International Journal of Education and Practice
2025 Vol. 13, No. 1, pp. 327-344
ISSN(e): 2310-3868
ISSN(p): 2311-6897
DOI: 10.18488/61.v13i1.4060

Comparing two psychometric approaches: The case of item analysis for a classroom test in mathematics

Mark Lester B. Garcia[1]+
Kevin Carl P. Santos[2]
Catherine P. Vistro-Yu[3]

[1,3]Mathematics Department, Ateneo de Manila University, Philippines.
[1]Email: mark.garcia@student.ateneo.edu
[3]Email: cvistro-yu@ateneo.edu
[2]College of Education, University of the Philippines, Diliman, Philippines.
[2]Email: kpsantos1@up.edu.ph

(+ Corresponding author)

Article History

Received: 28 May 2024
Revised: 23 December 2024
Accepted: 7 January 2025
Published: 31 January 2025

Keywords

Classical test theory
Classroom assessment
Item analysis
Item response theory
Mathematics assessment
Psychometric approaches
Summative test.

ABSTRACT

Multiple-choice and constructed-response item formats are widely used in Philippine classrooms where tests are primarily summative. Conducting item analysis provides teachers the professional possibility to independently improve their test items while also allowing them to make informed decisions about test improvement and ensure that their assessments provide meaningful scores. Comparisons of item analysis approaches have been extensively used in standardized tests but not so much in small-scale testing such as classroom assessments. Hence, this research reports an empirical study where a summative test for classroom use is analyzed using psychometric techniques from the Classical Test Theory (CTT) and the Item Response Theory (IRT). This study employs an action research design where the primary author being a teaching practitioner-researcher, reports how an item analysis of his teacher-made summative test was conducted to improve the test items. Results from this study show that interpretations of the item parameter estimates from both psychometric approaches, CTT and IRT agree on most of the recommendations for the items. The items where the two approaches provide contrasting recommendations call for a further qualitative evaluation of the flagged items putting the teacher in a judicious position to discern which recommendations to follow regarding the concerned items. In the context of this study, an example of a guiding insight in deciding would be reflecting on the learning objectives measured by the test and the context in which the test was created.

Contribution/Originality: This empirical action research is about how item analysis using different psychometric approaches is feasible in a small-scale setting such as in a classroom assessment in the Philippine context. It demonstrates how a teacher-practitioner can make informed decisions about improving one's teacher-made test items geared towards the improvement of one's assessment practices and the teaching-learning cycle in the context of mathematics test items.

1. INTRODUCTION

1.1. Classroom Assessments, Item Quality and Test Validity

In classroom assessments, test development is crucial in bridging the gap between student progress and curricular goals. Assessment is an integral part of the educational process that is concerned with the measurement and appraisal of student achievement, testing becomes instrumental in measuring the knowledge and skills that students are supposed to acquire (Ghaicha, 2016). Hence, classroom teachers rely heavily on tests not only to gauge and monitor learning but also to provide feedback to students about mastery of the lesson. Additionally, results from tests that are concerned with the performance of the class help inform the teacher about his or her teaching, completing the teaching-learning feedback loop. In test development, evaluating item quality can serve as a tool that provides an item writer (or the teacher in the case of classroom assessments) a clear image of how well-constructed the test items are in terms of measuring student learning. Evaluating the quality of test items can be done prior to and after test administration and sometimes even during the test especially in research settings. An ante hoc appraisal of such items includes expert review while post hoc analysis of the items is conducted using statistical analysis based on item responses (Albano & Rodriguez, 2018). Statistical analysis in the form of item analysis is utilized in improving both the validity and reliability of assessment tools. One important use of validity is that it reinforces the legitimacy of scores and grades (Brown & Abdulnabi, 2017).

Ghaicha (2016) upholds that classroom assessments must follow test development, test validation and test evaluation models that are theoretically founded and must also be aligned with quality control criteria. Brown and Abdulnabi (2017) warn of the consequences of poorly performing items as they "lead to inappropriate decisions about student ability and consequent decisions" (p. 2); hence, quality assurance of test items is warranted as it ensures that test scores are meaningful and accurate. Item analysis as a means of evaluating item quality aims to identify items to be reused, revised, or discarded from the item bank which ultimately leads to improved item quality and test validity (Haladyna & Rodriguez, 2021). Validity is all about the extent to which evidence on student learning supports test score interpretations for their intended uses (Plake, Huff, Reshetar, Kaliski, & Chajewski, 2016). Validity is considered one of the foundations of testing and is deemed to be the most fundamental consideration in making tests along with reliability and fairness (American Educational Research Association, 2014).

1.2. Item Analysis through Classical Test Theory and Item Response Theory

Martinková and Drabinová (2018) strongly recommend that routine psychometric analysis must be enforced in test development. Two influential theories in the field of psychometrics are the Classical Test Theory (CTT; Crocker and Algina (1986)) and the Item Response Theory (IRT). CTT is the traditional psychometric model that assumes that a test-taker's score is interpreted as a sum of his or her true ability and an unobserved measurement error (Brown & Abdulnabi, 2017). On the other hand, IRT is otherwise known as the latent trait theory. It assumes that an examinee possesses a certain quantity of the latent variable or construct being measured (Berezner & Adams, 2017). One of the main differences between the two theories is the type of test model used. As CTT assumes that test scores on a single scale characterize linearity, IRT assumes the opposite. This means that if students A, B, C and D obtained the scores 9, 8, 3 and 2 respectively in a 10-point test, the difference between students A and B's abilities is not considered the same as those of students C and D. IRT also expressed the probability that a student answers an item correctly as a function of the examinee's ability and the item's characteristics.

Aside from the difference in the use of linear models (in CTT) and nonlinear models (in IRT) (Hambleton & Jones, 1993), CTT and IRT largely differ in their assumptions behind measurement precision. According to Jabrayilov, Emons, and Sijtsma (2016) a common estimate of the measurement precision in CTT is assumed to be equal for all individuals regardless of their abilities whereas in IRT the measurement precision depends on the latent-attribute value. The two theories also differ in terms of the perspectives they offer on the relationship between test items and examinee abilities. IRT has the capacity to measure and provide estimates for examinee ability and item difficulty as parameters and on the same scale- a feat that CTT is unable to do (Hori, Fukuhara, & Yamada, 2022) because it does not examine item-ability relationships. Another difference lies in the presumption behind test items where they are not expected to perform in the same way across different tests and across different test-taker sample properties which are termed item invariance and person invariance respectively (Wu, Tam, & Jen, 2016). This difference was highlighted in Hambleton and Jones's (1993) comparison of CTT and IRT where it was stated that item and person statistics are sample dependent in CTT which is not the case for IRT as long as the model fits the data. Both CTT and IRT have item statistics (or item parameters) for difficulty and discrimination. The difficulty index *p* in CTT is obtained by computing the proportion of examinees who answered the item correctly whereas the difficulty parameter *b* in IRT is estimated based on a probabilistic function such that a given value of $b = 0.2$ means that approximately 50% of examinees whose proficiency is 0.2 will be able to answer the item correctly (DeMars, 2010). As for the item discrimination estimates, CTT uses item total correlation *r* whereas IRT represents this as the *a*-parameter which is often referred to as the slope because it shows the relationship between the rise in probability of a correct response given any increase in the examinee's proficiency (ibid.).

