



Contents lists available at [Journal ELORA](#)

JRTI (Jurnal Riset Tindakan Indonesia)

ISSN: 2502-079X (Print) ISSN: 2503-1619 (Electronic)

Journal homepage: <https://jrti.eloracenter.org/jrti>



Classical test theory and item response theory validation of a building design and modeling fundamentals test in vocational education

Yolanda Putri Nami, Fahmi Rizal^{1*)}, M. Giatman, Ari Syaiful Rahman Arifin

Department of Technology and Vocational Education, Graduate School, Universitas Negeri Padang, Indonesia

Article Info

Article history:

Received Jul 10th, 2026
Revised Aug 21th, 2026
Accepted Sept 29th, 2026

Keyword:

Test instrument,
Vocational education,
Classical test theory,
Item response theory,
2PL

ABSTRACT

Classical test theory (CTT) and item response theory (IRT) rest on different measurement assumptions: CTT describes items relative to the tested group, whereas IRT places items and persons on a common latent ability scale. Using both frameworks therefore offers complementary validity evidence for classroom tests in vocational education, where such evidence is rarely reported. This research and development study developed a test for the Fundamentals of Building Design, Modeling, and Information subject (Dasar-dasar Desain Pemodelan dan Informasi Bangunan, Basic DPIB) and examined its items with CTT and the two-parameter logistic (2PL) IRT model. Forty multiple-choice items were written from a competency-based blueprint and judged by five experts (Aiken's $V = 0.80-0.95$). After a tryout with 30 students, 32 items were administered to 103 DPIB students selected by proportional stratified random sampling. CTT showed high split-half reliability (Spearman-Brown = 0.95), good-to-very-good discrimination for 30 items, and 13 items with at least one poorly functioning distractor; no item was classified as difficult. The 2PL model produced positive discrimination ($a = 0.848$ to 4.833) and mostly negative difficulty ($b = -2.597$ to 0.034), and test information peaked at $\theta \approx -0.98$. Both frameworks identified the same easiest, hardest, and most discriminating items, but they contributed different evidence: CTT pinpointed distractors to revise, whereas IRT showed that measurement precision is concentrated below average ability. The test is therefore better suited to identifying students who have not yet mastered basic competencies than to distinguishing high achievers; adding difficult items and recalibrating with a larger sample are required before high-stakes use.



© 2026The Authors.

This is an open access article under the CC BY-NC-SA license
(<https://creativecommons.org/licenses/by-nc-sa/4.0>)

Corresponding Author:

Fahmi Rizal,
Universitas Negeri Padang
Email fahmi@ft.unp.ac.id

Introduction

Vocational education primarily aims to provide students with the knowledge and skills that are required for the needs of the industry. Generally, vocational high school (SMK) graduates in Indonesia have higher education related characteristics that are in line with work requirements but the issue of a smoother move from school to the job market still needs to be addressed (Ariansyah et al., 2024). The mismatch between school learning and industry requirements has long been identified as a structural challenge of Indonesian vocational education (Suharno et al., 2020). As a result, performance assessment of students during vocational learning should be

supported by an instrument that would allow the collection of authentic data on the level of competency mastery by the pupils.

Tests serve a vital role in assessment. When done right, they can also facilitate the learning process. Murphy et al. (2023) revealed that testing can be a method for retrieving and practicing which in fact strengthens the learning instead of only being a tool for grading. Consequently, the goodness of a test instrument will greatly contribute to the assessment quality of learning.

The Building Information Modeling and Design (DPIB) study program introduces the Basic DPIB subject as a gateway to deeper learning of specific skills. The topics presented in this study are designed to give the learner an overview of the fundamentals of construction activity, occupational health, safety, and environment (K3LH), technical drawing, materials, energy efficiency, water management, and the implementation of sustainable building concepts. Apart from that, DPIB studies have revealed that students' mastery of the essential content is one of the factors which influences their academic performance and preparation for work-based learning (Manurung & Mulyana, 2024). Because the DPIB program prepares students for building information modeling (BIM) practice, which the construction industry increasingly requires of its workforce (Sacks & Pikas, 2013), weaknesses in these foundational competencies are likely to carry over into advanced BIM learning.

