together, bilingualism matters in spelling achievement; not because there is a general difficulty in learning to spell in an L2, but due to the lower SES, and because the Austrian (and similarly the German) education system does not take typical manifestations of dual language exposure into account.

1.4 The current study

As outlined above, there is evidence that bilingualism and SES are associated with spelling (see Figure 1, path c on the left; e.g., Breit et al., 2016; Henschel et al., 2023; Lenhart et al., 2019; Niemietz et al., 2023). There is also evidence that bilingualism and SES are associated with spelling precursors (see Figure 1, path a on the left). While low-SES children are considered to lag behind their peers in lexical and grammatical abilities (Gilkerson et al., 2017; Pace et al., 2017), for bilinguals differential effects are expected, specifically, lower language skill levels (grammar and vocabulary; Czapka et al., 2019; Schulz and Grimm, 2019; Seifert et al., 2019). Since findings regarding a potential bilingual advantage in phonological awareness are inconsistent (e.g., Antoniou, 2019; Cat et al., 2018; Ross and Melinger, 2017, see Section 1.3), we had no specific hypothesis in this regard. Finally, there is a large body of research highlighting the importance of language skills (grammar and vocabulary) and phonological information processing (Bigozzi et al., 2016; Ennemoser et al., 2012; von Goldammer, 2010; Kim et al., 2013; Pinto et al., 2015) to the development of spelling skills (see path b on the left side of Figure 1). Integrating these three paths into a single combined model produces a mediation model (MacKinnon, 2013; shown on the right-hand side of Figure 1) that – to the best of our knowledge – our study was the first to test. This model assumes that parental education (as a measure of SES) and bilingual status are associated with language skills and strongly with specific precursors of phonological information processing at the time of school enrollment (path a); these skills in turn predict spelling skills assessed at the end of grade 1 (path b). Path c' reflects the effect of bilingual status and parental education on spelling that is not accounted for by the mediators. Notably, we expected that – in line with a mediation hypothesis – there are significant indirect effects of bilingual status and parental education on spelling. Considering similar research into the prediction of reading (Li et al., 2023), we hypothesized that at least for parental education the c'-path is significant.

2 Method

2.1 Participants and recruitment

This study used data from two cohorts (2021/22 and 2022/23) of first graders from Upper Austrian schools, mainly from the school district Linz-Land. Recruitment proceeded as follows: First, we obtained written permission from the regional school board for Upper Austria to contact schools for participation. Second, we informed the head teachers of 11 schools about the project and invited them to participate. Since all head teachers agreed to join the project, we asked the first-grade teachers to distribute letters to parents that included general information about the project, a consent form, and a written questionnaire for them to complete and return to the school. 1,070 questionnaires were collected. For this study, the analysis excluded 58 children for the following reasons: 20 children were ill during both testing periods (i.e., at school entry and at the end of grade 1), and parental consent was not received for 38 children. Thus, data on 1,012 children (from a total of 61 classes) were used for this study. The sample comprised 48.5% girls, which reflects the proportion of girls among Upper Austrian first graders [46.7%; $\chi^2(1) = 0.736$, $p = 0.391$].

73.0% of the children grew up monolingually with Austrian German. This proportion of monolinguals corresponds to the rate of Upper Austrian first graders that speak exclusively German in their everyday life [78.9%; $\chi^2(1) = 0.014$, $p = 0.907$]. Reference values for the

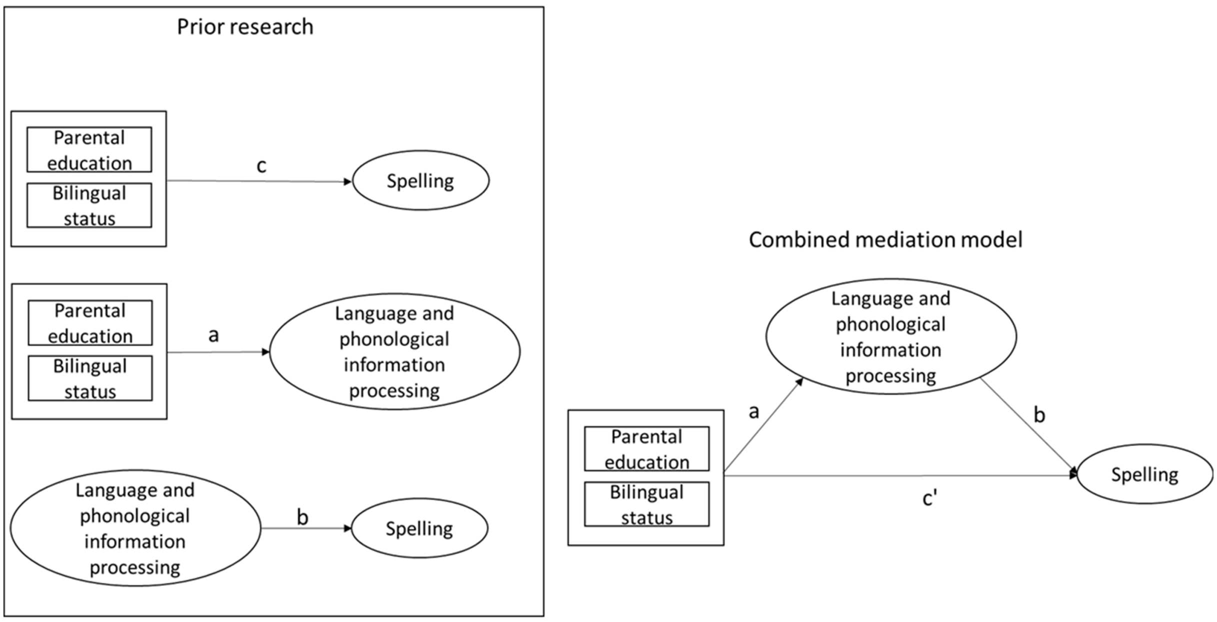

FIGURE 1
Prior research on the associations between spelling and (i) SES/parental education, bilingual status, (ii) spelling precursors and a combined mediation model.

percentages of girls and bilinguals, respectively, were taken from the Austrian School Statistics (Statistik Austria, 2024).

The overall sample consisted of parents from all educational backgrounds. Using the highest level of parental education (i.e., that of the parent with the higher level of education), in 44.6% of families at least one parent had a university degree. About a fifth (19.6%) had a university entrance qualification. Another 30% reported a vocational education or training, and in 5.9% of the families neither parent had a qualification beyond compulsory schooling. Compared to the Upper Austrian parent population of 4th graders in the school year 2017/18,[1] parents with a university degree were overrepresented (44.6% vs. 26.5%), and parents with vocational education or training were underrepresented (30.0% vs. 45.8%; $\chi^2(3) = 170.85$, $p < 0.001$). The sample was thus representative in terms of child gender and bilingual status, but not in terms of parental education, which is probably due to the sample including mostly schools in urban areas, where more parents with a higher level of education live (see also Elliott, 2018; Yulianti et al., 2023).

2.2 Procedure

The analyses described are part of a longitudinal project (SCHNAPP) that aims to assess and monitor early reading and writing development of children in primary schools. Two cohorts were assessed (a) at the very beginning of the first year of schooling (in 2021 and 2022) in terms of precursors of spelling (language and phonological information processing) and (b) at the end of first grade (in 2022 and 2023) by a spelling test.

The Schnapp spelling test (Schöfl, Steinmair, et al., 2023) was administered in class via tablet PCs. The items of this test are embedded in a cloze task, and sentences have been pre-recorded and presented as audio via headphones. Children were asked to write down the target words on the tablet using a digital pen. Unlimited repetitions and corrections were allowed. The mean duration was 15 min (SD = 7.2). Teachers were present but were asked not to interact with the children during the test. Written feedback about the performance of each class and a video-explanation were given to head teachers at the end of the assessment phase.

Oral language-related precursors and child-related information were assessed in a one-on-one setting (i.e., individual administration). Informed consent was gained from all parents, and only children whose parents had given permission participated. The assessments were carried out by trained students and project staff. The test was embedded in a magical story, presented on a tablet, to keep the children's attention and maintain motivation. Assessors adapted to the working pace of the children; the test took 19 min on average (SD = 5.2). A questionnaire was developed to collect information from parents about the child's age and whether s/he was growing up monolingually or bilingually, attested language difficulties, language use within the family, and the parent's socio-economic background (assessed via parental education).

2.3 Measures

2.3.1 Spelling

The app-based Schnapp spelling test (Schöfl et al., 2023b) comprises 22 words in 6 hierarchically organized categories of increasing spelling complexity. All items correspond to spoken and

1 We thank the IQS (Federal Institute for Quality Assurance in the Austrian School System) for providing detailed population data.

written trochees to meet the canonical prosodic (Wiese, 2006) and graphemic word shapes of German (Fuhrhop and Peters, 2023). The accuracy of children's spellings was coded in binary as correct (1) or incorrect (0). Correct means that the child's realization corresponds to the orthographic norm (i.e., whole-word accuracy). Any deviation from the normed spelling was coded as incorrect. A recent validation study (Schöfl et al., 2023b) using a Rasch modeling approach provided evidence of the unidimensionality of the spelling test and reported on its good reliability (0.86). In the current study, we were interested in the effects of multilingualism and SES on spelling. To accurately analyze spelling differences between mono- and bilingual children and among children with different parental educational backgrounds, the issue of measurement invariance of the test must be addressed (Bauer, 2023). For example, measurement invariance would be violated if – given the same latent spelling proficiency – bilingual children spelled a word (in)correctly more often than monolingual children (i.e., differential item functioning, DIF). Not adequately addressing DIF – for instance, by using a simple sum score for the spelling test or by using a factor score based on the 22 test items (given the high number of items, modeling spelling as a latent variable is not feasible) – will bias estimates of the association between spelling and bilingualism and SES, respectively (see, e.g., Bauer, 2023; Curran et al., 2018). Thus, we applied moderated nonlinear factor analysis (MNLFA; Bauer, 2017, 2023), a procedure that allows comprehensive evaluation of measurement invariance (DIF), to estimate covariate informed scale scores as the outcome variable. For details, see Supplementary material.

2.3.2 Language

Two aspects of language performance were used: (1) *Receptive vocabulary* was assessed using the digital version of the Graz Vocabulary Test (GraWo; Seifert et al., 2017), which comprises 30 matching tasks in which the child selects a picture that matches the audio-presented word. Reliability estimates provided for the paper form of GraWo (Cronbach's alpha = 0.89 at the end of first grade, retest reliability $r_{tt} = 0.93$) indicate good reliability. Internal consistency for the digital version estimated for our sample was high at 0.821. (2) *Morphosyntactic skills* were assessed by the German adaptation of the LITMUS-SRT (Hamann et al., 2013; Hamann and Abed Ibrahim, 2017 for German), a sentence-repetition task designed with a focus on multilingual children. The test comprises 15 items of varying complexity. Items were scored as correct (=1) when children repeated the sentence correctly. Internal consistency was acceptable at 0.750.

2.3.3 Phonological information processing

Phonological information processing was measured using the following subtests of the SCHNAPP Assessment (for more information see Schöfl et al., 2022; Schöfl et al., 2023b). (1) *Phonological awareness* was assessed with 10 items: children were asked to identify the initial phoneme of a word. The task used high-frequency words from the childLex database for the 6–8 years age group (Schroeder et al., 2015). A letter was presented visually and simultaneously as a speech sound, and children had to choose from three pictures that with the same initial phoneme ("Which word begins with I for Ines: Hase, Igel, Spiegel?"). Internal consistency was relatively low at 0.66. (2) *RAN* was assessed using two stimuli: objects and digits. In the object condition, five high-frequency monosyllabic words (hand, cow, tree, mouse, ice) were presented visually and orally, and the child named them as quickly as possible. The digit condition followed a similar procedure with monosyllabic digits (1, 2, 3, 6, 8). Recently collected data has confirmed the effects of both measures assessed at the beginning of grade 1 on reading skills at the end of grade 2 (Schöfl et al., 2023a). The time (measured in seconds) children needed to repeat the 5 words was recorded as a measure of RAN. The RAN modes correlated with $r = 0.603$ ($p < 0.001$). The mean of both measures was used in the subsequent analyses. (3) *Letter knowledge* was measured by displaying all the letters in random order on the screen, where each page contained three to four capital letters. Children, despite not having learned these letters at school, were asked to name any they recognized. The number of correctly named letters was recorded. (4) *Phonological working memory* was assessed using subtests of a broad-range intelligence test battery [IDS-2; Intelligence and Development Scales for Children and Adolescents (Grob et al., 2009)]. The children were asked to repeat increasingly long, and therefore increasingly difficult, digit-letter sequences in reverse order. The task ended after three unsolved or incorrectly solved tasks. Grob et al. (2009) reported high reliability (Cronbach's alpha = 0.89 at the end of the first grade, retest reliability $r_{tt} = 0.93$).

To test the assumed differentiation between language and phonological information processing skills, we applied a two-factor CFA model with (i) receptive vocabulary and morphosyntactic skills loading on the language factor and (ii) phonological awareness, RAN, letter knowledge and phonological working memory loading on the phonological information processing factor. CFI (0.952) and SRMR (0.049) indicated acceptable and good fits, respectively, whereas RMSEA (0.085, 90%-CI [0.067, 0.105]) and the ratio of χ^2/df (65.573/8 = 8.2) indicated poor model fit. The modification index suggests that allowing the errors of morphosyntactic skills and phonological working memory to covary would improve fit. Given that both tasks required children to repeat a stimulus (sentence, digit-letter sequence), error covariance is justified. The modified model yielded an overall acceptable fit (χ^2/df = 44.107/7 = 6.3; CFI = 0.969; RMSEA = 0.073 90%-CI [0.053, 0.094], SRMR = 0.039). Reliability – MacDonalds ω calculated based on the parameter estimates of the CFA (Hayes and Coutts, 2020) – indicated relatively low internal consistency for phonological information processing (ω = 0.590) and relatively high reliability for language skills (ω = 0.791). The latent correlation between two factors was $r = 0.497$ ($p < 0.001$).

2.4 Analysis

We applied a structural equation modeling (SEM) approach (e.g., González-Valenzuela et al., 2023) using Mplus 8 (Muthén and Muthén, 1998-2017) to test the mediation model outlined above. Notably, as the current study used clustered data (i.e., children in classes), variance in the spelling test might be due not only to bilingualism, parental education and precursors of spelling, but also to differences in the first year of schooling (e.g., differences in instruction and other teacher effects). However, our hypotheses focused on the individual level (level 1): We sought to explain individual differences in spelling that are not due to class-level variables (level 2). Thus, we applied multilevel SEM (see, e.g., Preacher et al., 2010), where we focused on the mediation

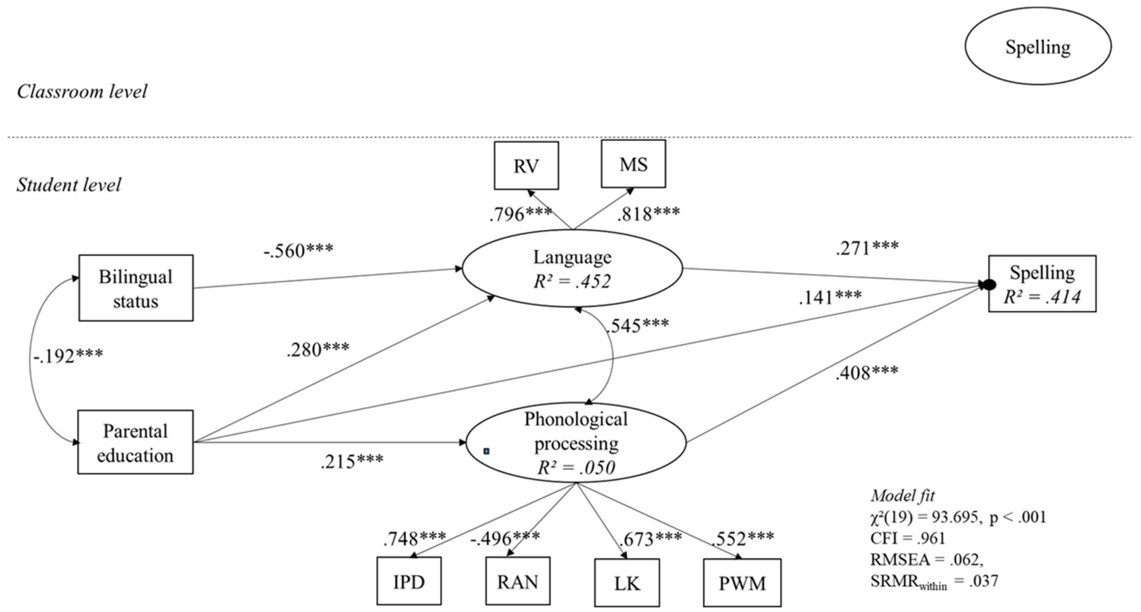

FIGURE 2
Mediation model (Standardized Coefficients). MS, morphosyntactic skills; LK, letter knowledge; RV, receptive vocabulary; IPD, initial phoneme detection; RAN, rapid automatized naming; PWM, phonological working memory. Only significant effects are shown. The graphic representation of the model follows the Mplus User's Guide (Muthén and Muthén, 1998-2017), where the black dot shown at the within level and the ellipse at between level represent the random intercept for spelling (i.e., intercepts are allowed to vary between classes). ***$p < 0.001$.

process at level 1 to explain variation in spelling within school classes[2]. Notably, although classes differed in terms of parental background, proportion of bilinguals, language skills and phonological information processing (see Supplementary Table A1), these differences were not due to schooling, as these variables were assessed at the very beginning of formal schooling and some are time-invariant (parental education, bilingualism). Therefore, these variables were not decomposed into individual and class-level components, but treated as individual-level variables.[3] The corresponding model is shown in Figure 2. The statistical significance of the indirect effects – supporting the proposed mediation hypotheses – was calculated using the delta method (MacKinnon, 2013). The preferable bias corrected bootstrap approach is not available for multilevel models in Mplus 8.

The rate of missing data ranged from 0% (bilingual status) to 16.4% (spelling). Little's Missing Completely at Random (MCAR; see, e.g., Enders, 2010) test as implemented in SPSS 29 was significant [$\chi^2(80) = 302.049$, $p < 0.001$], which indicates that nonresponse depended on the observed variables [i.e., Missing at Random (MAR)]. In detail, supplementary analyses indicated systematic dropout at time 2 (i.e., the spelling assessment). Children with missing data on the spelling test were more likely bilingual, had parents with lower education, and showed comparably lower scores on the subtests for language skills and phonological information processing. Since these results suggested a MAR mechanism, we applied a Full Information Maximum Likelihood Estimation to appropriately deal with missing values.

3 Results

The correlations of all variables are shown in Table 1.[4] Overall, we found that all variables were significantly correlated with spelling, with the highest correlations for morphosyntactic skills ($r = 0.458$, $p < 0.001$) followed by initial phoneme detection ($r = 0.397$, $p < 0.001$) and letter knowledge ($r = 0.390$, $p < 0.001$). The lowest correlation was found for bilingual status ($r = -0.197$; $p < 0.001$). Taken together, children achieved higher spelling scores at the end of grade 1 if they had better language and phonological information processing skills at

2 Notably, using a SEM-multilevel mediation approach (Preacher et al., 2010) would also allow testing for mediation at the classroom level. However, the mediation hypotheses outlined in this paper refer to the individual level. Mediation at the classroom level is not tenable because there is no basis for the a-paths at the classroom level: Since bilingual status, SES, and the mediators were assessed at the very beginning of formal schooling, any associations between these variables at the classroom level cannot be due to classroom-level mechanisms; it is therefore untenable to hypothesize that class differences in spelling precursors are affected by classroom SES and the proportion of bilingual children.

3 Following the argument that variation in the predictors and mediators cannot be due to schooling, we used no centering for these variables. Thus, for these variables the total covariance (i.e., at the within and between levels) is captured in the regression slopes. This issue has been discussed as conflation of within and between effects in the methodological literature (e.g., Preacher et al., 2010). However, if we consider the a-paths of the mediation model, it is not tenable to talk of conflation of individual and class-level effects, as the cause for the between-level covariation of the variables precedes school enrollment. The between-level variation is likely due to social and ethnic population differences in the catchment areas of the schools (Biedermann et al., 2016). Nonetheless, we also report results of a strictly within-level mediation in the Supplementary Tables A3, A4, using group-mean centering for the predictors and mediators (see also Preacher et al., 2010).

4 For correlations at the within and between levels see Supplementary Table A1.

TABLE 1 Correlations and descriptive statistics of all variables.

	BS	PE	MS	LK	RV	IPD	RAN	PWM	Spelling
PE	−0.192***								
MS	−0.487***	0.356***							
LK	−0.003	0.159***	0.214***						
RV	−0.513***	0.260***	0.663***	0.207***					
IPD	−0.023	0.151***	0.304***	0.517***	0.322***				
RAN	0.077*	−0.094**	−0.139**	−0.389***	−0.132**	−0.342***			
PWM	−0.131***	0.158***	0.389***	0.317***	0.345***	0.413***	−0.279***		
Spelling	−0.197***	0.331***	0.458***	0.390***	0.353***	0.397***	−0.312***	0.354***	
M	0.270	3.028	9.575	14.064	20.930	8.287	34.966	3.919	−0.099
SD	0.444	0.990	4.056	8.445	5.267	1.882	9.619	2.027	1.063

Correlations for spelling are based on the within class variance of spelling. For all other variables the total variance was used. BS, bilingual status; PE, parental education; MS, morphosyntactic skills; LK, letter knowledge; RV, receptive vocabulary; IPD, initial phoneme detection; RAN, rapid automatized naming; PWM, phonological working memory. ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$.

TABLE 2 Standardized indirect, direct and total effects of bilingual status and parental education on spelling.

	Indirect effects			Direct effect	Total effect
	Via language	Via phonological information processing	Total		
	β (SE)	β (SE)	β (SE)	β (SE)	β (SE)
Bilingual status	−0.152** (0.050)	−0.012 (0.018)	−0.164** (0.056)	0.024 (0.056)	−0.140*** (0.036)
Parental education	0.076** (0.026)	0.088*** (0.018)	0.163*** (0.027)	0.141*** (0.039)	0.305*** (0.036)

***$p < 0.001$, **$p < 0.01$.

school entry, grew up monolingually, and/or if they grew up in high-SES families. Further, SES was significantly correlated with all language and phonological information processing variables, with correlations ranging from $r = -0.094$ ($p < 0.01$) for RAN to $r = 0.663$ ($p < 0.001$) for receptive vocabulary. All language and phonological information processing variables correlated significantly. Finally, the negative correlation of $r = -0.193$ ($p < 0.001$) indicates that – as expected – parents of bilingual children had a significantly lower SES.

The results of the mediation model are shown in Figure 2[5] and Table 2. Overall, the model showed an acceptable fit [$\chi^2(19) = 93.69$, $p < 0.001$; CFI = 0.961, RMSEA = 0.062, SRMR = 0.032]. There was a strong effect ($\beta = -0.560$, $p < 0.001$) of bilingual status on language skills, which indicates that the monolingual children outperformed their bilingual peers in our language measures at school entry. The effect of bilingual status on phonological information processing, however, was close to zero and not significant ($\beta = -0.029$, $p > 0.05$). SES significantly predicted language skills ($\beta = 0.280$, $p < 0.001$) and phonological information processing ($\beta = 0.215$, $p < 0.001$). The higher the parental SES, the better were the spelling precursors of their children. Both language skills ($\beta = 0.270$, $p < 0.001$) and phonological information processing ($\beta = 0.408$, $p < 0.001$) predicted spelling at the end of grade 1, where the latter precursor turned out to be the strongest. Focusing on the mediation hypotheses for bilingual status, we found a non-significant direct effect close to zero on spelling ($\beta = 0.024$, $p > 0.05$). The indirect effect, however, was statistically significant ($\beta_{\text{indirect\_total}} = -0.164$, $p < 0.01$). This effect can be attributed almost completely to language skills as a mediator ($\beta_{\text{indirect}} = -0.152$, $p < 0.01$), which supports the hypothesis that bilingual children's lower spelling scores are due to less well developed language skills. Regarding parental SES, the results support partial mediation. There were significant indirect effects of parental SES on spelling via language skills ($\beta_{\text{indirect}} = -0.076$, $p < 0.01$) and via phonological information processing ($\beta_{\text{indirect}} = -0.088$, $p < 0.01$). However, even after controlling for the mediators, there was a direct effect on spelling ($\beta = 0.141$, $p < 0.001$). Thus, as proposed by Pace et al. (2017), there are apparently further mechanisms by which SES affects children's spelling.

Finally, since we tested mediation in a multilevel setting, focusing only on the individual level, we additionally report intraclass correlation (ICC) coefficients for spelling. In a null-model (i.e., no predictors), the ICC was 0.190. Thus, at the end of grade one, about 19% of the variance in spelling skills can be attributed to class differences. In the full mediation model, ICC was 0.09, indicating that

5 For detailed results see Supplementary Table A2. Notably, the standardized loadings on the phonological information processing factor were somewhat smaller than those on the language factor, which implies a somewhat smaller model-based reliability (McDonald's ω=0.71) for phonological information processing, than language (McDonald's ω=0.79). To evaluate whether reliability affected the parameter estimates (we thank the anonymous reviewer for the suggestion), we ran a small Monte Carlo simulation adapted from Thoemmes et al. (2010), where we varied the reliability of the latent mediator. The results (see Supplementary material) showed that the coefficients of the a-, b-, and c'-paths, as well as the indirect paths, are not affected by the reliability of the mediator.

about half of the class differences were due to differences in class composition in terms of bilingual status, parental education and spelling precursors. The remaining 9% of the variance were explained by other variables, such as differences in instruction method or personal characteristics of the teacher.

4 Discussion

This study took two well-documented sets of factors that influence children's spelling skills in primary school (i.e., oral language abilities and socio-demographic factors) and brought them together in a joint mediation model. In detail, we tested whether spelling differences related to SES and bilingual status are mediated by oral language-related precursors. First, our study supports prior findings which suggest that children with lower SES (Hindman et al., 2010; Morgan et al., 2015; Von Stumm et al., 2020) and bilingual children achieve lower spelling accuracy (Henschel et al., 2023; Breit et al., 2016; Lenhart et al., 2019). Second, our results are in line with studies that show an association between spelling and its precursors: Oral language-related precursors are best known to predict spelling, especially early word-spelling competencies (Caravolas et al., 2001; Niolaki et al., 2020; Treiman et al., 2023). Our findings corroborate the results published in an extensive body of literature. Both language skills and phonological information processing are powerful predictors of spelling at the end of grade 1. Interestingly, we found that phonological information processing abilities are a better predictor of spelling than language skills (vocabulary and grammar). Though this accords with the findings of some studies (e.g., Kim et al., 2013), other results contradict it (Von Goldammer et al., 2010).

Extending prior research (Segerer et al., 2013; Seifert et al., 2019 in the context of reading), our results suggest that the lower spelling abilities of bilingual children compared to their monolingual peers are due to lower language abilities, more specifically, in terms of vocabulary and grammar. In detail, the mediation model supports the notion that effects of bilingual status on spelling are fully mediated by language skills. Due to dual language exposure, bilinguals (as a group) may have limited oral lexical (Bialystok et al., 2010; Klassert et al., 2014, for German) and grammatical knowledge in the societal language, which in turn translates into difficulties dealing with the phonological and graphemic structure of words. Notably, we found no effect of bilingual status on phonological information processing, which matches recent findings that call into question a bilingual advantage in this area (Antoniou, 2019; Cat et al., 2018; Paap et al., 2014; Ross and Melinger, 2017). Limiting our findings in the process, this study took a binary perspective on bilingualism, differentiating only between monolinguals and bilinguals. Considering the language biography (i.e., measures such as age of onset of L2 use or language dominance; Grimm and Cristante, 2022; Thordardottir, 2015; Thordardottir and Brandeker, 2013) and the L1 system could help to better understand the role of bilingualism in spelling attainment.

The effect of SES on spelling is mediated by language and by phonological information processing. Notably, we also found a direct effect of SES on spelling. This suggests that, in addition to the two mediators under investigation, other mechanisms are also responsible for spelling difficulties of children growing up in low-SES families (see, e.g., Pace et al., 2017). It has been argued that children's later academic achievement is affected by parental involvement, which is likely to vary with SES, with less conducive forms of involvement being more prevalent among low-SES parents (Pomerantz and Grolnick, 2017; Weber et al., 2021); this negatively affects motivation in the educational context. As hypothesized by Pace et al. (2017), SES may affect spelling via other skills, such as processing efficiency and fast-mapping skills, which in turn also affect language development and spelling skills. Future research should thus additionally focus on further mediators that may account for the shared variance of SES and spelling. We did not consider further components of SES, such as income and occupational status (Conger and Donnellan, 2007), that may share unique variance with spelling above and beyond that of education (and may also confound the associations between bilingual status, spelling precursors, and spelling); nor did we consider potential correlations of SES and language biography with spelling in the bilingual group (i.e., age of onset to German, exposure to German and the L1). Hence, studying the variation of SES within the bilingual group would help to understand the role of SES in spelling development.

