ISSN: 2790-4407

Journal of Computer Adaptive Testing in Africa

2025; 4; 20-53

Enhancing Test Quality in Lesotho Basic Education Through Classical Test and Item Response Theory Analyses

[a]Lefa Clement Thamae*

[a]Department of Educational Foundations, Faculty of Education, National University of Lesotho, Roma, Lesotho

Abstract

This study developed and psychometrically evaluated a Grade 6 mathematics test for Lesotho's Cambridge curriculum using Classical Test Theory (CTT) and Item Response Theory (IRT). A quantitative psychometric design involving 200 learners from Cambridge international schools was employed. Data were analysed using jMetrik through CTT indices and 1PL and 2PL IRT models. Results revealed substantial psychometric weaknesses, including extreme item difficulty variation, weak and negative discrimination indices, low reliability coefficients ($\alpha = 0.1975$), and several misfitting items. The 2PL model demonstrated superior fit over the 1PL model based on deviance, AIC, and BIC statistics. While CTT identified global weaknesses in reliability and item discrimination, IRT provided richer diagnostic evidence regarding item functioning and conditional measurement precision across ability levels. The study concludes that integrating CTT and IRT strengthens assessment validity, reliability, fairness, and evidence-based test development in Lesotho basic education. The study recommends rigorous item piloting, routine use of CTT and IRT analyses, and enhanced psychometric training for teachers and assessment practitioners in Lesotho.

Keywords: Test quality, Lesotho basic education, classical test theory (CTT), item response theory (IRT), integration

National University of Lesotho, Department Educational Foundations, Faculty of Education, e-mail: makauclement@gmail.com

To cite this article:

Thamae, L. C. (2025). Enhancing Test Quality in Lesotho Basic Education Through Classical Test and Item Response Theory Analyses. *Journal of Computer Adaptive Testing in Africa, 4,* 20-53. https://doi.org/10.71291/pdrgvp19

Received: 3 September 2025
Accepted: 20 December 2025
Published: 23 December 2025

INTRODUCTION

Educational assessment plays a central role in monitoring learner achievement, informing instructional practices, and supporting educational decision-making. The quality of any educational assessment depends largely on the psychometric soundness of its test items, including their reliability, validity, difficulty, discrimination, and fairness (Geleta, Fisseha, & Zenebe, 2024). In Lesotho, basic education assessment practices remain predominantly grounded in Classical Test Theory (CTT), where teachers and assessment bodies commonly rely on total scores, percentages, item difficulty indices, discrimination indices, and reliability coefficients to evaluate assessment quality. Although CTT has contributed significantly to educational measurement, it possesses limitations such as sample dependency, test dependency, and the inability to estimate conditional measurement precision across varying ability levels. In contrast, Item Response Theory (IRT) provides stronger psychometric evidence by estimating item parameters and learner ability on a common latent trait scale while also evaluating measurement precision through item and test information functions. However, the application of IRT within Lesotho basic education remains limited, especially in the context of Lesotho.

In the context of Lesotho's basic education, assessment practices remain predominantly grounded in CTT. At the classroom level, teachers commonly evaluate assessments using conventional indices such as total scores, percentages, *p*-value, discrimination indices, and reliability coefficients derived from CTT procedures. Assessments developed by teachers are commonly evaluated using basic indices such as total scores, percentages, item difficulty, and reliability coefficients derived from CTT approaches (Ayanwale, Chere-Masopha, & Morena, 2022). Similarly, the Examinations Council of Lesotho (ECoL), which administers national assessments and examinations, primarily relies on classical item analysis procedures for evaluating examination quality (Ayanwale et al., 2022). Also, teacher training institutions in Lesotho appear to emphasise traditional measurement approaches, with limited exposure to modern psychometric frameworks such as IRT. Consequently, assessment practices within Lesotho continue to focus largely on observed scores rather than latent trait estimation and conditional measurement precision (Ayanwale et al., 2022).

Although CTT has contributed significantly to educational assessment, it possesses several limitations in contemporary psychometric evaluation. Item statistics and learner scores are sample and test dependent, meaning that item difficulty, discrimination, and observed scores may vary across different groups and test forms (Ayanwale et al., 2022; Magno, 2009). In addition, CTT

assumes equal measurement precision across all ability levels by treating reliability as a single overall estimate (DeMars, 2010). Consequently, CTT provides limited evidence regarding item functioning and measurement precision across the latent trait continuum (Cook & Pitoniak, 2025; Rusch et al., 2017). In contrast, IRT estimates item difficulty and discrimination while locating both learners and items on a common latent trait scale (Chen et al., 2021; Yang & Kao, 2014). Furthermore, IRT evaluates measurement precision across ability levels through item and test information functions, thereby providing more detailed evidence regarding score reliability and item functioning (Mair, 2018). For this reason, many contemporary assessment systems increasingly integrate both CTT and IRT to achieve more comprehensive psychometric evaluation (Mitee, 2019; Rezae et al., 2018; Siregar & Panjaitan, 2021).

Despite the advantages of IRT, its application within Lesotho basic education remains extremely limited. Most classroom assessments, national examinations, and teacher-training programmes continue to rely almost exclusively on CTT-based methods (Ayanwale et al., 2022). This creates a significant gap in assessment practice because educational decisions may be based on tests whose psychometric properties are not fully evaluated using modern measurement approaches. Furthermore, little empirical research in Lesotho has systematically compared CTT and IRT within the same educational assessment context.

LITERATURE REVIEW

Recent psychometric research increasingly supports the integration of Classical Test Theory (CTT) and Item Response Theory (IRT) in strengthening educational assessment quality (Hu et al., 2021). For instance, Setiawati et al. (2023) investigated the comparability of CTT and IRT item parameters using the Differential Aptitude Test. Their findings revealed strong relationships between item difficulty and discrimination estimates generated by the two frameworks, particularly under the 1PL and 2PL models. The study concluded that CTT and IRT provide complementary psychometric evidence and can jointly improve the evaluation of assessment quality. These findings suggest that combining traditional and modern psychometric approaches may produce more comprehensive information regarding item functioning and learner performance.

Likewise, Santoso et al. (2024) validated a light phenomena conceptual assessment using both CTT and IRT analyses. The study demonstrated that the integration of the two frameworks strengthened

evidence regarding validity, item functioning, and learner ability estimation across different proficiency levels. Although some problematic items were identified, the combined application of CTT and IRT enabled deeper psychometric evaluation than the use of either framework independently. The researchers further argued that integrating evidence from both approaches improves the interpretation of learner achievement and supports more defensible educational decisions.

Similarly, El-Hamamsy et al. (2023) developed and validated the competent Computational Thinking Test for primary school learners using CTT, IRT, Differential Item Functioning, and measurement invariance analyses. The findings showed that integrating these psychometric approaches enhanced assessment reliability, validity, and fairness across learner groups. Furthermore, the study demonstrated that IRT-based ability estimation provided a clearer interpretation of learner performance across grades and ability levels, while CTT statistics remained useful for preliminary item screening and overall test evaluation.

In another recent study, Li et al. (2025) conducted a psychometric evaluation of the Chinese version of the BENEFITS-CCCSAT scale using both CTT and IRT methods. The study found that reliability estimates derived from both frameworks produced converging evidence regarding scale quality, while IRT provided more detailed information concerning item discrimination and measurement precision. The researchers concluded that integrating CTT and IRT improves psychometric robustness and supports more accurate score interpretation and assessment decisions.

Although these international studies demonstrate the advantages of combining CTT and IRT, very limited research has examined the joint application of these frameworks within the context of Lesotho basic education. Assessment practices in Lesotho continue to rely predominantly on traditional CTT procedures, particularly within classroom assessments and national examinations administered by the Examinations Council of Lesotho (Ayanwale et al., 2022). Consequently, limited evidence exists regarding whether mathematics assessment items satisfy modern psychometric requirements related to reliability, validity, discrimination, model fit, and measurement precision across learner ability levels. Therefore, the present study is necessary because it seeks to address this empirical and methodological gap by systematically comparing CTT and IRT analyses within Lesotho basic education using mathematics checkpoint items. The study further aims to determine how integrating evidence from the two frameworks can improve test quality, learner performance

interpretation, and educational assessment practices in Lesotho. Thus, the study aims to answer the following research questions.