Although the importance of sound assessment is widely acknowledged, a review of recent literature reveals a clear gap. Studies of DPIB learning have focused mainly on instructional models and have reported learning outcomes measured with teacher-made tests whose psychometric properties are not documented (Manurung & Mulyana, 2024). Psychometric studies in Indonesian vocational schools have examined other domains, such as accounting (Purnama, 2017; Purnama & Alfarisa, 2020), while CTT–IRT comparisons in Indonesia have mostly addressed general subjects such as biology (Nisfatulsanah & Sugiharto, 2024) or aptitude tests (Setiawati et al., 2023). Internationally, dual-framework analyses have been applied mainly to national examinations and general-education tests (Awopeju & Afolabi, 2016; Ayanwale et al., 2022; Butakor, 2022), and BIM-related education research has concentrated on curriculum and industry requirements rather than on measurement (Sacks & Pikas, 2013). To the best of our knowledge, no validated instrument is available for measuring mastery of Basic DPIB competencies, and no study has examined how the evidence produced by CTT and IRT supports decisions about such a vocational classroom test. This gap matters because teachers need to know not only whether a test is reliable but also for which students it measures precisely.

The development of the instrument in this study is grounded in evidence-centered design (ECD), which links the competencies about which claims are made (competency model), the observable behaviors that count as evidence (evidence model), and the tasks that elicit those behaviors (task model) (Arieli-Attali et al., 2019). Validation follows the argument-based approach, in which different sources of evidence, namely expert judgment of content, item statistics, and score precision, support specific interpretations and uses of test scores (Cook et al., 2015). Within this framework, content validity is established through expert judgment quantified with Aiken's V index (Aiken, 1985; Kania et al., 2024; Nabil et al., 2022), and empirical evidence on the items is obtained from CTT and IRT.

In the stage of empirical analysis, Classical Test Theory (CTT) gives insights about test reliability, questions' levels of difficulty, item discrimination, item-total correlations, and how good/ineffective alternatives are as distractors. At the same time, Item Response Theory (IRT) permits one to obtain estimates of item parameters such as difficulty and discrimination of items on a latent ability scale (Gyamfi & Acquaye, 2023). Moreover, it is quite typical from the findings in meta-studies that CTT as well as IRT are the two leading methods in measuring instruments and that the selection of model and the understanding of parameters and information functions are critical aspects (Kılıç et al., 2023). The two frameworks rest on different assumptions. CTT decomposes an observed score into a true score and error at the level of the whole test, so its item statistics depend on the group tested, whereas IRT models the probability of a correct response as a function of latent ability and item parameters, so that measurement precision can vary along the ability scale (Bichi & Talib, 2018; Fan, 1998). Empirical comparisons show that CTT and IRT item statistics are often highly correlated (Fan, 1998; Setiawati et al., 2023), yet the frameworks differ in the decisions they support, particularly for individual-level interpretation (Jabrayilov et al., 2016). Using both is therefore not redundant: each supplies a different piece of the validity argument.

Purnama and Alfarisa's (2020) research proved that, apart from item analysis, we can use classical test theory and item response theory techniques to obtain information about item characteristics from a vocational high school test. According to Nisfatulsanah and Sugiharto's (2024) paper, CTT and IRT offer distinct but mutually reinforcing viewpoints during scale development. Accordingly, this study develops a Basic DPIB test and examines its characteristics with both CTT and the 2PL IRT model, treating the two frameworks as complementary sources of validity evidence.

The study aimed at: (1) Developing a DPIB basic test instrument using competencies and analyzing learner needs as a base; (2) Assessing the quality of different test items with Classical Test Theory (CTT); (3) Explaining the parameters of those items according to the 2-PL IRT model; and (4) Involving a comparison of the information from the two approaches. The contribution of this study is twofold: it provides a psychometrically documented test for Basic DPIB competencies, and it shows how the complementary evidence from CTT and IRT can be translated into concrete decisions about the use and revision of a vocational classroom test.

Method

This investigation was a Research and Development (R&D) study that utilized a quantitative approach. Instrument development followed eight stages: determining specifications, drafting a blueprint, item construction, expert evaluation, pretest, item analysis, revision, and test compilation. These stages follow the ECD logic of moving from competency claims to evidence and tasks (Arieli-Attali et al., 2019). Specifications were determined by mapping the learning outcomes of the Basic DPIB subject onto competency indicators covering construction fundamentals, occupational health, safety, and environment (K3LH), technical drawing, building materials, energy efficiency, water management, and sustainable building concepts. The blueprint (schematic plan) specified, for each indicator, the cognitive level and the number of items, so that every item could be traced to one indicator. Items were then written as five-option multiple-choice questions following established item-writing guidelines (Haladyna et al., 2002). The stages, samples, and outputs at each decision point are summarized in the Development Stages table below.

