Development of a Computer-Based Chemistry Misconception Detector Integrated with Item Response Theory

Achmad Rante Suparman*, Murtihapsari, and Apriani Sulu Parubak

Department of Chemistry Education, Faculty of Teacher Training and Education, Universitas Papua, Manokwari Papua Barat, Indonesia
Email: a.rante@unipa.ac.id (A.R.S.); m.murtihapsari@unipa.ac.id (M.); a.parubak@unipa.ac.id (A.S.P.)
*Corresponding author

Manuscript received January 15, 2025; revised April 1, 2025; accepted May 7, 2025; published November 13, 2025

Abstract—**This study developed a computer-based chemistry misconception detection media integrated with Item Response Theory (IRT). Involving 595 students in small and large-scale trials, small-scale trials involving 55 students, and large-scale trials involving 540 students. The schools involved are schools with accreditation ratings A and B. This media is designed to identify misconceptions quickly and accurately. Analysis using R studio shows that in the Confirmatory Factor Analysis (CFA), all loading factors are positive, indicating that all questions have measured the appropriate factors and Root Mean Square Error of Approximation (RMSEA), Standardized Root Mean Square Residual (SRMR), Goodness-of-Fit Index (GFI), Normed Fit Index (NFI), Comparitive Fit Index (CFI), Incremental Fit Index (IFI), Relative Fit Index (RFI), and Tucker–Lewis Index (TLI) are proven to be fit, this indicates that the instrument construct is proven valid. The estimated value of the construct reliability using Construct Stability (CR) obtained a value > 0.5, indicating that the construct is proven reliable so that the developed detection media is proven valid and reliable. The item response theory analysis used is the Partial Credit Model (PCM) because it has a smaller Akaike Information Criterion (AIC) value. Further study in hierarchical cluster analysis using the silhouette method utilizing R Studio showed two clusters: high and low. The correlation between school accreditation and student ability based on high and low-ability clusters was analyzed using the Spearman rank correlation test and obtained a rho value of −0.006748223, which indicates a fragile relationship between the accreditation and cluster variables.**

Keywords—**confirmatory factor analysis, item response theory, misconception, R studio, silhouette method**

I. Introduction

Chemistry is a science that underlies various natural phenomena and is crucial in multiple fields such as health, pharmacy, and industry. A strong understanding of chemical concepts is essential for success in these fields [1]. However, learning chemistry is often colored by misunderstandings (misconceptions) experienced by students [2–5]. These misconceptions can hinder the correct understanding of chemical concepts and result in learning difficulties in the future [6].

Chemical misconceptions can have significant negative impacts on students. First, misconceptions can lead to an incorrect understanding of chemical concepts that are not scientific facts. This can hinder students' ability to solve chemistry problems and understand more complex concepts [7]. Second, misconceptions can cause frustration and reluctance in students to learn chemistry. Third, misconceptions can have long-term impacts and hinder students' career development in chemistry-related fields [8].

Misconceptions in chemistry learning are common and can hinder students' in-depth understanding. These misconceptions are often formed due to daily experiences, initial intuition, or even errors in the delivery of material by teachers. Determining student misconceptions is considered necessary to correctly arrange subsequent knowledge in building knowledge [9]. If not identified and appropriately addressed, misconceptions can accumulate and hinder the mastery of more complex chemical concepts [10, 11]. Misconceptions in chemistry not only impact students' conceptual understanding but can also hinder students' interest in learning this subject. This can have implications for the choice of study majors at a higher level, as well as the readiness of human resources in science and technology.

In general, some misconceptions in chemistry include widespread misunderstandings about metal bonds and metal properties in textbooks and online resources in chemical education [12]. Students who have a misunderstanding of dynamic equilibrium will tend to have a misunderstanding about the concept of shifting chemical equilibrium on the concentration factor [13]. The students maintain misunderstandings about the law of conservation of mass, which shows a failure to achieve meaningful learning [14]. There are eight misunderstandings in the isomer [15]. There is a significant misunderstanding of molecular movements, gas pressure, entropy, and temperature changes in the thermodynamic system [16].

In particular, students' misconceptions in the field of chemistry related to the chemical material studied in high school include chemical equilibrium material; students believe that the concentration of products and reactants is the same at equilibrium [17]; in the material on chemical bonds, students have misconceptions about distinguishing covalent, polar, and nonpolar bonds [8], the material on thermochemistry, students experience misconceptions in describing submicroscopic representations of a phenomenon to differentiate the concepts of temperature and heat [18], in ionic bonds, students experience misconceptions due to students' misunderstandings in thinking about ionic bonds formed by cations and anions. Meanwhile, in the concept of ionic bond formation, there are exceptions where the interaction between cations and anions does not always form ionic bonds, such as in the HCl compound. The interaction between H^+ ions and Cl^- ions does not produce ionic bonds but covalent bonds [19]. The material with the most misconceptions is chemical equilibrium, and Indonesia ranks first in the problem of misconceptions in chemistry [1].

Early detection of student misconceptions is crucial in the learning process. By knowing students' misconceptions, teachers can design effective learning strategies to correct and strengthen correct understanding. However, manual misconception detection often takes a lot of time and resources. Conventional misconception detection methods, such as written tests, often only measure the final learning outcomes without revealing the students' thinking processes underlying their answers. This makes it difficult for teachers to identify the root causes of misconceptions [20].

Developing efficient and accurate misconception detection instruments is needed to overcome the existing challenges. One promising approach is to utilize computer technology [21]. Information and communication technology development has opened up new educational opportunities, including misconception detection. Developing computer-based misconception detection software can be an efficient and effective solution to overcome these challenges. Several previous studies have proposed various approaches, such as computer-based two-tier multiple-choice diagnostic tests to identify misconceptions on chemical bonds [22], Three-Tier Diagnostic and Multiple Choice Tests in chemical concepts with response change analysis [23], Online Cognitive Diagnostic Assessment with sequential multiple-choice items in time problems [24], and computer-assisted diagnostic assessment in L2 collaborative learning [25]. Although these approaches have been developed, there are still limitations regarding accuracy and flexibility in analyzing students' misconceptions.

One of the instruments widely used in detecting misconceptions is a diagnostic test based on the "five-tier diagnostic." This instrument is a development of two-tier, three-tier, and four-tier diagnostic tests previously used to detect misconceptions. The five-tier diagnostic test consists of five levels, namely: (1) students' answers to a question, (2) the level of confidence in the answer, (3) the reasons students chose the answer, (4) the level of confidence in the reasons given, and (5) students' recognition of learning experiences relevant to the concept [26, 27]. Misconceptions can be detected using a five-tier diagnostic test because it can analyze correct or incorrect answers and consider the student's confidence level and learning experience. In addition, this instrument provides in-depth insight into students' thinking patterns, allowing educators to design more specific interventions to address misconceptions.

Although previous studies have significantly contributed to developing computer-based misconception detection instruments, several vital gaps still need to be addressed. First, many existing studies still focus on general misconception detection without providing an in-depth analysis of students' level of understanding of various aspects of the tested chemical concepts. Second, integrating computer-based misconception detection instruments with strong psychometric theories such as Item Response Theory (IRT) is still limited, even though this integration can provide more accurate and in-depth information about the characteristics of test items and students' abilities [28–30]. Third, few studies have explicitly utilized the results of misconception detection to provide personalized feedback to students and recommendations for adaptive learning strategies for teachers. This study attempts to fill these gaps by developing a computer-based chemical misconception detector that comprehensively integrates IRT to produce a more in-depth analysis of students' understanding and provide personalized feedback and learning recommendation features. Thus, this study is expected to significantly improve the effectiveness of diagnostic assessments and learning interventions in chemistry education.

The detection research proposes a more comprehensive approach by creating a computer-based misconception detection model integrated with Item Response Theory (IRT). The computer-based approach offers high flexibility in designing questions that can be adjusted to each student's ability level and learning style [31–33]. Unlike other methods that only rely on the analysis of right or wrong answers, this model provides more in-depth diagnostic information, including mapping the level of students' understanding of specific concepts. In addition, the developed software also has a personalization feature, where the system can provide automatic feedback to students and learning recommendations tailored to the misconceptions detected. This allows teachers to design more targeted remedial learning strategies. Thus, this study contributes significantly to improving the effectiveness of diagnostic assessments in chemistry education and enriching the study of technology-based misconception detection.

Item Response Theory (IRT) is a statistical framework that helps develop and analyze test instruments. IRT allows us to measure students' abilities more accurately and independently of the characteristics of the test items [34–36]. By integrating IRT into a misconception detection instrument, we can obtain more detailed information about the level of item difficulty, the discriminatory power of the test items, and the ability of students to answer the test items. Computer-based misconception detection instruments can be integrated into the overall learning cycle. Thus, teachers can use the results of data analysis to improve learning designs and provide timely interventions. Integrating IRT into a misconception detection instrument can add more detailed information about the level of item difficulty, the discriminatory power of the test items, and the ability of students to answer the test items.

This study aims to develop a computer-based chemistry misconception detector integrated with Item Response Theory that can identify students' misconceptions quickly and accurately, provide specific feedback to students regarding their misconceptions, assist teachers in designing effective remedial learning, and improve the quality of chemistry learning assessments, can be an effective tool for teachers in improving the quality of chemistry learning and helping students achieve a better understanding of chemistry concepts.

To achieve the research objectives, several main questions focused on the development process, the validity and reliability of the item instrument, and students' analytical abilities based on the approach used. The research questions raised include:

1) How is the process of developing a web-based misconception detection media integrated with Item Response Theory (IRT)?
2) To what extent is the validity and reliability of the chemical misconception diagnostic instrument items developed based on the Confirmatory Factor Analysis (CFA) and Partial Credit Model (PCM) 1 Parameter Logistic (PL) analysis?
3) To what extent do the chemical misconception diagnostic instrument test items meet the item fit criteria and show optimal information function based on the Partial Credit Model (PCM) 1 PL analysis?