These results have several implications: First, our findings show the ongoing need for systematic oral language training for bilingual children and children from low-SES backgrounds that starts in preschool and continues throughout primary school. Providing sufficient and high-quality language training for these children should be a key concern of the educational system. Second, given that phonological awareness was a predictor of spelling achievement for all children, phonological awareness should be part of the training for preschool and school-aged children. Since our mediation model summarizes several subcomponents of phonological information processing into a single dimension, we cannot formulate clear suggestions for specific components teachers should focus on. Based on previous studies, letter knowledge seems to be a relevant factor in first graders (Elbro and Scarborough, 2004; Muter et al., 2004; Schulte-Körne et al., 2006; Foulin, 2005). Given the spelling system of German, letters should be taught in a way that allows insights to be gained into their function and graphotactic restrictions within the word. Finally, even for the regular graphemic structures implemented in the items of the SCHNAPP test, we found negative effects of SES and lower oral language achievement. This suggests that children from low-SES backgrounds and with lower language abilities have difficulties inferring basic graphemic principles of (Austrian) German independently (Bredel et al., 2017), and calls for alternative approaches to spelling instruction, for example, as suggested by Röber (2013), Bredel (2010), or Krauß (2010). These approaches explicitly address the role of syllabic and morphological principles in German language orthography and take these principles as a basis for written language instruction from the first grade onwards.

In conclusion, our study confirms the central roles of phonological awareness, oral language abilities and SES in children's spelling achievement, both as direct and indirect predictors. Importantly, our mediation model shows that negative effects of bilingualism disappear once these two factors are taken into account. Future studies will have to show whether explicit and systematic instruction based on graphemic principles indeed reduces effects of SES and oral language abilities.

Finally, there are several limitations of our study that should also be considered in future research. First, our mediation model focuses on the student level (i.e., our aim was to explain variation in spelling abilities within classes). However, there is a long tradition of research on the effects of class and school composition on student achievement (for meta-analyses see, e.g., van Ewijk and Sleegers, 2010a,b). Future research should therefore extend the analysis model to include class composition effects. Second, although the time sequence of the measurements justifies the mediation model tested, there are alternative models that may explain the covariance between variables. For example, Mehta et al. (2005) considered reading, spelling, writing, and phonological awareness as indicators of a unidimensional construct literacy. Testing such competing models would require repeated assessment of phonological information processing, language skills, and spelling over time, as this would allow application of statistical models that strengthen causal inference and make it possible to take a closer look on (bidirectional) developmental relationships between the variables (see, e.g., Lüdtke and Robitzsch, 2022; Mulder and Hamaker, 2021). Third, our analyses did not consider if language skills, phonological information processing, and SES may have different associations with spelling in mono- and bilingual children. Future research should also address such moderation effects. Lastly, although we tested the mediation model using a relatively large sample that is representative of the intake of the Austrian school system in terms of gender and the proportion of bilingual children, especially children of tertiary educated parents were overrepresented, which ultimately limits the generalizability of our results.

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 participants/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

MS: Data curation, Project administration, Writing – original draft, Writing – review & editing. CW: Data curation, Formal analysis, Methodology, Writing – original draft, Writing – review & editing. GS: Writing – original draft, Writing – review & editing. SZ: Writing – original draft, Writing – review & editing. AG: Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. All costs related to this research project were borne by the University of Education Upper Austria and Research Institute for Developmental Medicine (RID), Johannes Kepler University Linz. The tablets were provided by the project team with funds from the University of Education Upper Austria. Article processing charge was funded by the Johannes Kepler University Linz Open Access Publishing Fund.

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/feduc.2024.1383421/full#supplementary-material

References

Aitchison, J. (1997). Wörter im Kopf: Eine Einführung in das mentale Lexikon (M. Wiese, Trans.). Konzepte der Sprach- und Literaturwissenschaft, vol. *56*. Tübingen: Niemeyer.

Antoniou, M. (2019). The advantages of bilingualism debate. *Ann. Rev. Linguist.* 5, 395–415. doi: 10.1146/annurev-linguistics-011718-011820

Baddeley, A. (1992). Working memory. *Science (New York, N.Y.)* 255, 556–559. doi: 10.1126/science.1736359

Bauer, D. J. (2017). A more general model for testing measurement invariance and differential item functioning. *Psychol. Methods* 22, 507–526. doi: 10.1037/met0000077

Bauer, D. J. (2023). Enhancing measurement validity in diverse populations: modern approaches to evaluating differential item functioning. *Br. J. Math. Stat. Psychol.* 76, 435–461. doi: 10.1111/bmsp.12316

Berglez, A. (2003). Prävention von Lese-Rechtschreibschwierigkeiten. Ein Training der Benennungsgeschwindigkeit: Dissertation Universität Bielefeld.

Berninger, V. W., Abbott, R. D., Swanson, H. L., Lovitt, D., Trivedi, P., Lin, S.-J. C., et al. (2010). Relationship of word- and sentence-level working memory to reading and writing in second, fourth, and sixth grade. *Lang. Speech Hear. Serv. Sch.* 41, 179–193. doi: 10.1044/0161-1461(2009/08-0002)

Bialystok, E. (2011). Reshaping the mind: the benefits of bilingualism. *Can. J. Experiment. Psychol.* 65, 229–235. doi: 10.1037/a0025406

Bialystok, E., and Luk, G. (2012). Receptive vocabulary differences in monolingual and bilingual adults. *Biling. Lang. Congn.* 15, 397–401. doi: 10.1017/S136672891100040X

Bialystok, E., Luk, G., Peets, K. F., and Sujin, Y. A. N. G. (2010). Receptive vocabulary differences in monolingual and bilingual children. Biling.: Lang. *Cogn.* 13, 525–531. doi: 10.1017/S1366728909990423

Bialystok, E., Majumder, S. H. I. L. P. I., and Martin, M. M. (2003). Developing phonological awareness: is there a bilingual advantage? *Appl. Psycholinguist.* 24, 27–44. doi: 10.1017/S014271640300002X

Biedermann, H., Weber, C., Herzog-Punzenberger, B., and Nagel, A. (2016). "Auf die Mitschüler/innen kommt es an? Schulische Segregation – Effekte der Schul- und Klassenzusammensetzung in der Primarstufe und der Sekundarstufe I" in Nationaler Bildungsbericht Österreich 2015, Band 2: Fokussierte Analysen bildungspolitischer Schwerpunktthemen. eds. M. Bruneforth, F. Eder, K. Krainer, C. Schreiner, A. Seel and C. Spiel (Graz: Leykam), 133–174.

Bigozzi, L., Tarchi, C., Pezzica, S., and Pinto, G. (2016). Evaluating the predictive impact of an emergent literacy model on dyslexia in Italian children: a four-year prospective cohort study. *J. Learn. Disabil.* 49, 51–64. doi: 10.1177/0022219414522708

Blom, E., Küntay, A. C., Messer, M., Verhagen, J., and Leseman, P. (2014). The benefits of being bilingual: working memory in bilingual Turkish-Dutch children. *J. Exp. Child Psychol.* 128, 105–119. doi: 10.1016/j.jecp.2014.06.007

Bradley, R. H., and Corwyn, R. F. (2002). Socioeconomic status and child development. *Annu. Rev. Psychol.* 53, 371–399. doi: 10.1146/annurev.psych.53.100901.135233

Bredel, U. (2010). Der Schrift vertrauen. Wie Wörter und ihre Strukturen entdeckt werden können. *Praxis Deutsch* H.221, 14–21.

Bredel, U. (2015). "Grammatik und Literalität" in Deutschunterricht in Theorie und Praxis: Band 6. Deutschunterricht in Theorie und Praxis: Sprachreflexion und Grammatikunterricht. ed. H. Gornik (Baltmannsweiler: Schneider Verlag Hohengehren GmbH), 160–173.

Bredel, U., Fuhrhop, N., and Noack, C. (2017). Wie Kinder lesen und schreiben lernen. Tübinger: Narr Francke Attempo Verlag GmbH + Co.

Breit, S., Bruneforth, M., and Schreiner, C. (2016). Standardüberprüfung 2015 Deutsch, 4. Schulstufe: Bundesergebnisbericht. BMBWF.

Brizzolara, D., Chilosi, A., Cipriani, P., Di Filippo, G., Gasperini, F., Mazzotti, S., et al. (2006). Do phonologic and rapid automatized naming deficits differentially affect dyslexic children with and without a history of language delay? A study of Italian dyslexic children. *Cognit. Behav. Neurol.* 19, 141–149. doi: 10.1097/01.wnn.0000213902.59827.19

Caravolas, M., Hulme, C., and Snowling, M. J. (2001). The foundations of spelling ability: evidence from a 3-year longitudinal study. *J. Mem. Lang.* 45, 751–774. doi: 10.1006/jmla.2000.2785

Clahsen, H. (1986). Die Profilanalyse: Ein linguistisches Verfahren für die Sprachdiagnose im Vorschulalter. Berlin: Logotherapia. 3.

Clark, E. V. (1995). The lexikon acquisition. Cambridge studies in linguistics 65. Cambridge: Cambridge University Press.

Conger, R. D., and Donnellan, M. B. (2007). An interactionist perspective on the socioeconomic context of human development. *Annu. Rev. Psychol.* 58, 175–199. doi: 10.1146/annurev.psych.58.110405.085551

Coulmas, F. (2003). Writing systems: An introduction to their linguistic analysis. Cambridge textbooks in linguistics. Cambridge: Cambridge University Press.

Curran, P. J., Cole, V. T., Bauer, D. J., Rothenberg, W. A., and Hussong, A. M. (2018). Recovering predictor-criterion relations using covariate-informed factor score estimates. *Struct. Equ. Model. Multidiscip. J.* 25, 860–875. doi: 10.1080/10705511.2018.1473773

Czapka, S., Klassert, A., and Festman, J. (2019). Executive functions and language: their differential influence on mono- vs. multilingual spelling in primary school. *Front. Psychol.* 10:97. doi: 10.3389/fpsyg.2019.00097

Ehri, L. C. (2017). "Orthographic mapping and literacy development revisited" in Studies in written language and literacy. Theories of Reading development. eds. K. Cain, D. L. Compton and R. K. Parrila (Amsterdam: John Benjamins Publishing Company), 127–145.

Elbro, C., and Scarborough, H. S. (2004). "Early identification" in Handbook of Children's literacy. eds. T. Nunes and P. Bryant (Dordrecht: Kluwer), 339–359.

Elliott, J. (2018). Family engagement in urban education: the impact and factors affecting engagement. Master of Education Applied Research Projects. Available at: https://digitalcommons.cedarville.edu/education_research_projects/5 (Accessed March 18, 2024).

Enders, C. K. (2010). Applied missing data analysis. Methodology in the social sciences. New York: The Guilford Press.

Engelkamp, J., and Rummer, R. (1999). "Die Architektur des mentalen Lexikons" in Enzyklopädie der Psychologie: Themenbereich C Theorie und Forschung; Ser. 3 Sprache; Bd. 2. Sprachrezeption. ed. A. D. Friederici (Göttinger: Hogrefe Verl. für Psycholgie), 155–201.

Ennemoser, M., Marx, P., Weber, J., and Schneider, W. (2012). Spezifische Vorläuferfertigkeiten der Lesegeschwindigkeit, des Leseverständnisses und des Rechtschreibens. *Zeitschrift Entwicklungspsychologie Pädagogische Psychologie* 44, 53–67.

Evertz, M., and Primus, B. (2013). The graphematic foot in English and German. *Writ. Syst. Res.* 5, 1–23. doi: 10.1080/17586801.2013.765356

Fernald, A., Marchman, V. A., and Weisleder, A. (2013). Ses differences in language processing skill and vocabulary are evident at 18 months. *Dev. Sci.* 16, 234–248. doi: 10.1111/desc.12019

Foorman, B., Coyne, M., Denton, C. A., Dimino, J., Hayes, L., Justice, L., et al. (2016). Foundational skills to support Reading for understanding in kindergarten through 3rd grade. NCEE 2016–4008, U.S. Department Of Education. Available at: https://ies.ed.gov/ncee/wwc/Docs/practiceGuide/wwc_foundationalreading_040717.pdf (Accessed March 11, 2024).

Foulin, J. N. (2005). Why is letter-name knowledge such a good predictor of learning to read? *Read. Writ.* 18, 129–155. doi: 10.1007/s11145-004-5892-2

Fricke, S., Szczerbinski, M., Fox-Boyer, A., and Stackhouse, J. (2016). Preschool predictors of early literacy Acquisition in German-Speaking Children. *Read. Res. Q.* 51, 29–53. doi: 10.1002/rrq.116

Fuhrhop, N. (2018). "19. Graphematik des Deutschen" in Grammatiktheorie und Empirie in der germanistischen Linguistik. eds. A. Wöllstein, P. Gallmann, M. Habermann and M. Krifka (Berlin, Boston: De Gruyter), 587–616.

Fuhrhop, N., and Peters, J. (2023, 2023). Einführung in die Phonologie und Graphematik. 2nd Edn.

Georgiou, G. K., Parrila, R., and Kirby, J. R. (2009). RAN components and Reading development from grade 3 to grade 5: what underlies their relationship? *Sci. Stud. Read.* 13, 508–534. doi: 10.1080/10888430903034796

Gilkerson, J., Richards, J. A., Warren, S. F., Montgomery, J. K., Greenwood, C. R., Kimbrough Oller, D., et al. (2017). Mapping the early language environment using all-day recordings and automated analysis. *Am. J. Speech Lang. Pathol.* 26, 248–265. doi: 10.1044/2016_AJSLP-15-0169

Gindri, G., Keske-Soares, M., and Mota, H. B. (2007). Memória de trabalho, consciência fonológica e hipótese de escrita [working memory, phonological awareness and spelling hypothesis]. *Profono* 19, 313–322. doi: 10.1590/S0104-56872007000300010

Gogolin, I. (2008). Der monolinguale Habitus der multilingualen Schule. Internationale Hochschulschriften, vol. *101*. Münster: Waxmann Verlag.

González-Valenzuela, M.-J., Lopez-Montiel, D., Díaz-Giraldez, F., and Martín-Ruiz, I. (2023). Word writing and cognitive predictors in Spanish at the age of seven. *Int. J. Early Childhood*. doi: 10.1007/s13158-023-00359-w

Goswami, U., Grabherr, B., Ziegler, J. C., Dalton, L., and Schneider, W [Wolfgang] (2003). Nonword reading across orthographies: how flexible is the choice of reading units? Appl. Psycholinguist., 24, 235–247. doi: 10.1017/S0142716403000134

Goswami, U., Ziegler, J. C., Dalton, L., and Schneider, W. (2001). Pseudohomophone effects and phonological recoding procedures in Reading development in English and German. *J. Mem. Lang.* 45, 648–664. doi: 10.1006/jmla.2001.2790

Grimm, A., and Cristante, V. (2022). Deutsch als Zweitsprache - DaZ. Kurze Einführungen in die germanistische Linguistik: Band 28. Heidelberg: Universitätsverlag Winter.

Grob, A., Meyer, C., and Hagmann-von Arx, P. (2009). Intelligence and Development Scales (IDS). Intelligenz- und Entwicklungsskalen für Kinder von 5–10 Jahren. Bern: Hans Huber.

Grosjean, F. (2012). "Bilingualism: a short introduction" in The psycholinguistics of bilingualism. eds. F. Grosjean and P. Li. *1st* ed (New Jersey: John Wiley & Sons), 5–25.

Cat, C., Gusnanto, A., and Serratrice, L. (2018). Identifying a threshold for the executive function advantage in bilingual children. *Stud. Second. Lang. Acquis.* 40, 119–151. doi: 10.1017/S0272263116000486

Hamann, C., and Abed Ibrahim, L. (2017). Methods for Identifying specific language impairment in bilingual populations in Germany. *Front. Commun.* 2:16. doi: 10.3389/fcomm.2017.00016

Hamann, C., Chilla, S., Ruigendijk, E., and Abed Ibrahim, L. (2013). A German sentence repetition task: Testing bilingual Russian/German children.

Hayes, A. F., and Coutts, J. J. (2020). Use omega rather than Cronbach's alpha for estimating reliability. But…. *Commun. Methods Meas.* 14, 1–24. doi: 10.1080/19312458.2020.1718629

Henschel, S., Heppt, B., and Weirich, S. (2023). "Zuwanderungsbezogene Disparitäten" in IQB-Bildungstrend 2022: Sprachliche Kompetenzen am Ende der 9. Jahrgangsstufe im dritten Ländervergleich. eds. P. Stanat, S. Schipolowski, R. Schneider, S. Weirich, S. Henschel and K. A. Sachse (Münster: Waxmann), 299–344.

Hindman, A. H., Skibbe, L. E., Miller, A., and Zimmerman, M. (2010). Ecological contexts and early learning: contributions of child, family, and classroom factors during head start, to literacy and mathematics growth through first grade. *Early Child. Res. Q.* 25, 235–250. doi: 10.1016/j.ecresq.2009.11.003

Hoff, E. (2003). The specificity of environmental influence: socioeconomic status affects early vocabulary development via maternal speech. *Child Dev.* 74, 1368–1378. doi: 10.1111/1467-8624.00612

Huttenlocher, J., Waterfall, H., Vasilyeva, M., Vevea, J., and Hedges, L. V. (2010). Sources of variability in children's language growth. *Cogn. Psychol.* 61, 343–365. doi: 10.1016/j.cogpsych.2010.08.002

Kauschke, C. (2000). "Der Erwerb des frühkindlichen Lexikons: Eine empirische Studie zur Entwicklung des Wortschatzes im Deutschen" in Tübinger Beiträge zur Linguistik. Series A, language development, vol. *27* (Tübingen: G. Narr).

Kim, Y.-S., Apel, K., and Al Otaiba, S. (2013). The relation of linguistic awareness and vocabulary to word reading and spelling for first-grade students participating in response to intervention. *Lang. Speech Hear. Serv. Sch.* 44, 337–347. doi: 10.1044/0161-1461(2013/12-0013)

Klassert, A., Gagarina, N., and Kauschke, C. (2014). Object and action naming in Russian- and German-speaking monolingual and bilingual children. *Biling. Lang. Congn.* 17, 73–88. doi: 10.1017/S1366728913000096

Korhonen, T. T. (1995). The persistence of rapid naming problems in children with reading disabilities: a nine-year follow-up. *J. Learn. Disabil.* 28, 232–239. doi: 10.1177/002221949502800405

Krauß, A. (2010). "Orthographieerwerb von Beginn an. Ein silbenorientiertes Konzept für den Anfangsunterricht" in Reihe Germanistische Linguistik: Vol. 289.

Schriftsystem und Schrifterwerb: Linguistisch - didaktisch - empirisch (pp. 133–150). eds. U. Bredel, A. Müller and G. Hinney (Berlin: De Gruyter).

Landerl, K., Ramus, F., Moll, K., Lyytinen, H., Leppänen, P. H. T., Lohvansuu, K., et al. (2013). Predictors of developmental dyslexia in European orthographies with varying complexity. *J. Child Psychol. Psychiatry* 54, 686–694. doi: 10.1111/jcpp.12029

Landerl, K., and Wimmer, H. (2008). Development of word reading fluency and spelling in a consistent orthography: an 8-year follow-up. *J. Educ. Psychol.* 100, 150–161. doi: 10.1037/0022-0663.100.1.150

Lenhart, J., Marx, P., Segerer, R., and Schneider, W [Wolfgang] (2019). Rechtschreibung ohne Schreiben. Diagnostica, 65, 216–227. doi: 10.1026/0012-1924/a000229

Lervåg, A., and Hulme, C. (2010). Predicting the growth of early spelling skills: are there heterogeneous developmental trajectories? *Sci. Stud. Read.* 14, 485–513. doi: 10.1080/10888431003623488

Levelt, W. J. M. (1998). Speaking: From intention to articulation (5. Print). ACL-MIT press series in natural-language processing. Cambridge Mass: MIT Press.

Li, J., Peng, P., Ma, X., Ding, N., and Zhao, J. (2023). How does family socioeconomic status influence Children's Reading ability? Evidence from Meta-analytic structural equation modeling. *Educ. Psychol. Rev.* 35:119. doi: 10.1007/s10648-023-09834-1

Lüdtke, O., and Robitzsch, A. (2022). A comparison of different approaches for estimating cross-lagged effects from a causal inference perspective. *Struct. Equ. Model. Multidiscip. J.* 29, 888–907. doi: 10.1080/10705511.2022.2065278

Maas, U. (2015). "Laute und Buchstaben - zu den phonographischen Grundlagen des Schrifterwerbs" in Schriftsprach- und Orthographieerwerb: Erstlesen, Erstschreiben. eds. C. Röber and H. Olfert (Baltmannsweiler: Schneider Verlag), 135–161.

MacKinnon, D. P. (2013). Introduction to statistical mediation analysis: [CD included]. New York: Routledge.

Mähler, C., and Schuchardt, K. (2014). "Die Bedeutung des Arbeitsgedächtnisses für den Schriftspracherwerb" in Legasthenie und Dyskalkulie: Neue Methoden zur Diagnostik und Förderung // Legasthenie und Dyskalkulie: neue Methoden zu Diagnostik und Förderung: Mit 21 Tabellen. eds. G. Schulte-Körne, K. Brandelik and B. Estner (Bochum: Winkler), 77–86.

Malling, A. S. B., Juul, H., Gejl, A. K., Damsgaard, L., Wienecke, J., and Nielsen, A.-M. V. (2022). Word Reading, letter knowledge, and memory skills in Danish children (6-year-olds). *Scand. J. Educ. Res.* 66, 1237–1252. doi: 10.1080/00313831.2021.1983646

Martínez, N., and Goikoetxea, E. (2019). Predictors of Reading and spelling words change as a function of syllabic structure in Spanish. *Psicol. Educ.* 26, 37–48. doi: 10.5093/psed2019a20

Mayer, A. (2016). Lese-Rechtschreibstörungen (LRS). München, Basel: Ernst Reinhardt Verlag.

Mehta, P. D., Foorman, B. R., Branum-Martin, L., and Taylor, W. P. (2005). Literacy as a unidimensional multilevel construct: validation, sources of influence, and implications in a longitudinal study in grades 1 to 4. *Sci. Stud. Read.* 9, 85–116. doi: 10.1207/s1532799xssr0902_1

Morgan, P. L., Farkas, G., Hillemeier, M. M., Hammer, C. S., and Maczuga, S. (2015). 24-month-old children with larger Oral vocabularies display greater academic and behavioral functioning at kindergarten entry. *Child Dev.* 86, 1351–1370. doi: 10.1111/cdev.12398

Mulder, J. D., and Hamaker, E. L. (2021). Three extensions of the random intercept cross-lagged panel model. *Struct. Equ. Model. Multidiscip. J.* 28, 638–648. doi: 10.1080/10705511.2020.1784738

Müller, A. (2010). Rechtschreiben lernen: Die Schriftstruktur entdecken - Grundlagen und Übungsvorschläge//Die Schriftstruktur entdecken; Grundlagen und Übungsvorschläge; [mit CD-ROM (3. Auflage 2019, entspricht der 3. überarbeiteten Auflage 2019//1.Aufl.)]. Praxis Deutsch. Klett|Kallmeyer; Kallmeyer in Verb. mit Klett; Klett.

Muter, V., Hulme, C., Snowling, M. J., and Stevenson, J. (2004). Phonemes, rimes, vocabulary, and grammatical skills as foundations of early reading development: evidence from a longitudinal study. *Dev. Psychol.* 40, 665–681. doi: 10.1037/0012-1649.40.5.665

Muthén, L. K., and Muthén, B. O. (1998-2017). Mplus user's guide. *Eighth* Edn. Los Angeles, CA: Muthén & Muthén.

Neef, M. (2015). Writing systems as modular objects: proposals for theory design in grapholinguistics. *Open Linguist.* 1. doi: 10.1515/opli-2015-0026

Niemietz, J., Jindra, C., Schneider, R., Schumann, K., Schipolowski, S., and Sachse, K. A. (2023). "Soziale Disparitäten" in IQB-Bildungstrend 2022: Sprachliche Kompetenzen am Ende der 9. Jahrgangsstufe im dritten Ländervergleich. eds. P. Stanat, S. Schipolowski, R. Schneider, S. Weirich, S. Henschel and K. A. Sachse (Münster: Waxmann), 261–298.

Niklas, F., Wirth, A., Mues, A., and Birtwistle, E. (2023). "Children's early spelling development in the context of the home literacy environment" in Routledge international handbook of visual-motor skills, handwriting, and spelling. eds. Y. Ye, T. Inoue, U. Maurer and C. McBride (Abingdon: Routledge), 366–378.

Niolaki, G. Z., Vousden, J., Terzopoulos, A. R., Taylor, L. M., Sephton, S., and Masterson, J. (2020). Predictors of single word spelling in English speaking children: a cross sectional study. *J. Res. Read.* 43, 577–596. doi: 10.1111/1467-9817.12330

Noble, K. G., McCandliss, B. D., and Farah, M. J. (2007). Socioeconomic gradients predict individual differences in neurocognitive abilities. *Dev. Sci.* 10, 464–480. doi: 10.1111/j.1467-7687.2007.00600.x

OECD (2023). PISA 2022 results (volume I): The state of learning and equity in education. Paris: OECD Publishing.

Oller, K., and Eilers, R. (2002). Language and literacy in bilingual children. Blue Ridge Summit: Multilingual Matters.

Paap, K. R., Johnson, H. A., and Sawi, O. (2014). Are bilingual advantages dependent upon specific tasks or specific bilingual experiences? *J. Cogn. Psychol.* 26, 615–639. doi: 10.1080/20445911.2014.944914

Pace, A., Luo, R., Hirsh-Pasek, K., and Golinkoff, R. M. (2017). Identifying pathways between socioeconomic status and language development. *Ann. Rev. Linguist.* 3, 285–308. doi: 10.1146/annurev-linguistics-011516-034226

Pan, S. C., Rickard, T. C., and Bjork, R. A. (2021). Pan, Rickard, and Bjork (2021) does spelling still matter—And if so, how should it be taught?

Paradis, J., and Genesee, F. (1996). Syntactic Acquisition in Bilingual Children. *Stud. Second. Lang. Acquis.* 18, 1–25. doi: 10.1017/S0272263100014662

Paradis, J., Genesee, F., and Crago, M. (2021). Dual language development & disorders: a handbook on bilingualism and second language learning. *3rd* Edn. Newburyport: Brookes Publishing.

Pfost, M. (2015). Children's phonological awareness as a predictor of Reading and spelling. *Zeitschrift Entwicklungspsychol. Pädagog. Psychol.* 47, 123–138. doi: 10.1026/0049-8637/a000141

Pinto, G., Tarchi, C., and Bigozzi, L. (2015). The relationship between oral and written narratives: a three-year longitudinal study of narrative cohesion, coherence, and structure. *Br. J. Educ. Psychol.* 85, 551–569. doi: 10.1111/bjep.12091

Pomerantz, E. M., and Grolnick, W. S. (2017). "The role of parenting in children's motivation and competence: what underlies facilitative parenting?" in Handbook of competence and motivation: theory and application. eds. A. J. Elliot, C. S. Dweck and D. S. Yeager (The Guilford Press), 566–585.

Preacher, K. J., Zyphur, M. J., and Zhang, Z. (2010). A general multilevel SEM framework for assessing multilevel mediation. *Psychol. Methods* 15, 209–233. doi: 10.1037/a0020141

Ptok, M., Berendes, K., Gottal, S., Grabherr, B., Schneeberg, J., and Wittler, M. (2008). Phonologische Verarbeitung. *Monatsschr. Kinderheilkd.* 156, 860–866. doi: 10.1007/s00112-008-1770-3

Röber, C. (2013). Die Leistungen der Kinder beim Lesen- und Schreibenlernen. Grundlagen der Silbenanalytischen Methode: Ein Arbeitsbuch mit Übungsaufgaben. Schneider-Verlag Hohengehren.

