REID (Research and Evaluation in Education), 11(2), 2025, 183-198
Available online at: http://journal.uny.ac.id/index.php/reid

Score conversion methods with modern test theory approach: Ability, difficulty, and guessing justice methods

Siti Nurjanah*; Muhammad Iqbal; Siti Nurul Sajdah; Yohana Veronica Feibe Sinambela; Shaufi Ramadhani
Universitas Negeri Yogyakarta, Indonesia
*Corresponding Author. E-mail: siti960pasca.2023@student.uny.ac.id

ARTICLE INFO

Article History
Submitted:
November 14, 2023
Revised:
November 25, 2025
Accepted:
December 5, 2025

Keywords
1-PL; item response theory; R program; Rasch model

Scan Me:

ABSTRACT

The one-parameter logistic (1-PL) model is widely used in Item Response Theory (IRT) to estimate student ability; however, ability-based scoring disregards item difficulty and guessing behavior, which can bias proficiency interpretations. This study evaluates three scoring alternatives derived from IRT: an ability-based conversion, a difficulty-weighted conversion, and a proposed guessing-justice method. Dichotomous responses from 400 students were analyzed using the Rasch (1-PL) model in the R environment with the ltm package. The 1-PL specification was retained to support a parsimonious and interpretable calibration framework consistent with the comparative scoring purpose of the study. Rasch estimation produced item difficulty values ranging from −1.03 to 0.18 and identified 268 unique response patterns. Ability-based scoring yielded only eight score distinctions, demonstrating limited discriminatory capacity. In contrast, the guessing-justice method produced a substantially more differentiated distribution, with approximately 70 percent of patterns consistent with knowledge-based responding and 30 percent indicative of guessing. The findings indicate that scoring models incorporating item difficulty and guessing behaviour provide a more equitable and accurate representation of student proficiency than traditional ability-based conversions. The proposed approach offers a practical and implementable alternative for classroom assessment and can be applied using widely accessible spreadsheet software such as Microsoft Excel.

This is an open access article under the **CC-BY-SA** license.

To cite this article (in APA style):
Nurjanah, S., Iqbal, M., Sajdah, S. N., Sinambela, Y. V. F., & Ramadhani, S. (2025). Score conversion methods with modern test theory approach: Ability, difficulty, and guessing justice methods. *REID (Research and Evaluation in Education)*, *11*(2), 183-198. https://doi.org/10.21831/reid.v11i2.67484

INTRODUCTION

Item analysis using Classical Test Theory (CTT) is relatively easy to conduct, but the use of Item Response Theory (IRT) is increasingly recommended in line with current technological developments (Batool et al., 2023; Hergesell, 2022). IRT addresses key limitations of CTT (Polat, 2022; Triono et al., 2020). A major limitation of CTT is its inability to distinguish examinee characteristics from item characteristics (Subali et al., 2020). Test developers face significant challenges in obtaining examinees for field tests of different instruments due to shortcomings that are specific to certain groups (group-dependent). Another limitation is score dependence on the specific test form, which makes it difficult to compare individuals who completed different test versions (Hambleton et al., 1991). While IRT is often regarded as superior to CTT, some studies report similar outcomes from both approaches, particularly when tests are short, unidimensional, and administered to homogeneous groups.

Person and item estimates obtained from CTT and IRT often show a considerable degree of similarity. The level of consistency in item statistics across different samples, which is often seen as the theoretical advantage of IRT models, was found to be comparable for both

measurement frameworks (Fan, 1998; Subali et al, 2020). CTT and IRT can be valuable in offering a quantitative evaluation of items and scales throughout the content validity phase. Items that do not provide sufficient information may be removed (Batool et al., 2023). To enhance the content validity of measures, it is advisable to use either the CTT or the IRT, depending on the individual circumstances and type of measure (Cappelleri et al., 2014). These similarities do not imply equivalence between the two frameworks; rather, they highlight that the advantages of IRT become meaningful when item difficulty, response behaviour, and ability variation influence score interpretation.

Due to these advantages, researchers have reported positive results when applying IRT methodologies (Triono et al., 2020). The results of the analysis using IRT proved to provide more information from the items (Gorter et al., 2020; Hu et al., 2021). IRT approaches can be used to construct different attachment scales that possess beneficial psychometric properties. These characteristics include unidimensionality, invariance, and high reliability (Fraley et al., 2000; Retnawati, 2014; Hambleton & Swaminathan, 1985). IRT is a remarkably effective method for creating, assessing, and improving questionnaires. It produces accurate, reliable, and relatively brief instruments that minimize response burden (Edelen & Reeve, 2007). Furthermore, the IRT method has also been proven to develop more efficient tests, fairer ability measurements, and more comprehensive item analysis (Hambleton et al., 1991; Chalmers, 2012; Embretson & Reise, 2000). IRT models estimate the probability of a correct response based on item difficulty, discrimination, and guessing parameters (Ariyadi, 2025). Guessing fairness refers to ensuring that students' scores are not distorted by guessing behavior, especially among low-ability students. In such contexts, CTT lacks the capacity to account for guessing and item-level properties. This distinction is central to the present study, which applies IRT not merely for item calibration but to explore score conversion methods that address fairness, particularly in relation to guessing behavior. This principle is applied through ability estimation that considers the overall response pattern.

CTT suffers from major limitations in failing to account for the psychometric properties of test items and respondents' behaviors. In terms of scoring, IRT-based approaches offer significant advantages over the traditional "number right" scoring methods commonly employed in CTT tests (Abedalaziz & Leng, 2018). Specifically, when estimating an examinee's score using IRT. Prior research examined the ability of IRT to identify bias in items. This certainly becomes an advantage in testing using IRT. It has been discovered that techniques that integrate data on the relationship between item responses and latent trait or observed score, along with data on the distribution of the latent trait or observed test score, are effective in detecting item bias (Mellenbergh, 1989; Stark et al., 2006). Indeed, earlier research has discovered numerous additional advantages of IRT. Nevertheless, the researchers focus on score conversion by employing the IRT approach to examine the feasibility of implementing it in the classroom.

