Full research paper

DEVELOPMENT OF COMPUTER-BASED CHEMICAL FIVE-TIER DIAGNOSTIC TEST INSTRUMENTS: A GENERALIZED PARTIAL CREDIT MODEL

Achmad Rante Suparman[1]✉
Eli Rohaeti[2]
Sri Wening[2]

[1]Universitas Papua, Indonesia

[2]Universitas Negeri Yogyakarta, Indonesia

✉ a.rante@unipa.ac.id

ABSTRACT

This study focuses on developing a five-tier chemical diagnostic test based on a computer-based test with 11 assessment categories with an assessment score from 0 to 10. A total of 20 items produced were validated by education experts, material experts, measurement experts, and media experts, and an average index of the Aiken test > 0.70 was obtained. The validation results were tested on 580 respondents and analyzed using the Generalized Partial Credit Model (GPCM) Item Response Theory (IRT) type. The results of the analysis show that all of the items meet the requirements to be said to be valid for the model; the evidence of the value this: RMSEA < 0.08, CFI > 0.87, SRMR < 0.10, GFI > 0.90, NFI > 0.90, NNFI > 0.90, IFI > 0.90, TLI > 0.90, and RFI > 0.90, and all items were obtained has a p.S_X2 value greater than 0.05 which indicates that all items developed are fit and by the GPCM model. The construct reliability (CR) value is 0.99, which suggests the construct is reliable. The most challenging item is item 9, and the most accessible item is item 4.

KEYWORDS

Five-tier, chemical diagnostic, computer-based test, Generalized Partial Credit Model

HOW TO CITE

Suparman A. R., Rohaeti E., Wening S. (2024) 'Development of Computer-Based Chemical Five-Tier Diagnostic Test Instruments: A Generalized Partial Credit Model', *Journal on Efficiency and Responsibility in Education and Science*, vol. 17, no. 1, pp. 91-105. http://dx.doi.org/10.7160/eriesj.2024.170108

Article history
Received
July 5, 2023
Received in revised form
August 19, 2023
Accepted
October 24, 2023
Available on-line
March 31, 2024

Highlights

- *The five-tier chemical diagnostic test is in the form of questions with five interrelated levels.*
- *The Generalized Partial Credit Model (GPCM) is a psychometric model in item response theory used to analyze polytomous data.*
- *The items developed have difficulty parameters ranging from -2 to +2, which indicates that the produced items are excellent and informative about students' abilities.*

INTRODUCTION

One of the factors that can influence learning is students' prior knowledge (Merriënboer and Bruin, 2014). Students' misunderstandings about a material can affect the following learning process, play a role in the formation of new knowledge, and can be an inhibiting factor in constructing actual knowledge for these students (Özmen, 2004). Correcting student conceptual errors must be implemented (Üce and Ceyhan, 2019). Teachers can make contextual errors experienced by students as a basis for starting learning so that the expected goals can direct the learning methods used. According to Barke, Hazari and Sileshi Yitbarek (2009), a good lesson is correcting misunderstandings in students and providing correct knowledge, not just providing knowledge without detecting student misunderstandings.

Chemistry is a compulsory subject in high school that studies matters relating to the matter and its changes. In chemistry, many concepts are macroscopic, microscopic, and symbols. According to Tien and Osman (2017), macroscopic chemical processes can be observed and felt by sensory motors; microscopic describe particles' arrangement, interaction, and movement, while representations in symbols, numbers, formulas, and equations are called chemical symbols. Treagust, Chittleborough and Mamiala (2003) state that the symbolic level is represented in chemical symbols, formulas, and reaction equations. Misunderstandings in schools can be caused by specific scientific terminology and language problems, especially substances, particles, and chemical symbols that must be distinguished (Barke et al., 2009). The symbolic level is a representation of chemistry, so the symbolic level must be understood so that students can broadly realize chemistry concepts. Wang et al. (2017) stated that the representation of chemical symbols is a medium for the transformation between

the actual phenomena of the macroscopic world and the sub-microscopic world. Chi et al. (2018) due to its abstract nature, many students struggle with learning and effectively utilizing these symbolic representations, which can lead to ongoing failure in subsequent chemistry learning. Taking the perspective of learning progressions, this study identifies how students' abilities in chemical symbol representation progress at different grade levels (Grade 10-12) also revealed that the representation of chemical symbols is widely used in chemistry learning. However, because of its abstract nature, many students need help to learn and use these symbolic representations effectively, causing difficulties in understanding chemistry.

Misconceptions resist change and hinder students' ability to understand scientific concepts and form new cognitive structures; therefore, misunderstandings about concepts must be corrected (Potvin, 2023). Some misconceptions students experience in studying chemistry include the following: many students still think that in an equilibrium system, the forward reaction rate differs from the reverse reaction rate (Harza, Wiji and Mulyani, 2021). Students assume that the volume of liquid mixed in liquid Solubility does not affect the density of the mixture (Kiray and Simsek, 2020). Students need help correctly abstracting the concept of acids and bases (Rusmini et al., 2021). Students need clarification about polarity and need to distinguish between covalent bonds and polar and nonpolar bonds (Derkach, 2021). Students write negative reaction equations (Widarti et al., 2021). Misconceptions about reaction rates are in the form of the assumption that activation energy is the amount of energy released during a reaction and that the catalyst does not affect the reaction mechanism (Jusniar et al., 2021). Students experience misconceptions about redox because they need to understand the reduction and oxidation of the term (Murniningsih, Muna and Irawati, 2020). Students must still clearly understand the primary variable's effect on the solution's boiling point (Llanos et al., 2021). Some students need clarification about the concept of rate constants (Lamichhane, Reck and Maltese, 2018).

Equating knowledge or cognitive structures, such as very complex chemistry, is not easy, and it is unsurprising that students from high school to university still need various clarifications (Vladusic, Bucat and Ozic, 2022). A diagnostic test can be used to find out whether students understand a concept correctly or not (Istiyono et al., 2023). In addition to diagnosing student errors in understanding concepts, another function of the diagnostic is to provide input to teachers in making decisions in learning (Wang et al., 2023). Diagnostic tests can be used to discover concepts truly understood by students, ideas only partially understood, and visions that students have misconceptions about. In understanding the level of student misconceptions, especially in the field of chemistry, several types of diagnostic tests can be used, such as a two-tier diagnostic test (Mutlu and Sesen, 2015), three-tier diagnostic test (Prodjosantoso, Hertina and Irwanto, 2019), four-tier diagnostic test (Dewi, Parlan and Suryadharma, 2020) (two and finally the five-tier diagnostic test (Putra, Hamidah and Nahadi, 2020).

The five-tier diagnostic test can be combined with a computer-based test to make it easier for students to take it and for teachers to check students' work. According to Pokorný (2023), teachers must integrate modern technology into teaching. Lowyck (2014) states that the basic principle in the interaction between technology and education is how technology can support individuals and groups to achieve learning goals. Groen and Eggen (2020) said that developing a test using a Computer Based Test is the first choice the developer must make. Currently, facilities in the form of computer technology due to the discovery of computer software for use in the classroom are giving positive results, one of which significantly affects the motivation to use it (Kimmons, Clark and Lim, 2017; Suparman, Rohaeti and Wening, 2023) teacher candidates and K-12 students in a state in the USA (n = 2261). Istiyono et al. (2020) state that using Computer Based Tests can save time, and the results obtained by students come out immediately after students complete the test. Mills and Breithaupt (2016) also argue various benefits of implementing Computer Based Tests in testing, including increased measurement accuracy and efficiency, convenience, speed of reporting results, increased access to information sources and tools, and ability to assess complex skills and experience of examinees.

METHODOLOGY

Research design

Two types of research and development models are used in this study: the design of the test instrument development model and the design of the media development model. The design of the test instrument development model used the Oriondo and Antonio test development model, and the media development model used the rapid prototyping model. The design of the test instrument development model and the creation of the media development model collaborate to make it more effective because there are stages in instrument development and media development that can be carried out simultaneously. According to Oriondo and Antonio (1984), the test development model consists of four stages: instrument design, instrument testing, empirical validity determination, and reliability determination. According to Martin and Betrus (2019), the rapid prototyping model consists of assessing needs and analyzing content, constructing a prototype, utilizing a prototype, and maintaining the system. The collaboration of the two models resulted in four stages, namely: (1) designing the test instrument and CBT media, (2) integrating the instrument into CBT, (3) testing the instrument using CBT, and (4) analyzing the results of the trial.

Analysis and sample

The research analysis uses the Generalized Partial Credit Model (GPCM) so that the research sample meets the minimum requirements for analysis with GPCM. According to Debelak, Stobl and Zeigenfuse (2022), for items 5 to 20, the required sample size is 500 to obtain an accurate estimate of the GPCM model. This research develops 20 items, so the subjects have at least 500 samples. The research subjects used were 580 students from 19 schools consisting of schools with a and B accreditation. The sample selection was based on the sampling area, so the sample consisted of

students from three regions of Indonesia: West Indonesia, Central Indonesia, and East Indonesia.

Data collection technique

The data collection technique was done through a five-tier chemical diagnostic test based on CBT. The five-tier chemical diagnostic test consists of five levels of questions that form a single unit. The first question is the central question; the second question is the level of confidence in answering the main question; the third question is the reason for choosing the answer to the main question; the fourth question is the level of confidence in the cause, and the fifth question is a chemical symbolic question related to the main question. Table 1 shows the categories and scoring of the five-tier diagnostic test resulting from the development and modification of Anam et al. (2019) and Bayuni, Sopandi and Sujana (2018), which consists of 32 answer patterns. The five-tier chemical diagnostic test is integrated with a computer-based test, and students do it online. Figure 1 shows the Computer-based Test flowchart used in this study.