1.3. IRT Models for Dichotomously Scored and Polytomously Scored Items

Despite the use of nonlinear models in IRT, these are later on linearized in such a way that examinee scores (or abilities) as well as item difficulty estimates are measured in log-odd units or logits and expressed as a continuous variable θ on the same scale. Thus, two pairs of data points on the θ scale that are of the same distance will reflect equal differences in logits (abilities or difficulties). IRT models vary in terms of the number of parameters being considered in each probabilistic function in computing for probability $P(\theta)$ that an examinee of ability θ will be able to answer a specified item correctly. For instance, the three-parameter logistic (or 3PL) model (Birnbaum, 1968) for dichotomously scored items takes into account the following three parameters: the two previously mentioned parameters which are item discrimination *a*, item difficulty *b*, and an additional parameter which is the guessing parameter *c*, as can be seen in Equation 1 (DeMars, 2010).

$$\mathrm{P}(\theta) = \mathrm{c}_i + (1 - \mathrm{c}_i)\frac{\mathrm{e}^{1.7\mathrm{a}_i(\theta - \mathrm{b}_i)}}{1 + \mathrm{e}^{1.7\mathrm{a}_i(\theta - \mathrm{b}_i)}} \qquad (1)$$

Apart from the 3PL model, another model for dichotomously scored items is the two-parameter logistic or 2PL model (Birnbaum, 1968) which is a constrained version of 3PL in the sense that *c* is set to zero, and in effect only considers discrimination and difficulty parameters (DeMars, 2010). More constrained versions of 2PL model are the one-parameter logistic or 1PL model and the Rasch model (Rasch, 1960) which only consider the difficult parameter but differ in their assumptions behind the discrimination parameter as the Rasch model sets *a* to 1 for all items while the 1PL model only assumes the *a* values to be equal without necessarily being equal to 1 (Kirkeby, 2019) although these models are mathematically equivalent (Bandalos & DeMars, 2018). As for polytomously scored items or items with more than two available response categories, IRT models such as the Graded Response Model (GRM; Samejima (1969)), Partial Credit Model (PCM; Masters (1982)) and Generalized Partial Credit Model (GPCM; Muraki (1992)) are used. Samejima's GRM is an extension of the 2PL model to polytomous items (Auné, Abal, & Attorresi, 2019) while Masters' PCM and Muraki's GPCM both belong to the family of Rasch models for polytomous items except that the discrimination parameters in GPCM are freely estimated (Paek & Cole, 2019).

1.4. Rationale for the Study

Classroom assessment (CA) is a growing field of research that has earned its right to own a separate identity in educational assessments. McMillan (2013) argued that CA has been identified as a research base and is "the most powerful type of measurement in education that influences student learning" (p. 4) serving as a tool to gather information needed in making well-supported inferences about student learning. Teacher judgment on student learning at the end of the teaching-learning cycle depends on the quality of assessments conducted (Moss, 2013) and ensuring assessment quality in the classroom is called for given that classroom assessments are the most common type of assessments used in educational systems (Ghaicha, 2016). Given the advances in the field of psychometrics and the rise of IRT and its extensive applications in educational assessment, a move to further widen its use to the confines of CA remains irresolute but is warranted.

Unexamined test item quality can lead to negative unintended consequences in assessment. Item analysis in CA is necessary as flawed items are found to be prevalent among teacher-made tests resulting in misleading insights about student performance (Brown & Abdulnabi, 2017). This pressing need is amplified as the use of multiple-choice items in assessments is common due to the ease and objectivity in scoring (Geisinger & Usher-Tate, 2016). In the Philippine context where assessments in classrooms have been found mostly summative (Griffin, Cagasan, Care, Vista, & Nava, 2016). Between CTT and IRT item analysis, the former remains the widely used and established procedure in scrutinizing test item quality post-test administration. Despite CTT having sufficient statistics to evaluate items (Brown & Abdulnabi, 2017) IRT has an advantage in terms of providing more information about test items making it a preferable scoring model (Faulkner-Bond & Wells, 2016). The trade-off is that CTT item analysis is simpler and intuitive compared to the theoretically complicated IRT method that requires complex statistical methods (ibid.).

1.5. Research Aims and Questions

Knowledge of various tools in item analysis can enable the teacher to make more informed decisions about improving the quality of his or her test items. This study compares CTT and IRT techniques in item analysis to prepare the way for the potential integration of both in the classroom because IRT models can be difficult for teachers to understand, especially as they have important teaching responsibilities. The outcomes of these statistical analyses can offer not only feedback on the quality of the items but also an opportunity to consider why certain items may perform well or poorly and how to improve those for the subsequent administration cycle without depending on external feedback. This study finds it imperative to answer the following research questions given the case laid out in the introduction:

1. What are the similarities and differences in the item analysis results when using CTT and IRT in a classroom assessment on polynomial functions?
2. What insights can be learned from item analysis in CTT and IRT that can be applied by a teacher in improving classroom assessments?

2. LITERATURE REVIEW

The range of applications of item analysis spans across different levels of assessments. It is used in internal assessments or those within institutions, external assessments such as standardized tests administered by external organizations, high-stakes tests, and even large-scale assessments (Berezner & Adams, 2017). Several studies have employed item analysis in math assessments in basic education such as assessments of grade school students' skills in whole number multiplication (Chai, Pang, & Chin, 2018) multiplicative reasoning (Johnson et al., 2018) operations on integers (Nurnberger-Haag, Kratky, & Karpinski, 2022) and even high-stakes mathematics assessments of grade 10 students (Okitowamba, 2015).

Other studies have also explored other IRT models for item analysis in tests including those that consist of mixed-item formats (tests with both multiple-choice items and constructed-response items). Chon, Lee, and Ansley (2007) determined the IRT model combination of best fit to perform an item analysis on the Iowa Tests of Basic Skills for reading comprehension and math with 500 examinees in the U.S. The model combinations formed from the dichotomous models (e.g., 1PL, 2PL, 3PL) and polytomous models (e.g., GRM, PCM, GPCM) were analyzed with the help of item fit statistics. Similar to Chon et al.'s (2007) study was that of Galli, Chiesi, and Primi (2011) who developed a scale to measure the mathematical ability needed by college students in Italy who are non-math majors for an introductory course in statistics. They used item responses from 788 psychology students and applied the 1PL, 2PL and 3PL models to responses in a 30-item test consisting of only multiple-choice questions. Yılmaz (2019) also performed similar comparisons among IRT model combinations to assess their fit in analyzing item responses to a mixed-format science test for elementary students in Turkey. The 25-item test designed for grade 4 students was taken by 2,351 examinees and was analyzed using combinations formed from 1PL/2PL/3PL models and GRM/GPCM. Some of the more recent studies involving item analysis have employed IRT approaches for varied purposes and were conducted on different scales. At the course level, Abdellatif (2023) showed how teacher-made test items for a medical course in Oman were analyzed using the Rasch model with the goal of comparing the psychometric analysis results of two tests with or without a test blueprint. Similarly, Arriza, Retnawati, and Ayuni (2024) reported mathematics test item analysis results from comparing the Rasch, 2PL and 3PL models on responses by grade 11 students in Indonesia to gauge which model depicted a better fit with the items compared to the rest of the models. Meanwhile, Akveld and Kinnear (2023) applied item analysis at the university level where they made a comparative item analysis of two mathematics diagnostic tests from two universities – one in the U.K. and another in Switzerland, for the purpose of improving the said tests so these can provide better information about students' abilities. IRT item analysis applications were also used in broader scopes such as the national level or educational system level as seen in the works of Tsigilis, Krousorati, Gregoriadis, and Grammatikopoulos (2023) and Fajobi (2024). Tsigilis et al. (2023) examined the factorial validity and measurement invariance of a US-based test – the Preschool Early Numeracy Skills Test–Brief Version, in the Greek educational context using the 2PL model in analyzing the responses of 906 preschool children in Greece to 20 items on numbering relations and arithmetic operations. According to Fajobi's (2024) work, a comparison of the psychometric qualities of the multiple-choice mathematics items constructed by the Nigerian national and West African regional examination councils for secondary school students was conducted and the results showed that the two sets of items were comparable.

