

Evaluating the Fairness of Fiqh Learning Outcome Scores: A Comparative Study between Classical Test Theory and Item Response Theory

Joko Subando¹, Minhayati Saleh², Siti Choiriyah³, Budiharjo⁴

¹ Postgraduate Program, Institut Islam Mambaul Ulum Surakarta, Indonesia

²Department of Mathematics Education, Faculty of Science and Technology Universitas Islam Negeri Walisongo, Indonesia

³Faculty of Tarbiyah, Universitas Islam Negeri Raden Mas Said Surakarta, Indonesia

⁴Sekolah Tinggi Islam Al Mukmin Ngruki Sukoharjo, Indonesia

¹Email: jokosubando@yahoo.co.id

Proceedings of the Kalijaga International Conference on Education 2026

Abstract. Educational measurement commonly applies Classical Test Theory (CTT) and Item Response Theory (IRT) to estimate students' achievement scores. CTT derives observed scores from the total number of correct responses, whereas IRT models students' latent ability by incorporating item characteristics. The one-parameter logistic (1PL) or Rasch model, in particular, estimates ability based solely on item difficulty under the assumption of equal item discrimination. Although Rasch-based approaches have been increasingly adopted in large-scale assessments in Indonesia, empirical evidence comparing Rasch- and CTT-based scoring in religious subject assessments remains limited. This study aims to examine differences in score estimates produced by CTT and the IRT 1PL (Rasch) model, with specific attention to the role of item difficulty. An evaluative quantitative design was employed involving 387 twelfth-grade students from MAN 1 Surakarta who completed a Fiqh achievement assessment in 2023. The dataset consisted of dichotomous responses to multiple-choice items. Students' ability scores were estimated using both CTT and the Rasch model, and differences between the two scoring approaches were analyzed using paired-sample t-tests, as both scores were derived from the same examinees. The results reveal a statistically significant difference between CTT- and Rasch-based score estimates. The mean Rasch score ($M = 71.61$) was significantly higher than the mean CTT score ($M = 59.18$), with $t = -13.298$ and $p < .001$. Further analysis indicates that 78% of students experienced lower scores under CTT, while only 21% benefited relative to Rasch-based estimates. These findings suggest that Rasch-based scoring provides a more equitable and measurement-consistent representation of students' latent ability than CTT, particularly in assessments containing items with varying levels of difficulty. The study underscores the importance of adopting modern measurement models to enhance fairness and interpretability in educational assessment.

Keywords: Classical Test Theory, Evaluating, Fairness, Item Response Theory

INTRODUCTION

Assessment of learning outcomes constitutes a fundamental component of educational systems, as it serves as the basis for evaluating academic achievement and informing various high-stakes decisions that influence students' future trajectories[1]. In this context, the issue of fairness in assessment has become increasingly critical, particularly when the resulting scores are used for ranking, graduation decisions, and instructional interventions[2], [3]. Assessment fairness is not only concerned with the procedures of test administration, but also with the extent to which the resulting scores accurately and proportionally represent students' abilities, free from bias introduced by the characteristics of the measurement instrument[4], [5].

One of the primary factors that may influence score fairness is the scoring approach employed in test analysis. In educational evaluation practice, two major approaches are widely utilized, namely CTT and IRT [6]. As a conventional approach, CTT calculates scores based on the total number of correct responses under the assumption that each test item carries equal weight. Consequently, this approach tends to overlook variations in item difficulty, such that students who correctly answer more difficult items do not necessarily receive proportionally higher scores. This condition may lead to distortions in the interpretation of ability and reduce the fairness of the resulting scores[7].

In contrast, IRT was developed to address these limitations by modeling the relationship between student ability and item characteristics. Through parameters such as item difficulty, discrimination, and guessing, IRT enables more precise estimation of ability and assigns differential weights to items based on their

informational value. Therefore, theoretically, IRT holds greater potential to produce fairer scores, as it accounts for both item complexity and variability in student ability[8], [9], [10].

Although numerous studies have compared CTT and IRT, most prior research has primarily focused on technical aspects such as parameter estimation accuracy and model superiority, without explicitly examining the implications of these approaches for score fairness experienced by students[7]. Furthermore, existing studies are generally conducted in large-scale assessment contexts, offering limited insight into how these approaches function in classroom assessment, which directly affects students' learning experiences[11].

These limitations become more salient when considered in relation to the characteristics of Fiqh as a subject in Madrasah Aliyah. Fiqh requires not only memorization but also deep conceptual understanding and legal reasoning, resulting in a wide range of item difficulty levels. In such a context, a scoring approach that is insensitive to differences in item difficulty may produce scores that do not accurately reflect students' true abilities[12]. Consequently, students who are capable of solving complex problems may receive scores comparable to those who can only answer basic items, thereby raising concerns regarding assessment fairness[13].

Based on the foregoing discussion, there is a pressing need to examine how differences in scoring approaches between CTT and IRT affect the fairness of learning outcome scores, particularly in the context of Fiqh instruction in madrasahs. This study aims to compare the score estimates generated by both approaches and to evaluate the extent to which each approach fairly represents students' abilities by taking into account item difficulty. Accordingly, this research not only contributes to the methodological discourse in educational assessment but also offers practical implications for the development of more equitable and representative classroom assessment systems.