This research was conducted in 2026. Expert validation involved three expert lecturers and two Basic DPIB subject teachers from SMK Negeri 1 Padang. The tryout involved 30 students who answered all 40 items. Items were selected for the main study on the basis of the tryout statistics: items with a discrimination index below [criterion to be completed by the authors] or a difficulty index outside [range to be completed by the authors] were removed, and among the remaining items those with the highest discrimination were retained while keeping every blueprint indicator represented, resulting in 32 items. The main study involved 103 students enrolled in the DPIB program. The population comprised [N] students distributed across [k] classes. Classes served as strata; the number of students drawn from each class was proportional to its size, and students were randomly selected within each class. The sample consisted of [n] male and [n] female students. The sample size was determined by the accessible population rather than by an a priori power analysis. Because stable 2PL estimates generally require larger samples (Şahin & Anıl, 2017; Schroeders & Gnambs, 2025), the IRT results are interpreted as exploratory, and this is acknowledged as a limitation.

The first assessment tool was a multiple-choice paper having five alternative answers. The content and measuring aspects were determined based mainly on what the learners were supposed to know, plus of what was in the test design blueprint that had been developed. Content validity information was gathered from the ratings of the five validators, who judged the relevance of each item to its indicator on a five-point scale (1 = not relevant, 5 = highly relevant); the ratings were analyzed with Aiken's V index, which expresses the degree of expert agreement and allows a significance-based cutoff for small panels (Aiken, 1985; Kania et al., 2024). In this work, the cutoff point for a score indicating content validity had taken the level of 0.78. Student responses from the fieldwork were used for the analyses. The analyses were carried out using classical test theory (CTT) with ANATES and the two-parameter logistic (2PL) item response theory (IRT) model with Stata. ANATES estimates reliability with the split-half method corrected by the Spearman–Brown formula; different CTT programs have been shown to produce consistent item indices (Nurjanah et al., 2024). The 2PL model was estimated with the `irt 2pl` command in Stata, which uses marginal maximum likelihood estimation with adaptive Gauss–Hermite quadrature.

The CTT analysis involved five major areas; reliability, item difficulty, discrimination, validity via item-total correlation and distractor quality. The 2PL IRT parameter model was employed to estimate item discrimination (a) and difficulty (b) parameters. Besides, Item Characteristic Curve (ICC), Item Information Function (IIF), Test Information Function (TIF), and Test Characteristic Curve (TCC) were also investigated. IRT method was employed as CTT complementary to the CTT approach since, in IRT, item characteristics can be identified through the connection of a person's latent ability to his or her responses (Gyamfi & Acquaye, 2023). The 2PL model was chosen over the Rasch (1PL) model because items written for different content areas are unlikely to discriminate equally, and the Rasch model assumes equal discrimination across items (Boone, 2016); comparative studies have also found the 2PL model to be the most informative about discrimination (Setiawati et al., 2023). The three-parameter (3PL) model was not used because its pseudo-guessing parameter is difficult to estimate in small samples (Şahin & Anıl, 2017), although it can yield better ability estimates when large samples are available (Iwintolu et al., 2024).

Table 1. Research Stages, Activities, and Outputs in Instrument Development and Validation

Stage	Activity	Output
1	Needs and competency analysis	Test specification and indicators
2	Blueprint development and drafting of 40 items	Initial instrument draft
3	Content validation by 5 validators	Content-validated instrument
4	Limited tryout with 30 students	Initial empirical data
5	Item analysis and selection	32 items for the main study
6	Main study with 103 students	32 × 103 response data
7	CTT and 2PL IRT analysis	Item characteristics and parameters
8	Revision/instrument assembly	Final instrument draft

Results and Discussions

Content Validity of the Instrument

Content validation involved 40 items validated by 5 validators who were 3 senior lecturers and 2 subject teachers of Basic DPIB from SMK Negeri 1 Padang. All items attained Aiken's V values between 0.80 and 0.95. There are several items having the minimum value ($V = 0.80$) but the maximum was Item 1 ($V = 0.95$). The mean Aiken's V score was 0.83125 which is much higher than the standard content-validity threshold of 0.78. The rationale is, all 40 items qualify as well-validated content. Aiken's V is used similarly as how expert-agreement is assessed by Aiken (1985). Its application is also well-documented content validity of educational instruments (Nabil et al., 2022). This range is comparable to that reported in recent instrument development studies using five-expert panels, for example 0.817–0.884 for a mathematical problem-solving test (Kania et al., 2024). Aiken's V, however, provides evidence of content representativeness only; it does not show how the items function empirically, which is why the tryout and the main study were necessary.

