Development of Four-Tier Diagnostic Test (FTDT) To Assess Readiness of Incoming Grade 11 Students in General Mathematics

Leonard G. Finez[1], Allen E. Pasia[2]

Laguna State Polytechnic University, San Pablo City, Laguna, Philippines[1]

Dr. Apolonio M. Lirio National High School, Tanauan City, Batangas, Philippines[2]

DOI: https://doi.org/10.47772/IJRISS.2026.1026EDU0373

Received: 27 May 2026; Accepted: 01 June 2026; Published: 22 June 2026

ABSTRACT

Readiness in General Mathematics is a critical factor in ensuring students' successful transition from Junior High School to Senior High School. However, traditional assessments often focus solely on the correctness of answers and fail to identify students' reasoning, confidence, and misconceptions. This study aimed to develop a Four-Tier Diagnostic Test (FTDT) integrated with the Certainty of Response Index (CRI) to assess the readiness of incoming Grade 11 students in General Mathematics. Specifically, it sought to establish the psychometric properties of the instrument and analyze students' readiness through multidimensional diagnostic classifications.

A descriptive-developmental research design was employed involving 100 Grade 10 students from Dr. Apolonio M. Lirio National High School. The instrument underwent expert validation and pilot testing. Data were analyzed using the Content Validity Index (CVI), Rasch Model, Readiness Index (RI), and frequency and percentage distribution.

Results showed that the FTDT demonstrated good content validity, with an S-CVI/Ave of 0.81. Rasch analysis revealed high person reliability for both the knowledge tier (0.883) and reasoning tier (0.884), indicating that the instrument could effectively differentiate students according to readiness levels. The Readiness Index revealed that 55% of students exhibited low readiness, 42% demonstrated moderate readiness, and only 3% achieved high readiness. Furthermore, the FTDT-CRI classification identified misconceptions, false negative responses, false positive responses, and lack of knowledge, highlighting significant gaps in prerequisite competencies.

The findings suggest that the FTDT integrated with CRI is a valid and reliable diagnostic instrument for assessing readiness in General Mathematics. The instrument provides a comprehensive assessment of students' knowledge, reasoning, and confidence, which may assist teachers in identifying learning gaps and designing targeted instructional interventions and bridging programs.

Keywords: Diagnostic assessment, Certainty of Response Index, Four-Tier Diagnostic Test, General Mathematics readiness, Rasch Model, Readiness Index

INTRODUCTION

Education plays a vital role in preparing learners for higher-level academic work and real-world problem-solving. In mathematics education, this preparation becomes particularly important as students transition to Senior High School, where they are expected to demonstrate readiness in prerequisite competencies necessary for General Mathematics.

Mathematical readiness extends beyond the ability to recall procedures or provide correct answers. It involves conceptual understanding, reasoning skills, and the ability to apply knowledge in different contexts. However, traditional assessments often focus solely on correctness and provide limited information about students'

reasoning, confidence, and misconceptions. Consequently, teachers may find it difficult to identify learning gaps and design appropriate interventions.

Diagnostic assessment offers a more comprehensive approach by evaluating learners' prior knowledge, strengths, and weaknesses before formal instruction begins. One promising diagnostic tool is the Four-Tier Diagnostic Test (FTDT), which consists of answer selection, confidence in the answer, reasoning selection, and confidence in reasoning. This structure enables the identification of misconceptions, partial understanding, false positive responses, false negative responses, and lack of knowledge, providing a deeper understanding of students' readiness.

Given the need for valid and reliable readiness assessments in General Mathematics, this study focuses on the development and validation of a Four-Tier Diagnostic Test (FTDT) integrated with the Certainty of Response Index (CRI). By incorporating correctness, reasoning, and confidence, the instrument aims to provide a comprehensive assessment of incoming Grade 11 students' readiness and support data-driven instructional planning and intervention.