Unlike prior research, the current study investigates the potential of employing the IRT approach using the R software to uncover score conversion methods, based on the findings obtained. Indeed, research utilizing the common IRT scoring practice of weighted or simple sum scores "does not take full advantage of item response theory models" and yields inconsistent results (Hambleton & Jones, 1993). Previous research has focused on comparing psychometric properties produced by CTT and IRT, evaluating item fit, test reliability, and detecting item bias using IRT procedures (Mellenbergh, 1989; Stark et al., 2006). Nevertheless, there is limited empirical investigation on how IRT-derived parameters can be directly leveraged for classroom score conversion and grading practices, especially using accessible tools such as R software. Previous studies have relied mainly on simple sum scores or weighted scores which do not fully utilize the advantages of IRT models and may produce inconsistent results (Hambleton & Jones, 1993). Polat (2022) examines how the results of IRT (or Rasch) estimates can be used for score conversion on formative tests or objective tests (e.g. multiple-choice) but only up to the power of differentiation, not up to guessing.

Hence, this study aims to implement a more sophisticated scoring approach by leveraging Item Response Theory (IRT) variables derived from the R program, including ability, item difficulty, and respondent score patterns. The objective of this study is to evaluate and compare the most optimal and equitable technique of converting scores among the three methods. The study's findings will propose an alternative approach for educational instructors to convert scores, enabling them to deliver more equitable and accurate assessments to their students. The novelty of this study lies in developing and comparing three score conversion methods that explicitly incorporate IRT parameters and introduce the concept of guessing fairness as a theoretical and practical contribution.

METHOD

Research Design and Data Source

This study employed a quantitative descriptive design using secondary data obtained from Mahmud (2021). The dataset comprised responses from 400 eighth-grade students to ten dichotomous multiple-choice items modeled on the TIMSS framework. The items targeted numerical content across knowledge, application, and reasoning domains, forming a coherent construct suitable for latent-trait modeling (Mullis & Martin, 2017). Using an established item set ensured that the comparison of scoring procedures was not influenced by item instability or developmental flaws.

Preliminary Analyses and Assumption Evaluation

Analysis using the IRT approach requires data pre-analysis before performing analysis with the 1-PL model. The analysis procedure includes: (1) examining the response data; (2) detecting outliers; (3) testing IRT assumptions, including dimensionality and local independence; (4) testing model fit with the AIC and BIC indices; and (5) 1-PL IRT model analysis, including item parameter (b) estimation and ability (θ) estimation.

Data screening confirmed complete responses and adequate variability across all items. Unidimensionality was assessed using parallel analysis, which indicated a single dominant factor. Anderson et al. (2017) demonstrated that when a dominant latent trait exists, unidimensional IRT models recover item and person parameters with minimal loss of precision even in the presence of minor secondary dimensions. Because unidimensionality was satisfied, the assumption of local independence was considered plausible (Retnawati, 2014).

Model Calibration

The item response data were calibrated using the Rasch (1-PL) model implemented in the ltm package in R (Rizopoulos, 2006). Calibration yielded estimates of item difficulty (b) and person ability (θ), which served as the basis for subsequent score conversions.

Model Selection and Justification for the Rasch Specification

Although 1-PL, 2-PL, and 3-PL models were fitted, the 1-PL model was retained despite slightly higher AIC and BIC values (AIC = 5281.387; BIC = 5325.293). The short and homogeneous ten-item instrument meets conditions under which Rasch estimation is known to be stable. Anderson et al. (2017) found that parameter estimates and examinee rank ordering remained highly consistent across 1-PL, 2-PL, and bifactor models, with standard error differences ranging from 0.00 to 0.03, and normative rank stability exceeding 65%, indicating minimal inferential gain from more complex models. van Rijn et al. (2016) reported 3-PL guessing parameters often produce unstable estimates even with very large samples, and that the practical improvement in model fit beyond the 2-PL model is typically negligible. Besides, inspection of item characteristic curves (ICCs) indicated that the 1-PL model provided the most coherent functional form.

Moreover, given that the study's purpose is comparative scoring analysis, model parsimony and interpretability were prioritized over marginal numerical fit advantages. The use of the Rasch model is consistent with the analytical purpose of this study, as the investigation focuses on how alternative scoring procedures operate when the underlying item parameters are derived from a one-parameter framework. Therefore, parameters with 2-PL or 3-PL models, which involve discrimination and pseudo-guessing parameters, are not required in the analysis process. The characteristics of the data obtained from ten items and the limited number of respondents, totaling 400, will result in less stable estimates using the 2-PL or 3-PL model (Hambleton & Swaminathan, 1985). In this context, Rasch calibration serves as a methodological baseline that allows the scoring conversions to be examined as potential remedies for the well-known limitations of the 1-PL model, rather than as outcomes which are intended to optimize latent-trait estimation.

Assessment of Item Fit and Validity

Item validity was assessed through Rasch item-fit diagnostics and visual examination of item characteristic curves (ICCs). All items exhibited acceptable fit patterns and were retained. van Rijn et al. (2016) found that even when statistical misfit occurs, the practical consequences for proficiency-level classifications and mean scores are typically negligible in well-constructed assessments.

Detection of Guessing Behavior

The guessing behavior was detected using a probability-based procedure, which is grounded in the Rasch model. For each respondent–item pair, the probability of a correct response was computed using the logistic function $P(\theta, b)$. The correct responses were classified as guess-based when the model-implied probability of a correct response was below 0.25, which represents the random-chance level for four-option items. Further, van Rijn et al. (2016) documented that low-probability correct responses are associated with systematic residual patterns suggestive of guessing, particularly at the lower end of the ability distribution. Guess-correct responses were awarded 25% of full credit, whereas correct responses above the threshold received full credit.

Scoring Procedures

Three scoring metrics were derived from the calibrated parameters. The first metric is the ability-based scoring, which transformed θ estimates into a 1–100 scale for interpretability. The second metric is the difficulty-based scoring, which converts item difficulty parameters into positive values following Baker's (2001) conventional range of −3 to 3. The item response theory (IRT) model with scoring methods using item difficulty parameters does not show an absolute scale, so the values obtained from IRT analysis results can be linearly transformed without changing their actual meaning. Hambleton and Swaminathan (1985) explain that adding numbers to all item difficulty parameters is a permissible linear transformation since this addition does not change the function of item characteristics or the probability of answering correctly. Based on this principle, the item difficulty values (IRT 1-PL output) are converted into scores by adding 3 to the difficulty value of each item to shift the range of item difficulty values, which are usually in the range of -3 to 3 (Baker, 2001). This method recognizes that more difficult questions should have a higher value, assuming that students who are able to answer difficult questions can also answer easy questions. Then, the third metric is the guessing-adjusted scoring, incorporating the probabilistic correction to mitigate score inflation from random success. All scoring procedures were implemented in Microsoft Excel to ensure transparency, replicability, and accessibility for practitioners.