Table 2. Summary of Content Validity Results for the Research Instrument

Indicator	Result
Number of items validated	40
Number of validators	5 (3 expert lecturers; 2 subject teachers)
Range of Aiken's V	0.80–0.95
Mean Aiken's V	0.83125
Criterion threshold	≥ 0.78
Decision	40 items valid in content

Determination of the Research Instrument

The first prototype of the test contained 40 questions. Upon content validation, each question was found fit for content validity and a set of them was given to 30 students as a pilot test. The initial tryout results showed that only 32 questions passed the criterion requirements for the main experiment and the rest of them (the 8 questions) weren't kept. Because of that, the main experiment employed 32 multiple-choice questions with 103 respondents participating. From these steps one can recognize that the instrument was gradually finalized based on the selection process instead of being simply copied from the initial list. The removal of eight items after the tryout is consistent with evidence that empirical item statistics often reject items that have passed qualitative review; Purnama (2017), for instance, found that only 24 to 27 of 37 reviewed vocational accounting items met 2PL quality criteria. Possible explanations for the eight rejected items include ambiguous wording, content that had not yet been taught to all classes, and distractors that were implausible to students, but the small tryout sample does not allow these explanations to be separated.

Results of the Classical Test Theory (CTT) Analysis

CTT analysis of the 32 items across 103 respondents resulted in a split-half correlation of 0.91, which corresponds to a Spearman–Brown corrected reliability coefficient of 0.95. It can be said that the result represents exceptional internal consistency, since a reliability coefficient of 0.95 or high can be described as very satisfactory. Nevertheless, high reliability should be interpreted in conjunction with item features, as the fact that reliability coefficient is high does not mean by itself that difficulty level and distractor functions of all items are appropriate. Very high reliability can also signal item redundancy or a narrowly defined construct (Taber, 2018). Because the items in this test were drawn from seven content areas, the coefficient should be read together with the item-level evidence below rather than as proof that the items measure a single narrow skill. The value is considerably higher than that reported for a teacher-made mathematics test in Ghana (0.61; Butakor, 2022), which suggests that systematic blueprinting and expert review contributed to score consistency.

Table 3. Results of Item Analysis, Reliability, Validity, and Distractor Effectiveness

Aspect	Result	Interpretation
Number of respondents	103	Analysis subjects
Number of items	32	Multiple-choice
Split-half	0.91	Very high
Reliability	0.95	Very high
Item validity	30 highly significant; 2 significant	All items significant
Discrimination index	13 very good; 17 good; 2 fair	Predominantly good to very good
Distractors	19 good to very good; 13 flagged for review	Some require revision

Item Difficulty Level

As for the difficulty level distribution, one item was very easy, 19 items were easy, and 12 items were moderate. Therefore, 62.5% of the items were either easy or very easy, and none were difficult. It seems the test was quite easy for students. A similar pattern to this one has been reported by Purnama and Alfarisa (2020) who concluded that the difficulty-level distribution is an important factor when vocational tests are being evaluated. Two explanations are plausible. The skew may reflect the true ability of students who had already studied the Basic DPIB material, or it may be a measurement artifact produced by items written mostly at lower cognitive levels and by distractors that were easy to eliminate. The present data cannot fully separate these explanations, but the distractor findings reported below suggest that the second explanation plays at least a partial role.

Table 4. Item Difficulty-Level Distribution of the Test

Category	Number	Percentage
Very easy	1	3.1%
Easy	19	59.4%
Moderate	12	37.5%
Difficult	0	0.0%
Total	32	100%

Discrimination Index and Distractors

The discrimination value indicated that 13 items were at the top or very good level, 17 items were in the good range, and 2 items were just fair. All items were above the poor/negative discrimination level. Item 8 revealed the highest point-biserial correlation with a value of 0.766. In comparison, Item 1 and Item 30 had a relatively low discrimination value. This discovery implies that nearly all items successfully differentiated students with high and low ability. This profile is considerably more favorable than that reported for a teacher-made mathematics test in which 62.5% of the items showed weak or negative discrimination (Butakor, 2022), and it is in line with the finding of Awopeju and Afolabi (2016) that well-constructed examination items tend to show adequate discrimination under both CTT and IRT.

The result of the distractor analysis revealed that 19 test items were good enough to very good in terms of functioning of the distractors, whereas in 13 test items at least one choice was found that required further investigation. These comments were mostly about the answer options which were hardly ever picked or which instead appealed to the high-ability group. Distractor data is a plus of Classical Test Theory that you don't obtain directly from the 2PL model, therefore, its value still holds for item- revision decisions. Non-functioning distractors are common in multiple-choice tests (Tarrant et al., 2009) and have been shown to be associated with easier items and lower discrimination (Rezigalla et al., 2024). This link offers a direct, content-preserving way to raise the difficulty of the test: replacing implausible options with distractors based on common student misconceptions, as recommended in evidence-based distractor development (Gierl et al., 2017; Haladyna et al., 2002).