4) How does the k-means clustering method perform in grouping students' abilities based on the level of understanding, application, and analysis in the chemical diagnostic test?

II. Methodology

This study aims to develop a computer-based diagnostic instrument that can significantly improve the quality of chemistry learning. This diagnostic instrument is integrated with Item Response Theory (IRT). It is expected to provide more detailed information about the ability profile and types of student misconceptions so that it can be used as a basis for designing more individual learning.

A. Research Design

This development research aims to produce a computer-based chemistry misconception detection media integrated with IRT. The development model combines the Spiral media development model [37] and the test instrument development model [38]. The Spiral Model focuses on developing interactive computer-based media, while the instrument development model is used to construct and evaluate IRT-based test instruments systematically. Both complement each other in the development process because each contributes to the media and evaluation instrument aspects needed in the final product. The spiral model development model consists of 4 stages: determine objectives, identify and resolve risks, develop and test, and plan the next iteration. The instrument test development model consists of 9 stages: compiling test specifications, writing test questions, reviewing test questions, conducting test trials, analyzing test items, improving tests, assembling tests, conducting tests, and interpreting tests. The results of the collaboration and modification were obtained in six stages: media and instrument design, limited trials, revisions, large-scale trials, analysis, and production of final products. Data were collected through observation, student response questionnaires, and test results. The categories and scores of the instrumented assessment in this study adopted the order and scores of the five-tier diagnostic [26]. Data analysis was carried out quantitatively using the R Program.

B. Sample

The sampling technique used purposive sampling to ensure representation from various regions in Indonesia, namely the western, central, and eastern parts [39]. School selection was done by considering the representation of the three central areas and ensuring diversity in educational conditions. Accessibility factors were also taken into account to ensure the smooth implementation of the research without ignoring the diversity of school characteristics. The selected schools covered various types, both public and private. They had diverse academic backgrounds to ensure that the research results could broadly represent the population of first-grade high school students in Indonesia.

Students who became respondents were selected based on inclusion criteria: first-grade high school students who had received previous chemistry lessons and had experience with computer-based tests. The sample selection process in each school was carried out by considering the variation in students' academic backgrounds to obtain more comprehensive data. With this method, the research sample reflects the diversity of educational conditions in Indonesia.

The research sample used in this study consisted of 595 students, a small-scale trial of 55 students, and a large-scale trial of 540 students. The small-scale test aims to see deficiencies or errors that need to be corrected before conducting a larger trial and to obtain initial feedback from users to make revisions. The large-scale test is to get more representative data to generalize the research results and ensure the product is ready for use. The number of samples of 540 has met the criteria for analysis using the Partial Credit Model (PCM) [40].

C. Data Analysis

Data analysis was conducted in several stages using R Studio: (1) Expert Validation: A panel of experts (chemistry educators and educational technologists) reviewed the initial test items to assess their content validity and alignment with the learning objectives. Aiken's V was used to measure the level of agreement among experts. (2) Interrater Reliability: Interrater reliability was assessed to ensure consistency among raters. (3) Confirmatory Factor Analysis (CFA): CFA was conducted to examine the test's factor structure and ensure that the items loaded appropriately on their respective factors. (4) IRT Analysis: The Partial Credit Model (PCM) was used to estimate item parameters (difficulty, discrimination) and student abilities. Item fit statistics were examined to identify and remove any misfitting items.

III. Result

This study uses six main stages in the development process, which result from the collaboration of two types of development, namely the development of the Spiral model media [37] and the test instrument development model [38]. The results of collaboration and modification include media and instrument design, limited trials, revisions, large-scale trials, analysis, and production of final products.

The detection media uses a web-based approach that allows users to access it easily. This website was developed using HyperText Markup Language (HTML), Cascading Style Sheets (CSS), and JavaScript for the interface display. In contrast, the backend was developed using Personal Home Page (PHP) and MySQL to manage user data and test results. The Bootstrap framework is applied to improve the display's responsiveness so that users can access this media via desktop or mobile devices.

The development of this website was carried out through several main stages, namely; (a) needs analysis, including activities to identify user needs, both from the student and teacher side, determining key features such as diagnostic tests, result analysis, and automatic feedback, and developing a conceptual model based on Item Response Theory (IRT); (b) system design, including the creation of a system architecture design that includes the front-end, backend, and database, designing the User Interface/ User Experience (UI/UX) using Figma to ensure an intuitive and user-friendly display, and determining the MySQL database structure to store test result data and student responses; (c) implementation, including activities to develop user interfaces using HTML, CSS, and JavaScript, build backends using PHP with the Laravel framework for development efficiency, and conduct internal testing of various input scenarios to ensure the system runs appropriately; (d) testing and evaluation, including activities to test functionality with user samples to identify potential bugs and fixes and compatibility testing on various devices and browsers to ensure accessibility.

The developed detection media has various main features that aim to improve accuracy in detecting student misconceptions. The intuitive user interface is designed for students and teachers to access test result information easily. Some of its main features include: (a) Administrator dashboard that functions to display user statistics (number of schools, students, teachers, and active question banks) and provides a summary of important information related to data management; (b) User management (schools, teachers, students), allows administrators to manage school, teacher, and student data and each user has access according to their role (administrator, teacher, student); (c) Question bank containing a collection of questions that can be used for diagnostic tests and allows questions to be managed based on specific topics or categories; (d) Test result reports and analysis, functions to provide information related to student diagnostic test results, can be accessed by teachers to see the development of student understanding, and can be accessed by students to find out the types of misconceptions.

Validation of media and instruments using Aiken's V coefficient. The Aiken Index is a measure used to assess the content validity of an instrument based on expert assessments [41]. This index is calculated by considering the proportion of expert agreement on a statement or instrument item with a particular scale. The Aiken index value ranges from 0 to 1, where a higher value indicates a more substantial level of validity. An index above 0.75 indicates the instrument has good content validity [42, 43]. In media validation, the aspects that are the focus of the assessment are the user-friendliness aspect, the application menu aspect, the appearance aspect, and the language aspect. The Aiken index value obtained is between 0.82 and 1; this indicates that all aspects assessed are proven to be good and valid. Meanwhile, in the validation of the instrument using R, the assessment focuses on the suitability of the material, indicators, objectives, language, answer keys, and the relevance of the items to the indicators and media obtained that all Aiken values from number 1 to 20 have values between 0.80 and 1. All Aiken index values are between 0.86 and 1; this shows that all aspects assessed are proven valid because the Aiken index value is more significant than 0.75.

The trial was carried out twice, namely a small-scale trial and a large-scale trial. The small-scale trial involved 55 students, and it aimed to test the feasibility and effectiveness of the chemical detection media and to obtain feedback and suggestions from students to improve and revise the product. After being revised, a large-scale trial was carried out to test the generalization and application of the chemical detection media and instruments. The results of large-scale trials involving 540 students are used to evaluate the validity, reliability, and usability of the instrument or media.

In instrument development, factor analysis can be used to identify latent constructs measured by a set of items and to assess the reliability and validity of a set of items. The chemistry materials used in this study cover topics taught in the early semesters of high school in Indonesia, including chemical equilibrium, covalent bonds, acid-base theory, atomic structure, periodic table of elements, and concepts of matter and their classification. The selection of these materials is also based on research results that identify that these concepts often cause misconceptions among students [1].

The validity proof of the chemistry construct instrument consists of three latent variables adapted from Bloom's taxonomy: understanding, applying, and analyzing. The selection of these three variables was adjusted to the cognitive level of high school students who were the research subjects. The first variable, understanding (MI), reflects the ability of test participants to recognize, explain, and interpret basic concepts in chemistry. The indicators of this variable consist of items Sy4, Sy5, Sy10, Sy11, Sy14, and Sy15. The second variable, applying (MN), refers to the ability of test participants to use chemical concepts or principles to solve new problems or situations. The indicators of this variable consist of items Sy6, Sy7, Sy8, Sy12, and Sy16. Next, the third variable, analyzing (MS), describes test participants' skills in describing information, comparing concepts, and drawing conclusions based on the data provided. The indicators of this variable consist of items Sy1, Sy2, Sy3, Sy9, Sy13, Sy17, Sy18, Sy19, and Sy20. The instrument's feasibility analysis was carried out through Confirmatory Factor Analysis (CFA) to ensure the suitability of the measurement model. CFA tests whether the indicators grouped based on their latent variables (constructs) are consistent in the construct. The results of the CFA analysis further show that each item in each factor provides a significant model fit, confirming the validity of the relationship between indicators and constructs. In addition, the output of the CFA model reveals that all loading factors are positive and significant, indicating that each item accurately measures the relevant factor [26, 44]. The results of the CFA analysis are shown in Fig. 1.

All loading factors have been positive, indicating that all items have measured the appropriate factors; however, only CFI and SRMR meet the criteria for instrument feasibility, so it is continued with modification indices provided in the CFA output. Modification indices are not a substitute for theory but help analyze structural relationships, especially in refining correlational assumptions [45]. Modification indices result according to Fig. 2.

Analysis of instrument feasibility results according to CFA was obtained in Table 1.

The eight categories in Table 1 are proven to be fit; this indicates that the instrument construct is proven valid (construct validity is met). The estimated value of construct reliability using Construct Reliability (CR) obtained a value of > 0.5, which indicates that the construct is proven to be reliable [26].

After being valid and reliable, the IRT model fit test is continued. There are several approaches commonly used in polytomous data, such as the Graded Response Model (GRM), Generalized Partial Credit Model (GPCM), Partial Credit Model (PCM), and PCM 1 Parameter Logistic (PCM 1 PL). However, in this study, only PCM Rasch and PCM 1 PL were considered because the characteristics of the test items and the purpose of measuring student ability emphasize graded responses without considering the differences in discrimination between response categories in a complex manner.