LITERATURE REVIEW

Diagnostic Assessment in Mathematics Education

Studies have consistently shown that readiness in mathematics is a multidimensional construct involving conceptual understanding, procedural fluency, reasoning skills, and confidence. Research by Kilpatrick et al. (2001), Carlson et al. (2010), and Duncan et al. (2007) emphasized that mastery of prerequisite competencies significantly influences students' success in advanced mathematics. Furthermore, studies have demonstrated that diagnostic assessments provide valuable information about learners' strengths, weaknesses, and misconceptions, enabling teachers to design targeted interventions before formal instruction begins. These findings highlight the importance of developing comprehensive diagnostic tools that measure multiple dimensions of mathematical readiness rather than relying solely on correct answers.

Mathematical Readiness and Prerequisite Competencies

Research suggests that readiness extends beyond content mastery and includes learners' ability to apply knowledge, reason mathematically, and demonstrate confidence in their responses. Studies have shown that students with strong conceptual understanding and procedural fluency perform better in higher-level mathematics. In the Philippine context, several studies reported persistent misconceptions in algebra and deficiencies in higher-order thinking skills among learners. These findings underscore the need for readiness assessments that capture cognitive and affective dimensions of learning to provide a more accurate representation of students' preparedness for General Mathematics.

Four-Tier Diagnostic Test (FTDT)

The Four-Tier Diagnostic Test (FTDT), developed by Caleon and Subramaniam (2010), was designed to improve traditional multiple-choice assessments by incorporating answer selection, confidence rating, reasoning selection, and confidence in reasoning. Numerous studies have demonstrated its effectiveness in identifying misconceptions, distinguishing true understanding from guessing, and providing detailed diagnostic information. Research conducted in mathematics and science education reported high validity and reliability of FTDT instruments, making them useful tools for assessing conceptual understanding and diagnosing learning difficulties. The FTDT has also been recognized for its ability to classify learners according to levels of understanding, misconceptions, and uncertainty, thereby providing richer diagnostic information than conventional assessments.

Certainty of Response Index (CRI)

The Certainty of Response Index (CRI), introduced by Hasan et al. (1999), measures learners' confidence in their responses. Studies have shown that CRI enhances diagnostic assessment by distinguishing between

correct answers based on genuine understanding and those resulting from guessing. When combined with diagnostic tests, CRI improves the identification of misconceptions, partial understanding, and lack of knowledge. Its integration provides additional information about students' cognitive and metacognitive processes, leading to more accurate assessment results. In mathematics education, CRI has been widely used to determine the degree of certainty associated with students' conceptual understanding and to identify areas requiring instructional intervention.

Misconceptions in Mathematics Learning

Misconceptions are persistent and incorrect understandings that influence how students interpret mathematical concepts and solve problems. Research indicates that misconceptions commonly arise when students rely heavily on procedures without fully understanding the underlying concepts. Diagnostic assessments have been found to be effective in identifying these misconceptions, particularly when reasoning and confidence measures are included. Identifying misconceptions allows teachers to implement focused interventions that promote conceptual change and improve mathematical understanding. Studies further suggest that misconceptions in foundational mathematical concepts can hinder students' readiness for more advanced topics, making early diagnosis essential for successful learning progression.

Rasch Analysis in Educational Measurement

Rasch analysis has become an important approach in educational research for evaluating the psychometric properties of assessment instruments. Developed from Item Response Theory (IRT), the Rasch Model transforms ordinal raw scores into interval-level measures, allowing for more precise estimation of both item difficulty and person ability. Researchers have utilized Rasch analysis to examine item fit, person fit, dimensionality, reliability, separation indices, and rating scale functioning. Unlike classical test theory, Rasch analysis provides sample-independent item estimates and item-independent person measures, enhancing the objectivity and generalizability of assessment results.

Several studies have highlighted the usefulness of Rasch analysis in mathematics education, particularly in validating diagnostic and achievement tests. Through fit statistics such as Infit Mean Square (MNSQ), Outfit MNSQ, and standardized Z-values, researchers can identify problematic items that do not function as intended. Rasch analysis also enables the construction of Wright Maps, which visually compare item difficulty and respondent ability on a common scale. These features make Rasch analysis a valuable tool for ensuring that assessment instruments accurately measure the intended constructs and provide meaningful diagnostic information.