Research Questions

This investigation is meant to address the following research questions:

i. To what extent do the items in Lesotho basic educational assessments satisfy the psychometric requirements and assumptions of CTT and IRT models?
ii. How do CTT item statistics compare with IRT item parameter estimates in evaluating the quality of assessment items?
iii. To what extent does the combined application of CTT and IRT improve the reliability and validity of Lesotho basic educational assessments?
iv. How does integrating CTT and IRT evidence enhance the interpretation of learner performance in Lesotho basic educational assessments?
v. To what extent can the combined use of CTT and IRT support more defensible decisions regarding test quality and educational assessment practices in Lesotho?

MATERIAL AND METHOD

The study adopted a quantitative psychometric research design under the post-positivist paradigm. The target population comprised Grade 6 learners from international schools in Lesotho offering the Cambridge Primary Mathematics curriculum. A sample of 200 learners was selected using stratified and random sampling procedures. The instrument consisted of teacher-developed mathematics multiple-choice items validated by four subject matter experts using Content Validity Index (CVI) and Content Validity Ratio (CVR) procedures. Thereafter, the instrument was piloted with 30 Grade 6 Cambridge learners under standardised conditions. Subsequently, item difficulty, discrimination indices, and distractor analyses were conducted within the CTT framework to assess psychometric quality. In addition, learner feedback on clarity and readability was collected. Based on both quantitative and qualitative evidence, poorly functioning items were revised, while some were discarded. Thus, the final 30-item test was considered valid, reliable, curriculum-aligned, and suitable for administration to the target population. The Data were analysed using jMetrik version 4.1.1 through both CTT and IRT frameworks. Specifically, CTT analyses examined item difficulty,

discrimination, reliability, and distractor functioning, while IRT analyses utilised 1PL and 2PL models to estimate item parameters and model fit. IRT assumptions of dimensionality, local independence, and monotonicity were also evaluated.

RESULTS

The results of this study are presented from the preliminary analysis to the targeted integration of CTT and IRT, starting from the IRT assumptions.

IRT Assumptions

The dataset consisting of 200 learners and 30 multiple-choice mathematics items was analysed to evaluate the core assumptions underlying IRT: dimensionality, local independence, and monotonicity.

The dimensionality assumption was evaluated using the inter-item correlation matrix and eigenvalue structure. Specifically, the first eigenvalue was 3.58, while the second eigenvalue was 1.77, producing a ratio of approximately 2.02. Furthermore, several eigenvalues exceeded 1.00, indicating that the test is not strongly unidimensional. Although a dominant latent dimension appears to exist, the relatively low first-to-second eigenvalue ratio suggests the presence of secondary dimensions or subskills within the mathematics test (Reckase, 1979; Reise & Revicki, 2015). Therefore, the data demonstrate weak-to-moderate unidimensionality rather than strict unidimensionality (Embretson & Reise, 2000). This implies that the 1PL and 2PL IRT models may still be applied cautiously, but multidimensional influences are likely present.

Furthermore, the local independence assumption was examined through inter-item correlations. Most item correlations were relatively low, supporting approximate local independence. However, several item pairs showed moderate residual associations, including Item 6 to Item 25 ($r = 0.352$), Item 13 to Item 18 ($r = 0.330$), and Item 14 to Item 30 ($r = 0.325$). These elevated correlations suggest possible local dependence caused by overlapping content, shared solution strategies, or similar wording structures (Chang, Tseng, & Lou, 2012; Hambleton, Swaminathan, Rogers, & Rogers, 1991). Consequently, the assumption of local independence is only partially satisfied.

The monotonicity assumption was assessed through item-total relationship patterns. Several items displayed weak or negative item-total correlations, including Item 9 ($r = -0.187$), Item 25 ($r = -0.123$), Item 2 ($r = -0.119$), and Item 29 ($r = -0.114$). Hence, negative relationships indicate that

higher-performing learners were not consistently more likely to endorse higher item categories, which violates monotonicity expectations (Meijer, Sijtsma, Smid, Philips, & Eindhoven, 1990; Mokken, 1971). As a result, this suggests that some items may contain ambiguous distractors, poor alignment with learner ability, or flawed scoring structures.

Overall, the dataset demonstrates partial satisfaction of IRT assumptions. The test exhibits a dominant but imperfect latent structure, moderate evidence of local dependence among some items, and violations of monotonicity for several items. These findings imply that the application of IRT models is possible but should be interpreted cautiously, particularly for problematic items exhibiting weak psychometric behaviour. Thereafter, model fit was assessed to determine the extent to which the selected IRT models adequately represented the observed response data.

Model Fit

According to psychometric literature, the best-fitting IRT model is the one that balances statistical fit, parsimony, and theoretical appropriateness for the data (De Ayala, 2022; Hambleton et al., 1991). Therefore, IRT models were compared using deviance or -2 Log-Likelihood, Akaike Information Criterion (AIC), and the Bayesian Information Criterion (BIC) to check the best-fitting model for the data. As a result, to compare the two IRT models, the Log-Likelihood values from the jMetrik analysis of these two IRT models were used. Tables 1 and 2 illustrate the last part of 1PL and 2PL logs, indicating log-likelihood values from jMetrik, respectively.

Table 1: 1PL log

[INFO] 07 Jun 2025 13:27:46,727

PROX CYCLE: 1 3.3672958300 NaN

PROX CYCLE: 2 0.6380645356 NaN

PROX CYCLE: 3 0.0476344677 NaN

PROX CYCLE: 4 0.0085670342 NaN

STARTING EM CYCLES...

Number of available processors = 4

EM CYCLE: 1 0.2753275077 -2589.9652020981 [0 0 0 0]

EM CYCLE: 2 0.0179179447 -2588.0894511535 [0 0 0 0]

EM CYCLE: 3 0.0149586615 -2587.8439748510 [0 0 0 0]

EM CYCLE: 4 0.0118741476 -2587.6905038052 [0 0 0 0]

EM CYCLE: 5 0.0093957055 -2587.5943060441 [0 0 0 0]

EM CYCLE: 6 0.0074325057 -2587.5340159361 [0 0 0 0]

EM CYCLE: 7 0.0058784191 -2587.4962288717 [0 0 0 0]

EM CYCLE: 8 0.0046488859 -2587.4725390536 [0 0 0 0]

EM CYCLE: 9 0.0036764577 -2587.4576797445 [0 0 0 0]

EM CYCLE: 10 0.0029066550 -2587.4483552671 [0 0 0 0]

EM CYCLE: 11 0.0022977902 -2587.4424989383 [0 0 0 0]

EM CYCLE: 12 0.0018163907 -2587.4388163764 [0 0 0 0]

EM CYCLE: 13 0.0014358009 -2587.4364971378 [0 0 0 0]

EM CYCLE: 14 0.0011349272 -2587.4350336501 [0 0 0 0]

EM CYCLE: 15 0.0008970836 -2587.4341078978 [0 0 0 0]

Elapsed time: 0 secs, 568 msecs

Table 2: 2PL log

```
[INFO] 07 Jun 2025 19:24:59,431
PROX CYCLE:    1    3.3672958300    NaN
PROX CYCLE:    2    0.6380645356    NaN
PROX CYCLE:    3    0.0476344677    NaN
PROX CYCLE:    4    0.0085670342    NaN
STARTING EM CYCLES...
Number of available processors = 4
EM CYCLE:    1    7.4050550364    -2569.9057994965 [0 1 0 0]
EM CYCLE:    2    12.7160425458    -2505.7128645651 [0 1 0 0]
EM CYCLE:    3    6.0435854743    -2486.6597760133 [0 1 0 0]
EM CYCLE:    4    2.2349735673    -2480.6496042099 [0 1 0 0]
EM CYCLE:    5    1.7851492975    -2478.6003473559 [0 2 0 0]
EM CYCLE:    6    0.4404222796    -2476.7694836537 [0 2 0 0]
EM CYCLE:    7    0.0899960689    -2476.1399361774 [0 2 0 0]
EM CYCLE:    8    0.0675889909    -2475.8881573183 [0 2 0 0]
EM CYCLE:    9    0.0508580801    -2475.7735979284 [0 2 0 0]
EM CYCLE:    10    0.0384106642    -2475.7163409644 [0 2 0 0]
EM CYCLE:    11    0.0289195835    -2475.6826787483 [0 2 0 0]
EM CYCLE:    12    0.0218792078    -2475.6639615127 [0 2 0 0]
EM CYCLE:    13    0.0166348146    -2475.6533133781 [0 2 0 0]
EM CYCLE:    14    1.4139578568    -2475.6470844006 [0 2 0 0]
EM CYCLE:    15    0.0097042678    -2475.6131690435 [0 2 0 0]
EM CYCLE:    16    0.0074323683    -2475.6105914393 [0 2 0 0]
EM CYCLE:    17    0.0057039101    -2475.6090532438 [0 2 0 0]
EM CYCLE:    18    0.0043839830    -2475.6080715684 [0 2 0 0]
EM CYCLE:    19    0.0033677141    -2475.6074688170 [0 2 0 0]
EM CYCLE:    20    0.0025880040    -2475.6070516961 [0 2 0 0]
EM CYCLE:    21    0.0020312415    -2475.6074561618 [0 2 0 0]
EM CYCLE:    22    0.0015574483    -2475.6071243318 [0 2 0 0]
EM CYCLE:    23    0.0011968379    -2475.6070678124 [0 2 0 0]
EM CYCLE:    24    0.0009245891    -2475.6069527556 [0 2 0 0]
Elapsed time: 0 secs, 744 msecs
```

