Prima: Jurnal Pendidikan Matematika
Vol. 9, No. 2, May 2025, pp. 256 - 268
P-ISSN: 2579-9827, E-ISSN: 2580-2216
Web: http://jurnal.umt.ac.id/index.php/prima

DEVELOPMENT OF A FOUR-TIER DIAGNOSTIC TEST TO DETECT MIDDLE SCHOOL STUDENTS' MISCONCEPTIONS IN ALGEBRAIC THINKING

Aristiawan[1], Hestu Wilujeng[2], Moh Rizky Yoga Erwanto[3], Wasilatul Murtafiah[4]
[1]Tadris IPA, IAIN Ponorogo, Jawa Timur, Indonesia
[2,3] Tadris Matematika, IAIN Ponorogo, Jawa Timur, Indonesia
[4]Pendidikan Matematika, Universitas PGRI Madiun, Jawa Timur, Indonesia
e-mail: aristiawan@iainponorogo.ac.id

Abstract

Misconceptions in algebra are a common issue encountered in mathematics learning. To address this problem, teachers and educators need tools that can accurately detect students' misconceptions. This study aims to develop a four-tier diagnostic test instrument to detect junior high school students' misconceptions in algebraic thinking. The type of research used is development research, adopting the Oriondo Dallo Antonio development model, which includes three stages: (1) test planning, (2) test trials, and (3) test trial result analysis. The trials were conducted in June 2024 with 118 seventh-grade students from junior high schools in Ponorogo. The data analysis used in this research includes several analyses, such as validity analysis using the product-moment correlation, reliability analysis using Cronbach's alpha, and item difficulty analysis using the Item Mean Difficulty (D) method under the classical test theory approach and item response theory. The results showed that the four-tier diagnostic test instrument for diagnosing junior high school students' misconceptions in algebraic thinking consists of 17 valid items divided into four algebraic ability indicators: analytical thinking, problem-solving, generalization, and mathematical modeling. The instrument's reliability is 0.759, categorized as reliable. In this study, the majority of the developed items fell into the medium difficulty category, totaling 16 items. Only one item was in the easy category, and none fell into the difficult category.

Keywords: algebraic thinking, four-tier diagnostic test, misconceptions

Abstrak

Miskonsepsi dalam aljabar adalah masalah yang sering dihadapi dalam pembelajaran matematika. Untuk mengatasi masalah ini, guru dan pendidik perlu memiliki alat yang mampu mendeteksi miskonsepsi siswa dengan tepat. Penelitian ini bertujuan mengembangkan instrumen four-tier diagnostic test untuk mendeteksi miskonsepsi siswa SMP dalam berpikir aljabar. Jenis penelitian adalah penelitian pengembangan dengan mengadopsi model pengembangan Oriondo Dallo Antonio yang meliputi tiga tahap, yaitu (1) perencanaan tes, (2) uji coba tes, dan (3) analisis hasil uji coba tes. Uji coba dilakukan pada Juni 2024 pada 118 siswa kelas 7 SMP yang berada di Ponorogo. Analisis data yang digunakan pada penelitian ini meliputi beberapa analisis, antara lain analisis validitas menggunakan korelasi product moment, analisis reliabilitas menggunakan cronbach alpha dan analisis tingkat kesulitan butir menggunakan metode Item Mean Difficulty (D) pada pendekatan teori tes klasik dan teori respon butir. Hasil penelitian menunjukkan bahwa Instrumen *four-tier diagnostic test* untuk mendiagnosa miskonsepsi siswa SMP dalam berpikir aljabar memiliki 17 butir soal valid yang terbagi pada 4 indikator kemampuan aljabar yaitu berpikir analitis, pemecahan masalah, generalisasi dan pemodelan matematis. Reliabilitas instrumen sebesar 0,759 yang terkategori reliabel. Dalam penelitian ini mayoritas soal yang dikembangkan berada pada kategori sedang yaitu sejumlah 16 soal. Hanya ada 1 soal pada kategori mudah, dan tidak ada yang masuk dalam kategori sulit.