Validation of Assessment Instruments

The development of quality assessment instruments requires evidence of validity and reliability. Content validity is commonly established through the Content Validity Index (CVI), while construct validity and reliability may be examined using Rasch analysis. The Rasch Model provides objective measurement by evaluating item difficulty, person ability, fit statistics, and reliability indices. Previous studies have demonstrated that Rasch analysis is an effective method for validating educational instruments because it provides more precise and invariant measurement compared to traditional approaches. These advantages make Rasch analysis particularly suitable for validating diagnostic instruments such as the FTDT. In addition, Rasch-generated reliability and separation indices help determine whether an instrument can effectively distinguish among different levels of learner ability and item difficulty, thereby strengthening the overall quality of the assessment.

METHODOLOGY

This study employed a descriptive-developmental research design. The descriptive component was utilized to describe the psychometric properties of the instrument and the readiness levels of the respondents, while the developmental component involved the construction, validation, pilot testing, and refinement of the Four-Tier Diagnostic Test (FTDT). The study focused on developing a diagnostic instrument capable of assessing

incoming Grade 11 students' readiness in General Mathematics through the integration of the Certainty of Response Index (CRI).

The respondents of the study were 100 Grade 10 students from Dr. Apolonio M. Lirio National High School selected through convenience sampling. The sample served as the pilot-testing group for the validation of the FTDT. While previous Rasch studies suggest that a sample size of 100 may provide stable item estimates for pilot testing, the findings should be interpreted within the context of a single-school sample and may not be generalized to all incoming Grade 11 students.

Prior to administration, the FTDT underwent a series of validation procedures. Face validation was conducted to evaluate the clarity, readability, and appropriateness of the items. Content validation was then performed by expert validators to determine the alignment of the items with the intended competencies and learning objectives. Revisions were made based on the validators' comments and recommendations. After validation, the revised instrument was subjected to pilot testing among the selected respondents to evaluate its psychometric properties.

The FTDT consisted of 50 items covering prerequisite competencies in General Mathematics, including algebra, functions, rational expressions, geometry, and statistics and probability. Each item was structured into four tiers: (1) answer selection, (2) confidence in the answer, (3) reasoning selection, and (4) confidence in the reasoning. The content of the items was developed based on the Grade 10 Mathematics curriculum and the prerequisite competencies identified in the General Mathematics curriculum guide. Distractors and reasoning options were constructed using findings from previous studies on mathematical misconceptions and common student errors reported in the literature. A Table of Specifications (TOS) was prepared to ensure alignment between the competencies and the test items.

Table 1. Distribution of FTDT Items by Competency

Competency Area	Number of Items
Algebra	10
Functions	10
Rational Expressions	10
Geometry	10
Statistics and Probability	10
Total	50

Permission to conduct the study was secured from the appropriate authorities before data collection. The validated FTDT was administered to the respondents, and their responses were collected, encoded, and organized for analysis. The gathered data were then subjected to statistical treatment to determine the validity, reliability, and diagnostic capability of the instrument.

The data were analyzed using the Content Validity Index (CVI) to establish content validity. Rasch analysis was employed to examine item difficulty, person ability, fit statistics, and reliability of the FTDT. The Readiness Index (RI) was used to determine the readiness levels of the respondents, while frequency and percentage distributions were utilized to classify students according to the FTDT-CRI diagnostic categories and describe the overall response patterns.

RESULTS

Content Validation of the FTDT

The CVI was used as a standard for establishing the content validity of the FTDT (i.e.., the ratings from seven raters) were utilized in this process by using the CVI to validate the content validity of the FTDT

Table 2. Scale-Level Content Validity Index (S-CVI)

Index	Value	Interpretation
S-CVI/Ave	0.81	Good

The FTDT obtained an S-CVI/Ave of 0.81, indicating acceptable content validity. According to Polit and Beck (2006), an S-CVI/Ave value of 0.80 or higher suggests that the items are generally relevant and representative of the construct being measured. This finding indicates that the FTDT adequately reflects the prerequisite competencies required in General Mathematics. Although several items required revision, the validation process helped improve item clarity, relevance, and alignment with the intended learning competencies. The result supports previous studies emphasizing the importance of expert validation in ensuring the quality and appropriateness of educational assessment instruments.