Using the final log-likelihood values from the jMetrik outputs, the model comparison statistics were computed as follows:

- 1PL final log-likelihood = −2587.4341078978
- 2PL final log-likelihood = −2475.6069527556

Given:

- Number of items = 30
- Sample size $(N) = 200$

Parameter counts:

- 1PL: $k = 30$ difficulty parameters
- 2PL: $k = 60$ difficulty + discrimination parameters

The formulas used were:

$$Deviance = -2LL$$
$$AIC = -2LL + 2k$$
$$BIC = -2LL + k\ln(N)$$

where:

- k= number of estimated parameters,
- $\ln(N)$= natural logarithm of the sample size.
- $\therefore \ln(200) = 5.2983$

Thus, in this study, 5.2983 acts as the penalty weight for model complexity in the BIC formula. The larger the sample size, the larger the ln(N) value becomes, and the more strongly BIC penalises models with many parameters (Schwarz, 1978). Consequently, BIC tends to favour simpler and more parsimonious models unless a more complex model substantially improves fit.

The computed results are presented in Table 3 below.

Table 3: Deviance, AIC, and BIC Results for Model Fit

Model	Log-Likelihood (LL)	Parameters (k)	Deviance (-2LL)	AIC	BIC
1PL	-2587.434	30	5174.868	5234.868	5393.817
2PL	-2475.607	60	4951.214	5071.214	5389.112

The results indicate that the 2PL model has substantially lower deviance (4951.214) and AIC (5389.112) values than the 1PL model, suggesting superior model fit to the data. Furthermore, the 2PL model also shows a slightly lower BIC value (5389.112) compared to the 1PL model (5393.817), despite the stronger penalty imposed for model complexity. Therefore, based on deviance, AIC, and BIC criteria, the 2PL model provides the best fit for the data because it captures variability in both item difficulty and item discrimination more effectively than the restrictive 1PL model. Thus, for further analysis and comparison of the two frameworks, 2PL was used, and results are presented from this IRT model.

Table 4 presents the CTT item analysis results generated using jMetrik. Specifically, the table reports the psychometric properties of the mathematics items, including item difficulty, item discrimination, distractor performance, reliability indices, and standard error of measurement. These statistics were examined to evaluate the quality, consistency, and preliminary functioning of the items before conducting more advanced latent-variable analyses.

Table 4: CTT Analysis Results

ITEM ANALYSIS
cammath1.CAMMATH1
June 7, 2025 12:38:15
==========

Item	Option (Score)	Difficulty	Std. Dev.	Discrimin.
item_1	Overall	0.2050	0.4047	0.1274
	1.0(1.0)	0.2050	0.4047	0.1274
	2.0(0.0)	0.3750	0.4853	-0.4744
	3.0(0.0)	0.1900	0.3933	-0.1112

	4.0(0.0)	0.2300	0.4219	-0.2020
item_2	Overall	0.1900	0.3933	0.1121
	1.0(0.0)	0.1300	0.3371	-0.1249
	2.0(1.0)	0.1900	0.3933	0.1121
	3.0(0.0)	0.1800	0.3852	-0.0999
	4.0(0.0)	0.5000	0.5013	-0.4830
item_3	Overall	0.1450	0.3530	0.0729
	1.0(0.0)	0.2250	0.4186	-0.0746
	2.0(0.0)	0.4350	0.4970	-0.4590
	3.0(1.0)	0.1450	0.3530	0.0729
	4.0(0.0)	0.1950	0.3972	-0.1636
item_4	Overall	0.2650	0.4424	0.0419
	1.0(0.0)	0.1750	0.3809	-0.3494
	2.0(1.0)	0.2650	0.4424	0.0419
	3.0(0.0)	0.3000	0.4594	-0.2650
	4.0(0.0)	0.2600	0.4397	-0.2004
item_5	Overall	0.5200	0.5009	0.0372
	1.0(1.0)	0.5200	0.5009	0.0372
	2.0(0.0)	0.1700	0.3766	-0.2612
	3.0(0.0)	0.1850	0.3893	-0.3176
	4.0(0.0)	0.1250	0.3315	-0.2464
item_6	Overall	0.1550	0.3628	0.1603
	1.0(0.0)	0.6100	0.4890	-0.4342
	2.0(1.0)	0.1550	0.3628	0.1603
	3.0(0.0)	0.1250	0.3315	-0.1814
	4.0(0.0)	0.1100	0.3137	-0.0901
item_7	Overall	0.2050	0.4047	-0.0070
	1.0(0.0)	0.2550	0.4370	-0.0940
	2.0(0.0)	0.2750	0.4476	-0.1512
	3.0(0.0)	0.2650	0.4424	-0.4764
	4.0(1.0)	0.2050	0.4047	-0.0070
item_8	Overall	0.2050	0.4047	0.1454

	1.0(1.0)	0.2050	0.4047	0.1454
	2.0(0.0)	0.2500	0.4341	-0.1276
	3.0(0.0)	0.3350	0.4732	-0.4594
	4.0(0.0)	0.2100	0.4083	-0.2429
item_9	Overall	0.1300	0.3371	0.0783
	1.0(0.0)	0.1500	0.3580	-0.0237
	2.0(0.0)	0.0950	0.2940	-0.0307
	3.0(1.0)	0.1300	0.3371	0.0783
	4.0(0.0)	0.6250	0.4853	-0.4977
item_10	Overall	0.2050	0.4047	0.0564
	1.0(1.0)	0.2050	0.4047	0.0564
	2.0(0.0)	0.4250	0.4956	-0.4279
	3.0(0.0)	0.1750	0.3809	-0.1319
	4.0(0.0)	0.1950	0.3972	-0.1636
item_11	Overall	0.1450	0.3530	0.2096
	1.0(0.0)	0.5800	0.4948	-0.4752
	2.0(0.0)	0.1300	0.3371	-0.2220
	3.0(1.0)	0.1450	0.3530	0.2096
	4.0(0.0)	0.1450	0.3530	-0.0386
item_12	Overall	0.1450	0.3530	0.2166
	1.0(0.0)	0.5650	0.4970	-0.3932
	2.0(0.0)	0.1300	0.3371	-0.1641
	3.0(1.0)	0.1450	0.3530	0.2166
	4.0(0.0)	0.1600	0.3675	-0.2361
item_13	Overall	0.4400	0.4976	-0.3295
	1.0(1.0)	0.4400	0.4976	-0.3295
	2.0(0.0)	0.1850	0.3893	-0.0740
	3.0(0.0)	0.2050	0.4047	-0.1675
	4.0(0.0)	0.1700	0.3766	-0.1293
item_14	Overall	0.0050	0.0707	0.0957
	1.0(0.0)	0.9400	0.2381	-0.4292