FINDINGS AND DISCUSSION

Item Difficulty Analysis Using the 1-PL IRT Model

Figure 1 shows the outcomes of item plots generated using the IRT-1 PL method in the R programming language. The item difficulty level was assessed by examining the 1-PL plot output generated by the R application, as depicted in Figure 1. Item 9 has the highest level of difficulty, whilst item 2 has the lowest level of difficulty. Figure 2 shows the overall level of item difficulty.

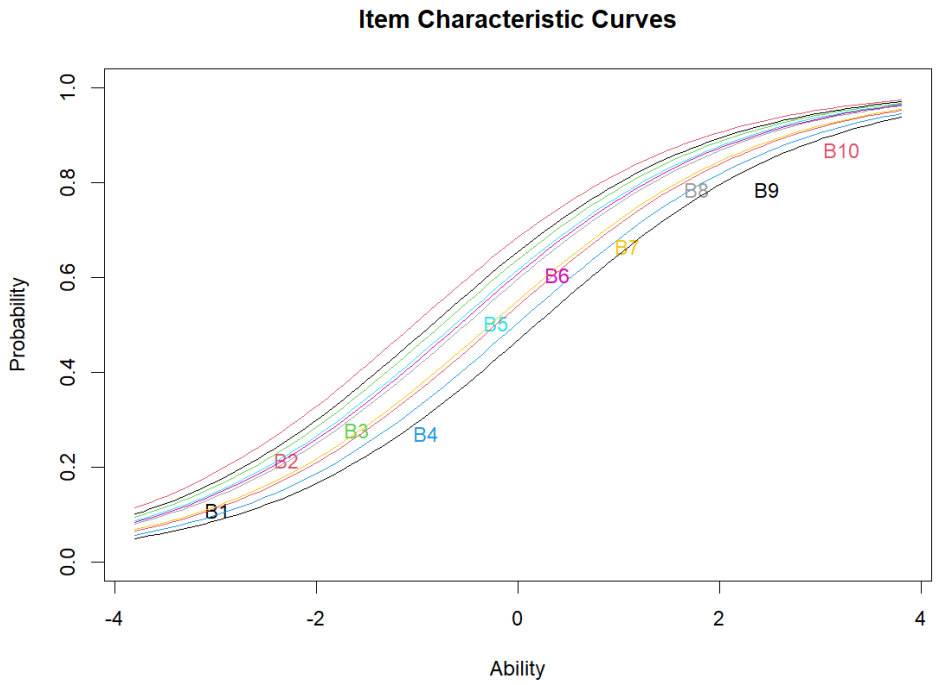

Figure 1. Output of 1-PL Plot from the R Program

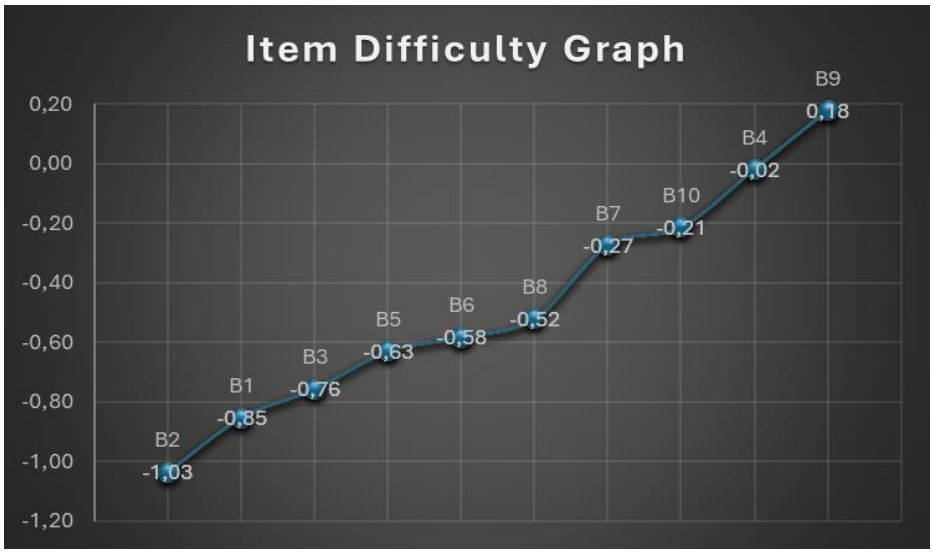

Figure 2. Item Difficulty Falls Between -1.03 and 0.18

According to Baker and Kim (2017), all items have an average level of difficulty based on their qualifications. Setiawati et al. (2023) explain the application of 1-PL where the difficulty index for items below -2 is included in the easy category, between -2 and +2 is categorized as medium and more than +2 is categorized as hard. This previous research supports the results of our findings, where from the results above, it is clear that Item 2 has the lowest difficulty index,

with an index of -1.03, and the item only has a positive index on question number 9, where all questions are included in the medium category. The evaluated items are considered satisfactory due to the absence of items classed as highly difficult or overly easy. The application of IRT analysis using the 1-parameter logistic (1-PL) model, facilitated by the R programming language, indicates that the 10 items under examination possess validity. As a result, the entirety of the item set can be utilized in the forthcoming scoring analysis.

Respondents' Answer Patterns

Through the R program for analysis, we found 268 respondent answer patterns out of a total of 400 responses. The researchers conducted score conversion and subsequent analysis to compare the three score conversion methods utilizing the 268 answer patterns. The items were renamed and reordered for the purpose of analysis. The items that were previously randomized based on their difficulty level, as depicted in Figure 1, were now rearranged in accordance with their difficulty level. The excel analysis has been sorted based on the level of difficulty, with items 1 to 10. Item 1 is the easiest item, while Item 10 is the most difficult one. The purpose of this sorting is to facilitate the researcher's analysis of the respondent's answer pattern. In fact, modeling of the application of item response theory has often been carried out, including by Zhou et al. (2023) where the aim of their study was to find out how students' abilities were based on tests that had been designed with the help of e-learning. This is relevant to research conducted by researchers where modeling can also be carried out on item response theory based on student answer patterns, and also sorted based on the level of difficulty that has been analyzed from item response theory.

