frontiers | Frontiers in Education

TYPE Original Research
PUBLISHED 17 September 2025
DOI 10.3389/feduc.2025.1644811

Measuring cognitive levels in high-stakes testing: A CDM analysis of a university entrance examination using Bloom's Taxonomy

Hamdollah Ravand[1], Reza Shahi[2], Farshad Effatpanah[3]* and Ali Moghadamzadeh[4]

[1]Department of English, Vali-e-Asr University of Rafsanjan, Rafsanjan, Iran, [2]Department of English, Ilam University, Ilam, Iran, [3]Faculty of Rehabilitation Sciences, TU Dortmund University, Dortmund, Germany, [4]Faculty of Psychology and Education, University of Tehran, Tehran, Iran

OPEN ACCESS

EDITED BY
Peida Zhan,
Zhejiang Normal University, China

REVIEWED BY
Kathy Ellen Green,
University of Denver, United States
Francis K. Awuah,
Kwame Nkrumah University of Science and Technology, Ghana

*CORRESPONDENCE
Farshad Effatpanah
✉ farshad.effatpanah@tu-dortmund.de

RECEIVED 10 June 2025
ACCEPTED 15 August 2025
PUBLISHED 17 September 2025

CITATION
Ravand H, Shahi R, Effatpanah F and Moghadamzadeh A (2025) Measuring cognitive levels in high-stakes testing: A CDM analysis of a university entrance examination using Bloom's Taxonomy.
Front. Educ. 10:1644811.
doi: 10.3389/feduc.2025.1644811

The aim of this study was twofold. First, this study explored the potential use of a high-stakes multiple-choice test for measuring cognitive complexity by using Bloom's Taxonomy and applying cognitive diagnostic models. Second, it investigated the interplay of cognitive complexity with gender and item difficulty. Data from 1,000 applicants to English PhD programs were analyzed. Six experts coded test items based on the cognitive levels they target. Q-matrices were constructed, one for each expert, specifying item-cognition relationships. The G-DINA model was used to assess these relationships. Based on the best-fitting Q-matrix, 27% of the items measured the lowest cognitive level (Remember), 50% measured Understand, and 23% measured Analyze levels. Test takers demonstrated mastery of these levels by 56, 39, and 28%, respectively. Findings indicated that the test primarily assessed lower levels of Bloom's Taxonomy. In addition, the results showed that male test takers outperformed female counterparts at higher levels. Furthermore, the analysis showed that cognitive complexity contributed to item difficulty. Finally, implications were discussed, and suggestions were made.

KEYWORDS

Bloom's Taxonomy, levels of cognition, MCQ, item difficulty, gender, cognitive diagnostic models

1 Introduction

The key role that high-stakes tests play in test takers lives necessitates a critical assessment to ensure that the tests accurately measure student abilities (Kubiszyn and Borich, 2024; French et al., 2023; Mason, 2007). Whereas high-stakes tests have traditionally been aligned with the rote memorization of facts, they should ideally make a switch to being comprehensive measures of student capability. This means shifting toward the assessment of students for their critical thinking and problem-solving capabilities, and that is achieved through focusing on the cognitive complexity of test items. This issue has accordingly received extensive research attention (e.g., Ehrich et al., 2021; Noroozi and Karami, 2022), and frameworks like the one employed in international assessments such as Trends in International Mathematics and Science Study (TIMSS) 2019 exemplify this focus by incorporating a cognitive dimension alongside the traditional content dimension. This dual focus ensures that in this way a more holistic evaluation of the students' capabilities takes place, preparing them for the demands of the 21st century more effectively (Gorin, 2006; Shavelson, 2010).

Bloom's Taxonomy provides a well-established framework for classifying and categorizing cognitive objectives in education and assessment (Alshurafat et al., 2024; Anderson and Krathwohl, 2001; Krathwohl, 2002, Krathwohl and Anderson, 2009; Pellegrino et al., 2016; Ramirez, 2017). This framework subdivides the cognitive domains into six categories of Remembering, Comprehending, Applying, Analyzing, Evaluating, and Creating (Anderson and Krathwohl, 2001; Das et al., 2022). These levels range from lower-order thinking (LOT) skills to higher-order thinking (HOT) skills (Freahat and Smadi, 2014; Krause et al., 2021). According to Muhayimana et al. (2022), LOT skills require recalling and comprehending information (remembering and applying knowledge), while HOT skills demand deeper analysis, application, and evaluation (analyzing, evaluating, and creating). Assessing HOT skills is crucial as they reflect critical thinking, problem-solving, and innovation—capabilities essential for preparing 21st-century learners to address complex socio-scientific challenges (Rahayu and Alsulami, 2024). The emphasis on specific HOT skills may vary across different subject domains, reflecting the nature of authentic problems encountered in each discipline. A large body of research has investigated cognitive levels in high-stakes tests across various contexts (e.g., Baghaei et al., 2021; DeWitt et al., 2013; Jung Lim, 2014; Ho, 2022). However, most studies focused solely on identifying Bloom's Taxonomy in analyzed tests rather than assessing how effectively the tests assess the targeted cognitive skills. Furthermore, nearly all of these studies used a qualitative method and descriptive statistics as the means of analyzing the data of their research.

In addition, for decades, Multiple-Choice Questions (MCQs) have been the dominant format in large-scale testing (Tractenberg et al., 2013). While MCQs are undoubtedly effective in assessing factual knowledge and information recall, there's a growing body of studies suggesting that MCQs may also measure complex capabilities beyond simple recall (e.g., Cecilio-Fernandes et al., 2018; Crowe et al., 2008; Douglas et al., 2012; Jensen et al., 2014; Karpen and Welch, 2016; Kim et al., 2012; Kıyak et al., 2022; Thompson and O'Loughlin, 2015; Thompson et al., 2016; Yeong et al., 2020). However, this evidence remains inconclusive, and significant gaps in our understanding persist.

Moreover, research on the impact of gender on cognitive abilities is complex and ongoing (e.g., Ernawati and Baharullah, 2020; Migliaccio et al., 2009; Nasution et al., 2023; Ryan and Chiu, 2001). However, findings are inconsistent: some studies suggest a potential male advantage in assessments of HOT skills (e.g., Amin et al., 2024; Lager et al., 2024; Wright et al., 2016), while others report no significant differences or even better performance by females in specific contexts (e.g., Aldila et al., 2013; Araiku et al., 2019; Bastick, 2002). These inconsistencies highlight the need for further exploration considering factors beyond gender alone, such as the specific cognitive skills assessed, subject matter, and test format.

Furthermore, recent development of Cognitive Diagnostic Models (CDMs) provides a promising approach to analyze cognitive processes that are measured by tests (DiBello and Stout, 2007). Unlike traditional assessments that view measuring learning as the sole purpose of assessment, CDMs go deep into investigating underlying knowledge and cognitive strategies needed to do certain tasks correctly (Ravand, 2016). The rich diagnostic information provided can be used to improve learning and instruction through pinpointing areas where students might struggle (Leighton and Gierl, 2007). In addition, CDMs show great potential for changing assessment practices in the humanities by allowing the light to be shed on the hard-to-measure cognitive skills that students acquire in these disciplines.

This study aims to address existing gaps by integrating CDMs with Bloom's Taxonomy to analyze the cognitive processes assessed by the Iranian National PhD Entrance Exam (INPEE), a high-stakes MC test designed to evaluate the readiness of Iranian Master graduates for PhD programs. The general English section of the INPEE, encompassing grammar, vocabulary, and reading comprehension, plays a pivotal role in determining candidates' success.

By examining the types of knowledge evaluated in this section, the study seeks to identify the cognitive levels measurable through MCQs in the field of English Language Teaching. Additionally, it aims to demonstrate the application of CDMs in analyzing exams in alignment with desired learning outcomes, potentially revealing the genuine competencies and knowledge of English students in a high-stakes testing context. Furthermore, the research also compares the performance of male and female applicants across different levels of cognitive complexity and explores the relationship between item difficulty and cognitive complexity.

2 Literature review

Analyzing cognitive levels of test items is crucial for effective teaching, learning and testing. Measuring the complexity of thinking processes that test takers need to answer questions can help educators to create test items that clearly reflect the students' abilities and inform instructional practices. A large array of research studies supports the importance of this approach (Bezuidenhout and Alt, 2011; Borda et al., 2018; Das et al., 2020; Gorin, 2007; Mustafidah and Pinandita, 2022). This approach sufficiently assists in determining students' abilities and weaknesses and thus makes assessment fairer and broadens the learner's apprehension of the subject matter. Hence, it ultimately improves learning outcomes when assessment is aligned to the cognitive level.

2.1 Bloom's Taxonomy

Bloom's Taxonomy is a framework to analyze and classify cognitive objectives in different educational fields. Bloom et al. (1956) proposed the framework, which according to Anderson and Krathwohl (2001) has established itself as a necessary tool for evaluating educational and testing objectives. To shed light on the learning process and the range of cognitive abilities among students, researchers have thoroughly examined its use in various educational settings (e.g., Crowe et al., 2008; Elim, 2024; Granello, 2001; Kahn, 2012; Momen et al., 2022; Scott, 2003; Swart, 2009). This taxonomy has been utilized to analyze the cognitive level of textbooks (Khoy, 2025; Mizbani et al., 2023; Razmjoo and Kazempourfard, 2012; Rena et al., 2023; Parsaei et al., 2017; Zorluoglu et al., 2020), final exam questions (Chang and Chung, 2009; Febriyana and Harjanto, 2023; Ebadi and Shahbazian, 2015; Khorsand, 2009; Shahbazian, 2016; Tangsakul et al., 2017), university entrance exams (Aydin and Birgili, 2023; Han and Xiang, 2025; Razmjoo and Madani, 2013; Rezaee and Golshan, 2016). In addition, it has been used to enhance the development of intelligent assessment tools (Jaramillo and Cadavid, 2015; Kosorus and Küng, 2014; Krouska et al., 2018; Tíjaro-Rojas et al., 2016; Ying and Yang, 2008), and finally to improve teaching practices and student performance (e.g., Butler, 2018; Niyibizi et al., 2018; Nkhoma et al., 2017; Rentmeester, 2018). Extending this framework to language assessment, several studies have explored the cognitive processes involved in language exams like IELTS and TOEFL (e.g., Baghaei et al., 2020, 2021; Bax, 2013; Moslehi and Razmjoo, 2021; NamazianDoost and HayaviMehr, 2017). These studies primarily focused on the reading and listening sections, employing methods like eye-tracking, retrospective questionnaires and content analysis.

