MALAYSIAN JOURNAL OF LEARNING AND INSTRUCTION
https://e-journal.uum.edu.my/index.php/mjli

How to cite this article:
Solikhin, F., Wiyarsi, A., Muhson, A., & Retnawati, H. (2026). Psychometric evaluation and item response theory validation of a higher-order thinking skills test on buffer solution concepts in chemistry. *Malaysian Journal of Learning and Instruction*, *23*(2), 142-165. https://doi.org/10.32890/mjli2026.23.2.7

PSYCHOMETRIC EVALUATION AND ITEM RESPONSE THEORY VALIDATION OF A HIGHER-ORDER THINKING SKILLS TEST ON BUFFER SOLUTION CONCEPTS IN CHEMISTRY

[1]Febrian Solikhin, [2]Antuni Wiyarsi, [3]Ali Muhson & [4]Heri Retnawati
[1,3&4]Educational Research and Evaluation Department, Graduate School, Universitas Negeri Yogyakarta, Yogyakarta, Indonesia
[1]Chemistry Education Study Program, Faculty of Teacher Training and Education, Universitas Bengkulu, Bengkulu, Indonesia
[2]Chemistry Education Department, Faculty of Mathematics and Natural Science, Universitas Negeri Yogyakarta, Indonesia

[1]Corresponding author: febriansolikhin.2023@student.uny.ac.id

Received: 5/1/2025 Revised: 25/5/2026 Accepted: 27/7/2026 Published: 31/7/2026

ABSTRACT

Purpose – This study aims to evaluate and validate a higher-order thinking skills (HOTS) test instrument for buffer solution concepts using the generalized partial credit model (GPCM). While HOTS is critical in chemistry, there is a lack of specialized instruments validated against modern psychometric theories to accurately assess students' analytical and evaluative abilities with regard to buffer solutions.

Methodology – A quantitative research approach was employed. The instrument used was a test instrument with an 8-item polytomous instrument. It was administered to 288 high school students in Bengkulu City, Indonesia. Data were analyzed using an item response theory (IRT) - GPCM to estimate item parameters (discrimination and difficulty) and test information function.

Findings – The results confirmed that 7 out of 8 items exhibited good fit to the GPCM. These items cover a wide range of difficulty levels, effectively distinguishing between students with varying abilities in HOTS. The test provided maximum information at slightly above-average abilities ($\theta \approx 0.49$), and remained across

the low to high ability range, making it a precise tool for mapping a wide variation in student abilities in HOTS, although less than optimal for truly extreme abilities.

Novelty – This study bridges modern psychometrics and chemistry pedagogy by utilizing the GPCM to map student cognitive load during multi-step buffer solution reasoning. Unlike previous assessments limited to a standard ability range, the instrument used crossed the cognitive threshold and offered teachers a high-precision assessment tool.

Significance – The study contributes to the field of chemistry education by providing a psychometrically sound instrument for diagnosing HOTS. Unlike traditional classical test theory, the application of the GPCM allows teachers to identify specific conceptual gaps in student reasoning regarding buffer solutions, enabling more targeted instructional interventions.

Keywords: Higher-order thinking skills, buffer solutions, chemistry education, item response theory, generalized partial credit model, test validation.

INTRODUCTION

Higher order thinking skills or HOTS can be used as an important learning strategy in growing and developing critical and analytical thinking skills. HOTS is one of the skills needed today (Chamidah et al., 2024; Pratama & Retnawati, 2018). This ability can be integrated into various pedagogical approaches, such as problem-based, inquiry-based, and discovery-based learning. Learning by integrating this ability can train students in critical thinking, creativity, and problem solving (Muntari et al., 2021; Wildan et al., 2021). HOTS can reflect the way students think critically and logically (Shanmugam & Rajoo, 2020). Therefore, at the secondary school level, the development of these abilities is focused on learning (Rahman, 2019). It is aimed at preparing students to engage effectively in academic and employment activities after graduation (Algouzi et al., 2023; Hai & Thi, 2021). In addition, this ability can be used as an indicator to build a skilled and qualified workforce (Ismail et al., 2022).

Every learning process requires appropriate and accurate assessment, including the use of HOTS assessment. This assessment is used to determine students' HOTS achievement (Saepuzaman et al., 2022). The development and implementation of effective assessment tools for HOTS are essential for improving the quality of education across various disciplines, including chemistry. In chemistry teaching and learning, buffer solutions are one of the most conceptually challenging and cognitively demanding topics for students. A recent cross-grade study identified buffer solutions as one of the most challenging topics for eleventh-grade students (Pratomo et al., 2025). The difficulty in mastering the concept of buffer solutions stems not only from the calculations, but also from the reasoning required when using this concept.

One of the main challenges in assessing HOTS is the need for authentic assessment aligned with the competencies students are expected to achieve. However, it appears to be difficult for educators to find suitable assessment alternatives in evaluating this ability (Saepuzaman et al., 2022; Widiantie et al., 2022). HOTS assessment includes complex cognitive processes that go beyond memorization and the requirements for application, analysis, and evaluation of information in a meaningful context (Phakiti, 2018). HOTS can

be mapped using Bloom's Revised Taxonomy, a two-dimensional framework which explicitly links higher-order cognitive processes (specifically, analyze – C4, evaluate – C5, and create – C6) with the conceptual and procedural knowledge dimensions inherent in the concept (Anderson & Krathwohl, 2001). These cognitive levels and types of knowledge serve as a reference for developing HOTS-based test instruments, specifically for the buffer solutions concept.

Based on the concept of integration in the cognitive process, the 'analyze' level requires students to break down complex chemical systems into their component parts, specifically distinguishing between weak acids/bases and their conjugate pairs, and to determine how these components interact when small amounts of strong acids or bases are added. The 'evaluate' level requires students to make critical judgments based on scientific criteria, such as assessing and criticizing the efficiency of a system's buffer capacity from a titration curve or a non-ideal chemical environment, while the 'create' level requires students to combine these chemical principles to form a functional whole, which involves creating alternative chemical formulations or planning appropriate laboratory procedures to design buffer solutions with specific, targeted pH values (Anderson & Krathwohl, 2001).

HOTS-based instruments are still rarely found in daily tests or semester exams. This is because many teachers are not used to developing this type of instrument (Saepuzaman et al., 2022). HOTS is one of the 21st century skills that need to be developed. Every instrument need good validity and reliability, so that the instruments developed are suitable for use, including the HOTS-based instrument (Triastuti & Laksono, 2021). The instrument development process requires data analysis to reflect the characteristics of each of the item developed. One of the theories that can be used to do this is the item response theory or IRT. The IRT emerged as a statistical approach that allows the modeling of the relationship between individuals' latent traits and their item responses at test time. This framework provides advantages over classical test theory (CTT). The IRT can provide an understanding of item characteristics, including difficulty and discrimination parameters, and is useful in the development of valid and reliable instruments (Hendrickson & Mislevy, 2005; Nunes et al., 2017; Reise, 2016).