	2.0(0.0)	0.0400	0.1965	0.1771
	3.0(1.0)	0.0050	0.0707	0.0957
	4.0(0.0)	0.0150	0.1219	0.1114
item_15	Overall	0.9300	0.2558	-0.1362
	1.0(0.0)	0.0350	0.1842	-0.1109
	2.0(0.0)	0.0200	0.1404	-0.0981
	3.0(1.0)	0.9300	0.2558	-0.1362
	4.0(0.0)	0.0150	0.1219	0.0737
item_16	Overall	0.1000	0.3008	0.1179
	1.0(0.0)	0.1200	0.3258	-0.0773
	2.0(0.0)	0.6700	0.4714	-0.4461
	3.0(0.0)	0.1100	0.3137	-0.0542
	4.0(1.0)	0.1000	0.3008	0.1179
item_17	Overall	0.0750	0.2641	0.2276
	1.0(0.0)	0.0950	0.2940	-0.0385
	2.0(0.0)	0.7550	0.4312	-0.5192
	3.0(0.0)	0.0750	0.2641	0.0394
	4.0(1.0)	0.0750	0.2641	0.2276
item_18	Overall	0.2050	0.4047	0.1454
	1.0(1.0)	0.2050	0.4047	0.1454
	2.0(0.0)	0.1650	0.3721	-0.0545
	3.0(0.0)	0.2150	0.4119	-0.1072
	4.0(0.0)	0.4150	0.4940	-0.5751
item_19	Overall	0.1150	0.3198	0.1523
	1.0(0.0)	0.1750	0.3809	-0.0438
	2.0(0.0)	0.1000	0.3008	-0.0666
	3.0(0.0)	0.6100	0.4890	-0.5078
	4.0(1.0)	0.1150	0.3198	0.1523
item_20	Overall	0.0800	0.2720	0.1340
	1.0(0.0)	0.7700	0.4219	-0.3565
	2.0(1.0)	0.0800	0.2720	0.1340
	3.0(0.0)	0.0900	0.2869	-0.1188
	4.0(0.0)	0.0600	0.2381	-0.0834

item_21	Overall	0.1950	0.3972	0.0178
	1.0(0.0)	0.3050	0.4616	-0.2882
	2.0(0.0)	0.2950	0.4572	-0.2229
	3.0(1.0)	0.1950	0.3972	0.0178
	4.0(0.0)	0.2050	0.4047	-0.2417
item_22	Overall	0.1250	0.3315	0.0639
	1.0(0.0)	0.2000	0.4010	-0.0491
	2.0(0.0)	0.5200	0.5009	-0.4648
	3.0(1.0)	0.1250	0.3315	0.0639
	4.0(0.0)	0.1550	0.3628	-0.1210
item_23	Overall	0.1800	0.3852	0.0425
	1.0(0.0)	0.1900	0.3933	-0.0541
	2.0(1.0)	0.1800	0.3852	0.0425
	3.0(0.0)	0.4450	0.4982	-0.4819
	4.0(0.0)	0.1850	0.3893	-0.1312
item_24	Overall	0.1000	0.3008	0.0943
	1.0(0.0)	0.0900	0.2869	-0.1956
	2.0(0.0)	0.0850	0.2796	-0.0589
	3.0(1.0)	0.1000	0.3008	0.0943
	4.0(0.0)	0.7250	0.4476	-0.3371
item_25	Overall	0.0900	0.2869	0.0152
	1.0(0.0)	0.1150	0.3198	0.0780
	2.0(1.0)	0.0900	0.2869	0.0152
	3.0(0.0)	0.1100	0.3137	-0.1043
	4.0(0.0)	0.6850	0.4657	-0.4433
item_26	Overall	0.1450	0.3530	-0.0961
	1.0(0.0)	0.4950	0.5012	-0.4374
	2.0(1.0)	0.1450	0.3530	-0.0961
	3.0(0.0)	0.1500	0.3580	-0.0869
	4.0(0.0)	0.2100	0.4083	-0.0049

item_27	Overall	0.1050	0.3073	0.1013
	1.0(0.0)	0.0850	0.2796	-0.1230
	2.0(0.0)	0.0350	0.1842	0.0995
	3.0(1.0)	0.1050	0.3073	0.1013
	4.0(0.0)	0.7750	0.4186	-0.4189
item_28	Overall	0.4000	0.4911	-0.2647
	1.0(0.0)	0.2000	0.4010	-0.1601
	2.0(1.0)	0.4000	0.4911	-0.2647
	3.0(0.0)	0.2000	0.4010	-0.0548
	4.0(0.0)	0.2000	0.4010	-0.2563
item_29	Overall	0.1350	0.3426	-0.0496
	1.0(0.0)	0.1250	0.3315	-0.1217
	2.0(0.0)	0.1250	0.3315	-0.2400
	3.0(1.0)	0.1350	0.3426	-0.0496
	4.0(0.0)	0.6150	0.4878	-0.2411
item_30	Overall	0.1000	0.3008	0.0554
	1.0(0.0)	0.6350	0.4826	-0.4555
	2.0(1.0)	0.1000	0.3008	0.0554
	3.0(0.0)	0.1600	0.3675	0.0053
	4.0(0.0)	0.1050	0.3073	-0.1005

==========

TEST LEVEL STATISTICS

==========

Number of Items = 30

Number of Examinees = 200

Min = 1.0000

Max = 13.0000

Mean = 6.0400

Median = 6.0000

Standard Deviation = 2.2042

Interquartile Range = 3.0000

Skewness = 0.4965

Kurtosis = 0.2964

KR21 = 0.0073

==========

RELIABILITY ANALYSIS

==========

Method	Estimate	95% Conf. Int.	SEM
Guttman's L2	0.2656	(0.1112, 0.4046)	1.8936
Coefficient Alpha	0.1975	(0.0287, 0.3494)	1.9795
Feldt-Gilmer	0.3125	(0.1679, 0.4426)	1.8322
Feldt-Brennan	0.2003	(0.0322, 0.3517)	1.9760
Raju's Beta	0.1975	(0.0287, 0.3494)	1.9795

==========

RELIABILITY IF ITEM DELETED

==========

Item	L2	Alpha	F-G	F-B	Raju
item_1	0.2361	0.1642	0.3026	0.1665	0.1642
item_2	0.2411	0.1698	0.3197	0.1722	0.1698
item_3	0.2533	0.1830	0.3111	0.1858	0.1830
item_4	0.2614	0.1916	0.2885	0.1944	0.1916
item_5	0.2638	0.1940	0.3569	0.1974	0.1940
item_6	0.2290	0.1576	0.3170	0.1599	0.1576
item_7	0.2753	0.2068	0.3207	0.2098	0.2068
item_8	0.2327	0.1583	0.4366	0.1608	0.1583
item_9	0.2513	0.1820	0.3213	0.1846	0.1820
item_10	0.2571	0.1869	0.3189	0.1896	0.1869
item_11	0.2169	0.1444	0.3962	0.1465	0.1444
item_12	0.2166	0.1424	0.3248	0.1445	0.1424
item_13	0.3723	0.3251	0.3636	0.3285	0.3251
item_14	0.2613	0.1928	0.3172	0.1954	0.1928
item_15	0.2909	0.2265	0.3105	0.2296	0.2265
item_16	0.2445	0.1740	0.3164	0.1765	0.1740
item_17	0.2263	0.1536	0.4244	0.1559	0.1536
item_18	0.2307	0.1583	0.3707	0.1605	0.1583
item_19	0.2362	0.1640	0.3553	0.1664	0.1640
item_20	0.2441	0.1725	0.3784	0.1753	0.1725
item_21	0.2677	0.1989	0.3069	0.2019	0.1989
item_22	0.2559	0.1859	0.3147	0.1885	0.1859
item_23	0.2608	0.1913	0.2916	0.1942	0.1913
item_24	0.2502	0.1795	0.3399	0.1823	0.1795

item_25	0.2659	0.1978	0.2909	0.2005	0.1978
item_26	0.2935	0.2284	0.2975	0.2312	0.2284
item_27	0.2475	0.1775	0.2840	0.1801	0.1775
item_28	0.3561	0.3020	0.3457	0.3058	0.3020
item_29	0.2811	0.2154	0.2968	0.2184	0.2154
item_30	0.2576	0.1887	0.3099	0.1914	0.1887

==========

L2: Guttman's lambda-2

Alpha: Coefficient alpha

F-G: Feldt-Gilmer coefficient

F-B: Feldt-Brennan coefficient

Raju: Raju's beta coefficient

CONDITIONAL STANDARD ERROR OF MEASUREMENT

==========

True Score	CSEM
0	0.0
1	0.86
2	1.19
3	1.43
4	1.62
5	1.78
6	1.91
7	2.02
8	2.11
9	2.19
10	2.25
11	2.30
12	2.34
13	2.37
14	2.39
15	2.39
16	2.39
17	2.37
18	2.34
19	2.30
20	2.25

21	2.19
22	2.11
23	2.02
24	1.91
25	1.78
26	1.62
27	1.43
28	1.19
29	0.86
30	0.0

Elapsed time: 0 secs, 90 msecs

The table below indicates the results of the 2PL model from jMetrik software.