Answer	Confidence Level of Answers	Reason	Reason Confidence Level	Chemical Symbolic Knowledge	Category	Score
Correct	Sure	Correct	Sure	Correct	Understand	10
Correct	Not sure	Correct	Sure	Correct	Understand but lack confidence	9
Correct	Sure	Correct	Not sure	Correct		
Correct	Not sure	Correct	Not sure	Correct		
Correct	Sure	Correct	Sure	Wrong	Type 1 (lack of knowledge)	8
Correct	Not sure	Correct	Not sure	Wrong		
Correct	Sure	Correct	Not sure	Wrong		
Correct	Not sure	Correct	Sure	Wrong		
Correct	Sure	Wrong	Not sure	Correct	Type 2 (lack of knowledge)	7
Correct	Not sure	Wrong	Not sure	Correct		
Correct	Sure	Wrong	Sure	Correct		
Correct	Not sure	Wrong	Sure	Correct		
Correct	Sure	Wrong	Not sure	Wrong	Type 3 (lack of knowledge)	6
Correct	Not sure	Wrong	Sure	Wrong		
Correct	Sure	Wrong	Sure	Wrong		
Correct	Not sure	Wrong	Not sure	Wrong		
Wrong	Not sure	Correct	Sure	Correct	Type 4 (lack of knowledge)	5
Wrong	Not sure	Correct	Not sure	Correct		
Wrong	Sure	Correct	Not sure	Correct		
Wrong	Sure	Correct	Sure	Correct		
Wrong	Not sure	Correct	Sure	Wrong	Type 5 (lack of knowledge)	4
Wrong	Sure	Correct	Not sure	Wrong		
Wrong	Not sure	Correct	Not sure	Wrong		
Wrong	Sure	Correct	Sure	Wrong		
Wrong	Not sure	Wrong	Not sure	Correct	Guess knowledge	3
Wrong	Sure	Wrong	Not sure	Correct	Partial misconception	2
Wrong	Not sure	Wrong	Sure	Correct		
Wrong	Sure	Wrong	Sure	Correct		
Wrong	Sure	Wrong	Not sure	Wrong	Complete misconception	1
Wrong	Not sure	Wrong	Sure	Wrong		
Wrong	Sure	Wrong	Sure	Wrong		
Wrong	Not sure	Wrong	Not sure	Wrong	No knowledge	0

Table 1: Categories And Scoring Of the Five-Tier Diagnostic Test

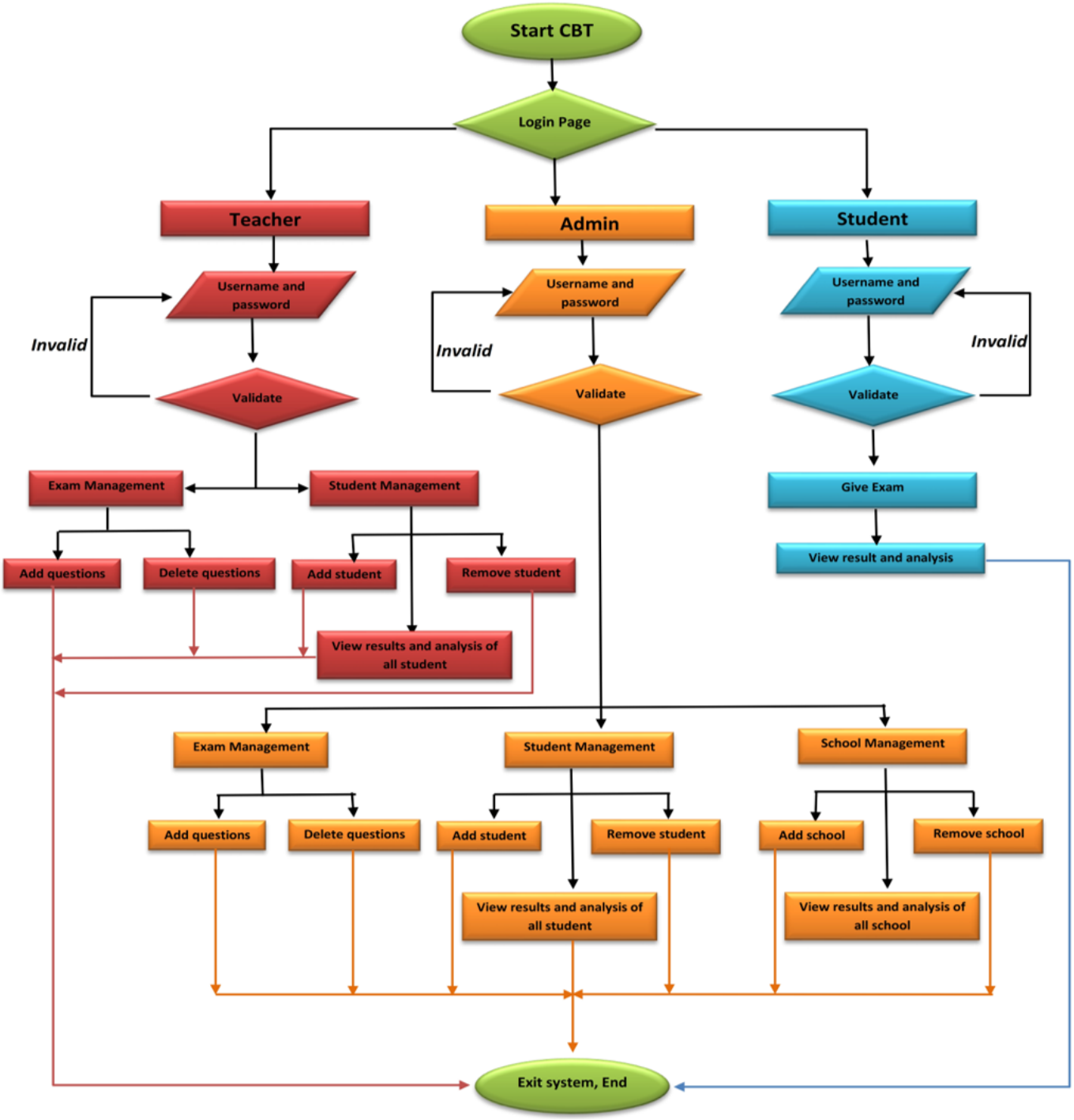

Figure 1: Flowchart Computer Based Test

RESULTS

The five-tier chemical diagnostic test grid contains indicators of a chemical material developed and continued in writing chemical questions. The suitability of the indicators with the items can be seen from the validation results carried out by experts. Content validation performed by experts was calculated using Aiken's V formula according to Table 2. Aiken suggested that valid instruments must have a validity range between 0.7 and 1. A validity range of 0.7 indicates that the set of tools is quite good, while a range of 0.9 means high validity (Aiken, 1980).

Table 2 shows that all item items are declared valid because the material, grid, indicators, and item items are appropriate or essential. The analysis results show that the instruments developed are crucial and by the curriculum, as evidenced by the average value of the Aiken index test > 0.70. The valid instruments were then tested on 580 respondents. The results of the instrument testing were analyzed using Confirmatory Factor Analysis (CFA). Figure 2 shows the Confirmatory Factor Analysis plot.

The feasibility analysis of the instrument obtained from the CFA analysis by Table 3.

Item number	Aiken index	Information	Item number	Aiken index	Information
Q1	0.93	Valid	Q11	0.96	Valid
Q2	0.96	Valid	Q12	1.00	Valid
Q3	0.96	Valid	Q13	0.93	Valid
Q4	1.00	Valid	Q14	1.00	Valid
Q5	1.00	Valid	Q15	0.96	Valid
Q6	0.93	Valid	Q16	0.93	Valid
Q7	0.96	Valid	Q17	0.96	Valid
Q8	1.00	Valid	Q18	0.96	Valid
Q9	0.86	Valid	Q19	1.00	Valid
Q10	1.00	Valid	Q20	0.86	Valid

Table 2: Categories And Scoring Of the Five-Tier Diagnostic Test

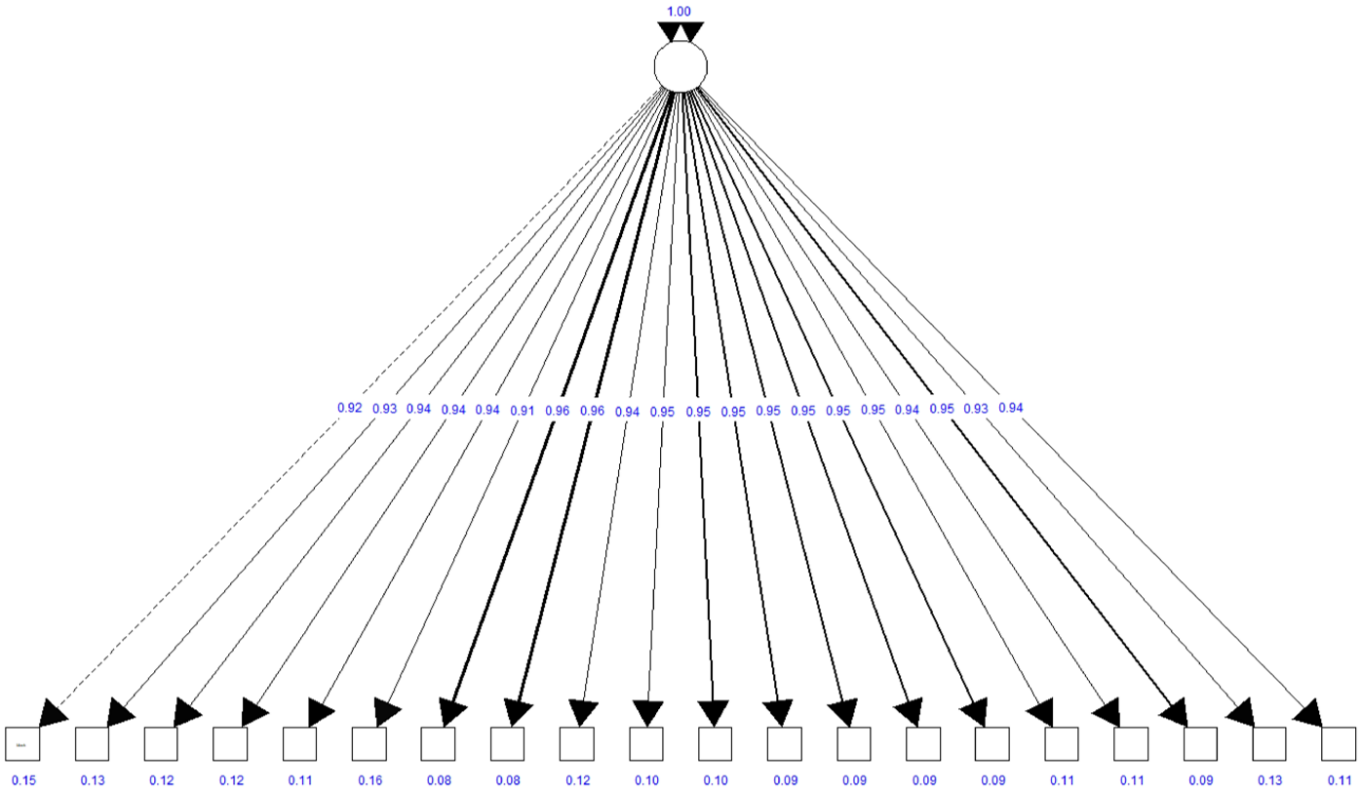

Figure 2: Confirmatory Factor Analysis (CFA)