Currently, the IRT has evolved to accommodate complex items, such as polytomous items. These responses are characterized by having multiple response categories, rather than just binary choices. This model in polytomous IRT provides a framework for understanding responses to statistically modeled items to generate valid inferences about latent properties, such as student understanding or psychological constructs (Kim & Wilson, 2019; Penfield, 2014). The application of this latest model is also relevant for educational contexts with assessments designed to measure complex constructs. The polytomous IRT model provides a framework for analyzing data from items with three or more response options, offering a more detailed understanding of the relationship between latent traits and observed responses (Penfield, 2014; Stanke & Bulut, 2019; Sun et al., 2013). In addition, the polytomous IRT model is also able to provide detailed psychometric properties of the test items. This is because polytomous IRT can provide information about the components of difficulty, discrimination power, and response category function (Kim & Wilson, 2019; Penfield, 2014).

Polytomous models can be analyzed using the following: Partial credit model (PCM), graded response model (GRM), generalized partial credit model (GPCM) and nominal response model (NRM). The NRM model enables analysis of items with unordered response categories, providing flexibility in modeling

different types of categorical data (Preston & Reise, 2013; Qiu & Torre, 2023). The PCM model allows assessment of items that lack a strict ordinal structure and accommodates a wide variety of item formats (Kim & Wilson, 2019). The characteristics of this model allow for the calibration of test items that are free from the influence of the student sample used (Masters, 1982). Another model, the GRM can be used to analyze Likert-type scales, where respondents give ratings to categories (Stanke & Bulut, 2019). This model is specifically designed for items with ordered response categories. These categories reflect the level of agreement or frequency with the statement item. The model assumes that the probability of a particular response category is a function of the latent trait level and item parameters, including thresholds between response categories.

In contrast to the GRM, which focuses on cumulative probabilities for rigidly ordered response categories, the GPCM which was developed by Muraki (1992), offers a more flexible approach to measuring cognitive achievement because it allows item discrimination parameters to vary across items, unlike the standard PCM, which assumes constant discrimination. Based on the IRT framework by Muraki (1997), the formal expression for the probability of a learner with cognitive ability level (θ) obtaining a category score (x) on an item in the GPCM can be stated as in Equation 1:

$$P_{ix}(\theta) = \frac{\exp(\sum_{j=1}^{x} \alpha_i(\theta - \delta_{ij}))}{\sum_{r=0}^{m_i}\left[exp \sum_{j=1}^{r} \alpha_i(\theta - \delta_{ij})\right]} \qquad (1)$$

Where $\sum_{j=1}^{0} \alpha_i(\theta - \delta_{ij}) = 0$, α_i represents the item discrimination parameter i, while δ_{ij} is the difficulty parameter for j the response category on item i. For comparison, the formal expression for the GRM by Samejima (1997) operates through the cumulative probability of students scoring equal to or higher than category x, which is as formulated in Equation 2:

$$P_{ix}^{*}(\theta) = \frac{\exp[\alpha_i(\theta - \beta_{ij})]}{1 + \exp[\alpha_i(\theta - \beta_{ij})]} \qquad (2)$$

Although both models accommodate varying discrimination parameters, the GPCM is considered the most theoretically appropriate for the polytomous response of a cognitive test instrument. With reference to the view of Embretson and Reise (2000), the GPCM is specifically designed to estimate transition probabilities between scores on polytomous tests of academic ability or achievement. In the context of chemical reasoning on buffer solutions, the use of the GPCM provides a strong theoretical foundation because it is able to objectively assign partial score credits to each stage of the student problem solving process, where each increase in response category reflects a higher depth of conceptual understanding of the phenomenon.

Despite the importance of developing HOTS in chemistry learning, assessing these skills in complex topics presents significant diagnostic challenges. Traditional assessments rely largely on the CTT, which is sample-dependent and fails to provide accurate measurements across different student ability levels. As a result, teachers often lack objective data to identify the precise cognitive threshold of struggling students. Meanwhile, HOTS item inherently involve multi-step reasoning that can result in partial understanding. Therefore, a key issue of concern in the present research is the lack of psychometrically validated assessment tools specifically designed to map student responses to HOTS questions in buffer solutions without the confounding effect of sample bias. To address this measurement gap, this study aims to evaluate

and validate a HOTS test instrument on the buffer solutions concept using the GPCM. Specifically, the research objectives were formulated as follows: estimate HOTS item parameters, including item difficulty and discrimination index, by applying the polytomous GPCM framework; determine item fit and measurement precision across the latent trait continuum (θ) through the item and test information function (TIF); and estimate students' latent ability levels (θ) based on their HOTS performance, thus allowing for proper calibration between students' competencies and item difficulty on the same scale. By achieving these objectives, this study provides a powerful modern psychometric evaluation tool that offers in-depth diagnostic insights to higher-order chemistry teaching and assessment.

METHODOLOGY

Using a quantitative research approach, the present study has been conducted by testing and analyzing the responses for HOTS-based test instruments. To ensure data representation and avoid selection bias, cluster random sampling was used. The sampling procedure was carried out through several stages. The target population in this study comprised all registered public senior high schools in Bengkulu City, Indonesia. From a total population 11 schools, a sampling frame was used to determine the number of schools that met the criteria of having implemented the same national curriculum and had completed the topic of buffer solutions. In the first stage of cluster randomization, five (5) schools were randomly selected to represent the variety of schools in the region. Next, from the selected schools, a second-level randomization was conducted to determine which 11th-grade mathematics and natural science classes would be involved as participants in the study. From this procedure, eight (8) classes from five (5) different schools were finally selected, resulting in a total of 288 students as the final sample. According to prior studies, the lower bound for getting the parameter estimation accuracy to be stable in the GPCM is 250 (de Ayala, 2022; Finch & French, 2019). Distribution of participants included 93 male students (32.30%) and 195 female students (67.70%). All came from A-accredited public senior high school, with 108 students (37.5%) from schools in the city center and 180 students (62.5%) from schools outside the city center.

The instrument used was a HOTS-based test instrument with eight (8) open-ended questions. Examples of items and rubrics in this instrument can be seen in Table 1.

Table 1

Examples of the HOTS-based Test Instrument Items and Rubric

<table>
<tr><td>Indicator</td><td>Analyzing the minimal pH changes of a solution upon the addition of small amounts of acid or base to identify its buffering components and justify its mechanism</td></tr>
<tr><td>Item Code</td><td>Item2</td></tr>
<tr><td>Cognitive Level</td><td>C4 – Analyzing</td></tr>
<tr><td>Item</td><td>A student makes a buffer solution by mixing 200 mL of 0.2 M CH_3COOH (Ka= 2×10^{-5}) and 50 mL of 0.2 M NaOH. Then a little acid and a little base are added to the solution mixture to obtain the following pH.
<table><tr><td>Treatment</td><td>pH</td></tr><tr><td>Before addition</td><td>4.22</td></tr><tr><td>After addition a little acid</td><td>4.05</td></tr><tr><td>After addition a little acid</td><td>4.32</td></tr></table>Based on the pH data above, identify the components of the buffer solution and give reasons.</td></tr>
</table>

(continued)

Rubric	0	No answer / wrong answer.
	1	Able to make reactions with or without the amount of 'initial', 'reacted', 'remainder', but there are still errors and no explanation.
	2	Able to make reactions correctly accompanied by the correct amount of 'initial', 'reacted', 'remainder', but there is no explanation or does not make reactions and cannot explain the components of the buffer solution.
	3	Able to make reactions accompanied by the correct amount of 'initial', 'reacted', 'remainder' and can explain the components, but there are still a few errors in identifying them.
	4	Able to make reactions accompanied by the correct amount of 'initial', 'reacted', 'remainder', can explain the components, and can identify the components of the buffer solution correctly.