2.2 Multiple-choice questions for measuring cognitive level

MCQs are a widely used assessment tool, but their ability to measure complex thinking skills remains a topic of debate. While some studies highlight their potential for assessing higher-order cognition, others emphasize significant limitations. Proponents argue that well-constructed MCQs can effectively measure Application, Analysis, and Synthesis. Jensen et al. (2014) emphasize that with careful construction, MCQs can assess Application, Analysis, and even Synthesis of Knowledge—key components of HOT. Similarly, Kıyak et al. (2022) advocate for improving MCQ design to test Critical Thinking rather than being limited to Factual Knowledge. Kim et al. (2012) found that 13% of MCQs in their study reached the highest levels of Bloom's Taxonomy, suggesting that although challenging, it is possible to create MCQs that evaluate advanced cognitive processes. Zaidi et al. (2018) proposed categorizing MCQs into lower and higher cognitive levels, asserting that they can align with up to four levels of Bloom's Taxonomy. Ten Cate et al. (2018) further supported this view by stating that well-crafted MCQs engage learners in Applying and Synthesizing knowledge, which are essential higher-order cognitive functions.

However, other studies highlight limitations in using MCQs for higher-order assessment. Tarrant and Ware (2008) note that most MCQs predominantly focus on Factual Recall rather than Application or Evaluation. Tarrant et al. (2009) reinforce this idea, pointing out that due to logistical and practical constraints, MCQs often remain confined to testing basic knowledge. Similarly, Stanger-Hall (2012) reported that MCQs indeed hinders Critical Thinking. Zheng et al. (2008) conclude that MCQs are mostly effective only at the first two levels of Bloom's Taxonomy ('Apply' and 'Analyse'), falling short in evaluating more complex skills like 'Evaluate' and 'Create'. Harland and Wald (2021), along with Scouller (1998), argue that essay-based assessments allow for better demonstration of higher-order cognition compared to MCQs, which often promote surface-level memorization rather than deep understanding. Liu et al. (2024) found that while many educators believe MCQs can effectively test middle-level cognitive functions like Application and Analysis, there is skepticism about their capacity to measure the highest levels of Bloom's Taxonomy—Evaluation and Creation.

Despite these contrasting perspectives, there appears to be agreement on one key point: the effectiveness of MCQs in measuring higher-order cognitive skills depends heavily on the quality of question design and alignment with intended learning outcomes. Studies such as Ali and Ruit (2015), Billings et al. (2016), and Kibble (2017) support the potential of MCQs to test problem-solving and critical thinking when questions are constructed with deliberate attention to cognitive complexity. Choudhury and Freemont (2017) demonstrate that with proper framing, MCQs can assess higher cognitive functions. Moreover, Zaidi et al. (2016, 2017, 2018) developed frameworks at the University of Michigan Medical School to help faculty write MCQs targeting higher-order cognition. Santen et al. (2019) also encourage efforts to develop MCQs that test clinical reasoning and decision-making skills, which are essential in medical education. On the other hand, Monrad et al. (2021) caution that without shared understanding and training among educators, the classification and design of higher-order MCQs remain inconsistent and subjective and the effectiveness of this approach remains unclear.

2.3 Cognitive level and item difficulty

Item difficulty estimation has been a central focus in educational measurement, particularly in the context of MC items. Some studies have investigated the connection between cognitive processing models and the difficulty level of items. Embretson and Wetzel (1987) were among the first to propose a cognitive processing model for paragraph comprehension items, identifying decision processes such as falsification and confirmation as strong predictors of item difficulty. Similarly, Gorin and Embretson (2006) found that response-decision processes in Graduate Record Examinations (GRE) paragraph comprehension sections had a stronger association with item difficulty than general text-comprehension ability, reinforcing the role of higher-order cognitive strategies.

Kirsch et al. (2017) introduced a construct-based assessment model that defines operationalized variables tied to task characteristics, including "processes and strategies." These strategies involve locating, cycling, integrating, and generating information, as well as the abstractness or concreteness of required knowledge and the plausibility of distractors. This model underscores how cognitive strategies and semantic complexity directly affect item difficulty. Ferrara et al. (2011) further defined "item demands" as the knowledge, comprehension, and cognitive processes required for answering correctly, highlighting the necessity of aligning item design with cognitive expectations. Hsu et al. (2018) also supported this connection and reported that there is a link between cognitive processing models and item difficulty.

Rush et al. (2016) found that higher cognitive skills are needed for more difficult items in veterinary exams. They also identified pitfalls like implausible distractors that decrease item quality and discrimination. Their study emphasizes the importance of well-designed MCQs with effective distractors to accurately assess HOT (Chang and Chung, 2009; Febriyana and Harjanto, 2023; Ebadi and Shahbazian, 2015; Khorsand, 2009; Shahbazian, 2016; Tangsakul et al., 2017). Researchers also have investigated the connection between Bloom's Taxonomy and item quality. Testa et al. (2018) reported that higher levels of cognitive processing enable test developers to produce more functioning distractors. They also reported that there is a significant relation between distractor efficiency and item difficulty. However, contrasting with the aforementioned studies, some research has reported no consistent relationship between cognitive complexity levels, as specified in Bloom's Taxonomy, and item difficulty or discrimination. For instance, Kibble and Johnson (2011) found no significant correlation between the cognitive level of MCQs and their difficulty or discrimination indices. Similarly, Tan and Othman (2013) observed that items categorized into different Bloom's Taxonomy levels did not exhibit a strong relationship with item difficulty.

2.4 Gender differences in cognition

The relationship between gender and cognitive abilities is a complex and continually evolving area of research with studies often yielding contradictory results. While some studies report male advantages in higher-order cognitive tasks across various domains (Amin et al., 2024; Lager et al., 2024; Wright et al., 2016), others indicate female advantages (Aldila et al., 2013; Bastick, 2002) or find no significant differences (Araiku et al., 2019). The pertinent literature reveals several lines of inquiry, each characterized by its specific research purpose. The first comprises studies exploring how gender influences cognitive performance at varying levels of Bloom's Taxonomy. For instance, Bastick (2002) specifically explored whether different objective test question formats might favor males or females and if responses to questions assessing abilities at different levels of Bloom's cognitive domain varied by gender. Each subtest comprised six questions, with each question designed to target a specific level of Bloom's Taxonomy. A comparison of mean male and female scores across three subtest formats revealed only one statistically significant advantage: females excelled in matching questions. This advantage was attributed to significant female strengths at the Analysis and Synthesis levels within matching questions. Aldila et al. (2013) investigated the cognitive and attitudinal outcomes of secondary school students learning about global warming using the Student Team Achievement Division (STAD) method, based on gender. Across all cognitive levels (C1 to C6) of Bloom's Taxonomy, female students consistently outperformed male students, achieving higher N-Gain values in Remembering, Understanding, Applying, Analyzing, Evaluating, and Creating. In another study, Wright et al. (2016) examined how the cognitive difficulty and format of exams affect student performance in introductory biology courses, focusing on gender and socioeconomic status (SES) differences. They found that male students outperformed female students on exams that tested higher levels of Bloom's Taxonomy, a difference observed even after controlling for prior academic ability. This performance gap increased as the Bloom's level of the exam increased. Moreover, Araiku et al. (2019) provided further insights into gender differences in mathematical abilities across various levels of Bloom's Taxonomy. Their research, involving 156 junior high school students and analyzed using a two-way analysis of variance (ANOVA) with *post hoc* tests, indicated no significant overall performance difference between male and female students. However, male students specifically outperformed female students at the C1 (Remember) level. This suggests that gender might influence performance at certain cognitive levels but not others. The study underscores the importance of considering the interaction between gender and cognitive levels when designing educational strategies, highlighting that while gender can be a factor at specific Bloom's Taxonomy levels, it does not necessarily impact overall academic achievement. This nuanced understanding is vital for educators aiming to cater to the diverse needs of male and female students in mathematics. Amin et al. (2024) also offered valuable insights into gender differences within Bloom's Taxonomy, particularly among prospective teachers in Pakistan. They discovered that female participants demonstrated stronger mastery of lower-level skills, such as Remembering and Understanding, whereas male participants showed greater proficiency in HOT skills, like Analyzing and Evaluating.

The second line of research assesses gender differences across distinct cognitive domains (e.g., memory, spatial skills) and consistency of performance. For example, Lager et al. (2024)'s comprehensive analysis of gender differences in operational and cognitive abilities revealed that male candidates scored significantly higher on mental spatial ability, memory retention, abstract problem-solving, multitasking ability, and manual spatial ability. In contrast, female candidates scored higher on perceptual speed. The study also highlighted that correlations between different cognitive abilities were significantly stronger among female candidates, indicating a more homogeneous performance profile. This implies that females may exhibit a more consistent application of cognitive skills across various domains, a relevant consideration when applying Bloom's Taxonomy in educational settings.

The third line encompasses studies examining how gender differences manifest in social cognition, particularly attributional complexity. While most educational assessments tend to focus on specific cognitive skills, research in social psychology suggests that gender differences in thinking may also stem from broader socio-cognitive patterns. For instance, Foels and Reid (2010) found that women tend to show higher levels of attributional complexity than men. This concept refers to the tendency to consider multiple factors and perspectives when interpreting social behavior. People who are high in attributional complexity are more likely to think in nuanced and flexible ways—they avoid oversimplifying others' actions or relying on stereotypes. Instead of assuming someone behaves a certain way because that is just who they are," they take into account the person's background, the situation, and even timing. Although the focus of Foels and Reid's study was on social dominance orientation, their findings offer meaningful insights into gender-related patterns in complex thinking. These insights could be relevant to understanding how men and women approach cognitively demanding tasks, such as those found at the higher levels of Bloom's Taxonomy.

Despite the mentioned studies, the relationship between gender and cognitive complexity in educational assessments remains unclear. While research in other disciplines hints at potential variations, these findings have not been consistently replicated in high-stakes language testing. This underscores the need for further investigations tailored specifically to language assessments to gain a clearer picture.