Table 5: 2PL Model Analysis Results

IRT ITEM CALIBRATION

cammath1.CAMMATH1

June 7, 2025 19:24:58

MMLE ITEM PARAMETER ESTIMATES

Item	Code	Apar (SE)	Bpar (SE)	Cpar (SE)	Upar (SE)
item_1	L2	0.67 (0.19)	2.20 (0.60)		
item_2	L2	0.58 (0.19)	2.66 (0.83)		
item_3	L2	0.43 (0.19)	4.28 (1.87)		
item_4	L2	0.32 (0.15)	3.27 (1.58)		
item_5	L2	-0.25 (0.15)	0.32 (0.59)		
item_6	L2	1.14 (0.25)	1.82 (0.32)		
item_7	L2	0.24 (0.11)	5.65 (2.43)		
item_8	L2	0.46 (0.18)	3.05 (1.14)		
item_9	L2	0.69 (0.23)	2.99 (0.89)		
item_10	L2	0.47 (0.18)	3.02 (1.12)		
item_11	L2	1.21 (0.26)	1.83 (0.31)		

item_12	L2	1.01 (0.24)	2.06 (0.41)
item_13	L2	-1.86 (0.27)	-0.19 (0.10)
item_14	L2	1.13 (0.67)	5.20 (2.62)
item_15	L2	0.08 (0.08)	-33.92 (37.38)
item_16	L2	0.81 (0.26)	3.02 (0.84)
item_17	L2	1.43 (0.34)	2.26 (0.38)
item_18	L2	0.80 (0.20)	1.91 (0.45)
item_19	L2	0.67 (0.23)	3.28 (1.04)
item_20	L2	0.47 (0.22)	5.35 (2.41)
item_21	L2	0.33 (0.17)	4.41 (2.19)
item_22	L2	0.69 (0.23)	3.05 (0.91)
item_23	L2	0.48 (0.19)	3.33 (1.26)
item_24	L2	0.76 (0.25)	3.16 (0.93)
item_25	L2	0.71 (0.26)	3.53 (1.16)
item_26	L2	0.30 (0.19)	6.00 (3.64)
item_27	L2	0.79 (0.25)	2.99 (0.83)
item_28	L2	0.03 (0.03)	13.27 (15.65)
item_29	L2	0.37 (0.18)	5.20 (2.50)
item_30	L2	0.64 (0.24)	3.66 (1.27)

ITEM FIT STATISTICS

Item	S-X2	df	p-value
item_1	6.8485	9.0000	0.6529
item_2	15.9362	9.0000	0.0682
item_3	9.2553	9.0000	0.4141
item_4	11.0444	9.0000	0.2727
item_5	15.1483	9.0000	0.0869
item_6	12.8920	9.0000	0.1676
item_7	11.8095	9.0000	0.2243
item_8	3.9137	9.0000	0.9170
item_9	12.0165	9.0000	0.2124
item_10	7.7791	9.0000	0.5565
item_11	24.8809	9.0000	0.0031
item_12	10.7478	9.0000	0.2934
item_13	8.0097	9.0000	0.5332
item_14	13.1202	8.0000	0.1078
item_15	15.4990	9.0000	0.0781

item_16	13.5619	9.0000	0.1388
item_17	12.9381	9.0000	0.1654
item_18	6.3305	9.0000	0.7064
item_19	14.2951	9.0000	0.1122
item_20	12.9400	9.0000	0.1653
item_21	11.7983	9.0000	0.2249
item_22	14.6746	9.0000	0.1003
item_23	8.6674	9.0000	0.4685
item_24	6.1102	9.0000	0.7288
item_25	19.9314	9.0000	0.0183
item_26	11.4308	9.0000	0.2473
item_27	19.5953	9.0000	0.0206
item_28	21.1541	9.0000	0.0120
item_29	10.1963	9.0000	0.3348
item_30	12.7596	9.0000	0.1738

==========

LATENT DISTRIBUTION

==========

Point	Density
-3.00000000	5.55191583e-04
-2.87500000	8.01512645e-04
-2.75000000	1.13917916e-03
-2.62500000	1.59399823e-03
-2.50000000	2.19582531e-03
-2.37500000	2.97798075e-03
-2.25000000	3.97612596e-03
-2.12500000	5.22651898e-03
-2.00000000	6.76361810e-03
-1.87500000	8.61707301e-03
-1.75000000	1.08082310e-02
-1.62500000	1.33463838e-02
-1.50000000	1.62250764e-02
-1.37500000	1.94188738e-02
-1.25000000	2.28810251e-02
-1.12500000	2.65424535e-02
-1.00000000	3.03124333e-02

-0.87500000	3.40811849e-02
-0.75000000	3.77244328e-02
-0.62500000	4.11097562e-02
-0.50000000	4.41043303e-02
-0.37500000	4.65834566e-02
-0.25000000	4.84391305e-02
-0.12500000	4.95878314e-02
0.00000000	4.99767536e-02
0.12500000	4.95878314e-02
0.25000000	4.84391305e-02
0.37500000	4.65834566e-02
0.50000000	4.41043303e-02
0.62500000	4.11097562e-02
0.75000000	3.77244328e-02
0.87500000	3.40811849e-02
1.00000000	3.03124333e-02
1.12500000	2.65424535e-02
1.25000000	2.28810251e-02
1.37500000	1.94188738e-02
1.50000000	1.62250764e-02
1.62500000	1.33463838e-02
1.75000000	1.08082310e-02
1.87500000	8.61707301e-03
2.00000000	6.76361810e-03
2.12500000	5.22651898e-03
2.25000000	3.97612596e-03
2.37500000	2.97798075e-03
2.50000000	2.19582531e-03
2.62500000	1.59399823e-03
2.75000000	1.13917916e-03
2.87500000	8.01512645e-04
3.00000000	5.55191583e-04

==========

Mean = 0.0000

Std. Dev. = 0.9887

2 secs, 137 msecs Elapsed time: 2 secs, 174 msecs

From the two tables above, results are presented according to the five research questions and interpreted using both CTT and IRT frameworks.

Research Question 1: To what extent do the items in Lesotho basic educational assessments satisfy the psychometric requirements and assumptions of CTT and IRT models?

Under the CTT framework, item difficulty indices ranged from 0.005 to 0.930, showing wide variation in difficulty, with most items classified as difficult because many *p*-values fell outside the recommended range. Several items demonstrated negative discrimination values, including items 13, 26, and 28, indicating problematic functioning and possible construct irrelevance. Reliability analysis revealed very low internal consistency, with Cronbach's alpha (α) = 0.1975 and KR-21 = 0.0073, suggesting that the test failed to meet acceptable psychometric standards.

In contrast, within the IRT 2PL model, most items showed acceptable fit statistics, although items 11, 25, 27, and 28 significantly misfit the model. Discrimination parameters ranged from -1.86 to 1.43, while difficulty parameters ranged from -33.92 to 13.27. Extreme estimates for items 15 and 28 suggested unstable calibration and poor item behaviour. Overall, the findings revealed weaknesses in test construction and item quality despite acceptable IRT fit for some items. **Figures 1** to **4** illustrate the item characteristic curves (ICCs) of items that indicate misfit (items 13 and 25) and extreme estimates (items 15 and 28), respectively.