Results of the 2PL IRT Model Analysis

The 2PL IRT estimation of the parameters gave discrimination (a) and difficulty (b) for 32 items. No a value was negative, the minimum a value was 0.848 while the maximum was 4.833. Item 8 had the highest a value, namely 4.833. In Item 1 the minimum a value, 0.848, could be found. Regarding the b parameter, the values were between -2.597 and 0.034. The more negative the b value, the lower the ability required for a 50% probability of answering the item correctly, that is, the easier the item; higher b values indicate relatively more difficult items. Discrimination values above 4, such as that of Item 8, should be interpreted with caution because, in samples of about 100 respondents, large a estimates tend to be inflated and imprecise (König et al., 2020; Şahin & Anıl, 2017).

Table 5. Summary of 2PL IRT Item Parameter Estimates

Parameter	Minimum	Maximum	Notable items
a (discrimination)	0.848	4.833	Highest: Item 8; lowest: Item 1

Parameter	Minimum	Maximum	Notable items
b (difficulty)	-2.597	0.034	Easiest: Item 1; relatively most difficult: Item 10

Results of b values point out that most of the tested items cluster mainly in the lower level of item difficulty. For instance, Item 1 ($b = -2.597$) appeared to be the most accessible item whereas Item 10 ($b = 0.034$) was relatively the most demanding item in the test. Item 12 ($b = -0.006$), Item 27 ($b = -0.021$) and Item 32 ($b = -0.114$) were among the few items that are located around or near to the average item difficulty ($b = 0$). These findings align with those of Classical Test Theory (CTT) which has also mainly revealed easy and intermediate level items as the predominant item types in the sample test.

In the ICC, item curves typically rise towards the right direction. The slope of a curve is the a parameter whereas the shifting or the position of a curve relative to the ability scale is the b parameter. Item 8 has the steepest curve since it has the highest a parameter. On the contrary, Item 1 is located to the further most left side since it has the most negative value of b parameter. It can be inferred from the IIF and TIF that the distribution of the amount of information provided by the assessment was uneven throughout the ability spectrum. The TIF peaked at a theta of about -0.98 with the value being about 29.55, which implies that the measurement device was able to give most detailed feedback at the lower-to-moderate level of ability. Because the standard error of measurement equals $1/\sqrt{I(\theta)}$, the peak information corresponds to a standard error of about 0.18 on the θ scale, whereas precision decreases markedly for students above average ability. Information functions of this kind indicate where on the ability scale new items are needed (Zanon et al., 2016), and they express the same targeting problem that Rasch-based studies display as a mismatch between item and person distributions (Boone, 2016).

This result is significant as the information function depicts the accuracy of measurement at different ability levels, whereas CTT basically provides a summarization of the group's empirical scores. The IRT literature highlights that besides their usage in scoring, IRT models are also used to estimate item characteristics through difficulty and discrimination parameters and, in the end, they can give information in a detail way of the items-latent ability relation (Gyamfi & Acquaye, 2023). In addition, calibrated item parameters form the basis for building item banks and computerized adaptive tests, as demonstrated in the development of an IRT-based adaptive physics test in Indonesia (Istiyono et al., 2020).

Comparison of CTT and IRT

Table 6. Comparison of CTT and 2PL IRT Results for the 32-Item Test

Aspect	CTT (ANATES)	IRT 2PL (Stata)	Synthesis
Number of respondents	103	103	Same
Number of items	32	32	Same
Reliability	0.95	Not a primary 2PL parameter	CTT provides reliability information
Difficulty	1 very easy; 19 easy; 12 moderate	$b = -2.597$ to 0.034 ; mostly negative	Both indicate a tendency toward easy items
Relatively easiest item	Item 1; 86.41% correct	Item 1; $b = -2.597$	Consistent
Relatively most difficult item	Item 10; 46.60% correct	Item 10; $b = 0.034$	Consistent
Highest discrimination	Item 8; $rpb = 0.766$	Item 8; $a = 4.833$	Substantively consistent
Distractors	Directly analyzed	Not directly analyzed in 2PL	CTT more informative
Information by ability level	Not directly displayed	ICC, IIF, TIF, TCC	IRT more informative