To ensure methodological transparency and illustrate the cognitive demands of the instrument, sample items for levels C5 – Evaluate and C6 – Create, with their corresponding scoring rubrics are provided in Appendix 2. This instrument has been declared valid based on the results of V Aiken by four experts in prior research (Suswanti et al., 2024). The magnitude of V Aiken and the distribution of cognitive levels can be seen in Table 2.

Table 2

Distribution of Cognitive Level and Aiken's V Value of Each Item

Item Code	Cognitive Level	V Aiken
Item1	C-4	0.86
Item2	C-4	0.86
Item3	C-6	0.95
Item4	C-4	0.95
Item5	C-4	0.97
Item6	C-5	0.86
Item7	C-5	0.86
Item8	C-4	0.88

The resulting data were then analyzed using polytomous IRT with the help of R-Studio with the 'mirt' package (Chalmers, 2012). This data analysis includes the following: model fit test, data adequacy test, assumption test, item parameter estimation, ability (θ) parameter estimation, and IRT reliability. Other packages used in this data analysis are as follows: dplyr (Wickham et al., 2026), car (Fox et al., 2023), psych (Revelle, 2023), CTT (Willse, 2026), TAM (Robitzsch et al., 2023), ltm (Rizopoulos, 2006), FactoMineR (Lê et al., 2008), factoextra (Kassambara & Mundt, 2020), openxlsx (Schauberger & Walker, 2025), and ggplot2 (Wickham, 2016).

Model fit testing uses three main models, namely the PCM, the GPCM, and the GRM. This is because the scoring used was ordinal data. The NRM model was not tested because it is designed for items with unordered alternative answers (Retnawati, 2014). A model is considered to have a better fit if it demonstrates lower AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) values, a higher LL (Log-Likelihood), and a greater number of fitting items (Dai et al., 2021; Kang et al., 2009; Orlando & Thissen, 2000). Subsequently, item fit was estimated only under the most suitable model. Item fit used a p-value reference that was more than 0.05. The number of items that match was then tested for

data adequacy. The data adequacy test used the Kaiser-Meyer-Olkin (KMO) test and this was followed by the Bartlett test to test data homogeneity. This test was carried out in order to provide a dimensionality assumption test.

The assumption test consists of the assumptions of unidimensionality, local independence, and parameter invariance. The unidimensionality assumption was evaluated using the formed eigenvalue. The dominance of a single factor was met if the main eigenvalue was able to explain a minimum of 20% of the variance, meanwhile, the local independence assumption was evaluated through the residual correlation index, Q3, which was below the threshold 0.50 (Harsana et al., 2024). If the eigenvalues are met, the data can be estimated using a suitable model. Item parameter estimation includes the parameters of discrimination, difficulty level, category characteristic curve (CCC), item information function (IIF), and test information function (TIF), while the ability parameter estimation includes the ability of the students who take this test. According to Baker and Kim (2017), the discrimination index can be categorized into very low (0.01-0.34); low (0.35-0.64); medium (0.65-1.34); high (1.35-1.69); and very high (more than 1.70). Meanwhile, according to Georgiev (2008), difficulty levels can be classified into very easy (less than -2); easy (-2 to -1); moderate (-1 to +1); difficult (+1 to +2); and very difficult (more than +2).

RESULTS AND DISCUSSION

Model Fit Test

The polytomous responses obtained were then tested for fit using the following three IRT models: the PCM, the GPCM, and the GRM. This test is used to get a model that best fit the empirical data and instrument characteristics. The comparison of model fit test results is as shown in Table 3.

Table 3

Comparison of Model Fit Test Results

Model	AIC	BIC	LL	Number of Fit Items
PCM	**6636.94**	**6757.81**	-3328.47	6
GPCM	6644.58	6791.10	**-3282.29**	**7**
GRM	6646.54	6793.06	-3283.27	7

Notes. AIC= Akaike Information Criterion; BIC= Bayesian Information Criterion; and LL= Log-Likelihood.

As can be seen from Table 3, the PCM meets goodness of fit when viewed in terms of AIC and BIC, while the GPCM fulfills it in terms of the LL and the number of fit items. According to Kang et al. (2009) and Orlando and Thissen (2000), the smaller AIC, BIC, the larger LL and the maximum number of fit items can be used as model fit selection. The model fit indices provide information about the model's fit relative (Nye et al., 2020). Numerically, the two models differ, but are not significant based on ANOVA testing. The final model selection is not only based on macro indices, but also considers the micro level (item fit), especially when the test length is short.

Based on the statistical evaluation of fit item, the GPCM can rescue one of the misfit items in the PCM, namely the only item with C6-level. The empirical response pattern on this item cannot be well explained by the PCM due to the assumption that all items have uniform discrimination power. This indicates that variation in discrimination power between items is a characteristic of the empirical data of this instrument. Substantively, the HOTS-based instrument involves more complex problem-solving and analysis, so the instruments can inherently produce variations in discrimination power (Brookhart, 2010; Hillsley, 2025). According to Muraki (1992), this is consistent with the characteristics of the GPCM, where the item step parameters of this model are decomposed into location and threshold parameters, following the rating scale formulation of Andrich (1978). Forcing the use of the PCM for global parsimony will actually maintain estimation bias in item parameters and reduce the validity of the overall measurement of student ability.

The GPCM model also provides RMSEA results of 0.06; CFI 0.96; and TLI 0.95. All three are also determinants of the right model. According to prior research, there are other considerations in determining a suitable model, namely from the RMSEA, TLI and CFI values (Lubbe, 2023; Maydeu-Olivares & Joe, 2014; Xia & Yang, 2019). The model is considered fit if the RMSEA is <0.08 and the CFI and TLI are>0.90 (Hartono et al., 2023). Based on this statement, the data analyzed fits the GPCM model. This stage also tested the suitability of items. Table 4 shows that there are seven (7) items that fit. The p-values for these seven items can be seen in Table 4.

Table 4

Item Suitability Test Results

Item Code	P-Value	Information
Item1	0.49	Fit
Item2	0.21	Fit
Item3	0.06	Fit
Item4	0.15	Fit
Item5	0.01	Misfit
Item6	0.08	Fit
Item7	0.34	Fit
Item8	0.19	Fit

As can be seen from Table 4, one item was misfit and was excluded from further analysis. All items in this study were developed based on a blueprint mapping of each item to buffer solution sub-topic domain and the HOTS cognitive level (see Appendix). This was to ensure that the set represented the full scope of the sub-topic domain and HOTS cognitive levels. The exclusion of Item5 did not compromise the overall structural coverage of the construct, as the representation of sub-topic domains and cognitive levels remained consistent with the test blueprint. However, given the reduced number of remaining items in specific blueprint cells, particularly at cognitive levels C5 and C6, measurement precision at these levels requires cautious interpretation.