Kata kunci: berpikir aljabar, four-tier diagnostic test, miskonsepsi

INTRODUCTION

Algebraic thinking is an essential component of mathematics education at the secondary and advanced levels (Jahudin & Siew, 2023; Pitta-Pantazi, Chimoni, & Christou, 2020). Algebra serves as a bridge between concrete concepts in arithmetic and more complex

Received: October 24, 2024; Revised: May 7, 2025; Accepted: May 27, 2025

mathematical abstractions, such as equations, functions, and variables (Deskins, 1995; Kaput, 2018; Stillwell, 2001). Mastery of algebra is not only crucial for academic success in mathematics but also plays a significant role in developing logical and analytical thinking skills relevant to problem-solving in various fields (Harti & Agoestanto, 2019; Jahudin & Siew, 2023). However, despite algebra being central to the mathematics curriculum, many students struggle to understand fundamental algebraic concepts (Ndemo & Ndemo, 2018; Welder, 2012). This is due to various factors, including misconceptions or incorrect understanding of certain concepts (Booth, McGinn, Barbieri, & Young, 2017).

Misconceptions in algebra are a frequent issue in mathematics education. Misconceptions occur when students have an incorrect or flawed understanding of a concept, leading them to consistently make mistakes when applying that knowledge (Hasan, Bagayoko, & Kelley, 1999). In algebra, students often misinterpret mathematical symbols, such as the equal sign, or generalize operational rules that apply to numbers into the context of variables without proper understanding (Wilujeng, Kusumah, & Darhim, 2019). For example, students might believe that the distributive property of multiplication can always be applied without recognizing its limitations. Such misconceptions can hinder students' progress in grasping more complex concepts later on.

To address this issue, teachers and educators need tools that can accurately detect students' misconceptions. Traditional diagnostic instruments, such as multiple-choice tests or open-ended questions, are often insufficient for revealing deep-seated misconceptions. This is because students may be able to answer correctly without truly understanding the concept or, conversely, make mistakes that do not reflect actual misconceptions. Therefore, a more comprehensive and effective instrument is needed to explore students' understanding. One promising solution is the development of the Four-Tier Diagnostic Test (Fariyani & Rusilowati, 2015).

The Four-Tier Diagnostic Test is an assessment model designed to provide a deeper insight into students' understanding (Pujayanto et al., 2018). This test consists of four assessment levels, each providing different information about students' thinking. The first level is a multiple-choice question that requires students to select an answer to a problem. The second level measures students' confidence in the answer they chose. The third level asks students to provide a reason for their choice, while the fourth level evaluates students' confidence in the reason they provided. By using these four tiers, teachers can identify whether students understand a concept correctly or if they are merely guessing, as well as assess the level of confidence students have in their own understanding (Fariyani & Rusilowati, 2015).

The main advantage of the Four-Tier Diagnostic Test is its ability to differentiate whether a student has a misconception or simply lacks understanding (Caleon & Subramaniam, 2010). For instance, students who select the correct answer but are unsure of their reasoning, or those who are confident in an incorrect answer, can be easily identified. This information is invaluable for teachers, as it allows them to design more targeted learning interventions. Thus, the Four-Tier Diagnostic Test serves as a guide for teachers to understand students' thought processes and to implement strategies to address existing misconceptions.

In the context of algebra, applying the Four-Tier Diagnostic Test is highly relevant because students often struggle to grasp concepts such as variables, equations, and functions. Many students lack a solid understanding of how variables function in an equation or how to manipulate equations correctly (Egodawatte, 2011). These types of misconceptions are often difficult to detect using traditional instruments, but with the Four-Tier Diagnostic Test, teachers can more easily identify which areas are causing confusion and how confident students are in their understanding. For example, students who may choose the correct answer in a question about linear equations but provide an incorrect explanation or are unsure of their answer can be given special attention for deeper conceptual understanding.