Conversion of Respondent Scores Based on Ability ($\boldsymbol{\theta}$)

The respondents' scores were converted using the Microsoft Excel application by inputting the previously acquired difficulty level value (b) using the R program. The ability ($\boldsymbol{\theta}$) of each responder was computed using the 1-PL formula. The θ value is subsequently transformed into a numerical number within the range of 1 to 100. According to Xia et al. (2019), assessment grounded in IRT is essential, as student scores are generated on a 0–100 scale based on estimated ability and adjusted according to the standard deviation. The preference of this range was based on its prevalence as the score range most frequently employed in educational institutions. Utilizing this method for value conversion yields a reduced degree of variability and an extensive spectrum of values across abilities. Among the 268 answer patterns, there are only eight distinct value distributions. Indeed, this outcome lacks representativeness in assessing the respondent's proficiency and fails to differentiate across respondents. Furthermore, while using this scoring method in educational institutions, it lacks contextualization and pertinence. This limitation is further compounded by the fact that many teachers feel more confident in constructing test items than in utilizing assessment results for instructional decision-making (Koloi-Keaikitse, 2017). This indicates a critical need for enhancing teachers' competencies in interpreting and leveraging theta scoring results to inform pedagogical practices and improve learning outcomes.

Conversion of Respondent Scores Based on Difficulty

This method employs a scoring system that transforms the difficulty levels of items into corresponding scores. The score of each item varies based on the item's level of difficulty. The level of difficulty of an item directly correlates with its score, resulting in a higher score for more difficult items. This method is unrelated to the amount of correct answers. The influence on the respondent's score is that it is feasible for a responder who answers fewer tough questions correctly to receive a higher score than a respondent who answers more easy questions correctly. The scoring approach presented is more equitable than the ability-based scoring method as it incorporates a higher level of granularity, with item scores being contingent upon their respective

difficulty levels. This strategy is more effective in representing learners' abilities. It is anticipated that learners who can respond to difficult questions will also be capable of answering easier questions, as it necessitates a greater level of competency to overcome difficult question types. In some instances, when participants respond properly to difficult questions but wrongly to easier ones, it is presumed that they were influenced by external factors, such as being misled by the placement of incorrect response choices, crossing out options, or mistakenly selecting the wrong answer. It is considered that those who can handle complex analytical requirements when answering challenging questions are also capable of answering simpler questions that simply require basic analysis.

Table 1. Conversion of Item Difficulty to Item Score

No.	Item	Difficulty	Difficulty Classification	Difficulty Conversion	Correct Scores
1	Q2	-1.03	Average	1.97	7.77
2	Q1	-0.85	Average	2.15	8.49
3	Q3	-0.76	Average	2.24	8.87
4	Q5	-0.63	Average	2.37	9.37
5	Q6	-0.58	Average	2.42	9.56
6	Q8	-0.52	Average	2.48	9.80
7	Q7	-0.27	Average	2.73	10.78
8	Q10	-0.21	Average	2.79	11.01
9	Q4	-0.02	Average	2.98	11.79
10	Q9	0.18	Average	3.18	12.57

The conversion of item difficulty into item score (Table 1) is accomplished by adding each item difficulty value to the number 3. According to Baker's (2021) theory, item difficulty typically falls within the range of -3 to 3. Researchers cannot definitively assert that this conversion is the most appropriate, as there is a possibility that an item with a difficulty of -3, when increased by +3, would yield a score of 0. Thus, if this item is incorrect, it will not have any impact on the ultimate score. Subsequently, this approach garnered criticism due to the lack of contribution to the score. Consequently, it is argued that items devoid of significance should be outright eliminated. Nevertheless, the researcher in this instance made an approximation by using reliable rules (such as those proposed by Baker) to estimate the score, which was deemed to be more accurate. Indeed, things of low quality (with extremely low difficulty) will be excluded from the score conversion process. Although no items in this analysis possess this property, all items are valid with an average level of difficulty. If there are any, the researchers will eliminate them as invalid items before proceeding to the score conversion phase. Metsämuuronen (2023) concluded that weighting items based on difficulty produces more accurate scoring, where harder items receive greater weight than easier ones. Therefore, the higher the coefficient value of the difficulty level, the higher the scoring weight will be given compared to questions with a lower level of difficulty.

The difficulty-based assessment method provides a fairer alternative for teachers in evaluating students. In classroom implementation, teachers can use this method to design a more objective assessment system, where students who are able to answer difficult questions correctly receive rewards commensurate with their cognitive efforts (Brookhart, 2013). As a concrete example, teachers can implement a question weighting system in daily tests or midterm exams, where questions with high difficulty levels are given greater weight. To facilitate the implementation of this weighting system more efficiently, weighted classification models such as TFPOS-IDF and Word2vec have proven effective in identifying and classifying questions that measure various cognitive levels, making assessment fairer and more accurate toward higher-order thinking skills (Mohammed & Omar, 2020). By utilizing these classification models, teachers can more easily construct balanced and representative assessment instruments for all cognitive levels they wish to measure, ranging from basic knowledge to complex analysis and evaluation capabilities.

Conversion of Respondents' Scores Considering Guessing Justice

This method accounts for the possibility that a respondent may correctly guess the answer based on the pattern of their responses to each item. This analysis differs from previous research that aimed to reduce the impact of guessing on scores by using unique test formats (Pollard, 1989). The premise underlying our analysis is that a correct response obtained through random guessing does not accurately represent the respondent's true underlying ability. This assumption is grounded in the understanding that successfully answering an item by chance alone, rather than through the demonstration of the required proficiency, does not provide a valid indication of the examinee's actual knowledge or competence in the measured domain. Ideally, these guessed items should not be included in the respondent's total score, as the final score is meant to reflect the respondent's actual skill or competence.

The literature remains divided on whether examinees should be mandated to provide responses for all test items, even through blind guessing, or be discouraged from guessing by penalizing incorrect answers (Burton, 2002). However, it is important to recognize that examinees can be in one of three subjective states when confronted with a test question: (1) full knowledge, (2) partial knowledge, or (3) absence of knowledge, though this self-assessment may not always be accurate (Ben-Simon et al., 1997). Correct responses can arise from true knowledge, random guessing, or a combination of both (partial knowledge) (Abu-Ghazalah et al., 2023). Therefore, it is crucial to implement assessment methods that can distinguish between correct answers stemming from genuine proficiency and those obtained through guessing, in order to ensure that the final scores accurately reflect examinees' actual competence levels.