Rasch Reliability and Construct Validity

Table 3. Model Fit Statistics for the Knowledge Tier

Indicator	Value	Interpretation
Person Reliability	0.883	High
MADaQ3	0.118	Acceptable
p-value	< .001	Model fits the data

Note. MADaQ3 = Mean of absolute values of centered Q_3 statistic with p-value obtained using Holm adjustment; Ho = the data fit the Rasch model.

Table 4. Model Fit Statistics for the Reasoning Tier

Indicator	Value	Interpretation
Person Reliability	0.884	High
MADaQ3	0.119	Acceptable
p-value	< .001	Model fits the data

Note. MADaQ3 = Mean of absolute values of centered Q_3 statistic with p-value obtained using Holm adjustment; Ho = the data fit the Rasch model.

Table 3 and 4 shows the person reliability indices obtained through Rasch analysis. The knowledge tier achieved a reliability coefficient of 0.883, while the reasoning tier obtained 0.884. These results indicate that both tiers consistently differentiated respondents according to their readiness levels.

Wright Map Findings

Figure 1. Wright Map for Knowledge Tier

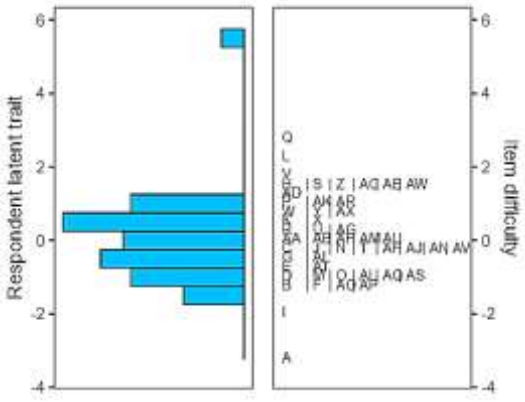
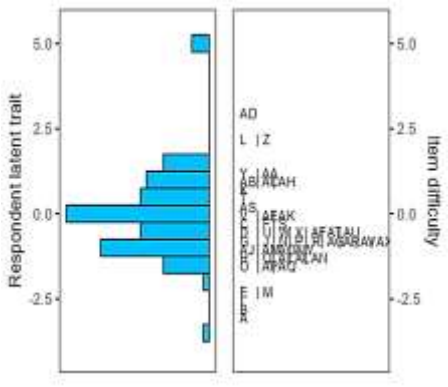


Figure 2. Wright Map for Reasoning Tier



Figures 1 and 2 present the Wright Maps for the knowledge and reasoning tiers. Respondents were generally clustered between −2 and +1 logits, indicating low to moderate readiness levels. Item difficulties ranged from approximately −3 to +2.8 logits, demonstrating a wide distribution of difficulty levels.

Readiness Index

Table 5. Distribution of Students Based on Readiness Index (RI)

Readiness Level	Frequency	Percentage	Interpretation
High Readiness (80–100%)	3	3%	Ready for SHS Mathematics
Moderate Readiness (50–79%)	42	42%	Needs instructional support
Low Readiness (<50%)	55	55%	Needs intervention
Total	100	100%	

Table 5 shows the readiness levels of the respondents. The majority of students (55%) demonstrated low readiness, while 42% demonstrated moderate readiness. Only 3% of the respondents were classified as highly ready.

FTDT-CRI Classification

Table 6. Mean Distribution of Response Categories per Item

Category	Mean	Interpretation
High Readiness	29.88	Majority demonstrated mastery
Partial Understanding	7.28	Correct but unsure
False Positive	10.46	Possible guessing
False Negative	22.5	Difficulty in application
Misconception	13.3	Incorrect understanding
Lack of Knowledge	16.58	No prior knowledge

Table 6 presents the mean percentage of responses classified under each FTDT-CRI diagnostic category across all test items. High readiness obtained the highest mean percentage (29.88%), indicating that a considerable proportion of responses reflected mastery of prerequisite competencies. However, false negative responses

(22.50%), lack of knowledge (16.58%), and misconceptions (13.30%) were also prominent, suggesting that many students experienced difficulties in applying concepts correctly or possessed inaccurate conceptual understanding despite demonstrating partial knowledge.