METHODS

Study area

This study is classified as an evaluative research employing a quantitative approach. The primary objective is to examine differences in students' scoring outcomes between the CTT and IRT approaches, particularly in the context of item difficulty analysis and the fairness of learning outcome scores in the subject of Fiqh. The participants of this study comprised all Grade XII students at Madrasah Aliyah Negeri (MAN) 1 Surakarta who undertook the Fiqh assessment in 2023. A total of 387 students were involved, and the data analyzed consisted of students' responses to a multiple-choice test instrument in Fiqh.

Procedures

The analytical procedures included: 1) Collecting students' examination data, 2) Calculating student scores using both CTT and IRT approaches, 3) Analyzing item difficulty using CTT (p-values) and IRT (b-parameters), 4) Testing differences in scores between CTT and IRT using a paired sample t-test, and 5) Interpreting the results to assess the extent to which differences in scoring approaches affect fairness in measurement.

Data analysis

Data analysis was conducted using two main approaches: 1) CTT Analysis, which estimates item difficulty based on the proportion of students who answered correctly (p-value) and assigns total scores based on the number of correct responses (fixed scoring)[14], [15], 2) IRT Analysis, applying the one-parameter logistic model (1PL/Rasch) to estimate item difficulty (b-parameter) and item information, and to generate student scores proportionally by accounting for item characteristics[16], [17], 3) Statistical Test of Score Differences: To determine the significance of differences between CTT- and IRT-based scores, a paired sample t-test was employed[18]. This test was conducted on each student's scores derived from both scoring approaches to examine statistically significant differences in mean scores. CTT data processing was carried out using Microsoft Excel and SPSS, while IRT analysis was performed using Winsteps software[19]. The paired sample t-test was conducted using SPSS version 25.

RESULTS AND DISCUSSION

Within CTT framework, item difficulty is determined based on the proportion of students who answer an item correctly (referred to as the p-value). The higher the p-value, the easier the item. Based on the item difficulty analysis using the CTT approach for 40 test items, the distribution of item difficulty was found to be varied and could be identified in detail at the individual item level.

Items classified as very difficult (index 0.00–0.25)[20], [21] totaled 10 items, representing 25% of the test, specifically items 2, 3, 4, 5, 6, 19, 23, 27, 31, and 38. These items exhibited very low success rates,

indicating high cognitive demands or potential issues related to item wording or misalignment between content and students' ability levels. The difficult category (index 0.26–0.50) constituted the largest group, comprising 13 items (33%), namely items 1, 8, 20, 21, 22, 26, 28, 29, 30, 32, 35, 36, and 39. The dominance of difficult items suggests that the instrument tends to measure moderate to higher-order abilities.

The moderate (or easy) category (index 0.51–0.75) [20], [21] included 11 items (28%), specifically items 10, 13, 16, 17, 18, 24, 25, 33, 34, 37, and 40. These items were relatively accessible to students with average ability levels. Meanwhile, the very easy category (index 0.76–1.00) [20], [21] consisted of 6 items (15%), namely items 7, 9, 11, 12, 14, and 15. Very easy items function to reinforce students' confidence and serve as a balancing component within the test. Overall, this distribution indicates that the instrument encompasses a reasonably comprehensive range of difficulty levels. However, the proportion of very difficult and difficult items still requires further evaluation in order to achieve a more balanced and optimal measurement structure (see Table 1).

Table 1. Item Difficulty Level Based on Classical Test Theory (CTT)

Range of Proportion (p)	Category	Frequency	Percentage
0.00 – 0.25	Very Difficult	10	25%
0.26 – 0.50	Difficult	13	33%
0.51 – 0.75	Easy	11	28%
0.76 – 1.00	Very Easy	6	15%
		40	100%

Based on the item difficulty analysis using the IRT approach, referring to the difficulty parameter (b), the distribution of the 40 test items exhibited a wide range of difficulty levels, identifiable at the individual item level. In the very difficult category ($b > +1.5$), there was 1 item (3%), specifically item 27. This item could only be answered correctly by students with very high ability, representing measurement at the upper end of the ability scale [22], [23].

The difficult category ($+0.5 < b \leq +1.5$) [24], [25] was the most dominant, consisting of 25 items (63%), namely items 2, 3, 8, 19, 20, 21, 22, 23, 28, 29, 31, 32, 35, 38, and 39, as well as items with b-values approaching the upper limit such as 1.06, 0.98, 0.72, and 1.22. The predominance of this category indicates that the instrument focuses primarily on measuring moderate to higher-level abilities. The moderate category ($-0.5 \leq b \leq +0.5$) [24], [25] included 10 items (25%), specifically items 1, 10, 13, 16, 17, 24, 25, 26, 30, 36, and 40. These items align with the average ability of the students and are theoretically the most informative for ability estimation. The very easy category ($b < -0.5$) [24], [25] consisted of 4 items (10%), namely items 4, 5, 6, and 14. These items could be correctly answered by nearly all students and serve to measure basic-level ability [22], [23]. Overall, this distribution indicates that the instrument encompasses a full range of item difficulty; however, it is still dominated by difficult items, highlighting the need for balancing to achieve more evenly distributed measurement of ability (see Table 2).

Table 2. Item Difficulty Level Based on Item Response Theory (IRT)

Range of b-Parameter	Category	Frequency	Percentage
$b > +1,5$	Very Difficult	1	3%
$+0,5 < b \leq +1,5$	Difficult	25	63%
$-0,5 \leq b \leq +0,5$	Easy	10	25%
$b < -0,5$	Very Easy	4	10%
		40	100%

Based on the comparison of item difficulty between CTT and IRT approaches, a distinction can be made between items categorized consistently and inconsistently across the two methods. This analysis is essential for evaluating the stability of item characteristics across measurement models.