Research on the development and application of the Four-Tier Diagnostic Test has been conducted in various science fields, such as physics, chemistry, and biology, with promising results (Afif, Nugraha, & Samsudin, 2017; Diani, Alfin, Anggraeni, Mustari, & Fujiani, 2019; Fakhriyah & Masfuah, 2021; Habiddin & Page, 2019). In these fields, the instrument has proven capable of revealing misconceptions that are difficult to detect with conventional tests. However, in the field of algebra, research on developing and using this instrument is still limited. Given the importance of algebra in the mathematics curriculum and the high prevalence of misconceptions in this area, developing a Four-Tier Diagnostic Test specifically for algebra is highly relevant. With a valid and reliable instrument to diagnose algebraic misconceptions, teachers will find it easier to understand the challenges students face and can design more effective teaching strategies.

The purpose of this research is to develop a Four-Tier Diagnostic Test instrument specifically designed to diagnose students' misconceptions in algebraic concepts. Through the development of this instrument, it is hoped that teachers will gain deeper insights into students' understanding and be able to help them correct their conceptual errors. Additionally, this research aims to test the validity and reliability of the developed instrument so that it can be widely used in mathematics education contexts.

METHODS

This research is a research and developmental study (RnD). The development model adopted in this study follows the Oriondo Dallo Antonio development model, which consists of three stages: (1) test planning, (2) test trials, and (3) analysis of the test trial results. In the test planning stage, the researchers developed 20 items that included four indicators of algebraic thinking: analytical thinking, problem-solving, generalization, and mathematical modeling. The design of the four-tier diagnostic test instrument was then validated by subject matter experts and measurement experts (expert judgment) before being used for the trial phase.

The test trials were conducted in June 2024. The subjects of this trial were 118 seventh-grade students from MTs N 2 Ponorogo. The trial data were analyzed to determine the validity, reliability, and item difficulty level of the test. A test item is considered valid if it has a product-moment correlation value greater than 0.3. The item is deemed reliable if it has a Cronbach's alpha value greater than 0.7. Meanwhile, the item difficulty level was analyzed using the classical test theory approach. The following table classifies the item difficulty categories based on classical test theory.

Table 1. Item Difficulty Categories

Category	Criteria
Difficult	< 0,3
Medium	0,3 – 0,7
Easy	> 0,7

RESULTS AND DISCUSSION

In this research, a test was developed to diagnose misconceptions in algebraic thinking. The test includes indicators of algebraic thinking, consisting of analytical thinking, problem-solving, generalization, and mathematical modeling. These indicators are reflected in the material on the System of Linear Equations in Two Variables.

A high-quality instrument must meet several criteria: it must be valid, reliable, and have good item parameters (Aristiawan & Istiyono, 2020; Mehrens & Lehmann, 1991). The validity of an instrument reflects the degree of accuracy or correctness of the instrument. Validity indicates how well the instrument measures what it is supposed to measure according to the objectives and concepts being assessed (Miller, Linn, Gronlund, & Linn, 2009). In this research, the validity analysis used is product-moment validity. Below is the result of the validity analysis using product-moment correlation.

Table 2. Product Moment Validity Test Results

Categori	Criteria	Item
≥ 0,3	Valid	1, 2, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 20
< 0,3	Invalid	3, 18, 19

Based on the table above, it is evident that there are three items with a correlation value of less than 0.3, and these items are therefore considered invalid. The test items that are declared valid have a significant contribution to measuring the construct, meaning that these items can accurately measure students' analytical thinking, problem-solving, generalization, and mathematical modeling skills.