2.5 Cognitive diagnostic models

CDMs are designed to provide detailed insights into learners' mastery of specific skills or attributes, offering a more fine-grained perspective than traditional psychometric models such as item response theory or classical true score theory (Rupp and Templin, 2008). One of the key distinctions among CDMs lies in how they handle the interaction of the skills required to answer an item correctly. Based on these assumptions, CDMs are typically divided into specific and general models (Ravand et al., 2024).

Specific CDMs assume a single type of relationship among attributes for all items in a test. These relationships generally fall into three categories: conjunctive, disjunctive, and additive. In conjunctive models, the idea is that all of the required attributes must be present for a correct response. A well-known example is the *Deterministic Inputs, Noisy "And" gate* model (DINA; Junker and Sijtsma, 2001). This model operates on an all-or-nothing principle: if even one essential attribute is missing, the chance of answering the item correctly is low, regardless of how many other attributes are mastered. On the other hand, disjunctive models assume that mastering just one of the required attributes may be enough. The *Deterministic Inputs, Noisy "Or" gate* model (DINO; Templin and Henson, 2006) falls into this category. In this case, having any one of the relevant attributes can boost the probability of success, even if the others are not mastered. Additive models take a more gradual approach. In these models, each mastered attribute contributes independently to the probability of answering correctly. That is, the more attributes a learner has mastered, the better their chances—without any single skill being strictly necessary. Examples of additive models include the *Additive Cognitive Diagnosis Model* (*A*-CDM; de la Torre, 2011), the *Linear Logistic Test Model* (LLTM; Maris, 1999), and the *Reduced Reparameterized Unified Model* (RRUM; Hartz, 2002).

While specific CDMs impose the same interaction rule across all items, general CDMs allow for more flexibility. The *Generalized DINA* model (G-DINA; de la Torre, 2011), for instance, does not assume that all items behave the same way. Instead, it allows each item to follow its own pattern of attribute interaction—whether conjunctive, disjunctive, or additive. This flexibility makes general CDMs especially useful when test items are expected to involve different types of cognitive processing.

3 The present study

This study addresses several notable gaps in the existing literature on cognitive assessment in high-stakes language testing. While CDMs have been increasingly applied in language assessments to provide detailed insights into learners' cognitive processes, there remains a paucity of research integrating CDMs with Bloom's Taxonomy to analyze the cognitive demands of MCQs in such contexts. Moreover, prior studies examining the alignment of test items with Bloom's Taxonomy have predominantly employed qualitative and descriptive methodologies, lacking the quantitative rigor that CDMs can offer.

This study employs the integration of CDMs with Bloom's Taxonomy to quantitatively assess the cognitive complexity of MCQs in the INPEE, specifically within the English language section. By doing so, it seeks to provide a more detailed understanding of the cognitive skills assessed, moving beyond traditional item analysis methods. Furthermore, the research explores the relationship between cognitive complexity and item difficulty, as well as potential gender differences in performance across different cognitive levels. The research questions guiding this study are:

> *RQ1:* To what extent can MCQs in INPEE effectively assess higher-order cognitive skills as defined by Bloom's Taxonomy?
>
> *RQ2:* How does the cognitive complexity of MCQs influence item difficulty?
>
> *RQ3:* Is there any significant difference between male and female test takers' performance on items tapping into different levels of Bloom's Taxonomy?

4 Method

4.1 Data and participants

The data for this study were obtained from the Iranian Measurement Organization, a testing body responsible for administering the Iranian PhD entrance examination in March 2015. It included item responses from 1,000 Iranian Master of Arts (MA) holders in English majors. The item responses for the General English (GE) section, comprising 30 MCQs, were used for analysis.

Participants consisted of 653 females and 347 males, representing 65.2% and 34.7% of the sample, respectively. Their ages ranged from 23 to 62 years, with a mean age of 33.10 years. The GE section formed part of the INPEE comprehensive assessment, which also included sections on content knowledge and educational aptitude, totaling 150 MCQs.

In the present study, six expert judges—three researchers specializing in Bloom's Taxonomy and three English Language Teaching (ELT) professors with over five years of experience in language instruction and assessment—coded the items according to the level of the taxonomy each measured. The panel comprised three male and three female experts with substantial academic and practical backgrounds in education and assessment, including professors of neurology, science education, biology education, and ELT.

4.2 Q-matrix construction

At the core of any CDM lies the Q-matrix (Tatsuoka, 1983), which specifies the relationship between test items and the cognitive attributes they are intended to measure. Accurate specification of the Q-matrix is crucial, as it directly affects the precision of test takers' classifications (Ravand, 2016). The Q-matrix is structured as a binary table, with rows representing items and columns representing attributes. A value of 1 at the intersection of an item and an attribute indicates that the item measures that attribute; a value of 0 indicates that it does not.

While aggregating all expert Q-matrices into a consensus matrix was considered, we found that doing so would have led to an overly inclusive Q-matrix in which nearly all items were linked to the first three cognitive attributes—Remember, Understand, and Analyze. This would have resulted in substantial attribute overlap, reducing the model's diagnostic precision by limiting its ability to distinguish between latent classes (DeCarlo, 2011; Chiu et al., 2009).

Therefore, to construct and validate the Q-matrix in the present study, a mixed-methods approach was adopted. Each expert judge was provided with a copy of the test and a table listing items in the rows and Bloom's Taxonomy levels in the columns. They were asked to examine each item, identify the highest level of the taxonomy it targeted, and mark a 1 at the corresponding intersection. This process resulted in six initial Q-matrices, one from each expert (see Table 1 for a sample Q-matrix).

The initial Q-matrices were then subjected to empirical validation using the procedure proposed by Chiu et al. (2009), de la Torre and Chiu (2016) as implemented in the GDINA package in R (R Core Team, 2024). This procedure suggested modifications to the Q-matrices, either by removing an attribute from an item (changing a 1 to a 0) or by adding an attribute (changing a 0 to a 1). To ensure validity, we required convergence from at least two sources—expert judgment and empirical validation—for each Q-matrix entry. Although formal inter-rater reliability statistics were not computed, this structured approach ensured that the final Q-matrix reflected expert consensus while preserving model identifiability and interpretability (Madison and Bradshaw, 2015).

The modified Q-matrices were reanalyzed, and model fit indices—AIC, BIC, and log likelihood—were compared. The Q-matrix with the lowest values for these indices was selected for use in subsequent analyses.

TABLE 1 The final Q-matrix.

Items	Remember	Understand	Analyze
1	0	1	0
2	0	1	0
3	1	0	0
4	1	0	0
5	1	0	0
6	1	0	0
7	1	0	0
8	1	0	0
9	0	1	0
10	0	1	0
11	0	1	0
12	0	1	0
13	0	1	0
14	0	1	0
15	0	1	0
16	0	1	0
17	0	1	0
18	0	1	0
19	0	1	0
20	0	1	0
21	0	0	1
22	0	0	1
23	1	0	0
24	0	1	0
25	1	0	0
26	0	0	1
27	0	0	1
28	0	0	1
29	0	0	1
30	0	0	1

4.3 Data analysis

Data were analyzed using the GDINA package (Ma and de la Torre, 2020) in R. To assess the appropriateness of the Q-matrices, the analysis was conducted six times, each time using one of the six Q-matrices developed by the expert judges. Model fit was evaluated using two relative fit indices: Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). In addition, for the model based on the best-fitting Q-matrix, four absolute fit indices—MX2, MADcor, and SRMSR—were examined. According to Ravand (2016), non-significant MX2 values and MADcor values below 0.05 suggest a well-fitting model. Similarly, Maydeu-Olivares (2013) indicates that SRMSR values below 0.05 reflect a negligible degree of model misfit.

Furthermore, the average mastery level of participants on each attribute was calculated, and the prevalence of attribute profiles was reported. To address the second research question, a multigroup G-DINA analysis was conducted to compare attribute mastery and profile prevalence between male and female participants. Additionally, three independent-samples *t*-tests were performed to compare male and female performance on the three attributes identified through the Q-matrix specification and validation process.

5 Results

5.1 Q-matrix validation

It should be noted that the initial list of the levels of cognition according to Bloom's Taxonomy included six levels. However, during their evaluation, the expert judges concurred that the Apply, Evaluate, and Create levels had not been addressed in the exam items. Consequently, these three levels were excluded from the initial Q-matrices before the analysis.

Following each analysis, the software generated a proposed Q-matrix suggesting modifications to the initial one. To accept these software-generated changes, they needed concurrence from at least two of the expert Q-matrices. If this agreement existed and the change was accepted, the G-DINA analysis was repeated with the modified Q-matrix. It should be noted that when there was substantive reason to incorporate the modifications suggested by the program, each modification was accommodated at a time and the log-likelihood of the new model was compared against that of the base model, that is the model with no modification. After an analysis, the log-likelihood of the new Q-matrix, which differed in a single attribute from the original one, was subjected to a likelihood ratio test (LRT) to evaluate its impact on the overall model fit. As these changes pertained to a single item at a time, the degrees of freedom were limited to one. Therefore, if the result of the LRT exceeded 3.85, it indicated that the change would deteriorate the model fit and was, therefore, rejected. Conversely, if the result was less than 3.85, the modification was accepted. This iterative process continued until the Q-matrix was thoroughly refined. In this manner, each initial Q-matrix developed by each expert judge was analyzed by the software, modified if appropriate, evaluated via the LRT.

Then, the six final Q-matrixes were compared using AICs and BICs to determine the best model for the final Q-matrix of the study. In the interest of space, we do not include the details of Q-matrix validation here.

5.2 Model fit and final Q-matrix

Table 2 displays the AICs and BICs for the models with the six modified Q-matrices in the descending order of their AIC

TABLE 2 Relative fit indices of the models with the final Q-matrices.