DISCUSSION

Content Validation of the FTDT

The S-CVI/Ave value of 0.81 indicates acceptable content validity, suggesting that the FTDT adequately represents the prerequisite competencies required in General Mathematics. According to Polit and Beck (2006), an S-CVI/Ave value of 0.80 or higher is considered acceptable for newly developed instruments. The relatively low S-CVI/UA value indicates that certain items required refinement to achieve stronger agreement among validators. These findings demonstrate the importance of expert review in improving the quality and relevance of assessment items.

Rasch Reliability and Construct Validity

The high person reliability values for both the knowledge and reasoning tiers indicate that the FTDT can consistently differentiate students according to readiness level. Most items also demonstrated acceptable fit statistics, suggesting that they functioned as intended within the Rasch measurement model. The presence of only a few misfitting items indicates that the instrument generally possesses acceptable construct validity, although several items require revision to improve measurement accuracy.

Wright Map Findings

The Wright Maps revealed a mismatch between respondent ability and item difficulty. Most respondents were clustered between −2 and +1 logits, indicating low to moderate readiness levels, whereas several items were located above the average respondent ability. This suggests that a number of prerequisite competencies assessed by the FTDT were beyond the current mastery level of many students. The maps also showed clustering of items near the mean difficulty level, indicating potential redundancy among certain items, as well as gaps at the extreme ends of the scale. These findings suggest the need for additional items targeting very low- and very high-ability respondents to improve measurement precision and ensure more balanced coverage of the readiness continuum.

Readiness Index

The finding that only 3% of students demonstrated high readiness indicates substantial deficiencies in prerequisite mathematical competencies. The large proportion of students classified as low readiness suggests the presence of learning gaps that may hinder performance in General Mathematics. These results emphasize the importance of early diagnostic assessment and intervention before formal instruction begins.

FTDT-CRI Classification

The FTDT-CRI classification revealed that misconceptions and false negative responses were common among respondents. False negative responses suggest that students may possess partial understanding but experience difficulty applying concepts correctly. Meanwhile, misconceptions reflect strongly held incorrect understandings that may interfere with future learning. These findings demonstrate that readiness extends beyond answer correctness and involves reasoning and confidence. Consequently, diagnostic assessments that incorporate confidence measures provide more meaningful information than traditional assessments alone.

CONCLUSION

This study was limited to 100 Grade 10 students from a single public secondary school. Although the sample size was adequate for pilot testing and Rasch analysis, the findings may not be generalizable to all incoming Grade 11 students. Furthermore, the study focused primarily on content validity and Rasch-based

psychometric evaluation. Criterion-related validity, differential item functioning, and cognitive interviews were not conducted. Future studies may involve larger and more diverse samples, examine subgroup differences, and collect additional validity evidence to further strengthen the FTDT.

This study successfully developed and validated a Four-Tier Diagnostic Test (FTDT) integrated with the Certainty of Response Index (CRI) to assess the readiness of incoming Grade 11 students in General Mathematics. The results established that the instrument possesses acceptable content validity and high reliability, indicating that it is capable of providing consistent and meaningful measurements of students' readiness. The Rasch analysis further confirmed that the majority of the test items functioned appropriately and contributed to the measurement of the intended construct.

The findings revealed that most students demonstrated low to moderate levels of readiness, with only a small proportion exhibiting high readiness. This suggests that many incoming Grade 11 students have not yet fully mastered the prerequisite competencies required for success in General Mathematics. The Wright Map analysis further showed that several test items were more difficult than the average ability level of the respondents, highlighting existing gaps in foundational mathematical knowledge and skills.