In the context of diagnostic tests, validity is more important compared to regular summative or formative tests. This is because the main goal of diagnostic tests is to detect misconceptions, which requires instruments that can accurately and specifically measure errors in students' understanding (Santos et al., 2020; Smith, Cerhan, & Ivnik, 2003). Without adequate validity, diagnostic tests may fail to detect conceptual errors or may provide a misleading picture of students' understanding. For example, if items designed to measure mathematical modeling end up testing students' ability to perform simple calculations, the test results will not provide clear insights into the students' difficulties in conceptualizing Linear Equation Systems.

By ensuring validity, instrument developers can have confidence that the test can indeed detect existing misconceptions and provide information that can be used to design appropriate learning interventions. Good validity ensures that the findings from this test can be trusted and significantly contribute to efforts to improve student learning.

Reliability refers to the consistency, accuracy, and precision of a measurement tool or a series of measurements (Anastasi, 1976). It indicates how dependable the measurement results are. A measurement result is considered reliable if repeated measurements on the same group of subjects yield relatively consistent results, as long as the aspect being measured has not changed in the subject. Below are the results of the reliability analysis

Table 2. Reliability Test Result

Cronbach's Alpha	N of Items
0,759	17

Based on the table above, the Cronbach's alpha value is 0.759. This figure is greater than 0.7, indicating that the test is reliable. This means that the test instrument has a good level of consistency when used on a population of students with varying abilities. High reliability ensures that test results remain relatively stable when repeated under the same conditions (Santos et al., 2020). This consistency of results is crucial in the context of diagnostic tests, where measuring misconceptions must be accurate and trustworthy. In reliable instruments, students with the same ability are expected to yield similar results in retesting, minimizing external factors that could influence test results.

Item parameters are numerical values that describe the characteristics of a test item, such as item difficulty, discriminative power, and guessing. In this research, only the item difficulty was analyzed.

The analysis of item parameters was conducted using the classical test theory (CTT) approach and item response theory (IRT). In the classical test theory approach, the analysis of difficulty level is conducted using the Item Mean Difficulty (D) method. This method utilizes the normalized average score relative to the maximum possible score for an item. It was chosen because it provides standardized results (values between 0 and 1), facilitating comparisons between items and offering a more comprehensive understanding of item difficulty based on the average score relative to the maximum score (McCowan & McCowan, 1999). This makes the method more practical and informative for testing and data analysis as a whole compared to other methods.

The results of the item difficulty analysis using the classical test theory approach can be seen in the Figure 1.