Ross, J., and Melinger, A. (2017). Bilingual advantage, bidialectal advantage or neither? Comparing performance across three tests of executive function in middle childhood. *Dev. Sci.* 20. doi: 10.1111/desc.12405

Rothweiler, M. (1993). Der Erwerb von Nebensätzen im Deutschen: Eine Pilotstudie. *Linguistische Arbeiten* 302. doi: 10.1515/9783111355856

Rowe, M. L. (2008). Child-directed speech: relation to socioeconomic status, knowledge of child development and child vocabulary skill. *J. Child Lang.* 35, 185–205. doi: 10.1017/S0305000907008343

Schatschneider, C., Fletcher, J. M., Francis, D. J., Carlson, C. D., and Foorman, B. R. (2004). Kindergarten prediction of Reading skills: a longitudinal comparative analysis. *J. Educ. Psychol.* 96, 265–282. doi: 10.1037/0022-0663.96.2.265

Schmitterer, A. M., and Schroeder, S. (2019). Effects of reading and spelling predictors before and after school entry: evidence from a German longitudinal study. *Learn. Instr.* 59, 46–53. doi: 10.1016/j.learninstruc.2018.09.005

Schneider, W. (2009). "The development of reading and spelling: relevant precursors, developmental changes, and individual differences" in Human development from early childhood to early adulthood: Findings from a 20 year longitudinal study. eds. W. W. Schneider and M. Bullock (London: Psychology Press), 199–220.

Schneider, W., Roth, E., Küspert, P., and Ennemoser, M. (1998). Kurz- und langfristige Effekte eines Trainings der sprachlichen (phonologischen) Bewußtheit bei unterschiedlichen Leistungsgruppen: Befunde einer Sekundäranalyse. *Zeitschrift Entwicklungspsychol. Pädagog. Psychol.* 30, 26–39.

Schnitzler, C. (2008). Phonologische Bewusstheit und Schriftspracherwerb: 31 Tabellen. Forum Logopädie: Thieme.

Schöfl, M., Seifert, S., Steinmair, G., and Weber, C. (2023a). Are both object and alphanumeric rapid automatized naming measures required to predict word Reading fluency in German prereaders? *J. Psychoeduc. Assess.* 41, 811–827. doi: 10.1177/07342829231185375

Schöfl, M., Steinmair, G., Holzinger, D., and Weber, C. (2022). Predicting word Reading deficits using an app-based screening tool at school entry. *Front. Pediatr.* 10:863477. doi: 10.3389/fped.2022.863477

Schöfl, M., Steinmair, G., Zepnik, S., and Gnambs, T. (2023b). Entwicklung eines digitalen Rechtschreibtests für die erste Klasse Grundschule: Dimensionalität und Reliabilität des SCHNAPP-Rechtschreibtests. *Lernen Lernstörungen* 12, 188–200. doi: 10.1024/2235-0977/a000404

Schöfl, M., Steinmair, G., Zepnik, S., and Weber, C. (under review). Using an app-based screening tool to predict deficits in written word spelling at school entry. *Front. Educ.*

Schroeder, S., Würzner, K.-M., Heister, J., Geyken, A., and Kliegl, R. (2015). childLex–Eine lexikalische Datenbank zur Schriftsprache für Kinder im Deutschen. *Psychol. Rundsch.* 66, 155–165. doi: 10.1026/0033-3042/a000275

Schulte-Körne, G., Warnke, A., and Remschmidt, H. (2006). Zur Genetik der Lese-Rechtschreibschwäche [Genetics of dyslexia]. *Zeitschrift Kinder Jugendpsychiatrie Psychotherapie* 34, 435–444. doi: 10.1024/1422-4917.34.6.435

Schulz, P. (2007). "Erstspracherwerb Deutsch: Sprachliche Fähigkeiten von eins bis zehn" in Entwicklungslinien der Grundschulpädagogik: Vol. 4. Diagnostik und Förderung im Elementarbereich und Grundschulunterricht: Lernprozesse wahrnehmen, deuten und begleiten. eds. U. Graf and E. Moser Opitz (Baltmannsweiler: Schneider Verlag Hohengehren), 67–85.

Schulz, P., and Grimm, A. (2019). The age factor revisited: timing in acquisition interacts with age of onset in bilingual acquisition. *Front. Psychol.* 9:2732. doi: 10.3389/fpsyg.2018.02732

Segerer, R., Marx, A., Stanat, P., Schneider, W, and Roick, T., & Marx, P. (2013). Determinanten der Lesekompetenz bei Jugendlichen nicht deutscher Herkunftssprache. Zur Bedeutung der Spracherwerbsreihenfolge im Falle von Mehrsprachigkeit. PISA 2009 - Impulse für die Schul- und Unterrichtsforschung, Zeitschrift für Pädagogik, Beiheft

Seifert, S., Paleczek, L., and Gasteiger-Klicpera, B. (2019). Rezeptive Wortschatzleistungen in der Grundschule: Unterschiede zwischen Kindern mit Deutsch als Erst- und Zweitsprache. *Empirische Sonderpädagogik* 4, 259–278.

Seifert, S., Paleczek, L., Schwab, S., and Gasteiger-Klicpera, B. (2017). GraWo – Grazer Wortschatztest: Hogrefe.

Silverman, R. D., Coker, D., Proctor, C. P., Harring, J., Piantedosi, K. W., and Hartranft, A. M. (2015). The relationship between language skills and writing outcomes for linguistically diverse students in upper elementary school. *Elem. Sch. J.* 116, 103–125. doi: 10.1086/683135

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

Skowronek, H., and Marx, H. (1989). "The Bielefeld longitudinal study on early identification of risks in learning read and write: theoretical background and first results" in Children at risk: Assessment, longitudinal research and intervention. eds. M. Brambring, F. Lösel and H. Skowronek, 268–294.

Statistik Austria. (2024). STATcube – Statistische Datenbank von STATISTIK AUSTRIA: Schulstatistik; Schüler:innen ab 2006. Available at: https://statcube.at (Accessed February 17, 2024).

Steinbrink, C., and Klatte, M. (2008). Phonological working memory in German children with poor reading and spelling abilities. *Dyslexia* 14, 271–290. doi: 10.1002/dys.357

Strattman, K., and Hodson, B. W. (2005). Variables that influence decoding and spelling in beginning readers. *Child Lang. Teach. Therapy* 21, 165–190. doi: 10.1191/0265659005ct287oa

Thoemmes, F., MacKinnon, D. P., and Reiser, M. R. (2010). Power analysis for complex mediational designs using Monte Carlo methods. *Struct. Equ. Model.* 17, 510–534. doi: 10.1080/10705511.2010.489379

Thordardottir, E. (2015). 12. Proposed diagnostic procedures for use in bilingual and cross-linguistic contexts. In S. Armon-Lotem, JongJ. de and N. Meir (Eds.), Communication disorders across languages. Assessing multilingual children: Disentangling bilingualism from language impairment // assessing multilingual children: Disentangling bilingualism from language impairment (pp. 331–358). Blue Ridge Summit: Multilingual Matters

Thordardottir, E., and Brandeker, M. (2013). The effect of bilingual exposure versus language impairment on nonword repetition and sentence imitation scores. *J. Commun. Disord.* 46, 1–16. doi: 10.1016/j.jcomdis.2012.08.002

Treiman, R., Hulslander, J., Olson, R. K., Samuelsson, S., Elwér, Å., Furnes, B., et al. (2023). Predicting later spelling from kindergarten spelling in U.S., Australian, and Swedish children. *Sci. Stud. Read.* 27, 428–442. doi: 10.1080/10888438.2023.2186234

Tunmer, W. E., and Hoover, W. A. (1992). "Cognitive and linguistic factors in learning to read" in Reading acquisition. eds. P. B. Gough, L. C. Ehri and R. Treiman (Lawrence Erlbaum Associates, Inc), 175–214.

van Ewijk, R., and Sleegers, P. (2010a). 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

van Ewijk, R., and Sleegers, P. (2010b). Peer ethnicity and achievement: a meta-analysis into the compositional effect. *Sch. Eff. Sch. Improv.* 21, 237–265. doi: 10.1080/09243451003612671

Veldhuis, T. M. (2015). Effects of literacy, typology and frequency on children's language segmentation and processing units [Ph.D Dissertation]. University of Tilburg

Vellutino, F. R., Fletcher, J. M., Snowling, M. J., and Scanlon, D. M. (2004). Specific reading disability (dyslexia): what have we learned in the past four decades? *J. Child Psychol. Psychiatry* 45, 2–40. doi: 10.1046/j.0021-9630.2003.00305.x

Goldammer, A.von. (2010). Von der Sprache zur Schriftsprache: Diagnostische und prognostische Validität der Erfassung von Vorläuferkompetenzen der Schriftsprache im Vorschulalter.

Goldammer, A.Von, Mähler, C., Bockmann, A.-K., and Hasselhorn, M. (2010). Vorhersage früher Schriftsprachleistungen aus vorschulischen Kompetenzen der Sprache und der phonologischen Informationsverarbeitung. *Zeitschrift Entwicklungspsychol. Pädagog. Psychol.*, 42, 48–56. doi: 10.1026/0049-8637/a000005

Von Stumm, S., Rimfeld, K., Dale, P. S., and Plomin, R. (2020). Preschool verbal and nonverbal ability mediate the association between socioeconomic status and school performance. *Child Dev.* 91, 705–714. doi: 10.1111/cdev.13364

Wagner, R. K., and Torgesen, J. K. (1987). The nature of phonological processing and its causal role in the acquisition of reading skills. *Psychol. Bull.* 101, 192–212. doi: 10.1037/0033-2909.101.2.192

Weber, C., Helm, C., and Kemethofer, D. (2021). Are social and ethnic Reading inequalities increasing during school closures?—the mediating role of parental involvement in distance learning. *Front. Educ.* 6:737064. doi: 10.3389/feduc.2021.737064

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

Wiese, R. (2006). The phonology of German. Oxford: Oxford University Press.

Wolf, M., and Bowers, P. G. (1999). The double-deficit hypothesis for the developmental dyslexias. *J. Educ. Psychol.* 91, 415–438. doi: 10.1037/0022-0663.91.3.415

Yulianti, K., Denessen, E., and Droop, M. (2023). The effects of parental involvement on children's education: a study in elementary schools in Indonesia. *Int. J. Parents Educ.* 10, 14–32. doi: 10.54195/ijpe.14123

Zaretsky, E., and Lange, B. P. (2017). Sociolinguistic factors associated with the subjectively and objectively measured language development in German preschoolers in three follow-up studies. *Linguistics* 55:0.1515/ling-2016-0036. doi: 10.1515/ling-2016-0036

Ziegler, J. C., and Goswami, U. (2006). Becoming literate in different languages: similar problems, different solutions. *Dev. Sci.* 9, 429–436. doi: 10.1111/j.1467-7687.2006.00509.x

frontiers | Frontiers in Education

TYPE Original Research
PUBLISHED 30 October 2024
DOI 10.3389/feduc.2024.1380295

OPEN ACCESS

EDITED BY
Christoph Weber,
University of Education Upper Austria, Austria

REVIEWED BY
John Romig,
University of Texas at Arlington, United States
Diana Alves,
University of Porto, Portugal

*CORRESPONDENCE
Erica S. Lembke
✉ lembkee@missouri.edu

RECEIVED 01 February 2024
ACCEPTED 30 September 2024
PUBLISHED 30 October 2024

CITATION
Lembke ES, McMaster KL, Duesenberg-Marshall MD, McCollom E, Choi S, Shanahan E, An J, Sussman-Dawson K and Birinci S (2024) Data based individualization in early writing: the importance and measurement of implementation fidelity. *Front. Educ.* 9:1380295. doi: 10.3389/feduc.2024.1380295

Data based individualization in early writing: the importance and measurement of implementation fidelity

Erica S. Lembke[1]*, Kristen L. McMaster[2], McKinzie D. Duesenberg-Marshall[1], Elizabeth McCollom[3], Seohyeon Choi[2], Emma Shanahan[4], Jechun An[2], Katya Sussman-Dawson[5] and Seyma Birinci[2]

[1]University of Missouri, Columbia, MO, United States, [2]University of Minnesota Twin Cities, St. Paul, MN, United States, [3]Southern Methodist University, Dallas, TX, United States, [4]The University of Texas at Austin, Austin, TX, United States, [5]University of Missouri–St. Louis, St. Louis, MO, United States

In this paper we describe the process of monitoring fidelity of implementation for a teacher-implemented early writing intervention. As part of a large, federally funded project, teachers who worked with students in grades 1 through 3 in schools across two states in the US were recruited and then randomly assigned to implementation and control conditions. Using Data-Based Individualization (DBI) as a framework for best practice in assessment and intervention, teachers in the implementation group received professional development on early writing intervention and assessment and then implemented these practices with their students who had significant writing challenges. Coaches, who were part of the research project, supported teachers and also observed teachers in both the implementation and control conditions at least twice during the course of the 20-week study. This paper focuses on the results of the fidelity measures that were administered throughout the project. An overview of the importance of fidelity checks is followed by a description of the fidelity tools used, as well as data from those tools. Areas of strength and challenge for teachers when implementing early writing assessment and intervention and engaging in data-based decision making with fidelity are discussed, along with recommendations regarding the practical and research importance of fidelity checks.

KEYWORDS

writing, data-based, intervention, fidelity, elementary

Introduction

Early intervention in writing instruction is of increasing importance (Berninger et al., 2008). This is underscored by the current writing performance of students in the United States. Data from national assessments in the U.S. consistently show that ∼75% of fourth, eighth, and 12th-grade students perform below proficient levels in writing (Irwin et al., 2021). Through early intervention, educators can help lay a strong foundation for developing writing and literacy skills. This foundation supports overall academic success and equips students with essential communication and critical thinking skills that extend beyond school.

Intervening in students' writing development during their formative years is crucial in part because it is related to their reading skills (Fitzgerald and Shanahan, 2000; Kendeou et al., 2009; Kim, 2020; Kim et al., 2015; National Early Literacy Panel, 2008). The interplay between reading and writing underscores the significance of early intervention because improvements in one domain can catalyze advancements in the other (Wanzek et al., 2017). Furthermore, early intervention in writing harnesses the power of writing as a learning tool. It enables children to establish connections between ideas, fostering critical thinking and encouraging them to explore and apply new concepts (McMaster et al., 2017). Early intervention not only can empower children to approach their education with confidence and curiosity, it may also reduce the number of students who need special education services under Specific Learning Disabilities (SLD; Lovett et al., 2017). Moreover, this early exposure to writing instruction sets the stage for students to excel in school and in the diverse life contexts they will encounter, from postsecondary education to the professional world (Diamond et al., 2008; Graham and Perin, 2007).

While the majority of students respond positively to standard intervention protocols, a small proportion requires more intensive, individualized instruction (Wanzek and Vaughn, 2009). This need for precision in implementation becomes particularly challenging for educators, given the potential hurdles arising from insufficient preparation (e.g., Boardman et al., 2005; Roehrig et al., 2008; Stecker et al., 2005). This is compounded by evidence that teachers often struggle to implement interventions as prescribed (Groskreutz et al., 2011; Noell et al., 2000; Sanetti et al., 2015, 2018). Thus, the overarching success of data-based individualization in early writing research and instruction lies in ensuring fidelity in its implementation, bridging the gap between intention and execution for the benefit of individual students.

Supporting teacher implementation of writing interventions

To address underachievement in writing, it is crucial for teachers to identify writing difficulties early and use systematic data-based approaches to align students' needs with evidence-based interventions. Data-based individualization (DBI) emerges as a systematic approach for customizing educational strategies for students with significant learning needs (Deno and Mirkin, 1977; Fuchs et al., 2010; National Center on Intensive Intervention, 2024). DBI involves targeted instruction, weekly progress monitoring, and data-based decision-making.

Data-based individualization (DBI), as outlined by Deno and Mirkin (1977), Fuchs et al. (2010), and the National Center on Intensive Intervention (2024), provides a structured framework to support teachers in tailoring instruction for students facing challenges in specific academic skills. DBI encompasses a systematic approach, employing evidence-based diagnostic assessments, targeted interventions, and consistent progress monitoring. This process enables educators to make data-driven decisions pertaining to student development and to implement necessary adjustments in their instructional strategies (National Center on Intensive Intervention, 2024). To ensure effective, targeted instruction and inform adaptations that maximize skill development, DBI is a vital tool for educators (Lembke et al., 2018). Research suggests DBI offers structure and guidance for teachers and has been effective at improving student outcomes as compared to business-as-usual or interventions implemented without considerations of student progress (Jung et al., 2018; Stecker et al., 2005).

The Early Writing Project

In the area of early writing and DBI framework utilization, tools were developed to support teachers' implementation of early writing interventions. This larger study was called Data-Based Individualization–Tools, Learning, Coaching ("DBI-TLC"). DBI-TLC is a comprehensive professional development system that incorporates several components to facilitate student growth in early writing. The DBI-TLC system was implemented through *The Early Writing Project*, an Institute of Education Sciences (IES)-funded project to support educators' implementation of intensive early writing interventions and adherence to the DBI process. *The Early Writing Project* has been implemented in various educational environments, with evidence of feasibility and usability (Lembke et al., 2018; Poch et al., 2020). It has enhanced teacher-related outcomes on DBI in early writing and has shown promising signs of improved student outcomes (McMaster et al., 2020). Multiple fidelity tools were created as part of the project to monitor and support effective implementation.

Learning modules

One key element of DBI-TLC is the training of teachers through learning modules in which they are provided with extensive training and practice of various tools, including instructional materials, assessment resources, and data analysis techniques. These tools assist teachers in gaining a deep understanding of their students' writing abilities, allowing for tailored and precise instruction. DBI-TLC emphasizes the importance of continuous learning and knowledge-building for educators through these learning modules. It recognizes that teachers require a solid foundation of expertise and skills to effectively support their students (McMaster et al., 2020). Therefore, DBI-TLC offers comprehensive educational resources to equip teachers with the necessary knowledge and competencies.

Coaching

Furthermore, ongoing coaching support is integral to the DBI-TLC process. Coaches provide educators with the guidance and assistance they need to adapt and improve their instructional practices. This coaching and support ensures that teachers can effectively address their students' intensive early writing needs over time. However, it is important to note that coaching support is not only an integral part of the DBI process, but also integral to maintain treatment integrity of the intervention and assessment. Continual support for implementation of interventions through coaching, ongoing PD, and performance feedback has been shown to encourage the maintenance of high levels of intervention fidelity (Codding et al., 2008; Saunders et al., 2021). Thus, fidelity is a crucial component of the Early Writing Project and the DBI-TLC materials to support teacher assessment, instructional, and decision-making alignment.

Role of fidelity

The first step in the DBI process, as outlined by the National Center for Intensive Instruction (intensiveintervention.org), is to ensure that the chosen and validated intervention is delivered with fidelity (National Center on Intensive Intervention, 2024). Fidelity, also known as implementation fidelity, in academic intervention is essential for effective educational practices, ensuring that prescribed procedures are closely followed (Durlak and DuPre, 2008; McMaster et al., 2017; Sanetti and Collier-Meek, 2015). Procedural fidelity pertains to how closely educators adhere to intended programs, assessments, and implementation plans (e.g., dosage, frequency) including student responsiveness and program differentiation (Dane and Schneider, 1998). Within intervention implementation, there is a widespread consensus on three factors that influence intervention outcomes: (1) adherence to the prescribed intervention, (2) exposure or the frequency and duration of the intervention, and (3) the quality of how well the intervention was delivered (Sanetti and Kratochwill, 2009).

When interventions and assessments are implemented with fidelity, intervention teams can make more accurate and informed decisions regarding individual students' progress and their need for intervention modifications or intensifications (Dane and Schneider, 1998; O'Donnell, 2008; Swanson et al., 2013). Without fidelity, it can become challenging to establish a clear link between student outcomes and the provided instruction, which hinders the ability to assess the effectiveness of interventions, emphasizing the importance of collecting this data in education research (Durlak, 2015; Guo et al., 2016; Sanetti and Kratochwill, 2009; Webster-Stratton et al., 2011).

Research findings indicate that when implemented with fidelity, interventions show average effect sizes that are two to three times greater than when fidelity is not monitored (Durlak and DuPre, 2008). Fidelity is vital for identifying whether a student requires more intensive support and forms the foundation for the DBI process (Lemons et al., 2014). Without implementation fidelity, the collected data may not effectively support appropriate individualization, as the lack of progress may be attributed to an ineffective intervention or an intervention not appropriately delivered (Collier-Meek et al., 2013; McKenna et al., 2014). Furthermore, fidelity data can indicate difficulties practitioners or interventionists have during the intervention implementation process (Knoche et al., 2010). This data can serve as a facilitator in promoting dialogue regarding potential barriers and modifications needed to best support the generalizability of the intervention (Gadke et al., 2021; Kimber et al., 2019; William McKenna and Parenti, 2017).

In educational research, the reporting of implementation fidelity is of increasing importance. A review of special education research conducted between 1995 to 1999 found 18% of studies reported implementation fidelity (Gresham et al., 2000), while in a review of general and special education research conducted between 2005 to 2009 found 47% studies conducted fidelity checks (Swanson et al., 2013). Further, in the field of school psychology, a review of literature from 1995 to 2008 found 37% reported treatment fidelity (Sanetti et al., 2011), and a more recent study of the same school psychology journals found that now a majority of the research incorporated treatment fidelity (72.8%; Sanetti et al., 2020). While it is gaining traction in educational research, making fidelity monitoring accessible and easily implemented for teachers is of utmost importance (McKenna et al., 2014).

We believe it's critically important to highlight fidelity to readers because of the strength of intervention that can be fully realized when implemented as intended. In this study, we did a thorough job of identifying areas where having fidelity data would be important, creating measures, and collecting and assessing data. One of the important considerations for our research team was not only the data, but how we used that data to provide feedback to teachers to improve their practice. This provision of feedback is a critical piece of data-based decision making. In this paper, we highlight the different types of fidelity data that we collected and how that data was collected in order to demonstrate how we used the data both as a way to support the internal validity of our study as well as to support teachers as part of the professional development we were providing.

Fidelity in *The Early Writing Project*

Given that the fidelity of implementation of a validated intervention program serves as the foundation of the DBI process and promotes the generalizability of an intervention, it is imperative to have fidelity checks integrated into the process (McMaster et al., 2020). To support teachers' fidelity to DBI implementation, fidelity assessment instruments, modified from the Accuracy of Implementation Rating Scales (AIRS; Fuchs et al., 1984), were created by principal investigators (PIs). These fidelity forms were designed to assess essential teacher and student actions expected following the Professional Development learning modules, including implementation in CBM Administration, Writing Instruction, and Decision Making. The forms (provided in the Appendices) were created to provide feedback to teachers, but perhaps more importantly, to ensure that the research procedures were implemented as intended. Fidelity observations were completed for each DBI component as well as for each TLC component. In the methods section, we describe the development of the fidelity tools as they were integrated into *The Early Writing Project*. While we focus on early writing in this paper, the fidelity checks we describe would be important to implement in any high-quality study where implementation as intended is critical. In this paper, we focus on two of the three factors that influence intervention outcomes: (1) adherence to the prescribed intervention and (2) the quality of how well the intervention was delivered (Sanetti and Kratochwill, 2009). It was beyond the scope of this project to collect exposure data systematically.

In an effort to highlight the importance of fidelity, in this paper, we report the results of the fidelity of implementation checklist data that was collected during observations conducted for *The Early Writing Project*. This fidelity data included teacher implementation of CBM administration and writing instruction; teacher decision making; implementation of professional development learning modules; and implementation of coaching. The research questions that guided this portion of the study included the following:

1. At what level did teachers implement DBI including CBM, writing instruction, and DBDM, and did fidelity vary by cohort?

2. At what level of fidelity was professional development and coaching implemented?

Method

Setting and participants

The Early Writing Project study was a randomized control trial conducted in 23 urban, suburban, and rural public school districts in two Midwestern states. The study took place over three academic years with three cohorts of teacher participants: 2018–2019 (Cohort 1), 2019–2020 (Cohort 2), and 2021–2022 (Cohort 3). Cohort 3 was postponed a year due to the COVID-19 pandemic. The methods and results of this randomized control trial are reported in detail in McMaster et al. (2024).[1]

To be eligible for participation in the study, teachers needed to provide direct writing instruction to students with significant writing difficulties. We only included teachers who had at least 2 years of teaching experience, given the time commitment that project participation required. Out of 154 teachers who were recruited and randomized to treatment and control conditions across 3 years, a total of 132 teachers completed the study (n = 68 treatment; n = 64 control). Of the 154 randomized teachers, most were female (96.1%), White (92.2%), and special educators (89.6%). Teachers had an average of 11.35 years of teaching experience (SD = 8.49). For Cohorts 1 and 3, there were no statistically significant differences in teacher demographics between the treatment and control groups. In contrast, for Cohort 2, the control group had significantly more experience, measured by years in the current position and years teaching elementary school, compared to the treatment group. The difference in hours of professional development (PD) in CBM (Curriculum-Based Measurement; M range = 1.7–2.3) and data-based decision making (DBDM; M range = 2.1–2.5) between the treatment and control groups was not significant. However, there was a significant difference in PD in Writing assessment/instruction/intervention favoring the treatment group for Cohort 2 (M range = 1.1–1.9). Teacher participants nominated students for participation in grades 1–3 who had significant writing difficulties, access to the general education curriculum, functional English skills, and the ability to write one or more letters of the alphabet. The research team screened nominated students using two Curriculum-Based Measurement (CBM) tasks: word dictation and picture word, each administered for 3 min. Word dictation is designed to evaluate transcription skills at the word level. Through individual administration, the examiner dictates phonetically regular, grade 1 words from the Common Core State Standards for students to write. The examiner may only repeat each word once during the administration of the task. This task is scored using correct letter sequences (CLS), which requires that any two adjacent letters be placed correctly according to the correct spelling of the word. Picture word is designed to measure a students' sentence-level transcription and text generation skills and can be administered in a group or individually. Prior to administration of this CBM, the examiner models the task for the student(s). After practice, the students are instructed to write as many sentences as they can using the picture and corresponding word to generate ideas for their writing. There is a total of 12 picture prompts on each picture word CBM. Students are given 3 min for this task. The task is scored using correct word sequences (CWS), which is when any two words are spelled and used correctly in a sentence. We selected two to three students who scored the lowest on these measures among their teachers' nominated students. Of 365 eligible students who completed pretesting, 309 completed their participation in the study. Of eligible students, most were male (67.4%), White (29.3%), in second grade (55.3%), or received special education services (84.9%). Students' most frequent primary disabilities were other health impairment (21.1%), autism (20.0%) or specific learning disability (19.7%). On average, students scored in the "low" range (M = 68.67, SD = 14.22) on the Kaufman Test of Educational Achievement—Third Edition (KTEA-3; Kaufman and Kaufman, 2014) Written Expression subtest.

Study design and conditions

The Early Writing Project study was a randomized control trial where teachers were randomly assigned within school to either a treatment or business-as-usual (BAU) control condition, and students were nested within teachers.

Treatment teachers received a DBI-TLC package which included instruction, assessment, and decision-making tools at the beginning of their participation to the study. They also received four training sessions, as well as bi-weekly coaching. Coaching meetings started around mid to late August and ended around late March or early April. Treatment teachers delivered interventions using mini-lessons and materials provided in DBI-TLC package, administered CBM and graphed student data weekly. Treatment teachers were instructed to implement writing interventions three times per week for at least 20–30 min per session. Teachers made an instructional decision for each of their target students after they collected eight progress monitoring data points, and they repeated decision-making every 4–6 weeks after first decision.

Business-as-usual teachers in the control condition did not have access to DBI-TLC during the study and continued implementing their typical direct writing instruction for their participating students. After the study was completed, BAU teachers received all DBI-TLC materials and participated in one 8-h training session that included abbreviated content from all learning modules.

DBI-TLC implementation fidelity

Given the multi-component nature of DBI-TLC, we measured teachers' fidelity of implementation for multiple study procedures. In this paper, we report fidelity data we collected related to teachers' writing instruction, CBM administration, and Data Based Decision Making. Additionally, we evaluated the fidelity of DBI-TLC procedures by measuring the research team's learning module and coaching fidelity. For each fidelity category, we

1 McMaster, K. L., Lembke, E. S., Shanahan, E., Choi, S., An, J., Schatschneider, C., et al. (2024). *Supporting Teachers' Data-Based Individualization of Early Writing Instruction: An Efficacy Trial* (under review).

outline its conceptualization, how it was measured, and procedures for measurement in the sections below. For fidelity forms, see Appendices A through G.