The comparison indicates that CTT and IRT offer distinct yet mutually reinforcing features. CTT is especially good with diagnosing the results based on group responses, like reliability and distractor functions. The 2PL IRT item analysis gives the finer detail of how the items, in turn, position, and their ability to discriminate along the ability scale as well as the measurement precision via the information function. The main patterns about difficulty and the comparative level of discrimination were essentially consistent in both methods. This result is in keeping with the one of the findings of the research done by Nisfatulsanah and Sugiharto (2024) that CTT and IRT give different but complementary outcome in instrument-quality analysis. The convergence in item ordering agrees with Fan (1998) and Setiawati et al. (2023), who found strong correlations between CTT and IRT item statistics, and with Awopeju and Afolabi (2016), who reported comparable CTT- and IRT-based parameter estimates in a mathematics examination. The divergence in the kind of evidence supplied agrees with Jabrayilov et al. (2016), who showed that the advantages of IRT emerge mainly for individual-level decisions, and with Ayanwale et al. (2022), who argued that relying on CTT alone limits test development practice. The apparent tension between the epistemological assumptions of the two frameworks is resolved when each is linked to a different inference in the validity argument (Cook et al., 2015): CTT supports inferences about the

consistency of total scores and the functioning of answer options, whereas IRT supports inferences about the precision with which students at different ability levels are measured (Gyamfi & Acquaye, 2023; Kılıç et al., 2023).

The study suggests a main point is that item difficulty levels should be carefully considered. While reliability and discrimination were mostly satisfactory, the lack of difficult items results in the instrument not being the best way to differentiate the highest-level students. Items which could be of medium-to-high difficulty should be included in future development. Meanwhile, item with issues about the distractors should undergo checks on the wording, the quality of options, and the correct answer.

The main contribution of this study to vocational assessment practice therefore lies not in the individual statistics but in the decision rule they jointly produce. CTT is used to repair items, by revising keys and distractors, while IRT is used to decide where on the ability scale new items should be added. For the present instrument, this rule indicates that the test is currently suitable for identifying students who have not yet mastered basic DPIB competencies, for example to target remedial instruction before work-based learning, but not for ranking or certifying high-achieving students. This interpretation-and-use statement is the kind of explicit, evidence-based claim that argument-based validation requires (Cook et al., 2015), and it has seldom been made for vocational classroom tests (Purnama & Alfarisa, 2020).

Conclusions

This study produced a 32-item Basic DPIB test whose content validity was supported by expert judgment (mean Aiken's $V = 0.83$) and whose scores were highly consistent (split-half reliability = 0.95). Its main contribution lies in what the two psychometric frameworks jointly reveal about the intended use of the test. CTT and IRT converged on the easiest item (Item 1), the relatively most difficult item (Item 10), and the most discriminating item (Item 8), which indicates that the ordering of the items is robust to the choice of framework. The frameworks diverged in the evidence they supplied: CTT identified 13 items with poorly functioning distractors, which is directly actionable for item revision, whereas the 2PL information function showed that measurement precision peaks below average ability ($\theta \approx -0.98$). The test is therefore best suited to diagnosing students who have not yet mastered basic competencies, rather than to distinguishing high achievers. For vocational assessment practice, the study demonstrates a practical sequence in which CTT is used to repair items and IRT is used to decide where new items are needed on the ability scale. These conclusions are bounded by the limitations described below, chiefly the modest IRT calibration sample and the absence of difficult items. The next development cycle should add items targeted above average ability, revise the flagged distractors, and recalibrate the test with a larger, multi-school sample before it is used for high-stakes decisions. This study has several limitations. First, the 2PL parameters were estimated from 103 respondents, which is below the sample sizes usually recommended for stable 2PL calibration (Şahin & Anıl, 2017; Schroeders & Gnams, 2025); extreme discrimination estimates such as $a = 4.833$ for Item 8 may therefore be inflated, and the parameters should be re-estimated with a larger sample or with small-sample methods such as Bayesian hierarchical estimation (König et al., 2020). Second, the participants were drawn from a limited population of DPIB students, so the sampling frame restricts generalization to other schools and regions, and selection bias cannot be excluded. Third, the skewed difficulty distribution may reflect the true ability level of the students, a measurement artifact of items written at lower cognitive levels and of non-functioning distractors, or both; the present design cannot separate these explanations. Fourth, the 32 items were selected on the basis of a small tryout ($n = 30$), which may have capitalized on chance. Finally, the assumptions of unidimensionality and local independence, as well as model fit, should be examined formally in future calibrations, and alternative IRT models should be compared.

Acknowledgments

This paper would like to acknowledge the help of the validators, subject teachers, students, the school, and the Universitas Negeri Padang who have supported the implementation of this study.