Some studies performed both CTT and IRT item analysis on the same set of examinee responses. Such a study was conducted by Magno (2009) on a chemistry test for junior high school students in the Philippines (n = 219), where he compared the difficulty estimates, internal consistency, and measurement errors for the CTT and IRT approaches. A more recent study by Ayanwale, Adeleke, and Mamadelo (2018), parallel to Magno's (2009) work, applied CTT and IRT item analysis to an assessment with a much larger scope. Item statistics were estimated for math test items in a national basic education certificate examination taken by 978 examinees in Nigeria.

These studies have shown different possible applications of CTT and IRT analysis for a wide range of assessments across different subject or content areas. However, most of these used commercially available software such as PARSCALE (Muraki & Bock, 1997) IRTPRO (Item Response Theory for Patient-Reported Outcomes) (Cai, Du Toit, & Thissen, 2011) and jMetrik (Meyer, 2014) in performing item analysis. Among those that incorporated IRT item analysis, only one of the existing IRT models has been deliberately chosen by the researchers without any discussion on how the IRT model was selected. Other studies that were previously mentioned claimed that other IRT models are superior in terms of providing a better fit while the Rasch model seemed to be a very practical and convenient model to use. It can also be noted that large sample sizes and a high number of test items were used but these simply do not apply to classroom settings specifically in teacher-made summative tests given that the class size may be smaller and the test may be shorter in length. Finally, there are mixed results when it comes to estimating IRT parameters separately or simultaneously for dichotomous and polytomous items while some of these studies show that there are more mismatched items when analyzed from a CTT perspective as compared to IRT. A separate calibration entails estimating item parameters for items of different formats (e.g., multiple-choice items and constructed-response items) in two separate analyses whereas a simultaneous calibration computes item parameters in a test regardless of item format all at once.

3. METHODOLOGY

3.1. Theoretical Framework

The design of this study is backed by Evidence-Centered (assessment) Design or ECD which is a test development strategy that considers test score interpretability (Mislevy, Steinberg, & Almond, 2003). The overall goal of this study is to apply CTT and IRT item analysis to the same set of teacher-made summative test items for classroom use, compare the results and draw insights that may prove useful for teachers who actively seek ways to link their assessment practices with theoretical and research-based perspectives. This essentially provides evidence of test validity if the test items are found to be of acceptable quality post-analysis. As for those items that have been identified otherwise, they provide feedback to the teacher regarding their level of quality for his or her future reference. This application of ECD is desirable because it provides valuable evidence supporting test score interpretations for intended uses (Plake et al., 2016). ECD is not limited to a specific type of assessments as it is a principled assessment design focusing on evidentiary arguments for various types of assessments including classroom tests (Shute, Leighton, Jang, & Chu, 2016). This study aims to focus on the backend activity or component of assessment which is rarely carried out by teachers because they are already busy with other assessment-related tasks such as grading papers, computing grades and providing feedback on students' learning. However, there are many aspects to ECD including domain analysis, domain modelling and the development of the assessment framework.

3.2. Research Design and Data Collection Procedure

This study has an exploratory qualitative research design – more specifically action research where the lead author is positioned as a teacher practitioner-researcher. Being practitioner-based research that is essentially and necessarily self-reflexive (Cain, 2011) action research is a practical and systematic method that is useful in investigating one's teaching (Nolen & Putten, 2007) and improving teaching practice, and in turn students' learning outcomes (Professional Learning and Leadership Development Directorate, New South Wales Department of Education and Training [DET], 2010 as cited in Scanlon (2018)). Action research forges stronger action-reflection as well as theory-practice ties that provide practical solutions to pressing issues (Bradbury, 2015; as cited in Bradbury, Lewis, and Embury (2019)). It is also an inquiry process that promotes self-reflectiveness and can be done at a micro level within an individual class with the classroom as a common starting point (Bradbury et al., 2019). This action research simply explores the lead author's quest in conducting an item analysis that banks on theoretical foundations to improve the test items he has created for a summative test in a classroom setting. Thus, the variables inherently examined are the item parameters namely item difficulty and item discrimination which were estimated using CTT and IRT approaches.

The data set consists of the item responses of 90 junior high school examinees to a summative test consisting of 15 multiple-choice items and five constructed-response items created by the same teacher. These students were enrolled in one of the specialized public high schools in the Philippines. On the other hand, the item writer is a high school math teacher with 10 years of teaching experience, having taught math subjects ranging from elementary algebra to differential calculus. The test covers an entire chapter on polynomial functions as part of a course on advanced algebra, where students are expected to identify their characteristics, describe and find zeros, graph a given function, analyze a given graph and apply various theorems related to the characteristics of polynomial functions. The duration of the discussion on this topic is estimated to be 12 learning hours which includes time spent on formative assessments. Additionally, each multiple-choice (MC) item has four response options, only one of which is the correct answer.

3.3. Data Analysis

All calculations related to item indices and parameter estimates were done with the aid of R which is free open-source statistical software with packages for different computational purposes including psychometric packages. Among the existing and available R packages for item analysis, the mirt package (Chalmers, 2012) was used for IRT item analysis. It primarily uses marginal maximum likelihood estimation as an estimation method and offers flexible parameter estimation features (Chalmers, 2015). The mirt package also performs better compared to other

IRT packages such as sirt (Robitzsch, 2022) and TAM (Robitzsch, Kiefer, & Wu, 2022) in terms of performance in point estimate and standard error recovery, and in terms of program running time where it had the shortest estimation time for dichotomous models (Kim & Paek, 2017). The R instructions outlined in the book by Desjardins and Bulut (2018) were implemented for the CTT item analysis.

Separate and simultaneous calibrations of item statistics and parameters were conducted using R for the determination of the IRT model to be used. The dichotomous models considered were only Rasch, 1PL, and 2PL as 3PL had to be excluded due to the large sample size required which is around 500 which is a highly unlikely scenario to occur in a high school classroom setting. Meanwhile, the polytomous models used were GRM, PCM, and GPCM. The item statistics used in determining model fit are as follows: the Akaike Information Criterion (AIC; Akaike (1974)) values, the Bayesian Information Criterion (BIC; Schwarz (1978)) values, the Root Mean Square Error of Approximation (RMSEA), and M2 which is a limited information test statistic by Maydeu-Olivares and Joe (2006).

4. RESULTS AND DISCUSSION

4.1. Selection of IRT Model

Running the item response data set in R using the mirt package reveals the item fit statistics AIC, BIC, RMSEA and M2. A benchmark for determining acceptable fit using AIC and BIC values is having lower values as those with higher AIC and BIC values are indicative of more complex models (Kirkeby, 2019). According to Brown and Abdulnabi (2017) models whose AIC values differ by more than 10 indicate that the model with the smaller AIC value has superior fit to the data. Similarly, the acceptable range of values for RMSEA is any value less than 0.06 (Revicki, Chen, & Tucker, 2015). Tables 1 and 2 below show the results for separate and simultaneous calibration respectively along with the 95% confidence intervals for RMSEA. The results of the separate calibration show that the Rasch model is the best-fit model for dichotomous items while 1PL is the worst fit model due to the extremely large AIC and BIC values and an RMSEA above the baseline value of 0.06. As a result, the 1PL model has been excluded from the simultaneous calibration.