Writing instruction fidelity

In the pilot phase of DBI-TLC, we assessed teachers' implementation fidelity using an adapted version of the Accuracy of Implementation Rating Scales (AIRS), originally created by Fuchs et al. (1984), which measured adherence to core DBI procedures: writing instruction, CBM administration, and DBDM. However, we found that the pilot writing instruction fidelity tool did not meaningfully differentiate between the instruction of DBI and control teachers, suggesting that the intended impacts of DBI-TLC on instruction were not being adequately measured (Lembke et al., 2018). Evaluating writing instruction in the context of DBI presented a unique challenge, which was determining how to best capture fidelity when teachers were delivering individualized instruction. Additionally, our observations of DBI-TLC teachers' writing instruction suggested that our measure did not adequately capture dimensions of instructional quality, such as meaningful teacher-student interactions and student engagement. As a result, we revisited and refined our measure for the efficacy phase of our study.

The final writing instruction fidelity tool (see Appendix A) was a checklist of observed steps for adherence components of explicit and intensive writing instruction. These adherence components were scored using a scale of 0 to 2 (for single event observations: "not in-place" to "in-place and effective," and for reoccurring observations: "rarely or never observed" to "all or almost all of observed instances"). Total fidelity scores were observed item points divided by possible item points multiplied by 100. The writing instruction of both treatment and control teachers was observed, as eligible teacher participants needed to provide direct writing instruction to their students with significant writing difficulties. One aim of the updated fidelity to tool was to capture flexible, adapted writing instruction fidelity. Meaning, we accounted for the individualized nature of our treatment teachers' instruction (e.g., Johnson and McMaster, 2013). In the primary writing study where we collected data for the present study, teachers created individualized lesson plans by selecting from researcher-created activities. To evaluate fidelity to the individualized aspects of instruction, we created two to three fidelity items for each instructional activity. Observers would then use those items as part of their fidelity evaluation. We also included items intended to measure instructional quality, in addition to adherence, which were also scored on a scale of 0 to 2 (for single event observations: "not in-place" to "in-place and effective," and for reoccurring event observations: "rarely or never" to "all or almost all of observed instances"). These included items related to how students engaged with the instruction and the instructional environment.

Research team members evaluating teachers' writing instruction fidelity first reached 85% agreement with an expert evaluator on the research team. Fidelity of writing instruction was observed both informally and formally. These observations happened in person or, when necessary, using teacher-recorded videos of their writing instruction. Coaches informally observed the writing instruction of their assigned teachers once a month and provided feedback to support the teachers' implementation during coaching meetings. Graduate Research Assistants (GRAs) and Project Coordinators (PCs) also conducted formal observations twice, once in fall and once in spring, with teachers they were not coaching to ensure teachers were delivering intensive intervention as they had been trained during the Learning Modules. After observing, researchers reviewed writing instruction fidelity scoring with teachers and provided feedback on how to improve fidelity, if needed.

Control observations

Project staff observed control teachers twice during the study using the same fidelity of writing instruction tool used with treatment teachers and took detailed field notes. Observed activities ranged considerably (e.g., writing days of the week, sentence construction, writing responses to reading). A more detailed description of control teacher instruction is available in McMaster et al. (2024) (see text footnote 1).

Students were typically taught in small groups, for 8–50 min ($M = 25$ min). Fidelity of control teachers' writing instruction ranged from 38% to 97% ($M = 75\%$, $SD = 13.24$) across cohorts, sites, and time points.

CBM administration fidelity

The purpose of collecting CBM administration fidelity data was to ensure that treatment teachers were collecting reliable progress monitoring data for their students. The CBM administration fidelity tool was a checklist of steps, measuring adherence, required to administer one of three CBM for writing tasks (see Appendix B–D). Observations were conducted in person, via videorecording, or via audio recording. Items were scored as 0 (not observed) or 1 (observed). In cases of audio recording, where steps could not be observed (e.g., "Has materials on hand"), evaluators would mark the item as "N/A" for not applicable. Teachers also had the option of using shortened CBM directions once their students were familiar with the procedure—in these cases, items not included in shortened directions were also scored as "N/A." Total fidelity scores were observed item points divided by possible item points multiplied by 100.

Before assessing teachers' CBM administration fidelity, researchers were trained on how to administer the measures using the fidelity form. Researchers needed to demonstrate 85% fidelity during an administration rehearsal, and then again during several actual administrations with a student during pretesting for the study. Researchers formally observed teachers twice, once in the fall and once in the spring. After observing, researchers reviewed CBM administration fidelity scoring with teachers and provided feedback on how to increase fidelity, if necessary. If teachers demonstrated <80% fidelity during the administration, another CBM administration was observed to determine if fidelity had improved.

DBDM fidelity

Treatment teachers in the current study collected weekly CBM data and made one to three decisions per student over the course of the 20-week study, depending on the number of decision-making opportunities. The purpose of collecting decision-making data was to ensure that treatment teachers were making timely and/or appropriate decisions in response to student data. As they were making instructional decisions, teachers completed an online Qualtrics survey to record what decisions they made and why. A copy of these Google Sheets and a copy of progress monitoring graph for the corresponding student and decision was used to calculate DBDM fidelity. A fidelity tool (see Appendix E) created by the Principal Investigators was used to assess each teacher's decision-making fidelity. The form included the percentage of timely decisions, which measured DBDM adherence, appropriate decisions, which measured DBDM quality, and overall fidelity for each decision-making opportunity. *Timely* decisions were defined as those made after the first eight data points were collected and every 4–6 data points after that, and decisions were deemed *appropriate* if teachers raised the goal if the trend line was steeper than the goal line, continued with the same instruction if the trend line was on track with the goal, or changed instruction if the trend line was flatter than the goal line. All of the guidelines around timely and appropriate decisions were taught and practiced during the professional development learning modules and supported through coaching.

Fidelity of learning modules

Teachers participated in four learning modules throughout the study. During module development, key components that must be addressed through explanation, modeling, guided practice, and application during training were identified. While the trainings were delivered, an observer noted whether each component was observed or not. Fidelity of learning module adherence (see Appendix F for example) was recorded as the percentage of components observed during the trainings.

Fidelity of coaching

During biweekly coaching sessions with treatment teachers, coaches used a coaching conversation form to guide their meeting, following a general coaching procedure. All coaches audio recorded their two coaching sessions for fidelity purposes. Project Investigators (PIs) listened to the recordings and checked the fidelity of adherence to key components of coaching on a scale of 0–1 (see Appendix G), including rapport-building, review of objectives, review of DBI steps and goals, discussion of student data, and planning for next steps. Additionally, the PIs scored coaches on a scale of 1–4 for components of coaching quality, including how well the coach fostered a collaborative conversation with the teacher, how responsive the coach was to the teacher's needs, and how well the coach made appropriate and logical connections between challenges, solutions, and next steps (see Appendix G). Total coaching fidelity scores were calculated by dividing the number of observed points by the number of possible points and multiplied by 100.

Results

Our first research question addressed the level at which teachers implemented DBI including writing instruction, CBM, and DBDM, and whether that fidelity varied by cohort.

Table 1 shows the results of fidelity observations for writing instruction. When comparing the results by experimental condition, on average, treatment teachers demonstrated noticeably higher overall writing instruction fidelity (M = 90.77%, range = 66%−100%, SD = 6.45%) compared to that of control teachers (M = 74.55%, range = 38%−97%, SD = 13.30%). Lower scores were generally observed in the lesson setup and wrap-up components, though control teachers also demonstrated lower fidelity levels in other components. When examining cohorts and experimental conditions together, Cohort 1 treatment teachers demonstrated an average writing instruction fidelity of 91.37%, Cohort 2 reached 89.76%, and Cohort 3 achieved 91.41%. In contrast, control teachers in each cohort showed lower fidelity levels: 73.64% for Cohort 1, 72.25% for Cohort 2, and 77.26% for Cohort 3. These results reveal that, within the same experimental condition, the variation in fidelity levels across cohorts was negligible, with treatment teachers consistently maintaining high levels of overall fidelity in writing instruction.

Table 2 presents the results of fidelity observations for CBM administration conducted with treatment teachers. On average, when collapsing across cohorts and fidelity observation times, treatment teachers administered various CBM tasks with 92.56% fidelity (range = 59%−100%, SD = 10.2%). We did not observe any significant patterns or differences based on cohort or observation timeframe.

Table 3 displays the results of fidelity observations for decision-making, which were assessed only for treatment teachers. When collapsing across cohorts and fidelity evaluation time points, the overall fidelity for decision-making was 86.77% (range = 0%−100%, standard deviation = 27.2%). Breaking down the results into the adherence (i.e., timely) and quality (i.e., appropriate) of decision-making, fidelity of timely decision-making averaged at 85.95% (range = 0%−100%, SD = 33.05%), while the fidelity of appropriate decision-making averaged at 92.14% (range = 0%−100%, SD = 25%). Treatment teachers' fidelity in making appropriate and timely decisions showed a medium-size correlation (r = 0.42), each of which was highly correlated with overall decision-making fidelity (rs = 0.66 and 0.92, respectively). For treatment teachers, the correlations between their fidelity in CBM administration, writing instruction, and decision-making were marginal (−0.03 to 0.17) and not statistically significant, suggesting that a teacher showing high fidelity in one area did not necessarily show high fidelity in another.

Our second research question addressed the level of fidelity that professional development and coaching was implemented. In addition to evaluating the fidelity of teachers, it was crucial to assess the fidelity of the learning modules and coaching, as detailed in McMaster et al. (2024) (see text footnote 1). This was important to assure that our professional development and supports provided to teachers were implemented as intended. In addition, it was important to have uniformity in delivery of professional development and coaching given that the study was implemented across two sites.

Regarding the fidelity of the learning modules, for each module, assigned GRAs or PCs assessed the fidelity of presentation components and teacher activities using the research-developed checklist (see Appendix F). The score was determined by dividing the number of observed items by the total presentation

TABLE 1 Teacher writing instruction (WI) fidelity scores by cohort and condition.

	Time 1				Time 2				Time 1 + Time 2			
	n	*M*	*SD*	Range	*n*	*M*	*SD*	Range	*n*	*M*	*SD*	Range
All treatment	68	88.80	7.12	66–100	63	92.88	4.87	77–100	131	90.77	6.45	66–100
All control	58	73.12	14.57	38–96	40	76.62	11.05	50–97	98	74.55	13.30	38–97
All cohort 1	38	83.94	11.69	47–100	21	93.00	3.75	86–100	59	87.16	10.54	47–100
All cohort 2	48	78.79	14.91	38–97	45	84.73	12.41	50–98	93	81.66	14.01	38–98
All cohort 3	40	82.70	13.45	50–100	37	85.16	11.24	60–100	77	83.88	12.42	50–100
Treatment cohort 1	24	89.95	4.62	80–100	21	93	3.75	86–100	45	91.37	4.46	80–100
Treatment cohort 2	25	87.12	7.06	67–97	25	92.40	4.50	80–98	50	89.76	6.44	67–98
Treatment cohort 3	19	89.57	9.41	66–100	17	93.47	6.56	77–100	36	91.41	8.31	66–100
Control cohort 1	14	73.64	13.04	47–94	–	–	–	–	14	73.64	13.04	47–94
Control cohort 2	23	69.73	16.00	38–94	20	75.15	12.54	50–95	43	72.25	14.59	38–95
Control cohort 3	21	76.47	13.70	50–96	20	78.10	9.42	60–97	41	77.26	11.69	50–97
Total	126	81.58	13.62	38–100	103	86.57	11.16	50–100	229	83.82	12.79	38–100

TABLE 2 CBM administration fidelity scores of treatment teachers by cohort.

	Time 1				Time 2				Time 1 + Time 2			
	n	*M*	*SD*	Range	*n*	*M*	*SD*	Range	*n*	*M*	*SD*	Range
Cohort 1	24	92.08	8.46	73–100	24	91.95	10.56	64–100	48	92.02	9.46	64–100
Cohort 2	25	90.00	14.34	59–100	25	92.24	10.29	63–100	50	91.12	12.40	59–100
Cohort 3	19	96.89	3.74	90–100	19	93.42	9.31	61–100	38	95.15	7.21	61–100
Total	68	92.66	10.48	59–100	68	92.47	9.99	61–100	136	92.56	10.20	59–100

Only CBM administration fidelity scores of treatment group were measured in the larger project.

components/activities (23 for Module 1, 30 for Module 2, 30 for Module 3, and 28 for Module 4) and then multiplying the result by 100. Across different learning modules, sites, and cohorts, adherence fidelity scores ranged from 96 to 100%.

The fidelity of coaching was also assessed and monitored using a research-developed checklist (see Appendix G). The checklist included the observation of adhering to each specified component in the coaching conversation form (0 = not observed; 1 = observed), as well as an evaluation of coaching quality during the sessions on a 4-point scale (with 4 indicating high quality). The overall fidelity scores, calculated by dividing the total points awarded by the total possible score (23, including 11 binary observation items and three items rated on a 4-point scale), and then multiplied by 100%, ranged from 83 to 100% across different cohorts and sites. When separating adherence and quality of coaching, the average scores were 88% and 95%, respectively.

Discussion

While the majority of students respond to standard interventions, some students do not and require intensive, individualized instruction (Wanzek and Vaughn, 2009). Unfortunately, research suggests that teachers often have difficulty implementing interventions as prescribed (Noell et al., 2000; Sanetti et al., 2015, 2018). Data-based individualization (DBI) offers a systematic approach that educators can use to customize educational strategies for students who are non-responsive to intervention (Deno and Mirkin, 1977; Fuchs et al., 2010; National Center on Intensive Intervention, 2024). DBI encompasses targeted instruction, along with weekly progress monitoring and data-based decision-making. It is important to not only measure fidelity for intervention, but also for CBM administration, data-based decision-making, and any related PD or coaching provided to teachers. Assessing intervention fidelity also has implications for special education determination in the US. Many states in which an instructional model (i.e., Multi-Tiered Systems of Support or Response-to-Intervention) is used to identify students for special education under the category of a specific learning disability (SLD), a required component of eligibility is documenting that research-based interventions have been delivered with fidelity and that the student did not make adequate progress (Fletcher et al., 2007; Fuchs and Fuchs, 2006).

In the current study, we monitored the fidelity of two groups: teachers participating in our study and members of our team delivering professional learning and coaching. First and foremost, it was important for us to monitor the fidelity with which our teachers were implementing DBI for several reasons: (1) we wanted to support the teachers in implementing the best possible framework for their students; (2) we wanted the teachers to implement well

TABLE 3 Teacher decision making (DM) fidelity scores by cohort.

	Time 1				Time 2				Time 1 + Time 2			
	n	*M*	*SD*	Range	*n*	*M*	*SD*	Range	*n*	*M*	*SD*	Range
DM_timely												
Cohort 1	24	91.66	28.23	0–100	24	87.50	26.58	0–100	48	89.58	27.20	0–100
Cohort 2	24	79.16	41.48	0–100	19	94.73	15.76	50–100	43	86.04	33.31	0–100
Cohort 3	19	84.21	37.46	0–100	11	72.72	46.70	0–100	30	80.00	40.68	0–100
Total	67	85.07	35.90	0–100	54	87.03	29.43	0–100	121	33.05	85.95	0–100
DM_appropriate												
Cohort 1	24	95.83	20.41	0–100	24	83.33	31.85	0–100	48	89.58	27.20	0–100
Cohort 2	24	91.66	28.23	0–100	19	97.36	11.47	50–100	43	94.18	22.38	0–100
Cohort 3	19	94.73	22.94	0–100	11	90.90	30.15	0–100	30	93.33	25.37	0–100
Total	67	94.02	23.87	0–100	54	89.81	26.38	0–100	121	92.14	25.00	0–100
DM_Sum												
Cohort 1	24	93.75	22.42	0–100	24	81.25	28.78	0–100	48	87.50	26.29	0–100
Cohort 2	24	83.33	31.85	0–100	19	92.10	18.73	50–100	43	87.20	26.93	0–100
Cohort 3	19	86.84	32.66	0–100	11	81.81	25.22	50–100	30	85.00	29.79	0–100
Total	67	88.05	28.98	0–100	54	85.18	25.00	0–100	121	86.77	27.20	0–100

and with support so that they could sustain implementation after the study was complete; (3) to sufficiently test the use of DBI-TLC vs. business as usual, we needed to make sure that teachers were implementing the framework as intended. While it took some planning and a few trials to establish a good fidelity routine (including both the measures utilized and the system used for observation and feedback), it was time well spent in that the teachers demonstrated overall strong levels of fidelity for writing instruction, decision making, and CBM administration. Perhaps most important, when teachers did not demonstrate fidelity that was above 80% for a given observation, the coach would provide feedback to the teacher to help remediate the particular area of need and then a second observation was conducted after a brief time.

This pattern of observation and feedback could be replicated in a school if a fidelity checklist was developed (this could consist of the 5–10 top elements that are critical for the intervention or process to be effective) and either a video is collected for later observation, or a direct observation is conducted. A peer colleague could utilize the checklist to provide brief feedback on strengths and areas of need for the instruction. A cycle of observation and sharing feedback could be established with teacher pairs throughout the year. While this observation and feedback would take time, it would be time well spent in terms of enhancing instructional practices.

Our second set of fidelity observations were around our team's delivery of professional development through learning modules and our monitoring of our own coaching practices. These areas were important to maintain the integrity of the research practices but were also important so that we maximized teachers' time during the professional development and delivered coaching that was of the highest quality. Our teachers, as with most teachers, were very busy with lots of in school and outside of school obligations and activities. We wanted to provide coaching that was targeted and helpful, and that could support and improve their practices. So developing an effective coaching routine and training our coaches to utilize this routine was an important way to maximize teachers' time and provide high quality feedback.

Limitations

Several limitations of the current study highlight the need for more research on the fidelity of DBI-TLC. First, teachers were mostly female and white and students were mostly male and white, which limits the generalizability of the findings to other teacher and student populations. Furthermore, the current teacher sample was limited to those with at least 2 years of teaching experience. Less experienced teachers may have different needs and/or respond differently to coaching than more experienced teachers. Lastly, while the current study measured many aspects of fidelity, there were revisions to the initial writing instruction fidelity tool that capture instructional quality which may be more subjective than components of intervention adherence. For instance, there may be qualitative aspects that coaches had to make a decision on as they observed that may be more subjective. It was also not possible, given the nature of the project, to capture elements of exposure, which is a critical part of intervention fidelity (Sanetti and Kratochwill, 2009).

Directions for future research

Future studies should seek a more diverse sample that represents different racial, ethnic, and gender across both teachers and students to evaluate the effectiveness of DBI-TLC across different demographic groups. In addition, future studies may also want to include novice teachers to measure the effects across teachers with varying levels of experience. All treatment teachers in this study received bi-weekly coaching from a member of the research team. Follow-up studies may also want to explore ways in which to identify the appropriate frequency and duration of coaching for each teacher. For instance, teachers with a good understanding of DBI may need fewer coaching sessions than teachers with a more limited understanding. Another critical component that researchers may want to examine in the future is whether the skills and practices learned can be sustained without researcher or coaching support. Coaching in schools can be costly (Knight, 2012); therefore, it is important to explore the sustainability of DBI without coaching support.

Implications for practice

There are several implications for practice that can be drawn from the current study including those that relate to teacher PD, writing assessment, and writing intervention.

First, in the current study, treatment teachers were observed using the writing instruction fidelity tool that captured both individualized and flexible writing instruction components. Fidelity of writing instruction should include components of adherence (i.e., to what extent was the component implemented) and quality (i.e., how well the component was implemented), according to research (Collier-Meek et al., 2013; Sanetti and Collier-Meek, 2015). In the current study, treatment teachers were observed by their coach who utilized a fidelity tool, which was then used to inform feedback on what components of the intervention were delivered with fidelity and which could be improved. Overall, treatment teachers had higher levels of fidelity for writing intervention than control teachers. Therefore, one assumption might be that educators may want to consider utilizing fidelity checks as an opportunity to inform feedback on teacher instructional practices, especially when new instructional practices or interventions are introduced. It is important to remember that these teachers in the treatment group had also received Professional Development and Coaching, which could positively impact their implementation.

Second, intensive writing intervention should be informed by consistent CBM progress monitoring data. This allows teachers to monitor student progress and make appropriate instructional adaptations based on the data in a timely manner. In the current study, teacher CBM administration and scoring fidelity were measured. This was especially important given that all of the Early Writing Project CBMs are teacher-administered, scored, and graphed, unlike many reading and math CBMs, which are often computer-based. In the current study, teachers' fidelity was measured for both timely and appropriate decisions. Treatment teachers, with the support of their coach, were able to reach acceptable levels of fidelity for both timely and appropriate decisions, with teachers scoring higher for appropriate decisions, on average, compared to timely decisions. It is possible that real-life barriers, such as snow days, student absences, competing priorities, etc. somewhat hindered teachers' abilities to make timely decisions. Schools will want to plan ahead for these potential barriers to consistent progress monitoring and data-based decision-making.

Finally it is important for schools to invest in ongoing professional development (PD) opportunities for teachers that focus on evidence-based instructional practices. During PD, it is important to ensure that all components of the training are delivered each time; therefore, it may be useful for those delivering PD to create a fidelity checklist for an observer use to measure fidelity.

Data availability statement

The datasets presented in this article are not readily available. Requests to access the datasets should be directed to Erica Lembke, lembkee@missouri.edu.

Ethics statement

The studies involving humans were approved by University of Missouri IRB, Columbia, MO and University of Minnesota IRB, Minneapolis, MN. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

EL: Writing – original draft, Writing – review & editing. KM: Writing – original draft, Writing – review & editing. MD-M: Writing – original draft, Writing – review & editing. EM: Writing – original draft, Writing – review & editing. SC: Writing – original draft, Writing – review & editing. ES: Writing – original draft, Writing – review & editing. JA: Writing – original draft, Writing – review & editing. KS-D: Writing – original draft, Writing – review & editing. SB: Writing – original draft, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. The research reported here was funded by the Institute of Education Sciences, U.S. Department of Education, through Grant R324A170101 to the University of Minnesota.

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

Author disclaimer

The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/feduc.2024.1380295/full#supplementary-material

References

Berninger, V. W., Nielsen, K. H., Abbott, R. D., Wijsman, E., and Raskind, W. (2008). Writing problems in developmental dyslexia: under-recognized and under-treated. *J. Sch. Psychol.* 46, 1–21. doi: 10.1016/j.jsp.2006.11.008

Boardman, A. G., Argüelles, M. E., Vaughn, S., Hughes, M. T., and Klingner, J. (2005). Special education teachers' views of research-based practices. *J. Spec. Educ.* 39, 168–180. doi: 10.1177/00224669050390030401

Codding, R. S., Livanis, A., Pace, G. M., and Vaca, L. (2008). Using performance feedback to improve treatment integrity of classwide behavior plans: an investigation of observer reactivity. *J. Appl. Behav. Anal.* 41, 417–422. doi: 10.1901/jaba.2008.41-417

Collier-Meek, M. A., Fallon, L. M., Sanetti, L. M., and Maggin, D. M. (2013). Focus on implementation: assessing and promoting treatment fidelity. *Teach. Except. Child.* 45, 52–59. doi: 10.1177/004005991304500506

Dane, A. V., and Schneider, B. H. (1998). Program integrity in primary and early secondary prevention: are implementation effects out of control? *Clin. Psychol. Rev.* 18, 23–45. doi: 10.1016/S0272-7358(97)00043-3

Deno, S. L., and Mirkin, P. K. (1977). *Data-Based Program Modification: A Manual.* Reston, VA: Council for Exceptional Children.

Diamond, K. E., Gerde, H. K., and Powell, D. R (2008). Development in early literacy skills during the pre-kindergarten year in Head Start: relations between growth in children's writing and understanding of letters. *Early Child. Res. Q.* 23, 467–478. doi: 10.1016/j.ecresq.2008.05.002

Durlak, J. A. (2015). Studying program implementation is not easy but it is essential. *Prev. Sci.* 16, 1123–1127. doi: 10.1007/s11121-015-0606-3

Durlak, J. A., and DuPre, E. P. (2008). Implementation matters: a review of research on the influence of implementation on program outcomes and the factors affecting implementation. *Am. J. Community Psychol.* 41, 327–350. doi: 10.1007/s10464-008-9165-0

Fitzgerald, J., and Shanahan, T. (2000). Reading and writing relations and their development. *Educ. Psychol.* 35, 39–50. doi: 10.1207/S15326985EP3501_5

Fletcher, J. M., Lyon, G. R., Fuchs, L. S., and Barnes, M. A. (2007). *Learning Disabilities: From Identification to Intervention.* New York, NY: Guilford.

Fuchs, D., and Fuchs, L. S. (2006). Introduction to response to intervention: what, why, and how valid is it?. *Read. Res. Q.* 41, 93–99. doi: 10.1598/RRQ.41.1.4

Fuchs, D., Fuchs, L. S., and Stecker, P. M. (2010). The "blurring" of special education in a new continuum of general education placements and services. *Except. Child.* 76, 301–323. doi: 10.1177/001440291007600304

Fuchs, L. S., Deno, S. L., and Mirkin, P. K. (1984). The effects of frequent curriculum-based measurement and evaluation on pedagogy, student achievement, and student awareness of learning. *Am. Educ. Res. J.* 21, 449–460. doi: 10.3102/00028312021002449

Gadke, D. L., Kratochwill, T. R., and Gettinger, M. (2021). Incorporating feasibility protocols in intervention research. *J. Sch. Psychol.* 84, 1–18. doi: 10.1016/j.jsp.2020.11.004

Graham, S., and Perin, D. (2007). A meta-analysis of writing instruction for adolescent students. *J. Educ. Psychol.* 99:445. doi: 10.1037/0022-0663.99.3.445

Gresham, F. M., MacMillan, D. L., Beebe-Frankenberger, M. E., and Bocian, K. M. (2000). Treatment integrity in learning disabilities intervention research: do we really know how treatments are implemented? *Learn. Disab. Res. Pract.* 15, 198–205. doi: 10.1207/SLDRP1504_4

Groskreutz, N. C., Groskreutz, M. P., and Higbee, T. S. (2011). Effects of varied levels of treatment integrity on appropriate toy manipulation in children with autism. *Res. Autism Spectr. Disord.* 5, 1358–1369. doi: 10.1016/j.rasd.2011.01.018

Guo, Y., Dynia, J. M., Logan, J. A., Justice, L. M., Breit-Smith, A., Kaderavek, J. N., et al. (2016). Fidelity of implementation for an early-literacy intervention: dimensionality and contribution to children's intervention outcomes. *Early Child. Res. Q.* 37, 165–174. doi: 10.1016/j.ecresq.2016.06.001

Irwin, V., Zhang, J., Wang, X., Hein, S., Wang, K., Roberts, A., et al. (2021). *Report on the Condition of Education 2021 (NCES 2021-144). U.S. Department of Education.* Washington, DC: National Center for Education Statistics. Available at: https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2021144 (accessed July 23, 2024).

Johnson, L. D., and McMaster, K. L. (2013). "Adapting research-based practices with fidelity: flexibility by design," in *Evidence-Based Practices*, 1st Edn, Vol. 26, eds. B. G. Cook, M. Tankersley, and T. J. Landrum (Leeds: Emerald Group Publishing Limited), 65–91. doi: 10.1108/S0735-004X(2013)0000026006

Jung, P. G., McMaster, K. L., Kunkel, A. K., Shin, J., and Stecker, P. M. (2018). Effects of data–based individualization for students with intensive learning needs: a meta–analysis. *Learn. Disabil. Res. Pract.* 33, 144–155. doi: 10.1111/ldrp.12172

Kaufman, A. S., and Kaufman, N. L. (2014). *Kaufman Test of Educational Achievement,* 3rd Edn. Bloomington, MN: NCS Pearson.

Kendeou, P., Van den Broek, P., White, M. J., and Lynch, J. S. (2009). Predicting reading comprehension in early elementary school: the independent contributions of oral language and decoding skills. *J. Educ. Psychol.* 101:765. doi: 10.1037/a0015956

Kim, Y. S., Al Otaiba, S., and Wanzek, J. (2015). Kindergarten predictors of third grade writing. *Learn. Individ. Differ.* 37, 27–37. doi: 10.1016/j.lindif.2014.11.009

Kim, Y. S. G. (2020). "Interactive dynamic literacy model: an integrative theoretical framework for reading-writing relations," in *Reading-Writing Connections. Literacy Studies*, Vol. 19, eds. R. A. Alves, T. Limpo, and R. M. Joshi (Cham: Springer). doi: 10.1007/978-3-030-38811-9_2

Kimber, M., Barac, R., and Barwick, M. (2019). Monitoring fidelity to an evidence-based treatment: practitioner perspectives. *Clin. Soc. Work J.* 47, 207–221. doi: 10.1007/s10615-017-0639-0

Knight, J. (2012). *High-Impact Instruction: A Framework for Great Teaching.* Corwin Press.