Moreover, the FTDT-CRI classification revealed the presence of misconceptions, false negative responses, false positive responses, and lack of knowledge among students. These findings indicate that students' readiness extends beyond answer correctness and involves reasoning ability and confidence in their understanding. The prevalence of misconceptions and false negative responses suggests that many students experience difficulties in applying mathematical concepts despite possessing partial or perceived understanding.

Overall, the FTDT integrated with CRI provides a multidimensional assessment of readiness by simultaneously measuring knowledge, reasoning, and confidence. As a diagnostic instrument, it offers valuable information that may assist teachers in identifying learning gaps, designing targeted instructional interventions, and implementing readiness enhancement programs to support students' success in General Mathematics.

Practical Implications

The FTDT may serve as a diagnostic tool for mathematics teachers at the beginning of Grade 11. The readiness classifications can be used to identify students requiring remediation, bridging activities, or enrichment programs. Students classified under misconceptions may benefit from conceptual change strategies, while those exhibiting false negative responses may require opportunities to strengthen the application of their knowledge through guided practice and problem-solving activities. The results may also assist teachers in grouping learners for differentiated instruction and in designing targeted interventions that address specific learning gaps before formal instruction in General Mathematics begins.

REFERENCES

1. Aksoy, G. (2024). Development of a four-tier diagnostic test to identify misconceptions in chemistry. Journal of Chemical Education Research, 101(2), 345–356.
2. Aristiawan, A., et al. (2025). Development of a four-tier diagnostic test in algebra to identify students' misconceptions. International Journal of Mathematics Education, 18(1), 45–60.
3. Bernardo, A. B. I. (2002). Language and mathematical problem solving among bilinguals. The Journal of Psychology, 136(3), 283–297.
4. Bernardo, A. B. I., & Ismail, R. (2020). Mathematics achievement and higher-order thinking skills of Filipino students. Asia Pacific Education Review, 21(3), 417–429.
5. Black, P., & Wiliam, D. (1998). Assessment and classroom learning. Assessment in Education, 5(1), 7–74.
6. Bond, T. G., & Fox, C. M. (2015). Applying the Rasch model: Fundamental measurement in the human sciences (3rd ed.). Routledge.
7. Boone, W. J., Staver, J. R., & Yale, M. S. (2014). Rasch analysis in the human sciences. Springer.
8. Brookhart, S. M. (2013). How to create and use rubrics for formative assessment and grading. ASCD.