Number	Category Name	Acceptance Category According to Theory	Analysis Results	Information
1	RMSEA	RMSEA < 0.08 (Cornick, 2015)	0.07	Fit
2	CFI	CFI > 0.87 (Dagnall et al., 2018)	0.99	Fit
3	SRMR	SRMR < 0.10 (Dagnall et al., 2018)	0.01	Fit
4	GFI	GFI > 0.90 (Kwahk and Lee, 2008)	0.95	Fit
5	NFI	NFI > 0.90 (Kwahk and Lee, 2008)	0.98	Fit
6	NNFI	NNFI > 0.90 (Kwahk and Lee, 2008)	0.97	Fit
7	IFI	IFI > 0.90 (Marsh, Balla and Mcdonald, 1988)	0.99	Fit
8	TLI	TLI > 0.90 (Marsh et al., 1988)	0.97	Fit
9	RFI	RFI > 0.90 (Marsh et al., 1988)	0.97	Fit

Table 3: Instrument Feasibility Analysis

Proving the assumptions of item response theory

The proof of the assumptions of the theory response items consists of three: unidimensional tests, local independence, and parameter invariance. Unidimensional is the ability of a question to measure only one ability. The test is unidimensional if the items are statistically dependent on the entire population (Crocker and Algina, 2008). The unidimensional assumption can be seen from the scree plot exploratory factor analysis shown in Figure 3. In Figure 3, it can be seen that there is one

factor that is measured in the chemical five-tier diagnostic test instrument. The steepness of the graph on one element is enough to prove unidimensional assumptions (Linden, 2018; Linden and Hambleton, 1997; Suparman, Rohaeti and Wening, 2022).

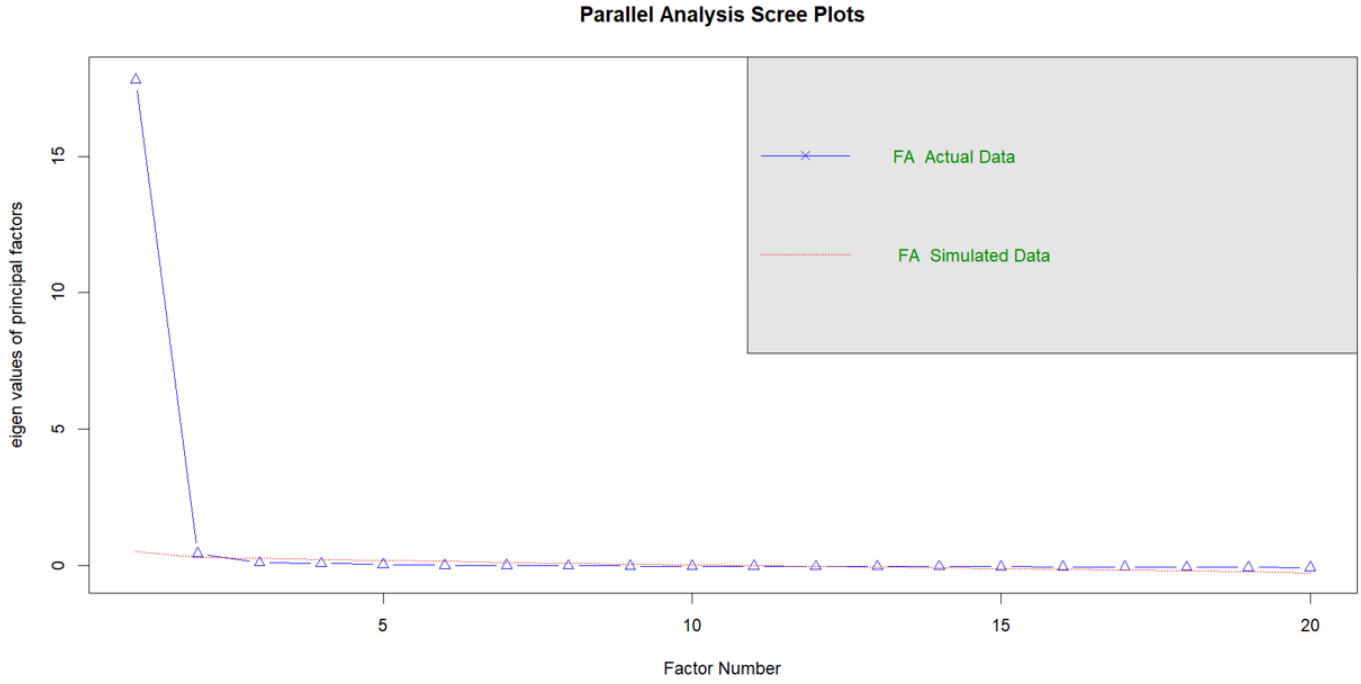

Figure 3: Scree Plot Exploratory Factor Analysis

The second assumption is local independence. Local independence is fulfilled if the students' answers are independent of their answers to other questions. The premise of local independence will automatically be proven if the unidimensional test has been established (Demars, 2010; Hambleton, 2006; Hambleton, Swaminathan and Rogers, 1991).

The third assumption is parameter invariance. Parameter invariance indicates that question parameters do not depend on the sample of examinees (Rupp and Zumbo, 2006) the equality of item and examinee parameters from different examinee populations or measurement conditions. In this article, using the well-known fact that item and examinee parameters are identical only up to a set of linear transformations specific to the functional form of a given IRT model, violations of these transformations for unidimensional IRT models are investigated using analytical, numerical, and visual tools. Because item parameter drift (IPD). In GPCM, there are two parameters, so parameter invariance also consists of item parameter invariance and ability parameter invariance. There are two item invariances, namely item parameter invariance based on differential power, according to Figure 4, and item parameter invariance based on difficulty, according to Figure 5. Ability parameter invariance based on odd and even items is shown in Figure 6.

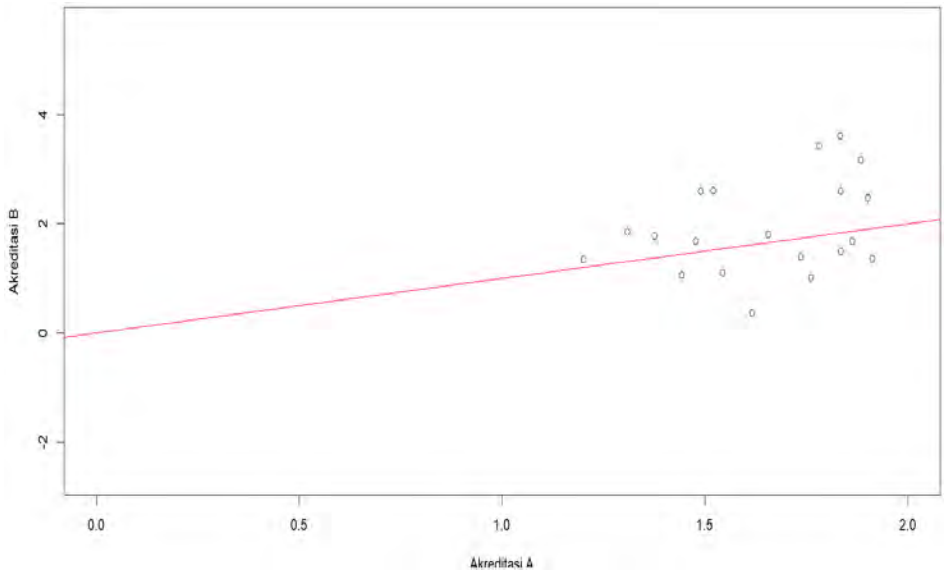

Figure 4: Invariance of Item Parameters Based on Differential Power

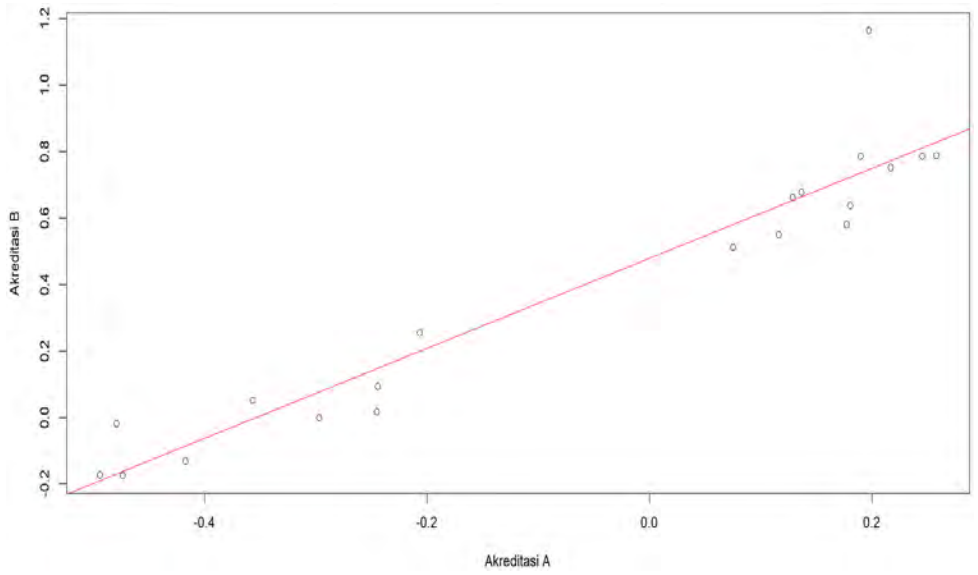

Figure 5: Invariance of Item Parameters by Difficulty

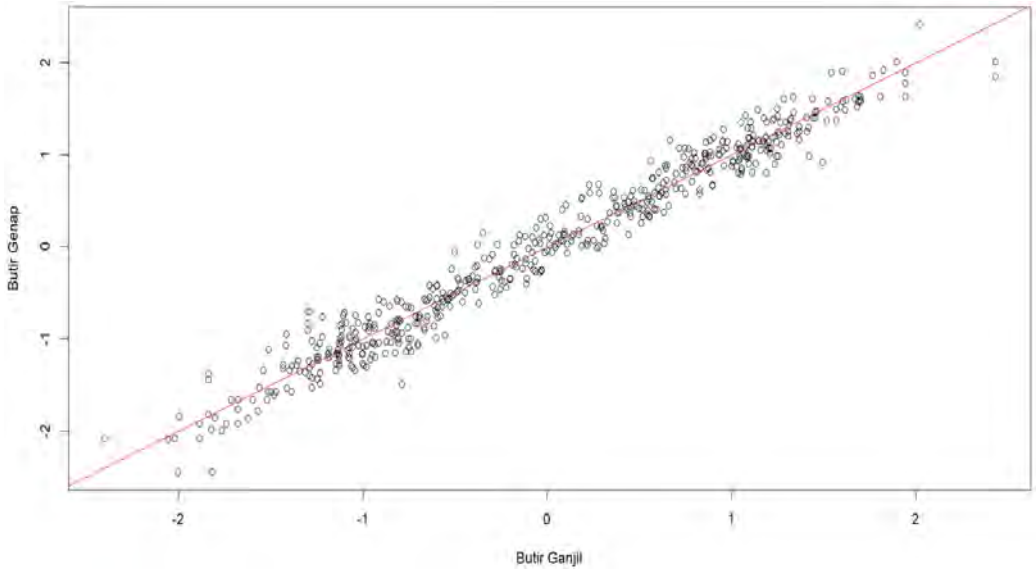

Figure 6: Invariance of Capability Parameters

Item analysis

The GPCM model has discrimination parameters and difficulty parameters. The discrimination parameter measures how well an item differentiates between people with different ability levels, while the difficulty parameter measures how difficult a question (Muraki, 1992). The value of the discrimination parameter in this study is shown in Table 4, and the difficulty value is shown in Table 5.