Table 1. Model fit statistics for separate calibration of MC and CR items.

Format	Model	AIC	BIC	RMSEA	95% CI	M2
MC	Rasch	1248.680	1286.177	0.000	(0, 0.054)	87.759
MC	1PL	127567.400	127637.400	0.275	Not available (NA)	
MC	2PL	1263.469	1333.463	0.000	(0, 0.056)	74.437
CR	GRM	655.100	692.597	0.000	NA	Error[1]
CR	PCM	655.552	683.050	0.000	(0, 0.012)	0.721
CR	GPCM	655.782	693.280	0.000	NA	Error[1]

Note: [1]Computing the M2 statistic using the mirt package in R displays this error message: “Error: M2 statistic cannot be calculated due to too few degrees of freedom”.

Table 2. Model fit statistics for simultaneous calibration of MC and CR items.

Combinations	AIC	BIC	RMSEA	95% CI	M2
Rasch + GRM	1885.414	1957.908	0.040	(0, 0.065)	183.756
Rasch + PCM	1881.129	1943.624	0.034	(0, 0.061)	183.368
Rasch + GPCM	1885.856	1958.351	0.040	(0, 0.065)	183.749
2PL + GRM	1897.829	2005.320	0.033	(0, 0.061)	161.564
2PL + PCM	1895.834	1990.826	0.033	(0, 0.061)	166.733
2PL + GPCM	1898.312	2005.804	0.034	(0, 0.062)	162.017

Table 1 shows that the Rasch model is the best-fitting model for dichotomous items. Among the polytomous items, which have roughly the same AIC values, the PCM turned out to be the model of best fit. The M2 statistic did not seem to be a reasonable statistic for comparison, as some calculations in R produced errors in the GRM and GPCM. Regarding the comparison of the model combinations in the simultaneous calibration, Rasch + PCM yielded the lowest values for AIC and BIC, making it the overall best-fitting model. As results from the separate calibration endorse the use of Rasch and PCM, the Rasch + PCM model combination was used as the final IRT model.

4.2. Results from CTT and IRT Item Analyses

The item indices and parameters in the succeeding tables were computed in R. Table 3 shows the difficulty and discrimination indices computed using CTT approach while Table 4 shows the *a*- and *b*-parameters computed using the IRT approach.

Table 3. Summary of item indices in CTT item analysis.

Items	Diff.	Disc.	Items	Diff.	Disc.
MC01	0.844	0.416	MC11	0.311	0.391
MC02	0.578	0.288	MC12	0.733	0.506
MC03	0.811	0.459	MC13	0.933	0.353
MC04	0.789	0.218	MC14	0.767	0.305
MC05	1.000	NA	MC15	0.367	0.337
MC06	0.689	0.400	CR01	0.972	0.232
MC07	0.711	0.418	CR02	0.906	0.330
MC08	0.956	0.203	CR03	0.528	0.641
MC09	0.767	0.460	CR04	0.722	0.557
MC10	0.900	0.211	CR05	0.422	0.636

Table 4. Summary of item parameters in IRT item analysis.

Items	*a*	*b*	Items	*a*	*b*	b_1	b_2
MC01	1	-1.906	MC11	1	0.908	NA	NA
MC02	1	-0.361	MC12	1	-1.153	NA	NA
MC03	1	-1.649	MC13	1	-2.919	NA	NA
MC04	1	-1.496	MC14	1	-1.353	NA	NA
MC05	Error		MC15	1	0.626	NA	NA
MC06	1	-0.909	CR01	1	NA	-2.008	-3.640
MC07	1	-1.029	CR02	1	NA	-0.548	-2.945
MC08	1	-3.367	CR03	1	NA	0.281	-0.481
MC09	1	-1.353	CR04	1	NA	-0.418	-1.251
MC10	1	-2.452	CR05	1	NA	-0.404	1.144

After the item indices or parameters were computed, items that failed to meet acceptable item indices or parameters were flagged for further inspection. Robitzsch's (2022) scoping review of literature offers guidelines regarding difficulty and discrimination indices for CTT item analysis where the discrimination index (*D*) must fall within the range $D > 0.2$ while the difficulty index (*P*) must fall within the range $0.3 < P < 0.7$. Since CR items are worth more than 1 point each, the item indices for Constructed-Response (CR) items were adjusted by scaling. This was done by dividing each item index by the maximum score so that the resulting indices fall in the interval 0 to 1, as seen in Table 3. In both CTT and IRT item analyses, R failed to compute the item statistics (hence displayed an error message) of item MC05 as all test-takers were able to answer this item correctly.

In the CTT item analysis, recommended actions for the given item discrimination indices are interpreted as follows: 0 to 0.19 – reject, 0.2 to 0.39 – revise, and 0.4 to 1.0 – retain (Ebel & Frisbie, 1991; Wu & Adams, 2007). As for the item difficulty indices, those in the interval 0 – 0.25 were considered difficult, those in 0.26 – 0.75 were considered moderately difficult and those in 0.76 and above were easy. Obon and Rey (2019) (as cited in Robitzsch (2022)) recommend that moderately difficult items be retained while difficult and easy items be considered for revision or rejection.

For the IRT item analysis, the *a*-parameters were automatically set to 1 based on the underlying assumptions behind the Rasch model and the PCM, which is a Rasch model extended to polytomous items. Since there are three response categories for the CR items (i.e., 2, 1, 0), then there are two *b*-parameters (labeled as b_1 and b_2) for every CR item. Thus, the *b*-values are not applicable (labeled as NA) to CR items while the b_1- and b_2-values are not applicable to MC items. DeMars (2010) states that *b*-parameters ranging from -2 to 2 indicate fairly acceptable difficulty parameters for the intended test population, while De Ayala (2022) considers difficulty parameters in the interval 0.8 to 2.5 as good values.

The researchers cautiously identified indices that are either beyond or critically on the borderline values in flagging items to arrive at recommendations on retaining, revising or rejecting the specific item based on the recommendations from the aforementioned references. The summary of recommended actions based on the computed indices or parameters is summarized in the decision matrix below.

Table 5. Decision matrix on flagging items based on CTT and IRT item indices or parameters.

CTT item discrimination index	Recommended decision	CTT item difficulty index	Recommended decision	IRT item difficulty parameter	Recommended decision
Less than 0.2	Reject	At most 0.25	Revise or reject	At most -2	Revise or reject
0.2 to 0.4	Revise	Between 0.25 and 0.75	Retain	Between -2 and 2	Retain
Greater than 0.4	Retain	At least 0.75	Revise or reject	At least 2	Revise or reject

It can be noted that some items will be flagged for revision or replacement based on either or both the difficulty and discrimination indices upon applying the guidelines for CTT item analysis in Table 5 to the items based on the results in Table 3. In instances where either index recommends revision but the other is within acceptable range (e.g., MC02, MC15), the item would be retained as it is. For items where both indices recommend revision (e.g., MC04, MC10), the course of action for such items would be to revise them. But for items where one index recommends revision and the other recommends replacement (e.g., MC08, CR01), the final recommendation for CTT item analysis would be to consider revising the item. As for the IRT item analysis, items that have difficulty parameters in the interval -3 to -2 or 2 to 3 (e.g., MC10, CR02), the recommendation for these items would be to revise them, while for those with extreme parameters that are less than -3 or greater than 3, they would be flagged for replacement. Table 6 shows the summary of the recommendations for each item in both CTT and IRT item analyses.