Experts	AIC	BIC
Expert 1	27867.94	28854.4
Expert 2	27505.97	28482.62
Expert 3	27388.44	28384.72
Expert 4	27073.09	29207.97
Expert 5	27011.37	28498.42
Expert 6	27170.48	28421.95

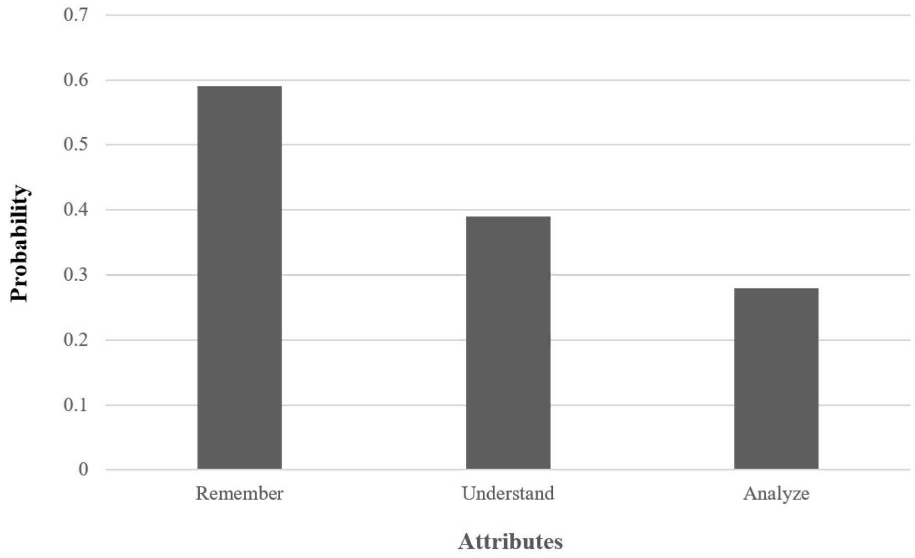

FIGURE 1
Percentage of levels mastered by test takers.

values. The model with the smallest AIC and BIC is considered the best-fitting model; however, no single model met both AIC and BIC criteria. While both AIC and BIC penalize model complexity, BIC's penalty also considers the sample size, resulting in a larger penalty. Therefore, the Q-matrix developed by Expert 3, with a BIC of 28384.72, was selected as the final Q-matrix for further analyses. In addition, the absolute fit indices indicated that the model fits the data (maxX2 = 12, $p = 0.15$, MADcor = 0.035, SRMSR = 0.048).

As Table 1 shows, eight items measured Remember (27%), 15 items measure Understand (50%), and 7 items tap into Analyze (23%). Thus, the answer to the first research question is that items of INPEE mostly measure lower levels of Bloom's Taxonomy, with no items measuring the two highest levels: Evaluating and Creating.

To address the second research question, the average mastery probabilities of the attributes were calculated. As shown in Figure 1, Remember was the easiest attribute, with 56% of the test takers mastering it, whereas Analyze was the most difficult, with only 28% of the test takers mastering it.

Table 3 shows the attribute profiles, or latent classes, to which the test takers belong. For three attributes, CDMs estimate $2^3 = 8$ profiles. The first profile [000] represents the percentage of test takers who have mastered none of the attributes, while the last profile [111] represents those who have mastered all the attributes. The third column includes the number of test takers belonging to each attribute profile. As the table shows, the most populated attribute profiles were [000], with about 38% of the test takers; [111], with about 22%; [011], those who mastered Attributes 2 and 3 but not 1, with about 14%; and [100], those who have mastered only Attribute 1, with about 12%. Very low probabilities of attribute profiles such as [010] indicated that the mastery of the second attribute without mastery of the first attribute is almost improbable. Similarly, profiles [001] and [101] suggested that the mastery of the third attribute without mastering the first two attributes has a low probability. This supports the hierarchical relationships among the cognitive levels of Bloom's Taxonomy.

TABLE 3 Attribute profiles of the entire group.

Profiles	Class.prob	Class.expfreq
000	0.384	383.59
100	0.122	122.15
010	0.034	34.24
110	0.072	72.32
001	0.020	19.73
101	0.007	6.60
011	0.137	137.01
111	0.224	224.35

To address the third research question, a multigroup G-DINA model was run to estimate the mastery probabilities of the attributes and attribute profiles separately for males and females. Two models were compared: one assuming item parameter invariance between males and females, and the other assuming non-invariance. Table 4 presents the relative and absolute fit indices for both models. While the AIC of the invariant model was lower, indicating a better fit, the BIC of the non-invariant model was lower. BIC imposes a heavier penalty for model complexity, thus better balancing fit and parsimony (Schwarz, 1978). Additionally, the non-invariant model had a non-significant max X2 and smaller MADcor, and SRMSR, values, suggesting a better overall fit. Therefore, the subsequent analyses were based on the non-invariant model.

As shown in Figure 2, the order of attribute difficulty across groups mirrors that of the entire sample, with attributes becoming increasingly difficult at higher Bloom's Taxonomy levels. Notably, at the lowest level, the percentage of test takers mastering Remember was almost the same for males and females. However, as we progress through higher levels of cognition, male test takers increasingly outperformed females.

TABLE 4 Relative and absolute fit indices of the multigroup models.

Models	Npars	Nobs	AIC	BIC	maxX2	p_maxX2	MADcor	SRMSR
Model 1	210	1,000	27,258	28,289	36.4003	0.000	0.035	0.048
Model 2	398	1,000	27,369	28,122	14.4086	0.110	0.020	0.031

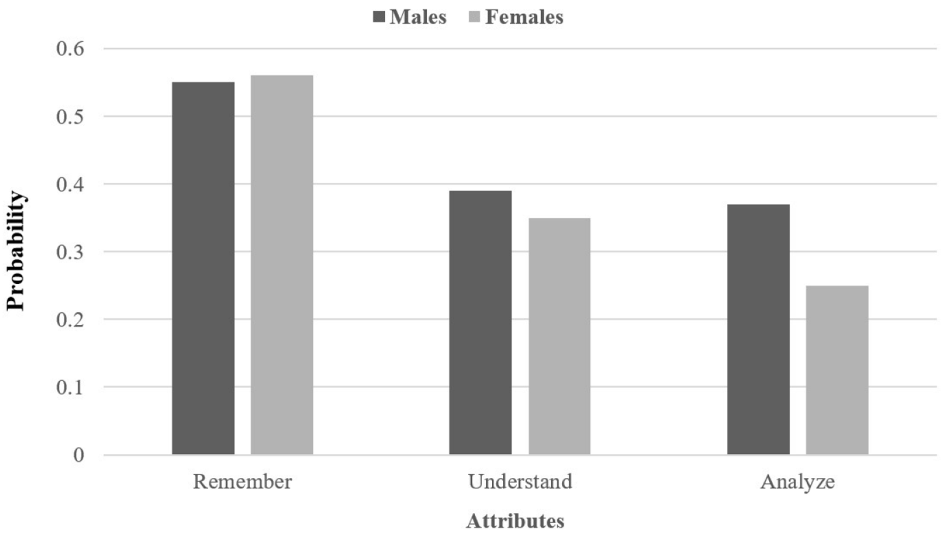

FIGURE 2
Attribute profiles of the entire group.

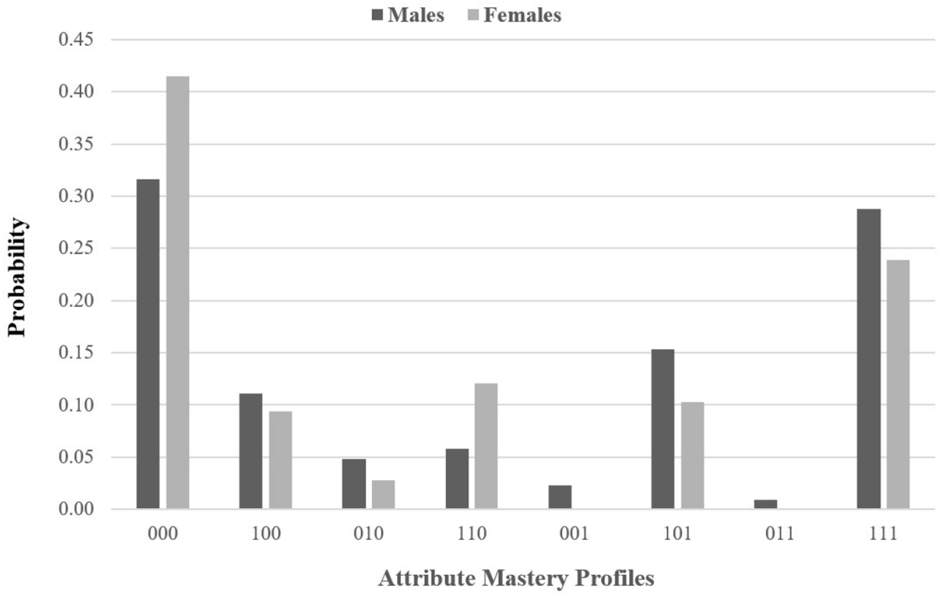

FIGURE 3
Mastery probabilities by group.

TABLE 5 Independent samples *t*-test of gender differences across cognitive levels.

Variables	*t* statistic	Degrees of freedom	*p*-value	Mean (Females)	Mean (Males)	Mean difference	Std. error difference	Eta squared
Remember	−1.114	709.217	0.265	1.891	2.023	0.132	0.118	0.005
Understand	−3.371	667.714	0.001	3.538	4.063	0.526	0.153	0.050
Analyze	−10.865	569.009	0.000	2.115	4.179	2.064	0.190	0.644

TABLE 6 Regression coefficients: levels of cognition predicting item difficulty.

Coefficients[a]						
Model		Unstandardized coefficients		Standardized coefficients	*t*	Sig.
		B	Std. error	Beta		
1	(Constant)	0.773	0.243		3.186	0.004
	remember	−2.090	0.412	−0.739	−5.078	0.000
	understand	0.929	0.430	0.372	2.159	0.040
	analyze	0.929	0.430	0.314	2.159	0.040

[a]Dependent variable: item difficulty.

According to Figure 3, attribute profiles of the two groups showed that the most populated profiles are [000] and [111]. About 32% of males were non-masters of all attributes, compared to about 41% of females. Conversely, for profile [111], about 29% of males mastered all attributes, compared to about 24% of females. There is a hierarchy of subskills for both males and females, as indicated by the low probability of profiles such as [010], [001], and [011]. The only exception is the profile [101], which represents test takers who have mastered the third attribute without mastering the second attribute.

Furthermore, the performance of male and female test takers was compared across items measuring each attribute. As Table 5 shows, consistent with the results in Figure 3, the gap between male and female performance widened as cognitive levels increased. At the lowest level, Remember, the difference was not statistically significant. For Understand items, the performance difference was statistically significant, but the eta squared indicated a small effect size (η2 = 0.01 for small, 0.06 for medium, and 0.14 for large, per Cohen's guidelines). However, for Analyze items, the difference was both statistically significant and, according to eta squared, the effect size was large.