Data Adequacy Test

This test was carried out by finding the correlation matrix between items and then finding the Kaiser-Meyer-Olkin (KMO) value and p-value for Bartlett's test. The overall KMO value obtained was 0.8 and the p-

value of the Bartlett test was 0.000. This proves that the data obtained is in good condition, so it can be continued to the next stage. According to prior research, a KMO value of > 0.5 indicates that the data is acceptable for further analysis (Cheng & Choy, 2007).

Assumption Test

The first assumption test is unidimensionality. This assumption refers to the number of dimensions formed from the data. The percentage of dimensions in this dataset is as shown in Figure 1.

Figure 1

Scree Plot of Variance Percentage Explained

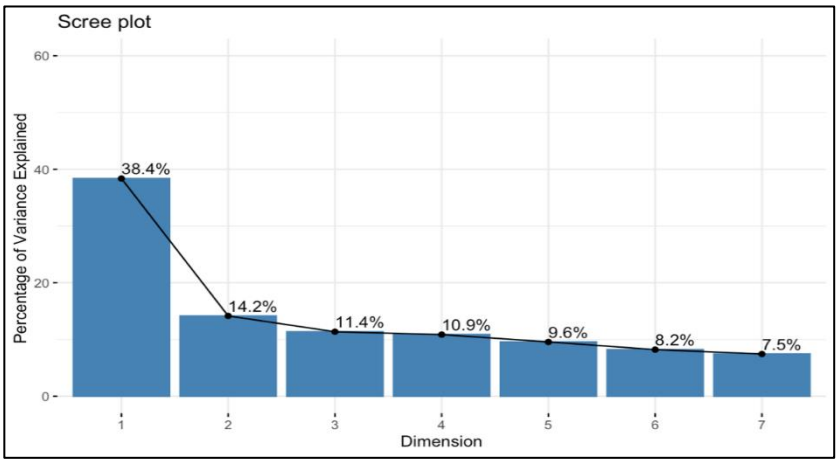


Figure 1 shows that the results obtained are the first dimension, which had a percentage of variance explained by 38.4%, with an eigenvalue of 2.68. This indicates that the analyzed data was unidimensional. According to Harsana et al. (2024), unidimensionality is met if the first-dimension accounts for at least 20%. In addition, Swaminathan et al. (2006) also explain that the percentage of the largest eigenvalue in unidimensional IRT is at least 25%.

The second assumption is local independence. Local independence asserts that responses to individual items are independent (Reise, 2014; Stojanov, 2018). The correlation matrix that uses residuals with Q3 can be seen in Table 5.

Table 5

Q3 Correlation Matrix

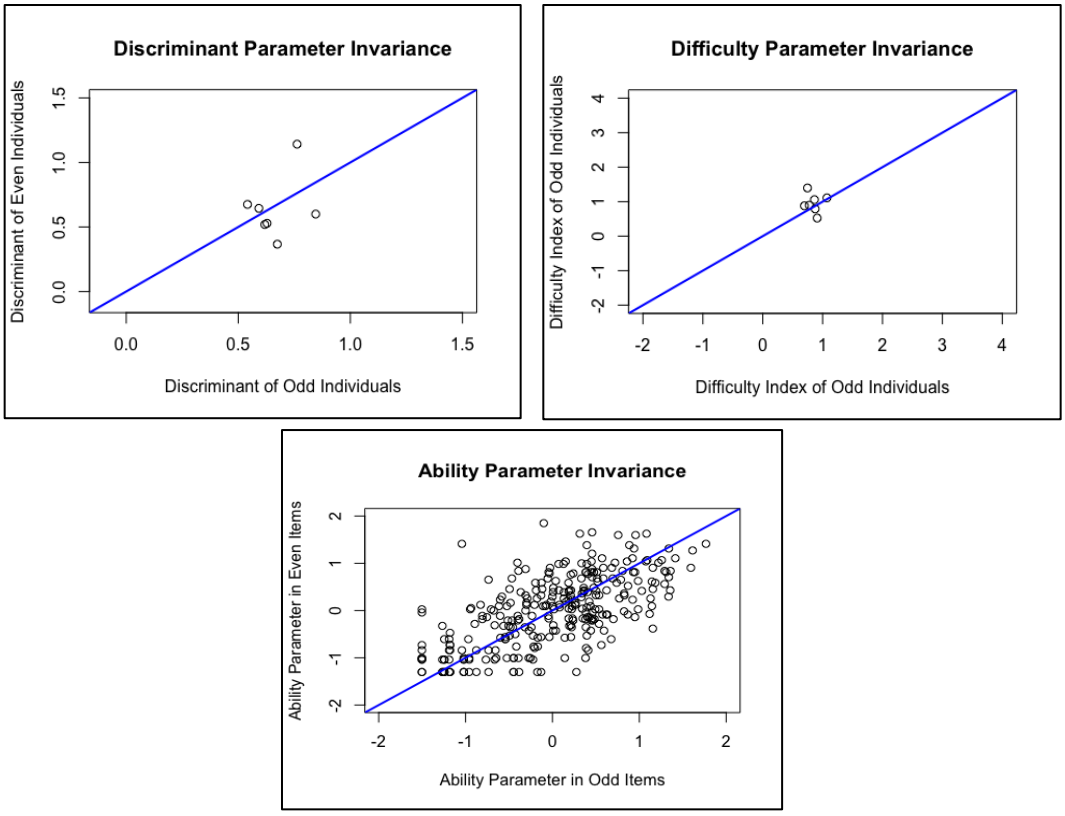
Item Code	Item1	Item2	Item3	Item4	Item6	Item7	Item8
Item1	1.00	-0.01	-0.17	-0.07	-0.19	-0.15	-0.15
Item2	-0.01	1.00	0.01	-0.27	-0.25	-0.19	-0.11
Item3	-0.17	0.01	1.00	-0.14	-0.13	-0.24	-0.12
Item4	-0.07	-0.27	-0.14	1.00	-0.17	-0.19	-0.15
Item6	-0.19	-0.25	-0.13	-0.17	1.00	0.09	-0.08
Item7	-0.15	-0.19	-0.24	-0.19	0.09	1.00	0.01
Item8	-0.15	-0.11	-0.12	-0.15	-0.08	0.01	1.00

As can be seen from Table 5, the correlation matrix generally shows that the correlation was not high, although there was a negative correlation. If there is a rather high item correlation, it indicates that the items being measured are the same (Chen & Thissen, 1997). According to Harsana et al. (2024), this assumption is met if the correlation between items is less than 0.5. It is usually met when the unidimensionality assumption has also been met (Hambleton et al., 1991; Retnawati, 2014).

The third assumption is parameter invariance. This parameter invariance refers to the invariance of item parameter, including discriminant and difficulty invariance and ability invariance. The results of testing this assumption can be seen in Figure 2.

Figure 2

Parameter Invariance Testing



As can be seen from Figure 2, the assumption of parameter invariance was satisfied because the test points lay around the line. This test categorizes the data into two different groups. The invariance of this discriminant parameter indicates how well the items distinguish individuals with different abilities. This parameter indicates how sharply items can distinguish between individuals with low and high abilities. The invariance of the difficulty parameter describes how well items can be done with various categories, while the invariance of the ability parameter expresses the ability of individuals to do different items. Based on the test, this instrument meets all assumptions, so it can proceed with the estimation of each parameter.