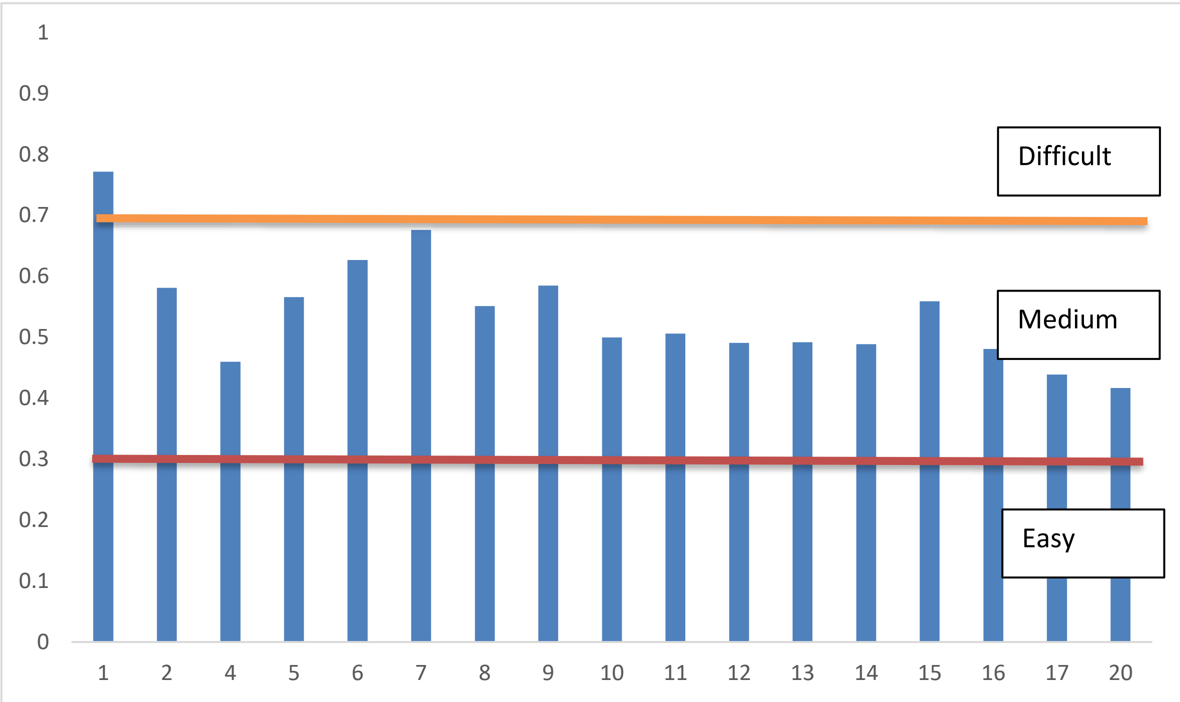

Figure 1. Difficulty Level using CTT

From the Figure 1, it can be concluded that the majority of the developed items are categorized as medium difficulty, totaling 16 items. There is only 1 item in the easy category, and none fall into the difficult category.

The item difficulty analysis using the IRT approach was conducted using the Graded Response Model (GRM). In GRM, each item does not have a single difficulty level. Instead, item difficulty is typically interpreted through thresholds or category boundaries. Each item with polytomous scoring has several thresholds that indicate the points at which respondents are more likely to move from one score category to the next. Below are the results of the item difficulty analysis.