Items categorized as consistent are those assigned to the same difficulty category under both CTT and IRT. This group includes items 1, 8, 14, 16, 18, 20, 21, 22, 26, 27, 28, 29, 30, 32, 33, 34, 35, 36, 37, and 39. For instance, items 1 and 8 were classified as difficult under both approaches, while items 16, 18, 33, 34, and 37 consistently fell within the easy category. Item 27 also demonstrated strong consistency, being categorized as very difficult in both CTT and IRT. Such consistency indicates that these items possess relatively stable difficulty characteristics that are minimally affected by differences in model assumptions.

Conversely, several items exhibited inconsistent categorization between CTT and IRT. These items include 2, 3, 4, 5, 6, 7, 9, 10, 11, 12, 13, 15, 17, 19, 23, 24, 25, 31, 38, and 40. The most pronounced inconsistencies were observed in items 4, 5, and 6, which were classified as very difficult under CTT but very easy under IRT. This discrepancy highlights that the proportion of correct responses (CTT) and the

item's position on the latent ability scale (IRT) can yield substantially different interpretations. Additionally, several items, such as 2, 3, 19, 23, 31, and 38, shifted from very difficult in CTT to difficult in IRT, indicating differences in model sensitivity to student ability (see Table 3).

Table 3. Difference of Difficulty Level between CTT and IRT

Item Number	CTT Difficulty Level	Description	IRT Difficulty Level	Description
1	0,50	Difficult	0,05	Difficult
2	0,16	Very Difficult	1,88	Difficult
3	0,17	Very Difficult	1,79	Difficult
4	0,00	Very Difficult	-7,2	Very Easy
5	0,00	Very Difficult	-7,2	Very Easy
6	0,00	Very Difficult	-5,99	Very Easy
7	0,80	Very Easy	-1,45	Easy
8	0,37	Difficult	0,64	Difficult
9	0,77	Very Easy	-1,22	Easy
10	0,56	Easy	-0,19	Difficult
11	0,82	Very Easy	-1,64	Easy
12	0,89	Very Easy	-2,14	Easy
13	0,56	Easy	-0,22	Difficult
14	0,95	Very Easy	-2,96	Very Easy
15	0,77	Very Easy	-1,28	Easy
16	0,63	Easy	-0,48	Easy
17	0,53	Easy	-0,06	Difficult
18	0,69	Easy	-0,84	Easy
19	0,19	Very Difficult	1,66	Difficult
20	0,29	Difficult	1,06	Difficult
21	0,31	Difficult	0,98	Difficult
22	0,35	Difficult	0,72	Difficult
23	0,21	Very Difficult	1,5	Difficult
24	0,54	Easy	-0,14	Difficult
25	0,54	Easy	-0,11	Difficult
26	0,45	Difficult	0,26	Difficult
27	0,10	Very Difficult	2,46	Very Difficult
28	0,30	Difficult	1,01	Difficult
29	0,26	Difficult	1,22	Difficult
30	0,45	Difficult	0,28	Difficult
31	0,20	Very Difficult	1,6	Difficult
32	0,27	Difficult	1,17	Difficult
33	0,67	Easy	-0,72	Easy
34	0,71	Easy	-0,92	Easy
35	0,41	Difficult	0,52	Difficult
36	0,47	Difficult	0,17	Difficult
37	0,72	Easy	-0,92	Easy
38	0,17	Very Difficult	1,77	Difficult
39	0,38	Difficult	0,62	Difficult
40	0,53	Easy	-0,06	Difficult

Overall, these results confirm that, although some items demonstrate good consistency, the differing paradigms between CTT and IRT can lead to divergent classifications of item difficulty. IRT provides a more varied and individualized classification, allowing for more precise identification of items that are not proportionate to students' abilities. In contrast, CTT relies more heavily on the distribution of participants' performance; therefore, if the majority of students have either low or high ability, the resulting classification may become biased.

Item Score Weights

In the CTT approach, each item is assigned a fixed score weight according to policy or evaluator agreement. As shown in Table 3, items classified as easy or moderate were typically assigned a weight of 1.5, particularly items 1 through 11. Subsequent items (12–40), which appear to require higher-order cognitive skills or are more content-significant, were assigned a weight of 2 per item.

These weights are fixed and do not account for item difficulty relative to students' abilities. Under CTT, all students receive the same score if they answer the same item correctly, regardless of whether the item is easy or difficult. While this simplifies scoring, it does not provide additional credit for more challenging items.

In contrast, IRT assigns variable item weights based on item difficulty and other item characteristics. The IRT approach provides scores that are proportional to the informational value of each item. A student who correctly answers difficult items receives higher scores than a student who only answers easy items correctly.

The highest scores were assigned to difficult items, such as item 2 (3.21), item 3 (3.19), item 38 (3.18), item 19 (3.14), item 31 (3.12), and item 23 (3.09). Conversely, very easy or non-informative items (e.g., items with extreme negative parameter estimates or invalid items) were assigned low weights, such as items 4 and 5 (0.32) and item 6 (0.70). Most items had scores ranging from 2.00 to 3.00, depending on their discrimination and difficulty parameters, with the total score summing to 100 points.