Item Parameter Estimation

This test can produce item characteristics derived from the data obtained. Discriminant values and item difficulty levels can be seen in Table 6.

Table 6

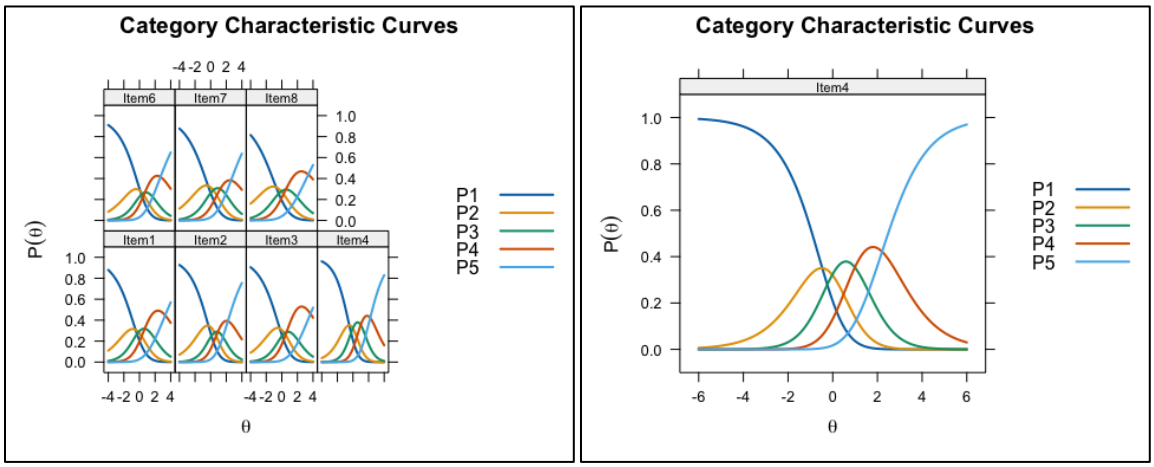
Discriminant Index (a) and Difficulty Level (b) of Each Item

Item Code	a	Category	b1	b2	b3	b4	b_{mean}	Category
Item1	0.59	Low	-0.50	-0.12	0.67	3.29	0.92	Moderate
Item2	0.69	Moderate	-0.28	0.49	0.93	2.20	0.90	Moderate
Item3	0.62	Low	-0.26	0.35	0.61	3.67	1.17	Difficult
Item4	0.92	Moderate	-0.47	-0.02	0.99	2.21	0.79	Moderate
Item6	0.58	Low	0.10	0.40	0.68	2.69	1.01	Difficult
Item7	0.56	Low	-0.33	0.34	1.23	2.60	0.99	Moderate
Item8	0.50	Low	-0.73	-0.06	0.51	3.39	0.89	Moderate

This estimation test also produces category characteristics curves (CCC). This CCC is a graph of each category on each item. The CCC for this item estimation can be seen in Figure 3.

Figure 3

Category Characteristic Curves for Each Item and Item4

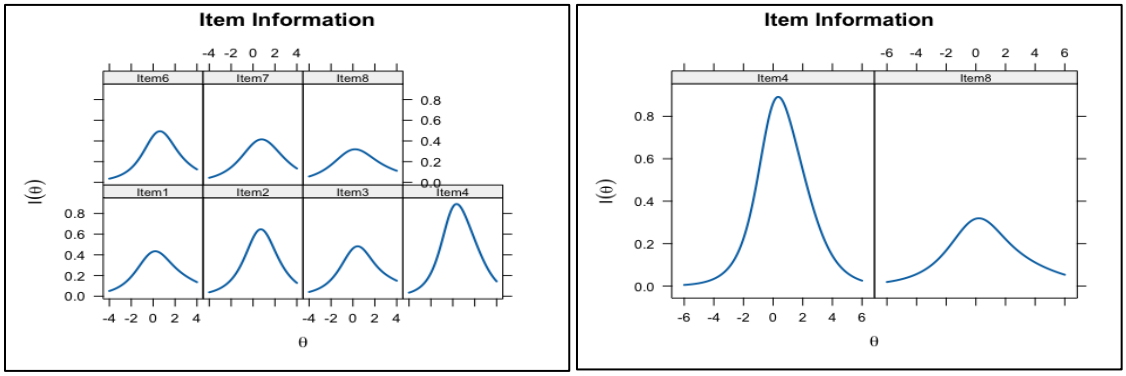


Notes. θ = latent ability level; P(θ) = probability of obtaining a given score category at ability level θ; P1 = probability of obtaining a score of 0, P2 = probability of a score of 1, P3 = probability of a score of 2, P4 = probability of a score of 3, and P5 = probability of a score of 4.

In addition to the characteristic curves, this analysis also produces item information curves (IIC). This information curve item can be seen in Figure 4.

Figure 4

Item Information Curve for Each Item and Comparison between Two Items



Item information can also produce the proportion of each item that determines the amount of information provided by the item in a certain ability range relative to the total information provided by the item at all ability levels. In this case, the ability range used was from -4 to 4. The proportion of each item and the test in the ability range of -4 to 4 can be seen in Table 7.

Table 7

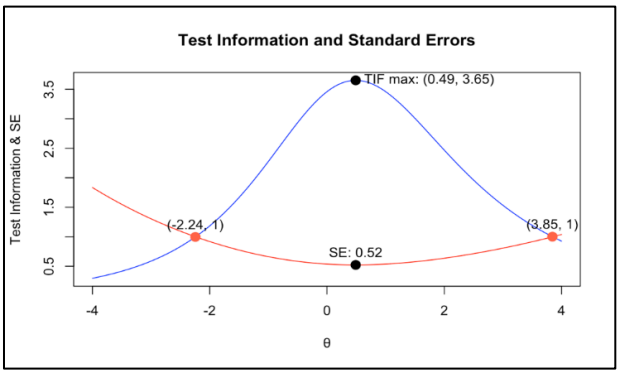
Information on Each Item and Test

Item Code	Info	Total Info	Proportion	Percentage (%)
Item1	2.01	2.38	0.84	84%
Item2	2.53	2.77	0.91	91%
Item3	2.08	2.48	0.83	83%
Item4	3.46	3.66	0.94	94%
Item6	2.04	2.33	0.87	87%
Item7	1.91	2.23	0.85	85%
Item8	1.60	1.99	0.80	80%
Overall	15.64	17.86	0.87	87%

Besides the item information, the other curves generated were the test information (TIF) and standard error (SE). The information for the entire test and the standard error can be seen in Figure 5.

Figure 5

Test Information Curve and Standard Error



The seven retained items function as unified measurement units. The low-to-moderate discrimination power of individual items was compensated for by a well-distributed range of difficulty thresholds, spanning from below-average to well-above-average ability level, enabling the instrument to effectively map student ability without relying on a single cutoff point. This pattern aligned with the test information (TIF) curve in Figure 5, which exceeded the standard error (SE) of measurement across the target ability range (Harsana et al., 2024; Pardede et al., 2023). The cutoff points indicate that this precision did not apply evenly across the ability spectrum. The instrument remained reliable for both below-average and high-ability students within the range, but was less than optimal for those at the extremes beyond this range. Item8, for instance, yielded the least information individually, yet remained supported by the instrument's cumulative precision (Tan, 2024).