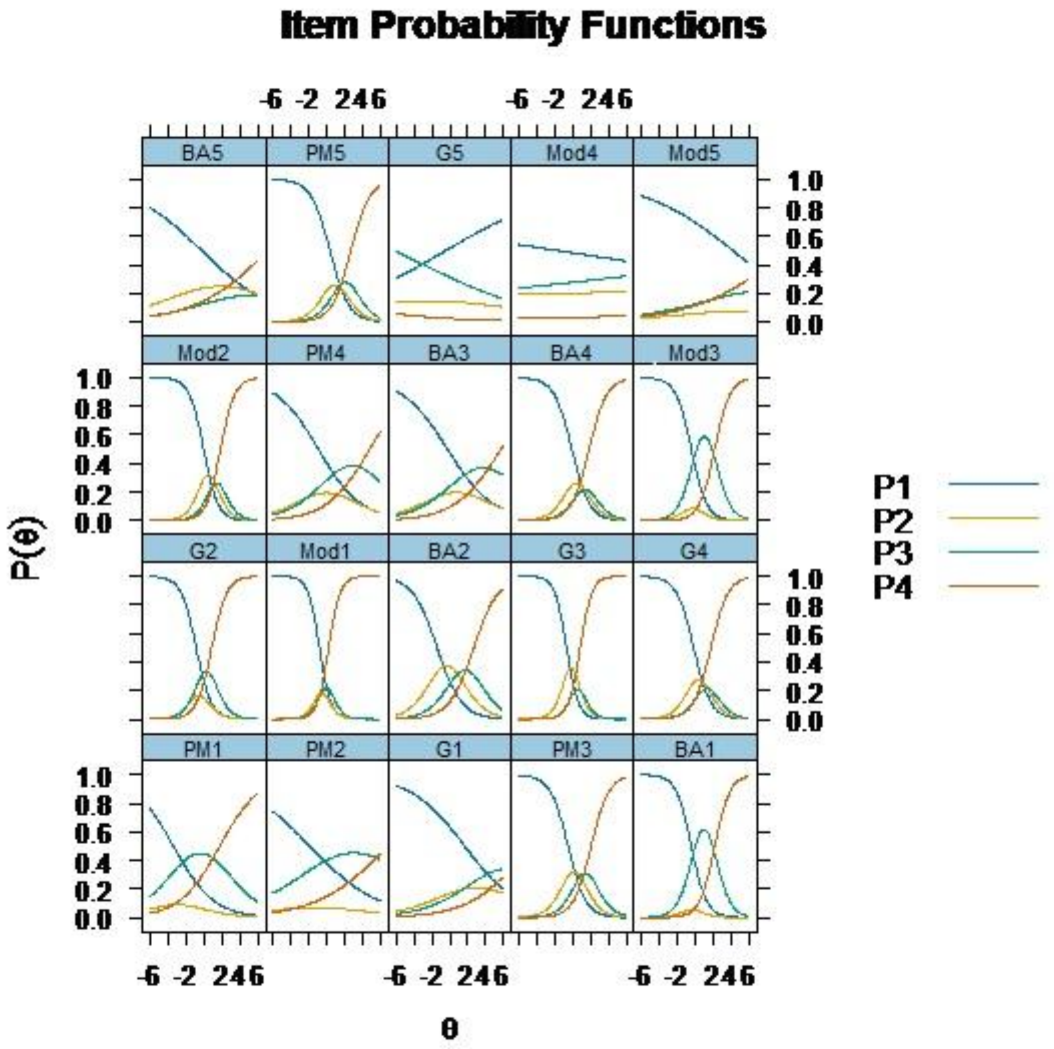

Figure 2. Difficulty Level using IRT

Fugure 2 shows that most of the items have functioning score categories, as they exhibit clear thresholds that allow for a gradual transition in responses based on students' abilities. Based on the analysis using the Graded Response Model (GRM), the majority of items have thresholds around the average ability level ($\theta \approx 0$), with some items indicating a tendency to be easier. Therefore, in general, the items fall into the moderate category, leaning toward easy.

Based on the analysis using classical test theory and item response theory, the majority of the developed items have a difficulty level classified as medium. The use of items with medium difficulty in diagnostic tests is considered appropriate because this level of difficulty ensures that students with lower abilities are not overly challenged while allowing students with higher abilities to engage without finding the questions too easy (Fariyani & Rusilowati, 2015).

Moderate difficulty levels are very important in the context of diagnostic tests, especially when the main goal is to identify students' misconceptions (Smolkowski & Cummings, 2015). Items with moderate difficulty ensure that students with varying ability levels, both low and high, can answer in appropriate proportions. Items that are too easy will likely be answered correctly by most students, making them ineffective for diagnosing deep misconceptions. Conversely, items that are too difficult may frustrate students with lower abilities, and the results may reflect students' incapacity rather than misconceptions (Borko et al., 1992). The selection of most items in the moderate category supports the objectives of the diagnostic instrument, which seeks to measure students' abilities more comprehensively.

The level of difficulty of items in the moderate category also has direct implications for the learning process. Items with moderate difficulty reflect concepts that have already been studied and are expected to be mastered by students at a certain level. This indicates that the instrument used in this study is quite representative in measuring students' understanding of the material on Linear Equation Systems (SPLDV). If most items are in the easy category, it could indicate that the material may be too easy, or that the instruction given has been very effective, allowing students to master the material well. Conversely, if many items fall into the difficult category, it may indicate that the material taught has not been well understood by most students. In this context, items with moderate difficulty provide more balanced feedback for teachers and students about the extent to which their understanding has developed and which areas need further improvement.

Furthermore, the moderate difficulty level of the items allows teachers to identify students across different ability spectrums. Students who can easily answer moderately difficult items may require additional challenges to encourage them to reach a higher level of understanding. Meanwhile, students who struggle with moderately difficult items may need reinforcement of fundamental concepts before they can progress further.