9. Bujang, M. A., et al. (2024). Sample size guidelines for pilot studies. Malaysian Journal of Medical Sciences, 31(1), 1–10.
10. Caleon, I. S., & Subramaniam, R. (2010). Development and application of a multi-tier diagnostic test. International Journal of Science Education, 32(7), 939–961.
11. Capuno, R., et al. (2021). Confidence and problem-solving performance in mathematics. International Journal of Educational Research, 105, 101714.
12. Carlson, M. P., et al. (2010). The calculus concept readiness (CCR) instrument. International Journal of Mathematical Education in Science and Technology, 41(7), 833–855.
13. Chen, Y., et al. (2014). Sample size requirements for Rasch model stability. Educational Measurement: Issues and Practice, 33(4), 14–24.
14. Chi, M. T. H. (2005). Commonsense conceptions of emergent processes. Journal of the Learning Sciences, 14(2), 161–199.
15. Chick, H. L. (2004). Cognitive aspects of algebraic understanding. In K. Stacey et al. (Eds.), The future of the teaching and learning of algebra (pp. 12–29). Springer.
16. Department of Education. (2016). Senior high school curriculum guide. DepEd.
17. Desstya, A., et al. (2025). Validity of four-tier diagnostic tests. Journal of Educational Assessment, 19(2), 112–125.
18. Duncan, G. J., et al. (2007). School readiness and later achievement. Developmental Psychology, 43(6), 1428–1446.
19. Engzell, P., Frey, A., & Verhagen, M. D. (2021). Learning loss due to school closures during the COVID-19 pandemic. Proceedings of the National Academy of Sciences, 118(17), e2022376118.
20. Fadhilatullathifi, N., et al. (2020). Application of four-tier diagnostic tests in calculus. Journal of Mathematics Education, 11(3), 567–580.
21. Gurel, D. K., Eryilmaz, A., & McDermott, L. C. (2015). A review and comparison of diagnostic instruments. Eurasia Journal of Mathematics, Science and Technology Education, 11(5), 989–1008.
22. Habiddin, & Page, E. M. (2019). Development of four-tier diagnostic test instruments. Journal of Science Education, 20(3), 45–55.
23. Hasan, S., Bagayoko, D., & Kelley, E. L. (1999). Misconceptions and the certainty of response index (CRI). Physics Education, 34(5), 294–299.
24. Harjono, A., et al. (2021). Development of diagnostic tests in science education. Journal of Science Education, 12(4), 455–470.
25. Hwang, G. J., et al. (2021). Enhancing students' problem-solving ability in mathematics. Computers & Education, 168, 104209.
26. Istiyono, E., et al. (2023). Development of four-tier diagnostic tests in physics education. Journal of Physics Education Research, 19(1), 010101.
27. Kilpatrick, J., Swafford, J., & Findell, B. (2001). Adding it up: Helping children learn mathematics. National Academy Press.
28. Kiray, S. A., & Simsek, U. (2021). The effect of diagnostic tests on conceptual understanding. International Journal of Science Education, 43(8), 1325–1342.
29. Kuhfeld, M., et al. (2020). Projecting the potential impact of COVID-19 school closures. Educational Researcher, 49(8), 549–565.
30. Linacre, J. M. (1994). Sample size and item calibration stability. Rasch Measurement Transactions, 7(4), 328.
31. Mamolo, L. (2017). Development of an achievement test in General Mathematics. Visayas State University.
32. Mejias, S. (2019). Early assessment and intervention in mathematics readiness. Journal of Educational Psychology, 111(5), 789–802.
33. National Council of Teachers of Mathematics. (2000). Principles and standards for school mathematics. NCTM.
34. Niss, M., & Højgaard, T. (2021). Competencies and mathematical learning. Springer.
35. Organisation for Economic Co-operation and Development. (2019). PISA 2018 results (Volume I): What students know and can do. OECD Publishing.
36. Orale, R., & Uy, M. (2018). Algebra misconceptions among Filipino students. Asia Pacific Journal of Education, 38(2), 217–233.

37. Piaget, J. (1952). The origins of intelligence in children. International Universities Press.
38. Polit, D. F., & Beck, C. T. (2006). The content validity index. Research in Nursing & Health, 29(5), 489–497.
39. Polit, D. F., Beck, C. T., & Owen, S. V. (2007). Is the CVI an acceptable indicator? Research in Nursing & Health, 30(4), 459–467.
40. Rahmatiah, R., & Nurhayati, N. (2022). Confidence-based assessment in mathematics learning. Journal of Mathematics Education, 13(1), 45–58.
41. Tall, D. (1993). Advanced mathematical thinking. Kluwer Academic Publishers.
42. UNESCO. (2019). Global education monitoring report 2019: Migration, displacement and education. UNESCO.
43. Villegas, F. (2020). Developmental research design in education. Philippine Journal of Education, 99(1), 12–20.
44. Villegas, J. H. (2021). Designing project-based formative assessment in Science 10 in a modular distance learning modality (Master's thesis, Laguna State Polytechnic University, San Pablo City Campus).
45. Vygotsky, L. S. (1978). Mind in society: The development of higher psychological processes. Harvard University Press.
46. Widodo, A., et al. (2023). Improving diagnostic assessment using CRI. Journal of Educational Evaluation, 17(2), 89–102.
47. Wright, B. D., & Stone, M. H. (1979). Best test design. Mesa Press.
48. Yusoff, M. S. B. (2019). ABC of content validation. Education in Medicine Journal, 11(2), 49–54.
49. Zakaria, E., & Zainal, M. (2020). Validity of diagnostic tests in mathematics education. International Journal of Instruction, 13(3), 67–82.