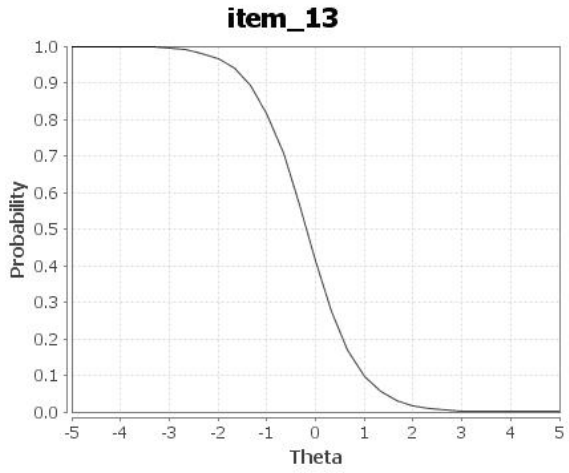

Figure 1: ICC for item 13

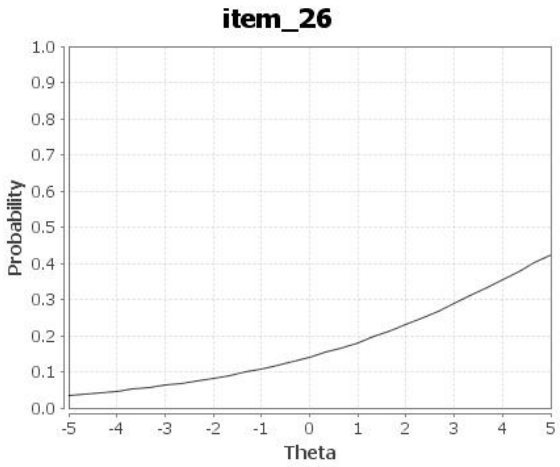

Figure 2: ICC for item 26

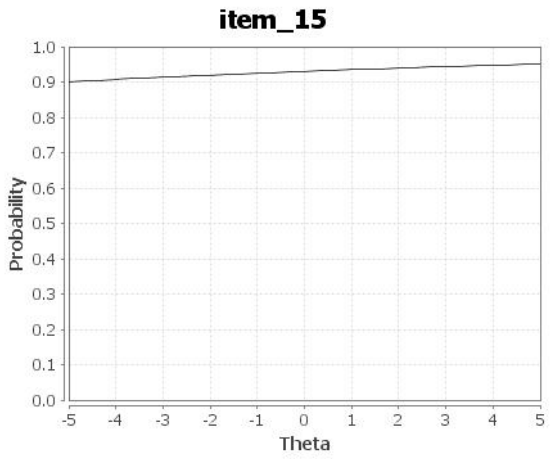

Figure 3: ICC for item 15

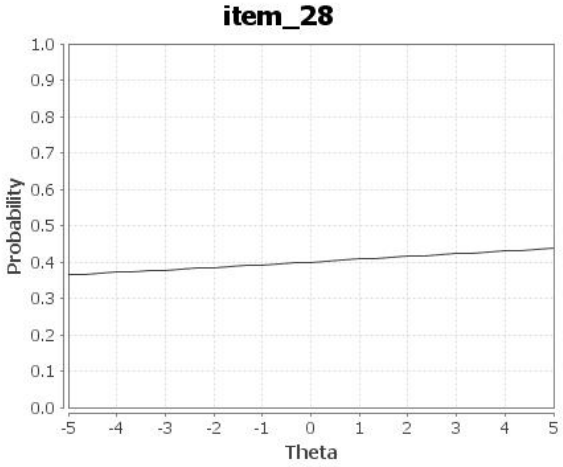

Figure 4: ICC for item 28

Research Question 2: How do CTT item statistics compare with IRT item parameter estimates in evaluating the quality of assessment items?

The comparison between CTT and IRT showed consistent identification of problematic items. Item 13 demonstrated negative discrimination under both CTT (-0.3295) and IRT (a = -1.86), confirming severe item malfunctioning. Similarly, item 28 showed negative CTT discrimination and extremely poor IRT parameters, including very low discrimination (a = 0.03) and excessively high difficulty (b = 13.27), indicating weak differentiation among learners. While CTT identified weak items mainly through low or negative discrimination indices, IRT provided more detailed diagnostic information by estimating item discrimination and difficulty independently. Items 11, 12, and 17 showed strong IRT discrimination values (a > 1.00), suggesting effective differentiation of learner ability despite only moderate CTT discrimination indices. Overall, IRT provided richer item-level diagnostic evidence than CTT across varying ability levels.

Research Question 3: To what extent does the combined application of CTT and IRT improve the reliability and validity of Lesotho basic educational assessments?

The integration of CTT and IRT provided complementary evidence of reliability and validity. CTT showed poor overall internal consistency, with coefficient alpha = 0.1975 and Guttman's L2 = 0.2656. However, IRT analysis indicated that several items demonstrated acceptable discrimination

and fit, suggesting that some parts of the test measured learner ability more effectively than reflected by CTT reliability estimates.

On the other hand, IRT underscored the Test Information Function (TIF) that indicated that measurement precision was strongest at moderate-to-high ability levels, particularly for items 6, 11, 12, and 17 with high discrimination values ($a > 1.00$). In contrast, CTT assumed constant measurement error, although the Conditional Standard Error of Measurement (CSEM) varied across score levels and exceeded 2.30 in the middle range, indicating inconsistent precision (Huebner & Skar, 2021).

Overall, combining CTT and IRT improved the evaluation of both overall test consistency and conditional measurement precision across ability levels. Figure 5 indicates an example of ICC illustrating a strong discrimination.

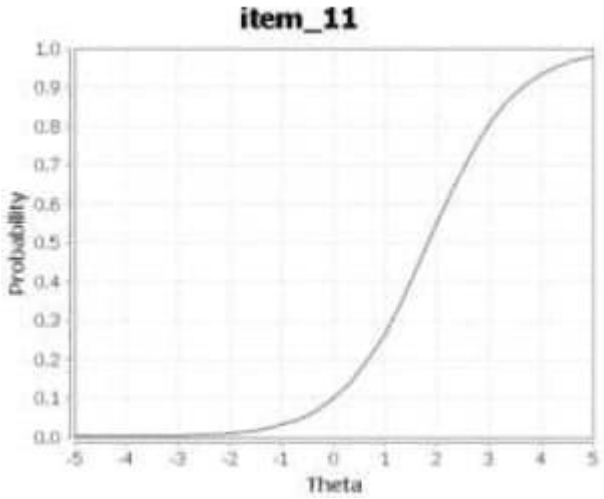

Figure 5: ICC with item discrimination value = 1.21

Research Question 4: How does integrating CTT and IRT evidence enhance the interpretation of learner performance in Lesotho basic educational assessments?

The combined use of CTT and IRT enhanced score interpretation by providing both observed-score and latent-trait perspectives. CTT results showed that the mean test score was 6.04 out of 30,

indicating generally low learner performance. However, IRT extended interpretation beyond raw scores by locating learners and items on a common latent ability scale.

The latent distribution showed an approximately normal distribution with mean = 0.000 and standard deviation = 0.9887, indicating that learner abilities were distributed around average proficiency (De Ayala, 2022; Embretson & Reise, 2000). Furthermore, the item difficulty parameters demonstrated that many items targeted higher ability levels, which partly explains the low observed performance under CTT.

Thus, integrating CTT and IRT evidence enabled a more comprehensive interpretation of learner achievement by linking raw-score performance with latent proficiency estimation and item functioning characteristics.

Research Question 5: To what extent can the combined use of CTT and IRT support more defensible decisions regarding test quality and educational assessment practices in Lesotho?

The findings demonstrated that combining CTT and IRT strengthened evidence-based decision-making regarding assessment quality. Hence, CTT identified low reliability and weak item discrimination, highlighting concerns about the consistency of the examination. Simultaneously, IRT identified specific misfitting items and unstable parameter estimates that could compromise fairness and validity if retained in operational testing.

The integration of both frameworks allowed problematic items such as 13, 25, 27, and 28 to be identified consistently across methods, supporting stronger item revision and test development decisions. Additionally, IRT-based information regarding item functioning across ability levels provided a stronger basis for defensible educational decisions than reliance on classical indices alone.

Comparison of CTT Internal Reliability and IRT Reliability Through Test Information

CTT internal reliability estimates were notably low, with Cronbach's alpha = 0.1975 and KR-21 = 0.0073, indicating weak score consistency across the test (Tavakol & Dennick, 2011; Zakariya, 2022). These coefficients suggest that the examination produced substantial measurement error and limited dependability.