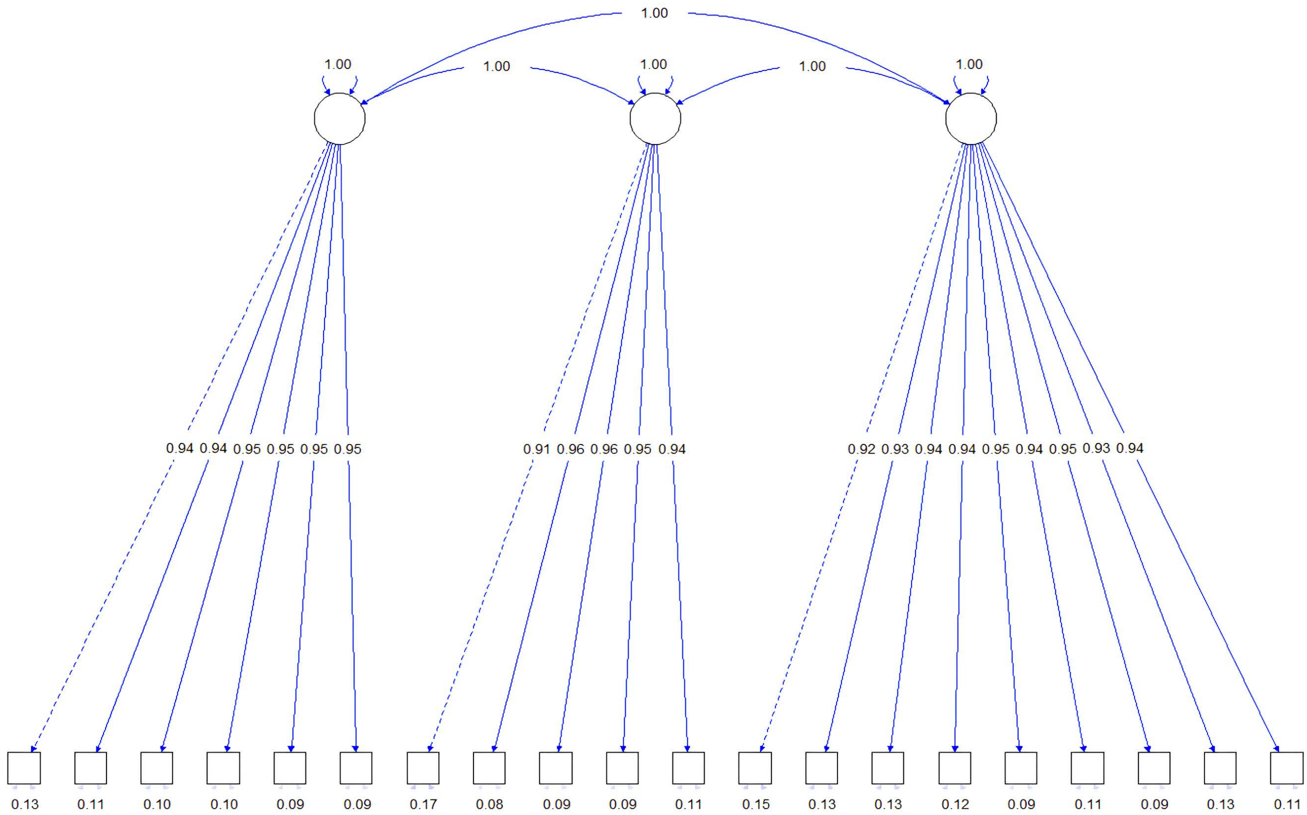


Fig. 1. Confirmatory factor analysis.

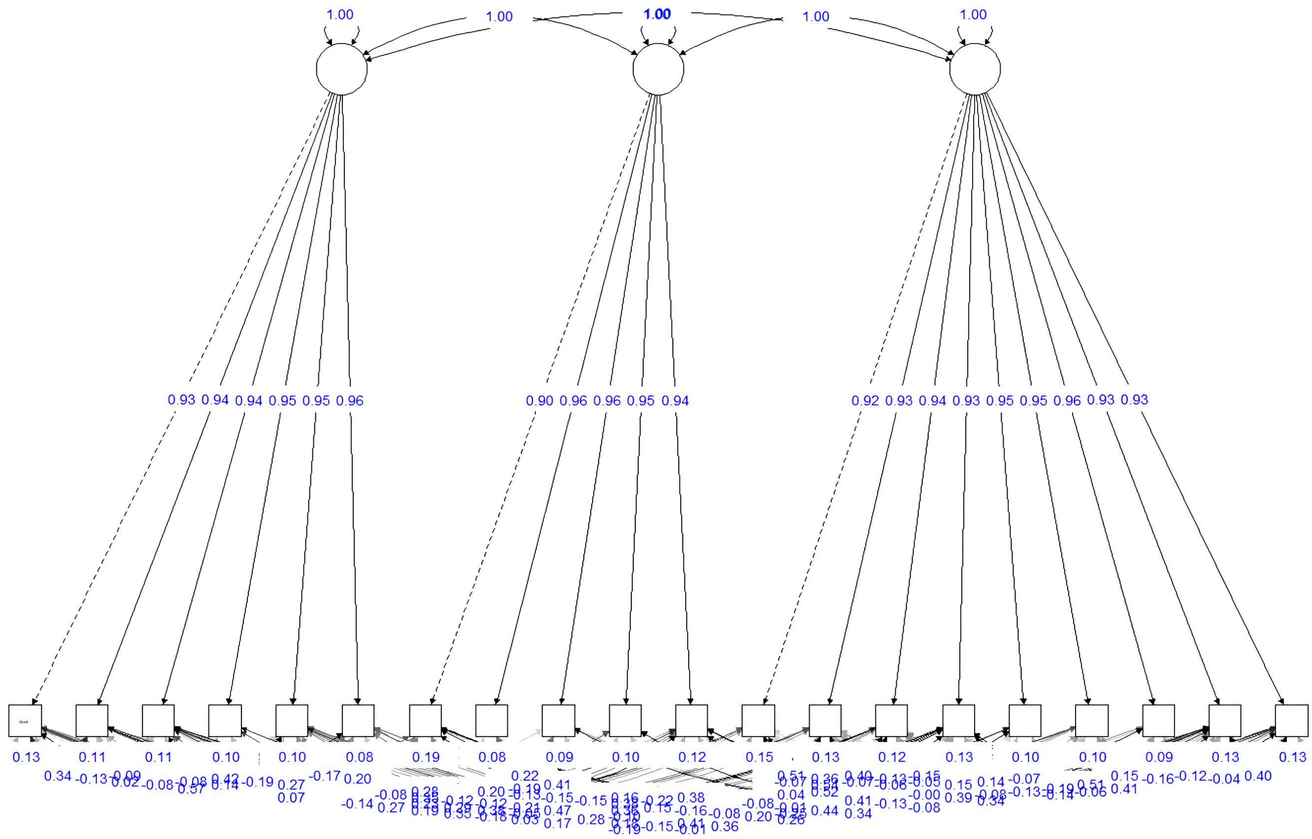


Fig. 2. Confirmatory factor analysis after modification indices.

GPCM and GRM offer greater flexibility because they allow different discrimination parameters for each item, but these models are not appropriate for the design of this study. GRM is generally used for ordered response items or in an ordinal form [50], while the items in this study do not always have a consistent category order. Meanwhile, GPCM allows different discrimination parameters for each item [51] but risks causing overparameterization and unnecessary estimation complexity in this context. In contrast, PCM and PCM 1 PL are more suitable for developing diagnostic instruments, such as in this study, where informative but straightforward parameter interpretation is a priority.

Therefore, these two models are considered and compared empirically in the research.

Table 1. Instrument feasibility results

Fit Category	Results	Description
RMSEA < 0.08 [46]	0.076	Fit
SRMR < 0.10 [47]	0.009	Fit
GFI > 0.90 [48]	0.951	Fit
NFI > 0.90 [48]	0.988	Fit
CFI > 0.87 [47]	0.991	Fit
IFI > 0.90 [49]	0.991	Fit
RFI > 0.90 [49]	0.968	Fit
TLI > 0.90 [49]	0.975	Fit

PCM is used in this study to analyze test takers' responses to polytomous scale questions, where each item has more than two score categories. This model allows for calculating the probability of participants answering each category more accurately than the dichotomous model. PCM is used to estimate participants' abilities based on the responses given at each score level in the test item, therefore providing more detailed information about the level of concept mastery by test takers [52, 53]. The difference between PCM Rasch and PCM 1 PL is that in PCM Rasch, the item discrimination parameter is determined with a value of 1, while PCM 1 PL is not necessarily 1 [54, 55]. The model with the smallest AIC value will be selected, or the model that fits the data [35, 56, 57] will be selected. The AIC result value will be according to the R output according to Table 2.

Table 2. Determination of the type of PCM

	Akaike Information Criterion (AIC)	Bayesian Information Criterion (BIC)	Log.lik
PCM Rasch	35899.32	36771.92	−17749.66
PCM 1 PL	35313.95	36190.92	−17455.98

Table 2 shows that PCM 1 PL has a smaller AIC value than PCM Rasch, so item analysis will be continued using PCM 1 PL. The results of the R analysis using the PCM Rasch model obtained a discriminant value of items 1 to 20 of 1.546. The discrimination parameter value is greater than 0, and less than 2 is a good item category [51, 58, 59].

The difficulty parameters of each item are shown in Table A1, which is the value of the difficulty parameter for each item. Each item is analyzed based on several threshold parameters denoted by b1 to b10. These parameters indicate the ability scale (θ) point at which respondents have an equal probability of moving from one answer category to the next. For example, b1 represents the transition limit from category 0 to 1, b2 from category 1 to 2, and so on up to b10, which means the transition from category 9 to 10. The higher the threshold value, the higher the ability level required to reach that category. Conversely, a low threshold value indicates that respondents with low ability can already choose that category. In addition, column b in Table A1 reflects the average of all threshold values in one item. The difficulty value has a range of $-3 \leq b \leq +3$. A higher b value means that the question is more complex and requires a higher level of ability to answer correctly. In comparison, a lower b value means that the question is more straightforward and requires a lower level of ability to answer correctly [60].