Knoche, L. L., Sheridan, S. M., Edwards, C. P., and Osborn, A. Q. (2010). Implementation of a relationship-based school readiness intervention: a multidimensional approach to fidelity measurement for early childhood. *Early Child. Res. Q.* 25, 299–313. doi: 10.1016/j.ecresq.2009.05.003

Lembke, E. S., McMaster, K. L., Smith, R. A., Allen, A., Brandes, D., Wagner, K., et al. (2018). Professional development for data-based instruction in early writing: tools, learning, and collaborative support. *Teach. Educ. Spec. Educ.* 41, 106–120. doi: 10.1177/0888406417730112

Lemons, C. J., Kearns, D. M., and Davidson, K. A. (2014). Data-based individualization in reading: intensifying interventions for students with significant reading disabilities. *Teach. Except. Child.* 46, 20–29. doi: 10.1177/0040059914522978

Lovett, M. W., Frijters, J. C., Wolf, M., Steinbach, K. A., Sevcik, R. A., Morris, R. D., et al. (2017). Early intervention for children at risk for reading disabilities: the impact of grade at intervention and individual differences on intervention outcomes. *J. Educ. Psychol.* 109, 889–914. doi: 10.1037/edu0000181

McKenna, J. W., Flower, A., and Ciullo, S. (2014). Measuring fidelity to improve intervention effectiveness. *Interv. Sch. Clin.* 50, 15–21. doi: 10.1177/1053451214532348

McMaster, K. L., Kunkel, A., Shin, J., Jung, P., and Lembke, E. S. (2017). Early writing intervention: a best evidence synthesis. *J. Learn. Disabil.* 51, 1–18. doi: 10.1177/0022219417708169

McMaster, K. L., Lembke, E. S., Shin, J., Poch, A. L., Smith, R. A., Jung, P. G., et al. (2020). Supporting teachers' use of data-based instruction to improve students' early writing skills. *J. Educ. Psychol.* 112:1. doi: 10.1037/edu0000358

National Center on Intensive Intervention (2024). Washington, DC: Office of Special Education Programs; U.S. Department of Education.

National Early Literacy Panel (2008). *Developing Early Literacy: Report of the National Early Literacy Panel.* Washington, DC: National Institute for Literacy. Available at http://www.nifl.gov/earlychildhood/NELP/NELPreport.html (accessed August 25, 2024).

Noell, G. H., and Witt, J. C. LaFleur, L. H., Mortenson, B. P., Ranier, D. D., LeVelle, J. (2000). Increasing intervention implementation in general education following consultation: a comparison of two follow-up strategies. *J. Appl. Behav. Anal.* 33, 271–284. doi: 10.1901/jaba.2000.33-271

O'Donnell, C. L. (2008). Defining, conceptualizing, and measuring fidelity of implementation and its relationship to outcomes in K−12 curriculum intervention research. *Rev. Educ. Res.* 78, 33–84. doi: 10.3102/0034654307313793

Poch, A. L., McMaster, K. L., and Lembke, E. S. (2020). Usability and feasibility of data-based instruction for Students with intensive writing needs. *Elem. Sch. J.* 121, 197–223. doi: 10.1086/711235

Roehrig, A. D., Bohn, C. M., Turner, J. E., and Pressley, M. (2008). Mentoring beginning primary teachers for exemplary teaching practices. *Teach. Teach. Educ.* 24, 684–702. doi: 10.1016/j.tate.2007.02.008

Sanetti, L. M. H., Charbonneau, S., Knight, A., Cochrane, W. S., Kulcyk, M. C., Kraus, K. E., et al. (2020). Treatment fidelity reporting in intervention outcome studies in the school psychology literature from 2009 to 2016. *Psychol. Sch.* 57, 901–922. doi: 10.1002/pits.22364

Sanetti, L. M. H., and Collier-Meek, M. A. (2015). Data-driven delivery of implementation supports in a multi-tiered framework: a pilot study. *Psychol. Sch.* 52, 815–828. doi: 10.1002/pits.21861

Sanetti, L. M. H., Collier-Meek, M. A., Long, A. C., Byron, J., and Kratochwill, T. R. (2015). Increasing teacher treatment integrity of behavior support plans through consultation and implementation planning. *J. Sch. Psychol.* 53, 209–229. doi: 10.1016/j.jsp.2015.03.002

Sanetti, L. M. H., Gritter, K. L., and Dobey, L. M. (2011). Treatment integrity of interventions with children in the school psychology literature from 1995 to 2008. *Sch. Psych. Rev.* 40:72. doi: 10.1080/02796015.2011.12087729

Sanetti, L. M. H., and Kratochwill, T. R. (2009). Toward developing a science of treatment integrity: Introduction to the special series. *Sch. Psych. Rev.* 38:445.

Sanetti, L. M. H., Williamson, K. M., Long, A. C., and Kratochwill, T. R. (2018). Increasing in-service teacher implementation of classroom management practices through consultation, implementation planning, and participant modeling. *J. Posit. Behav. Interv.* 20, 43–59. doi: 10.1177/1098300717722357

Saunders, A. F., Wakeman, S., Cerrato, B., and Johnson, H. (2021). Professional development with ongoing coaching: a model for improving educators' implementation of evidence-based practices. *Teach. Except. Child.* doi: 10.1177/00400599211049821

Stecker, P. M., Fuchs, L. S., and Fuchs, D. (2005). Using curriculum-based measurement to improve student achievement: review of research. *Psychol. Sch.* 42, 795–819. doi: 10.1002/pits.20113

Swanson, E., Wanzek, J., Haring, C., Ciullo, S., and McCulley, L. (2013). Intervention fidelity in special and general education research journals. *J. Spec. Educ.* 47, 3–13. doi: 10.1177/0022466911419516

Wanzek, J., Petscher, Y., Otaiba, S. A., Rivas, B. K., Jones, F. G., Kent, S. C., et al. (2017). Effects of a year long supplemental reading intervention for students with reading difficulties in fourth grade. *J. Educ. Psychol.* 109:1103. doi: 10.1037/edu0000184

Wanzek, J., and Vaughn, S. (2009). Students demonstrating persistent low response to reading intervention: three case studies. *Learn. Disabil. Res. Pract.* 24, 151–163. doi: 10.1111/j.1540-5826.2009.00289.x

Webster-Stratton, C., Reinke, W. M., Herman, K. C., and Newcomer, L. L. (2011). The incredible years teacher classroom management training: the methods and principles that support fidelity of training delivery. *Sch. Psych. Rev.* 40, 509–529. doi: 10.1080/02796015.2011.12087527

William McKenna, J., Parenti, M. (2017). Fidelity assessment to improve teacher instruction and school decision making. *J. Appl. Sch. Psychol.* 33, 331–346. doi: 10.1080/15377903.2017.1316334

frontiers | Frontiers in Education

TYPE Brief Research Report
PUBLISHED 30 October 2024
DOI 10.3389/feduc.2024.1376434

Check for updates

OPEN ACCESS

EDITED BY
Christoph Weber,
University of Education Upper Austria, Austria

REVIEWED BY
Patrícia Botelho Silva,
Mackenzie Presbyterian University, Brazil
Ferenc Kemény,
University of Graz, Austria

*CORRESPONDENCE
Joana Wolfsperger
✉ Joana.Wolfsperger@edu.lmu.de

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

RECEIVED 25 January 2024
ACCEPTED 01 October 2024
PUBLISHED 30 October 2024

CITATION
Wolfsperger J and Mayer A (2024) A brief research report on the efficacy of a RAN training in elementary school age children. *Front. Educ.* 9:1376434.
doi: 10.3389/feduc.2024.1376434

A brief research report on the efficacy of a RAN training in elementary school age children

Joana Wolfsperger*† and Andreas Mayer†

Faculty of Psychology and Educational Sciences, Department of Education and Rehabilitation, Institute for Speech and Language Pathology, Ludwig-Maximilians-University Munich, Munich, Germany

Numerous studies could prove that children with reading disorders show RAN deficits and that RAN is especially strongly connected to the automatization of word recognition. In contrast intervention studies so far have provided little convincing evidence for the trainability of RAN or even a transfer effect on reading competence. The aim of this study is therefore to evaluate the efficacy of an adaptive, software-based RAN-training in a group of 57 children aged 9.3 years (SD 1.10 years) with a RAN-Deficit and/or difficulties with word-reading-speed and to determine if there is a transfer effect to reading speed of words and pseudowords as well as to reading comprehension on the word-, sentence and text level. The software-based training of RAN letters, numbers and colors is an intervention over 18 training days. Children were encouraged to name visually presented items in a serial format as fast as a red rectangle jumping from item to item. The speed of the red rectangle was successively increased with each successful training session. Univariate ANOVAs with repeated measures showed that children had a significantly higher naming speed for letters and numbers both immediately after completion of the intervention and in the follow-up measures. A transfer of the training effect to reading speed for words, reading comprehension on the word- and sentence-level, but not to reading speed for pseudowords and text comprehension, could be demonstrated. Due to the missing of a control-group comparison, it remains to be proven in subsequent research whether this effect is a maturation-related change or a training effect.

KEYWORDS

Rapid Automatized Naming (RAN), software-based training, elementary school age, reading fluency, reading comprehension

1 Theoretical positioning

1.1 "Phonological-core-deficit-hypothesis" and "double-deficit-hypothesis"

Developmental Dyslexia, i.e., a specific disorder in reading that cannot be accounted for by low IQ, poor educational opportunities or obvious sensory or neurological damage (World Health Organization, 2008), affects about 6–8% of the population (Shaywitz et al., 1990; Galuschka and Schulte-Körne, 2016).

Reading Researchers generally agree that deficits in the processing of phonological information underlie a large portion of specific reading disorders (Galuschka et al., 2020). According to Wagner and Torgesen (1987), the abilities to perceive, consciously manipulate, store, and process information about the sound structure of language or to access phonological representations in long-term memory during the production and processing of spoken and written language are allocated to phonological processing. Accordingly, phonological processing includes phonological awareness, phonological working memory, and Rapid Automatized Naming (RAN). In German-speaking countries, when considering the fundamental (meta-)linguistic-cognitive competencies for successful literacy acquisition, the construct of phonological awareness, i.e., the conscious identification, analysis, synthesis, and manipulation of linguistic units at the sublexical level (Mayer, 2021), is the focus of attention, while RAN-deficits and resulting difficulties with automatized word recognition and reading fluency are hardly considered. RAN is the ability to visually process a sequence of simultaneously visible familiar symbols (e.g., letters, numbers, colors) as quickly as possible, and to name the corresponding word (or phone) (Mayer, 2021).

While the importance of phonological awareness on the acquisition of reading and spelling is not disputed by proponents of the "double-deficit hypothesis," Wolf and Bowers (1999) for example postulate that phonological awareness and RAN each make specific contributions to explaining differences in written language skills despite correlative relationships of moderate magnitude, and serve largely independently as predictors of reading and spelling difficulties (Mayer, 2014; Torppa et al., 2013; Vaessen and Blomert, 2010; Swanson et al., 2003; Wolf et al., 2002). The influence of RAN on reading should not be reduced to its phonological component, i.e., the speed of access to phonological representations in long-term memory; rather, a naming speed deficit represents a second "core-deficit" of reading difficulties that is at least partly independent of phonological skills (Wolf et al., 2002).

Evidence for this assumption can be derived from empirical surveys of the "double-deficit hypothesis." Children with below-average performance in both functions ("double deficit") usually experience significantly greater difficulties with reading as compared to children with isolated weaknesses in phonological awareness or naming speed (Furnes et al., 2019; Araújo and Faísca, 2019; Ozernov-Palchik et al., 2017; Cronin, 2013; Mayer, 2014; Torppa et al., 2013; Kirby et al., 2003). Moreover, the two functions appear to be associated with different written literacy subskills to varying degrees. Research findings from countries with different levels of orthographic transparency suggest that phonological awareness is primarily correlated with reading accuracy and spelling, while RAN is the best predictor for the development of reading automaticity and reading fluency for words and pseudowords (da Silva et al., 2020; Furnes et al., 2019; Mayer, 2018; Moll et al., 2014; Kirby et al., 2003, 2010; Georgiou et al., 2008, 2009; Wimmer and Mayringer, 2002).

The fact that naming speed is largely independent of phonological information processing is justified in the theory by the complexity of the RAN-construct. In addition to the speed of access to the phonological lexicon, visual processing speed and discriminative ability as well as executive functions (attention control) are involved. Finally, a finely tuned, highly automated interplay of the underlying perceptual, lexical, phonological, and cognitive components as well as the fastest possible execution of the subprocesses is necessary (Wolf and Bowers, 1999).

Research studies have shown that children with reading difficulties perform significantly worse on RAN Tests than do children with typical reading and writing skills (Araújo and Faísca, 2019; Mayer, 2014, 2018).

In a meta-analysis that included data from a total of 22,418 children and 216 comparisons between children with average and below-average reading and spelling skills, Araújo and Faísca (2019) found an average effect size of $d = 1.19$, meaning that children with reading difficulties performed more than one standard deviation worse on RAN-Tests than children of the same age with typical reading skills.

Given that word recognition speed differentiates best between poor and average readers in countries with relatively transparent orthographies, and that impairments in reading fluency are the core problem of German-speaking children with poor reading skills (Wimmer, 1993), which in turn are significantly related to reading comprehension, RAN is crucial for the early detection, diagnosis and intervention of (impending) reading and spelling difficulties in German-speaking countries.

1.2 Explaining the relationship between RAN and reading

Various hypotheses explain the relationship between RAN and literacy, suggesting multiple factors may be involved (Kirby et al., 2010). The phonological deficit hypothesis explains RAN's influence on reading via its component of speed of access to phonological representations in long-term memory.

An alternative explanation interprets the (impaired) speed of access to phonological entries as the consequence of a global processing-speed-deficit (Kail and Hall, 1994), a specific visual processing-speed-deficit (Stainthorp et al., 2010) or an impaired "timing mechanism" (Wolf and Bowers, 1999), which sets a limit to rapid and automated execution and fine-tuned interplay of underlying executive functions. Finally Bowers et al. (1994) suggest that a RAN deficit affects reading through slower visual processing of the single letters in a word or delayed activation of the corresponding phonemes. This delay hinders the recognition and storage of frequently co-occurring letters as orthographic patterns, crucial for word recognition. Affected children may remain at the stage of phonological recoding and require specific support focusing on holistic processing of orthographic patterns for an automation of word recognition.

Assuming that RAN and phonological awareness are two functions that influence word recognition largely independently and that RAN cannot be reduced to the phonological components, it can be expected that children whose reading difficulties are associated with a RAN deficit do not benefit sufficiently from phonologically oriented support, but rather need a RAN-training.

1.3 Approaches for training RAN

To date however, there is hardly any empirical evidence for the effectiveness of training RAN.

For example, Berglez (2003) attempted to improve preschoolers' naming speed by asking them to name printed colors and simple objects on the same templates at high frequency as quickly as possible. The fact that the naming speed of the trained children and a control group improved to a comparable extent over a period of 6 months is interpreted as evidence that the improvement is less a result of the training but more a consequence of developmental progress due to maturation. A similar conclusion was drawn by de Jong and Vrielink (2004), who failed to improve rapid letter naming in a group of Dutch first graders after 10 days of daily training comparable to Berglez (2003). Conrad and Levy (2011) were also unable to demonstrate an increase in RAN, let alone effects on reading and spelling performance, by training children with a naming speed deficit through high-frequency letter naming.

Initial evidence for the trainability of RAN with positive effects on reading fluency are provided by Vander Stappen and van Reybroeck (2018) and Pecini et al. (2019). Vander Stappen and van Reybroeck (2018) trained 18 s-grade children from Belgium on the naming speed of objects, while training in the control group ($n = 18$) focused on phonological awareness. In the RAN training group, participants were motivated in a playful way to name printed objects of different word lengths and syllable structures of varying complexity as quickly as possible. The children in the RAN training group improved their naming performance significantly more than the control group. Moreover, a transfer of RAN training to reading speed was demonstrated.

Pecini et al. (2019) offered a RAN-training for objects and colors with the help of a software-based program to 22 children with impaired naming speed and reading difficulties between the second and the fifth grades. The intervention was delivered exclusively at home under the supervision of caregivers over a three-month period, with sessions occurring three to five times per week and lasting approximately 10 minutes. In a comparison group ($n = 22$) students were asked to read texts with the syllables of each word visually highlighted one at a time. The results demonstrate the effectiveness of the RAN training. Both groups were able to significantly increase their reading speed at the word, sentence, and text levels. There were no detectable differences between the two conditions. The authors interpret their finding as evidence of the effectiveness of RAN training, whose particular advantages include the simple implementation in a family context.

2 Summary and aims of the study

With the above in mind, an adaptive, software-based training program for the improvement of naming speed was developed in cooperation with the company Trout GmbH (Kassel, Germany) for the research project documented here. It was carried out with 57 students with reading difficulties and a naming-speed-deficit between January and July 2023, and it was evaluated with regard to its effectiveness and its impact on word recognition and reading comprehension on the word-, sentence- and text-level.

The intervention study sought to answer the following research questions:

(1) Can the adaptive, software-based training improve naming speed?
(2) To what extent do improvements in RAN performance positively impact reading speed and reading comprehension?

3 Method

3.1 Sample

Three special education centers (schools for children with special educational needs in the areas of learning, language, emotional, and social development) and four elementary schools from a metropolitan area in Southern Germany agreed to participate in the study. The training group sample consists of students for whom parental consent was obtained: 27 children (47.4%) from elementary schools and 30 children (52.6%) from special education centers participated in the training. The children mainly attended grades 2 (elementary school) and 3 (special education center). Students from grades 4 and 5 ($N = 10$) accounted for a total percentage of 17.6%.

The students (54.5% female) had a mean age of 9.3 years (7.2–12.1 years, SD: 1.5 years).

3.2 Design

The study features a pre-/post-/follow-up design. At the first measurement date in December 2022 167 children were tested for RAN and reading speed using the standardized Test for Phonological Awareness and Naming Speed (TEPHOBE, Mayer, 2020) as well as the subtests real word reading and pseudoword reading of the Salzburg Reading and Spelling Test (SLRT II, Moll and Landerl, 2014). Reading comprehension at the word-, sentence-, and text-level was assessed using ELFE II (Lenhard et al., 2022) (see Measurements, Chapter 3.3).

All children who scored below-average (PR < 16) in at least one of the three RAN tests (letters, numbers, colors) and/or reading speed of real words in the pretest were included in the training group. The training was carried out twice a week and it took place at school during class hours. A training session usually lasted 10–15 min.

The posttest was administered immediately after the completion of the 18 training sessions. In the posttest, RAN, the reading speed, and the reading comprehension were tested once again. Three children had to leave the program early because they changed schools or moved. A follow-up test was conducted in July 2023. The purpose of this test was to demonstrate a possible long-term effect of the training. It once again assessed RAN, reading speed, and reading comprehension. 32 children took part in the follow-up test.

Both the measurement and the training were carried out by university students of special education and speech therapy (master's degree), who were thoroughly familiarized with the diagnostic tools and the training program.

The study was approved by the ethics committee of LMU Munich and by the responsible Bavarian district and school authorities.

3.3 Measures

To evaluate the potential effects of the training, performance in the rapid naming of letters, numbers, and colors was assessed using TEPHOBE (Test for Phonological Awareness and Naming Speed, Mayer, 2020). The children had to name five different numbers, letters, and colors each repeated 10 times and all presented simultaneously in a serial format as quickly as possible. The interpretation of the results is based solely on the time taken by the child to complete each subtest. Reliability for the naming speed tests in the TEPHOBE is in the satisfactory to good range (Cronbach's alpha = 0.78 and 0.86, respectively).

Both reading speed and reading comprehension were measured in order to evaluate the possible transfer of the training to reading skills.

The two "One-minute reading fluency tests" from the "Salzburg Reading and Spelling Test" (SLRT II, Moll and Landerl, 2014) were used to measure reading speed for words and pseudowords. In both subtests, students have 1 min each to read aloud a list of 156 (pseudo-) words. The list of pseudowords is designed to reveal possible difficulties in the use of phonological recoding, while the automatization of word recognition is measured via the reading speed for real words.

Reading speed [number of (pseudo-)words read correctly per minute] is the key metric for scoring and interpreting both reading tests. The parallel test reliability for the one-minute reading fluency tests is between $r = 0.90$ and 0.98 in the different age groups.

ELFE II (Lenhard et al., 2022) was used to assess reading comprehension, at word-, sentence-, and text-level in grades 1 to 7. Word comprehension is tested by assigning as many as possible correct words from four alternatives to a picture in 3 min. For the sentence comprehension test, children are presented with a sentence in which one word is missing. The gap has to be completed from a choice of five different alternatives (time allowed: 3 min). The text comprehension test contains short factual and narrative texts. Questions on the content of these texts must be answered in a single-choice format (time allowed: 7 min).

Both the values for split-half reliability and the retest reliability are in the "*very good*" range, with values between $r = 0.81$ and 0.96 for the word and sentence comprehension test and the overall test. The values for text comprehension are in a satisfactory to good range.

3.4 Training

The training measure used in this study is an adaptive, software-based RAN training in which the usual RAN matrices are presented on the screen in a serial format (five lines of ten items each). The letter category uses 13 different elements (M, T, E, S, P, F, U, O, R, A, L, N, B), the number category uses the numbers 0 to 9, and the color category uses red, green, blue, yellow, black, purple, and brown.

The intervention comprises 18 sessions lasting approximately 10–15 min each, including an assessment of the children's baseline performance in naming letters, numbers and colors on day 1.

The first 2 days of the training (day 2 and 3) familiarize the students with the format of the training: the children had to name the templates with all three categories on her own pace, while the number of different symbols per category successively were increased. From the following session on (day 4–20), participants complete three rounds per category with the software generating a new matrix for each trial, such that the order of the symbols is constantly changing during training.

In order to motivate the children to continuously speed up their processing, the presentation time of each symbol is reduced by 3 msec/item after each successful trial. For this purpose, the symbols are framed by a red rectangle that jumps (increasingly faster) from item to item, thus guiding the children's naming speed.

To determine changes in naming speed and reading competence as a result of the intervention, repeated measures ANOVAs were conducted with the three test times (pretest, posttest, follow-up test) as factors.

4 Results

4.1 Descriptive statistics

First, descriptive statistics for naming speed, reading speed, and reading comprehension performance at the time of the pre-, post- and follow-up test are presented to characterize the group (Table 1).

Deviations in the number of students participating at the time of the pretest are due to illness or refusal to complete individual tasks.

For naming speed, only raw values are available. TEPHOBE does not provide Percentile Ranks or T-Scores for children from the middle of second grade onwards. However, a comparison with the orientation values published by van Ermingen-Marbach et al. (2015) for the naming speed of children with an average age of 9.7 years (SD: 0.54) (RAN letters: 2.18 items/s [SD: 0.37], RAN numbers 1.93 items/s [SD: 0.39], RAN colors: 1.02 items/s [SD: 0.18]) shows that the values achieved by the participants in this study are clearly below average. Therefore, it can be assumed that this is indeed a sample with a RAN deficit.

Furthermore, at the time of the pretest, the group exhibited significantly below-average performance in terms of reading speed (words: PR 7.56 [SD: 7.08], pseudowords: PR 12.33 [13.55]) and reading comprehension (T-score: 30.91–37.29 [SD: 5.1–7.42]).

4.2 Inferential statistics

The results of the univariate repeated measures ANOVAs indicate significant differences for RAN letters [F (2, 60) = 35.16, $p < 0.01$, $\eta^2 = 0.54$], RAN numbers [F (2, 60) = 22.06, $p < 0.01$, $\eta^2 = 0.42$] but not for RAN colors [F (2, 60) = 2.99, $p = 0.06$] for the three test times. Bonferroni-corrected post-hoc tests show significantly ($p < 0.01$) higher values for RAN letters ($M_{Diff} = 0.51$ and 0.43; 95%-CI [0.66, 0.36], and [0.55, 0.30]) and RAN numbers ($M_{Diff} = 0.32$ and 0.27; 95%-CI [0.44, 0.21], and [0.38, 0.26]) at the posttest and follow-up test compared to the pretest, while no significant differences were found between the posttest and follow-up test. Thus, the participating students significantly improved their naming speed, for the categories of letters and numbers, and this effect persisted at the follow-up assessment. Furthermore, the calculated effect sizes (Cohen's d) indicate substantial differences between the pretest and posttest as well as between the pretest and the follow-up test. A large effect was found for RAN letters ($d = 1.22$–1.30) and RAN numbers ($d = 0.91$–0.98), while only a small effect was found for RAN colors ($d = 0.28$–0.37) (Table 1).

Additionally, a repeated measures ANOVA shows statistically significant differences in reading speed for real words across the three test times [F (2, 60) = 33.81, $p < 0.01$, $\eta^2 = 0.53$]. The results of the Bonferroni-corrected post-hoc tests indicate a significant difference between the pretest and posttest, as well as between the pretest and follow-up test ($M_{Diff} = 6.16$ and 6.36, 95%-CI [9.13, 3.19], and [9.13, 3.58], $p < 0.01$), while no significant differences were found between the posttest and follow-up test. No changes were found in the reading speed of pseudowords [F (2, 60) = 2.56, $p = 0.08$]. The calculated effect sizes (Cohen's d) indicate a medium to large effect for reading speed

TABLE 1 Descriptive statistics of the pre- post- and follow-up-test.

	Pretest (n = 57)	Posttest (n = 54)	Follow-up (n = 31)	Cohen's d (pre- post-test)	Cohen's d (pre- follow-up-test)
	MW (SD)	MW (SD)	MW (SD)		
RAN[1]					
RAN letters[2]	1.31 (0.32)	1.89 (0.43)	1.81 (0.29)	1.37**	1.22**
RAN numbers[2]	1.27 (0.33)	1.59 (0.43)	1.56 (0.26)	0.98**	0.91**
RAN colors[2]	0.75 (0.20)	0.83 (0.27)	0.90 (0.38)	0.28*	0.37*
Reading speed[3]					
Real words[4]	18.70 (10.41)	25.58 (10.77)	27.39 (10.87)	0.54**	1.04**
Pseudowords[4]	16.37 (7.29)	20.11 (6.63)	20.77 (6.28)	0.47**	0.37*
Reading comprehension[5]					
Word comprehension[4]	22.20 (9.27)	29.76 (9.66)	31.94 (9.12)	1.17**	1.32**
Sentence comprehension[4]	5.30 (3.57)	8.83 (6.83)	8.94 (4.93)	0.63**	1.00**
Text comprehension[4]	2.38 (3.03)	3.7 (5.33)	2.71 (1.85)	0.25[n.s.]	.00[n.s.]
Total score[6]	31.70 (5.10)	33.77 (6.06)	33.10 (6.70)	0.29**	.03[n.s]

[1]RAN Test (TEPHOBE, Mayer, 2020), [2]Items/s, [3]Salzburger Lesetest (SLRT II, Moll and Landerl, 2014), [4]Raw Scores, [5]ELFE II (Lenhard et al., 2022), [6]T-score, **$p < 0.001$, *$p < 0.05$, [n.s.]non significant.

of real words ($d = 0.54$–1.04) but only a small effect for reading speed of pseudowords ($d = 0.37$–0.48) (Table 1).

Regarding word comprehension, the repeated measures ANOVA also showed a statistically significant change [F (2, 60) = 27.97, $p < 0.01$, $\eta^2 = 0.65$], with Bonferroni-corrected post-hoc tests indicating significantly higher values for the posttest and follow-up test compared to the pretest ($M_{Diff} = 6.4$ and 8.7, 95%-CI [9.5, 3.2], and [11.7, 5.7]) with a large effect size ($d = 1.17$–1.32) (Table 1).

According to sentence comprehension, the repeated measures ANOVA also indicates statistically significant changes [F (2, 60) = 7.6, $p < 0.01$, $\eta^2 = 0.21$], with Bonferroni-corrected post-hoc tests showing significantly higher values for the posttest and follow-up test compared to the pretest ($M_{Diff} = 5.5$ and 3.4, 95%-CI [6.5, 0.5], and [4.9, 1.8]). The calculated effect sizes indicate a medium to large effect for sentence comprehension ($d = 0.63$–1.00) (Table 1).

No changes were found for text comprehension across the three test times.

5 Discussion

We could show that 18 training sessions conducted twice a week, each lasting 10 to 15 min, significantly improved the RAN performance of children with reading difficulties and a naming-speed-deficit. In particular, the improvements for RAN letters and digits, which are particularly relevant to the acquisition of literacy, showed large effects ($d = 1.37$ and $d = 0.98$). Moreover, repeated measures of ANOVA revealed significant gains in reading speed for words and comprehension at the word- and sentence-level between pre-, post- and follow-up tests.