Finally, to examine the relationship between item difficulty and cognitive complexity, a multiple linear regression analysis was conducted, regressing item difficulty on the three levels of cognitive complexity specified in the Q-matrix: Remember, Understand, and Analyze. The model accounted for approximately 45% of the variance in item difficulty ($R^2 = 0.45$). As shown in Table 6, the regression coefficient for Remember was negative and statistically significant ($\beta = -0.739$, $p < 0.001$), indicating that items measuring lower-order cognitive skills (i.e., Remember) tended to be easier. In contrast, both Understand ($\beta = 0.372$, $p = 0.040$) and Analyze ($\beta = 0.314$, $p = 0.040$) had positive and statistically significant regression coefficients, suggesting that items targeting higher-order cognitive processes were generally more difficult. These findings support the validity of the Q-matrix classifications in terms of their impact on item difficulty.

6 Discussion

One of the primary objectives of higher education is to cultivate thoughtful citizens equipped with problem-solving skills. Additionally, the nature and cognitive complexity of tests significantly impact both what is assessed and what is taught and learned (Martinez, 1999). This study aimed to investigate the cognitive levels addressed by the ubiquitous MCQs in large-scale tests, specifically focusing on the INPEE. Furthermore, the study compared the performances of male and female test takers on items targeting different levels of Bloom's Taxonomy.

The results of the study showed that 27% of the items measured Remember, 50% measured Understand, and 23% of the items tapped into Analyze. While this finding, in line with previous studies (Ali and Ruit, 2015; Choudhury and Freemont, 2017; Billings et al., 2016; Jensen et al., 2014; Kibble, 2017), suggests that MCQs can be effective in assessing various cognitive levels, including some complex and higher-order levels, they mostly assess lower-order cognitive levels (about 77%). This finding agrees with some previous studies (Kim et al., 2012; Knecht, 2001; Tarrant and Ware, 2008; Tarrant et al., 2009). A smaller proportion of items assessed Analyze, a higher-order cognitive skill. This finding aligns with Liu et al. (2024) who suggested that MCQs can measure Analyze. However, this level is not the highest level, and some other studies (Douglas et al., 2012; Cecilio-Fernandes et al., 2018; Crowe et al., 2008; Jensen et al., 2014; Karpen and Welch, 2016; Kim et al., 2012; Kıyak et al., 2022; Thompson et al., 2016; Thompson and O'Loughlin, 2015) suggest that MCQs can be designed to measure more complex cognitive processes. Nevertheless, the overall pattern of results in this high-stakes entrance exam indicates that MCQs in this context primarily assess LOT.

The findings showed that levels of cognitive complexity explained 45% of the variance in item difficulty. This finding aligns with previous research linking cognitive processing models to the level of item difficulty in MCQ tests (Embretson and Wetzel, 1987; Gorin and Embretson, 2006; Kirsch et al., 2017). The negative relationship between "Remember" and item difficulty supports the notion that items requiring simple recall are generally easier. This is consistent with the intuitive understanding of cognitive load and item difficulty. Conversely, the positive relationship between Understand, Analyze, and item difficulty suggests that items demanding Comprehension are more challenging. This accords with some studies that reported higher levels of cognitive processing require more difficult items (Rush et al., 2016; Testa et al., 2018). As Rush et al. (2016) documented, items requiring higher order level of thinking (Bloom's Level IV) consistently showed lower correct response rates and higher discrimination indices than Recall-based questions (Levels I–II). This occurs because higher order items—such as interpreting complex texts, inferring implicit relationships, or evaluating rhetorical strategies—demand integration of multiple information sources, increasing cognitive load.

However, some studies reported no relationship between cognitive complexity level and item difficulty (e.g., Kibble and Johnson, 2011; Tan and Othman, 2013; Tractenberg et al., 2013). This discrepancy might be attributed to several factors, including differences in subject matter. Another finding of the study was that the percentage of test takers mastering the Remember level was almost identical for males and females. However, regarding Understand, while the performance difference was statistically significant, the effect size was small. In addition, for Analyze items, the difference was both statistically significant and the effect size was large. These findings indicate that as cognitive demands increased, male test takers exhibited significantly higher performance.

In addition, the findings showed that Analyze is not the strongest predictor of difficulty, especially since it is a higher-order skill. This finding is somewhat unexpected. This can be explained by the nature of the "Analyze" items. As CDM results show, the mastery of Analyze is highly dependent on "Understand" [profile (011) = 13.7% vs. (001) = 2.0%]. The Analyze items often implicitly measure comprehension (Understand). If test takers lack foundational comprehension, they fail Analyze items regardless of analytical ability. This dependency artificially inflates the predictive role of Understand in the regression model while suppressing the unique contribution of Analyze, even though items classified as Analyze may still be cognitively demanding.

The observed gender disparity in higher-order cognitive skills, particularly at the Analyze level where males demonstrated significantly stronger performance ($\eta^2 = 0.644$), aligns with several studies reporting male advantages in complex thinking tasks. Our CDM results revealing higher male representation in the full-mastery profile [111] (29% vs. 24%) and lower representation in the non-mastery profile [000] (32% vs. 41%) further substantiate this pattern. These findings resonate with Wright et al. (2016), who documented widening male advantages at higher Bloom's levels in biology education, and Amin et al. (2024), who reported greater male proficiency in analytical and evaluative skills among prospective teachers. The hierarchical attribute structure observed in both genders (with low probabilities for profiles like [010] and [011]) parallels Lager et al.'s (2024) finding of strong intercorrelations between cognitive abilities, though their work additionally identified specific male advantages in abstract problem-solving and spatial abilities.

However, these results contrast with research reporting female cognitive advantages. Our null finding at the Remember level aligns with Araiku et al. (2019), who reported no overall gender differences in mathematics despite identifying male advantages at the Remember level. The significant female disadvantage in our Analyze items diverges markedly from Aldila et al. (2013) finding of consistent female superiority across all Bloom's levels in science education. Similarly, while Bastick (2002) found female strengths specifically in analysis/synthesis within matching formats, our humanities-focused assessment revealed opposite patterns for analytical skills. These contradictions may stem from cultural, socioeconomic, and disciplinary factors influencing the observed outcomes.

7 Conclusion

This study investigated the cognitive demands of MCQs in the INPEE and explored gender differences in performance across different cognitive levels. The results indicate that MCQs predominantly assess lower-order cognitive skills. Moreover, the study found a significant relationship between cognitive complexity and item difficulty. A notable gender gap emerged, with males outperforming females at higher cognitive levels.

This study has implications for researchers, test developers, instructors, and assessment practitioners. This study provides a methodology for empirically and objectively assessing of cognitive complexity of MCQs. This provide a quantitative and objective analysis by using CDM that can be adopted for other tests and fields. In addition, this study's findings underscore the necessity of revising large-scale assessments to better reflect HOT skills and recommends integrating intelligent assessment tools to achieve these goals. The analysis of cognitive levels in assessments is crucial for creating well-balanced evaluations that accurately measure a broad spectrum of cognitive skills. To promote HOT, test developers should carefully consider the cognitive demands of items and strive for a balanced representation of cognitive levels. For doing so, they should implement structured protocols to engineer balanced cognitive coverage. First, they should adopt a cognitive blueprint defining explicit target distributions aligned with the test objectives. Second, they should revise item development through specialized writer training focused on crafting higher-order MCQs—using scenario-based stems, misconception-driven distractors, and multi-step reasoning tasks—coupled with dual independent Bloom's classification during item review. Third, they should embed validation mechanisms to flag cognitive imbalances in draft tests and apply post-hoc CDMs to audit alignment between intended and measured attributes, mastery gaps, and complexity-difficulty relationships.

Moreover, measurement of cognitive complexity benefits both teaching and learning. Analyzing cognitive levels in assessments plays a vital role in measuring educational success. These analyses can help identify areas of both strength and weakness, informing assessment practices to promote fairer evaluations that consider students' varying cognitive skill sets. Additionally, analyzing cognitive levels can lead to more positive outcomes in problem-based learning environments by encouraging educators to design learning activities that promote understanding of underlying principles. By measuring cognitive complexity in assessments, educators gain valuable insights that can improve teaching, learning, and assessment practices, ultimately leading to a more well-rounded educational experience for students.

While our study highlights the need to examine different item formats and cognitive complexity across subject domains, future research in these areas may face challenges related to construct comparability and the classification of cognitive skills across disciplines. For instance, applying Bloom's Taxonomy to fields such as mathematics, sciences, or the humanities may yield different interpretations of higher-order skills, complicating cross-disciplinary comparisons. Moreover, item formats such as essays or open-ended questions—while better suited to assess HOT—require subjective scoring and are prone to inter-rater variability, calling for rigorous rater training and rubric validation.

Another important limitation is that the current study focused only on three levels of Bloom's Taxonomy (Remember, Understand, Analyze), as the other levels (Apply, Evaluate, and Create) were not represented in the test. This restriction narrows the cognitive scope of the analysis and limits our ability to draw conclusions about how the test assesses HOT skills more broadly.

In addition, the study infers a hierarchical relationship among cognitive attributes—most notably, that successful performance on items tapping Analyze presupposes mastery of Understand. While this assumption aligns with theoretical expectations, it was not independently tested through hierarchical CDMs or longitudinal modeling. Future research should consider empirically evaluating such hierarchical structures to avoid oversimplified interpretations of cognitive processing.

Another methodological consideration is the operationalization of distractor quality. Future research aiming to explore how distractor plausibility interacts with cognitive complexity and item difficulty must establish objective indicators for distractor functioning, which may involve integrating psychometric analyses (e.g., distractor discrimination) with cognitive models.

Additionally, collecting complementary qualitative data through interviews or think-aloud protocols with students and educators can provide deeper insights into the cognitive demands of test items. However, these methods are resource-intensive and may limit generalizability unless carefully sampled and triangulated with quantitative findings.

Data availability statement

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

Ethics statement

The studies involving humans were approved by Vali-e-Asr University of Rafsanjan, Rafsanjan, Iran (Ethical Approval No. IR. VAUR. REC.1399.123). 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. The participants provided their written informed consent to participate in this study.

Author contributions

HR: Resources, Conceptualization, Writing – review & editing, Methodology, Formal analysis, Writing – original draft. RS: Writing – original draft, Writing – review & editing. FE: Writing – review & editing, Writing – original draft. AM: Writing – review & editing, Data curation.