These characteristics can only be adequately captured by a measurement model that treats each response category as a separate unit of information. The IRT model for polytomous responses allows predicting the probability of a respondent being in a particular category based on their level of a latent trait, while simultaneously generating a characteristic curve for each item (Stanke & Bulut, 2019). The selection of the GPCM as the final model is a logical consequence of the variation in α values across items, a situation that the PCM cannot account for, because the GPCM treats the discrimination of each item independently (Jo et al., 2023). The GPCM estimates the discrimination parameter separately for each item, allowing for the adequate modeling of differences in items' ability to distinguish between individual ability levels (Muraki, 1992) and provides more accurate information about the performance of each item within the measurement framework (Penfield, 2010). The proportion of information generated by each item serves as the basis for assessing item performance and the overall test performance across a given ability range (Mellenbergh, 2014), and on this basis, these parameters can be used to refine items to better align with the intended construct (Bolt & Adams, 2016; García-Pérez, 2014).

Assessing HOTS regarding buffer solutions cannot be adequately achieved through dichotomous measures or polytomous models that assume uniform item discrimination. This validated instrument advances the HOTS assessment in buffer solution topics within chemistry education by providing a more precise mapping of student abilities at the upper end of spectrum, while also diagnosing cognitive transition points that have previously been difficult to capture with earlier HOTS instruments. Compared with prior research on HOTS assessment in chemistry education, which has generally utilized the PCM (Marlina et al., 2025; Rampean & Rohaeti, 2025), the GPCM approach in the present study expanded the parameterization to become more flexible and could provide psychometric insights into students' cognitive structures when they were engaged in multi-step reasoning on buffer solutions. This difference was also evident in the range of abilities successfully mapped; prior studies generally reported student ability distributions concentrated in the range of -2 to +2, whereas the instrument in this study penetrated higher cognitive thresholds. This finding is in line with results in Saepuzaman et al. (2022), who reported that HOTS ability estimates using the GPCM could cover a range beyond -2 to 2, suggesting that the flexibility of the GPCM parameterization contributed to the instrument's accuracy in reaching advanced students, not just students in the typical ability range. Specifically, within chemistry education, this provides teachers with a measurement tool that more precisely differentiates student performance regarding the staged reasoning involved in buffer solutions. The tool goes beyond merely confirming basic conceptual understanding to diagnose the transition toward the evaluative and creative reasoning that characterizes HOTS in chemistry, especially in buffer solution topics. For the broader field of educational measurement, these findings suggest that

polytomous IRT models with flexible parameterization, such as the GPCM, merit consideration as a standard for developing HOTS instruments, given their capacity to capture variations in item discrimination often overlooked by simpler models. This advantage has also been confirmed in a direct comparison of the GPCM with other polytomous IRT models using real test data (Wallmark et al., 2024).

This widespread difficulty pattern was consistent with the instrument's content coverage, which encompassed all subtopics of buffer solutions (see Appendix 1). Items at the creation level (C6) and evaluation level (C5) were more difficult than items at the analysis level (C4), confirming that C6 and C5 occupied a higher cognitive position than C4 (Anderson & Krathwohl, 2001). However, this taxonomic hierarchy does not necessarily guarantee a uniform order of difficulty. The result reveals that Item8, while at the C5 level, was actually easier than items at the C4 level items. This difference suggests that cognitive level alone is not sufficient to predict the empirical difficulty of an item. The other factors outside the taxonomy also play a role in determining the level of difficulty of HOTS items perceived by students, i.e the level of interactivity and contextual familiarity (Clarke & Roche, 2018; Sweller, 1994). Each topic has intrinsic properties that cause differences in the level of element interactivity and cognitive load (Sweller, 1994), while the familiar and contextual problems have been shown to help students produce correct answers (Salgado, 2017). This is because it is closely related to the problem solver's prior knowledge (Chi et al., 2023). Some C4 level questions in this instrument actually had a deeper analytical reasoning structure, thus increasing cognitive load, while C5 level questions were structured in a context that was more familiar with students' reasoning patterns, making them easier to answer even though they were at a higher formal taxonomic level. The familiar context was present through laboratory practical experiences, which had become part of the students' prior knowledge, making it easier for them to answer items framed within that context. These two factors indicate that cognitive processes in chemistry learning do not always proceed linearly following a taxonomic hierarchy, but are instead strongly influenced by the complexity of the content and the familiarity of the context associated with each question.

The HOTS-based instrument can serve as a diagnostic map that shows which students have mastered procedural competencies, but have not yet achieved reasoning, so that teachers can follow up through targeted chemistry learning (Su, 2021). This instrument reveals where students' difficulty thresholds lie, not just that students struggle with the topic in general, providing evidence that teachers can use to redesign instructional sequences in buffer solutions learning. The third threshold (b3) was only slightly above average, but the fourth threshold (b4) was significantly above it, resulting in many students stopping at a score of 3 due to difficulty justifying their written answer in writing. This finding provides a specific signal that students need to be encouraged to analyze, evaluate, and create more deeply about buffer solution data. The shift from calculation skills to justification skills cannot be treated as an automatic effect of more calculation practice. Bridging tasks therefore, are needed that explicitly train students to compare two equally correct calculations and articulate the reasoning behind them. This is because HOTS is not simply memorization, but rather involves mental processes embedded in problem-solving (Maryani et al., 2022). This kind of information is much more useful for teachers than the final score alone. Teachers can provide feedback on each assessment (Black & Wiliam, 1998). The feedback not only identifies the need for remediation, but also identifies the appropriate form of remediation.

At the curriculum level, the validation approach applied to this buffer solution instrument could also serve as a consideration for curriculum designers in science, particularly chemistry, when developing HOTS

assessments for other chemistry topics. This was empirically confirmed by Rohim et al. (2025), who found a significant increase in HOTS in chemistry learning through a curriculum-oriented scientific approach. Thinking skills such as HOTS need to be clearly outlined in curriculum documents so that teachers understand what is being measured, how to teach them, and how to assess them (FitzPatrick & Schulz, 2015). They are most effectively taught when integrated into curriculum design and the teaching processes (Alkhatib, 2019; Vidergor, 2018; Yosepha et al., 2023), including intentionally presenting contexts close to students' experiences when developing new HOTS questions and assignments, as demonstrated by the familiarity context finding above. At the same time, teachers need to understand the limits of this HOTS instrument. For example, for students with significantly below-average abilities, the instrument needs to be complemented by other, more sensitive diagnostic approaches, such as structured interviews or similar approaches, to prevent overused beyond its empirically tested precision range. Integrating HOTS into the chemistry curriculum and instruction can help foster the development of knowledge in students that is more constructive and applicable across various contexts (Su, 2021).