Item number	Discriminant Value	Item number	Discriminant Value
Q1	1.1	Q11	1.9
Q2	1.3	Q12	2.0
Q3	1.4	Q13	2.0
Q4	1.4	Q14	1.8
Q5	1.6	Q15	1.9
Q6	1.1	Q16	1.7
Q7	2.0	Q17	1.6
Q8	1.9	Q18	1.8
Q9	1.4	Q19	1.4
Q10	1.8	Q20	1.5

Table 4: Discriminant Parameter Values For Each Item

Item Number	Parameter Difficulty										
	b	b1	b2	b3	b4	b5	b6	b7	b8	b9	b10
Q1	-0.347	-2.491	-2.119	-1.263	-1.199	-0.367	-0.116	0.293	1.000	1.263	1.525
Q2	-0.292	-2.637	-1.977	-1.353	-1.179	-0.226	-0.060	0.473	0.987	1.296	1.752
Q3	-0.310	-2.562	-1.981	-1.406	-1.014	-0.412	-0.008	0.235	0.920	1.488	1.643
Q4	-0.362	-2.901	-1.881	-1.425	-1.079	-0.396	-0.131	0.442	0.769	1.321	1.660
Q5	-0.197	-2.831	-1.755	-1.217	-0.954	-0.230	0.010	0.478	1.010	1.461	2.059
Q6	0.399	-1.267	-1.059	-0.962	-0.390	0.261	0.386	0.513	1.756	2.190	2.563
Q7	-0.039	-2.350	-1.573	-1.137	-0.589	-0.141	-0.011	0.688	0.992	1.568	2.161
Q8	-0.104	-2.349	-1.655	-1.209	-0.820	-0.126	-0.019	0.496	1.012	1.507	2.119
Q9	0.431	-1.382	-0.985	-0.940	-0.437	0.253	0.330	0.872	1.609	2.079	2.915
Q10	0.307	-1.848	-1.291	-0.844	-0.538	-0.052	0.328	0.811	1.478	2.132	2.893
Q11	0.277	-1.811	-1.171	-0.730	-0.542	-0.244	0.395	0.843	1.351	2.018	2.664
Q12	0.344	-1.862	-1.203	-0.878	-0.435	0.016	0.445	0.934	1.531	2.179	2.715
Q13	0.333	-1.900	-1.211	-0.819	-0.596	0.107	0.519	0.927	1.465	2.126	2.707
Q14	0.308	-1.854	-1.159	-0.936	-0.305	-0.221	0.410	0.728	1.333	2.047	3.032
Q15	0.233	-1.919	-1.514	-0.920	-0.445	-0.002	0.305	0.795	1.155	1.948	2.923
Q16	0.394	-1.431	-1.076	-0.990	-0.443	0.142	0.424	0.812	1.637	2.081	2.782
Q17	-0.165	-2.491	-1.723	-1.295	-0.920	-0.243	0.149	0.369	1.168	1.233	2.103
Q18	-0.120	-2.248	-1.742	-1.208	-0.742	-0.348	0.062	0.466	1.003	1.405	2.153
Q19	0.420	-1.153	-1.088	-1.060	-0.431	0.286	0.525	0.624	1.741	1.989	2.764
Q20	0.371	-1.416	-0.988	-0.903	-0.659	0.383	0.103	0.881	1.535	2.184	2.592

Table 5: the Value of the Difficulty Parameter For Each Item

Apart from the discrimination parameter and the difficulty parameter, the subsequent analysis tests the item fit of the test items that have been developed. Item fit is an essential consideration in developing and using IRT-based tests. An item is called fit if it has a p.S_X^2 value > 0.05 (Dewanti, Hadi and Nu'man, 2021). The qualified items for the instruments developed using the GPCM model are according to Table 6.

Item	S_X^2	df.S_X^2	RMSEA.S_X^2	p.S_X^2
Q1	189.061	166	0.015	0.106
Q2	171.598	148	0.017	0.090
Q3	163.302	147	0.014	0.169
Q4	149.883	150	0.000	0.487
Q5	142.553	143	0.000	0.495
Q6	181.653	166	0.013	0.192
Q7	125.160	127	0.000	0.530
Q8	140.389	130	0.012	0.252
Q9	161.297	149	0.012	0.232
Q10	148.672	137	0.012	0.234
Q11	158.512	136	0.017	0.091
Q12	157.265	132	0.018	0.066
Q13	151.752	132	0.016	0.115
Q14	164.438	138	0.018	0.062
Q15	143.844	131	0.013	0.209
Q16	166.823	142	0.017	0.076
Q17	140.685	141	0.000	0.492
Q18	142.472	134	0.010	0.292
Q19	179.288	157	0.016	0.107
Q20	140.152	141	0.000	0.504

Table 6: Test Instrument Fit Items

DISCUSSION

The nine categories in Table 3 show that the instrument construct meets the fit category so that it can be concluded that the instrument construct is proven valid (construct validity is fulfilled). According to Figure 2, the CFA model's output shows that all loading factors are positive and significant; each item measures the relevant factor (Kwahk and Lee, 2008). The standardized loading factor obtained also ranges from 0.90 to 0.96, which indicates that these items significantly contribute to factor construction or purification of construct validity; according to Igbaria et al. (1997), loading factor > 0.3 is significant, loading factor > 0.4 is more important, and loading factor ≥ 0.5 is very significant.

Calculating the estimated value of construct reliability uses construct reliability (CR) using factor loading values and unique error indexes obtained from the CFA results. The construct is reliable if the CR obtained is more significant than 0.70 (Ghozali and Fuad, 2008). The calculation results show that the CR is 0.99; this indicates that the construct is proven reliable.

Before being analyzed with GPCM, it must first test the assumptions of response theory consisting of unidimensional tests, local independence, and parameter invariance. The purpose of the unidimensional assumption test is to count the number of items or questions designed to measure a test construct that genuinely represents one dimension or the construct that is structured and does not take advantage of other unrelated dimensions. Unidimensional assumptions are essential to ensure that the scores generated from items are meaningful and reliable and that measured construct representations are valid (Hambleton and Swaminathan, 1985; Hambleton et al., 1991; Linden and Hambleton, 1997) "Fundamentals of Item Response Theory" introduces the basics of item response theory (IRT). Scree plot exploratory factor analysis in Figure 3 shows that one factor is measured in the chemical five-tier diagnostic test instrument. This shows that one dominant factor is obtained to fulfill the unidimensional assumption.

The second assumption of IRT is local independence. Local independence is an illustration if the correlation between pairs of items is only caused by the main trait or ability that is measured by a series of test items and is not influenced by some traits or abilities that are not modeled that affect the two items (Demars, 2010). According to Hambleton et al. (1991) "Fundamentals of Item Response Theory" introduces the basics of item response theory (IRT) the assumption of local independence will be automatically proven if the unidimensional test has been proven. This means that this chemical five-tier diagnostic test instrument meets the assumption of local independence.

The third assumption of IRT is the invariance of item and capability parameters. Parameter invariance is a concept in the measurement field that refers to a parameter model of consistency or similarity across groups or subpopulations (Millsap and Kwok, 2004). The invariance of the item parameters is seen from the level of difficulty and differential power because, in the GPCM IRT model, there are two parameters, namely differential power and difficulty level, so it is necessary to look at the invariance of item parameters from difficulty level and differential power. The item invariance parameters for the difficulty level and differential power are based on the distribution of students in schools with a and B accreditation. The invariance of item parameters based on the questions' differential power and difficulty level are shown in Figures 4 and 5. In addition to the item invariance parameters, the invariance parameters are also determined by dividing even and odd questions. The invariance of the ability parameters is shown in Figure 6. If each point is close to slope line 1, then this indicates that there is no parameter variation (Drasgow and Mattern, 2006; Hambleton, 2006; Hambleton and Swaminathan, 1985; Hambleton et al., 1991; Linden and Hambleton, 1997) "Fundamentals of Item Response Theory" introduces the basics of item response theory (IRT). Figures 4, 5, and 6 show that the close points are with the red line, which is slope 1. This indicates that there is no variance in the estimation result parameters.

The IRT model used in this study is the GPCM model. GPCM is an IRT polytomous model used to estimate the probability of a person responding to a test item at a certain difficulty level. The GPCM is more flexible than the Rasch model, as it allows for different levels of difficulty between items and different ability levels between people. The GPCM model has parameters different from the Rasch model; the GPCM model has discriminant parameters and parameter difficulties (Muraki, 1992).

The discrimination parameter measures how well an item discriminates between people with different ability levels. Hambleton et al. (1991) "Fundamentals of Item Response Theory" introduces the basics of item response theory (IRT) state that good grains have discrimination parameters greater than zero and less than or equal to 2. Table 4 shows that all discrimination parameters have a value of $1.1 \leq a \leq 2$, meaning that all of these items are good because they can distinguish between students with high ability and those with low ability.

The difficulty parameter in GPCM measures how difficult a question is (Muraki, 1992). Hambleton et al. (1991) "Fundamentals of Item Response Theory" introduces the basics of item response theory (IRT state that a good item difficulty index is -2 to +2. Based on Table 5, it was found that all items with difficulty parameters b1 to b10 had values from most minor to most prominent, and the average b was in the range -2 to +2; this shows that the items developed were excellent and informative about students' abilities.