Respondents were justified in making reasonable guesses by taking multiple factors into consideration. The investigated assessment is a multiple-choice examination comprising four alternative answer choices. Regardless of the respondent's lack of knowledge about the required abilities for an item, they still have a 25% chance of answering correctly (the element of guessing is actually a disadvantage of multiple-choice tests). With multiple answer options, an examinee can correctly guess a difficult item through random guessing, even if they don't actually have the required proficiency to answer it correctly (Lin, 2018). A score difference of 0.25 or greater, even if it is minimal, such as 0.31 (a score difference of 0.6), is considered indicative of the respondent's knowledge and will enhance the likelihood of answering correctly. Under these circumstances, the respondents utilized their knowledge to consider their responses to the question, indicating that their answers were not mere guesses. Their ability to answer correctly on a given item was attributed to the contribution of their knowledge in their cognitive processes. This indicates that the assessment framework for scoring under investigation is designed to accommodate and account for partial knowledge exhibited by the examinees, where the assessors could implement more stringent selection criteria, but would need to have a strong justification and rationale for doing so. Thus, if a respondent answers properly on the n^{th} item, it is regarded as a guess only if the difficulty of the item does not align with the respondent's ability.

For instance, in Table 2, a responder with a θ value of -0.5 would be categorized as guessing if they answered item 10 correctly. This is because the likelihood of answering item 10 correctly is lower than the minimal probability of 0.25. Thus, individuals who have an ability score of ≥ 0.00 are deemed not to be guessing on the 10 items, since their ability is regarded as surpassing the overall difficulty of the items. Although the response pattern of individuals with the characteristic $\theta \geq 0.00$ may appear unusual (e.g. 1111110001), it is not deemed a random guess for item 10 because it is believed that they possess the ability to answer that particular item. However, the inability of respondents with high ability to answer easy items correctly was attributed to other influencing factors, as previously indicated. The author acknowledges the limits of the analysis in not examining answer patterns that depart from the proposed framework. An incorrect answer may be due to guessing (uninformed), an inaccurate belief (misinformation), or other construct-irrelevant factors such as poor item construction, test fatigue, or other human errors (Abu-Ghazalah et al., 2023). These alternative explanations for incorrect responses high-

light the need for a more comprehensive understanding of the factors that can influence examinee performance, beyond just the binary distinction between guessing and genuine knowledge. Further exploration of these nuanced factors could lead to more insightful interpretations of assessment data and inform the development of enhanced evaluation approaches.

Table 2. The Correspondence Between Ability (θ) and the Probability of Answering Correctly on the nth Item

θ	P1	Q1	P2	Q2	P3	Q3	P4	Q4	P5	Q5
	0.71	0.29	0.65	0.35	0.61	0.39	0.55	0.45	0.53	0.47
-0.50	**P6**	**Q6**	**P7**	**Q7**	**P8**	**Q8**	**P9**	**Q9**	**P10**	**Q10**
	0.51	0.49	0.40	0.60	0.38	0.62	0.31	0.69	0.24	0.76

After evaluating the respondents' guesses for each item, the correct score obtained by guessing was determined. The score obtained by guessing correctly is calculated as 25% of the points obtained by answering honestly (Table 3). This is because 25% represents the minimal probability of guessing correctly in an item. Considering the multiple-choice format with four alternative options per item, the minimal expected score that can be achieved through random guessing is 25%, as the probability of selecting the correct answer by chance alone is 1 out of 4, or 25% (Lau et al., 2011). This assessment method exhibits subjectivity as it adjusts to real-world circumstances. If a student receives a score of 0 for guessing, many students would likely protest. This is because the students' answer is true for the nth item, yet they are still awarded a score of 0.

Table 3. A Correct Score with a Guess Attempt is Worth 25% of the Correct Score with an Honest Attempt

Item	Honest Correct Score	Correct Guess Score
1	7.77	1.94
2	8.49	2.12
3	8.87	2.22
4	9.37	2.34
5	9.56	2.39
6	9.80	2.45
7	10.78	2.69
8	11.01	2.75
9	11.79	2.95
10	12.57	3.14

The investigation using the third method revealed 187 patterns of truthful responses (70% proportion) and 81 patterns of responses based on guessing (30%). Hence, the vast majority of the responses provided in this examination were honest and deemed to accurately reflect the ability to be evaluated.

Comparison of Score Conversion Methods Based on Ability, Difficulty, and Guessing Justice

This study emphasizes the contribution of developing and comparing three IRT-based score conversion approaches, which have not been widely discussed in previous research. Unlike the common practice of relying solely on ability (θ) or difficulty (b) scores, this study incorporates guessing justice, resulting in a conversion model that can distinguish between answers based on conceptual mastery and answers resulting from guessing. The integration of these three IRT parameters ability, difficulty, and guessing patterns makes this study one of the first to offer a more comprehensive, fair, and relevant score conversion framework for application in school assessments. Thus, this study makes a new contribution to the development of modern score conversion methods that are more targeted to the needs of learning evaluation.

The value conversion for the three methods was finally carried out using the Norm-Referenced Assessment (NRA) and Criterion-Referenced Assessment (CRA) approaches. The NRA and CRA represent two broad categories of assessment approaches utilized in educational settings. The NRA evaluates student performance relative to that of a peer group, whereas CRA evaluates performance against predetermined standards or learning criteria (Wallace & Ng, 2023).

While NRA and CRA represent widely adopted assessment approaches, both methods have faced various criticisms and debates within the educational community. NRA is considered more relevant to real-world evaluation and more equitable (Ertoprak & Dogan, 2016), whereas CRA is deemed fairer in assessing based on survey data (Wallace & Ng, 2023). NRA is also seen as more familiar to external stakeholders in terms of grading purposes, making it more suitable for summative assessment (Lok et al., 2016). Yeoh and Woods (2006) have demonstrated that NRA grading can be effectively implemented using Fuzzy methods for formative assessment. Despite the diverse critiques aimed at both assessment approaches, the researchers have chosen to utilize NRA and CRA because they are the most commonly used assessment methods in classroom settings (Wallace & Ng, 2023). This widespread adoption of NRA and CRA makes them representative and relevant for the intended scoring purposes of the study.