References

- Aiken, L. R. (1985). Three coefficients for analyzing the reliability and validity of ratings. *Educational and Psychological Measurement*, 45(1), 131–142. <https://doi.org/10.1177/0013164485451012>
- Ariansyah, K., Wismayanti, Y. F., Savitri, R., Listanto, V., Aswin, A., Ahad, M. P. Y., & Cahyarini, B. R. (2024). Comparing labor market performance of vocational and general school graduates in Indonesia: Insights from stable and crisis conditions. *Empirical Research in Vocational Education and Training*, 16, 5. <https://doi.org/10.1186/s40461-024-00160-6>

- Arieli-Attali, M., Ward, S., Thomas, J., Deonovic, B., & von Davier, A. A. (2019). The expanded evidence-centered design (e-ECD) for learning and assessment systems: A framework for incorporating learning goals and processes within assessment design. *Frontiers in Psychology*, 10, 853. <https://doi.org/10.3389/fpsyg.2019.00853>
- Awopeju, O. A., & Afolabi, E. R. I. (2016). Comparative analysis of classical test theory and item response theory based item parameter estimates of senior school certificate mathematics examination. *European Scientific Journal*, 12(28), 263. <https://doi.org/10.19044/esj.2016.v12n28p263>
- 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>
- Bichi, A. A., & Talib, R. (2018). Item response theory: An introduction to latent trait models to test and item development. *International Journal of Evaluation and Research in Education*, 7(2), 142–151. <https://doi.org/10.11591/ijere.v7i2.12900>
- Boone, W. J. (2016). Rasch analysis for instrument development: Why, when, and how? *CBE—Life Sciences Education*, 15(4), rm4. <https://doi.org/10.1187/cbe.16-04-0148>
- Butakor, P. K. (2022). Using classical test and item response theories to evaluate psychometric quality of teacher-made test in Ghana. *European Scientific Journal*, 18(1), 139. <https://doi.org/10.19044/esj.2022.v18n1p139>
- Cook, D. A., Brydges, R., Ginsburg, S., & Hatala, R. (2015). A contemporary approach to validity arguments: A practical guide to Kane's framework. *Medical Education*, 49(6), 560–575. <https://doi.org/10.1111/medu.12678>
- Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement*, 58(3), 357–381. <https://doi.org/10.1177/0013164498058003001>
- Gierl, M. J., Bulut, O., Guo, Q., & Zhang, X. (2017). Developing, analyzing, and using distractors for multiple-choice tests in education: A comprehensive review. *Review of Educational Research*, 87(6), 1082–1116. <https://doi.org/10.3102/0034654317726529>
- Gyamfi, A., & Acquaye, R. (2023). Parameters and models of Item Response Theory (IRT): A review of literature. *Acta Educationis Generalis*, 13(3), 68–78. <https://doi.org/10.2478/atd-2023-0022>
- Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, 15(3), 309–333. https://doi.org/10.1207/S15324818AME1503_5
- Istiyono, E., Dwandaru, W. S. B., Setiawan, R., & Megawati, I. (2020). Developing of computerized adaptive testing to measure physics higher order thinking skills of senior high school students and its feasibility of use. *European Journal of Educational Research*, 9(1), 91–101. <https://doi.org/10.12973/eu-jer.9.1.91>
- Iwintolu, R. O., Opesemowo, O. A. G., & Adetutu, P. O. (2024). Effect of 2-PL and 3-PL models on the ability estimate in mathematics binary items. *Journal on Efficiency and Responsibility in Education and Science*, 17(3), 257–272. <https://doi.org/10.7160/eriesj.2024.170308>
- Jabrayilov, R., Emons, W. H. M., & Sijtsma, K. (2016). Comparison of classical test theory and item response theory in individual change assessment. *Applied Psychological Measurement*, 40(8), 559–572. <https://doi.org/10.1177/0146621616664046>
- Kania, N., Kusumah, Y. S., Dahlan, J. A., Nurlaelah, E., Gurbuz, F., & Bonyah, E. (2024). Constructing and providing content validity evidence through the Aiken's V index based on the experts' judgments of the instrument to measure mathematical problem-solving skills. *REID (Research and Evaluation in Education)*, 10(1), 64–79. <https://doi.org/10.21831/reid.v10i1.71032>
- König, C., Spoden, C., & Frey, A. (2020). An optimized Bayesian hierarchical two-parameter logistic model for small-sample item calibration. *Applied Psychological Measurement*, 44(4), 311–326. <https://doi.org/10.1177/0146621619893786>
- Kılıç, A. F., Koyuncu, İ., & Uysal, İ. (2023). Scale development based on Item Response Theory: A systematic review. *International Journal of Psychology and Educational Studies*, 10(1), 209–223. <https://doi.org/10.52380/ijpes.2023.10.1.982>
- Manurung, G. V., & Mulyana, R. (2024). Pengaruh penerapan metode AIR (Auditory, Intelektual, Repetisi) terhadap hasil belajar Dasar DPIB siswa SMK. *Jurnal Pendidikan Teknik Bangunan*, 4(2), 115–124. <https://doi.org/10.17509/jptb.v4i2.75533>
- Murphy, D. H., Little, J. L., & Bjork, E. L. (2023). The value of using tests in education as tools for learning—not just for assessment. *Educational Psychology Review*, 35, 89. <https://doi.org/10.1007/s10648-023-09808-3>
- Nabil, N. R. A., Wulandari, I., Yamtinah, S., Retno, S. D. A., & Ulfa, M. (2022). Analisis indeks Aiken untuk mengetahui validitas isi instrumen asesmen kompetensi minimum berbasis konteks sains kimia. *PAEDAGOGIA*, 25(2), 184–191. <https://doi.org/10.20961/paedagogia.v25i2.64566>