Table 6. Summary of recommendations for all items based on CTT and IRT item analyses.

Items	CTT diff.	CTT disc.	CTT	IRT	Items	CTT diff.	CTT disc.	CTT	IRT
MC01	Revise	*	*	*	MC11	*	Revise	*	*
MC02	*	Revise	*	*	MC12	*	*	*	*
MC03	Revise	*	*	*	MC13	Reject	Revise	Revise	Reject
MC04	Revise	Revise	Revise	*	MC14	Revise	Revise	Revise	*
MC05	Reject	Reject	Reject	Reject	MC15	*	Revise	*	*
MC06	*	*	*	*	CR01	Reject	Revise	Reject	Reject
MC07	*	*	*	*	CR02	Revise	Revise	Revise	Revise
MC08	Reject	Revise	Reject	Reject	CR03	*	*	*	*
MC09	Revise	*	*	*	CR04	*	*	*	*
MC10	Revise	Revise	Revise	Revise	CR05	*	*	*	*

Note: * = Retain.

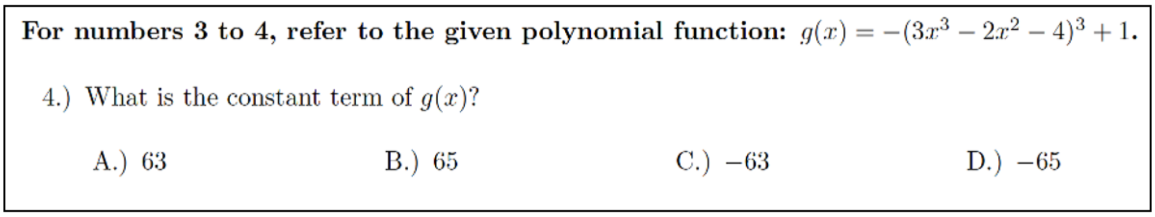
Consistent and inconsistent recommendations both arise for the flagged items when the results from CTT and IRT item analyses are placed side by side. There are a total of eight flagged items from among the 20 items, and CTT and IRT item analyses mostly agree on their recommendations. However, there is a contrast between the recommendations for three of the flagged items (MC04, MC13, and MC14). The recommendations from both approaches for the flagged items are summarized in Table 7.

Table 7. Comparison of recommendations for the flagged items.

Flagged items	CTT	IRT	Recommendations
MC04	Revise	Retain	Disagree
MC05	Reject	Reject	Agree
MC08	Reject	Reject	Agree
MC10	Revise	Revise	Agree
MC13	Revise	Reject	Disagree
MC14	Revise	Retain	Disagree
CR01	Reject	Reject	Agree
CR02	Revise	Revise	Agree

4.3. Implications from the Results

Items where CTT and IRT item analyses offer different recommendations certainly provide an opportunity for reflection to the teacher or item writer on how to proceed with the final item analysis. Further qualitative evaluation of the items is still much warranted while the differing recommendations do not necessarily clash (for instance, revising versus retaining the item). Figures 1 to 3 show such items (MC04, MC13 and MC14).

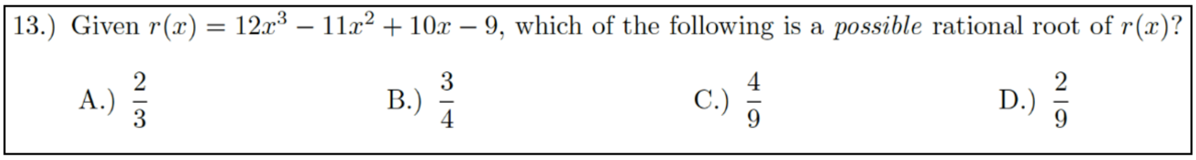


For numbers 3 to 4, refer to the given polynomial function: $g(x) = -(3x^3 - 2x^2 - 4)^3 + 1$.

4.) What is the constant term of $g(x)$?

A.) 63 B.) 65 C.) -63 D.) -65

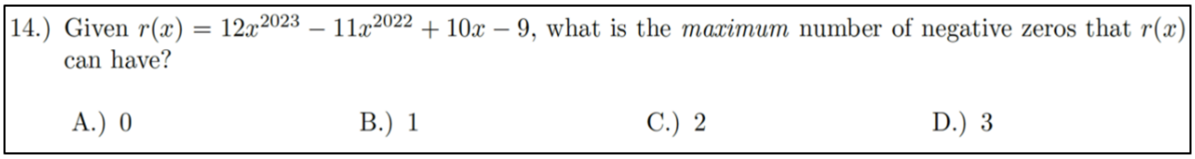
Figure 1. Item MC04



13.) Given $r(x) = 12x^3 - 11x^2 + 10x - 9$, which of the following is a *possible* rational root of $r(x)$?

A.) $\frac{2}{3}$ B.) $\frac{3}{4}$ C.) $\frac{4}{9}$ D.) $\frac{2}{9}$

Figure 2. Item MC13



14.) Given $r(x) = 12x^{2023} - 11x^{2022} + 10x - 9$, what is the *maximum* number of negative zeros that $r(x)$ can have?

A.) 0 B.) 1 C.) 2 D.) 3

Figure 3. Item MC14

Each of these items measures different skills related to polynomial functions which are described as follows:

- Item MC04: Computing the constant term of a polynomial function given that the function is not written in expanded form (which is useful in determining the tail-end behavior of the graph without having to graph the function).
- Item MC13: Applying the rational root theorem in finding possible rational roots given a polynomial function.
- Item MC14: Applying the Descartes' rules of signs in finding the maximum number of negative zeros of a given polynomial function.

As CTT and IRT item analyses are merely tools that produce item statistics, it is ultimately up to the teacher or item-writer to make the final decision regarding the retention, revision, or rejection of the concerned test items. Thus, on a personal level, he or she is urged to reflect critically on the skills tested by such items in relation to the lesson objectives and curricular objectives of the department or school. For instance, MC04 may be retained, as it is the only item that measures the skill of identifying the constant term of polynomial functions that are written in non-expanded form (combination of sum form and exponential form). However, a much more important (but still related) skill that may be tested in lieu of this is identifying the tail-end behavior of the graph of *g(x)* and hence can be a point for improvement upon revising the item, although this skill is also tested in MC07 (the polynomial in the function is expressed in product form). The case is different with MC13 where CTT prescribes that it be revised while IRT recommends replacing it. Rejecting the item would mean looking for an alternative that still measures the skill of identifying possible rational roots. However, it may also be revised in such a way that the response options are sets of possible rational roots instead of single numerical values which provides more insight as to whether the student is able to identify all possible rational roots of a given polynomial function.

4.4. Comparisons of Results with those from Related Studies

As a summary of the results of this study, the key findings are as follows: (1) In terms of the selection of the IRT model to be used in the item analysis, it was found that among the combinations of dichotomous and polytomous IRT models, the Rasch + PCM combination emerged as the most suitable model based on the AIC, BIC, RMSEA, and M2 values during the separate and simultaneous calibrations of the MC and CR items. (2) Among the 20 items in the mathematics summative test, 8 of these have been flagged for revision or replacement. The agreement rate among the recommendations by the CTT and IRT approaches is fairly high at 62.5% (five out of eight items). This means that out of the 20 items in the test, 15% (three out of 20 items) require further attention and scrutiny from the teacher.