Thus, the results of the present study are in line with the first promising results on the trainability of naming speed. While the effectiveness of RAN training in the studies by Vander Stappen and van Reybroeck (2018) and Pecini et al. (2019) were limited to naming speed for colors and objects, to our knowledge the present study is the first to show that naming speed for alphanumeric symbols can also be trained.

Studies that failed to show positive effects of RAN training took a methodological approach which asked children to name templates with visual symbols (letters, numbers, colors, objects) frequently (at their own pace) without supporting them in processing the symbols successively faster (de Jong and Vrielink, 2004; Berglez, 2003). In contrast the software used in this study tried to motivate children to improve their speed of visual processing and phonological activation. Three points are particularly important here.

First of all, the software supported attention-focusing, inhibition and control (i.e., "executive functions"), highlighting each item with a red rectangle jumping from item to item. This is important, because according to Wolf and Bowers (1999), RAN tasks require continuously focusing the attention on one symbol, while the information of the distracters must be inhibited.

Secondly, new RAN-matrices are generated for each session by the software in order to maintain the children's attention during the intervention. In the studies of de Jong and Vrielink (2004) and Berglez (2003) children had to always name the same templates during the whole training.

Finally, to speed up visual and phonological processing the speed of the jumping red rectangle was increased after each successful session. A similar program with comparable results was used in the study by Pecini et al. (2019).

As far as highlighting items is concerned, the developed RAN-training is comparable to the text fading paradigm, which has shown its effectiveness in aiding reading development (Korinth and Nagler, 2021; Nagler et al., 2015; Breznitz, 1987). The text fading paradigm is a method designed to help children progressively read words, sentences, and texts faster by gradually fading out letters or syllables of a word in the reading direction (adjusted to the current reading speed). Comparable to our RAN-training, the fading speed is increased slightly after each successful reading attempt.

The advantage of the RAN training (at least for the categories of colors and objects) can be seen in the possibility to use the RAN training with younger, i.e., preschool children than text fading programs, which require basic reading skills. Therefore, the RAN-training should be seen as a preventative method alongside phonological interventions for at-risk children to help develop reading and spelling.

As the training does not place any major demands on practicability, it is conceivable that the training could also be carried out outside of school settings. In the study by Pecini et al. (2019) for example, the entire measure was carried out exclusively at home with parental support.

However, RAN training should not be interpreted as a new magic bullet for preventing reading and spelling disorders. Rather, it is a new, easy-to-implement intervention that could be used to supplement phonological awareness instruction in preschool or explicit reading instruction in school-age children. RAN training practices the linguistic and cognitive processes reading is based on (attention, visual processing speed, speed of access to phonological representations). However, students learn to read exclusively in the symbol system of the written language and not through naming digits and letters.

6 Limitations

The study's evidence is limited to the progress of one group of children with reading difficulties and a naming-speed-deficit without a control group comparison, so we cannot be sure if the progress is the consequence of our training or due to maturation. Future studies must compare the development of a training group with a non-supported control group to confirm the method's efficacy. Furthermore, the calculations do not yet consider whether there are differences in children's performance depending on the type of school. It is conceivable that increases in performance can be explained by the specific support given (particularly for children in special education centers).

Nonetheless, RAN training's theoretical importance should not be underestimated, as it integrates sub-functions related to automatic word recognition. Measures aimed at integrating these sub-functions should be interpreted as an opportunity to provide preventative support in the automation of word recognition for children with impaired naming speed, for whom phonologically oriented support measures alone are obviously not sufficient. However, despite the promising results of the present study, affected children also need specific support in the development of appropriate reading fluency through measures aimed at the holistic and simultaneous processing of orthographic patterns and frequently occurring words (e.g., Mayer, 2002).

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 humans were approved by Ethics Committee of the Faculty of Psychology and Education, LMU Munich. The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation in this study was provided by the participants' legal guardians/next of kin. Written informed consent was obtained from the minor(s)' legal guardian/next of kin for the publication of any potentially identifiable images or data included in this article.

Author contributions

JW: Writing – review & editing, Writing – original draft. AM: Writing – review & editing, Writing – original draft.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Acknowledgments

We would like to thank the children and their families as well as the schools for their participation in this research project.

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

Araújo, S., and Faísca, L. (2019). A Meta-analytic review of naming-speed deficits in developmental dyslexia. *Sci. Stud. Read.* 23, 349–368. doi: 10.1080/10888438.2019.1572758

Berglez, A. (2003). Prävention von Lese-Rechtschreibschwierigkeiten. Ein Training der Benennungsgeschwindigkeit. Bielefeld: Universität Bielefeld.

Bowers, P. G., Golden, J. O., Kennedy, A., and Young, A. (1994). "Limits upon orthographic knowledge due to processes indexed by naming speed" in The varieties of orthographic knowledge. ed. V. W. Berninger (Dordrecht: Springer Netherlands), 173–218. doi: 10.1007/978-94-017-3492-9_6

Breznitz, Z. (1987). Increasing first graders' reading accuracy and comprehension by accelerating their reading rates. *J. Educ. Psychol.* 79, 236–242. doi: 10.1037/0022-0663.79.3.236

Conrad, N. J., and Levy, B. A. (2011). Training letter and orthographic pattern recognition in children with slow naming speed. *Read. Writ. Interdiscip. J.* 24, 91–115. doi: 10.1007/s11145-009-9202-x

Cronin, V. S. (2013). RAN and double-deficit theory. *J. Learn. Disabil.* 46, 182–190. doi: 10.1177/0022219411413544

da Silva, P. B., Engel de Abreu, P. M. J., Laurence, P. G., Nico, M. Â. N., Simi, L. G., Tomás, R. C., et al. (2020). Rapid automatized naming and explicit phonological processing in children with developmental dyslexia: a study with Portuguese-speaking children in Brazil. *Psychology* 11:928. doi: 10.3389/fpsyg.2020.00928

de Jong, P. F., and Vrielink, L. O. (2004). Rapid automatic naming: easy to measure, hard to improve (quickly). *Ann. Dyslexia* 54, 65–88. doi: 10.1007/s11881-004-0004-1

Furnes, B., Elwér, Å., Samuelsson, S., Olson, R. K., and Byrne, B. (2019). Investigating the double-deficit hypothesis in more and less transparent orthographies: a longitudinal study from preschool to grade 2. *Sci. Stud. Read.* 23, 478–493. doi: 10.1080/10888438.2019.1610410

Galuschka, K., Görgen, R., Kalmar, J., Haberstroh, S., Schmalz, X., and Schulte-Körne, G. (2020). Effectiveness of spelling interventions for learners with dyslexia: a meta-analysis and systematic review. *Educ. Psychol.* 55, 1–20. doi: 10.1080/00461520.2019.1659794

Galuschka, K., and Schulte-Körne, G. (2016). The diagnosis and treatment of Reading and/or spelling disorders in children and adolescents. *Dtsch. Arztebl. Int.* 113, 279–286. doi: 10.3238/arztebl.2016.0279

Georgiou, G. K., Parrila, R., and Kirby, J. R. (2009). RAN components and Reading development from grade 3 to grade 5: what underlies their relationship? *Sci. Stud. Read.* 13, 508–534. doi: 10.1080/10888430903034796

Georgiou, G. K., Parrila, R., Kirby, J. R., and Stephenson, K. (2008). Rapid naming components and their relationship with phonological awareness, orthographic knowledge, speed of processing, and different Reading outcomes. *Sci. Stud. Read.* 12, 325–350. doi: 10.1080/10888430802378518

Kail, R., and Hall, L. K. (1994). Processing speed, naming speed, and Reading. *Dev. Psychol.* 30, 949–954. doi: 10.1037/0012-1649.30.6.949

Kirby, J. R., Georgiou, G. K., Martinussen, R., and Parrila, R. (2010). Naming speed and Reading: from prediction to instruction. *Read. Res. Q.* 45, 341–362. doi: 10.1598/RRQ.45.3.4

Kirby, J. R., Parrila, R. K., and Pfeiffer, S. L. (2003). Naming speed and phonological awareness as predictors of Reading development. *J. Educ. Psychol.* 95, 453–464. doi: 10.1037/0022-0663.95.3.453

Korinth, S. P., and Nagler, T. (2021). Improving reading rates and comprehension? Benefits and limitations of the reading acceleration approach. *Lang. Ling Comp.* 15:e12408. doi: 10.1111/lnc3.12408

Lenhard, W., Lenhard, A., and Schneider, W. (2022). ELFE II. Ein Leseverständnistest für Erst- bis Siebtklässler. 5. *aktualisierte Auflage* Edn. Göttingen: Hogrefe.

Mayer, A. (2002). Blitzschnelle Worterkennung. Dortmund: Borgman media.

Mayer, A. (2014). Früherkennung und Prävention von Lese-Rechtschreibschwierigkeiten in inklusiven Kontexten. Unveröffentlichte Habilitationsschrift: Universität zu Köln.

Mayer, A. (2018). Benennungsgeschwindigkeit und Lesen. *Forschung Sprache*, 6, 20–43.

Mayer, A. (2020). TEPHOBE. Test zur Erfassung der phonologischen Bewusstheit und der Benennungsgeschwindigkeit. *4. Auflage* Edn. München: Reinhardt Verlag.

Mayer, A. (2021). Lese-Rechtschreibstörungen. *2. Auflage* Edn. München: Reinhardt Verlag.

Moll, K., and Landerl, K. (2014). SLRT II. Lese- und Rechtschreibtest. Weiterentwicklung des Salzburger Lese- und Rechtschreibtests (SLRT). 2., korrigierte Auflage mit erweiterten Normen 2014. Hogrefe, Göttingen.

Moll, K., Ramus, F., Bartling, J., Bruder, J., Kunze, S., et al. (2014). Cognitive mechanisms underlying reading and spelling development in five European orthographies. *Learn. Instr.* 29, 65–77. doi: 10.1016/j.learninstruc.2013.09.003

Nagler, T., Korinth, S., Linkersdörfer, J., Lonnemann, J., Rump, J., Hasselhorn, M., et al. (2015). Text-fading based training leads to transfer effects on children's sentence reading fluency. *Front. Psychol.* 6, 1–8. doi: 10.3389/fpsyg.2015.00119

Ozernov-Palchik, O., Norton, E. S., Sideridis, G., Beach, S. D., Wolf, M., Gabrieli, J. D. E., et al. (2017). Longitudinal stability of pre-reading skill profiles of kindergarten children: implications for early screening and theories of reading. *Dev. Sci.* 20:e12471. doi: 10.1111/desc.12471

Pecini, C., Spoglianti, S., Bonetti, S., di Lieto, M. C., Guaran, F., Martinelli, A., et al. (2019). Training RAN or reading? A telerehabilitation study on developmental dyslexia. *Dyslexia* 25, 318–331. doi: 10.1002/dys.1619

Shaywitz, S. E., Shaywitz, B. A., Fletcher, J. M., and Escobar, M. D. (1990). Prevalence of reading disability in boys and girls. *JAMA J. Am. Med. Assoc.* 264, 998–1002. doi: 10.1001/jama.1990.03450080084036

Stainthorp, R., Stuart, M., Powell, D., Quinlan, P., and Garwood, H. (2010). Visual processing deficits in children with slow RAN performance. *Sci. Stud. Read.* 14, 266–292. doi: 10.1080/10888431003724070

Swanson, H. L., Trainin, G., Necoechea, D. M., and Hammill, D. D. (2003). Rapid naming, phonological awareness, and reading: a meta-analysis of the correlation evidence. *Rev. Educ. Res.* 73, 407–440. doi: 10.3102/00346543073004407

Torppa, M., Parrila, R., Niemi, P., Lerkkanen, M., Poikkeus, A., and Nurmi, J. (2013). The double deficit hypothesis in the transparent Finnish orthography: a longitudinal study from kindergarten to grade 2. *Read. Writ. Interdiscip. J.* 26, 1353–1380. doi: 10.1007/s11145-012-9423-2

Vaessen, A., and Blomert, L. (2010). Long-term cognitive dynamics of fluent reading development. *J. Exp. Child Psychol.* 105, 213–231. doi: 10.1016/j.jecp.2009.11.005

van Ermingen-Marbach, M., Verhalen, N., Grande, M., Heim, S., Mayer, A., and Pape-Neumann, J. (2015). Orientierungswerte für die Benennungsgeschwindigkeit bei leseunauffälligen Kindern im Alter von 9 bis 11 Jahren. *Sprache Stimme Gehör.* 38, e28–e32. doi: 10.1055/s-0034-1375625

Vander Stappen, C., and van Reybroeck, M. (2018). Phonological awareness and rapid automatized naming are independent phonological competencies with specific impacts on word Reading and spelling: an intervention study. *Front. Psychol.* 9:320. doi: 10.3389/fpsyg.2018.00320

Wagner, R. K., and Torgesen, J. K. (1987). The nature of phonological processing and its causal role in the Acquisition of Reading Skills. *Psychol. Bull.* 101, 192–212. doi: 10.1037/0033-2909.101.2.192

Wimmer, H. (1993). Characteristics of developmental dyslexia in a regular writing system. *Appl. Psycholinguist.* 14, 1–33. doi: 10.1017/S0142716400010122

Wimmer, H., and Mayringer, H. (2002). Dysfluent Reading in the absence of spelling difficulties: a specific disability in regular orthographies. *J. Educ. Psychol.* 94, 272–277. doi: 10.1037/0022-0663.94.2.272

Wolf, M., and Bowers, P. G. (1999). The double-deficit hypothesis for the developmental dyslexias. *J. Educ. Psychol.* 91, 415–438. doi: 10.1037/0022-0663.91.3.415

Wolf, M., Goldberg O'Rourke, A., Gidney, C., Lovett, M. W., Cirino, P., and Morris, R. (2002). The second deficit: an investigation of the independence of phonological and naming-speed deficits in developmental dyslexia. *Read. Writ. Interdiscip. J.* 15, 43–72. doi: 10.1023/A:1013816320290

World Health Organization (2008). World health statistics. Available at:https://www.who.int/publications/i/item/9789241563598. Accessed August 06, 2024.

frontiers | Frontiers in Education

TYPE Original Research
PUBLISHED 04 November 2024
DOI 10.3389/feduc.2024.1378493

Check for updates

OPEN ACCESS

EDITED BY
Adrian Pasquarella,
University of Delaware, United States

REVIEWED BY
Diana Alves,
University of Porto, Portugal
Bita Moradi,
University of Delaware, United States

*CORRESPONDENCE
Martin Schöfl
✉ martin.schoefl@jku.at

RECEIVED 29 January 2024
ACCEPTED 30 September 2024
PUBLISHED 04 November 2024

CITATION
Schöfl M, Steinmair G, Zepnik S and Weber C (2024) Using an app-based screening tool to predict deficits in written word spelling at school entry. *Front. Educ.* 9:1378493.
doi: 10.3389/feduc.2024.1378493

Using an app-based screening tool to predict deficits in written word spelling at school entry

Martin Schöfl[1,2]*, Gabriele Steinmair[1], Sabine Zepnik[1] and Christoph Weber[1,2]

[1]Department of Educational Sciences, University of Education Upper Austria, Linz, Austria, [2]Research Institute for Developmental Medicine, Johannes Kepler University Linz, Linz, Austria

Introduction: The first year of schooling is crucial for the further development of spelling abilities in children, which makes early assessment and intervention essential. The aim of this study was to develop and validate an efficient and cost-free screening tool for identifying spelling problems in community school settings around the time of school entry.

Methods: A broad range of precursors of spelling (vocabulary, grammar, letter knowledge, phonological awareness, phonological working memory, rapid automatized naming) were assessed in 522 Austrian first graders (6–7 years of age) in the first weeks of schooling. At the end of first grade, spelling abilities were assessed by newly developed spelling tasks based on the trochaic foot. By applying logistic regression with the least absolute shrinkage and selection operator (LASSO), we aimed to select a set of important predictors of spelling problems at the end of grade 1 (i.e., scoring below the 16th percentile in the spelling test).

Results: Our analysis identified letter knowledge (i.e., an aspect of phonological information processing) and sentence repetition (i.e., a measure of grammatical knowledge) as important predictors of spelling problems. The screening tool has acceptable diagnostic accuracy [area under the curve (AUC) = 0.0.725 and DeLong 95% CI (0.666, 0.784)]. Further analyses indicated that the AUC differs neither between boys and girls nor between children with and without German as their first language.

Discussion: These results suggest that administering the screening tool during the first weeks of schooling is a valid approach to identifying spelling deficits, which in turn enables early targeted pedagogical interventions. Practical implications for spelling instructions are discussed.

KEYWORDS

Schnapp spelling test, multilingual first graders, phonological processing, language abilities, graphematic trochaic foot

Introduction

The term *spelling* is used ambiguously: on the one hand, spelling is often associated with the term *oral spelling*, which means memorizing a word and naming its individual letters (Treiman and Bourassa, 2000). On the other hand, it is used to describe the conversion of spoken words by young children into phonemes and thence into a chain of letters (Treiman et al., 2019). It is not clear whether the latter is also related to knowledge about an orthographic form (Treiman et al., 2019). Since spelling generally deals with correctness (Treiman et al., 2023), which also matches our approach, we use the term spelling when we refer to *writing* *down words correctly*, and include implicit and systematic knowledge of the graphematic word structure in German (Bredel, 2015a; Primus, 2010) as a basic principle.

Spelling prediction relies predominantly on phonological information processing (PI) and language comprehension. Embracing a comprehensive approach, we adopt Wagner and Torgesen's (1987) extended definition, which includes various components for word inscription. Phonological information processing, as defined by Wagner and Torgesen (1987), p. 192, entails the utilization of phonological cues—that is, the auditory aspect of language—in both written and oral language processing. This includes sublexical processes, such as letter knowledge (LK), rapid automatized naming (RAN), phonological awareness (PA), and phonological working memory (PWM) (Niolaki et al., 2020; González-Valenzuela et al., 2023). The related research literature (for an overview see Appendix 1) considers as relevant not only precursors related to phonological information processing, but also lexical precursors, such as visual short-term memory (VSTM), visual attention span (VAS) (Niolaki et al., 2020; Niolaki et al., 2024), oral language skills, including vocabulary (V) and grammar (GR) (Kim et al., 2013; von Goldammer et al., 2010; Fricke et al., 2016), and early oral spelling skills (ESS) (Treiman et al., 2023), which are considered foundational for orthographic spelling. We discern that, within English-language research, phonological awareness emerges as a key predictor. For other writing systems, however, the significance of RAN and oral language proficiency is increasingly acknowledged.

Learning to spell is a slow process; correct spelling usually requires effort, regardless of the orthography in question, and some children need extra support (Nag et al., 2019). Because spelling continues to be a highly valued skill (Pan et al., 2021), it is important to find sensitive, accurate methods for assessing children's spellings (Treiman and Kessler, 2004) or—even earlier—precursors for word-writing development and underlying cognitive processes that children use when they write (Caravolas et al., 2001; Kessler et al., 2013; Treiman et al., 2016). Since early orthographic knowledge predicts later orthographic skills (Zarić et al., 2020), early support is essential. Diagnostic and interventional approaches are normally based on theoretical models. Written language acquisition is usually described using so-called phase models (Frith, 1986; for German, e.g., Thomé, 2006; Scheerer-Neumann, 2006). These rely strongly on phoneme-grapheme relations, and require children to listen carefully and write down exactly what they have heard before orthographic writing can be developed. Our approach, however, builds on the alternative graphematical theory introduced by Eisenberg (1989) and Maas (1992) in which the phonological suprasegmental unit—the syllable—is given preference to single letter-sound relations within the written language. This view has been established within graphematics and expanded to the graphematic foot, a sequence of two syllables, a strong and a weak or reduced syllable (Primus, 2010; Evertz, 2016; Maas, 2022; Eisenberg, 2020; Fuhrhop and Peters, 2023). A didactical implementation of this graphematical theory suggests that children are introduced to the graphematic structure based on the syllable and the trochaic foot, and from the beginning of grade 1, they are thus enabled to discover patterns and regularities (Röber, 2013; Bredel, 2016; Bredel et al., 2017).

Existing screening tools

Only a few screening tools exist for use with German orthography: The "LRS-Screening" (Endlich et al., 2019) used in the year preceding formal schooling employs 14 subtests to forecast word-spelling deficits by the end of first grade. Notably, it has a sensitivity of 0.74, a specificity of 0.68, and a positive predictive value (PPV) of 0.27. The "LRS-Screening" is administered individually in a traditional paper-and-pencil format and takes approximately 30 min plus additional scoring time. It lacks an engaging cover story, which potentially limits its appeal. Another screening tool commonly used prior to school entry is the "Bielefelder Screening" (Jansen et al., 1999). Although its manual reports promising predictive values, independent studies have failed to replicate these findings. The Phonological Awareness-Reading and Spelling Screening (PB-LRS; 80) is designed for group assessment during the final year of kindergarten. It boasts a sensitivity of 63%, a specificity of 87%, and a PPV of 36%, but it requires approximately 60 min to administer. Another established tool, the "A Tour of Hörhausen" (Martschinke et al., 2001), comprises phonological assessments individually at the beginning and midpoint of first grade. Prognostic validity studies based on a sample of 375 children revealed a specificity exceeding 80%, with sensitivity ranging from 38 to 48%. Assessment takes approximately 40 min. The only app-based screening is a tool called "Förderorientieres Schuleingangsscreening" (BMBWF, 2024), which is supported by the Austrian government and consists of tasks that concern phonological awareness and letter knowledge (Jöbstl et al., 2022). This forms part of a broader battery of seven modules that also address precursors of numeracy development, graphomotor skills and executive functions. This assessment is of acceptable validity and presented in a motivating format, but lacks important precursors for written language, such as vocabulary and grammar.

In summary, existing screening tools for predicting word spelling difficulties in German are time-consuming to administer or exhibit limited predictive efficacy, as detailed in comprehensive reviews (Marx and Lenhard, 2010). Language as a main factor is, in many cases, not one of the components screened.

The present study

The objective of the present study was to develop and validate an efficient school-entry screening tool for identifying spelling problems at the end of grade one in community school settings. In order to enable broad implementation, the tool had to be administrable by teachers and feasible in terms of interpretation and integration into everyday school life. Therefore, to facilitate assessment and interpretation of results, and following the general trend in psychological and educational assessment towards increasing integration of digital media in practice (Schaumburg, 2015), we developed an app-based screening approach, which in Austria can also be administered in the final year of kindergarten (Thompson, 2015). Unlike existing screening tools, this study considered a broad set of precursors assessed at school entry phonological information processing (initial phoneme detection, letter knowledge, and phonological working memory) and language (vocabulary and grammar) as predictors of spelling. Focusing on the validity of the selected set of predictors, this study evaluated diagnostic accuracy—broadly defined as the ability of the screening to discriminate between children with and without spelling problems. To investigate the generalizability of the screening, an examination was undertaken of the potential for variation in diagnostic accuracy according to sociodemographic variables. These include the child's gender and first language, both of which are known to be associated with spelling as well as spelling precursors. The extant research demonstrates that children from low socioeconomic status (SES) backgrounds exhibit lower verbal academic performance than their peers from high-SES families (Sirin, 2005; White, 1982). Similarly, there is substantial evidence indicating that children from low-SES environments demonstrate weaker orthographic skills, particularly in spelling, in comparison to their higher-SES counterparts (Breit et al., 2016; Niemietz et al., 2023 for German-speaking countries). Large-scale international assessments such as PISA (OECD, 2023) indicate that students who have acquired the language of instruction as a second language, frequently due to a migration background, frequently demonstrate suboptimal performance in spelling. Specifically, research findings suggest that both second-language German learners (Lenhart et al., 2019) and students with a migration background (Henschel et al., 2023) exhibit comparatively weaker spelling abilities than their peers.

In summary, the following research questions guided this study:

> Research question 1: Which predictors assessed at school entry can identify spelling problems at the end of grade 1?
>
> Research question 2: What is the diagnostic accuracy of the selected set of predictors?
>
> Research question 3: Does the diagnostic accuracy of the selected set of predictors vary between sociodemographic groups?

Methods and materials

Sample and recruitment

This prospective study followed children from the beginning to the end of first grade and is part of a longitudinal and ongoing project called SCHNAPP (Detection of risks factors for reading and writing development, see Schöfl et al., 2022). The sample was recruited from various community-based schools in the district of the Upper Austrian capital. The principals of 14 schools were telephoned and then visited in person, and invited to participate in the study project. All of them agreed to participate and informed their teachers of first grade classes ($n = 32$). Parents of 557children (96.7%) provided permission for their child to participate in the study.

Precursors of spelling—phonological information processing and language—were assessed in the first weeks of first grade, before the formal teaching of reading and spelling had begun. In detail, individualized screening of precursors started in autumn 2022 within the first 2 weeks of the school year. Within 3 weeks, 85% of the sample had been assessed. In the two subsequent weeks, children who had been ill or unable to attend within the first testing period were surveyed. Spelling was assessed by a test administered in a classroom setting at the end of first grade. Twenty children were excluded from the analysis because they were listed as (preschoolers[1]) in their first year of learning and did not advance to the first grade at the end of the school year. Eight children were diagnosed during the school year as having special educational needs (SEN). The spelling test was not administered to these children as they had not yet acquired the requisite knowledge of the alphabet. A total of 7 children who were part of the target population for the screening could not be tested (4 children changed school and 3 children were ill, at the first and second test date). Thus, our study used data from $n = 522$ children (=557 (with parental consent) − 20 (preschoolers) − 8 (SEN students) − 7 (missing tests)).[2] For these children all variables used in the current study were available. Figure 1 shows the recruitment pathways and timeline.

The distribution of the children in the final study sample was heterogeneous. The proportion of girls in the sample was 46.7%, which equals the proportion of girls among Upper Austrian first graders in 2022 [$\chi^2(1) = 0.00$, $p > 0.05$; Statistik Austria, 2024]. 73.2% of the children assessed spoke German exclusively as their first language, which also corresponds to the population value (i.e., Upper Austrian first graders) of 73.1% [$\chi^2(1) = 0.01$, $p > 0.05$; Statistik Austria, 2024]. The sample had parents from all educational backgrounds. In about four out of 10 families (42.7%) at least one parent held a university degree. In 21.3% of the families, the highest educational degree was a university entrance qualification. About a third (30.1%) reported a vocational education or training, and in 5.9% of the families neither parent had any qualification beyond compulsory schooling. Note that, comparison of the proportion of parents with the highest educational level in our sample with that of the Upper Austrian parent population of 4th graders in the school year 2017/18[3] shows that families in which at least one parent has a university degree are overrepresented (42.7% vs. 26.5%) and parents with vocational education or training are underrepresented [30.1% vs. 45.8%; $\chi^2(3) = 77.50$, $p < 0.001$]. Thus, our study sample was representative in terms of gender and first language, but not in terms of parental education, which is probably due to the sample having been recruited from many schools in urban areas, where more parents with higher education live (see also Elliott, 2018, p. 245; Yulianti et al., 2023).

Procedure

Before summer vacation, teachers were informed by the research team about the procedure and received an information letter for the parents or legal guardians, including consent forms and questions about parents' educational background, children's first language, and

1 These children were not taught according to the first-grade elementary school curriculum, which means that the children did not learn to write within the study period.

2 Notably, from a missing data perspective (Enders, 2010), the seven cases that were excluded from the study (children who were ill and children who changed school) could be regarded as unit non-response, making an adequate treatment of missing data necessary. However, given the small number of cases with missing values (1.3%) and analyses indicating that missing cases did not significantly differ in study variables from the other cases (p-values >0.088), we decided to exclude these cases (i.e., listwise deletion).

3 We thank the IQS (Federal Institute for Quality Assurance in the Austrian School System) for providing detailed population data.