Table 4. Item Score Weights between CTT and IRT

Item Number	Item Score Weight Based on CTT	Item Score Weight Based on IRT
1	1,5	2,63
2	1,5	3,21
3	1,5	3,19
4	1,5	0,32
5	1,5	0,32
6	1,5	0,70
7	1,5	2,15
8	1,5	2,82
9	1,5	2,23
10	1,5	2,55
11	1,5	2,09
12	2	1,93
13	2	2,54
14	2	1,67
15	2	2,21
16	2	2,46
17	2	2,60
18	2	2,35
19	2	3,14
20	2	2,95
21	2	2,93
22	2	2,84
23	2	3,09
24	2	2,57
25	2	2,58
26	2	2,70
27	2	3,40
28	2	2,94
29	2	3,00
30	2	2,70
31	2	3,12
32	2	2,99
33	2	2,39
34	2	2,32

Item Number	Item Score Weight Based on CTT	Item Score Weight Based on IRT
35	2	2,78
36	2	2,67
37	2	2,32
38	2	3,18
39	2	2,81
40	2	2,60
	75	100,00

Score Differences Between CTT and IRT

Assessment of 387 students using CTT approach revealed substantial disparities in score gains and losses. Only a small proportion of students benefited, while the majority experienced score reductions. A total of 301 students (78%) experienced a decrease in scores under the CTT approach, indicating that their scores were lower compared to those obtained using an alternative method, such as IRT. This finding suggests that CTT does not adequately reward students' efforts in answering difficult items and overlooks individual ability variations. Consequently, most students are disadvantaged because CTT assigns equal weight to all items without considering item difficulty or discrimination.

Conversely, only 83 students (21%) benefited from CTT scoring. These students likely answered relatively easy items correctly and received full credit as if they had answered more difficult items. This outcome indicates that CTT may advantage students with average or lower performance who happen to answer simpler items correctly, without accurately reflecting their true abilities. Interestingly, only 3 students (1%) received scores that could be considered fair, showing no meaningful difference between CTT and alternative approaches. This demonstrates that CTT rarely produces neutral or balanced results, and tends to distort scores for the majority of students.

Based on this distribution, it can be concluded that CTT does not provide fair assessment for most students. Given the high proportion of students experiencing score reductions and the low proportion benefiting, the CTT approach requires reevaluation—especially in the context of subjects such as Fiqh, which demand proportional understanding, reasoning, and application of concepts relative to item difficulty. As an alternative, IRT is recommended because it provides more equitable, accurate, and representative assessment outcomes that better reflect students' actual abilities.

Table 5. Score Differences Between CTT and IRT

Participant Number	CTT Score	IRT Score	Difference	Description
200001	79	77	2	Gain
200002	91	90	1	Gain
200003	71	92	-21	Loss
200004	77	90	-13	Loss
200005	62	74	-13	Loss
200006	74	77	-4	Loss
200007	70	84	-14	Loss
200008	82	81	1	Gain
200009	81	86	-5	Loss
200011	69	74	-5	Loss
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...
200391	98	82	16	Gain
200392	93	87	6	Gain
200393	95	87	9	Gain
200394	95	78	17	Gain
200395	90	81	9	Gain
200396	93	88	5	Gain
200397	98	59	39	Gain
200398	90	89	1	Gain
200399	90	87	3	Gain
200400	93	79	14	Gain

Participant Number	CTT Score	IRT Score	Difference	Description
200401	93	37	56	Gain
200402	88	65	23	Gain
190464	90	68	22	Gain

Student Score Gains

Based on Table 6, the highest score gain was achieved by a single student, with a difference of +56 points. The majority of students experienced score gains ranging from +1 to +10 points, indicating that most gains were small to moderate. The most frequent gains occurred at +5 points (6 students), +6 points (7 students), and +3 and +14 points (5 students each). Only one student experienced an extreme gain of +56 points, which can be considered an anomaly or outlier. Overall, the CTT approach does not consistently confer substantial score gains and, in fact, may result in score reductions for the majority of other students.

Table 6. Recapitulation of Student Score Gains Using the CTT Approach

Total Assessment Gains Using the CTT Approach	Frequency	Percent	Valid Percent	Cumulative Percent
1	4	4.8	4.8	4.8
2	4	4.8	4.8	9.6
3	5	6.0	6.0	15.7
4	2	2.4	2.4	18.1
5	6	7.2	7.2	25.3
6	7	8.4	8.4	33.7
7	3	3.6	3.6	37.3
8	3	3.6	3.6	41.0
9	5	6.0	6.0	47.0
10	2	2.4	2.4	49.4
11	1	1.2	1.2	50.6
14	5	6.0	6.0	56.6
15	1	1.2	1.2	57.8
16	4	4.8	4.8	62.7
17	4	4.8	4.8	67.5
19	2	2.4	2.4	69.9
20	4	4.8	4.8	74.7
21	2	2.4	2.4	77.1
22	3	3.6	3.6	80.7
23	2	2.4	2.4	83.1
24	1	1.2	1.2	84.3
25	1	1.2	1.2	85.5
27	2	2.4	2.4	88.0
28	1	1.2	1.2	89.2
30	1	1.2	1.2	90.4
32	1	1.2	1.2	91.6
36	2	2.4	2.4	94.0
39	1	1.2	1.2	95.2
47	1	1.2	1.2	96.4
49	1	1.2	1.2	97.6
56	2	2.4	2.4	100.0
Total	83	100.0	100.0	

Student Score Losses

The data indicate the amount of lost or unearned points experienced by students due to the use of the CTT approach in assessing Fiqh test items. All recorded scores were negative, meaning that students' scores under CTT were lower than those obtained using an alternative approach, such as IRT. A total of 301 students experienced score reductions of varying degrees, ranging from a loss of 1 point to 62 points.