The results from this study seemingly deviate from the related studies that have been mentioned. For instance, Galli et al.'s (2011) findings revealed that among the IRT models for dichotomous items, the 2PL model was the most suitable model to analyze the scale that they developed as compared to the 1PL and 3PL models. However, results from this study concur with findings from Arriza et al.'s (2024) study where they found that the Rasch model had the best fit compared to the 2PL and 3PL models although their approach in model fit testing was to compute the significance values of the probabilities associated with the chi-squares of the items' IRT logistic parameters. Chon et al. (2007) results showed that the 2PL/3PL model combined with any polytomous model had fewer misfitting items than the 1PL model combined with any polytomous model and further concluded that the overall best model was 2PL/3PL + GPCM as for the findings related to IRT model combinations. Yılmaz (2019) concluded that the combination 3PL + GRM produced the best item fit statistics.

Finally, in terms of the comparison of the CTT and IRT approaches in item analysis, Magno (2009) found that significantly more items mismatched in terms of their difficulty in CTT compared to IRT, with no mismatched items. On the other hand, after comparing both approaches, Ayanwale et al. (2018) found that (a) the CTT approach classified 33 of the 60 items as poor, while IRT classified only 12 of the items as poor, thus depicting a much lower agreement rate; and (b) the mean differences between item parameters were statistically significant.

5. CONCLUSION

One key finding that emerged from the results of this empirical action-research study is that the CTT and IRT approaches mostly agree on the recommendations on the items based on their psychometric qualities namely difficulty and discrimination. Item parameter estimates only provide a guide on further action points to be done but it is ultimately up to the teacher to rationalize such actions. In the case of this study, the results enabled the teacher practitioner-researcher to review the items against the learning objectives measured by the test and items as well as the skills that the test-takers were expected to demonstrate. The teacher has the agency to reconcile the differing recommendations of the item analysis and act upon what he or she believes serves the best interest of the future test-takers keeping in mind greater concerns such as item validity and overall test validity given that the recommendations by the item analyses results are not definitive. Iterations of the current assessment produce data that not only give a depiction of students' learning outcomes but also provide teachers an image of how effective their pedagogy and curriculum materials are at present (Mellati & Khademi, 2018). The second key finding shows that all of the items warranting the teacher's attention are multiple-choice items. This concretizes how constructing multiple-choice items may be more difficult given that the writer of the item has to pay attention to the response options on top of the other item components as well as how the distractors compete with the correct response in filtering students who are able to demonstrate the skill required by that certain item.

The teacher may check for redundant items or check the effectiveness of the distractors for multiple-choice items upon further reflection of the test items. Identifying redundant items can help the teacher improve the test reliability by weeding out extra items that do not measure a skill different from those measured by the rest. Additionally, since teachers do not work in isolation, it is also highly recommended that the item writer discuss his or her findings with other co-teachers who are making tests for the same subject matter in the hopes of discovering potential common trends among item structures and examinee attributes that possibly contribute to the poor functioning of the items.

This will also allow teachers to develop both generic and teacher-specific solutions to improve the quality of individual items and their tests.

The findings of this paper suggest that in terms of practice and policy recommendations, teachers must be given opportunities to advance technical knowledge in their assessment practices as item analysis requires comprehensive understanding of statistical tools and models. Applying such tools in their practice will enable them to be independent educational practitioners who are able to pinpoint problematic items in their own tests and rationalize possible factors behind them. Institutions may thus consider departmental initiatives such as establishing a professional development (PD) program that promotes teachers' self-examination of test results after administration with the ultimate goal of improving their test items founded on theoretical knowledge and backed by research. Such a process must be driven by intent, purpose, structure, and goals, most especially in improving and upskilling teaching and assessment practices as professional development empowers mathematics teachers by encouraging one's active involvement and sense of ownership in personal growth (Roesken, 2011).

For future studies, a deeper level that will aid in the decision-making process of item development of MC items would be differential distractor functioning (DDF). An analysis of DDF allows one to evaluate the effectiveness of the distractors based on statistical thresholds. Moreover, it is recommended that future studies investigate the viability of and the effect of carrying out CTT and IRT item analyses for smaller sample sizes, especially since class sizes in high schools are generally small.

Funding: This study received no specific financial support.
Institutional Review Board Statement: The Ethical Committee of the Ateneo de Manila University, Philippines has granted approval for this study on 30 March 2022 (Ref. No. SOSEREC_22_001).
Transparency: The authors state that the manuscript is honest, truthful, and transparent, that no key aspects of the investigation have been omitted, and that any differences from the study as planned have been clarified. This study followed all writing ethics.
Competing Interests: The authors declare that they have no competing interests.
Authors' Contributions: All authors contributed equally to the conception and design of the study. All authors have read and agreed to the published version of the manuscript.

REFERENCES

Abdellatif, H. (2023). Test results with and without blueprinting: Psychometric analysis using the Rasch model. *Educación Médica*, *24*(3), 100802. https://doi.org/10.1016/j.edumed.2023.100802

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716-723. https://doi.org/10.1109/TAC.1974.1100705

Akveld, M., & Kinnear, G. (2023). Improving mathematics diagnostic tests using item analysis. *International Journal of Mathematical Education in Science and Technology*, 1-28. https://doi.org/10.1080/0020739x.2023.2167132

Albano, A. D., & Rodriguez, M. C. (2018). Item development research and practice in S. N. Elliott, R. J. Kettler, P. A. Beddow, & A. Kurz (Eds.), Handbook of accessible instruction and testing practices. In (pp. 181–198): Springer International Publishing. https://doi.org/10.1007/978-3-319-71126-3_12.

American Educational Research Association. (2014). *American psychological association & national council on measurement in education standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Arriza, L., Retnawati, H., & Ayuni, R. T. (2024). Item analysis of high school specialization mathematics exam questions with Item response theory approach. *Barekeng: Journal of Mathematics and Its Application*, *18*(1), 151–162. https://doi.org/10.30598/barekengvol18iss1pp0151-0162

Auné, S. E., Abal, F. J. P., & Attorresi, H. F. (2019). Application of the graded response model to a scale of empathic behavior. *International Journal of Psychological Research*, *12*(1), 49–56. https://doi.org/10.21500/20112084.3753

Ayanwale, M. A., Adeleke, J. O., & Mamadelo, T. I. (2018). An assessment of item statistics estimates of basic education certificate examination through classical test theory and item response theory approach. *International Journal of Educational Research Review*, *3*(4), 55-67.

Bandalos, D. L., & DeMars, C. E. (2018). Item response theory in D. L. Bandalos, measurement theory and applications for the social sciences. In (pp. 403-445): New York: Guilford Press.

Berezner, A., & Adams, R. J. (2017). Why large-scale assessments use scaling and item response theory in P. Lietz, J. C. Cresswell, K. F. Rust, & R. J. Adams (Eds.), Implementation of large-scale education assessments. In (1st ed., pp. 323–356): Hoboken, NJ: Wiley Online Library.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability in F. M. Lord & M. R. Novick (Eds.), Statistical theories of mental test scores. In (pp. 392-479). Reading, MA: Addison-Wesley.

Bradbury, H., Lewis, R., & Embury, D. C. (2019). Education action research: With and for the next generation. In C. A. Mertler (Ed.), The Wiley handbook of action research in education. In (1st ed., pp. 5–28): Hoboken, NJ: Wiley Blackwell.

Brown, G. T. L., & Abdulnabi, H. H. A. (2017). Evaluating the quality of higher education instructor-constructed multiple-choice tests: Impact on student grades. *Frontiers in Education*, *2*, 1-12. https://doi.org/10.3389/feduc.2017.00024

Cai, L., Du Toit, S. H. C., & Thissen, D. (2011). *IRTPRO: Flexible, multidimensional, multiple categorical IRT modeling computer software*. Chicago, IL: Scientific Software International.