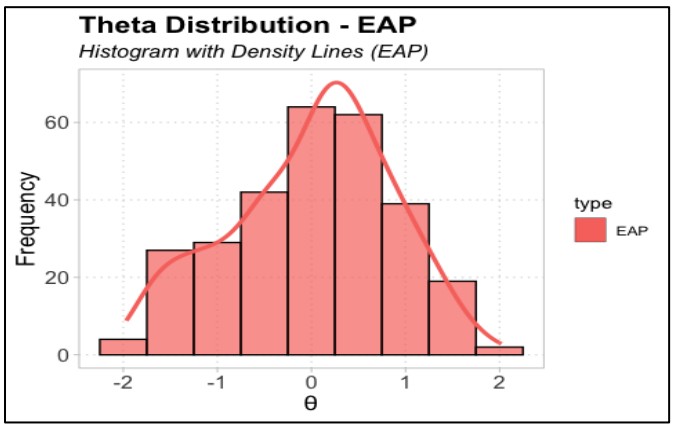
Moreover, these findings complement the literature on learning difficulties in chemistry, particularly buffer solutions (e.g. Pratomo et al., 2025), by pinpointing the transition point for students from procedural competence to integrative reasoning. This information is difficult to obtain from typical qualitative or quantitative approaches. This polytomous IRT methodology has the potential to be replicated to other chemistry topics with similar structures, where procedural fluency and conceptual integration are often conflated in classroom assessments, such as stoichiometry, equilibrium, redox reactions, or other topics.

Ability Parameter Estimation

This estimation is a method for calculating a student's ability level in a dimension based on their response patterns to a series of items with specific characteristics. Each parameter in IRT provides a more accurate estimate of individual ability than conventional methods. This estimation considers the item's level of difficulty and the pattern of student responses, so that the results better reflect students' true ability. The EAP (Expected A Posteriori) estimation method is used because it tends to produce estimates with a considered prior distribution. This distribution illustrates the density of student ability. The distribution can be seen in Figure 6.

Figure 6

Distribution of Students' Ability



As can be seen from Figure 6, the EAP estimates show a smoother peak and more balanced distribution, without any upper-boundary (too easy) or lower-boundary (too difficult) effects. This pattern indicates that the instrument is able to representatively map variations in students' HOTS abilities, with the majority of students falling in the intermediate HOTS ability category, and therefore, still requiring instructional stimulation to break through the extreme reasoning threshold on the buffer solution material.

Empirical reliability was recorded at 0.741 and marginal reliability at 0.737. Empirical reliability refers to the variation in scores estimated in the sample data, while marginal reliability refers to the distribution of overall population ability. The closeness of two values indicates that the sample is fairly representative of the population. These two values are in the acceptable category because they exceed 0.7 (Sumintono, 2015; Tavakol & Dennick, 2011), confirming that this test is consistent and able to provide a good estimate in measuring certain abilities (Eliaumra et al., 2022; Pada et al., 2018). These values also indicate stable and consistent results.

CONCLUSION

Based on the discussions of the study results and the foregoing section, seven of the eight HOTS items on buffer solutions fit the GPCM, with Item4 showing the best measurement characteristics and Item8 the weakest. The HOTS-based buffer solution test instrument was considered reliable (0.741 for empirical reliability and 0.737 for marginal reliability) and provided optimal information cumulatively through the combination of all items, thereby proving effective for measuring student HOTS capabilities within the context of buffer solutions. Estimating student ability using the EAP method places the item difficulty parameters and student ability on the same logit scale, allowing for a direct calibration between student competence and item difficulty. Theoretically, the effectiveness of this instrument reveals that the empirical difficulty level of HOTS items is not always linear with the hierarchy in Bloom's Taxonomy, but is also conditioned by the complexity of the content's cognitive load and student familiarity with the laboratory context. Practically, this instrument serves as a reliable formative diagnostic map (assessment for learning) for the topic of buffer solutions, allowing teachers to detect a student's specific difficulty threshold, especially in the transition from procedural calculation skills to integrative reasoning skills, and in order to justify the student answers. Teachers can design more relevant learning interventions based on this mapping, and curriculum designers can use it to more purposefully integrate HOTS into chemistry instructional design.

Despite this significance, the present study has some limitations because the instrument's optimal precision could not reach students' cognitive abilities at truly extreme levels outside its empirical range, and the material coverage was still limited to buffer solutions. Furthermore, the number of remaining test items was relatively limited (7 items), with minimal representation at cognitive levels C5 and C6, thereby limiting the precision of HOTS estimates at these levels. Based on these limitations, the agenda for future research needs to be directed toward several developmental avenues. First, the calibration and revision of individually weak items, such as Item8. This can be achieved by deliberately manipulating the novelty of the context and the interactivity of its elements to increase the cognitive load of the questions. Second, the stability and invariance of the instrument's parameters should be tested through large-scale trials, more heterogeneous student demographic and across regions, alongside an increase in the number of items at

cognitive levels C5 and C6 to enhance the instrument's precision and generalizability. Finally, there is also the need to replicate this polytomous IRT-based methodology on other chemistry topics with similar characteristics, such as stoichiometry, equilibrium, redox reactions or the other related topics, to expand the literature on mapping chemistry learning difficulties accurately.

CONFLICT OF INTEREST STATEMENT

No potential conflict of interest was reported by the authors.

ACKNOWLEDGMENT

The authors would like to express their gratitude to (1) the Center for Higher Education Funding and Assessment (Pusat Pembiayaan dan Asesmen Pendidikan Tinggi – PPAPT), The Ministry of Higher Education, Science and Technology (Kemdiktisaintek, Indonesia), and the Head of the Higher Education Financing Center, Kemdiktisaintek, who through a decree provided a scholarship under Decree Number 01826/BPPT/BPI.06/9/2023; (2) Indonesia Endowment Fund for Education (Lembaga Pengelola Dana Pendidikan – LPDP), Ministry of Finance, Indonesia for granting the Indonesian Education Scholarship (Beasiswa Pendidikan Indonesia – BPI) to the first author to complete his study.

REFERENCES

Algouzi, S., Alzubi, A. A. F., & Nazim, M. (2023). Enhancing EFL students' critical thinking skills using a technology-mediated self-study approach: EFL graduates and labor market in perspective. *PLoS ONE, 18*(10), e0293273. https://doi.org/10.1371/journal.pone.0293273

Alkhatib, O. J. (2019). A framework for implementing higher-order thinking skills (problem-solving, critical thinking, creative thinking, and decision-making) in engineering humanities. *2019 Advances in Science and Engineering Technology International Conferences (ASET)*, 8714262. Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/ICASET.2019.8714232

Anderson, L. W., & Krathwohl, D. R., (Eds.) (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Longman.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*(4), 561–573. https://doi.org/10.1007/BF02293814

Baker, F. B. & Kim, S.-H. (2017). *The basics of item response theory: Statistics for social and behavioral sciences*. Springer International Publishing.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy and Practice, 5*(1), 7–74. https://doi.org/10.1080/0969595980050102

Bolt, D. M., & Adams, D. J. (2016). Exploring rubric-related multidimensionality in polytomously scored test items. *Applied Psychological Measurement, 41*(3), 163–177. https://doi.org/10.1177/0146621616677715

Brookhart, S. M. (2010). *How to assess higher-order thinking skills in your classroom*. ASCD.