CONCLUSION

The four-tier diagnostic test instrument designed to diagnose misconceptions among junior high school students in algebraic thinking includes four indicators of algebraic skills: analytical thinking, problem-solving, generalization, and mathematical modeling. Out of the

20 items developed, 17 were found to be valid. The reliability of the instrument in this study is 0.759, which falls within the reliable category. The majority of the items, totaling 16, are categorized as medium difficulty, with no items in the difficult category and only 1 item in the easy category.

The results of this study have significant implications for educational practice, particularly in the context of mathematics instruction and the use of diagnostic tests. Knowing that the majority of developed items have a moderate difficulty level along with good validity and reliability, teachers can use this test as a tool to identify students' misconceptions in algebraic thinking. Teachers can utilize the results of this diagnostic test to adjust their teaching methods. For example, if most students show misconceptions on items that test analytical thinking indicators, teachers can provide more detailed explanations and additional practice to strengthen students' abilities in analyzing algebraic problems. Additionally, this test can also be used as a formative assessment tool to provide immediate feedback to students and correct their errors before the final examination.

REFERENCES

Afif, N. F., Nugraha, M. G., & Samsudin, A. (2017). Developing energy and momentum conceptual survey (EMCS) with four-tier diagnostic test items. AIP Conference Proceedings, 1848(1). AIP Publishing. Retrieved from https://pubs.aip.org/aip/acp/article-abstract/1848/1/050010/760254

Anastasi, A. (1976). Psychological testing. Retrieved from https://psycnet.apa.org/record/1976-25154-000

Aristiawan, A., & Istiyono, E. (2020). Developing instrument of essay test to measure the problem-solving skill in physics. Jurnal Pendidikan Fisika Indonesia, 16(2), 72–82.

Booth, J. L., McGinn, K. M., Barbieri, C., & Young, L. K. (2017). Misconceptions and Learning Algebra. In S. Stewart (Ed.), And the Rest is Just Algebra (pp. 63–78). Cham: Springer International Publishing. doi: 10.1007/978-3-319-45053-7_4

Borko, H., Eisenhart, M., Brown, C. A., Underhill, R. G., Jones, D., & Agard, P. C. (1992). Learning to teach hard mathematics: Do novice teachers and their instructors give up too easily? Journal for Research in Mathematics Education, 23(3), 194–222.

Caleon, I. S., & Subramaniam, R. (2010). Do Students Know What They Know and What They Don’t Know? Using a Four-Tier Diagnostic Test to Assess the Nature of Students’ Alternative Conceptions. Research in Science Education, 40(3), 313–337. doi: 10.1007/s11165-009-9122-4

Deskins, W. E. (1995). Abstract algebra. Courier Corporation.

Diani, R., Alfin, J., Anggraeni, Y. M., Mustari, M., & Fujiani, D. (2019). Four-tier diagnostic test with certainty of response index on the concepts of fluid. Journal of Physics: Conference Series, 1155(1), 012078. IOP Publishing. Retrieved from https://iopscience.iop.org/article/10.1088/1742-6596/1155/1/012078/meta

Egodawatte, G. (2011). Secondary school students’ misconceptions in algebra. University of Toronto, Toronto.

Fakhriyah, F., & Masfuah, S. (2021). The development of a four tier-based diagnostic test diagnostic assessment on science concept course. Journal of Physics: Conference Series, 1842(1), 012069. IOP Publishing. Retrieved from https://iopscience.iop.org/article/10.1088/1742-6596/1842/1/012069/meta

Fariyani, Q., & Rusilowati, A. (2015). Pengembangan four-tier diagnostic test untuk mengungkap miskonsepsi fisika siswa sma kelas x. Journal of Innovative Science Education, 4(2). Retrieved from https://journal.unnes.ac.id/sju/jise/article/view/9903

Habiddin, H., & Page, E. M. (2019). Development and validation of a four-tier diagnostic instrument for chemical kinetics (FTDICK). Indonesian Journal of Chemistry, 19(3), 720–736.