Cain, T. (2011). Teachers' classroom-based action research. *International Journal of Research & Method in Education*, *34*(1), 3-16. https://doi.org/10.1080/1743727X.2011.552307

Chai, C. P., Pang, V., & Chin, K. E. (2018). *Using rasch analysis to examine the effects of Year 5 students' understanding of whole numbers multiplication*. Paper presented at the In Q. Zhang (Ed.), Pacific Rim Objective Measurement Symposium (PROMS) 2016 Conference Proceedings (pp. 227–242). Springer Singapore.

Chalmers, R. (2015). *Multidimensional item response theory workshop in R presentation slides York University*. Retrieved from https://philchalmers.github.io/mirt/extra/mirt-Workshop-2015_Day-1.pdf

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

Chon, K. H., Lee, W. C., & Ansley, T. N. (2007). *Assessing IRT model-data fit for mixed format tests (Report No. 26)*. Iowa, IA: Center for Advanced Studies in Measurement and Assessment.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Holt: Rinehart and Winston.

De Ayala, R. J. (2022). *The theory and practice of item response theory* (2nd ed.). New York: The Guilford Press.

DeMars, C. (2010). *Item response theory*: New York: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195377033.001.0001.

Desjardins, C. D., & Bulut, O. (2018). *Handbook of educational measurement and psychometrics using R*. Boca Raton: Chapman and Hall/CRC Press.

Ebel, R., & Frisbie, D. (1991). *Essentials of educational measurement* (5th ed.). New Jersey: Prentice-Hall.

Fajobi, O. O. (2024). Examining the psychometric properties of WAEC and NECO mathematics items for Nigerian secondary schools using a 3-parameter logistic model. *Unizik Journal of Educational Research and Policy Studies*, *17*(3), 302–314.

Faulkner-Bond, M., & Wells, C. S. (2016). A brief history of and introduction to item response theory in C. S. Wells, M. Faulkner-Bond (Eds.), Educational measurement from foundations to future. In (pp. 107-125). New York London: The Guilford Press.

Galli, S., Chiesi, F., & Primi, C. (2011). Measuring mathematical ability needed for "non-mathematical" majors: The construction of a scale applying IRT and differential item functioning across educational contexts. *Learning and Individual Differences*, *21*(4), 392-402. https://doi.org/10.1016/j.lindif.2011.04.005

Geisinger, K. F., & Usher-Tate, B. J. (2016). A brief history of educational testing and psychometrics in C. S. Wells, M. Faulkner-Bond (Eds.), Educational measurement from foundations to future. In (pp. 3-20). New York: The Guilford Press.

Ghaicha, A. (2016). Theoretical framework for educational assessment: A synoptic review. *Journal of Education and Practice*, *7*(24), 212-231.

Griffin, P., Cagasan, L., Care, E., Vista, A., & Nava, F. (2016). Formative assessment policy and its enactment in the Philippines in D. Laveault & L. Allal (Eds.), Assessment for learning: Meeting the challenge of implementation. In (Vol. 4, pp. 75–92): Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-39211-0_5.

Haladyna, T. M., & Rodriguez, M. C. (2021). Using full-information item analysis to improve item quality. *Educational Assessment*, *26*(3), 198-211. https://doi.org/10.1080/10627197.2021.1946390

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, *12*(3), 38-47.

Hori, K., Fukuhara, H., & Yamada, T. (2022). Item response theory and its applications in educational measurement Part I: Item response theory and its implementation in R. *Wiley Interdisciplinary Reviews: Computational Statistics*, *14*(2), e1531. https://doi.org/10.1002/wics.1531

Jabrayilov, R., Emons, W. H., & 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

Johnson, H. L., Tzur, R., Hodkowski, N. M., Jorgensen, C., Wei, B., Wang, X., & Davis, A. (2018). *A written, large-scale assessment measuring gradations in students' multiplicative reasoning in E. Bergqvist, M. Österholm, C. Granberg, & L. Sumpter (Eds.).* Paper presented at the Proceedings of the 42nd Conference of the International Group for the Psychology of Mathematics Education (Vol. 3, pp. 163-170). Umeå, Sweden: PM.

Kim, T., & Paek, I. (2017). A comparison of item parameter and standard error recovery across different R packages for popular unidimensional IRT models in L. A. Van Der Ark, M. Wiberg, S. A. Culpepper, J. A. Douglas, & W.-C. Wang (Eds.), Quantitative psychology, the 81st annual meeting of the psychometric society. In (pp. 421–430). Asheville, North Carolina: Springer International Publishing.

Kirkeby, K. (2019). *Gender differences on the revised sociosexual orientation inventory: A differential item functioning analysis.* Master's Thesis, Ball State University.

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.

Martinková, P., & Drabinová, A. (2018). Shiny item analysis for teaching psychometrics and to enforce routine analysis of educational tests. *The R Journal*, *10*(2), 503-515. https://doi.org/10.32614/RJ-2018-074

Masters, G. N. (1982). A rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149-174. https://doi.org/10.1007/BF02296272

Maydeu-Olivares, A., & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika*, *71*(4), 713-732. https://doi.org/10.1007/s11336-005-1295-9

McMillan, J. H. (2013). Why we need research on classroom assessment in J. H. McMillan (Ed.), SAGE handbook of research on classroom assessment. In (pp. 3–16): Thousand Oaks, CA: Sage Publications, Inc. https://doi.org/10.4135/9781452218649.n1.

Mellati, M., & Khademi, M. (2018). Exploring teachers' assessment literacy: Impact on learners' writing achievements and implications for teacher development. *Australian Journal of Teacher Education (Online)*, *43*(6), 1-18. https://doi.org/10.14221/ajte.2018v43n6.1

Meyer, J. P. (2014). *Applied measurement with jMetrik*. New York: Routledge.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). Focus article: On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, *1*(1), 3-62. https://doi.org/10.1207/S15366359MEA0101_02

Moss, C. M. (2013). Research on classroom summative assessment in J. H. McMillan (Ed.), SAGE handbook of research on classroom assessment. In (pp. 235–255): Thousand Oaks, CA: Sage Publications. https://doi.org/10.4135/9781452218649.n14.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*(2), 159-176. https://doi.org/10.1177/014662169201600206

Muraki, E., & Bock, R. D. (1997). *PARSCALE 3: IRT based item analysis and test scoring for rating-scale data computer program*. Chicago, IL: Scientific Software International.

Nolen, A. L., & Putten, J. V. (2007). Action research in education: Addressing gaps in ethical principles and practices. *Educational Researcher*, *36*(7), 401-407. https://doi.org/10.3102/0013189X07309629

Nurnberger-Haag, J., Kratky, J., & Karpinski, A. C. (2022). The integer test of primary operations: A practical and validated assessment of middle school students' calculations with negative numbers. *International Electronic Journal of Mathematics Education*, *17*(1), em0667. https://doi.org/10.29333/iejme/11471

Obon, A. M., & Rey, K. A. M. (2019). Analysis of multiple-choice questions (MCQs): Item and test statistics from the 2nd year nursing qualifying exam in a university in Cavite, Philippines. *Abstract Proceedings International Scholars Conference*, *7*(1), 499–511. https://doi.org/10.35974/isc.v7i1.1128

Okitowamba, O. (2015). *Tracking learners' performances in high-stakes grade 10 mathematics examinations*. Doctoral Dissertation University of the Western Cape, Bellville, South Africa.