The range of score loss was wide, from -1 to -62 points. The highest loss was experienced by one student, with a reduction of -62 points, representing the most extreme case in the dataset. The majority of students lost between -10 and -30 points, reflecting moderate to severe score reductions. The most frequently observed losses were: -25 points for 15 students (5%), -16 points for 17 students (5.6%), -30 points for 14 students (4.7%), and -17, -22, and -27 points each for 12 students (4%). Overall, more than half of the students (approximately 55%) experienced losses ranging from -10 to -30 points, indicating that CTT substantially lowered their scores compared to other methods.

Based on this distribution, it can be concluded that the majority of students experienced significant score losses when assessed using the CTT approach. These losses ranged from minor (-1 point) to extreme (-62 points). Over 50% of students lost between 10 and 30 points, demonstrating that CTT tends to disadvantage students in terms of accurate ability measurement, particularly when compared to more adaptive methods such as IRT. In the context of academic assessment—especially in subjects like Fiqh, which demand analytical reasoning and deep understanding—IRT is more appropriate, as it provides a fairer evaluation by accounting for item difficulty and students' response characteristics.

Table 7. Recapitulation of Scoring Losses Using the CTT Approach

Total Assessment Losses Using the CTT Approach	Frequency	Percent	Valid Percent	Cumulative Percent
-62	1	.3	.3	.3
-56	2	.7	.7	1.0
-53	1	.3	.3	1.3
-52	1	.3	.3	1.7
-49	1	.3	.3	2.0
-48	1	.3	.3	2.3
-47	1	.3	.3	2.7
-44	4	1.3	1.3	4.0
-43	1	.3	.3	4.3
-42	1	.3	.3	4.7
-41	1	.3	.3	5.0
-40	1	.3	.3	5.3
-39	1	.3	.3	5.6
-38	2	.7	.7	6.3
-37	4	1.3	1.3	7.6
-36	5	1.7	1.7	9.3
-35	5	1.7	1.7	11.0
-34	3	1.0	1.0	12.0
-33	5	1.7	1.7	13.6
-32	3	1.0	1.0	14.6
-31	5	1.7	1.7	16.3
-30	14	4.7	4.7	20.9
-29	4	1.3	1.3	22.3
-28	7	2.3	2.3	24.6
-27	12	4.0	4.0	28.6
-26	8	2.7	2.7	31.2
-25	15	5.0	5.0	36.2
-24	7	2.3	2.3	38.5
-23	9	3.0	3.0	41.5
-22	12	4.0	4.0	45.5
-21	7	2.3	2.3	47.8
-20	9	3.0	3.0	50.8
-19	5	1.7	1.7	52.5
-18	8	2.7	2.7	55.1
-17	12	4.0	4.0	59.1
-16	17	5.6	5.6	64.8
-15	9	3.0	3.0	67.8

Total Assessment Losses Using the CTT Approach	Frequency	Percent	Valid Percent	Cumulative Percent
-14	10	3.3	3.3	71.1
-13	11	3.7	3.7	74.8
-12	6	2.0	2.0	76.7
-11	9	3.0	3.0	79.7
-10	9	3.0	3.0	82.7
-9	3	1.0	1.0	83.7
-8	8	2.7	2.7	86.4
-7	8	2.7	2.7	89.0
-6	5	1.7	1.7	90.7
-5	6	2.0	2.0	92.7
-4	6	2.0	2.0	94.7
-3	3	1.0	1.0	95.7
-2	8	2.7	2.7	98.3
-1	5	1.7	1.7	100.0
Total	301	100.0	100.0	

Differences Between CTT and IRT Scores

The mean score under the IRT approach (71.61) was higher than that under CTT (59.18). This indicates that, when assessment is conducted using the IRT model, students tend to achieve higher scores because IRT weighting adjusts for item difficulty. The standard deviation for CTT was 17.06, whereas for IRT it was smaller, at 13.87. This suggests that CTT scores are more widely dispersed, while IRT scores are more concentrated around the mean. These findings demonstrate that IRT tends to provide more consistent and stable scoring, as it accounts for both item parameters and individual student ability.

Table 8. Description of CTT- and IRT-Based Scoring

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	CTT	59.1835	387	17.06328	.86738
Pair 1	IRT	71.6124	387	13.86705	.70490

Standard Error of the Mean and Paired Sample t-Test Results

The Standard Error of the Mean (SEM) for IRT was smaller (0.705) compared to CTT (0.867). SEM measures the precision of the mean estimate; the smaller the SEM, the more precise the estimation. This indicates that the mean score obtained using IRT is more representative of the population as a whole than the mean score from CTT. Based on these descriptive statistics, it can be inferred that IRT provides a higher and more consistent mean score compared to the CTT approach. Score variability in IRT is lower, reflecting a more stable and accurate scoring system in assessing student ability. The precision of the mean calculation in IRT is superior, as evidenced by the smaller SEM.