In contrast, the IRT Test Information Function (TIF) indicated that the assessment provided greater measurement precision for learners with moderate-to-high ability levels, mainly due to highly

discriminating items such as items 6, 11, 12, and 17 (a > 1.00). However, reliability was uneven across the ability continuum because several items showed weak or negative discrimination, reducing information for low-ability learners. Thus, Table 6 illustrates an interpretation of IRT-TIF across the learner ability continuum.

Table 6: IRT Test Information and Reliability Across Ability Levels

Ability Range (θ)	**Test Information (TIF Interpretation)**	**Reliability Interpretation**
Low ability (θ < -1)	Low information	Weak measurement precision
Average ability (θ ≈ 0)	Moderate information	Moderate reliability
Moderate-high ability (θ > 1)	High information	Strong measurement precision

Therefore, while CTT portrayed the test as generally unreliable, IRT demonstrated that measurement precision varied across ability levels, providing a more nuanced understanding of reliability. This illustrates the advantage of integrating CTT and IRT in educational assessment evaluation.

DISCUSSION

The results of the present study are consistent with existing psychometric literature showing that combining Classical Test Theory and Item Response Theory provides stronger evidence for evaluating educational assessments than using either framework independently. Studies comparing CTT and IRT frequently report that problematic items identified under CTT also tend to misfit under IRT models. For example, Idowu et al. (2011) found that items classified as "bad items" under CTT also failed to fit the Rasch model, demonstrating convergence between the two approaches. This aligns with the present study, where items 13, 25, 27, and 28 showed weak or negative discrimination under CTT and also demonstrated poor IRT fit or unstable parameter estimates.

The low internal consistency reliability observed in this study, Cronbach's alpha = 0.1975, suggests substantial measurement error and weak homogeneity among test items. Similar findings have been reported in educational measurement studies where poorly functioning items reduce test reliability and threaten validity. For instance, Cappelleri et al. (2014) explained that low reliability coefficients indicate insufficient consistency in measuring the intended construct and may undermine the

defensibility of score interpretation. The extremely low KR-21 estimate in the present study further supports the conclusion that the assessment lacked adequate internal consistency for educational decision-making.

Furthermore, the negative discrimination indices identified for several items are also consistent with prior research indicating item malfunctioning or construct-irrelevant variance. This is reported by Wongvorachan and Bulut (2025) that, items with low or negative discrimination parameters weaken measurement quality because they fail to differentiate high-performing learners from low-performing learners. Similarly, the current study found that item 13 and item 28 exhibited severe psychometric weaknesses across both CTT and IRT analyses, suggesting flawed item construction, ambiguous wording, or poorly functioning distractors.

Additionally, the IRT findings support previous literature showing that the 2PL model provides richer diagnostic information than CTT because it estimates both item discrimination and difficulty independently. For instance, a study conducted by Jumini and Retnawati (2022) on mathematics examinations using the 2PL model found that some items fit the model despite weak discrimination, while others displayed negative discrimination and poor functioning. This pattern resembles the present findings, where several items exhibited acceptable S-X^2 fit statistics despite weak classical discrimination values. The results, therefore, reinforce the argument that IRT can reveal conditional item functioning not fully observable under CTT (Brzezińska, 2020).

The present findings further support research indicating that the 2PL model may produce unstable estimates for problematic items or small samples. This is argued by Helbling et al. (2021) that although the 2PL model offers greater flexibility through variable discrimination parameters, it may generate volatile parameter estimates under non-optimal testing conditions. This observation is reflected in the current study, where some items produced extreme difficulty estimates such as $b = -33.92$ and $b = 13.27$, indicating calibration instability and potential item defects.

The comparison between CTT and IRT item statistics in this study also aligns with research showing moderate-to-high correspondence between the two frameworks. For instance, Abedalaziz and Leng (2013) reported significant positive relationships between CTT and IRT item difficulty and discrimination estimates across 1PL, 2PL, and 3PL models. Likewise, Umobong and Jacob (2016) found strong correlations between CTT-based and IRT-based item parameters and concluded that both frameworks can complement one another in evaluating educational tests. The present study

similarly found that items with weak discrimination under CTT generally demonstrated poor IRT functioning, indicating convergence between the frameworks.

The findings regarding the usefulness of IRT-based information functions are also supported by existing literature. IRT reliability differs fundamentally from CTT reliability because it varies across the latent trait continuum rather than assuming constant precision (Gordon, 2015). Fu et al. (2024) noted that item and test characteristic curves provide valuable diagnostic information about measurement precision at different ability levels. In the present study, the Test Information Function (TIF) suggested that the assessment measured moderate-to-high ability learners more precisely than low-ability learners, whereas CTT reliability provided only a single global estimate. This demonstrates the practical advantage of IRT for interpreting learner proficiency and evaluating conditional precision.

The overall findings, therefore, support the broader psychometric consensus that integrating CTT and IRT strengthens assessment validation. As a result, Santoso et al. (2024) concluded that CTT and IRT analyses complement one another because both frameworks can identify problematic items while simultaneously providing different but useful perspectives on measurement quality. Similarly, the current study demonstrated that integrating CTT and IRT improved understanding of item quality, learner performance, reliability, and score interpretation within Lesotho's educational assessment context.

LIMITATIONS

This study had several limitations to be acknowledged. First, the sample consisted of only 200 Grade 6 learners from international schools in Maseru district, limiting the generalisability of the findings to other school systems and regions in Lesotho. Again, the relatively small sample size may have contributed to unstable IRT parameter estimates for some items, particularly under the 2PL model. Furthermore, the study focused only on a teacher-developed mathematics multiple-choice test aligned with the Cambridge curriculum, excluding other subjects, assessment formats, and national curriculum contexts. The study also relied solely on CTT to analyse piloted items, limiting the evaluation to basic reliability and item statistics without providing deeper diagnostic information available through IRT. Moreover, the study was limited to CTT, 1PL, and 2PL analyses and did not examine more advanced psychometric procedures such as 3PL modelling, Differential Item

Functioning (DIF), or measurement invariance analyses, which could provide additional evidence regarding test fairness and validity. More importantly, although some IRT assumptions were violated during preliminary analysis, the study still compared IRT models, which may have affected item functioning because some items showed signs of multidimensionality rather than strict unidimensionality.

CONCLUSION

This study evaluated mathematics test items in Lesotho basic education using CTT and IRT, and the findings showed that the test failed to meet acceptable psychometric standards. Specifically, CTT revealed extreme variation in item difficulty, weak and negative discrimination indices, and very low reliability (Cronbach's alpha = 0.1975), whereas the 2PL IRT model identified four significantly misfitting items and unstable difficulty estimates for two items. Although the 2PL model fit better than the 1PL model, both frameworks consistently flagged problematic items, including items 13 and 28. Moreover, IRT provided richer diagnostic information through conditional measurement precision, while CTT highlighted overall unreliability. Consequently, integrating both approaches improved the interpretation of learner performance, as low raw scores were partly attributable to items targeting higher ability levels. Therefore, the teacher-developed test was psychometrically inadequate.

RECOMMENDATIONS

The study recommends improving assessment quality in Lesotho by strengthening item construction and removing poorly discriminating items. It further recommends integrating both Classical Test Theory (CTT) and Item Response Theory (IRT) in routine assessment analysis, since CTT alone may conceal item weaknesses. Again, teacher training should include psychometrics, item writing, and IRT applications. Furthermore, assessments should also undergo rigorous piloting and item calibration before large-scale administration. Moreover, future studies should use larger, more representative samples across Lesotho to improve generalisability. In addition, researchers should investigate Differential Item Functioning (DIF) and measurement invariance to evaluate fairness across learner groups. Finally, policymakers should establish national assessment systems that combine traditional and modern measurement theories for valid and equitable decision-making.

ACKNOWLEDGMENT

The author sincerely acknowledges Dr Kunle Ayanwale Musa for his invaluable guidance, mentorship, and encouragement throughout the development of this study. As his student in Psychometrics, the author greatly benefited from his scholarly support, research guidance, and commitment to advancing psychometric knowledge. This study originated from a term paper assignment under his instruction, which inspired the author to further develop the work into a full research study. The author is also deeply grateful to him for introducing the author to the Association of Computerized Adaptive Testing in Africa (ACATA) and for continuously encouraging scholarly writing and research in educational measurement and psychometrics.