Paek, I., & Cole, K. (2019). *Using R for item response theory model applications*. London: Routledge.

Plake, B. S., Huff, K., Reshetar, R. R., Kaliski, P., & Chajewski, M. (2016). Validity in the making: From evidenced-centered design to the validations of the interpretations of test performance in C. S. Wells, M. Faulkner-Bond (Eds.), Educational measurement from foundations to future. In (pp. 62-73). Washington, DC: The Guilford Press.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.

Revicki, D. A., Chen, W. H., & Tucker, C. (2015). Developing item banks for patient-reported health outcomes in S. P. Reise and D. A. Revicki (Eds.), Handbook of item response theory modeling. In (pp. 334-363). New York: Routledge.

Robitzsch, A. (2022). *Sirt: Supplementary item response theory models [Computer software manual]*. Retrieved from https://CRAN.R-project.org/package=sirt

Robitzsch, A., Kiefer, T., & Wu, M. (2022). *TAM: Test analysis modules [Computer software manual]*. Retrieved from https://CRAN.R-project.org/package=TAM

Roesken, B. (2011). Mathematics teacher professional development in B. Roesken, Hidden dimensions in the professional development of mathematics teachers. In (pp. 1–28): Sense Publishers. https://doi.org/10.1007/978-94-6091-433-1_1.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, *34*(S1), 1–97. https://doi.org/10.1007/BF03372160

Scanlon, L. (2018). *The role of research in teachers' work: Narratives of classroom action research*. London: Routledge.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461-464. https://doi.org/10.1214/aos/1176344136

Shute, V. J., Leighton, J. P., Jang, E. E., & Chu, M.-W. (2016). Advances in the science of assessment. *Educational Assessment, 21*(1), 34–59. https://doi.org/10.1080/10627197.2015.1127752

Tsigilis, N., Krousorati, K., Gregoriadis, A., & Grammatikopoulos, V. (2023). Psychometric evaluation of the preschool early numeracy skills test–brief version within the item response theory framework. *Educational Measurement: Issues and Practice, 42*(2), 32-41. https://doi.org/10.1111/emip.12536

Wu, M., & Adams, R. (2007). *Applying the rasch model to psycho-social measurement: A practical approach melbourne*. Australia: Educational Measurement Solutions.

Wu, M., Tam, H. P., & Jen, T.-H. (2016). *Educational measurement for applied researchers: Theory into practice*. Singapore: Springer Nature Singapore Pte Ltd.

Yılmaz, H. B. (2019). A comparison of IRT model combinations for assessing fit in a mixed format elementary school science test. *International Electronic Journal of Elementary Education, 11*(5), 539–545.

APPENDIX

Appendix 1 presents the summative test on polynomial functions which was analyzed using the CTT and IRT approaches.

Appendix 1. The summative test used in this study.

A.) MULTIPLE CHOICE. Write the CAPITAL LETTER corresponding to the correct answer in each item. (1 point each)

1.) Which of the following is an example of a polynomial function?

A.) $f(x) = x^3 + 2^x - 4$ C.) $f(x) = x^7 + 3x^6 - 5^4$
B.) $f(x) = x^{-5} + 6x^4 - 8x^3$ D.) $f(x) = x^x - x^2 - 12$

2.) The following statements are always true about a polynomial function, **EXCEPT** one. Which one is it?

A.) It has exactly one *y-intercept*. C.) Its *y-intercept* is the constant term.
B.) Its *range* is the set of real numbers. D.) Its *leading coefficient* is nonzero.

For numbers 3 to 4, refer to the given polynomial function: $g(x) = -(3x^3 - 2x^2 - 4)^3 + 1$.

3.) What is the leading term of g(x)?

A.) $-9x^9$ B.) $-9x^{27}$ C.) $-27x^9$ D.) $-27x^{27}$

4.) What is the constant term of g(x)?

A.) 63 B.) 65 C.) -63 D.) -65

For numbers 5 to 7, please refer to the given function: $h(x) = -3x(1 - x)^6(x + 2)^5(x - 3)^7$.

5.) The following are zeros of h(x), **EXCEPT** one. Which one is it?

A.) 0 B.) 1 C.) -2 D.) -3

6.) At what point will the graph of h(x) be tangent to the x-axis?

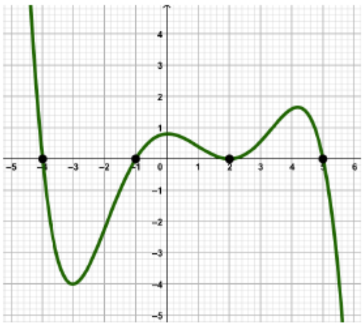
A.) (0, 0) B.) (1, 0) C.) (-2, 0) D.) (-3, 0)

7.) Which describes the tail-end behavior of the graph of h(x)?

A.) left tail down, right tail up C.) both tails up

B.) left tail up, right tail down D.) both tails down

For numbers 8 to 11, please refer to the given graph of a polynomial function Q(x) below.



8.) Which of the following statements is **NOT** true about the multiplicity of each zero?

A.) -4 has an odd multiplicity. C.) 2 has an even multiplicity.

B.) -1 has an odd multiplicity. D.) 5 has an even multiplicity.

9.) The degree of Q(x) is *m*. What is the *least* possible value of *m*?

A.) 4 B.) 5 C.) 6 D.) 7

10.) In which of the following intervals is $Q(x) < 0$?

A.) $x < -4$ B.) $-4 < x < -1$ C.) $-1 < x < 2$ D.) $2 < x < 5$

11.) Given: Q(x) is a *quintic* function whose leading coefficient is *k*. If $Q(-3) = -4$, find the value of *k*.

A.) $k = -\frac{1}{100}$ B.) $k = -\frac{1}{25}$ C.) $k = -\frac{1}{20}$ D.) $k = -\frac{1}{10}$

12.) If $p(x) = x^{100} + 2x^{99} - x^2 - 2x$, which of the following is **NOT** a factor of p(x)?

A.) $x + 2$ B.) $x + 1$ C.) $x - 1$ D.) $x - 2$

13.) Given $r(x) = 12x^3 - 11x^2 + 10x - 9$, which of the following is a *possible* rational root of r(x)?

A.) $\frac{2}{3}$ B.) $\frac{3}{4}$ C.) $\frac{4}{9}$ D.) $\frac{2}{9}$

14.) Given $r(x) = 12x^{2023} - 11x^{2022} + 10x - 9$, what is the *maximum* number of negative zeros that r(x) can have?

A.) 0 B.) 1 C.) 2 D.) 3

15.) Suppose that f(x) and g(x) are polynomial functions such that f(x) is an odd function, while g(x) is an even function. Which of the following results to an odd function?

A.) $x \cdot f(x)$ B.) $x \cdot g(x)$ C.) $f(g(x))$ D.) $g(f(x))$

B.) Do as indicated. (3 points each)

1.) Given $P(x) = 4x^5 - x^3 - 32x^2 + 8$:

a.) Prove that $(2x - 1)$ is a factor of $P(x)$.

b.) Prove that 4 is an upper bound.

c.) Find all zeros of $P(x)$.

d.) Sketch the graph of $P(x)$.

2.) Let $h(x) = 2x^3 + kx^2 - k^2x$, where k is a nonzero constant. When $h(x)$ is divided by $(x + 2)$ or by $(x - 1)$, the remainder is the same. Find all values of k.

Views and opinions expressed in this article are the views and opinions of the author(s), International Journal of Education and Practice shall not be responsible or answerable for any loss, damage or liability etc. caused in relation to/arising out of the use of the content.