The subsequent analysis is to test the fit items of the test items that have been developed. Item fit is an essential consideration in developing and using IRT-based tests. There are various statistics to determine item fit that can be used to assess item fits, such as infit statistics, outfit statistics, standardized residuals, S-X^2 fit index, and many other references. This study used the S-X^2 fit index to determine fit items because they correspond to polytomous items in educational and psychological research (Kang and Chen, 2011). An item is called fit if it has a p.S_X^2 value > 0.05 (Dewanti et al., 2021). Table 6 shows that all items from Q1 to Q20 have a p.S_X^2 value greater than 0.05; this indicates that all items developed in the five-tier chemical diagnostic test instrument fit and are by the GPCM model. Based on Table 6, all items are appropriate, but if you wish to change the number of items to be used,

removing items with a p.S_X^2 value close to 0.05 is better. For example, in this instrument, two items can be omitted, namely Q14 and Q16. Item Q14 has a p.S_X^2 value of 0.062, and Q16 has a p.S_X^2 value of 0.076. If you only want to use 18 of the 20 available items, you should delete items with a p.S_X^2 value close to 0.05.

Apart from item analysis, another essential thing to note in GPCM is test information and standard measurement errors. Test information refers to the ability of test items to distinguish between individuals with different levels of ability. In contrast, the standard error measurement refers to the amount of uncertainty associated with the estimation of individual ability, where the standard error measurement is the inverse of the square root of the test information so that the greater the information, the smaller the standard error and the greater the reliability (Demars, 2010). The information function of the test and standard error is presented in Figure 7.

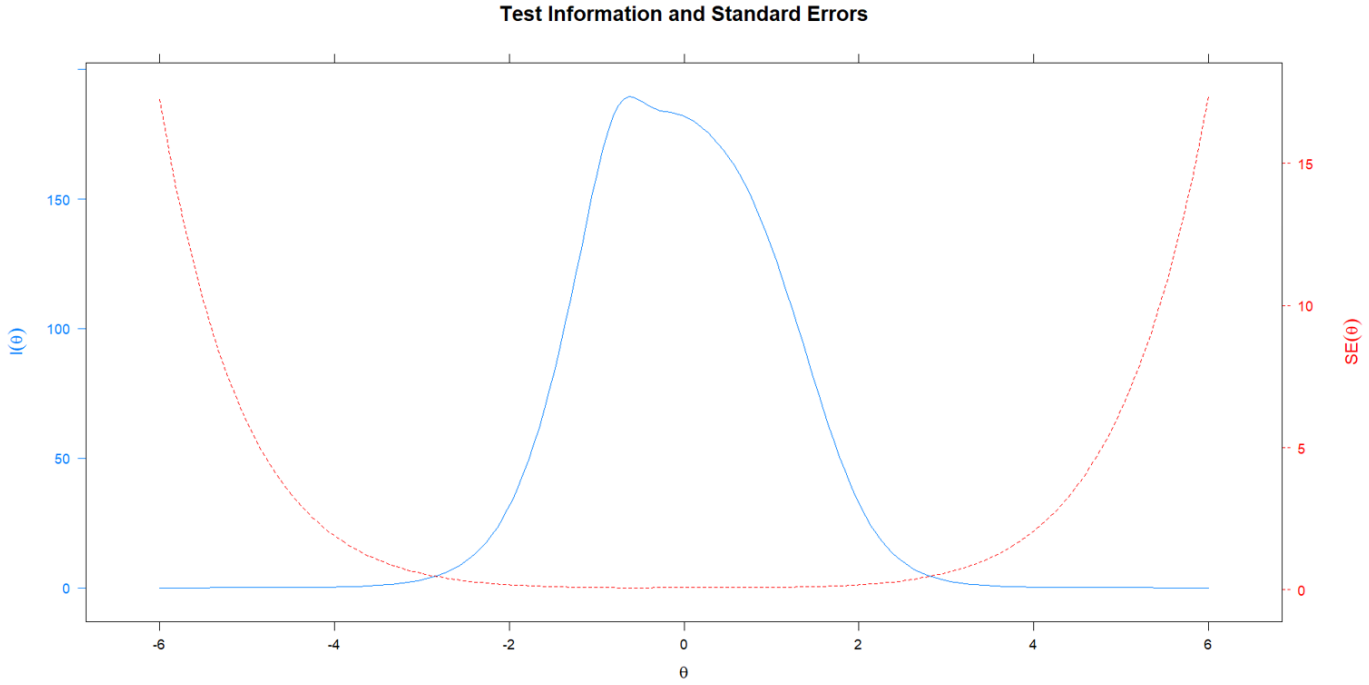

Figure 7: Information Function Test And Standard Error

Figure 7 shows that the theta obtained from the information function and standard error measurement on the CBT-based five-tier diagnostic test instrument has the intersection of the lines at the low limit with a theta of -2.85 and at the upper limit with a theta of +2.80. These results indicate that the CBT-based five-tier diagnostic test is suitable for students who have abilities between -2.85 and +2.80.

Table 5 shows that the most challenging item is item 9, and the easiest item is item 4. Item 4 contains questions about chemical problems related to electron configurations when electrons are released. This problem was made based on the consideration that there is a misconception among students who think that the 2 electrons released in $26Fe^{2+}$ come from the 3d orbital rather than the 4s orbital because they are farther from the nucleus (Kay et al., 2010). Item 9 contains questions about the analysis of the relationship between atomic number and the periodicity of elements (atomic radii) based on data on the periodicity of elements. This problem is made because all students believe that the size of the atomic radius increases down and to the right on the periodic table (Nicoll, 2001). Display of item 4 and item 9 on the computer-based test according to Figures 8 and 9.

The items' difficulty level analysis results can be interpreted into the Item Characteristic Curve (ICC). ICC is a graph that shows the relationship between the probability of correct answers from participants and their level of ability in a particular domain in the Item Response Theory model, which serves to determine the level of difficulty of a test item, determine the differentiability of a test item, and assess the quality of a test item. ICC shows the characteristics of the difficulty level of the items in the form of a curve of the relationship between the probability of answering correctly 50% and the level of student ability. The ICC item with the lowest difficulty level is item 4, which has a value of -0.362, and the greatest difficulty index is item 9, which is 0.431. Display ICC item 4 and item 9 according to Figures 10 and 11.

The ICC shown in Figures 10 and 11 shows that item 4 can be answered by students with a minimum θ ability of -2.901. In contrast, item 9 can be answered if they have a minimum θ ability of -1.382, meaning that to be able to work on item 9 a student must have higher abilities than when working on item 4.

Comparison of test results based on gender is analyzed using differential item functioning (DIF) according to Figure 12. DIF shows the probability that supports certain items between males and females (Tie, Chen and He, 2022).