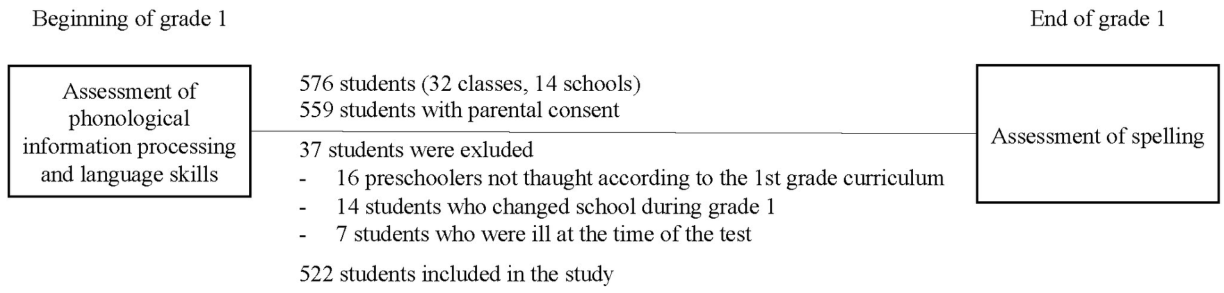

FIGURE 1
Recruitment pathways and timeline.

language use. Teachers entered the IDs on a digital platform for use on a tablet. The investigators, all of whom were student teachers, were enrolled in a student seminar where they learned about the tasks and the testing procedure and for which they received academic credits (towards their study program). Phonological information processing and language skills at the beginning of grade 1 were assessed one on one, with child and instructor seated opposite each other across a table. After a brief welcome, the child was handed a tablet running an app that introduced SCHWUPP the friendly dragon. App navigation was designed such that the child could use it independently. If needed, investigators could help, for instance, to move SCHWUPP across the screen. All instructions essential for the child were recorded as audio files, opened automatically, and repeated if necessary. A yellow background in the app signaled to the test administrator that the child was making test selections independently (e.g., in the phonological tasks). A gray background meant that the test administrator should take the tablet to read and then give instructions. The assessment, including all subtests, took an average of 15.4 min (SD = 4.3) per child. Pauses were possible but rarely needed. The spelling test is also tablet-based (see below). It was administered by the research team and student teachers in a classroom setting.

Measures

Phonological information processing

Letter knowledge

All 26 letters of the alphabet were presented as capital letters on screen in random order. Each page contained three to four letters, and each child was prompted, "I know you have not yet learned these letters at school. Maybe you already know one? If you do, please name it!" The child named letters and tapped on those that they knew. In the case of an error (wrong letter as chosen known), the investigator could correct the error by tapping again. Positive scores were given for letter names or sounds and ticked off on the tablet. Reliability (Cronbach's alpha) of this newly designed task was excellent at 0.96.

Initial phoneme detection

The task was introduced with three practice items, including feedback, followed by 10 test items. Tasks were constructed using high-frequency words from the childLex database (Schroeder et al., 2015) for the youngest age group (6–8 years). We presented each letter visually and as a speech sound simultaneously. The child then had to select from three pictures that which illustrated a word that started with the same first phoneme as the letter given ("Which word begins with I like Ines: Hase, Igel, Spiegel?") (Ines = German name "Ines," Hase = hare, Igel = hedgehog, Spiegel = mirror). Although internal consistency was relatively low at 0.66, this test has been shown to be an important predictor of reading in prior research (Schöfl et al., 2022). Therefore, we also used this subtest in the present study.

Phonological working memory

Phonological working memory was assessed by means of two subtests from a broad-range intelligence test battery [IDS-2; Intelligence and Development Scales for Children and Adolescents (Grob et al., 2009)] which tests the memory of letter-number sequences forward (letter-number span forward) and backward (letter-number span backward). Each child was asked to repeat a series of mixed digits and letters (3-A, 5-M-2) from beginning to end (forward condition) or vice versa (backward condition). The investigator tapped the correct solutions on the tablet. The difficulty level of the tasks was determined by the increasing length of the sequences, and the test was terminated after three unsolved or incorrectly solved tasks. Reliability of this test has been described as excellent; Cronbach's alpha was 0.89 (end of first grade). Retest reliability was rtt = 0.93 (first grade).

Rapid automated naming

To assess RAN, two different stimuli (objects and digits) were used. Analysis of our own data has recently shown effects of both measures assessed at the very beginning of schooling on reading assessed at the end of grade 1 (Schöfl et al., 2023a). The object condition was designed using five high-frequency monosyllabic words (cow, hand, ice, tree, and mouse). First, the items of the RAN tasks were presented visually and by playing an audio file via the app, and the task was repeated. After a training session, the test began: 30 items were presented on the screen in random order over six lines. The investigator pressed a button on the tablet to time the test. When the child reached the last item, the investigator stopped the clock, and the time distance was calculated and stored automatically. The RAN test in the digit condition was based on the work of Denckla and Rudel (1974) following Landerl et al. (2013) and was presented and rated analogously to the object condition with monosyllabic digits (2, 8, 1, 6, 3).

Language

Grammar

Morphosyntactic skills were assessed using an adapted sentence repetition task. The German version was constructed according to the Language Impairment Testing in Multilingual Children (LITMUS) principles by Ibrahim et al. (2018), following the COST Action IS0804. A block of 15 items representing morphosyntactic constructions with varying degrees of complexity was selected and scored according to whether the sentence was completed correctly. Internal consistency was high at 0.857.

Receptive vocabulary

Receptive vocabulary was assessed using the digital form of the Graz Vocabulary Test (GraWo; Seifert et al., 2019), which consists of 30 matching tasks. The child was asked to choose from four pictures the one that best matched the audio-presented word. Reliability data are given for the paper form of the GraWo (Cronbach's Alpha for end of first grade was 0.89; Retest reliability was rtt = 0.93).

Standardized assessment of spelling at the end of the first year

To identify students with spelling deficits, the newly developed and validated SCHNAPP spelling test (Schöfl et al., 2023b) was used. Within a magical cover story about SCHWUPP the dragon, 22 words must be written according to dictation. A stylus pen is used for input to the tablet. The target words and corresponding sentences are played using headphones. The audio files can be repeated upon an input command from the child. The next item is presented only when the child gives an input command, which allows the test to progress at the child's own pace. The vocabulary is based on the childLex corpus, a database of written language for children (Schroeder et al., 2015). In accordance with Schroeder et al. (2015), the words were divided into three groups according to frequency of use: >100 high-frequency, 10–100 medium-frequency, and 1–10 low-frequency. In relation to the writing system, we used a literature-derived hierarchy of vocabulary based on the canonical trochaic form typical of German (see Schöfl et al., 2023b). Schöfl et al. (2023b) showed that the test items conformed to the one-parameter Rasch model, and the test exhibited good reliability (0.86). Further, the test differentiates well in the low ability range, which makes it suitable for identifying spelling deficits. For the current study, test scores one standard deviation below the mean were considered as spelling deficits.

Sociodemographic measures

Mono- vs. bilingualism classification was based on a questionnaire completed by the parents. If they indicated that the first language was German only or that contact with German occurred from birth up to and including the age of 2, then children were classified as monolinguals. Children who had had contact with German only after the age of 2 years were classified as bilingual.

Analyses

First, we performed bivariate analyses. We report point-biserial correlations r_{pb} between the predictors and spelling as well as area under the curve (AUC) values based on receiver operator characteristic (ROC) curve analyses to evaluate the diagnostic accuracy of each predictor. AUCs are global measures of diagnostic accuracy. Values ≥0.9 are regarded as excellent, AUCs ≥0.8 and <0.9 as good, AUCs ≥0.7 and <0.8 as fair, and tests with AUCs <0.7 as poor (Swets, 1988). The pROC package (Robin et al., 2011) in R was used for ROC curve analysis.

Second, we applied multivariate method to simultaneously evaluate the predictive power of spelling precursors. We performed a logistic regression with adaptive variable selection to identify important predictors for inclusion in the screening. Specifically, we used the least absolute shrinkage and selection operator (LASSO; McNeish, 2015; Friedman et al., 2010; for an application of LASSO for the selection of screening variables, see Schöfl et al., 2022; Silverman et al., 2021) implemented in the glmnet R Package (Friedman et al., 2021). LASSO is a statistical learning procedure that serves two purposes: selection of important predictors and attenuation of overfitting. Notably, standard variable selection procedures such as backward and forward selection tend to increase overfitting, which limits the generalizability of the regression results and thus also of a screening algorithm to be developed because sample regression estimates are a blend of signal and noise; sample estimates are therefore greater than population estimates. To counter inflation of the sample estimates and improve generalizability of the results, LASSO applies a penalty term (λ) to the likelihood function. As in backward or forward selection procedures, unimportant predictors are excluded from the regression model. Note that LASSO does not allow any conclusions to be drawn about the statistical significance of a predictor. Although methods for calculating p-values for linear LASSO models have been developed, no such methods are available for logistic models. The predictors selected must be considered important irrespective of whether their regression estimates are statistically significant (McNeish, 2015; Silverman et al., 2021). We z-scored the predictors to make their effects comparable. Thus, the estimates are standardized. As noted, LASSO uses a penalty term (λ) to counter overfitting and zero-out predictors. Following the suggestions in the literature (James et al., 2013), we used 10-fold cross-validation to select the penalty term. Specifically, we chose the λ-value that resulted in the largest AUC (λ_{AUC_max}) to ensure the highest diagnostic accuracy of the prediction model. However, since the maximal AUC may in some cases insufficiently attenuate overfitting, we additionally report the results for a second λ value. We applied the one standard error rule (λ_{AUC_1SE}; selecting the highest value of λ at which the AUC is within one standard error of the greatest AUC; see, e.g., James et al., 2013, p. 214), which selects a model that has a comparable AUC to the model with the highest AUC, but which is more parsimonious (i.e., has a smaller number of predictors). Tenfold cross-validation randomly splits the data into 10 groups (data sets), where one set is used for validation and the remaining nine sets are used for training in the statistical learning procedure. Given this randomness, the results vary with each re-estimation of the model (James et al., 2013). To make our results reproducible and evaluate their robustness, we set five different random seeds and averaged the resulting estimates. For comparison we also report the results of a classical logistic regression using all predictors simultaneously. Finally, we used the estimates of the LASSO models to calculate the probability of spelling deficits at the end of grade 1 (i.e., spelling scores one SD below the mean). These scores, ranging from 0 to 1, were used as screening measures in subsequent analyses.

Third, by applying a bootstrapped test for the comparison of AUC values of paired ROC curves, we compared the diagnostic accuracies of the screening scores λ_{AUC_max} and λ_{AUC_1SE} (Robin et al., 2011).

Fourth, we evaluated whether the ROC curves differed (i) between girls and boys and (ii) between children with and without German as their first language. Differences in ROC curves and AUC values indicate variations in diagnostic accuracy and therefore limit the generalizability of screening (Youngstrom, 2014). We used a bootstrapped test for unpaired ROC curves to compare AUC values between two groups. Further, we applied the Venkatraman (2000) permutation test implemented in the pROC R-package (Robin et al., 2011), which compares actual ROC curves. Note that two ROC curves that do not differ in screening scores have the same sensitivity and specificity across groups, and thus no group-specific cut-off values are required.

Fifth, we used the R-OptimalCutpoints package (López-Ratón et al., 2014) to evaluate several cut-offs that can be used in practice. We report the following diagnostic accuracy statistics: sensitivity (Se; rate of children with spelling deficits identified correctly by the screening process), specificity (Sp; rate of children without spelling deficits identified correctly by the screening), positive predictive value (PPV; rate of screen-positives that show spelling deficits), negative predictive values (NPV; rate of screen-negatives that do not show spelling deficits), and diagnostic likelihood ratios for positive and negative screening results (DLR+ and DLR−, respectively). Unlike predictive values, DLR+ and DLR− are not associated with the prevalence of the problem (here: spelling deficits) under investigation (Pepe, 2004). DLR+ indicates the multiplicative change in the pre-screening odds of scoring one SD below the mean in the spelling test given a positive screening result (i.e., post-screening odds = DLR+ × pre-screening odds). DLR− is the change in the pre-screening odds of scoring one SD below the mean given a negative screening result (post-screening odds = DLR− × pre-screening odds). In the medical literature, DLR+ values ≥10 and DLR− ≤ 0.1 indicate great changes in pre-screening odds, DLR+ ≤ 10 and > 5, and DLR− > 0.1 and ≤ 0.2 indicate moderate changes, DLR+ ≤ 5 and > 2, and DLR− > 0.2 and ≤ 0.5 indicate small changes. DLR+ < 2 and DLR− > 0.5 are rarely important (Jaeschke et al., 1994).

Results

The results for research question 1 are shown in Table 1. The columns labeled bivariate analyses show that all subtests correlate significantly ($p < 0.001$) with spelling deficits, but the correlations are relatively small. The strongest correlation was found for sentence repetition ($r_{pb} = -0.236$) followed by letter knowledge ($r_{pb} = -0.218$) and initial phoneme detection ($r_{pb} = -0.203$). Children scoring higher on the sentence repetition test, the letter knowledge test, and the initial phoneme detection test at school entry were therefore less likely to show spelling deficits at the end of grade 1. The smallest correlation was found for the letter-number span forward ($r_{pb} = -0.116$). In line with the relatively low correlations, the AUC values for the single predictors are poor (i.e., <0.70). Thus, individually the subtests show relatively low diagnostic accuracy. To improve the diagnostic accuracy by using a set of important predictors simultaneously, we now focus on the results of the LASSO models, which are shown in the columns labeled LASSO models in Table 1. For the penalty term λ at which the AUC reaches its maximum, six out of the eight predictors emerge as important, with letter knowledge ($b = -0.389$) and sentence repetition ($b = -0.388$) being the strongest predictors. Initial phoneme detection ($b = -0.101$), RAN objects ($b = 0.138$), RAN digits ($b = 0.007$) and letter-number span backwards ($b = -0.079$) were also included in the prediction model. The more parsimonious λ_{AUC_1SE} model selected two predictors: sentence repetition ($b = -0.130$) and letter knowledge ($b = -0.076$). Further, we report the results of a standard logistic regression model with all predictors considered simultaneously for comparison (see Table 1, column logistic regression): These show that only letter knowledge ($b = -0.497$, $p < 0.01$) and sentence repetition ($b = -0.512$, $p < 0.01$) have significant effects on spelling deficits. However, when interpreting statistical significance, note that the collinearity of the predictors inflates the standard errors and therefore increases the p-values.

In regard to diagnostic accuracy (research question 2), we found that the six-predictor model resulted in an AUC of 0.736 [DeLong 95% CI (0.675; 0.797)], which suggests adequate diagnostic accuracy. The AUC for the two-predictor model was also fair and only slightly smaller than the AUC with six predictors [AUC = 0.725; DeLong 95% CI (0.666; 0.784)]. According to a bootstrapped test for paired ROC curves, the difference between the AUCs for λ_{AUC_max} and λ_{AUC_1SE} is not significant ($D = 0.919$, $p = 0.358$), which supports the use of the more parsimonious model with two predictors.

In order to provide information on the generalizability of the screening tool (research question 3), we report the results of analyses comparing ROC curves between groups. Table 2 provides AUC values for subgroups (girls vs. boys, and monolingual German-speaking children vs. children with another first language). The AUC based on the parsimonious λ_{AUC_1SE} model with three predictors was 0.746 [DeLong 95% CI (0.671; 0.821)] for boys and 0.692 [DeLong 95% CI (0.594; 0.791)] for girls. Both the bootstrapped test for unpaired ROC curves ($D = 0.855$, $p = 0.392$) and the Venkatraman test, which compares actual ROC curves ($E = 0.016$, $p = 0.475$), indicate that the ROC (AUC difference = 0.054) curves do not differ between boys and girls. Although the AUC difference between monolingual German-speaking children [AUC = 0.756; DeLong 95% CI (0.684; 0.829)] and children with a different first language [AUC = 0.658; DeLong 95% CI (0.540; 0.776)] is somewhat greater (AUC difference = 0.099), both tests again indicate that the ROC curves do not differ significantly ($D = 1.429$, $p = 0.153$; $E = 0.020$, $p = 0.207$). The same results (i.e., no significant differences in diagnostic accuracy between subgroups) were found for the λ_{AUC_max} screening score.

Finally, we evaluated various cut-offs for the screening scores (see Table 3). We report diagnostic accuracy statistics for cut-off values with sensitivity = 0.75, 0.80, 0.85, and 0.90. This range aims to capture various conditions (resources available to support students, etc.) at schools. A sensitivity of 0.75 results in a comparatively smaller number of screen-positives to be supported, at the cost of a high number of false negatives. For example, the parsimonious λ_{AUC_1SE} AUC screening score based on three predictors resulted, for a sensitivity of 0.75, in an empirically estimated SE of 0.750 [95% CI (0.644, 0.838)] and a rate of positive screening results of 42.3%. With this cut off, only 75% of the children with spelling deficits would be identified correctly as children in need of support. This cutoff yields a specificity of 0.639 [95% CI (0.592, 0.684)] and a PPV of 0.285 [95% CI (0.206, 0.408)], which implies that 28.5% of the positive screens finally show spelling deficits. The DLR+ of 2.079 [95% CI (1.745, 2.478)] and DLR− [0.391; 95% CI (0.268,

TABLE 1 Prediction of spelling deficits: bivariate and multivariate results.

	Bivariate analyses				Logistic regression			LASSO model	
		ROC-analysis						λ_{AUC_max}	λ_{AUC_1SE}
	r_{pb}	AUC	DeLong 95% CI		b	SE	p-value	b	b
Letter knowledge	−0.218***	0.677	0.616	0.738	**−0.497**	**0.164**	**0.002**	−0.389	−0.076
Initial phoneme detection	−0.203***	0.633	0.566	0.699	−0.145	0.145	0.319	−0.101	
Letter-number-span forward	−0.116**	0.606	0.537	0.676	0.149	0.146	0.307		
Letter-number-span backward	−0.193***	0.647	0.584	0.710	−0.124	0.146	0.395	−0.079	
RAN objects	0.175***	0.613	0.544	0.681	0.163	0.148	0.270	0.138	
RAN digits	0.169***	0.602	0.538	0.667	0.018	0.158	0.907	0.007	
Sentence repetition	−0.236***	0.683	0.622	0.744	**−0.512**	**0.182**	**0.005**	−0.388	−0.130
Receptive vocabulary	−0.178***	0.647	0.585	0.709	0.031	0.167	0.853		
Intercept					−1.915	0.146		−1.835	−1.661

TABLE 2 Comparison of ROC curves between subsamples.

λ_{AUC_max}	AUC	DeLong 95% CI		AUC difference	
				Value	Bootstrap (D)/Venkatraman (E)
Gender					
Boys	0.779	0.702	0.856	0.102	$D = 1.637$, $p = 0.102$
Girls	0.677	0.579	0.775		$E = 0.028$, $p = 0.103$
First language					
Monolingual German	0.760	0.688	0.832	0.079	$D = 1.110$, $p = 0.267$
Bilingual	0.681	0.557	0.804		$E = 0.016$, $p = 0.362$
Gender					
Boys	0.746	0.671	0.821	0.054	$D = 0.855$, $p = 0.392$
Girls	0.692	0.594	0.791		$E = 0.016$, $p = 0.475$
First language					
Monolingual German	0.756	0.684	0.829	0.099	$D = 1.429$, $p = 0.153$
Bilingual	0.658	0.540	0.776		$E = 0.020$, $p = 0.207$

0.570)] indicate small but important changes in the pre-screening odds. In contrast, a high sensitivity of 0.90 (i.e., a smaller number of false negatives) implies many screen-positives to be supported, which may overtax the resources available in schools. For the parsimonious λ_{AUC_1SE} AUC screening score, we found an empirically estimated SE of 0.905 [95% CI (0.821, 0.958)]. This cut off, however, resulted in 73.0% positive screening results, which means that about three out of four children would have to be supported in order to reach 90% of all children in need of support. Consequently, the specificity is very low at 0.306 [95% CI (0.263, 0.351)], and also the PPV [0.200; 95% CI (0.168, 0.375)] and the DLR+ [1.304; 95% CI (1.188, 1.431)] are low. Generally, the various cut-off values yield high proportions of students with a positive screening result (ranging from 43.5 to 72.3%) and thus low specificity values and PPVs, which indicate high rates of false positives.

Discussion

In international research, letter knowledge, oral skills (i.e., grammar and vocabulary), RAN, orthographic awareness, and phonological awareness have recently been identified as strong precursors of spelling (for a comprehensive review see Appendix 1). This is also the assumption of this study, however new is the for everyday school life suitable assessment procedure at the beginning to predict spelling at the end of the first year of primary school. By use of various precursor subtests, we designed a cost-free, tablet-based screening battery to predict written word spelling deficits in Austrian-German spelling. Particularly letter knowledge and grammar play a major part in this respect. The didactic implications of the screening variables letter knowledge and grammar are discussed below.

According to research question 1, we identified two models to predict German spelling ability in Austrian German: a broader model based on the variables letter knowledge, grammar, RAN digits, RAN objects, phoneme detection and phonological working memory. A second—even more robust—model highlights just the two most predictive variables, letter knowledge and grammar. Those predictors are in line with recent and prior research for German (Ennemoser et al., 2012; Fricke et al., 2016; Kastner-Koller et al., 2023). Other precursors, like RAN and phonological awareness are relevant but

TABLE 3 Diagnostic accuracy statistics for different cut-off values.

	Cut-off	Positive screens	SE	SP	PPV	NPV	DLR+	DLR−
λ_{AUC_max}								
SE = 0.75	0.141	50.9%	0.750 [0.644, 0.838]	0.539 [0.491, 0.586]	0.238 [0.205, 0.350]	0.918 [0.871, 0.932]	1.626 [1.386, 1.908]	0.464 [0.317, 0.679]
SE = 0.80	0.123	55.4%	0.798 [0.696, 0.877]	0.493 [0.445, 0.541]	0.232 [0.199, 0.354]	0.927 [0.881, 0.939]	1.574 [1.365, 1.814]	0.410 [0.266, 0.634]
SE = 0.85	0.106	62.8%	0.845 [0.750, 0.915]	0.411 [0.364, 0.459]	0.216 [0.184, 0.351]	0.933 [0.884, 0.944]	1.435 [1.272, 1.619]	0.377 [0.256, 0.629]
SE = 0.90	0.094	70.1%	0.905 [0.821, 0.958]	0.333 [0.293, 0.3807]	0.207 [0.175, 0.384]	0.948 [0.898, 0.957]	1.357 [1.233, 1.494]	0.286 [0.146, 0.560]
λ_{AUC_1SE}								
SE = 0.75	0.164	42.3%	0.750 [0.644, 0.838]	0.639 [0.592, 0.684]	0.285 [0.246, 0.408]	0.930 [0.889, 0.942]	2.079 [1.745, 2.478]	0.391 [0.268, 0.570]
SE = 0.80	0.160	47.7%	0.798 [0.696, 0.877]	0.571 [0.523, 0.618]	0.263 [0.229, 0.396]	0.936 [0.895, 0.947]	1.858 [1.595, 2.164]	0.355 [0.230, 0.546]
SE = 0.85	0.154	55.9%	0.857 [0.764, 0.924]	0.498 [0.450, 0.546]	0.247 [0.213, 0.399]	0.948 [0.907, 0.957]	1.706 [1.502, 1.939]	0.287 [0.169, 0.489]
SE = 0.90	0.145	73.0%	0.905 [0.821, 0.958]	0.306 [0.263, 0.351]	0.200 [0.168, 0.375]	0.944 [0.890, 0.954]	1.304 [1.188, 1.431]	0.311 [0.159, 0.611]

95% confidence intervals are in squared brackets.

lost importance in the given sample of first graders. Kastner-Koller et al. (2023) showed via the general school screening SES that a prediction of spelling at the beginning of second grade can be successful (42% predicted variance in writing one year after assessment in kindergarten). Here, working memory was identified as the most important predictor variable for reading, writing and arithmetic, followed by letter knowledge and phonological awareness. The authors proposed that phonological working memory is an underlying ability for different cognitive skills regarding speech.

For phonological awareness, we suppose that the specific predictive power of phonological awareness lies both in the younger sample age at the time of the initial survey (kindergarten) and in the small number of variables that were surveyed. Concerning the SES, grammar and vocabulary factors were not included. Furthermore, according to the authors (Kastner-Koller et al., 2023), the SES is decisive for predicting general school performance problems rather than specific performance deficits, such as difficulties in learning to write. Similar to the present study, Ennemoser et al. (2012) emphasized the importance of linguistic skills (i.e., grammar and vocabulary) for writing in the 2nd grade, whereas our results show the importance of grammar for subsequent writing as early as the 1st grade.

The predictive power of RAN for reading is a robust finding (see Landerl et al., 2022 for a review) and was recently shown in an Austrian sample of first graders (Schöfl et al., 2023a, 2023b). For writing, RAN also accounts for variance (Ennemoser et al., 2012; Kim et al., 2013; Lervåg and Hulme, 2010). Nevertheless, within our analyses model, RAN lost significance in predicting spelling at the end of first grade.

For research question 2, we found a reasonable diagnostic accuracy for a comprehensive and economical assessment battery. Other German batteries (Fricke et al., 2016) found comparable rates, for more details of effect sizes for spelling see Appendix 1.

As shown for research question 3, no significant differences in diagnostic accuracy were found between subgroups. There are no differences for boys and girls and for children with German as first or second language.

This result of missing gender disparities is in line with for example Dutch studies (Keuning and Verhoeven, 2008) for elementary school children from second grade on. For younger children, results are diverse, Allred (1990) found that girls from the first grade on are better spellers. For German language, recent studies on normative test batteries from first to seventh grade support our results of no gender disparities.

Concerning differences in spelling of mono- versus bilingual children, we found no significant group difference between mono- and bilinguals. We suppose that the missing differences are due to the provided vocabulary in the subtest. The words are based on typical word structures in German. Therefore, it is important to know these linguistically motivated structures on which the selected words are based. Knowledge about these structures could be independent of the children's individual language biography. To our knowledge further research is missing.

In summary the strongest predictors for word spelling skills at the end of first grade are letter knowledge and grammar. As there are no group differences concerning gender and language (children with German as L1 or L2), the didactic implications do consider the individual differences in spelling performance, but gender and/or language related disparities can be disregarded.

Didactic implications: letter knowledge

Early letter knowledge is a robust surface variable in spelling (Cordewener et al., 2012), comparable to rapid naming in reading

(Lervåg and Hulme, 2009). The strong relationship between letter knowledge and spelling skills might reflect an already existing access to written language itself. Children who know many letters might be able to show—at least implicitly—some elementary understanding of the basic phonological principle that is characteristic of alphabetical writing systems (Treiman, 2013).

Correlations between early letter knowledge and reading acquisition were described about 20 years ago (Foulin, 2005; Leppänen et al., 2008; Muter et al., 2004). Letter knowledge reflects seemingly implicit knowledge about the structure of written language. Simple training that focuses only on individual letters, however, is not effective for reading acquisition (in German: Scheerer-Neumann, 2006). Intervention studies instead support the linking hypothesis that it is the combination of letter knowledge, phonological exercises and subsequently the writing of syllables and words (Moraske et al., 2018). Isolated and exhaustive practice of letter knowledge at the preschool level is not a useful approach. Although the predictive power of phonological awareness is weaker than expected, the importance of PA for school exercises remains significant. In particular, the training of letters in combination with phonological awareness exercises should be emphasized (Moraske et al., 2018). To support children in this skill, Schnitzler (2008) suggested a variety of exercises, for instance, identifying words with the same initial phoneme or changing the initial phonemes in words. According to Mayer (2016), phonological awareness of smaller units, such as phonemes, develops during written language acquisition. In pratice, we suggest working with the initial phoneme at the beginning of written language acquisition rather than with various phoneme positions within the word which are more difficult to detect (Schnitzler, 2008). The combination of phonological awareness (i.e., detecting phonemes) and the basic alphabetic principle (i.e., letter-sound correspondence) is crucial for written language acquisition (Moraske et al., 2018). As the children are gradually introduced to letters and their names and sounds, initial phoneme detection is therefore to be systematically combined with the initial letter. This could be supported visually by the house-and-garage model mentioned above (Röber, 2013).

Regarding written language acquisition in grade 1, we infer that children with deficits in letter knowledge might benefit from integrated letter training that allows systematic insights not only into formal aspects, but also into aspects of the positioning of a letter within the syllable and the written canonical trochaic form of the German writing system. The letter <e> and its pronunciation depending on whether it is positioned in the strong or weak syllable might be a good example (Bredel, 2009; Fuhrhop and Peters, 2013). In practice, this could involve using use the house-and-garage model [originally designed for German trochaic words by Röber (2013)] to visualize for children the position of the latter within the syllable.