Funding

The author(s) declare that no financial support was received for the research 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.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Generative AI statement

The authors declare that Gen AI was used in the creation of this manuscript. Generative AI was used to edit the text for grammar and spelling.

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

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

Aldila, M., Sudargo Tapilouw, F., and Sanjaya, Y. (2013). Students' cognitive and attitude of secondary school in learning global warming using student team achievement division (STAD) based on gender. *J. Sci. Learn.* 1, 104–109. doi: 10.17509/jsl.v1i3.11793

Ali, S. H., and Ruit, K. G. (2015). The impact of item flaws, testing at low cognitive level, and low distractor functioning on multiple-choice question quality. *Perspect. Med. Educ.* 4, 244–251. doi: 10.1007/s40037-015-0212-x

Alshurafat, H., Alaqrabawi, M., and Al Shbail, M. O. (2024). Developing learning objectives for forensic accounting using bloom's taxonomy. *Account. Educ.* 33, 497–513. doi: 10.1080/09639284.2023.2222271

Amin, M., Naqvi, S. U. E. L., Amin, H., Kayfi, S. Z., and Amjad, F. (2024). Bloom's taxonomy and prospective teachers' preparation in Pakistan. *Qlantic J. Soci. Sci.* 5, 391–403. doi: 10.55737/qjss.791335465

Anderson, L. W., and Krathwohl, D. A. (2001). A taxonomy for learning, teaching, and assessing: A revision of bloom's taxonomy of educational objectives. New York: Longman.

Araiku, J., Sidabutar, R., and Mairing, J. P. (2019). Gender differences in mathematics ability of junior high school students based on bloom's taxonomy. *Jurnal Gantang* 4, 15–25. doi: 10.31629/jg.v4i1.969

Aydin, U., and Birgili, B. (2023). Assessing mathematical higher-order thinking skills: an analysis of Turkish university entrance examinations. *Educ. Assess.* 28, 190–209. doi: 10.1080/10627197.2023.2202311

Baghaei, S., Bagheri, M. S., and Yamini, M. (2020). Analysis of IELTS and TOEFL reading and listening tests in terms of revised bloom's taxonomy. *Cogent Educ.* 7:1720939. doi: 10.1080/2331186X.2020.1720939

Baghaei, S., Bagheri, M. S., and Yamini, M. (2021). Learning objectives of IELTS listening and reading tests: focusing on revised bloom's taxonomy. *Res. Eng. Lang. Pedag.* 9, 182–199. doi: 10.30486/relp.2021.1916940.1244

Bastick, T. (2002). Gender Differences for 6-12th Grade Students Over Bloom's Cognitive Domain. Paper presented at the Western Psychological Association, WPA 2002 Convention, Irvine, CA. USA.

Bax, S. (2013). The cognitive processing of candidates during reading tests: evidence from eye-tracking. *Lang. Test.* 30, 441–465. doi: 10.1177/0265532212473244

Bezuidenhout, M. J., and Alt, H. (2011). 'Assessment drives learning': do assessments promote high-level cognitive processing? *S. Afr. J. High. Educ.* 25, 1062–1076. doi: 10.10520/EJC37738

Billings, M. S., DeRuchie, K., Haist, S. A., Hussie, K., Merrell, J., Paniagua, M. A., et al. (2016). Constructing written test questions for the basic and clinical sciences. *4th* Edn. Philadelphia (PA): National Board of Medical Examiners.

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., and Krathwohl, D. R. (1956). *A Taxonomy of Educational Objectives: Handbook I The Cognitive Domain*. Longman, Green Co., New York.

Borda, M., Reyes-Ortiz, C., Pérez-Zepeda, M., Patiño-Hernández, D., Gómez-Arteaga, C., and Cano-Gutierrez, C. (2018). Educational level and its association with the domains of the Montreal cognitive assessment test. *Aging Ment. Health* 23, 1300–1306. doi: 10.1080/13607863.2018.1488940

Butler, A. C. (2018). Multiple-choice testing in education: are the best practices for assessment also good for learning? *J. Appl. Res. Mem. Cogn.* 7, 323–331. doi: 10.1016/j.jarmac.2018.07.002

Cecilio-Fernandes, D., Kerdijk, W., Bremers, A. J., Aalders, W., and Tio, R. A. (2018). Comparison of the level of cognitive processing between casebased items and non-case-based items on the interuniversity Progress test of medicine in the Netherlands. *J Educ Eval Health Prof.* 15:28. doi: 10.3352/jeehp.2018.15.28

Chang, W., and Chung, M. (2009). Automatic applying Bloom's Taxonomy to classify and analyze the cognition level of English question items. Paper presented at Joint Conferences on Pervasive Computing (JCPC), Taipei, Taiwan.

Chiu, C.-Y., Douglas, J., and Li, X. (2009). Cluster analysis for cognitive diagnosis: theory and applications. *Psychometrika* 74, 633–665. doi: 10.1007/s11336-009-9125-0

Choudhury, B., and Freemont, A. (2017). Assessment of anatomical knowledge: approaches taken by higher education institutions. *Clin. Anat.* 30, 290–299. doi: 10.1002/ca.22835

Crowe, A., Dirks, C., and Wenderoth, M. P. (2008). Biology in bloom: implementing bloom's taxonomy to enhance student learning in biology. *CBE Life Sci. Educ.* 7, 368–381. doi: 10.1187/cbe.08-05-0024

Das, S., Das Mandal, S. K., and Basu, A. (2022). Classification of action verbs of bloom's taxonomy cognitive domain: an empirical study. *J. Educ.* 202, 554–566. doi: 10.1177/00220574211002199

Das, S., Das Mandal, S. K., and Basu, A. (2020). Identification of cognitive learning complexity of assessment questions using multi-class text classification. *Contemp. Educ. Technol.* 12:ep275. doi: 10.30935/cedtech/8341

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika* 76, 179–199. doi: 10.1007/s11336-011-9207-7

de la Torre, J., and Chiu, C.-Y. (2016). A general method of empirical Q-matrix validation. *Psychometrika*, 81, 253–273. doi: 10.1007/s11336-015-9467-8

DeCarlo, L. T. (2011). On the analysis of fraction subtraction data: the DINA model,classification, latent class sizes, and the Q-matrix. *Appl. Psychol. Meas.* 35, 8–26. doi: 10.1177/0146621610377081

DeWitt, S. W., Patterson, N., Blankenship, W., Blevins, B., DiCamillo, L., Gerwin, D., et al. (2013). The lower-order expectations of high-stakes tests: a four-state analysis of social studies standards and test alignment. *Theory Res. Soc. Educ.* 41, 382–427. doi: 10.1080/00933104.2013.787031

DiBello, L. V., and Stout, W. (2007). Guest editors' introduction and overview: IRT-based cognitive diagnostic models and related methods. *J. Educ. Meas.* 44, 285–291. doi: 10.1111/j.1745-3984.2007.00039.x

Douglas, M., Wilson, J., and Ennis, S. (2012). Multiple-choice question tests: A convenient, flexible andeffective learning tool? A case study. *Innov. Educ. Teach. Int.* 49, 111–121. doi: 10.1080/14703297.2012.677596

Ebadi, S., and Shahbazian, F. (2015). Exploring the cognitive level of final exams in Iranian high schools: focusing on bloom's taxonomy. *J. Appl. Linguist. Lang. Res.* 2, 1–11.

Ehrich, J. F., Howard, S. J., Bokosmaty, S., and Woodcock, S. (2021). An item response modeling approach to cognitive load measurement. *Front. Educ.* 6:648324. doi: 10.3389/feduc.2021.648324

Elim, E. H. S. Y. (2024). Promoting cognitive skills in AI-supported learning environments: the integration of bloom's taxonomy. *Education* 3-13, 1–11. doi: 10.1080/03004279.2024.2332469

Embretson, S. E., and Wetzel, C. D. (1987). Component latent trait models for paragraph comprehension tests. *Appl. Psychol. Meas.* 11, 175–193. doi: 10.1177/014662168701100207

Ernawati, E., and Baharullah, B. (2020). Analysis of higher order thinking skills(hots) in mathematical problem solving based on revised blooms'taxonomy viewed from gender equality. *MaPan* 8, 315–328. doi: 10.24252/mapan.2020v8n2a10

Febriyana, F., and Harjanto, I. (2023). Cognitive levels of questions by Indonesian teachers of English. *J. Eng. Lang. Teach. Ling.* 8:2023. doi: 10.21462/jeltl.v8i2.1032

Ferrara, S., Svetina, D., Skucha, S., and Davidson, A. H. (2011). Test development with performance standards and achievement growth in mind. *Educ. Meas.: Issues Pract.* 30, 3–15. doi: 10.1111/j.1745-3992.2011.00218.x

Foels, R., and Reid, L. D. (2010). Gender differences in social dominance orientation: the role of cognitive complexity. *Sex Roles* 62, 684–692. doi: 10.1007/s11199-010-9775-5

Freahat, N. M., and Smadi, O. M. (2014). Lower-order and higher-order reading questions in secondary and university level EFL textbooks in Jordan. *Theory Pract. Lang. Stud.* 4, 1804–1813. doi: 10.4304/tpls.4.9.1804-1813

French, S., Dickerson, A., and Mulder, R. A. (2023). A review of the benefits and drawbacks of high-stakes final examinations in higher education. *High. Educ.* 88, 893–918. doi: 10.1007/s10734-023-01148-z

Gorin, J. S. (2006). Test design with cognition in mind. *Educ. Meas. Issues Pract.* 25, 21–35. doi: 10.1111/j.1745-3992.2006.00076.x

Gorin, J. S. (2007). Reconsidering issues in validity theory. *Educ. Res.* 36, 456–462. doi: 10.3102/0013189X07311607

Gorin, J. S., and Embretson, S. E. (2006). Item diffficulty modeling of paragraph comprehension items. *Appl. Psychol. Meas.* 30, 394–411. doi: 10.1177/0146621606288554

Granello, D. H. (2001). Promoting cognitive complexity in graduate written work: using bloom's taxonomy as a pedagogical tool to improve literature reviews. *Counsel. Educ. Superv.* 40, 292–307. doi: 10.1002/j.1556-6978.2001.tb01261.x

Han, C., and Xiang, J. (2025). Alignment analysis between China college entrance examination physics test and curriculum standard based on E-SEC model. *Int. J. Sci. Math. Educ.* 23, 215–234. doi: 10.1007/s10763-024-10468-0

Harland, T., and Wald, N. (2021). The assessment arms race and the evolution of a university's assessment practices. *Assess. Eval. High. Educ.* 46, 105–117. doi: 10.1080/02602938.2020.1745753

Hartz, S. M. (2002). A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality [Doctoral dissertation, University of Illinois at Urbana-Champaign]

Ho, P. J. (2022). Assessing the range of cognitive processes in the Hong Kong diploma of secondary education examination (HKDSE)'s English language reading literacy test. *Lang. Test. Asia* 12:18. doi: 10.1186/s40468-022-00167-4

Hsu, F. Y., Lee, H. M., Chang, T. H., and Sung, Y. T. (2018). Automated estimation of item difficulty for multiple-choice tests: an application of word embedding techniques. *Inf. Process. Manag.* 54, 969–984. doi: 10.1016/j.ipm.2018.06.007

Jaramillo, S. G., and Cadavid, J. M. (2015). "Selection of collaborative learning techniques using Bloom's taxonomy" in International Workshop on Social Computing in Digital Education (Cham: Springer), 1–11.