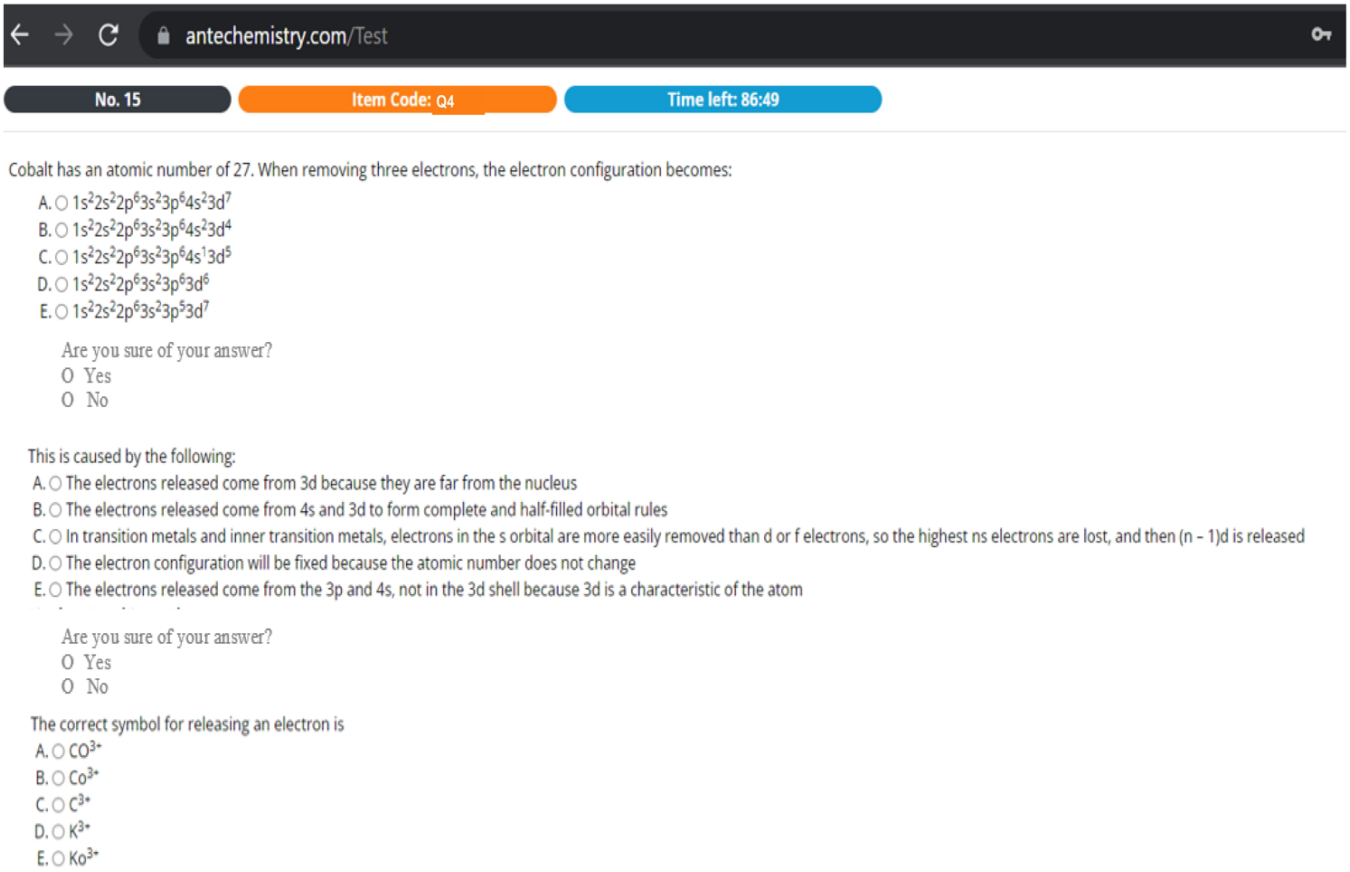

Figure 8: Display items 4

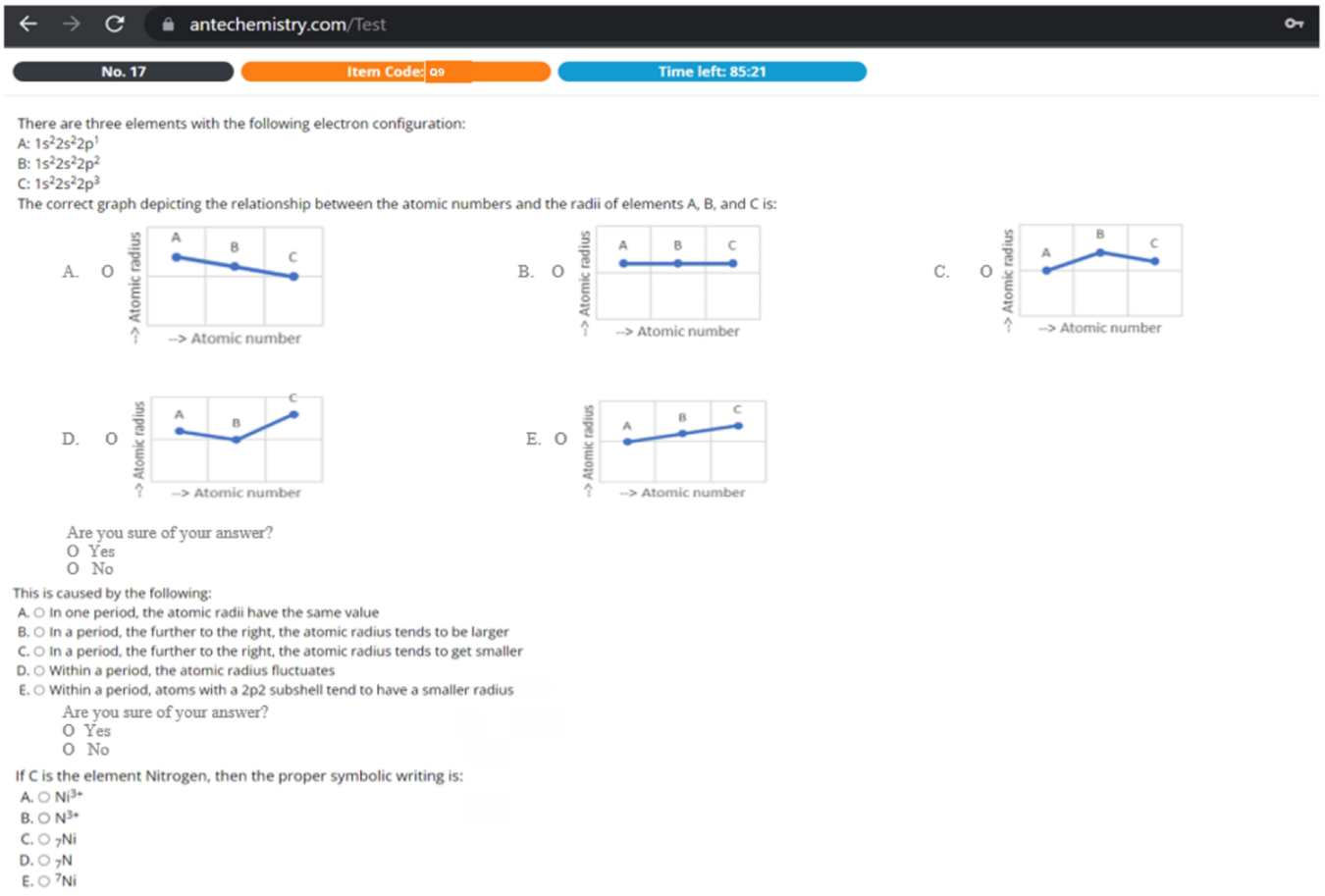

Figure 9: Display items 9

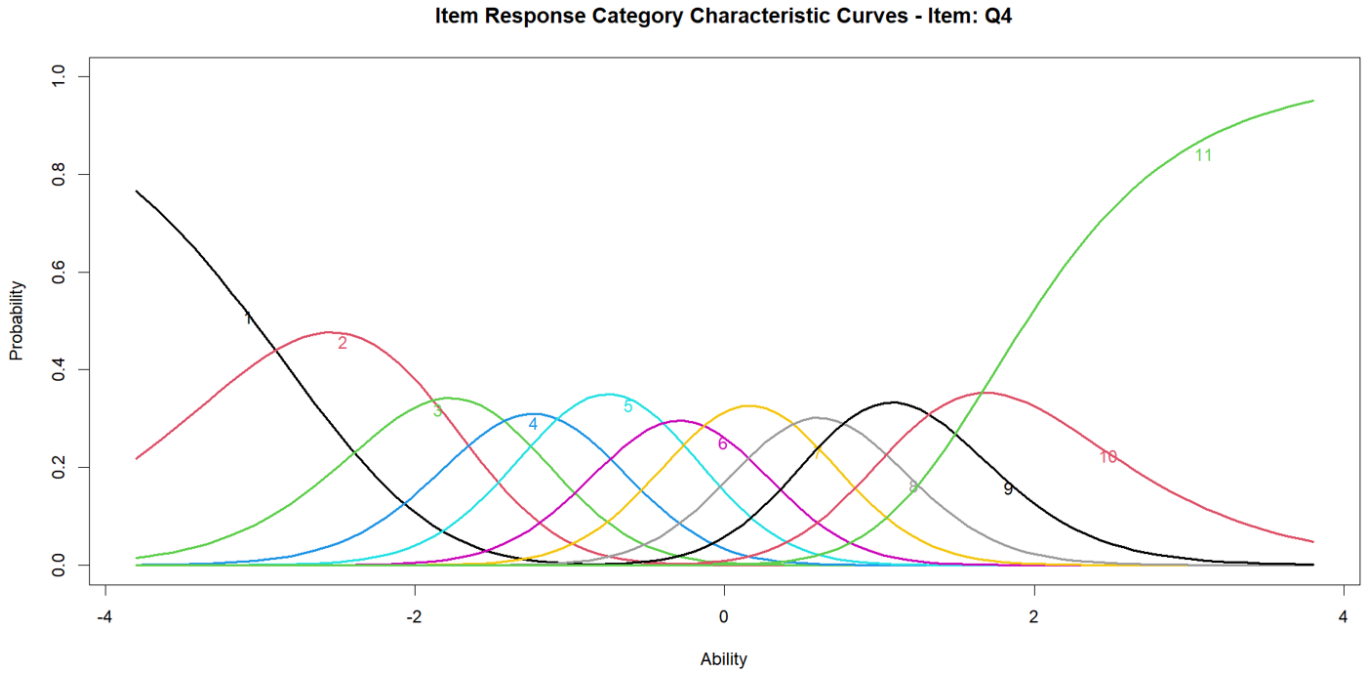

Figure 10: Item Characteristic Curve (ICC) item 4

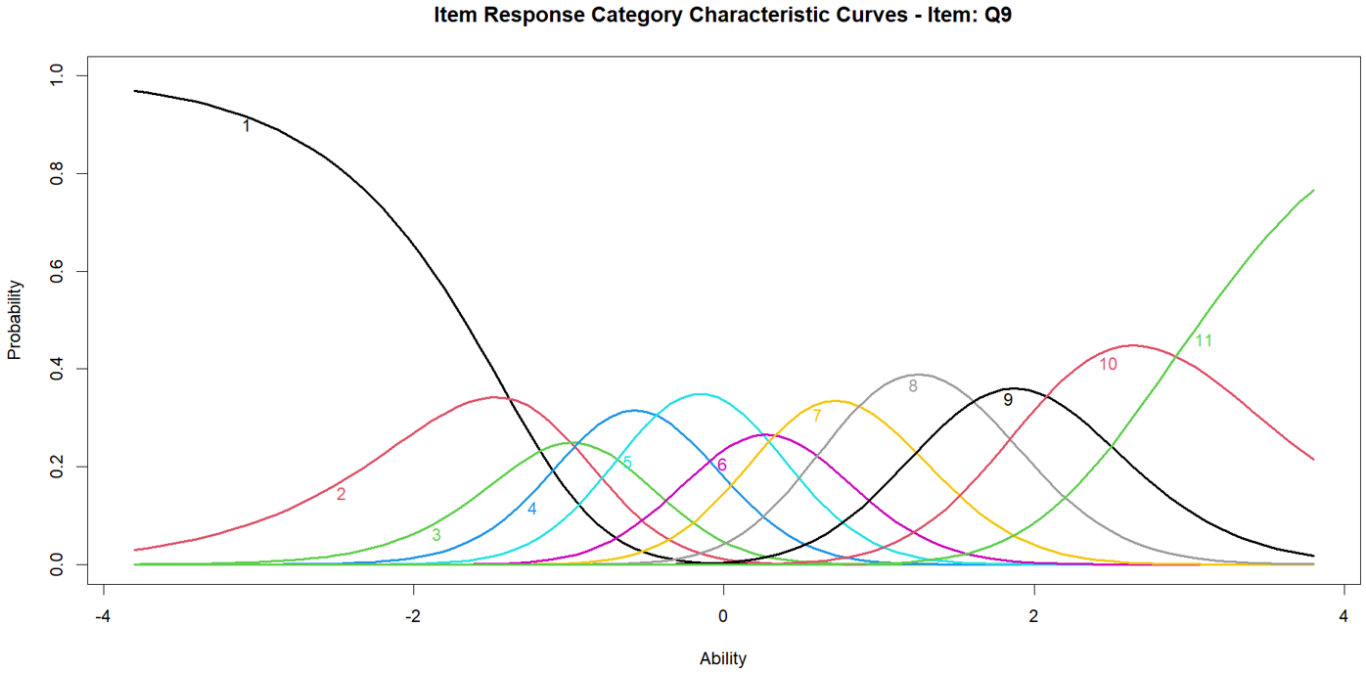

Figure 11: Item Characteristic Curve (ICC) item 9

Based on Figure 12, it can be seen that Q9 items are close to the upper limit, and Q4 items are close to the lower limit. Items close to the upper limit mean items have great difficulty, and items close to the lower limit indicate easy items. Figure 12 also shows that females find it easier to work on items in Q1, Q5, Q7, Q9, Q10, Q13, Q16, Q17, and Q20, while males find it easier to work on items in Q2, Q3, Q4, Q6, Q8, Q11, Q12, Q14, Q15, Q18, and Q19.

CONCLUSION

This research succeeded in developing a five-tier chemical diagnostic test instrument based on a computer-based test. The test instrument consists of five levels of questions: the first is the central question, the second is the confidence level, the third is the reason for the main question, the fourth is the confidence level for a reason, and the fifth is symbolic in chemistry related to the main question. The developed chemical test instruments cover three main chemistry sections: macroscopic, microscopic, and symbolic chemistry.