Harti, L. S., & Agoestanto, A. (2019). Analysis of algebraic thinking ability viewed from the mathematical critical thinking ablity of junior high school students on problem based learning. Unnes Journal Of Mathematics Education, 8(2), 119–127.

Hasan, S., Bagayoko, D., & Kelley, E. L. (1999). Misconceptions and the certainty of response index (CRI). Physics Education, 34(5), 294–299.

Jahudin, J., & Siew, N. M. (2023). An algebraic thinking skill test in problem-solving for seventh graders. Problems of Education in the 21st Century, 81(2), 223.

Kaput, J. J. (2018). Linking representations in the symbol systems of algebra. In Research issues in the learning and teaching of algebra (pp. 167–194). Routledge.

McCowan, R. J., & McCowan, S. C. (1999). Item Analysis for Criterion-Referenced Tests. Online Submission. Retrieved from https://eric.ed.gov/?id=ED501716

Mehrens, W. A., & Lehmann, I. J. (1991). Measurement and evaluation in education and psychology (4th ed). Fort Worth: Holt, Rinehart and Winston.

Miller, M. D., Linn, R. L., Gronlund, N. E., & Linn, R. L. (2009). Measurement and assessment in teaching (10th ed). Upper Saddle River, N.J: Merrill/Pearson.

Ndemo, Z., & Ndemo, O. (2018). Secondary school students' errors and misconceptions in learning algebra. Journal of Education and Learning (Edulearn), 12(4), 690–701.

Pitta-Pantazi, D., Chimoni, M., & Christou, C. (2020). Different Types of Algebraic Thinking: An Empirical Study Focusing on Middle School Students. International Journal of Science and Mathematics Education, 18(5), 965–984. doi: 10.1007/s10763-019-10003-6

Pujayanto, P., Budiharti, R., Radiyono, Y., Nuraini, N. R. A., Putri, H. V., Saputro, D. E., & Adhitama, E. (2018). Developing four tier misconception diagnostic test about kinematics. Jurnal Cakrawala Pendidikan, 37(2). Retrieved from https://journal.uny.ac.id/index.php/cp/article/view/16491

Santos, G.-M., Strathdee, S. A., El-Bassel, N., Patel, P., Subramanian, D., Horyniak, D., … Shoptaw, S. (2020). Psychometric properties of measures of substance use: A systematic review and meta-analysis of reliability, validity and diagnostic test accuracy. BMC Medical Research Methodology, 20(1), 106. doi: 10.1186/s12874-020-00963-7

Smith, G. E., Cerhan, J. H., & Ivnik, R. J. (2003). Diagnostic validity. In Clinical interpretation of the WAIS-III and WMS-III (pp. 273–301). Elsevier. Retrieved from https://www.sciencedirect.com/science/article/pii/B9780127035703500122

Smolkowski, K., & Cummings, K. D. (2015). Evaluation of Diagnostic Systems: The Selection of Students at Risk of Academic Difficulties. Assessment for Effective Intervention, 41(1), 41–54. doi: 10.1177/1534508415590386

Stillwell, J. (2001). Elements of algebra: Geometry, numbers, equations. Springer Science & Business Media.

Welder, R. M. (2012). Improving Algebra Preparation: Implications From Research on Student Misconceptions and Difficulties. School Science and Mathematics, 112(4), 255–264. doi: 10.1111/j.1949-8594.2012.00136.x

Wilujeng, H., Kusumah, Y. S., & Darhim, D. (2019). The students' achievement of algebraic thinking ability using Merrill's First Principles of Instruction. Journal of Physics: Conference Series, 1188(1), 012039. IOP Publishing. Retrieved from https://iopscience.iop.org/article/10.1088/1742-6596/1188/1/012039/meta