Furthermore, the paired sample t-test revealed a t-value of -13.298 with a significance level (p-value) of 0.000, indicating that the difference between CTT and IRT scores is statistically highly significant ($p < 0.05$). This finding demonstrates that the IRT approach consistently produces higher scores compared to CTT. This outcome can be explained by the fact that IRT assigns different weights to items based on their difficulty, allowing students who correctly answer more challenging items to earn higher scores. IRT scoring considers the interaction between student ability and item characteristics, rather than relying solely on the proportion of correct responses as in CTT. In contrast, CTT tends to equalize item weights, disregarding item difficulty or discrimination, so students who answer many easy items can receive the same score as those who answer difficult items—an approach that does not fairly reflect true ability.

Table 9. Results of the Differences in Scoring between CTT and IRT

		Paired Differences	Paired Differences	Paired Differences	Paired Differences	Paired Differences	t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference	95% Confidence Interval of the Difference	t	df	Sig. (2-tailed)
		Mean	Std.	Std. Error	Lower	Upper	t	df	Sig. (2-

		Paired Differences	Paired Differences	Paired Differences	Paired Differences	Paired Differences	t	df	Sig. (2-tailed)
			Deviation	Mean					tailed)
Pair 1	CTT IRT	-1.24289E1	18.38681	.93465	-14.26659	-10.59129	-13.298	386	.000

Results and Discussion should be written as a series of connecting sentences, however, for manuscript with long discussion should be divided into subtitles. Results should be clear and concise.

Discussion

This study demonstrates significant differences between CTT and IRT in analyzing item difficulty, scoring systems, and the fairness of student learning outcomes.

1. Item Difficulty: Conceptual Differences and Classification Results

CTT evaluates item difficulty based on the proportion of students answering correctly (p-value), whereas IRT employs the b parameter, which reflects the item's position along the latent ability scale. Under CTT, most items were classified as easy (60%), while IRT categorized the majority of items as difficult (63%). These differences reflect fundamental paradigm distinctions: CTT's difficulty estimates are sample-dependent, representing student performance rather than intrinsic item challenge[26], whereas IRT estimates item parameters independent of the sample, allowing for precise classification even when correct responses are scarce[8]. Theoretically, CTT is appropriate for small, homogeneous classroom tests but less suitable for large-scale or ability-based assessments[27]. IRT, in contrast, provides stable, precise, and unbiased information, supporting a shift toward latent trait-based evaluation—particularly relevant in subjects such as Fiqh that demand conceptual understanding and higher-order thinking[28].

2. Item Scoring: Fixed vs. Adaptive

In CTT, item scores are fixed, assigning equal weight to all correct responses regardless of difficulty. While simple and practical, this approach ignores critical item information[29]. Conversely, IRT applies adaptive scoring, weighting items according to difficulty (b), discrimination (a), and guessing (c, in 3PL models)[9], [17], [30]. Correct responses to difficult, highly discriminating items yield higher scores, whereas very easy or uninformative items are weighted lower[31]. Adaptive scoring rewards students proportionally for their performance, producing fairer and more representative assessments of individual ability. Theoretical implications indicate that IRT is psychometrically more valid for measuring actual ability and differentiating students accurately[32]. This aligns with modern measurement theory emphasizing measurement invariance, whereby scores should reflect ability rather than extraneous factors[33]. Practically, adaptive scoring incentivizes students to attempt challenging items, promoting higher-order thinking and increasing fairness in assessment[34].

3. Student Score Distribution: Fairness and Inequality

Analysis showed that 301 students (78%) experienced score losses under CTT, ranging from -1 to -62 points, whereas only 83 students (21%) gained minor benefits (1–10 points), and merely 1% had balanced scores. These findings highlight that CTT disproportionately disadvantages students capable of solving difficult items[35]. The uniform weighting in CTT ignores item difficulty and student ability variance, resulting in aggregated and non-proportional scores[26], [29]. High-ability students answering challenging items receive inadequate recognition, while low-ability students answering only easy items can still achieve full scores[36].

In contrast, IRT maps the interaction between student ability and item parameters individually, producing more precise and stable score distributions[31]. This enhances fairness, particularly for high-performing students who are disadvantaged by fixed-weight scoring. The findings suggest that both validity and fairness depend not only on item construction but also on the scoring model[37]. For subjects like Fiqh that require conceptual and analytical reasoning, IRT better reflects authentic student ability.

4. Statistical Comparisons: Significance of Differences

The mean IRT score (71.61) was significantly higher than the mean CTT score (59.18), with a paired sample t-test yielding $t = -13.298$, $p = 0.000$. IRT scores exhibited lower standard deviation and Standard Error of the Mean, indicating more stable and precise estimates of student ability[28]. These results demonstrate that IRT scoring reflects variations in student performance more accurately and proportionally, accounting for item difficulty and characteristics[26], [31].

CTT, being sample- and test-dependent, produces scores sensitive to population composition and item selection, whereas IRT provides invariant scores consistent across samples and test forms [16], [38].

Consequently, the observed differences are not merely statistical but reflect IRT's theoretical advantage in validity and reliability[39].

These findings underscore the importance of selecting a scoring model aligned with assessment goals. CTT's aggregate and fixed-weight scoring risks unfair or inaccurate educational decisions, particularly for medium- to high-ability students[40]. IRT, by contrast, allows for precise, fair, and proportionally valid measurement, supporting its use in classroom and large-scale assessments, adaptive testing, and diagnostic evaluation[28]. For subjects like Fiqh, where conceptual understanding and analytical reasoning are crucial, IRT provides a more accurate reflection of student competence. Educators and test developers are encouraged to adopt IRT-based assessment to enhance fairness, validity, and instructional insight.