The test results of the test instruments on a sample of 580 students showed that the test instruments developed had good validity and reliability and could distinguish between different student abilities based on each student's ability θ. Instrument development uses the IRT approach with the GPCM model for

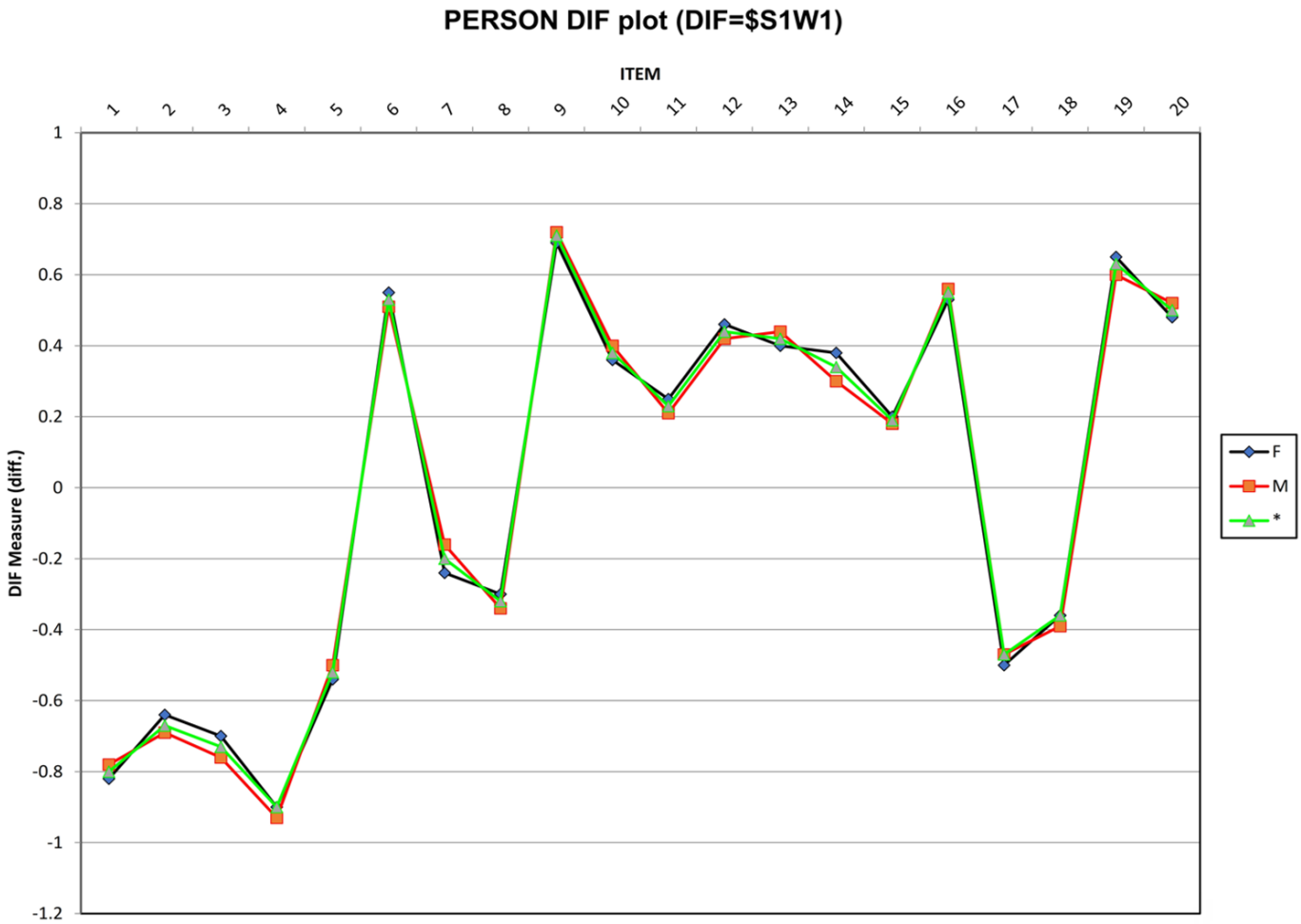

Figure 12: Differential Item Functioning (DIF) male and females

polytomous. This research makes an essential contribution to developing test instruments to detect students' misconceptions in chemistry and can be used as a reference in developing further misconception test instruments in the future.

This instrument consists of 20 items, and in one item, there are five levels of questions that can measure students' knowledge abilities which are divided into 11 ability categories consisting of no knowledge, complete misconception, partial misconception, guess knowledge, type 5 (lack of knowledge), type 5 (lack of knowledge), type 4 (lack of knowledge), type 3 (lack of knowledge), type 2 (lack of knowledge), type 1 (lack of knowledge), understand but lack confidence, and understand. This instrument has gone through content validity tests, constructs, criteria, and reliability tests. The reliability test results show a construct reliability value of 0.99, indicating that this instrument can be relied upon in measuring students' abilities, and the Aiken test score > 0.70 suggests that the instrument developed is essential and follows the curriculum.

The results of this study can be used as a reference for chemistry teachers to find out their students' abilities and find out where their students' misconceptions are. Future research can use this research as a reference in developing misconception instruments in other fields.

This study provides information about valid and reliable misconception instruments. The limitation of this research is that the research subjects are still in one country. This research can be expanded by using samples from various countries.

REFERENCES

Aiken, L. R. (1980) 'Content validity and reliability of single items or questionnaires', *Educational and Psychological Measurement*, Vol. 40, No. 4, pp. 955-959. https://doi.org/10.1177/001316448004000419

Anam, R. S., Widodo, A., Sopandi, W. and Wu, H. K. (2019) 'Developing a five-tier diagnostic test to identify students' misconceptions in science: an example of the heat transfer concepts', *Elementary Education Online*, Vol. 18, No. 3, pp. 1014-1029. https://doi.org/10.17051/ilkonline.2019.609690

Barke, H. D., Hazari, A. and Sileshi Yitbarek (2009) *Misconception in Chemistry*, Berlin: Springer.

Bayuni, T. C., Sopandi, W. and Sujana, A. (2018) 'Identification misconception of primary school teacher education students in changes of matters using a five-tier diagnostic test', *Journal of Physics: Conference Series*, Bandung, Vol. 1013, p. 012086 https://doi.org/10.1088/1742-6596/1013/1/012086

Chi, S., Wang, Z., Luo, M., Yang, Y. and Huang, M. (2018) 'Student progression on chemical symbol representation abilities at different grade levels (Grades 10-12) across gender', *Chemistry Education Research and Practice*, Vol. 19, No. 4, pp. 1055-1064. https://doi.org/10.1039/c8rp00010g

Cornick, J. E. (2015) 'Factor structure of the exercise self-efficacy scale', *Measurement in Physical Education and Exercise Science*, Vol. 19, No. 4, pp. 208-218. https://doi.org/10.1080/1091367X.2015.1074579

Crocker, L. and Algina, J. (2008) *Introduction to Classical and Modern Test Theory*, 2nd edition, Mason: Cengage Learning.

Dagnall, N., Denovan, A., Parker, A., Drinkwater, K. and Stephen Walsh, R. (2018) 'Confirmatory factor analysis of the inventory of personality organization-reality testing subscale', *Frontiers in Psychology*, Vol. 9, 1116, pp. 1-12. https://doi.org/10.3389/fpsyg.2018.01116

Debelak, R., Stobl, C. and Zeigenfuse, M. D. (2022) *An Introduction to the Rasch Model with Examples in R*, New York: CRC Press. https://doi.org/10.1201/9781315200620

Demars, C. (2010) *Item Response Theory*, New York: Oxford University Press.

Derkach, T. M. (2021) 'The origin of misconceptions in inorganic chemistry and their correction by computer modelling', *Journal of Physics: Conference Series*, Kryvyi Rih, Vol. 1840, pp. 1-13. https://doi.org/10.1088/1742-6596/1840/1/012012

Dewanti, S. S., Hadi, S. and Nu'man, M. (2021) 'The application of item response theory in analysis of characteristics of mathematical literacy test items', *İlköğretim Online*, Vol. 20, No. 1, pp. 1226-1237. https://doi.org/10.17051/ilkonline.2021.01.119

Dewi, F. C., Parlan, P. and Suryadharma, I. B. (2020) 'Development of four-tier diagnostic test for identifying misconception in chemical equilibrium', *AIP Conference Proceedings*, Vol. 2215, pp. 1-6. https://doi.org/10.1063/5.0000510

Drasgow, F. and Mattern, K. (2006) 'New tests and new items: Opportunities and issues', in Bartram, D. and Hambleton, R. K. (ed.) *Computer-Based Testing and the Internet*, Chichester: John Wiley & Sons Ltd.

Ghozali, I. and Fuad (2008) *Structural equation modeling: teori, konsep, dan aplikasi dengan Program Lisrel 8.80*, Semarang: Badan Penerbit Universitas Diponegoro.

Groen, M. M. van and Eggen, T. J. H. M. (2020) 'Educational Test Approaches: The Suitability of Computer-Based Test Types for Assessment and Evaluation in Formative and Summative Contexts', *Journal of Applied Testing Technology*, Vol. 21, No. 1, pp. 12-24.

Hambleton, R. K. (2006) 'Psychometric Models, Test Designs and Item Types for the Next Generation of Educational and Psychological Tests', in Bartram, D. and Hambleton, R. K. (ed.) *Computer Based Testing and the Internet*, Chichester: John Wiley & Sons Ltd.

Hambleton, R. K. and Swaminathan, H. (1985) *Item Response Theory: Principles and Applications*, New York: Springer Science+Business Media. https://doi.org/10.1007/978-94-017-1988-9

Hambleton, R. K., Swaminathan, H. and Rogers, H. J. (1991) *Fundamentals of Item Response Theory Library*, London: Sage Publications.