The most difficult question is Sy5, and the easiest question is Sy11. The Item Characteristic Curve (ICC) for items Sy5 and Sy11 is shown in Fig. 3 and Fig. 4.

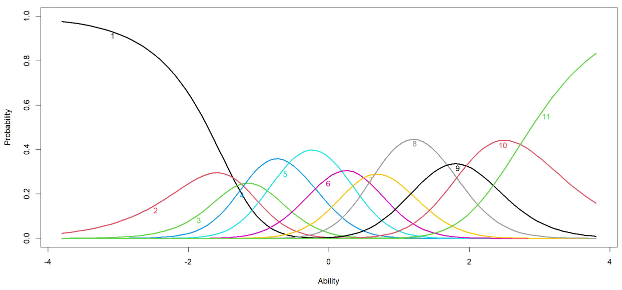


Fig. 3. Item response category characteristic curve-item: Sy5.

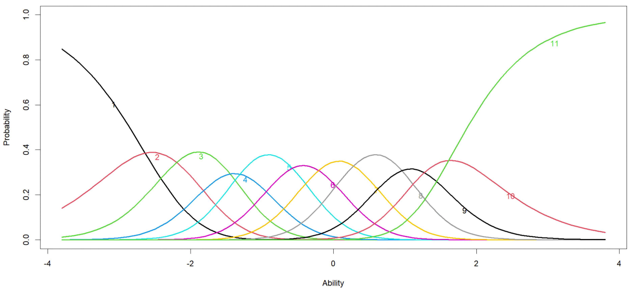


Fig. 4. Item response category characteristic curve-item: Sy11.

The following analysis tests the item fit of the developed test items. Item fit is an essential consideration in developing and using IRT-based tests. The item fit results are based on the S-X2 fit index, providing that an item is considered fit if p.S_X2 > 0.05 [61, 62]. The item fit results are in Table 3.

There are two types of information functions: the test and item information functions. Fig. 5 presents the test information function and standard error, and Figs. 6–7 present the Sy5 and Sy11 item information functions.

Table 3. Item fit instrument

Item	S_X2	df.S_X2	RMSEA.S_X2	p.S_X2
Sy1	191.779	145	0.024	0.056
Sy2	173.141	144	0.019	0.049
Sy3	161.54	143	0.015	0.138
Sy4	170.097	141	0.019	0.058
Sy5	159.832	145	0.013	0.189
Sy6	165.172	141	0.017	0.080
Sy7	149.26	140	0.011	0.281
Sy8	173.501	145	0.018	0.053
Sy9	142.779	143	0.000	0.489
Sy10	154.246	143	0.012	0.246
Sy11	144.778	147	0.000	0.536
Sy12	169.387	147	0.016	0.100
Sy13	170.873	141	0.019	0.054
Sy14	177.296	147	0.019	0.055
Sy15	161.975	140	0.016	0.099
Sy16	172.656	145	0.018	0.058
Sy17	141.034	141	0.001	0.483
Sy18	157.689	145	0.012	0.223
Sy19	171.268	139	0.020	0.053
Sy20	143.003	142	0.003	0.461

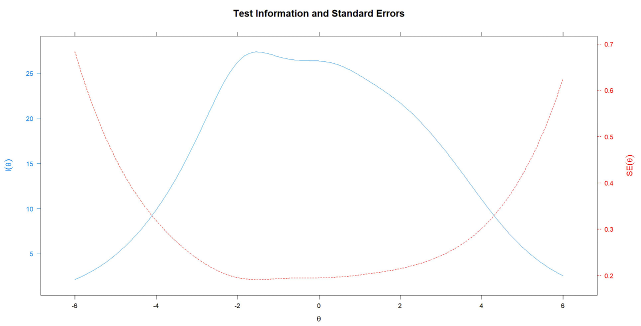


Fig. 5. Test information function and standard error.

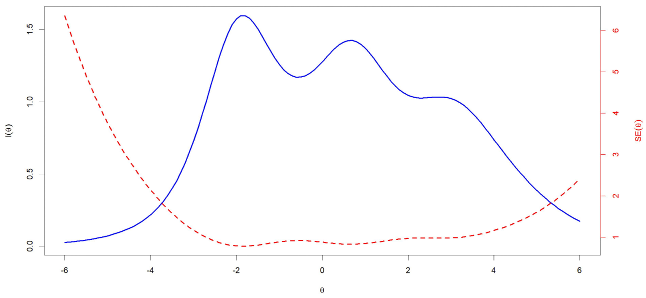


Fig. 6. Sy5 item information function.

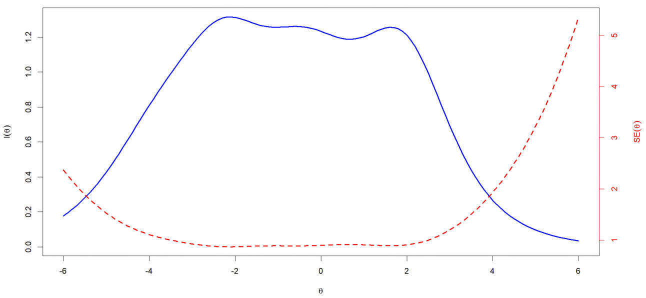


Fig. 7. Sy11 item information function.

The results of developing the integrated chemical diagnostic test item response theory can be accessed at https://pendeteksi.com/. Fig. 8 shows the initial display of the developed detection media.

Fig. 8. Initial display of the detection media.

The language used in the interface is Indonesian, as the primary users are Indonesian students and teachers. The main title reads "Pendeteksi Miskonsepsi Kimia Berbasis Komputer Terintegrasi Teori Respon Modern," which means "Computer-Based Chemistry Misconception Detector Integrated with Modern Response Theory." The subtitle states the objective: "It is hoped that the misconception detector can measure students' level of understanding in mastering chemical concepts."

This study uses the K-means clustering method to cluster the chemistry abilities of students, both from schools with A accreditation and students with B accreditation. The K-Means clustering method is used as an iterative-based approach commonly applied in data analysis. This algorithm functions to group a large number of samples into several clusters based on attribute similarities, thus allowing for more systematic identification of patterns in the data [63]. Clustering begins by separating the skills of students for the level of understanding questions (Sy4, Sy5, Sy10, Sy11, Sy14, Sy15), applying (Sy6, Sy7, Sy8, Sy12, Sy16), and analyzing (Sy1, Sy2, Sy3, Sy9, Sy13, Sy17, Sy18, Sy19, Sy20). The abilities of students from all schools are based on ability.

Analysis using K-means clustering requires several clusters based on existing data, and one method that can be used to determine the number of clusters is the silhouette method hierarchical clustering analysis using R. The silhouette method is popular because it is easy to interpret and relatively insensitive to noise [64]. The optimal number of clusters based on the silhouette method is shown in Fig. 9.

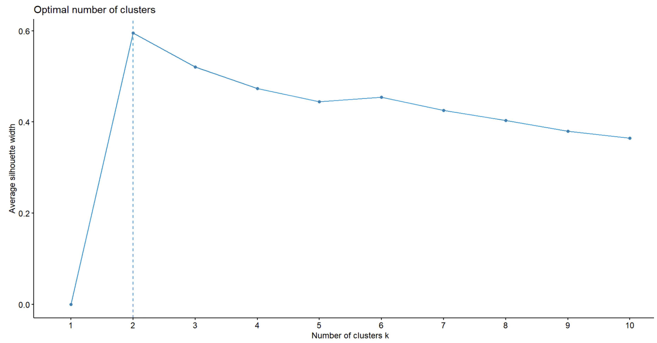


Fig. 9. Optimal number of clusters.

High and low clusters are distinguished based on cluster means, where positive values indicate high ability and lower or negative values indicate low ability. The results of the cluster mean analysis are shown in Table 4.

Table 4. Cluster means analisis

Cluster	Understanding	Applying	Analyzing	Percentage
1	−0.8183176	−0.822609	−0.8253577	49.31%
2	0.8278186	0.8246877	0.8637610	50.69%

The algorithm results are depicted or visualized using the Visio cluster, a function in the R package. The plot results are as per Fig. 10.

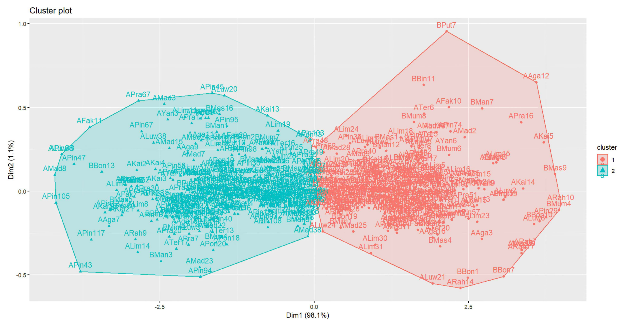


Fig. 10. Cluster plot visualization.

Further analysis results were performed using Spearman's rank. The selection of Spearman's rank analysis is based on its ability to measure monotonic relationships between variables without assuming a linear relationship. Spearman's rank is more flexible because it only considers the data's ranking, not its absolute value. In addition, this method is more resistant to outliers, which are often found in educational research [63]. The results of Spearman's rank are shown in Table 5.