Didactic implications: grammar

For German spelling, oral grammar competence is an important precursor (von Goldammer et al., 2011), which is also confirmed by in the Schnapp spelling test. This might be explained by the strong grammatical (syllabic, morphological and syntactical) principles in the German writing system that act in concert. Theoretical approaches to grammar teaching and recent writing system research in German assume a strong relationship between grammar and literacy skills in general (Bredel, 2015b for an overview of the relationship between grammar and written language). Furthermore, the ability to recognize certain patterns in written language units is important and develops during written language acquisition. Dealing with written language has a number of prerequisite metalinguistic skills. Learning an alphabetical writing system then requires the abilities to identify sounds and to recognize how these sounds are transferred into written language. The ability to recognize certain patterns in written language units is also important (Haueis, 2015). Sufficient knowledge of grammar helps to detect syllable and morphological patterns in German word spelling and later, when dealing with sentences, to detect syntactic patters, such as space between words, German capitalization, and punctuation, when dealing with sentences (Bredel, 2015c). Morphological awareness is often cited as important for German word spelling (Görgen et al., 2021). Morphemes are units that are not only relevant for stem consistency, but also for identifying inflections. The latter is often placed in the weak *schwa* syllable. Stable access to stem consistency and how inflection is managed might improve spelling skills in general.

Grammar as a key precursor for German spelling adds weight to the argument that grammatical structures are very important in the German writing system. For children who show deficits when screened, knowledge about the canonical trochaic foot and related patterns allows them to gain insights into morphological principles far more easily (Röber, 2013; Bredel et al., 2017). Furthermore, working with syntactical structures, such as German V2-sentences, helps children to understand how sentences are constructed. To support language development in a practical way, children can be introduced to the topological field model (Granzow-Emden, 2015). For first graders, pictures or word cards could be used to form a spoken sentence instead of writing down sentences in the model and reading them aloud. Visualizing sentence structure promotes metalinguistic skills in children (Topalovic and Michalak, 2015).

Synopsis and limitations

The screening showed acceptable diagnostic accuracy [AUC = 0.757, DeLong 95% CI (0.698; 0.816)]. However, to achieve the desired sensitivity (>0.80), we must accept a relatively high rate of false positives (specificity = 0.57), which results in about one out of two pupils being screened positive. With a focus on intervention, this large group of screen-positives might thus challenge or even overtax teachers and school resources in providing support for these children. The problem of a high false-positive rate is not unique to our screening test (Cogo-Moreira et al., 2023; Duff et al., 2015; Johnson et al., 2009). Future research should therefore aim to improve specificity, for instance, by adding further subtests to the screening tool (Johnson et al., 2009). For example, reading socialization via parents or friends, motivational aspects, and further predictive child-related variables could be considered (only language-related variables were included in our approach). Specific attention processes or visual approaches were not considered, as the requisite data were not collected in this study. Regarding phonological awareness, only one subtest was included in the Schnapp screening battery. A wider approach to phonological awareness including both easier subtests, such as rhyming and syllable counting, and more difficult ones (e.g., manipulating phonemes) following the hierarchical structure of phonological tasks by Schnitzler (2008) should be analyzed.

Some researchers (Kim et al., 2013; Stainthorp et al., 2013) found orthographic awareness to be a significant precursor of early spelling skills. Actually, we do not include any task to tap orthographic awareness. Subsequent studies should add this variable for screening purposes.

Finally, continuous surveillance of spelling development is required because there may be various pathways to spelling difficulties. Many children with early difficulties do not go on to develop spelling problems, in the same way that many children who initially screen negative may later demonstrate spelling difficulties later, analogously to early screening for reading problems (Duff et al., 2015).

In summary our didactical implication to support written language acquisition for children in primary school is related to didactical implementations (Röber, 2013; Röber et al., 2019; Bredel, 2016; Bredel et al., 2017 based on the graphematical theory Primus, 2010; Evertz, 2016; Maas, 2022; Eisenberg, 2020; Fuhrhop and Peters, 2023)—both mentioned above. Children gain a comprehensive insight into the patterns of written language such as different sounds depending on letter positioning within the graphematic foot, how strong and weak syllables are presented in written language and how morphological stems are derivated from the graphematic foot, which is the initial point for constructing sentences.

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 humans were approved by Regional School Board for Upper Austria. The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation in this study was provided by the participants' legal guardians/next of kin.

Author contributions

MS: Conceptualization, Data curation, Investigation, Project administration, Writing – original draft, Writing – review & editing. GS: Conceptualization, Investigation, Project administration, Writing – original draft, Writing – review & editing. SZ: Writing – original draft, Writing – review & editing. CW: Data curation, Formal analysis, Methodology, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. All costs related to this research project were borne by the University of Education Upper Austria and Research Institute for Developmental Medicine (RID), Johannes Kepler University Linz. The tablets were provided by the project team with funds from the University of Education Upper Austria. Article processing charge was funded by the Johannes Kepler University Linz Open Access Publishing Fund.

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/feduc.2024.1378493/full#supplementary-material

References

Allred, R. A. (1990). Gender differences in spelling achievement in grades 1 through 6. *J. Educ. Res.* 83, 187–193. doi: 10.1080/00220671.1990.10885955

Bredel, U. (2009). Orthographie als System—Orthographieerwerb als Systemerwerb. *Z. Lit. Linguist.* 39, 135–155. doi: 10.1007/BF03379820

BMBWF. (2024). Förderorientiertes Schuleingangsscreenin. Available at: https://ses.bmbwf.gv.at (Accessed October 9, 2024).

Bredel, U. (2015a). "Der Aufbau von elementarem Wissen über die Systematik der Orthographie deutscher Wörter" in Schriftsprach- und Orthographieerwerb: Erstlesen, Erstschreiben. eds. C. Röber and H. Olfert (Baltmannsweiler: Schneider Verlag Hohengehren), 255–279.

Bredel, U. (2015b). "Grammatik und Literalität" in Deutschunterricht in Theorie und Praxis: Band 6. Deutschunterricht in Theorie und Praxis: Sprachreflexion und Grammatikunterricht. ed. H. Gornik. *2. korrigierte Auflage* ed (Baltmannsweiler: Schneider Verlag Hohengehren GmbH), 160–173.

Bredel, U. (2015c). "Sprachreflexion und Orthographie" in Deutschunterricht in Theorie und Praxis: Band 6. Deutschunterricht in Theorie und Praxis: Sprachreflexion und Grammatikunterricht. ed. H. Gornik. *2. korrigierte Auflage* ed (Baltmannsweiler: Schneider Verlag Hohengehren GmbH), 266–281.

Bredel, U. (2016). "23. Schriftspracherwerb" in Handbuch Laut, Gebärde, Buchstabe. eds. U. Domahs and B. Primus (Berlin: De Gruyter), 436–454.

Bredel, U., Fuhrhop, N., and Noack, C. (2017). Wie Kinder lesen und schreiben lernen. Tübingen: Narr Francke Attempo Verlag GmbH + Co. KG.

Breit, S., Bruneforth, M., and Schreiner, C. (2016). Standardüberprüfung Deutsch, 4. Schulstufe. Salzburg: BMBWF.

Caravolas, M., Hulme, C., and Snowling, M. J. (2001). The foundations of spelling ability: evidence from a 3-year longitudinal study. *J. Mem. Lang.* 45, 751–774. doi: 10.1006/jmla.2000.2785

Cogo-Moreira, H., Molinari, G. L., Carvalho, C. A. F., Kida, A. S. B., Lúcio, P. S., and Avila, C. R. B. (2023). Cut-off point, sensitivity and specificity for screening the reading fluency in children. *CoDAS* 35:e20210263. doi: 10.1590/2317-1782/20232021263en

Cordewener, K. A. H., Bosman, A. M. T., and Verhoeven, L. (2012). Predicting early spelling difficulties in children with specific language impairment: a clinical perspective. *Res. Dev. Disabil.* 33, 2279–2291. doi: 10.1016/j.ridd.2012.07.003

Denckla, M. B., and Rudel, R. (1974). Rapid "automatized" naming of pictured objects, colors, letters and numbers by normal children. *Cortex* 10, 186–202. doi: 10.1016/s0010-9452(74)80009-2

Duff, F. J., Mengoni, S. E., Bailey, A. M., and Snowling, M. J. (2015). Validity and sensitivity of the phonics screening check: implications for practice. *J. Res. Read.* 38, 109–123. doi: 10.1111/1467-9817.12029

Eisenberg, P. (1989). "Die Schreibsilbe im Deutschen" in Schriftsystem und Orthographie. eds. P. Eisenberg and H. Günther (Berlin: De Gruyter), 57–84.

Eisenberg, P. (2020). Grundriss der deutschen Grammatik. Berlin: J.B. Metzler Stuttgart.

Elliott, J. (2018). "Family engagement in urban education: the impact and factors affecting engagement" in Master of Education Applied Research Projects, 5. Available at: https://digitalcommons.cedarville.edu/education_research_projects/5

Enders, C. K. (2010). Applied missing data analysis. New York: Guilford Press.

Endlich, D., Küspert, P., Lenhard, W., Marx, P., and Schneider, W. (2019). Laute, Reime, Sprache—Würzburger Screening zur Früherkennung von Lese-Rechtschreibschwierigkeiten. (Berlin: De Gruyter).

Ennemoser, M., Marx, P., Weber, J., and Schneider, W. (2012). Spezifische Vorläuferfertigkeiten der Lesegeschwindigkeit, des Leseverständnisses und des Rechtschreibens. *Z. Entwicklungspsychologie Padagog. Psychol.* 44, 53–67. doi: 10.1026/0049-8637/a000057

Evertz, M. (2016). "20. Graphematischer Fuß und graphematisches Wort" in Handbuch Laut, Gebärde, Buchstabe. eds. U. Domahs and B. Primus (De Gruyter), 377–397.

Foulin, J. N. (2005). Why is letter-name knowledge such a good predictor of learning to read? *Read. Writ.* 18, 129–155. doi: 10.1007/s11145-004-5892-2

Fricke, S., Szczerbinski, M., Fox-Boyer, A., and Stackhouse, J. (2016). Preschool predictors of early literacy acquisition in German-speaking children. *Read. Res. Q.* 51, 29–53. doi: 10.1002/rrq.116

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* 33, 1–22. doi: 10.18637/jss.v033.i01

Friedman, J., Hastie, T., Tibshirani, R., Narasimhan, B., Tay, K., Simon, N., et al. (2021). glmnet: LASSO and elastic-net regularized generalized linear models. Heidelberg: J.B. Metzlersche Verlagsbuchhandlung. R Core Team.

Frith, U. (1986). A developmental framework for developmental dyslexia. *Ann. Dyslexia* 36, 67–81. doi: 10.1007/BF02648022

Fuhrhop, N., and Peters, J. (2013). Einführung in die Phonologie und Graphematik: J.B. Metzler Stuttgart.

Fuhrhop, N., and Peters, J. (2023). Einführung in die Phonologie und Graphematik. *2nd* Edn: Baltmannsweiler: Schneider Verlag Hohengehren. J.B. Metzler Stuttgart.

González-Valenzuela, M.-J., Lopez-Montiel, D., Díaz-Giraldez, F., and Martín-Ruiz, I. (2023). Word writing and cognitive predictors in Spanish at the age of seven. *Int. J. Early Child.* doi: 10.1007/s13158-023-00359-w

Görgen, R., De Simone, E., Schulte-Körne, G., and Moll, K. (2021). Predictors of reading and spelling skills in German: the role of morphological awareness. *J. Res. Read.* 44, 210–227. doi: 10.1111/1467-9817.12343

Granzow-Emden, M. (2015). "Sprachstrukturen verstehen: Die Entwicklung grammatischer Kategorien" in Deutschunterricht in Theorie und Praxis: Band 6. Deutschunterricht in Theorie und Praxis: Sprachreflexion und Grammatikunterricht. ed. H. Gornik. *2. korrigierte Auflage* ed (Schneider Verlag Hohengehren GmbH), 213–241.

Grob, A., Meyer, C., and Hagmann-von Arx, P. (2009). Intelligence and Development Scales (IDS). Intelligenz- und Entwicklungsskalen für Kinder von 5–10 Jahren. Bern: Hans Huber.

Haueis, E. (2015). "Sprachreflexion und Schriftspracherwerb" in Deutschunterricht in Theorie und Praxis: Band 6. Deutschunterricht in Theorie und Praxis: Sprachreflexion und Grammatikunterricht. ed. H. Gornik. *2. korrigierte Auflage* ed (Baltmannsweiler: Schneider Verlag Hohengehren GmbH), 184–193.

Henschel, S., Heppt, B., and Weirich, S. (2023). "Zuwanderungsbezogene Disparitäten" in IQB-Bildungstrend 2022: Sprachliche Kompetenzen am Ende der 9. Jahrgangsstufe im dritten Ländervergleich. eds. P. Stanat, S. Schipolowski, R. Schneider, S. Weirich, S. Henschel and K. A. Sachse (Münster: Waxmann), 299–344. Waxmann

Ibrahim, L. A., Hamann, C., and Öwerdieck, D. (2018). Identifying specific language impairment (SLI) across different bilingual populations: German sentence repetition task (SRT). In A. B. Bertolini and M. J. Kaplan (Eds.), Proceedings of the 42nd annual Boston University Conference on Language Development. 1–14. Cascadilla Press.

Jaeschke, R., Guyatt, G. H., and Sackett, D. L. (1994). Users' guides to the medical literature. III. How to use an article about a diagnostic test. B. What are the results and will they help me in caring for my patients? The evidence-based medicine working group. *JAMA* 271, 703–707. doi: 10.1001/jama.1994.03510330081039

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). An introduction to statistical learning, vol. *103*. New York: Springer.

Jansen, H., Mannhaupt, G., Marx, H., and Skowronek, H. (1999). Bielefelder Screening zur Früherkennung von Lese-Rechtschreibschwierigkeiten (BISC). Göttingen: Hogrefe.

Jöbstl, V., Steiner, A. F., Kastner-Koller, U., Deimann, P., Kaltenberger, A., Wagner, V., et al. (2022). Entwicklung eines förderorientierten Schuleingangsscreenings: Erste Befunde zur prognostischen Validität. *Frühe Bildung* 11, 194–200. doi: 10.1026/2191-9186/a000585

Johnson, E. S., Jenkins, J. R., Petscher, Y., and Catts, H. W. (2009). How can we improve the accuracy of screening instruments? *Learn. Disabil. Res. Pract.* 24, 174–185. doi: 10.1111/j.1540-5826.2009.00291.x

Kastner-Koller, U., Kaltenberger, A., Jöbstl, V., Landerl, K., Steiner, A. F., Wagner, V., et al. (2023). Zur Vorhersage des schulischen Lern- und Arbeitsverhaltens durch domänenübergreifende Vorläuferfähigkeiten. *Lernen Und Lernstörungen* 12, 115–125. doi: 10.1024/2235-0977/a000412

Kessler, B., Pollo, T., Treiman, R., and Cardoso-Martins, C. (2013). Frequency analyses of prephonological spellings as predictors of success in conventional spelling. *J. Learn. Disabil.* 46, 252–259. doi: 10.1177/0022219412449440

Keuning, J., and Verhoeven, L. (2008). Spelling development throughout the elementary grades: the Dutch case. *Learn. Individ. Differ.* 18, 459–470. doi: 10.1016/j.lindif.2007.12.001

Kim, Y.-S., Apel, K., and Al Otaiba, S. (2013). The relation of linguistic awareness and vocabulary to word reading and spelling for first-grade students participating in response to intervention. *Lang. Speech Hear. Serv. Sch.* 44, 337–347. doi: 10.1044/0161-1461(2013/12-0013)

Landerl, K., Castles, A., and Parrila, R. (2022). Cognitive precursors of reading: a cross-linguistic perspective. *Sci. Stud. Read.* 26, 111–124. doi: 10.1080/10888438.2021.1983820

Landerl, K., Ramus, F., Moll, K., Lyytinen, H., Leppänen, P. H. T., Lohvansuu, K., et al. (2013). Predictors of developmental dyslexia in European orthographies with varying complexity. *J. Child Psychol. Psychiatry* 54, 686–694. doi: 10.1111/jcpp.12029

Lenhart, J., Marx, P., Segerer, R., and Schneider, W. (2019). Rechtschreibung ohne Schreiben. *Diagnostica* 65, 216–227. doi: 10.1026/0012-1924/a000229

Leppänen, U., Aunola, K., Niemi, P., and Nurmi, J.-E. (2008). Letter knowledge predicts grade 4 reading fluency and reading comprehension. *Learn. Instr.* 18, 548–564. doi: 10.1016/j.learninstruc.2007.11.004

Lervåg, A., and Hulme, C. (2009). Rapid automatized naming (RAN) taps a mechanism that places constraints on the development of early reading fluency. *Psychol. Sci.* 20, 1040–1048. doi: 10.1111/j.1467-9280.2009.02405.x

Lervåg, A., and Hulme, C. (2010). Predicting the growth of early spelling skills: are there heterogeneous developmental trajectories? *Sci. Stud. Read.* 14, 485–513. doi: 10.1080/10888431003623488

López-Ratón, M., Rodríguez-Álvarez, M. X., Suárez, C. C., and Sampedro, F. G. (2014). OptimalCutpoints: AnRPackage for selecting optimal cutpoints in diagnostic tests. *J. Stat. Softw.* 61, 1–36. doi: 10.18637/jss.v061.i08

Maas, U. (1992). Grundzüge der deutschen Orthographie. *De Gruyter.* Berlin: De Gruyter. doi: 10.1515/9783111376974

Maas, U. (2022). "Laute und Buchstaben—zu den phonographischen Grundlagen des Schrifterwerbs" in Deutschunterricht in Theorie und Praxis: Vol. 2. Schriftsprach- und Orthographieerwerb: Erstlesen, Erstschreiben. eds. C. Röber and H. Olfert, Baltmannsweiler: Schneider Verlag Hohengehren GmbH.113–139.

Martschinke, S., Kirschhock, E., and Frank, A. (2001). Der Rundgang durch Hörhausen: Erhebungsverfahren zur phonologischen Bewusstheit (1. Aufl.). Diagnose und Förderung im Schriftspracherwerb: Bd. 1. Donauwörth: Auer.

Marx, P., and Lenhard, W. (2010). "Diagnostische Merkmale von Screeningverfahren" in Frühprognose schulischer Kompetenzen. eds. M. Hasselhorn and W. Schneider (Göttingen: Hogrefe), 68–84.

Mayer, A. (2016). Lese-Rechtschreibstörungen (LRS). München, Basel: Ernst Reinhardt Verlag.

McNeish, D. M. (2015). Using LASSO for predictor selection and to assuage overfitting: a method long overlooked in behavioral sciences. *Multivar. Behav. Res.* 50, 471–484. doi: 10.1080/00273171.2015.1036965

Moraske, S., Wyschkon, A., Poltz, N., Kucian, K., von Aster, M., and Esser, G. (2018). LRS-Prävention bei Risikokindern: Langfristige Effekte bis in die 3. Klasse. *Lernen Und Lernstörungen* 7, 171–183. doi: 10.1024/2235-0977/a000213

Muter, V., Hulme, C., Snowling, M. J., and Stevenson, J. (2004). Phonemes, rimes, vocabulary, and grammatical skills as foundations of early reading development:

evidence from a longitudinal study. *Dev. Psychol.* 40, 665–681. doi: 10.1037/0012-1649.40.5.665

Nag, S., Vagh, S. B., Dulay, K. M., and Snowling, M. J. (2019). Home language, school language and children's literacy attainments: a systematic review of evidence from low- and middle-income countries. *Rev. Educ.* 7, 91–150. doi: 10.1002/rev3.3130

Niemietz, J., Jindra, C., Schneider, R., Schumann, K., Schipolowski, S., and Sachse, K. A. (2023). "Soziale Disparitäten" in IQB-Bildungstrend 2022: Sprachliche Kompetenzen am Ende der 9. Jahrgangsstufe im dritten Ländervergleich. eds. P. Stanat, S. Schipolowski, R. Schneider, S. Weirich, S. Henschel and K. A. Sachse (Paris: OECD Publishing), 261–298. doi: 10.1787/53f23881-en

Niolaki, G. Z., Papadimitriou, V., Terzopoulos, A. R., and Masterson, J. (2024). Greek-spelling predictors; an investigation of literacy- and cognitive-related factors. *J. Cogn. Psychol.* 36, 378–393. doi: 10.1080/20445911.2024.2322766

Niolaki, G. Z., Vousden, J., Terzopoulos, A. R., Taylor, L. M., Sephton, S., and Masterson, J. (2020). Predictors of single word spelling in English speaking children: a cross sectional study. *J. Res. Read.* 43, 577–596. doi: 10.1111/1467-9817.12330

OECD (2023). PISA 2022 results (volume I): the state of learning and equity in education: OECD Publishing.

Pan, S. C., Rickard, T. C., and Bjork, R. A. (2021). Does spelling still matter—and if so, how should it be taught? *Educ Psychol Rev* 33, 1523–1552. doi: 10.1007/s10648-021-09611-y

Pepe, M. S. (2004). The statistical evaluation of medical tests for classification and prediction: Oxford University Press.

Primus, B. (2010). "Strukturelle Grundlagen des deutschen Schriftsystems" in Schriftsystem und Schrifterwerb: linguistisch—didaktisch—empirisch. eds. U. Bredel, A. Müller and G. Hinney (Berlin: Walter de Gruyter GmbH & Co. KG), 9–45.

Röber, C. (2013). Die Leistungen der Kinder beim Lesen- und Schreibenlernen. Grundlagen der Silbenanalytischen Methode: Ein Arbeitsbuch mit Übungsaufgaben. Hohengehren: Schneider-Verlag Hohengehren.

Röber, C., Häusle, R., and Berchtold, M. (2019). "Erstklässler entdecken die Orthographie als Hilfe für das Rechtlesen und Rechtschreiben: Vorstellung des Schriftsprachstrukturierenden Konzeptes PALOPE" in Forschung & Entwicklung. Feldkirch: Forschungszeitschrift. 47–60.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., et al. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* 12:77. doi: 10.1186/1471-2105-12-77

Schaumburg, H. (2015). Chancen und Risiken digitaler Medien in der Schule: Medienpädagogische und -didaktische Perspektiven. Berlin: Bertelsmann Stiftung.

Scheerer-Neumann, G. (2006). "Entwicklung der basalen Lesefähigkeit" in Didaktik der deutschen Sprache—Band 1. eds. U. Bredel, H. Günther, P. Klotz, J. Ossner and G. Siebert-Ott (Stuttgart, Deutschland: UTB GmbH), 513–524.

Schnitzler, C. (2008). Phonologische Bewusstheit und Schriftspracherwerb: 31 Tabellen. Forum Logopädie. Stuttgart, New York, NY: Thieme.

Schöfl, M., Seifert, S., Steinmair, G., and Weber, C. (2023a). Are both object and alphanumeric rapid automatized naming measures required to predict word Reading fluency in German prereaders? *J. Psychoeduc. Assess.* 41, 811–827. doi: 10.1177/07342829231185375

Schöfl, M., Steinmair, G., Holzinger, D., and Weber, C. (2022). Predicting word reading deficits using an app-based screening tool at school entry. *Front. Pediatr.* 10:863477. doi: 10.3389/fped.2022.863477

Schöfl, M., Steinmair, G., Zepnik, S., and Gnambs, T. (2023b). Entwicklung eines digitalen Rechtschreibtests für die erste Klasse Grundschule: Dimensionalität und Reliabilität des SCHNAPP-Rechtschreibtests. *Lernen Und Lernstörungen* 12, 188–200. doi: 10.1024/2235-0977/a000404

Schroeder, S., Würzner, K.-M., Heister, J., Geyken, A., and Kliegl, R. (2015). childLex–Eine lexikalische Datenbank zur Schriftsprache für Kinder im Deutschen. *Psychol. Rundsch.* 66, 155–165. doi: 10.1026/0033-3042/a000275

Seifert, S., Paleczek, L., and Gasteiger-Klicpera, B. (2019). Rezeptive Wortschatzleistungen in der Grundschule: Unterschiede zwischen Kindern mit Deutsch als Erst- und Zweitsprache. *Empirische Sonderpädagogik* 4, 259–278. doi: 10.25656/01:18334

Silverman, R. D., McNeish, D., Speece, D. L., and Ritchey, K. D. (2021). Early screening for decoding- and language-related reading difficulties in first and third grades. *Assess. Eff. Interv.* 46, 99–109. doi: 10.1177/1534508419857234

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

Stainthorp, R., Powell, D., and Stuart, M. (2013). The relationship between rapid naming and word spelling in English. *J. Res. Read.* 36, 371–388. doi: 10.1111/jrir.12002

Statistik Austria. (2024). STATcube—Statistische Datenbank von STATISTIK AUSTRIA. Available at: https://statcube.at/

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science* 240, 1285–1293. doi: 10.1126/science.3287615

Thomé, G. (2006). "Entwicklung der basalen Rechtschreibkenntnisse" in Didaktik der deutschen Sprache—Band 1. eds. U. Bredel, H. Günther, P. Klotz, J. Ossner and G. Siebert-Ott (Stuttgart, Deutschland: UTB GmbH), 369–379.

Thompson, R. A. (2015). "Relationships, regulation, and early development" in Handbook of child psychology and developmental science: socioemotional processes. eds. M. E. Lamb and R. M. Lerner. *7th* ed (Oxford: John Wiley & Sons), 1–246. doi: 10.1002/9781118963418

Topalovic, E., and Michalak, M. (2015). "Sprachreflexion und Grammatik zwischen DaM und DaZ" in Grundlagen der Sprachdidaktik Deutsch als Zweitsprache. eds. M. Michalak and M. Kuchenreuther. *3. überarbeitete Auflage* ed (Baltmannsweiler: Schneider Verlag Hohengehren GmbH), 226–250.

Treiman, R. (2013). "Knowledge about letters as a foundation for reading and spelling" in Handbook of orthography and literacy. eds. R. M. Joshi and P. G. Aaron (New York: Routledge).

Treiman, R., and Bourassa, D. (2000). Children's written and oral spelling. *Appl. Psycholinguist.* 21, 183–204. doi: 10.1017/S0142716400002022

Treiman, R., Hulslander, J., Olson, R. K., Samuelsson, S., Elwér, Å., Furnes, B., et al. (2023). Predicting later spelling from kindergarten spelling in U.S., Australian, and Swedish children. *Reading* 27, 428–442. doi: 10.1080/10888438.2023.2186234

Treiman, R., Hulslander, J., Olson, R. K., Willcutt, E. G., Byrne, B., and Kessler, B. (2019). The unique role of early spelling in the prediction of later literacy performance. *Sci. Stud. Read.* 23, 437–444. doi: 10.1080/10888438.2019.1573242

Treiman, R., and Kessler, B. (2004). The case of case: children's knowledge and use of upper- and lowercase letters. *Appl. Psycholinguist.* 25, 413–428. doi: 10.1017/S0142716404001195

Treiman, R., Kessler, B., Decker, K., and Pollo, T. C. (2016). How do prephonological writers link written words to their objects? *Cogn. Dev.* 38, 89–98. doi: 10.1016/j.cogdev.2016.02.002

Venkatraman, E. S. (2000). A permutation test to compare receiver operating characteristic curves. *Biometrics* 56, 1134–1138. doi: 10.1111/j.0006-341x.2000.01134.x

von Goldammer, A., Mähler, C., Bockmann, A. K., Hasselhorn, M., and Hasselhorn, M. (2010). Vorhersage früher Schriftsprachleistungen aus vorschulischen Kompetenzen der Sprache und der phonologischen Informationsverarbeitung. *Z. Entwicklungspsychologie Padagog. Psychol.* 42, 48–56. doi: 10.1026/0049-8637/a000005

von Goldammer, A., Mähler, C., Hasselhorn, M., and Hasselhorn, M. (2011). Determinanten von Satzgedächtnis-Leistungen bei deutsch- und mehrsprachigen Vorschulkindern. *Z. Entwicklungspsychologie Padagog. Psychol.* 43, 1–15. doi: 10.1026/0049-8637/a000028

Wagner, R. K., and Torgesen, J. K. (1987). The nature of phonological processing and its causal role in the acquisition of reading skills. *Psychol. Bull.* 101, 192–212. doi: 10.1037/0033-2909.101.2.192

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

Youngstrom, E. A. (2014). A primer on receiver operating characteristic analysis and diagnostic efficiency statistics for pediatric psychology: we are ready to ROC. *J. Pediatr. Psychol.* 39, 204–221. doi: 10.1093/jpepsy/jst062

Yulianti, K., Denessen, E., and Droop, M. (2023). The effects of parental involvement on children's education. *Int. J. Parents Educ.* 10, 14–32. doi: 10.54195/ijpe.14123

Zarić, J., Hasselhorn, M., and Nagler, T. (2020). Orthographic knowledge predicts reading and spelling skills over and above general intelligence and phonological awareness. *Psychol. Educ.* 36, 21–43. doi: 10.25656/01:22749

Frontiers in Education

Explores education and its importance for individuals and society

A multidisciplinary journal that explores research-based approaches to education for human development. It focuses on the global challenges and opportunities education faces, ultimately aiming to improve educational outcomes.

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