Harza, A. E. K. P., Wiji, W. and Mulyani, S. (2021) 'Potency to overcome misconceptions by using multiple representations on the concept of chemical equilibrium', *Journal of Physics: Conference Series*, Jawa Barat, Vol. 1806, pp. 1-6. https://doi.org/10.1088/1742-6596/1806/1/012197

Igbaria, M., Zinatelli, N., Cragg, P. and Cavaye, A. L. M. (1997) 'Personal computing acceptance factors in small firms: A structural equation model', *MIS Quarterly*, Vol. 21, No. 3, pp. 279-305. https://doi.org/10.2307/249498

Istiyono, E., Dwandaru, W. S. B., Erfianti, L. and Astuti, W. (2020) 'Applying CBT in physics learning to measure students' higher order thinking skills', *Journal of Physics: Conference Series*, Yogyakarta, Vol. 1440, p. 012061. https://doi.org/10.1088/1742-6596/1440/1/012061

Istiyono, Edi, Dwandaru, W. S. B., Fenditasari, K., Ayub, M. R. S. S. N. and Saepuzaman, D. (2023) 'The development of a four-tier diagnostic test based on modern Test theory in physics education', *European Journal of Educational Research*, Vol. 12, No. 1, pp. 371-385. https://doi.org/10.12973/eu-jer.12.1.371

Jusniar, J., Effendy, E., Budiasih, E. and Sutrisno, S. (2021) 'Eliminating misconceptions on reaction rate to enhance conceptual understanding of chemical equilibrium using EMBE-R strategy', *International Journal of Instruction*, Vol. 14, No. 1, pp. 85-104. https://doi.org/10.29333/IJI.2021.1416A

Kang, T. and Chen, T. T. (2011) 'Performance of the generalized S-X2 item fit index for the graded response model', *Asia Pacific Education Review*, Vol. 12, No. 1, pp. 89-96. https://doi.org/10.1007/s12564-010-9082-4

Kay, C. C., Yiin, H. K., Chu, C. K. and Hong, K. Y. (2010) 'Misconceptions in the teaching of chemistry in secondary schools in Singapore & Malaysia', *Proceedings of the Sunway Academic Conference*, Petaling Jaya, No. 1996, pp. 1-10.

Kimmons, R., Clark, B. and Lim, M. (2017) 'Understanding web activity patterns among teachers, students and teacher candidates', *Journal of Computer Assisted Learning*, Vol. 33, No. 6, pp. 588-596. https://doi.org/10.1111/jcal.12202

Kiray, S. A. and Simsek, S. (2020) 'Determination and evaluation of the science teacher candidates' misconceptions about density by using four-tier diagnostic test', *International Journal of Science and Mathematics Education*, Vol. 19, No. 5, pp. 935-955. https://doi.org/10.1007/s10763-020-10087-5

Kwahk, K. Y. and Lee, J. N. (2008) 'The role of readiness for change in ERP implementation: Theoretical bases and empirical validation', *Information and Management*, Vol. 45, No. 7, pp. 474-481. https://doi.org/10.1016/j.im.2008.07.002

Lamichhane, R., Reck, C. and Maltese, A. V. (2018) 'Undergraduate chemistry students' misconceptions about reaction coordinate diagrams', *Chemistry Education Research and Practice*, Vol. 19, No. 3, pp. 834-845. https://doi.org/10.1039/c8rp00045j

Linden, W. J. van der (2018) *Handbook of Item Response Theory Volume Three: Applications*, Boca Raton: CRC Press.

Linden, W. J. and Hambleton, R. K. (1997) *Handbook of Modern Item Response Theory*, New York: Springer.

Llanos, J., Fernández-Marchante, C. M., García-Vargas, J. M., Lacasa, E., De La Osa, A. R., Sanchez-Silva, M. L., De Lucas-Concuegra, A., Garcia, M. T. and Borreguero, A. M. (2021) 'Game-based learning and just-in-time teaching to address misconceptions and improve safety and learning in laboratory activities', *Journal of Chemical Education*, Vol. 98, No. 10, pp. 3118-3130. https://doi.org/10.1021/acs.jchemed.0c00878

Lowyck, J. (2014) *Handbook of Research on Educational Communications and Technology* New York: Springer. https://doi.org/10.1007/978-1-4614-3185-5

Marsh, H. W., Balla, J. R. and Mcdonald, R. P. (1988) 'Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size', *Psychological Bulletin*, Vol. 103, No. 3, pp. 391-410. https://doi.org/10.1037/0033-2909.103.3.391

Martin, F. and Betrus, A. K. K. (2019) *Digital media for learning: Theories, processes, and solutions digital media for learning: Theories, processes, and solutions*, Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-030-33120-7

Merriënboer, J. J. G. van and Bruin, A. B. H. de (2014) 'Research Paradigms and Perspectives on Learning', in Spector, J. M., Merrill, M. D., Elen, J. and Bishop M. J. (ed.) *Handbook of Research on Educational Communications and Technology*, New York: Springer.

Mills, C. n. and Breithaupt, K. J. (2016) *Educational Measurement From Foundations to Future*, New York: The Guilford Press.

Millsap, R. E. and Kwok, O. M. (2004) 'Evaluating the impact of partial factorial invariance on selection in two populations', *Psychological Methods*, Vol. 9, No. 1, pp. 93-115. https://doi.org/10.1037/1082-989X.9.1.93

Muraki, E. (1992) 'A generalized partial credit model: Application of an EM algorithm', *Applied Psychological Measurement*, Vol. 16, No. 2, pp. 159-176. https://doi.org/10.1177/014662169201600206

Murniningsih, Muna, K. and Irawati, R. K. (2020) 'Analysis of misconceptions by four tier tests in electrochemistry, case study on students of the chemistry education study program UIN Antasari Banjarmasin', *Journal of Physics: Conference Series*, Vol. 1440, pp. 1-6. https://doi.org/10.1088/1742-6596/1440/1/012008

Mutlu, A. and Sesen, B. A. (2015) 'Development of a two-tier diagnostic test to assess undergraduates' understanding of some chemistry concepts', *Procedia - Social and Behavioral Sciences*, Vol. 174, pp. 629-635. https://doi.org/10.1016/j.sbspro.2015.01.593

Nicoll, G. (2001) 'A report of undergraduates' bonding misconceptions', *International Journal of Science Education*, Vol. 23, No. 7, pp. 707-730. https://doi.org/10.1080/09500690010025012

Oriondo, L. L. and Antonio, E. M. D. (1984) *Evaluating Educational Outcomes (Tests, Measurement and Evaluation)*, Manila: Rex Book Store.

Özmen, H. (2004) ‹Some student misconceptions in chemistry: A literature review of chemical bonding', *Journal of Science Education and Technology*, Vol. 13, No. 2, pp. 147-159. https://doi.org/10.1023/b:jost.0000031255.92943.6d

Pokorný, M. (2023) 'Experience with online learning of mathematics in primary education', *International Journal of Emerging Technologies in Learning (IJET)*, Vol. 18, No. 02, pp. 203-213. https://doi.org/10.3991/ijet.v18i02.35401

Potvin, P. (2023) 'Response of science learners to contradicting information: A review of research', *Studies in Science Education*, Vol. 59, No. 1, pp. 67-108. https://doi.org/10.1080/03057267.2021.2004006

Prodjosantoso, A. K., Hertina, A. M. and Irwanto (2019) 'The misconception diagnosis on ionic and covalent bonds concepts with three tier diagnostic test', *International Journal of Instruction*, Vol. 12, No. 1, pp. 1477-1488. https://doi.org/10.29333/iji.2019.12194a

Putra, A. S. U., Hamidah, I. and Nahadi (2020) 'The development of five-tier diagnostic test to identify misconceptions and causes of students' misconceptions in waves and optics materials', *Journal of Physics: Conference Series*, Vol. 1521, No. 2. https://doi.org/10.1088/1742-6596/1521/2/022020

Rupp, A. A. and Zumbo, B. D. (2006) 'Understanding parameter invariance in unidimensional IRT models', *Educational and Psychological Measurement*, Vol. 66, No. 1, pp. 63-84. https://doi.org/10.1177/0013164404273942

Rusmini, Suyono, Jatmiko, B. and Yonata, B. (2021) 'The diagnosis of misconception on the concept of acid-base theory in prospective teacher students used a three-tier test', *Journal of Physics: Conference Series*, Vol. 1899, No. 1, pp. 1-7. https://doi.org/10.1088/1742-6596/1899/1/012061

Suparman, A. R., Rohaeti, E. and Wening, S. (2022) 'Development of attitude assessment instruments towards socioscientific issues in chemistry learning', *European Journal of Educational Research*, Vol. 11, No. 4, pp. 1947-1958. https://doi.org/10.12973/eujer.11.4.1947

Suparman, A. R., Rohaeti, E. and Wening, S. (2023) 'Effect of computer based test on motivation: A meta-analysis', *European Journal of Educational Research*, Vol. 12, No. 4, pp. 1583-1599. https://doi.org/10.12973/eu-jer.12.4.1583

Tie, B., Chen, G. and He, J. (2022) 'Validation of the inflexible eating questionnaire in a large sample of Chinese adolescents: psychometric properties and gender-related differential item functioning', *Eating and Weight Disorders*, Vol. 27, No. 3, pp. 1029-1037. https://doi.org/10.1007/s40519-021-01239-9

Tien, L. T. and Osman, K. (2017) *Overcoming Students' Misconceptions in Science*, Singapore: Springer Nature. https://doi.org/10.1007/978-981-10-3437-4

Treagust, D., Chittleborough, G. and Mamiala, T. (2003) 'The role of submicroscopic and symbolic representations in chemical explanations', *International Journal of Science Education*, Vol. 25, No. 11, pp. 1353-1368. https://doi.org/10.1080/0950069032000070306

Üce, M. and Ceyhan, İ. (2019) ‹Misconception in chemistry education and practices to eliminate them: Literature analysis', *Journal of Education and Training Studies*, Vol. 7, No. 3, pp. 202. https://doi.org/10.11114/jets.v7i3.3990

Vladusic, R., Bucat, R. B. and Ozic, M. (2022) 'Understanding covalent bonding - a scan across the Croatian education system', *Chemistry Education Research and Practice*, Vol. 23, No. 4. https://doi.org/10.1039/D2RP00039C

Wang, F., Franco-Penya, H.-H., Kelleher, J. D., Pugh, J. and Ross, R. (2017) 'An analysis of the application of simplified silhouette to the evaluation of k-means clustering validity', *International Conference Machine Learning and Data Mining in Pattern Recognition*, Vol. 13, pp. 291-305. https://doi.org/10.1007/978-3-319-62416-7_21

Wang, Z., Yan, W., Zeng, C., Tian, Y. and Dong, S. (2023) 'A unified interpretable intelligent learning diagnosis framework for learning performance prediction in intelligent tutoring systems', *International Journal of Intelligent Systems*, Vol. 2023, pp. 1-20. https://doi.org/10.1155/2023/4468025

Widarti, H. R., Permanasari, A., Mulyani, S., Rokhim, D. A. and Habiddin (2021) 'Multiple representation-based learning through cognitive dissonance strategy to reduce student's misconceptions in volumetric analysis', *TEM Journal*, Vol. 10, No. 3, pp. 1263-1273. https://doi.org/10.18421/TEM103-33