Table 5. Spearman's rank test analysis results

	p-value		N	rho
	Cluster	Accreditation		
Cluster		0.8712	595	−0.006748223
Accreditation	0.8712			

Table 5 presents the results of the Spearman rank correlation test between the Cluster and Accreditation variables. The p-value of 0.8712 (> 0.05) indicates no statistically significant relationship between the two variables. The correlation coefficient (rho) is –0.0067, indicating a weak and negative association, effectively implying no ties. The term rho refers to Spearman's rank correlation coefficient, which measures the strength and direction of a monotonic association between two ranked variables. Spearman's rank correlation is a non-parametric method that does not assume a linear relationship or normal distribution, making it appropriate for the data used in this study. The value N denotes the total number of observations or samples included in the analysis, which is 595.

IV. Discussion

The media and instrument items have been proven valid and reliable based on the Aiken values obtained, so CFA analysis is continued based on the results of the work on a large-scale trial. Fig. 1 shows all positive and very significant loading factor values. However, several aspects of the fit category requirements are not yet appropriate, so it must be continued by modifying incidences to improve the results. The modification indices suggestions from R obtained 233 construct modifications, but in this study, only 95 construct modifications were carried out until a fit CFA model was obtained. Fig. 2 and Table 1 show that 20 items developed have measured the appropriate factors and produced suitable instruments. The standardized loading factor values obtained are between 0.91 and 0.95; this indicates that the construct validity criteria are met.

The IRT used in this study is polytomous and leads to PCM. PCM is used to analyze polytomous data, which refers to data with more than two response categories. There are two PCM criteria: the Rasch model PCM and the 1 PL model PCM. Rasch PCM is a type of Item Response Theory (IRT) model developed by Georg Rasch. 1 PL PCM assumes that each response category can be modeled as a separate item and the difficulty of each category can be estimated independently. In PCM, each item has a set of threshold parameters that determine the boundaries between each adjacent response category. These threshold parameters are estimated together with the test taker's ability parameters.

The comparison of AIC values shown in Table 2 shows that the appropriate analysis used is 1 PL PCM. This model estimates the probability of a test taker responding in each category, given their ability level and the difficulty threshold of the item. PCM has several advantages over other IRT models, such as the ability to estimate item difficulty at the category level and model items with different numbers of response categories [65].

The 1PL PCM assumes that the probability of a test taker giving a particular response is a function of the difference between the latent trait level and the item's difficulty. This means that the model assumes that all response options for an item have the same slope or discrimination parameter of 1.546. This PCM produces a difficulty parameter value for each item. The difficulty parameter in PCM measures how difficult an item is. The difficulty parameter can be defined as the point on the latent trait continuum where the probability of a respondent answering the item correctly is equal to the probability of the respondent answering the item incorrectly, thus finding the point on the latent trait continuum where the probability of a respondent answering the item correctly is 0.5 [51]. The difficulty parameter value is indicated by b, so from the Tabel A1 data, it can be seen that the most challenging question is the one with the highest difficulty parameter value, namely item Sy5 with a b value of 0.348, and the easiest with the lowest parameter, namely Sy11 with a b value of −0.453.

Fig. 3 and Fig. 4 show the Item Characteristic Curve (ICC). ICC is a graphical representation of the probability that an individual will answer an item correctly as a function of their ability. ICC is a valuable tool for understanding how an item measures a construct. Fig. 3 and Fig. 4 show that the X-axis (ability) measures an individual's level of latent ability. The further to the right, the higher the individual's ability. The Y-axis (probability) shows the probability or likelihood that a student with a particular ability level will answer an item (question) correctly or choose a specific option. A steeper curve indicates that the item is more sensitive to differences in individual ability, meaning that the item can distinguish well between individuals with high and low ability. The peak of the curve indicates the level of ability at which the probability of answering correctly is highest or is commonly called the item's difficulty level. A more sloping curve to the right indicates a more challenging item, while a more sloping curve to the left indicates a more accessible item.

Fig. 3 shows that item Sy5 is very difficult. Only students with very high ability have a high chance of answering this item correctly. Meanwhile, Fig. 4 shows that item Sy11 has a low difficulty level and can distinguish well between students with medium to high abilities.

After the difficult and easy items are seen, the next step is to analyze the overall item fit. Table 3 shows that all p.S_X2 values are more significant than 0.05, so all items developed in this detection media fit and are based on the Partial Credit Model (PCM) model. This shows that the PCM model can accommodate empirical data well, and no items show significant inconsistency with the model used. Although overall, all items fit, there is variation in the p.S_X2 value, which indicates different levels of fit between items. Some items, such as Sy9 (p = 0.489) and Sy11 (p = 0.536), have a very high level of fit, indicating that the PCM model is very appropriate for the participants' response patterns to the items. Conversely, there are items with p.S_X2 values approaching the lower limit of 0.05, such as Sy1 (p = 0.056) and Sy4 (p = 0.058). Although still meeting the fit criteria, values approaching this limit indicate that these items can show a lower fit to the model in different populations. Hence, they need to be monitored in further research.

In addition, the RMSEA.S_X2 values for all items are relatively small, with most of them below 0.02, indicating a low level of estimation error. This further strengthens the model used, which can describe the data well. However, one item with the highest RMSEA.S_X2 is Sy1 (0.024). Although still in the good category, this value indicates a slight deviation between the empirical data and the theoretical model. Therefore, this item can be evaluated further to ensure its stability under various measurement conditions. Overall, these results indicate that all items in the developed instrument fit well with the PCM model, with a low level of estimation error and no items showing significant mismatch. Thus, this instrument can accurately measure test takers' ability in this study. However, the finding that some items have p.S_X2 values approaching the fit limit or slightly higher RMSEA can be a consideration for further evaluation in broader applications.

The test information function shown in Fig. 5 shows that the X axis (θ) represents the level of individual latent ability. The further to the right, the higher the individual's ability. The Y axis (left/blue) shows the information the test provides at each ability level. The higher the curve, the more information the test provides about the individual's ability at that level. The Y axis (right/red) shows the magnitude of the standard error of the ability estimate at each ability level. The lower the curve, the more accurate the ability to estimate at that level. Fig. 5 shows that this test instrument suits students with theta abilities between −4 and +4.5.

Fig. 6 shows the test information function specifically for item Sy5, one of this instrument's most challenging items. It can be seen that to work on Sy5 items, a student must have a minimum ability of −3.5 and a maximum of theta +5.5. Meanwhile, Fig. 7 shows the test information function for Sy11 items, which is the easiest for this instrument. To work on this item, the minimum ability that students must have is −5.5 and a maximum of theta 4. This means that if a student has a theta of −5.5 to theta of −3.4, then the student can still work on Sy11 items but cannot work on Sy5 items.

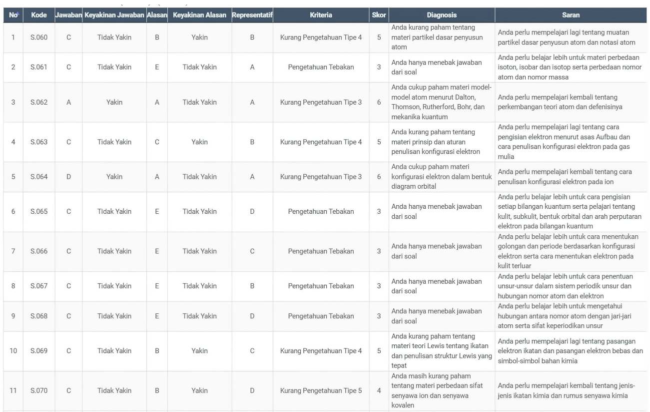
Table 4 shows that positive values are in cluster 2 and negative values are in cluster 1. This indicates that students in Cluster 2 have high abilities, and students in Cluster 1 have low skills. The division of students is based on the clusters obtained. Table 4 also shows that the number of students in the cluster with high chemical abilities and the cluster with low chemical abilities is not much different. The depiction or visualization of the algorithm results is according to Fig. 10.

Fig. 10 clarifies the two clusters of chemical abilities based on the existing data. The existence of two accreditations involved in this study, namely accreditation A and accreditation B, needs to be linked to student abilities in chemistry. The correlation between school accreditation and student abilities based on the high and low-ability clusters was analyzed using Spearman's rank correlation test. Spearman's rank correlation test was used because the data analyzed was ordinal [66]. High chemical ability is coded 2, the low chemical ability is coded 1, accreditation is coded 1, and accreditation B is coded 2. The analysis results with the Spearman's rank test are in Table 5.

The results of the Spearman test analysis in Table 5 show no significant correlation between school accreditation and students' chemistry ability, with a rho value of −0.0067 and a p-value of 0.8712. This finding indicates that the level of school accreditation is not correlated with students' chemistry ability in the context of this data. Several possible causes can be put forward. First, higher school accreditation does not necessarily reflect the quality of chemistry learning specifically because accreditation covers various administrative and managerial aspects that do not always directly impact student learning outcomes in a particular field of study. Second, student ability is likely influenced by individual factors such as interest, motivation to learn, and parental support, which are not directly related to the school's accreditation status. Third, the distribution of competent chemistry teachers and access to learning resources may be uneven but does not always depend on the level of accreditation, so the quality of learning may be similar even though the accreditation status is different. The absence of a correlation between students' chemistry abilities and accreditation emphasizes that this misconception detection media is effective for all students at schools with A and B accreditation.

The results of all item analyses entered into the detection media are precise, so IRT-based detection media are appropriate for detecting student misconceptions in chemistry. This chemical detection media, in addition to displaying values and theta, also displays diagnostic results obtained by students in Indonesian, such as Fig. 11.

Fig. 11 shows the diagnostic results obtained by students after working on the test using IRT-based chemical misconception detection media. The interface in Fig. 11 includes several columns with Indonesian headers, such as "No" (Number), "Kode" (Code), "Jawaban" (Answer), "Keyakinan Jawaban" (Answer Confidence), "Alasan" (Reason), "Keyakinan Alasan" (Reason Confidence), "Representatif" (Representativeness), "Kriteria" (Criteria), "Skor" (Score), "Diagnosis" (Diagnosis), and "Saran" (Recommendation). Each row represents a student's response to a test item, including their selected answer, confidence level, reasoning, and an automated diagnosis. For example, in row 1, a student with the code S.060 answered 'C' with low confidence and gave a reason labeled 'B' with high confidence. The system diagnosed the student with "Kurang Pengetahuan Tipe 4" (Lack of Knowledge Type 4) and recommended further study on fundamental particles and atomic notation.