A summary of the frequency of respondents' scores is presented in Table 4. The value conversion was ultimately conducted utilizing the NRA and CRA employing three distinct methods. Table 4 shows a summary of the distribution of scores provided by the respondents.

Table 4. Recapitulation of the Frequency of Respondents' Scores from the Three Assessment Methods

Ability			Difficulty			Guessing Justice		
Frequencies	**CRA**	**NRA**	**Frequencies**	**CRA**	**NRA**	**Frequencies**	**CRA**	**NRA**
A	1	1	A	10	10	A	10	10
B	9	0	B	67	58	B	67	65
C	211	31	C	108	102	C	88	86
D	41	189	D	65	71	D	69	70
E	6	47	E	18	27	E	34	37

Notes: Columns with yellow highlights indicate the mode

The assessment approach exhibits the poorest distribution of scores in the method of ability. The scores are highly focused exclusively on category C. The difficulty and guessing justice ways are evenly and reasonably distributed. The method of scoring based on the guessing justice method is the most optimal distribution.

Who Benefits from Each Scoring Method?

Each scoring method confers advantages to individuals with distinct characteristics. Implementing score conversion based on ability will be advantageous for respondents with poor ability. In contrast to difficulty-based approaches, low-ability participants will gain an advantage. This is because, even if they answer easy questions correctly, their scores will not differ significantly from those of high-ability participants who answer difficult questions correctly (Table 5). This is due to the fact that ability-based scoring disregards the level of difficulty of each question. Low-ability responders, who are more likely to guess, will gain an advantage compared to the guessing justice scoring method. Regardless of whether they make an appropriate guess or not, individuals are deemed capable of answering an item since the ability scoring technique disregards the factor of guessing. Responses that are answered properly through a guessing attempt will be acknowledged with an accurate score (retaining the maximum score).

The difficulty-based scoring method will be advantageous for respondents who possess a high level of skill (Table 6). Their endeavors to accurately respond to the questions will be compensated according to the score they ought to receive (the score adjusts to the level of difficulty). This method is more equitable than the ability-based scoring method. Nevertheless, this scoring

method still possesses disadvantages. Guessing may lead to the acquisition of accurate answers for difficult items, potentially resulting in an underestimation of one's actual proficiency.

Table 5. One of the Answer Patterns of Low Ability Respondents. Low-Ability Respondents Benefit the Most from the Ability-Based Assessment Method

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Teta	Judgement	Scoring Methods		
													Ability	Difficulty	Guessing Justice
242	0	1	1	0	0	0	0	0	0	1	-1.0	Guessing	58	30	17

Table 6. One of the Answer Patterns of High Ability Respondents. High Ability Respondents Benefit the Most from the Difficulty-Based Scoring Method

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Teta	Judgement	Score Based On...		
													Ability	Difficulty	Guessing Justice
P6	1	1	1	1	1	1	1	0	1	1	1.0	Knowledge-based responding	67	89	89

The implementation of the guessing justice scoring method will be advantageous for the rater/researcher as it will accurately reflect the respondent's competence. Conversely, this approach will greatly impact respondents who attempt to guess answers, particularly those with lower cognitive abilities, in a detrimental manner (Table 7). As the scoring of this method is still determined by the difficulty level of the item, respondents with high ability who make honest efforts will still be advantaged by this method. Hence, the individuals who derive advantages from this approach are typically identical to those who benefit from the difficulty-based approach. The guessing justice scoring method is the most equitable scoring approach when considering the accurate reflection of actual abilities.

Table 7. One of the Answer Patterns of Low Ability Respondents. Low-Ability Respondents are Most Disadvantaged by Assessment Methods Based on Guessing Justice

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Teta	Judgement	Score Based On...		
													Ability	Difficulty	Guessing Justice
P242	0	1	1	0	0	0	0	0	0	1	−1.0	Guessing	58	30	17

The analysis of who benefits from the different scoring methods provides valuable insights. However, the fundamental aim of a fair assessment approach is to minimize bias in the evaluation of diverse students (Zieky, 2016). Research has shown that teachers' conceptualizations of fair assessment vary, with some aligning more closely with the principle of equality, while others tend towards the notion of equity (Murillo & Hidalgo, 2018). This distinction is significant, as an equality-based approach may strive for a standardized, one-size-fits-all evaluation, whereas an equity-focused perspective recognizes the need to account for individual differences and provide tailored support to ensure all students have the opportunity to demonstrate their true potential. The challenge lies in striking the right balance between these principles of equality and equity when selecting the appropriate scoring method.

As the analysis has highlighted, different scoring approaches tend to advantage or disadvantage students with varying ability levels. To ensure a truly fair assessment, the researchers must carefully consider the intended purposes, the diversity of the student population, and the potential consequences of the evaluation outcomes. By deeply understanding these underlying principles, they can work towards developing and implementing assessment practices that promote inclusivity and empower all students to showcase their authentic competencies.

Further Evaluation: Remedial, Pass, or Enrichment?

Further assessment was conducted using scores, with the guessing justice method identified as the most equitable method. The researcher classifies the responders according to the pre-

defined categories in NRA and CRA. The judgment was made to provide remedial education for students in categories D and E, to allow students in category C to pass, and to offer enrichment programs for students in categories B and A. The pass group can opt for either remedial or enrichment activities based on their self-reflection. Table 8 shows the distribution of students in the categories of remedial, pass, or enrichment.

Table 8. Percentage of the Further Evaluation of The Respondents' Assessment Scores

Judgment	The Number of Respondents		Percentages	
Enrichment	77	75	29%	28%
Pass	88	86	33%	32%
Remedial	103	107	38%	40%

The implementation of IRT-based scoring offers practical benefits for differentiated instruction. Teachers can use these categorizations to design targeted interventions: the 38-40% of students requiring remedial support should receive instruction focused on foundational concepts, the 33-32% in the pass category can choose between consolidation or enrichment activities, while the 29-28% eligible for enrichment can engage with advanced problem-solving tasks and real-world applications of mathematical concepts.

The decision to provide remedial or enrichment programs is a critical one in educational settings, as high-performing students often seek competitive advantages through enrichment, while low-achievers use these programs to improve their performance and meet academic demands (Tan & Liu, 2023). This ability to leverage assessment results to design appropriate interventions is considered a sub-capability of assessment competence (Pardimin, 2018). By using the assessment data to inform targeted remedial support and enrichment opportunities, teacher can demonstrate a holistic approach to supporting student growth, but it is crucial that the implementation of these programs is guided by a deep understanding of each learner's unique needs and circumstances to ensure the interventions truly address the underlying factors contributing to their performance.