CONCLUSION

This study revealed significant differences between CTT and IRT approaches in analyzing item quality, scoring systems, and the fairness of student learning outcomes. Regarding item difficulty, CTT relies on the proportion of students answering correctly (p-value), whereas IRT uses the b parameter, which accounts for an item's position along the students' ability spectrum. The results indicate that under CTT, the majority of items were classified as easy (60%), whereas IRT categorized most items as difficult (38%) or very difficult (10%). Notable anomalies were observed in items 4, 5, and 6, which were rated as very difficult by CTT but very easy under IRT, highlighting CTT's limitation in detecting item misalignment.

In terms of scoring, CTT applies fixed weights to all items regardless of difficulty. In contrast, IRT employs adaptive weighting, assigning higher scores to difficult items and lower scores to very easy ones. This approach allows IRT to more fairly reward students for correctly answering challenging items. Regarding score fairness, 301 students (78%) experienced score losses under CTT, while only 83 students (21%) gained minor benefits. Only 1% of students received balanced scores. These findings indicate that CTT tends to disadvantage the majority of students, particularly those capable of solving difficult items.

Statistically, IRT produced a higher mean score with smaller standard deviation and standard error, demonstrating more stable and representative score estimation. Overall, these results confirm that IRT is superior in proportionally measuring student ability and should be considered for implementation in fair and accurate educational assessment systems.

REFERENCES

- [1] L. H. Schellekens, H. G. Bok, L. H. De Jong, M. F. Van der Schaaf, W. D. Kremer, and C. P. Van der Vleuten, "A scoping review on the notions of Assessment as Learning (AaL), Assessment for Learning (AfL), and Assessment of Learning (AoL)," *Studies in Educational Evaluation*, vol. 71, p. 101094, 2021.
- [2] A. Baniasadi, K. Salehi, E. Khodaie, K. Bagheri Noaparast, and B. Izanloo, "Fairness in classroom assessment: A systematic review," *The Asia-Pacific Education Researcher*, vol. 32, pp. 91-109, 2023.
- [3] R. D. Tierney, "Fairness in educational testing and assessment," *Routledge encyclopedia of education*. Routledge, vol. 10, 2022.
- [4] L. A. Suzuki, "Integrating cultural considerations and fairness in psychological assessment," *Fairness in Educational and Psychological Testing: Examining Theoretical, Research, Practice, and Policy Implications of the 2014 Standards*, p. 377, 2022.
- [5] R. Zwick, "Fairness in Educational Measurement," *Educational Measurement*, p. 383, 2025.
- [6] X. Tang, B. D. Schalet, J. D. Peipert, and D. Cella, "Does scoring method impact estimation of significant individual changes assessed by patient-reported outcome measures? Comparing classical test theory versus item response theory," *Value in Health*, vol. 26, pp. 1518-1524, 2023.
- [7] I. M. Fitriyah and H. Retnawati, "Analysis of the distractor of the multiple-choice test using classical test theory (CTT) and item response theory (IRT)," 2023.
- [8] A. Gyamfi and R. Acquaye, "Parameters and models of item response theory (IRT): A review of literature," *Acta Educationis Generalis*, vol. 13, pp. 68-78, 2023.
- [9] V. Mutiawani, A. M. Athaya, K. Saputra, and M. Subianto, "Implementing item response theory (irt) method in quiz assessment system," *TEM Journal*, vol. 11, 2022.
- [10] T. A. Ackerman and Y. Ma, "Examining differential item functioning from a multidimensional IRT perspective," *psychometrika*, vol. 89, pp. 4-41, 2024.
- [11] A. Mutholib, "Analisis Butir Soal Bahasa Arab Ujian Akhir Madrasah Berstandar Nasional (UAMBN) Bagi Siswa Madrasah Ibtidaiyah (MI) Tahun Pelajaran 2012-2013," *Arabia*, vol. 5, 2016.
- [12] M. S. Nawir, S. Nur, M. Abubakar, and A. Asmawati, "the analysis of item difficulty and discrimination of english summative teacher made test (tmt) at mas madani pao-pao gowa," *English Language, Linguistics, and Culture International Journal*, vol. 3, pp. 89-100, 2023.
- [13] D. Vlachopoulos and A. Makri, "A systematic literature review on authentic assessment in higher education: Best practices for the development of 21st century skills, and policy considerations," *Studies in Educational Evaluation*, vol. 83, p. 101425, 2024.