No	Kode	Jawaban	Keyakinan Jawaban	Alasan	Keyakinan Alasan	Representatif	Kriteria	Skor	Diagnosis	Saran
1	S.060	C	Tidak Yakin	B	Yakin	B	Kurang Pengetahuan Tipe 4	5	Anda kurang paham tentang materi partikel dasar penyusun atom	Anda perlu mempelajari lagi tentang muatan partikel dasar penyusun atom dan notasi atom
2	S.061	C	Tidak Yakin	E	Tidak Yakin	A	Pengetahuan Tebakan	3	Anda hanya menebak jawaban dari soal	Anda perlu belajar lebih untuk materi perbedaan isoton, isobar dan isotop serta perbedaan nomor atom dan nomor massa
3	S.062	A	Yakin	A	Tidak Yakin	A	Kurang Pengetahuan Tipe 3	6	Anda cukup paham materi model-model atom menurut Dalton, Thomson, Rutherford, Bohr, dan mekanika kuantum	Anda perlu mempelajari kembali tentang perkembangan teori atom dan defenisinya
4	S.063	C	Tidak Yakin	C	Yakin	B	Kurang Pengetahuan Tipe 4	5	Anda kurang paham tentang materi prinsip dan aturan penulisan konfigurasi elektron	Anda perlu mempelajari lagi tentang cara pengisian elektron menurut asas Aufbau dan cara penulisan konfigurasi elektron pada gas mulia
5	S.064	D	Yakin	A	Tidak Yakin	A	Kurang Pengetahuan Tipe 3	6	Anda cukup paham materi konfigurasi elektron dalam bentuk diagram orbital	Anda perlu mempelajari kembali tentang cara penulisan konfigurasi elektron pada ion
6	S.065	C	Tidak Yakin	E	Tidak Yakin	D	Pengetahuan Tebakan	3	Anda hanya menebak jawaban dari soal	Anda perlu belajar lebih untuk cara pengisian setiap bilangan kuantum serta pelajari tentang kulit, subkulit, bentuk orbital dan arah perputaran elektron pada bilangan kuantum
7	S.066	C	Tidak Yakin	E	Tidak Yakin	C	Pengetahuan Tebakan	3	Anda hanya menebak jawaban dari soal	Anda perlu belajar lebih untuk cara menentukan golongan dan periode berdasarkan konfigurasi elektron serta cara menentukan elektron pada kulit terluar
8	S.067	C	Tidak Yakin	E	Tidak Yakin	B	Pengetahuan Tebakan	3	Anda hanya menebak jawaban dari soal	Anda perlu belajar lebih untuk cara penentuan unsur-unsur dalam sistem periodik unsur dan hubungan nomor atom dan elektron
9	S.068	C	Tidak Yakin	E	Tidak Yakin	D	Pengetahuan Tebakan	3	Anda hanya menebak jawaban dari soal	Anda perlu belajar lebih untuk mengetahui hubungan antara nomor atom dengan jari-jari atom serta sifat keperiodikan unsur
10	S.069	C	Tidak Yakin	B	Yakin	C	Kurang Pengetahuan Tipe 4	5	Anda kurang paham tentang materi teori Lewis tentang ikatan dan penulisan struktur Lewis yang tepat	Anda perlu mempelajari lagi tentang pasangan elektron ikatan dan pasangan elektron bebas dan simbol-simbol bahan kimia
11	S.070	C	Tidak Yakin	B	Yakin	D	Kurang Pengetahuan Tipe 5	4	Anda masih kurang paham tentang materi perbedaan sifat senyawa ion dan senyawa kovalen	Anda perlu mempelajari kembali tentang jenis-jenis ikatan kimia dan rumus senyawa kimia

Fig. 11. Display of student achievement results.

This study makes it easier for teachers to detect the types of misconceptions experienced by their students. The detection media must be subjected to a Strengths, Weaknesses, Opportunities and Threats (SWOT) analysis to determine its effectiveness and potential application in an educational environment. SWOT analysis is a systematic approach to identifying strengths, weaknesses, opportunities, and threats.

This media has the advantage of accurately detecting misconceptions using the Item Response Theory (IRT) and Partial Credit Model (PCM) approaches. Integrating these two methods allows for a more in-depth analysis of student answers so that conceptual errors can be identified more precisely compared to conventional methods. In addition, this media is web-based, so users do not need to install additional software, which can improve its accessibility and operability.

The personalized feedback feature provided to students displays test score results and provides comprehensive diagnostic information on misconceptions experienced and recommendations for appropriate learning resources. On the other hand, teachers gain access to in-depth analysis reports on student misconception patterns, allowing them to adjust learning strategies more effectively. Another advantage is the system integration with R Studio, which allows for advanced statistical analysis to ensure the accuracy of diagnostic results and validation of instrument evaluations. The user-friendly interface is also designed to improve ease of navigation for students and teachers so that the assessment process can be carried out efficiently.

One of the weaknesses of this media is its dependence on a stable internet connection, which can be an obstacle in areas with limited digital infrastructure. In addition, although this system adopts the IRT approach, interpreting complex analysis results can be difficult for teachers who are not yet familiar with assessment models based on this modern theory. Therefore, special training is needed for educators so that they can optimally utilize this feature. Another technical aspect that needs to be considered is the dependence on the database, which can be at risk if there is a disruption to the server or potential data security threats.

In terms of opportunities, this media has great potential for further development in other disciplines, such as physics and biology, to detect misconceptions in various science concepts. In addition, this system can be integrated with the Learning Management System (LMS), which is currently widely used in online learning, thus allowing its use on a broader scale. Adaptation in the form of a mobile-friendly application can also be a strategic step to increase accessibility, especially for students who use mobile devices as the leading media in learning. In addition, this media can be developed as a means of broader educational research to apply the effectiveness of various learning methods in reducing misconceptions.

Several threats can be challenges in implementing this media. One of the main risks is the low level of adoption by schools and teachers if adequate socialization and training programs do not accompany it. The sustainability of the use of this media is highly dependent on teachers' readiness to interpret the results of the analysis and adapt learning strategies based on the feedback provided. In addition, dynamic changes in curriculum policies can affect the relevance of the exam materials developed in this media. Data security is also a significant concern, especially in protecting students' personal information from data leaks or covering up academic information.

Based on this SWOT analysis, computer-based chemistry misconception detection media has significant potential to improve the quality of diagnostic assessments in education. With the implementation of the right strategy, existing weaknesses and threats can be minimized so that this media can provide an optimal contribution to improving students' understanding of chemistry concepts. Furthermore, further development is needed to expand the scope of the field of science, increase system efficiency, and strengthen the security aspects and application of technology in a broader educational environment.

The findings of this study have significant practical implications for the development of computerized tests to detect misconceptions about chemistry concepts. By utilizing the PCM model, this system can measure students' abilities more accurately, allowing for a more in-depth diagnosis of misconceptions. The model used in this study can be implemented in a computer-based evaluation platform, allowing for automation in student response analysis and faster and more objective presentation of results. This is useful in modern learning environments, where data-driven formative assessments can help educators design more targeted interventions. In addition, the advantage of this medium in displaying individual feedback for students can improve learning effectiveness by providing recommendations for appropriate learning resources based on diagnostic results. Thus, further development of this system can support the personalization of learning and improve students' understanding of chemistry concepts more effectively.

V. Conclusion

This study successfully developed a computer-based misconception detection media in chemistry learning using information and communication technology to improve efficiency and accuracy in detecting student misconceptions. The instrument developed using Item Response Theory (IRT) can provide diagnostic information about students' understanding and identify their ability to work on questions.

A trial involving 595 students showed that the items of this instrument were valid and reliable, with an Aiken index value above the set threshold, so that it can be relied on in an educational context. Analysis using the Partial Credit Model (PCM) showed that all the items developed met the suitability criteria, meaning that this instrument can differentiate well between students with different abilities.

The results of this study confirm that computer-based misconception detection media can be a practical tool for teachers to improve the quality of learning and help students better understand chemical concepts. This approach also provides a more objective and measurable assessment model than traditional methods, allowing earlier detection of student misconceptions and providing a more detailed picture of their understanding of a concept.

VI. Implications

The model developed in this study can be applied in a computer-based evaluation platform, which allows automation in the analysis of student responses and faster and more objective presentation of results. This is useful in modern learning environments, where data-based formative assessments can help educators design more targeted interventions.

Further development of this system can expand the scope of the material analyzed, including other chemical concepts taught in grades XI and XII. The number of questions in this instrument also needs to be expanded to improve the accuracy of the diagnosis and the generalization of the results. A similar approach can be applied in chemistry and other fields of study that have identical characteristics in conceptual learning.

Appendix

Example Item

Lewis's theory states that covalent bonds are the sharing of a pair of electrons. If we attach the Lewis structure of NH3 to the fourth H atom, then what will happen is:

A. The formation of NH4 molecules through covalent bonds

B. The formation of NH4 molecules through coordinate covalent bonds

C. The formation of ammonium ions through covalent bonds

D. The formation of ammonium ions through coordinate covalent bonds

E. No bond occurs

Are you sure about your answer?

A. Sure

B. No

This is because:

A. The electrons carried by the fourth H atom will increase the total number of valence electrons around N to nine so that it is no longer an octet. This causes the bond between the N atom and the fourth H atom to only come from the free electron pair in NH3, forming a coordinate bond into an ammonium ion.