CONCLUSION

This research analyzes three assessment methods: ability-based, difficulty-based, and guessing justice-based. Among these three scoring methods, the guessing justice method has been found to be the most effective. The ability-based scoring technique is advantageous for respondents with low ability, whereas the difficulty-based scoring method and guessing justice are advantageous for respondents with high ability. Respondents who make conjectures about the correct answer will be placed in an unfavorable position when using the guessing justice scoring technique. The disadvantages and advantages discussed in this paper pertain to the final grade. The importance of this research is to provide an overview of giving student assessment scores more fairly by giving grades based on the parameters of ability, difficulty of the questions and also the level of student guessing, so that by combining all the scores based on coefficient values it will provide a more objective picture of student abilities. The researchers recommend further research to add a 4-parameter model as an addition to the assessment method, so that it can provide a more objective picture, and also the need for integration of CAT so that it can speed up the assessment process in schools.

ACKNOWLEDGMENT

We gratefully acknowledge the intellectual guidance and contribution of Dr. Haryanto, whose instruction and insights supported the development of this research and encouraged the use of R programming in our analysis. We also thank Mahmud for providing access to the dataset used as the source of the raw data analyzed in this research.

DISCLOSURE STATEMENT

The authors do not have any potential conflict of interest to disclose.

FUNDING STATEMENT

This work does not receive funding.

ETHICS APPROVAL

Ethical approval was not sought for this study as it relied exclusively on anonymized secondary data (Mahmud, 2021) and did not involve human subjects, intervention, or the collection of any identifiable personal information.

REFERENCES

Abedalaziz, N., & Leng, C. H. (2018). The relationship between CTT and IRT approaches in Analyzing Item Characteristics. *MOJES: Malaysian Online Journal of Educational Sciences*, *1*(1), 64–70. http://mojes.um.edu.my/index.php/MOJES/article/view/12857

Abu-Ghazalah, R. M., Dubins, D. N., & Poon, G. M. K. (2023). Dissecting knowledge, guessing, and blunder in multiple choice assessments. *Applied Measurement in Education*, *36*(1), 80–98. https://doi.org/10.1080/08957347.2023.2172017

Anderson, D., Kahn, J. D., & Tindal, G. (2017). Exploring the robustness of a unidimensional item response theory model with empirically multidimensional data. *Applied Measurement in Education*, *30*(3), 163–177. https://doi.org/10.1080/08957347.2017.1316277

Ariyadi, D. J. (2025). Application of Three-Parameter Logistic (3PL) item response theory in Learning Management System (LMS) for post-test analysis. *Journal of Informatics Development*, *3*(2), 33–46. https://doi.org/10.30741/jid.v3i2.1554

Baker, F. B. (2001). *The basics of item response theory* (2nd edition). ERIC Clearinghouse on Assessment and Evaluation. https://doi.org/10.1007/978-3-319-54205-8_1

Baker, F. B., & Kim, S.-H. (2017). *The basics of item response theory using R*. Springer. https://doi.org/10.1007/978-3-319-54205-8

Batool, I., Shah, A. A., & Naseer, S. (2023). Construction, analysis and calibration of multiple-choice questions: IRT versus CTT. *Archives of Educational Studies (ARES)*, *3*(2), 242–257. https://ares.pk/ojs/index.php/ares/article/view/69

Ben-Simon, A., Budescu, D. V, & Nevo, B. (1997). A comparative study of measures of partial knowledge in multiple-choice tests. *Applied Psychological Measurement*, *21*(1), 65–88. https://doi.org/10.1177/0146621697211006

Brookhart, S. M. (2013). The use of teacher judgement for summative assessment in the USA. *Assessment in Education: Principles, Policy & Practice*, *20*(1), 69–90. https://doi.org/10.1080/0969594X.2012.703170

Burton, R. F. (2002). Misinformation, partial knowledge and guessing in true/false tests. *Medical Education*, *36*(9), 805–811. https://doi.org/10.1046/j.1365-2923.2002.01299.x

Cappelleri, J. C., Jason Lundy, J., & Hays, R. D. (2014). Overview of classical test theory and item response theory for the quantitative assessment of items in developing patient-reported outcomes measures. *Clinical Therapeutics*, *36*(5), 648–662. https://doi.org/10.1016/j.clinthera.2014.04.006

Chalmers, R. P. (2012). Mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1-29. https://doi.org/10.18637/jss.v048.i06

Edelen, M. O., & Reeve, B. B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research*, *16*, 5–18. https://doi.org/10.1007/s11136-007-9198-0

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists* (1st ed.). Psychology Press. https://doi.org/10.4324/9781410605269

Ertoprak, D. G., & Dogan, N. (2016). A research on the classification validity of the decisions made according to norm and criterion-referenced assessment approaches. *Anthropologist*, *23*(3), 612–619. https://doi.org/10.1080/09720073.2014.11891981

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement*, *58*(3), 357–381. https://doi.org/10.1177/0013164498058003001

Fraley, R. C., Waller, N. G., & Brennan, K. A. (2000). An item response theory analysis of self-report measures of adult attachment. *Journal of Personality and Social Psychology*, *78*(2), 350–365. https://doi.org/10.1037/0022-3514.78.2.350

Gorter, R., Fox, J.-P., Riet, G. T., Heymans, M., & Twisk, J. (2020). Latent growth modeling of IRT versus CTT measured longitudinal latent variables. *Statistical Methods in Medical Research*, *29*(4), 962–986. https://doi.org/10.1177/0962280219856375

Hambleton, R. K., & Jones, R. W. (1993). An NCME instructional module on: Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, *12*(3), 38–47. https://doi.org/10.1111/j.1745-3992.1993.tb00543.x

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Springer Science. https://doi.org/10.1007/978-94-017-1988-9

Hambleton, R. K., Swaminathan, H., & Rogers, D. J. (1991). *Fundamentals of item response theory*. Sage publications. https://doi.org/10.2307/2075521

Hergesell, A. (2022). Using Rasch analysis for scale development and refinement in tourism: Theory and illustration. *Journal of Business Research*, *142*, 551–561. https://doi.org/10.1016/j.jbusres.2021.12.063