- [14] M. A. Ayanwale, J. Chere-Masopha, and M. C. Morena, "The classical test or item response measurement theory," 2022.
- [15] W. Vincent and S. K. S. Shanmugam, "The role of classical test theory to determine the quality of classroom teaching test items," *Pedagogia: Jurnal Pendidikan*, vol. 9, pp. 5-34, 2020.
- [16] G. Hamdu, F. Fuadi, A. Yulianto, and Y. Akhirani, "Items quality analysis using rasch model to measure elementary school students' critical thinking skill on stem learning," *JPI (Jurnal Pendidikan Indonesia)*, vol. 9, pp. 61-74, 2020.
- [17] W. C. Lee, S. Y. Kim, J. Choi, and Y. Kang, "IRT approaches to modeling scores on mixed-format tests," *Journal of Educational Measurement*, vol. 57, pp. 230-254, 2020.
- [18] J. Subando, *Teknik Analisis Data Kuantitatif Teori dan Aplikasi dengan SPSS: Penerbit Lakeisha*, 2021.
- [19] H. Untary, Risdianto, E., & Kusen, *Analisis Data Penelitian dengan Model rach dan Winstep*. Jakarta: Halaman Moeka Publishing., 2020.
- [20] J. M. Fiska, Y. Hidayati, N. Qomaria, and W. P. Hadi, "Analisis butir soal ulangan harian IPA menggunakan software Anates pada pendekatan teori tes klasik," *Natural Science Education Research*, vol. 4, pp. 65-76, 2021.
- [21] U. Nasyrah, B. Bustanur, and Z. Zulhaini, "Analisis Butir Soal Pilihan Ganda Menggunakan Teori Pengukuran Klasik (Pada Penilaian Akhir Tahun (PAT) Mata Pelajaran PAI dan BP Kelas VIII di SMPN 2 Kuantan Mudik)," *Al-Hikmah (Jurnal Pendidikan dan Pendidikan Agama Islam)*, vol. 6, pp. 56-66, 2024.
- [22] C. Z. L. Parisu, W. Ekadayanti, L. Sisi, and A. Juwairiyah, "Analisis Butir Soal Pengetahuan Dasar Matematika Menggunakan Pendekatan Rasch," *Science Tech: Jurnal Ilmu Pengetahuan dan Teknologi*, vol. 10, pp. 36-45, 2024.
- [23] J. Subando, M. Saleh, and N. Ngatimin, "The Analysis of Question Item on Tarikh Subject with Rasch Model Theory," *Journal of Research and Educational Research Evaluation*, vol. 11, pp. 166-177, 2022.
- [24] R. N. Amelia, "Identifikasi item fit dan person fit dalam pengukuran hasil belajar kimia," *Jurnal Ilmiah WUNY*, vol. 3, pp. 13-26, 2021.
- [25] A. Yusuf, M. Román-González, N. A. Atan, S. K. Behera, and N. M. Noor, "Development of a Quantum Literacy Test for K-12 Students: An Extension of the Computational Thinking Framework," *Education Sciences*, vol. 16, p. 31, 2025.
- [26] R. K. Hambleton, H. Swaminathan, and H. J. Rogers, *Fundamentals of item response theory vol. 2*: Sage, 1991.
- [27] A. Faria and G. L. Miranda, "Operationalising CTT and IRT in Spreadsheets: A Methodological Demonstration for Classroom Assessment," *Analytics*, vol. 5, p. 12, 2026.
- [28] R. E. Eghan, E. Osei-Sarpong, G. E. Awashie, R. N. Borkor, E. Yaokumah, and A. A. N'ganomah, "Item Response Theory for trait assessment in randomized item pool for computer based test," *Scientific African*, p. e03226, 2026.
- [29] A. J. Nitko and S. M. Brookhart, *Educational Assessment of Students*: Pearson Education, 2014.
- [30] J. Jumini and H. Retnawati, "Estimating item parameters and student abilities: An IRT 2PL analysis of mathematics examination," *Al-Ishlah: Jurnal Pendidikan*, vol. 14, pp. 385-398, 2022.
- [31] S. E. Embretson and S. P. Reise, *Item response theory for psychologists*: Psychology Press, 2013.
- [32] E. Rachmawati and S. Siswanto, "Developing an instrument to measure students' anti-corruption character using modern test theory," *Jurnal Penelitian dan Evaluasi Pendidikan*, vol. 29, pp. 26-42, 2025.
- [33] B. Ma, M. Krötz, V. Deutscher, and E. Winther, "Same Same yet Different: Equivalence of the Multidimensional Assessment of Vocational Competence Across Test Formats," *International Journal of Training and Development*, vol. 29, pp. 433-448, 2025.
- [34] H. Sulistyanto, S. Anif, S. Narimo, A. Sutopo, M. I. Haq, and G. A. Nasir, "Education application testing perspective to empower students' higher order thinking skills related to the concept of adaptive learning media," *Indonesian Journal on Learning and Advanced Education (IJOLAE)*, pp. 257-271, 2022.
- [35] S. Hadi, H. Retnawati, S. Munadi, E. Apino, and N. F. Wulandari, "The difficulties of high school students in solving higher-order thinking skills problems," *Problems of Education in the 21st Century*, vol. 76, p. 520, 2018.
- [36] T. Priyani and B. Sugiharto, "Analysis of biology midterm exam items using a comparison of the classical theory test and the Rasch model," *JPBI (Jurnal Pendidikan Biologi Indonesia)*, vol. 10, pp. 939-958, 2024.
- [37] S. Lane, "Validity, fairness, and technology-based assessment," in *Advancing natural language processing in educational assessment*, ed: Routledge, 2023, pp. 127-141.
- [38] E. Evita, E. D. K. Sari, E. Susieni, and B. Tamam, "analysis of item quality in islamic religious education subjects with rasch modeling," *Conciencia*, vol. 23, pp. 147-156, 2023.
- [39] C. D. Nye, S.-H. Joo, B. Zhang, and S. Stark, "Advancing and evaluating IRT model data fit indices in organizational research," *Organizational Research Methods*, vol. 23, pp. 457-486, 2020.
- [40] A. Robitzsch and O. Lüdtke, "Some thoughts on analytical choices in the scaling model for test scores in international large-scale assessment studies," *Measurement Instruments for the Social Sciences*, vol. 4, p. 9, 2022.