REFERENCES

Abedalaziz, N., & Leng, C. H. (2013). The relationship between CTT and IRT approaches in analysing item characteristics. *Malaysian Online Journal of Educational Sciences*, *1*(1), 64–70.

Ayanwale, M. A., Chere-Masopha, J., & Morena, M. C. (2022). The classical test or item response measurement theory: The status of the framework at the Examination Council of Lesotho. *International Journal of Learning, Teaching and Educational Research*, *21*(8), 384–406. https://doi.org/10.26803/ijlter.21.8.22

Brzezińska, J. (2020). Computation item response theory models in the measurement theory. *Communications in Statistics - Simulation and Computation*, *49*(12), 3299–3313. https://doi.org/10.1080/03610918.2018.1546399

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

Chang, C., Tseng, K., & Lou, S. (2012). A comparative analysis of the consistency and difference among in a web-based portfolio assessment environment for high school students. *Computers and Education*, *58*(1), 303–320. https://doi.org/10.1016/j.compedu.2011.08.005

Chen, Y., Li, X., Liu, J., & Ying, Z. (2021). Item response theory – A statistical framework for educational and psychological measurement. *Statistical Science*, *40*(2), 167–194.

Cook, L. L., & Pitoniak, M. J. (2025). *Educational measurement* (5th ed.). New York: Oxford University Press.

De Ayala, R. J. (2022). The theory and practice of item response theory. *Change for the Better: Personal Development through Practical Psychotherapy*, *77*. https://doi.org/10.4135/9781526438447.n17

DeMars, C. (2010). *Item response theory: Understandig statistics measurement*. New York: Oxford University Press.

El-Hamamsy, L., Zapata-C´aceres, M., Martin-Barroso, E., Mondada, F., Zufferey, J. D., Bruno, B., & Roman-Gonzalez, M. (2023). The competent computational Thinking test (cCTt): A valid, reliable and gender-fair test for longitudinal CT studies in grades 3-6. *Technology, Knowledge and Learning*, *30*(3), 1607–1661.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. New Jersey: Lawrence Erlbaum Associates.

Fu, J., Tan, X., & Kyllonen, P. C. (2024). Item and test characteristic curves of rank-2PL models for multidimensional forced-choice questionnaires. *Applied Measurement in Education*, *37*(3), 272–288. https://doi.org/10.1080/08957347.2024.2386939.Item

Geleta, K. T., Fisseha, M., & Zenebe, N. (2024). The relationship between the psychometric and performance properties of teacher-made tests and students ' academic performance in Ethiopian public universities: A baseline survey study. *Cogent Education*, *11*(1), 2298049. https://doi.org/10.1080/2331186X.2023.2298049

Gordon, R. A. (2015). Measuring constructs in family science: How can item response theory improve precision and validity? *Journal of Marriage and Family*, *77*(1), 147–176.

Hambleton, R. K., Swaminathan, H., Rogers, H. J., & Rogers, D. J. (1991). Fundamentals of item response theory. In *Choice Reviews Online* (Vol. 21). London: SAGE Publications. https://doi.org/10.2307/2075521

Helbling, L. A., Berger, S., & Verschoor, A. (2021). Flexibility at the price of volatility: Concurrent calibration in multistage tests in practice using a 2PL model. *Frontiers in Education*, *6*, 679864. https://doi.org/10.3389/feduc.2021.679864

Hu, Z., Lin, L., Wang, Y., & Li, J. (2021). The integration of classical testing theory and item response theory. *Psychology*, *12*(9), 1397–1409. https://doi.org/10.4236/psych.2021.129088

Huebner, A., & Skar, G. B. (2021). Conditional standard error of measurement: Classical test theory, generalizability theory and many-facet Rasch measurement with applications to writing assessment. *Practical Assessment, Research, and Evaluation*, *26*(14), 1–21. https://doi.org/10.7275/vzmm-0z68

Idowu, O. I., Eluwa, A. N., & Abang, B. K. (2011). Evaluation of mathematics achievement test: A comparison between classical test theory(CTT) and item Response theory(IRT). *Journal of Educational and Social Research*, *1*(4), 99–106.

Jumini, & Retnawati, H. (2022). Estimating item parameters and student abilities: An IRT 2PL analysis of mathematics examination. *Al-Ishlah: Jurnal Pendidikan*, *14*(1), 385–398. https://doi.org/10.35445/alishlah.v14i1.926

Li, C., Lin, Y., Tosun, B., Wang, P., Ye Guo, H., Ling, C. R., … Zhang, L. (2025). Psychometric evaluation of the Chinese version of the BENEFITS-CCCSAT based on CTT and IRT : a cross-sectional design translation and validation study. *Frontiers in Public Health*, *13*, 1532709. https://doi.org/10.3389/fpubh.2025.1532709

Magno, C. (2009). Demonstrating the difference between classical test theory and item response theory using derived test data. *The International Journal of Educational and Psychological Assessment*, *1*(1), 1–11.

Meijer, R. R., Sijtsma, K., Smid, N. G., Philips, & Eindhoven. (1990). Theoretical and empirical comparison of the Mokken and the Rasch approach to IRT. *Applied Psychological Measurement*, *14*(3), 283–298. https://doi.org/10.1177/014662169001400306

Mitee, T. L. (2019). Comparative study of classical test theory and item response theory using item analysis results of quantitative chemistry achievement test. *AJB-SDR*, *1*(1), 26–36.

Mokken, R. J. (1971). A theory and procedure of scale analysis: With applications in political research. In *Contemporary Sociology* (1st ed.). Mouton: Mouton & Co. https://doi.org/10.2307/2062452

Reckase, M. D. (1979). Unifactor latent trait model applied to multifactor tests: Results and implications. *Journal of Educational Statistics*, *4*(3), 207–230.

Reise, S. P., & Revicki, D. A. (2015). Handbook of item response theory modelling: Applications to typical performance assessment. In *Handbook of Item Response Theory Modeling* (1st ed.). New York: Routledge. https://doi.org/10.4324/9781315736013-27

Rezaee, R., Shafiayan, M., Jafari, P., & Zarifsanaiey, N. (2018). Invariance of item difficulty parameter estimates based on classical test theory and item response theory. *Journal of Advance Pharmacy Education and Research, 8, 156-161.*, *8*, 156–161.

Rusch, T., Lowry, P. B., Mair, P., & Treiblmaier, H. (2017). Breaking free from the limitations of classical test theory: Developing and measuring information systems scales using item response theory. *Information and Management*, *54*(2), 189–203. https://doi.org/10.1016/j.im.2016.06.005

Santoso, P. H., Istiyono, E., Haryanto, & Retnawati, H. (2024). Validating light phenomena conceptual assessment through the lens of classical test theory and item response theory frameworks. *Physics Education*, *59*(2), 1–19. https://doi.org/10.1088/1361-6552/ad183b

Schwarz, G. (1978). Estimating the dimension of model. *The Annals of Statistics*, *6*(2), 461–464.

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.

Siregar, H. N. I., & Panjaitan, A. (2021). A systematic literature review: The role of classical test theory and item response theory in item analysis to determine the quality of mathematics tests. *Jurnal Fibonaci*, *2*(2), 29–48. Retrieved from https://doi.org/10.24114/jfi.v2i%0AJurnal

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, *2*, 53–55. https://doi.org/10.5116/ijme.4dfb.8dfd

Umobong, M., & Jacob, S. S. (2016). A comparison of classical and item response theory person/item parameters of physics achievement test for technical schools. *African Journal of Theory and Practice of Educational Assessment*, *4*, 131.

Wongvorachan, T., & Bulut, O. (2025). Detecting construct-irrelevant variance: A comparison of network psychometrics and traditional psychometric methods using the HEXACO-PI dataset. *Psychology International*, *7*(4), 88. Retrieved from https://doi.org/10.3390/ psycholint7040088

Yang, F. M., & Kao, S. T. (2014). Item Response Theory for measurement validity. *Shanghai Archives of Psychiatry*, *26*(3), 171–177. https://doi.org/10.3969/j.issn.1002-0829.2014.03

Zakariya, Y. F. (2022). Cronbach's alpha in mathematics education research: Its appropriateness, overuse, and alternatives in estimating scale reliability. *Frontiers in Psychology*, *13*, 1–6. https://doi.org/10.3389/fpsyg.2022.1074430