Hu, Z., Lin, L., Wang, Y., & Li, J. (2021). The integration of classical testing theory and item response theory. *Psychology*, *12*, 1397–1409. https://doi.org/10.4236/psych.2021.129088

Koloi-Keaikitse, S. (2017). Assessment of teacher perceived skill in classroom assessment practices using IRT Models. *Cogent Education*, *4*(1), 1281202. https://doi.org/10.1080/2331186X.2017.1281202

Lau, P. N. K., Lau, S. H., Hong, K. S., & Usop, H. (2011). Guessing, partial knowledge, and misconceptions in multiple-choice tests. *Educational Technology and Society*, *14*(4), 99–110. https://www.jstor.org/stable/jeductechsoci.14.4.99

Lin, C.-K. (2018). Effects of removing responses with likely random guessing under Rasch measurement on a multiple-choice language proficiency test. *Language Assessment Quarterly*, *15*(4), 406–422. https://doi.org/10.1080/15434303.2018.1534237

Lok, B., McNaught, C., & Young, K. (2016). Criterion-referenced and norm-referenced assessments: Compatibility and complementarity. *Assessment and Evaluation in Higher Education*, *41*(3), 450–465. https://doi.org/10.1080/02602938.2015.1022136

Mahmud, M. N. (2021). *Diagnostik kesulitan belajar Matematika siswa SMP kelas VIII di Kota Baubau menggunakan soal-soal model TIMSS (Diagnostics of mathematics learning difficulties for class VIII*

junior high school students in Baubau City using TIMSS model questions). Thesis, Universitas Negeri Yogyakarta.

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research*, *13*(2), 127–143. https://doi.org/10.1016/0883-0355(89)90002-5

Metsämuuronen, J. (2023). Seeking the real item difficulty: Bias-corrected item difficulty and some consequences in Rasch and IRT modeling. *Behaviormetrika*, *50*(1), 121–154. https://doi.org/10.1007/s41237-022-00169-9

Mohammed, M., & Omar, N. (2020). Question classification based on Bloom's taxonomy cognitive domain using modified TF-IDF and word2vec. *PLOS ONE*, *15*(3), e0230442. https://doi.org/10.1371/journal.pone.0230442

Mullis, I. V. S., & Martin, M. O. (2017). TIMSS 2019 assessment frameworks. In *Hacking Connected Cars*. Boston College, TIMSS & PIRLS International Study Center. http://timssandpirls.bc.edu/timss2019/frameworks/

Murillo, F. J., & Hidalgo, N. (2018). Fair assessment conceptions of students. A phenomenographic study from teachers' perspective. *Revista Complutense de Educacion*, *29*(4), 995–1010. https://doi.org/10.5209/RCED.54405

Pardimin. (2018). Analysis of the Indonesia mathematics teachers' ability in applying authentic assessment. *Cakrawala Pendidikan*, *37*(2), 170–181. https://doi.org/10.21831/cp.v37i2.18885

Polat, M. (2022). Comparison of performance measures obtained from foreign language tests according to item response theory vs classical test theory. *International Online Journal of Education and Teaching*, *9*(1), 471–485. https://eric.ed.gov/?id=EJ1327729

Pollard, G. H. (1989). Scoring to remove guessing in multiple choice examinations. *International Journal of Mathematical Education in Science and Technology*, *20*(3), 429–432. https://doi.org/10.1080/0020739890200313

Retnawati, H. (2014). *Teori respons butir dan penerapannya (Item response theory and its application)*. Nuha Medika.

Rizopoulos, D. (2006). Itm: An R package for latent variable modeling and item response theory analyses. *Journal of Statistical Software*, *17*(5), 1–25. https://doi.org/10.18637/jss.v017.i05

Setiawati, F. A., Amelia, R. N., Sumintono, B., & Purwanta, E. (2023). Study item parameters of classical and modern theory of differential aptitude test: Is it comparable? *European Journal of Educational Research*, *12*(2), 1097–1107. https://doi.org/10.12973/eu-jer.12.2.1097

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology*, *91*(6), 1292–1306. https://doi.org/10.1037/0021-9010.91.6.1292

Subali, B., Kumaidi, & Aminah, N. S. (2020). The comparison of item test characteristics viewed from classic and modern test theory. *International Journal of Instruction, 14*(1), 647–660. https://doi.org/10.29333/IJI.2021.14139A

Tan, M., & Liu, S. (2023). A way of human capital accumulation: Heterogeneous impact of shadow education on students' academic performance in China. *SAGE Open*, *13*(4). https://doi.org/10.1177/21582440231207189

Triono, D., Sarno, R., & Sungkono, K. R. (2020). Item analysis for examination test in the postgraduate student's selection with classical test theory and Rasch measurement model.

2020 International Seminar on Application for Technology of Information and Communication (ISemantic), 523–529. https://ieeexplore.ieee.org/abstract/document/9234204/

van Rijn, P. W., Sinharay, S., Haberman, S. J., & Johnson, M. S. (2016). Assessment of fit of item response theory models used in large-scale educational survey assessments. *Large-Scale Assessments in Education*, *4*(1), 10. https://doi.org/10.1186/s40536-016-0025-3

Wallace, M. P., & Ng, J. S. W. (2023). Fairness of classroom assessment approach: Perceptions from EFL students and teachers. *English Teaching and Learning*, *47*(4), 529–548. https://doi.org/10.1007/s42321-022-00127-4

Xia, J., Tang, Z., Wu, P., Wang, J., & Yu, J. (2019). Use of item response theory to develop a shortened version of the. *Scientific Reports*, 9, 1764. https://doi.org/10.1038/s41598-018-37965-x

Yeoh, E.-T., & Woods, P. (2006). Formative assessment using norm-referenced fuzzy evaluations. *WSEAS Transactions on Information Science and Applications*, *3*(10), 1846–1850. https://www.wseas.us/e-library/conferences/2006cscc/papers/534-674.pdf

Zhou, Y., Suzuki, K., & Kumano, S. (2023). State-aware deep item response theory using student facial features. *Front. Artif. Intell., 6,* 1324279. https://doi.org/10.3389/frai.2023.1324279

Zieky, M. J. (2016). Developing fair tests. In *Handbook of test development* (2nd ed.) (pp. 81–99). Taylor and Francis. https://doi.org/10.4324/9780203102961-6