B. The electrons carried by the fourth H atom will increase the total number of valence electrons around N to nine so that it is no longer an octet. This results in no bonding.

C. The electrons carried by the fourth H atom will increase the total number of valence electrons around N to nine so that it is no longer an octet. This causes the bond between the N atom and the fourth H atom to come from 1 H atom and 1 N atom to form a covalent bond into an ammonium ion.

D. The electrons carried by the fourth H atom will fulfill the electron deficiency in NH3; the bond between the N atom and the fourth H atom comes from 1 H atom, and 1 N atom forms a covalent bond into an NH4 molecule.

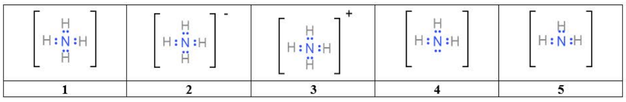
E. The electrons carried by the fourth H atom will fulfill the electron deficiency in NH3; the bond between the N atom and the fourth H atom comes from the lone pair of N electrons forming a coordinate covalent bond into an NH4 molecule

Are you sure about your answer?

A. Sure

B. No

From the following image, the correct writing of the ammonium ion is:



A. 1 B. 2 C.3 D.4 E.5

Table A1. Difficulty parameter value for each item

	b	b1	b2	b3	b4	b5	b6	b7	b8	b9	b10
Sy1	−0.425	−2.637	−2.195	−1.446	−1.309	−0.575	−0.208	0.281	0.957	1.272	1.613
Sy2	−0.378	−2.77	−2.117	−1.507	−1.301	−0.413	−0.136	0.443	0.968	1.292	1.765
Sy3	−0.401	−2.71	−2.129	−1.553	−1.161	−0.574	−0.112	0.215	0.903	1.456	1.654
Sy4	0.198	−1.98	−1.291	−0.835	−0.703	−0.383	0.363	0.815	1.313	2.008	2.671
Sy5	0.348	−1.56	−1.144	−1.085	−0.572	0.159	0.314	0.865	1.574	2.042	2.882
Sy6	0.297	−1.56	−1.263	−1.086	−0.533	0.109	0.382	0.641	1.647	2.092	2.544
Sy7	−0.137	−2.562	−1.751	−1.267	−0.704	−0.246	−0.127	0.69	0.939	1.535	2.124
Sy8	−0.203	−2.539	−1.84	−1.341	−0.967	−0.224	−0.118	0.461	0.986	1.469	2.082
Sy9	−0.291	−2.994	−1.924	−1.36	−1.101	−0.368	−0.07	0.451	0.993	1.432	2.03
Sy10	0.229	−2.015	−1.43	−0.966	−0.693	−0.151	0.281	0.774	1.452	2.121	2.917
Sy11	−0.453	−3.024	−2.045	−1.571	−1.221	−0.56	−0.225	0.397	0.763	1.3	1.66
Sy12	0.271	−2.041	−1.311	−1.006	−0.582	−0.071	0.405	0.89	1.504	2.197	2.725
Sy13	0.259	−2.085	−1.323	−0.926	−0.784	0.045	0.494	0.883	1.428	2.134	2.721
Sy14	0.232	−2.025	−1.278	−1.081	−0.426	−0.35	0.381	0.689	1.297	2.031	3.079
Sy15	0.156	−2.082	−1.695	−1.035	−0.576	−0.091	0.255	0.772	1.089	1.934	2.986
Sy16	0.318	−1.573	−1.203	−1.146	−0.588	0.066	0.39	0.781	1.617	2.054	2.778
Sy17	−0.259	−2.657	−1.892	−1.438	−1.066	−0.379	0.074	0.342	1.151	1.203	2.075
Sy18	−0.218	−2.42	−1.937	−1.342	−0.872	−0.487	−0.018	0.431	0.982	1.358	2.125
Sy19	0.333	−1.369	−1.247	−1.188	−0.564	0.182	0.498	0.654	1.683	1.955	2.725
Sy20	0.290	−1.576	−1.13	−1.053	−0.802	0.291	0.072	0.864	1.507	2.15	2.574

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Achmad Rante Suparman formulated the research problem, reviewed the relevant literature, and developed the conceptual framework. Murtihapsari was responsible for data collection, including coordinating with respondents, conducting field surveys, and ensuring ethical research procedures were followed. Apriani Sulu Parubak conducted the statistical analysis and interpreted the results using appropriate analytical tools. All authors critically reviewed and edited the manuscript, provided input on theoretical alignment, and approved the final version of the paper for submission.

FUNDING

This work was financially supported by the Ministry of Education, Culture, Research, and Technology for the National, Indonesia [Grant number: 076/E5/PG.02.00.PL/2024, dated June 11, 2024].

ACKNOWLEDGMENT

Gratitude is expressed to the Directorate of Research, Technology, and Community Service (DRTPM) of the Ministry of Education, Culture, Research, and Technology for the National Competitive Research grant funds under the Fundamental Research scheme for the 2024 fiscal year.

REFERENCES

[1] A. R. Suparman, E. Rohaeti, and S. Wening, "Student misconception in chemistry: A systematic literature review," *Pegem J. Educ. Instr.*, vol. 14, no. 2, pp. 238–252, 2024. doi: 10.47750/pegegog.14.02.28

[2] M. Aydeniz, K. Bilican, and Z. D. Kirbulut, "Exploring pre-service elementary science teachers' conceptual understanding of particulate nature of matter through three-tier diagnostic test," *Int. J. Educ. Math. Sci. Technol.*, vol. 5, no. 3, pp. 221–234, 2017.

[3] M. B. Nakhleh, "Why some students don't learn chemistry: Chemical misconceptions," *J. Chem. Educ.*, vol. 69, no. 3, pp. 191–196, 1992. doi: 10.1021/ed069p191

[4] J. R. B. Pérez *et al.*, "Student's misconceptions on chemical bonding : A comparative study between high school and first year university students," *Asian J. Educ. e-Learning*, vol. 05, no. 01, pp. 1–15, 2017.

[5] M. A. Uyulgan, N. Akkuzu, and Ş. Alpat, "Assessing the students' understanding related to molecular geometry using a two-tier diagnostic test," *J. Balt. Sci. Educ.*, vol. 13, no. 6, pp. 839–855, 2014.

[6] D. Treagust, G. Chittleborough, and T. Mamiala, "The role of submicroscopic and symbolic representations in chemical explanations," *Int. J. Sci. Educ.*, vol. 25, no. 11, pp. 1353–1368, 2003. doi: 10.1080/0950069032000070306

[7] R. Vladusic, R. B. Bucat, and M. Ozic, "Understanding covalent bonding—A scan across the Croatian education system," *Chem. Educ. Res. Pract.*, vol. 24, no. 1, pp. 108–131, 2023. doi: 10.1039/D2RP00039C

[8] T. M. Derkach, "The origin of misconceptions in inorganic chemistry and their correction by computer modelling," *Journal of Physics: Conference Series*, vol. 1840, no. 1, 012012, 2021.

[9] Y. Kökver, H. M. Pektaş, and H. Çelik, "Artificial intelligence applications in education: Natural language processing in detecting misconceptions," *Educ. Inf. Technol.*, vol. 30, no. 3, pp. 3035–3066, 2024.

[10] A. S. Härmälä-Braskén, K. Hemmi, and B. Kurtén, "Misconceptions in chemistry among Finnish prospective primary school teachers—A long-term study," *Int. J. Sci. Educ.*, vol. 42, no. 9, pp. 1447–1464, 2020.

[11] Y. Rahmawati *et al.*, "Students' conceptual understanding in chemistry learning using phet interactive simulations," *J. Technol. Sci. Educ.*, vol. 12, no. 2, pp. 303–326, 2022.

[12] A. P. Sutton, "Misconceptions about metals," *J. Chem. Educ.*, vol. 101, no. 7, pp. 2710–2715, 2024.

[13] D. Novita, Suyono, and Suyatno, "Misconceptions in dynamic equilibrium, their impact on misconceptions in chemical equilibrium due to concentration factors and their applied on prospective teachers," in *Proc. AIP Conf.*, 2024, 060031.

[14] E. Ö. Tokgöz and M. Sarıkaya. (March 2025). Experimental study on eliminating misconceptions regarding the law of conservation of mass in chemical reactions in gifted students. *J. Chem. Educ.*. [Online]. Available: https://pubs.acs.org/doi/abs/10.1021/acs.jchemed.4c01285

[15] M. Wu *et al.*, "Evaluating students' conceptual understanding of isomers based on a four-tier diagnostic tool in upper secondary schools," *Int. J. Sci. Math. Educ.*, vol. 23, pp. 907–947, 2024.

[16] B. Orazov *et al.*, "Exploring university students' misconceptions of the kinetic molecular theory of gases: A study from Kazakhstan," *Eurasia J. Math. Sci. Technol. Educ.*, vol. 21, no. 2, em2587, 2025.

[17] Y. Andriani, S. Mulyani, and W. Wiji, "Misconceptions and troublesome knowledge on chemical equilibrium," *Journal of Physics: Conference Series*, vol. 1806, 012184, 2021.

[18] H. Nasrudin and U. Azizah, "Overcoming studentâ€™s misconception through implementation of metacognitive skills-based instructional materials in energetics," *J. Pendidik. IPA Indones.*, vol. 9, no. 1, pp. 125–134, 2020. doi: 10.15294/jpii.v9i1. 21630

[19] A. K. Prodjosantoso, A. M. Hertina, and Irwanto, "The misconception diagnosis on ionic and covalent bonds concepts with three tier diagnostic test," *Int. J. Instr.*, vol. 12, no. 1, pp. 1477–1488, 2019.

[20] E. Istiyono *et al.*, "The development of a four-tier diagnostic test based on modern Test theory in physics education," *Eur. J. Educ. Res.*, vol. 12, no. 1, pp. 371–385, 2023.