Jensen, J. L., McDaniel, M. A., Woodard, S. M., and Kummer, T. A. (2014). Teaching to the test… or testing to teach: exams requiring higher-order thinking skills encourage greater conceptual understanding. *Educ. Psychol. Rev.* 26, 307–329. doi: 10.1007/s10648-013-9248-9

Jung Lim, H. (2014). Exploring the validity evidence of the TOEFL iBT reading test form A cognitive perspective. Michigan: Second Language Studies.

Junker, B. W., and Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Appl. Psychol. Meas.* 25, 258–272. doi: 10.1177/01466210122032064

Kahn, R. L. (2012). A taxonomy for choosing, evaluating, and integrating in-the-cloud resources in a university environment. *J. Educ. Technol. Syst.* 41, 171–181. doi: 10.2190/ET.41.2.e

Karpen, S. C., and Welch, A. C. (2016). Assessing the inter-rater reliability and accuracy of pharmacy faculty's bloom's taxonomy classifications. *Curr. Pharm. Teach. Learn.* 8, 885–888. doi: 10.1016/j.cptl.2016.08.003

Khorsand, N. (2009). Cognitive levels of questions used by Iranian EFL teachers in advanced reading comprehension tests (Unpublished M.A thesis). Shiraz University, Iran.

Khoy, B. (2025). Unlocking cognitive learning objectives: a comprehensive evaluation of how textbooks and syllabi align with revised bloom's taxonomy across disciplines. *Curr. Perspect.* 45:189202. doi: 10.1007/s41297-024-00295-2

Kibble, J. D. (2017). Best practices in summative assessment. *Adv. Physiol. Educ.* 41, 110–119. doi: 10.1152/advan.00116.2016

Kibble, J. D., and Johnson, T. (2011). Are faculty predictions or item taxonomies useful for estimating the outcome of multiple-choice examinations? *Adv. Physiol. Educ.* 35, 396–401. doi: 10.1152/advan.00062.2011

Kim, M. K., Patel, R. A., Uchizono, J. A., and Beck, L. (2012). Incorporation of bloom's taxonomy into multiplechoice examination questions for a pharmacotherapeutics course. *Am. J. Pharm. Educ.* 76, 1–8. doi: 10.5688/ajpe766114

Kirsch, I., Lennon, M., Yamamoto, K., and von Davier, M. (2017). Large-scale assessments of adult literacy. In *Advancing human assessment: Methodological, psychological, and policy contributions*. (eds.) R. Bennett and M. von Davier, (New York: Springer).

Kıyak, Y. S., Budakoğlu, I. İ., Bakan Kalaycıoğlu, D., Kula, S., and Coşkun, Ö. (2022). Can preclinical students improve their clinical reasoning skills only by taking case-based online testlets? A randomized controlled study. *Innov. Educ. Teach. Int.* 60, 1–10. doi: 10.1080/14703297.2022.2041458

Knecht, K. T. (2001). Assessing cognitive skills of pharmacy students in a biomedical sciences module using a classification of multiple-choice item categories according to bloom's taxonomy. *Am. J. Pharm. Educ.* 65:324.

Kosorus, H., and Küng, J. (2014). "Learning-oriented question recommendation using bloom's learning taxonomy and variable length hidden Markov models" in Transactions on large-scale data and knowledge-centered systems XVI. eds. H. Kosorus, J. Küng, H. Kosorus and J. Küng (Berlin, Heidelberg: Springer), 29–44.

Krathwohl, D. R. (2002). A revision of bloom's taxonomy: an overview. *Theory Pract.* 41, 212–218. doi: 10.1207/s15430421tip4104_2

Krathwohl, D. R., and Anderson, L. W. (2009). A taxonomy for learning, teaching, and assessing: A revision of bloom's taxonomy of educational objectives. New York: Longman.

Krause, U., Béneker, T., and van Tartwijk, J. (2021). Geography textbook tasks fostering thinking skills for the acquisition of powerful knowledge. *Int. Res. Geogr. Environ. Educ.* 31, 69–83. doi: 10.1080/10382046.2021.1885248

Krouska, A., Troussas, C., and Virvou, M.. (2018). Computerized adaptive assessment using accumulative learning activities based on revised bloom's taxonomy. Joint Conference on Knowledge-Based Software Engineering (pp. 252–258). Springer, Cham.

Kubiszyn, T., and Borich, G. D. (2024). Educational testing and measurement. Hoboken: John Wiley & Sons.

Lager, P., Alder, J., and Bostrom, L. (2024). Gender differences in operational and cognitive abilities: a study of aviation pilot candidates. *J. Cogn. Psychol.* 36, 123–134. doi: 10.3389/fpsyg.2024.1402645

Leighton, J., and Gierl, M. (Eds.) (2007). Cognitive diagnostic assessment for education: Theory and applications. New York: Cambridge University Press.

Liu, Q., Wald, N., Daskon, C., and Harland, T. (2024). Multiple-choice questions (MCQs) for higher-order cognition: perspectives of university teachers. *Innov. Educ. Teach. Int.* 61, 802–814. doi: 10.1080/14703297.2023.2222715

Ma, W., and de la Torre, J. (2020). GDINA: an R package for cognitive diagnosis modeling. *J. Stat. Softw.* 93, 1–26. doi: 10.18637/jss.v093.i14

Madison, M. J., and Bradshaw, L. (2015). The effects of Q-matrix design on classification accuracy in the log-linear cognitive diagnosis model. *Appl. Psychol. Meas.* 39, 87–103. doi: 10.1177/0013164414539162

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika* 64, 187–212. doi: 10.1007/BF02294568

Martinez, M. E. (1999). Cognition and the question of test item format. *Educ. Psychol.* 34, 207–218. doi: 10.1207/s15326985ep3404_2

Mason, E. J. (2007). Measurement issues in high stakes testing: validity and reliability. *J. Appl. Sch. Psychol.* 23, 27–46. doi: 10.1300/J370v23n02_03

Maydeu-Olivares, A. (2013). Goodness-of-fit assessment of item response theory models. *Measurement* 11, 71–101. doi: 10.1080/15366367.2013.831680

Migliaccio, B., Sheikh, O., Bateman, C., and Knopp, J. (2009). Gender differences in performance in principles of biochemistry based on bloom's taxonomy of question difficulty and study habits. *NCSU Undergrad Res J* 5, 76–83.

Mizbani, M., Salehi, H., Tabatabaei, O., and Talebinejad, M. (2023). Textbook evaluation based on bloom's revised taxonomy: Iranian senior high school textbook in focus. *Lang. Transl.* 13, 85–99. doi: 10.30495/ttlt.2023.698908

Momen, A., Ebrahimi, M., and Hassan, A. M. (2022). Importance and implications of theory of bloom's taxonomy in different fields of education. International conference on emerging technologies and intelligent systems (pp. 515–525). Springer International Publishing. Cham.

Monrad, S. U., Bibler Zaidi, N. L., Grob, K. L., Kurtz, J. B., Tai, A. W., Hortsch, M., et al. (2021). What faculty write versus what students see? Perspectives on multiple-choice questions using bloom's taxonomy. *Med. Teach.* 43, 575–582. doi: 10.1080/0142159X.2021.1879376

Moslehi, S., and Razmjoo, S. A. (2021). On the representation of bloom's revised taxonomy in TOEFL iBT and IELTS academic. *J. Eng. Lang. Teach. Learn.* 13, 173–200. doi: 10.22034/ELT.2021.46190.2391

Muhayimana, T., Kwizera, L., and Nyirahabimana, M. R. (2022). Using bloom's taxonomy to evaluate the cognitive levels of primary leaving English exam questions in Rwandan schools. *Curr. Perspect.* 42, 51–63. doi: 10.1007/s41297-021-00156-2

Mustafidah, H. S., and Pinandita, T. (2022). Natural language processing for mapping exam questions to the cognitive process dimension. *Int. J. Emerg. Technol. Learn.* 17, 4–16. doi: 10.3991/ijet.v17i13.29095

NamazianDoost, I., and HayaviMehr, M. (2017). A comparative study of critical thinking skills in high school and simulated IELTS reading comprehension questions. *Int. J. Eng. Lang. Teach.* 5, 35–69. doi: 10.37745/ijelt.13

Nasution, L., Rinjani, B. N. K. P., Hunaepi, H., and Samsuri, T. (2023). The analytical thinking ability of prospective science teachers: an overview of study programs and gender. *J. Penelitian Pendidikan IPA* 9, 1144–1150. doi: 10.29303/jppipa.v9iSpecialIssue.5256

Niyibizi, E., Sibomana, E., Niyomugabo, C., Yanzigiye, B., Jean de Dieu, A. N., and Perumal, J. (2018). Assessment in a Rwandan higher education institution: a quest for aligned assessment to promote socio-economic transformation. *Assess. Eval. High. Educ.* 43, 1166–1182. doi: 10.1080/02602938.2018.1436688

Nkhoma, M. Z., Lam, T. K., Sriratanaviriyakul, N., Richardson, J., Kam, B., and Lau, K. H. (2017). Unpacking the revised bloom's taxonomy: developing case-based learning activities. *Educ. Train.* 59, 250–264. doi: 10.1108/ET-03-2016-0061

Noroozi, S., and Karami, H. (2022). A scrutiny of the relationship between cognitive load and difficulty estimates of language test items. *Lang. Test. Asia* 12:13. doi: 10.1186/s40468-022-00163-8

Parsaei, I., Alemokhtar, M. J., and Rahimi, A. (2017). Learning objectives in ESP books based on bloom's revised taxonomy. *Beyond Words* 5, 14–22. doi: 10.33508/bw.v5i1.1112

Pellegrino, J. W., DiBello, L. V., and Goldman, S. R. (2016). A framework for conceptualizing and evaluating the validity of instructionally relevant assessments. *Educ. Psychol.* 51, 59–81. doi: 10.1080/00461520.2016.1145550

Rahayu, S., and Alsulami, N. M.. (2024). Assessing higher order thinking skills of the 21st century learners using socio-scientific issues as a context. AIP Conference Proceedings 070009. AIP Publishing LLC.