- Nisfatulsanah, A., & Sugiharto, B. (2024). Comparing classical test theory and item response theory for respiratory system question instruments. *PIONIR: Jurnal Pendidikan*, 13(3), 70–86. <https://doi.org/10.22373/pjp.v13i3.25435>
- Nurjanah, S., Iqbal, M., Zafrullah, Z., Mahmud, M. N., Seran, D. S. F., Suardi, I. K., & Arriza, L. (2024). Psychometric quality of multiple-choice tests under classical test theory (CTT): AnBuso, Iteman, and R. *Jurnal Penelitian dan Evaluasi Pendidikan*, 28(2), 161–172. <https://doi.org/10.21831/pep.v28i2.71542>
- Purnama, D. N. (2017). Characteristics and equation of accounting vocational theory trial test items for vocational high schools by subject-matter teachers' forum. *REID (Research and Evaluation in Education)*, 3(2), 152–162. <https://doi.org/10.21831/reid.v3i2.18121>
- Purnama, D. N., & Alfarisa, F. (2020). Karakteristik butir soal try out teori kejuruan akuntansi SMK berdasarkan teori tes klasik dan teori respons butir. *Jurnal Pendidikan Akuntansi Indonesia*, 18(1), 36–46. <https://doi.org/10.21831/jpai.v18i1.31457>
- Rezigalla, A. A., Eleragi, A. M. E. S. A., Elhussein, A. B., Alfaifi, J., ALGhamdi, M. A., Al Ameer, A. Y., Yahia, A. I. O., Mohammed, O. A., & Adam, M. I. E. (2024). Item analysis: The impact of distractor efficiency on the difficulty index and discrimination power of multiple-choice items. *BMC Medical Education*, 24, 445. <https://doi.org/10.1186/s12909-024-05433-y>
- Sacks, R., & Pikas, E. (2013). Building information modeling education for construction engineering and management. I: Industry requirements, state of the art, and gap analysis. *Journal of Construction Engineering and Management*, 139(11), 04013016. [https://doi.org/10.1061/\(ASCE\)CO.1943-7862.0000759](https://doi.org/10.1061/(ASCE)CO.1943-7862.0000759)
- Şahin, A., & Anıl, D. (2017). The effects of test length and sample size on item parameters in item response theory. *Educational Sciences: Theory & Practice*, 17(1), 321–335. <https://doi.org/10.12738/estp.2017.1.0270>
- Schroeders, U., & Gnams, T. (2025). Sample-size planning in item-response theory: A tutorial. *Advances in Methods and Practices in Psychological Science*, 8(1). <https://doi.org/10.1177/25152459251314798>
- Setiawati, F. A., Amelia, R. N., Sumintono, B., & Purwanta, E. (2023). Study item parameters of classical and modern theory of differential aptitude test: Is it comparable? *European Journal of Educational Research*, 12(2), 1097–1107. <https://doi.org/10.12973/eu-jer.12.2.1097>
- Suharno, Pambudi, N. A., & Harjanto, B. (2020). Vocational education in Indonesia: History, development, opportunities, and challenges. *Children and Youth Services Review*, 115, 105092. <https://doi.org/10.1016/j.childyouth.2020.105092>
- Taber, K. S. (2018). The use of Cronbach's alpha when developing and reporting research instruments in science education. *Research in Science Education*, 48(6), 1273–1296. <https://doi.org/10.1007/s11165-016-9602-2>
- Tarrant, M., Ware, J., & Mohammed, A. M. (2009). An assessment of functioning and non-functioning distractors in multiple-choice questions: A descriptive analysis. *BMC Medical Education*, 9, 40. <https://doi.org/10.1186/1472-6920-9-40>
- Zanon, C., Hutz, C. S., Yoo, H., & Hambleton, R. K. (2016). An application of item response theory to psychological test development. *Psicologia: Reflexão e Crítica*, 29, 18. <https://doi.org/10.1186/s41155-016-0040-x>