[21] A. R. Suparman, E. Rohaeti, and S. Wening, "Effect of computer based test on motivation: A meta-analysis," *Eur. J. Educ. Res.*, vol. 12, no. 4, pp. 1583–1599, 2023.

[22] G. R. Utami, H. Firman, and N. Nahadi, "Development of computer based two-tier multiple choice diagnostic test to identify misconceptions on chemical bonding," *J. Phys. Conf. Ser.*, vol. 1157, no. 4, 042033, 2019.

[23] S. Türkoguz, "Investigation of three-tier diagnostic and multiple choice tests on chemistry concepts with response change behaviour," *Int. Educ. Stud.*, vol. 13, no. 9, pp. 10–22, 2020.

[24] H. Chin and C. M. Chew, "Online cognitive diagnostic assessment with ordered multiple-choice items for word problems involving 'time'," *Educ. Inf. Technol.*, vol. 27, pp. 7721–7748, 2022.

[25] N. Pourdana, "Impacts of computer-assisted diagnostic assessment on sustainability of L2 learners' collaborative writing improvement and their engagement modes," *Asian-Pacific J. Second Foreign Lang. Educ.*, vol. 7, p. 11, 2022.

[26] A. R. Suparman, E. Rohaeti, and S. Wening, "Development of computer-based chemical five-tier diagnostic test instruments: A generalized partial credit model," *J. Effic. Responsib. Educ. Sci.*, vol. 17, no. 1, pp. 92–106, 2024.

[27] R. S. Anam *et al.*, "Developing a five-tier diagnostic test to identify students' misconceptions in science: An example of the heat transfer concepts," *Elem. Educ. Online*, vol. 18, no. 3, pp. 1014–1029, 2019.

[28] C. Chen *et al.*. (January 2025). Factor retention in exploratory multidimensional item response theory. *Educ. Psychol. Meas.* [Online]. Available: https://journals.sagepub.com/doi/abs/10.1177/001316 44241306680

[29] C. Anasi *et al.*, "Measuring innovation potential in ecuadorian ict companies: development and application of the CRI-IRT model," *Information*, vol. 16, no. 2, 146, 2025.

[30] U. Schroeders and T. Gnambs, "Sample-size planning in item-response theory: A tutorial," *Adv. Methods Pract. Psychol. Sci.*, vol. 8, no. 1, pp. 1–13, 2025

[31] A. R. Dewi, A. Saehu, and R. Budiman, "Students' perception of computer-based testing-edubox–and its impact on their english learning motivation," *Indones. J. Learn. Instr.*, vol. 4, no. 1, pp. 19–28, 2021. https://journal.uniku.ac.id/index.php/IJLI/article/view/4341.

[32] E. Istiyono *et al.*, "A four tier web based assessment with eight categories diagnostic," in *Proc. 2022 8th International Conf. on Education and Technology,* 2022, pp. 190–196.

[33] E. Istiyono *et al.*, "Applying CBT in physics learning to measure students' higher order thinking skills," *Journal of Physics: Conf. Series*, vol. 1440, 012061, 2020.

[34] S. E. Embretson and S. P. Reise, *Item Response Theory for Psychologists*, New York: Psychology Press, ch. 1, 2000.

[35] W. J. van der Linden, *Handbook of Item Response Theory Volume Three Applications*, Boca raton: CRC Press, ch. 1, 2018.

[36] S. J. Osterlind and Z. Wang, "Item Response Theory in Measurement, Assessment, and Evaluation for Higher Education," *Handbook on Measurement, Assessment, and Evaluation in Higher Education*, New York: Routledge, ch. 1, pp. 150–160, 2012.

[37] F. Martin and A. K. K. Betrus, *Digital Media for Learning: Theories, Processes, and Solutions*, Cham: Springer Nature Switzerland, 2019.

[38] D. Mardapi, *Techniques for Developing Test and Non-test Instruments*, Yogyakarta: Parama Publishing, 2018. (in Indonesian)

[39] K. Q. Magnone and E. J. Yezierski, "Beyond convenience: A case and method for purposive sampling in chemistry teacher professional development research," *J. Chem. Educ.*, vol. 101, no. 3, pp. 718–726, 2024. doi: 10.1021/acs.jchemed.3c00217.

[40] R. Debelak, C. Stobl, and M. D. Zeigenfuse, *An Introduction to The Rasch Model with Examples in R*, Boca raton: CRC Press, ch. 1, 2022.

[41] L. R. Aiken, "Content validity and reliability of single items or questionnaires," *Educ. Psychol. Meas.*, vol. 40, no. 4, pp. 955–959, 1980.

[42] P. Hidalgo-Díaz *et al.*, "Content validity of a measurement instrument for the reception of the 'virtual' mass," in *Proc. International Conf. on Communication and Applied Technologies*, 2023, pp. 69–79.

[43] L. R. Aiken, "Three coefficients for analyzing the reliability and validity of ratings," *Educ. Psychol. Meas.*, vol. 45, no. 1, pp. 131–142, 1985.

[44] A. R. Suparman, E. Rohaeti, and S. Wening, "Development of attitude assessment instruments towards socioscientific issues in chemistry learning," *Eur. J. Educ. Res.*, vol. 11, no. 4, pp. 1947–1958, 2022.

[45] H. Gatignon, *Statistical Analysis of Management Data*, New York: Springer, ch. 2, 2010.

[46] J. E. Cornick, "Factor structure of the exercise self-efficacy scale," *Meas. Phys. Educ. Exerc. Sci.*, vol. 19, no. 4, pp. 208–218, 2015.

[47] N. Dagnall *et al.*, "Confirmatory factor analysis of the inventory of personality organization-reality testing subscale," *Front. Psychol.*, vol. 9, 1116, 2018.

[48] K. Y. Kwahk and J. N. Lee, "The role of readiness for change in ERP implementation: Theoretical bases and empirical validation," *Inf. Manag.*, vol. 45, no. 7, pp. 474–481, 2008.

[49] H. W. Marsh, J. R. Balla, and R. P. Mcdonald, "Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size," *Psychol. Bull.*, vol. 103, no. 3, pp. 391–410, 1988. doi: 10.1037/0033-2909.103.3.391

[50] W. A. Sethar *et al.*, "Application of Item Response Theory (IRT)-Graded Response Model (GRM) to entrepreneurial ecosystem scale," *Sustainability*, vol. 14, no. 9, 5532, 2022. doi: 10.3390/su14095532

[51] E. Muraki, "A generalized partial credit model: Application of an EM algorithm," *Appl. Psychol. Meas.*, vol. 16, no. 2, pp. 159–176, 1992.

[52] G. Tutz, G. Schauberger, and M. Berger, "Response styles in the partial credit model," *Appl. Psychol. Meas.*, vol. 42, no. 6, pp. 407–427, 2018.

[53] S. G. Baek *et al.*, "Development and validation of a leadership competency scale for high school students using the partial credit

model and the confirmatory factor analysis," in *Proc. Pacific-Rim Objective Measurement Symposium (PROMS 2023)*, 2024, pp. 142–160.
[54] C. Fellinghauer, R. Debelak, and C. Strobl, "What affects the quality of score transformations? Potential issues in true-score equating using the partial credit model," *Educ. Psychol. Meas.*, vol. 83, no. 6, pp. 1249–1290, 2023.
[55] E. Istiyono, W. brams Dwandaru, and F. Rahayu. (June 2018). The developing of creative thinking skills test based on modern test theory in physics of senior high schools. *Cakrawala Pendidik.*. [Online]. (2). Available: https://www.neliti.com/publications/261608/the-developing-of-creative-thinking-skills-test-based-on-modern-test-theory-in-p#cite
[56] D. Rizopoulos, "LTM : An R package for latent variable modeling," *J. Stat. Softw.*, vol. 17, no. 5, pp. 1–25, 2006. doi: 10.18637/jss.v017.i05
[57] H. Sung and T. Kang, "Choosing a polytomous IRT model using bayesian model selection methods," in *Proc. National Council on Measurement in Education (NCME) Annual Meeting*, 2006, pp. 1–36.
[58] W. J. Linden and R. K. Hambleton, *Handbook of Modern Item Response Theory*, New York: Springer, ch. 1, 1997.
[59] R. K. Hambleton, H. Swaminathan, and H. J. Rogers, *Fundamentals of Item Response Theory Library*, London: Sage Publications, ch. 1, 1991.
[60] F. B. Baker and S. H. Kim, *The Basics of Item Response Theory Using R*, Switzerland: Springer International Publishing, ch. 1, 2017.
[61] S. S. Dewanti, S. Hadi, and M. Nu'man, "The application of item response theory in analysis of characteristics of mathematical literacy test items," *İlköğretim Online*, vol. 20, no. 1, pp. 1226–1237, 2021. doi: 10.17051/ilkonline.2021.01.119
[62] I. Paek and K. Cole, *Using R for Item Response Theory Model Applications*, New York: Routledge, ch. 1, 2020.
[63] Z. Wang, "Research on planting strategies based on k-means cluster analysis and spearman correlation analysis," in *Proc. 2024 IEEE 2nd International Conf. on Electrical, Automation and Computer Engineering (ICEACE)*, 2024, pp. 810–815.
[64] F. Wang *et al.*, "An analysis of the application of simplified silhouette to the evaluation of k-means clustering validity," in *Proc. Machine Learning and Data Mining in Pattern Recognition*, 2017, pp. 291–305.
[65] S. Krishnan and N. Idris, "Using partial credit model to improve the quality of an instrument," *Int. J. Eval. Res. Educ.*, vol. 7, no. 4, pp. 313–316, 2018.
[66] M. D. Richardson, K. E. Lane, and J. L. Flanigan, "Teachers' perceptions of principals' attributes," *Clear. House A J. Educ. Strateg. Issues Ideas*, vol. 69, no. 5, pp. 290–292, 1996. doi: 10.1080/00098655.1996.10114321