Ramirez, T. V. (2017). On pedagogy of personality assessment: application of bloom's taxonomy of educational objectives. *J. Pers. Assess.* 99, 146–152. doi: 10.1080/00223891.2016.1167059

Ravand, H. (2016). Application of a cognitive diagnostic model to a high-stakes reading comprehension test. *J. Psychoeduc. Assess.* 34, 782–799. doi: 10.1177/0734282915623053

Ravand, H., Effatpanah, F., Ma, W., de la Torre, J., Baghaei, P., and Kunina-Habenicht, O. (2024). Exploring interrelationships among L2 writing subskills: insights from cognitivediagnostic models. *Appl. Meas. Educ.* 37, 329–355. doi: 10.1080/08957347.2024.2424550

Razmjoo, S. A., and Kazempourfard, E. (2012). On the representation of bloom's revised taxonomy in interchange course books. *J. Teach. Lang. Skills* 4, 171–204.

Razmjoo, S. A., and Madani, H. (2013). A content analysis of the English section of university entrance exams based on bloom's revised taxonomy. *Int. J. Lang. Learn. Applied Linguistics World* 4, 105–129.

R Core Team. (2024). *R: A language and environment for statistical computing*. R foundation for statistical computing. Available at: https://www.R-project.org

Rena, I. P., Al-Baekani, A. K., and Kamil, A. B. (2023). An analysis of speaking activities in Indonesian ELT textbook based on cognitive domain of bloom's taxonomy revised. *IJET* 12, 39–47. doi: 10.15642/ijet2.2023.12.1.39-47

Rentmeester, C. (2018). Adding academic rigor to introductory ethics courses using bloom's taxonomy. *Int. J. Ethics Educ.* 3, 67–74. doi: 10.1007/s40889-017-0047-x

Rezaee, M., and Golshan, M. (2016). Investigating the cognitive levels of English final exams based on bloom's taxonomy. *Int. J. Educ. Investig.* 3, 57–68.

Rupp, A. A., and Templin, J. L. (2008). Unique characteristics of diagnostic classification models: a comprehensive review of the current state-of-the-art. *Measurement* 6, 219–262. doi: 10.1080/15366360802490866

Rush, B. R., Rankin, D. C., and White, B. J. (2016). The impact of item-writing flaws and item complexity on examination item difficulty and discrimination value. *BMC Med. Educ.* 16, 250–210. doi: 10.1186/s12909-016-0773-3

Ryan, K. E., and Chiu, S. (2001). An examination of item context effects, DIF, and gender DIF. *Appl. Meas. Educ.* 14, 73–90. doi: 10.1207/S15324818AME1401_06

Santen, S. A., Grob, K. L., Monrad, S. U., Stalburg, C. M., Smith, G., Hemphill, R. R., et al. (2019). Employing a root cause analysis process to improve examination quality. *Acad. Med.* 94, 71–75. doi: 10.1097/ACM.0000000000002439

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Stat.* 6:461464. Available at: http://www.jstor.org/stable/2958889

Scott, T. (2003). Bloom's taxonomy applied to testing in computer science classes. *J. Comput. Sci. Coll.* 19, 267–274. doi: 10.5555/948737.948775

Scouller, K. (1998). The influence of assessment method on students' learning approaches: multiple choice question examination versus assignment essay. *High. Educ.* 35, 453–472. doi: 10.1023/A:1003196224280

Shahbazian, F. (2016). Exploring the cognitive and knowledge dimension of National English Final Exam Items of Iranian high schools based on bloom's revised taxonomy (MA thesis). Kermanshah: Razi University.

Shavelson, R. J. (2010). On the measurement of competency. *Empir. Res. Vocat. Educ. Train.* 2, 41–63. doi: 10.1007/BF03546488

Stanger-Hall, K. F. (2012). Multiple-choice exams: an obstacle for higher-level thinking in introductory science classes. *CBE Life Sci. Educ.* 11, 294–306. doi: 10.1187/cbe.11-11-0100

Swart, A. J. (2009). Evaluation of final examination papers in engineering: a case study using bloom's taxonomy. *IEEE Trans. Educ.* 53, 257–264. doi: 10.1109/TE.2009.2014221

Tan, Y. T., and Othman, A. R. (2013). The relationship between complexity (taxonomy) and difficulty. *AIP Conf. Proc.* 15, 596–603. doi: 10.1063/1.4801179

Tangsakul, P., Kijpoonphol, W., Linh, D. N., and Kimura, N. L. (2017). Using bloom's revised taxonomy to analyze reading comprehension questions in team up in English 1–3 and grade 9 English O-net tests. *Int. J. Research Grantaalayah* 5, 31–41. doi: 10.29121/granthaalayah.v5.i7.2017.2106

Tarrant, M., and Ware, J. (2008). Impact of item-writing flaws in multiple-choice questions on student achievement in high-stakes nursing assessments. *Med. Educ.* 42, 198–206. doi: 10.1111/j.1365-2923.2007.02957.x

Tarrant, M., Ware, J., and Mohammed, A. M. (2009). An assessment of functioning and non-functioning distractors in multiple-choice questions: a descriptive analysis. *BMC Med. Educ.* 9, 1–8. doi: 10.1186/1472-6920-9-40

Tatsuoka, K. K. (1983). Rule-space: An approach for dealing with misconceptions based on item response theory. *J. Educ. Meas.* 20, 345–354. Available at: https://www.jstor.org/stable/1434951

Templin, J., and Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychol. Methods* 11, 287–305. doi: 10.1037/1082-989X.11.3.287

Ten Cate, O., Custers, E., and Durning, S. (Eds.) (2018). Principles and practice of case-based clinical reasoning education: A method for preclinical students. *15th* Edn. Cham: Springer International Publishing.

Testa, S., Toscano, A., and Rosato, R. (2018). Distractor efficiency in an item pool for a statistics classroom exam: assessing its relation with item cognitive level classified according to bloom's taxonomy. *Front. Psychol.* 9:1585. doi: 10.3389/fpsyg.2018.01585

Thompson, A. R., Kelso, R. S., Ward, P. J., Wines, K., and Hanna, J. B. (2016). Assessment driven learning: the use of higher-order and disciplineintegrated questions on gross anatomy practical examinations. *Medical Science Educator.* 26, 587–596. doi: 10.1007/s40670-016-0306-z

Thompson, A. R., and O'Loughlin, V. D. (2015). The blooming anatomy tool (BAT): A discipline-specific rubric for utilizing bloom's taxonomy in the design and evaluation of assessments in the anatomical sciences. *Anat. Sci. Educ.* 8, 493–501. doi: 10.1002/ase.1507

Tíjaro-Rojas, R., Arce-Trigatti, A., Cupp, J., Pascal, J., and Arce, P. E. (2016). A systematic and integrative sequence approach (SISA) for mastery learning: anchoring bloom's revised taxonomy to student learning. *Educ. Chem. Eng.* 17, 31–43. doi: 10.1016/j.ece.2016.06.001

Tractenberg, R. E., Gushta, M. M., Mulroney, S. E., and Weissinger, P. A. (2013). Multiple choice questions can be designed or revised to challenge learners' critical thinking. *Adv. Health Sci. Educ.* 18, 945–961. doi: 10.1007/s10459-012-9434-4

Wright, C. D., Eddy, S. L., Wenderoth, M. P., Abshire, E., Blankenbiller, M., and Brownell, S. E. (2016). Cognitive difficulty and format of exams predicts gender and socioeconomic gaps in exam performance of students in introductory biology courses. *CBE Life Sci. Educ.* 15:ar23. doi: 10.1187/cbe.15-12-0246

Yeong, F. M., Chin, C. F., and Tan, A. L. (2020). Use of a competency framework to explore the benefits of student-generated multiple-choice questions (MCQs) on student engagement. *Pedagogies* 15, 83–105. doi: 10.1080/1554480X.2019.1684924

Ying, M. H., and Yang, H. L. (2008). Computer-aided generation of item banks based on ontology and bloom's taxonomy. (International Conference on Web-Based Learning, pp. 157–166). Berlin, Heidelberg: Springer.

Zaidi, N. L. B., Grob, K. L., Monrad, S. M., Kurtz, J. B., Tai, A., Ahmed, A. Z., et al. (2018). Pushing critical thinking skills with multiple-choice questions: does bloom's taxonomy work? *Acad. Med.* 93, 856–859. doi: 10.1097/ACM.0000000000002087

Zaidi, N. L., Grob, K. L., Yang, J., Santen, S. A., Monrad, S. U., Miller, J. M., et al. (2016). Theory, process, and validation evidence for a staff-driven medical education exam quality improvement process. *Med. Sci. Educ.* 26, 331–336. doi: 10.1007/s40670-016-0275-2

Zaidi, N. B., Hwang, C., Scott, S., Stallard, S., Purkiss, J., and Hortsch, M. (2017). Climbing Bloom's taxonomy pyramid: Lessons from a graduate histology course. *Anat. Sci. Educ.* 10, 456–464. doi: 10.1002/ase.1685

Zheng, A. Y., Lawhorn, J. K., Lumley, T., and Freeman, S. (2008). Application of bloom's taxonomy debunks the" MCAT myth". *Science* 319, 414–415. doi: 10.1126/science.1147852

Zorluoglu, S. L., Kizilaslan, A., and Yapucuoglu, M. D. (2020). The analysis of 9th grade chemistry curriculum and textbook according to revised bloom's taxonomy. *Cypriot J. Educ. Sci.* 15, 9–20. doi: 10.18844/cjes.v15i1.3516