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

Chamidah, D., Suwono, H., Susilo, J., Sulisetijono, S., Munzil, M., & Zubaedah, S. (2024). Navigating the global landscape on project-based learning and 21st-century skills research (2020-2023): A bibliometric analysis. *Malaysian Journal of Learning and Instruction*, *21*(1), 217–247. https://doi.org/10.32890/mjli2024.21.2.7

Chen, W. H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, *22*(3), 265–289. https://doi.org/10.3102/10769986022003265

Cheng, T. C. E., & Choy, P. W. C. (2007). Measuring success factors of quality management in the shipping industry. *Maritime Economics and Logistics*, *9*(3), 234–253. https://doi.org/10.1057/palgrave.mel.9100179

Chi, S., Wang, Z., & Liu, X. (2023). Assessment of context-based chemistry problem-solving skills: Test design and results from ninth-grade students. *Research in Science Education*, *53*(2), 295–318. https://doi.org/10.1007/s11165-022-10056-8

Clarke, D., & Roche, A. (2018). Using contextualized tasks to engage students in meaningful and worthwhile mathematics learning. *Journal of Mathematical Behavior*, *51*, 95–108. https://doi.org/10.1016/j.jmathb.2017.11.006

Dai, S., Vo, T. T., Kehinde, O. J., He, H., Xue, Y., Demir, C., & Wang, X. (2021). Performance of polytomous IRT models with rating scale data: An investigation over sample size, instrument length, and missing data. *Frontiers in Education*, *6*, 1–18. https://doi.org/10.3389/feduc.2021.721963

de Ayala, R. J. (2022). *The theory and practice of item response theory* (2nd eds). The Guilford Press.

Eliaumra, E., Samaela, D. P., & Muhdin, N. K. (2022). Developing diagnostic test assessment to measure creative thinking skills of Biology preservice teacher students. *REID (Research and Evaluation in Education)*, *8*(2), 152–168. https://doi.org/10.21831/reid.v8i2.50885

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum Associates Inc. https://doi.org/10.4324/9781410605269

Finch, H., & French, B. F. (2019). A comparison of estimation techniques for IRT models with small samples. *Applied Measurement in Education*, *32*(2), 77–96. https://doi.org/10.1080/08957347.2019.1577243

FitzPatrick, B., & Schulz, H. (2015). Do curriculum outcomes and assessment activities in science encourage higher order thinking? *Canadian Journal of Science, Mathematics and Technology Education*, *15*(2), 136–154. https://doi.org/10.1080/14926156.2015.1014074

Fox, J., Weisberg, S., & Price, B. (2023). *car: Companion to Applied Regression*. https://cran.r-project.org/web/packages/car/index.html

García-Pérez, M. A. (2014). Multiple-choice tests: Polytomous IRT models misestimate item information. *The Spanish Journal of Psychology*, *17*, 88. https://doi.org/10.1017/sjp.2014.95

Georgiev, N. (2008). Item analysis of C, D and E series from Raven's standard progressive matrices with item response theory two-parameter logistic model. *Europe's Journal of Psychology*, *4*(3), e431. https://doi.org/10.5964/ejop.v4i3.431

Hai, T. N., & Thi, H. N. (2021). Training for critical thinking for high school students to meet the requirements of education innovation. *Journal of Science Educational Science*, *66*(1), 46–56. https://doi.org/10.18173/2354-1075.2021-0005

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item respons* (pp. 1–174). Sage Publication.

Harsana, F. N., Retnawati, H., Dewanti, S. R., & Lumenyela, R. A. (2024). Comparison of item characteristic analysis models of reading literacy test with polytomous item response theory. *REID (Research and Evaluation in Education)*, *10*(2), 214–226. https://doi.org/10.21831/reid.v10i2.77852

Hartono, W., Hadi, S., Rosnawati, R., & Retnawati, H. (2023). Exploration of student' cognitive mathematics ability diagnostic instruments: Validity, reliability, and item characteristics. *Pegem Journal of Education and Instruction*, *13*(3), 386–394. https://doi.org/10.47750/pegegog.13.03.39

Hendrickson, A., & Mislevy, R. J. (2005). Item response theory: Cognitive models. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (pp. 978–982). Wiley. https://doi.org/10.1002/0470013192.bsa108

Hillsley, K. (2025). Question format is the best predictor of item discrimination: A multivariable analysis. *Journal of Microbiology & Biology Education*, *26*(3). https://doi.org/10.1128/jmbe.00205-25

Ismail, S. N., Muhammad, S., Omar, M. N., & Shanmugam, K. S. (2022). The practice of critical thinking skills in teaching mathematics: Teachers' perception and readiness. *Malaysian Journal of Learning and Instruction*, *19*(1), 1–30. https://doi.org/10.32890/mjli2022.19.1.1

Jo, D., Im, S., Suh, D. E., Spencer, S. D., & Masuda, A. (2023). The personalized psychological flexibility index (PPFI): An item response theory analysis with racially diverse college students. *Journal of Psychopathology and Behavioral Assessment*, *45*(3), 829–843. https://doi.org/10.1007/s10862-023-10076-2

Kang, T., Cohen, A. S., & Sung, H. J. (2009). Model selection indices for polytomous items. *Applied Psychological Measurement*, *33*(7), 499–518. https://doi.org/10.1177/0146621608327800

Kassambara, A., & Mundt, F. (2020). *factoextra: Extract and visualize the results of multivariate data analyses* (1.0.7). CRAN. https://doi.org/10.32614/CRAN.package.factoextra

Kim, J., & Wilson, M. (2019). Polytomous item explanatory item response theory models. *Educational and Psychological Measurement*, *80*(4), 726–755. https://doi.org/10.1177/0013164419892667

Lê, S., Josse, J., & Husson, F. (2008). FactoMineR: An R package for multivariate analysis. *Journal of Statistical Software*, *25*(1), 1–18. https://doi.org/10.18637/jss.v025.i01

Lubbe, D. (2023). Advantages of using unweighted approximation error measures for model fit assessment. *Psychometrika*, *88*(2), 413–433. https://doi.org/10.1007/s11336-023-09909-6

Marlina, Hadi, S., & Rahim, A. (2025). The application of item response theory for development tests of high order thinking skills on chemical learning: Partial credit models study. *International Journal of Educational Reform*, *0*(0). https://doi.org/10.1177/10567879251323884

Maryani, I., Prasetyo, Z. K., Wilujeng, I., & Purwanti, S. (2022). Higher-order thinking test of science for college students using multidimensional item response theory analysis. *Pegem Egitim ve Ogretim Dergisi*, *12*(1), 292–300. https://doi.org/10.47750/pegegog.12.01.30

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

Maydeu-Olivares, A., & Joe, H. (2014). Assessing approximate fit in categorical data analysis. *Multivariate Behavioral Research*, *49*(4), 305–328. https://doi.org/10.1080/00273171.2014.911075

Mellenbergh, G. J. (2014). Item bias detection: Modern approaches. In N. Balakrishnan, T. Colton, B. Everitt, W. Piegorsch, F. Ruggeri & J. L. Teugels (Eds.), *Wiley StatsRef: Statistics Reference Online* (pp. 1–6). John Wiley & Sons. https://doi.org/10.1002/9781118445112.stat06390

Muntari, M., Hadisaputra, S., Haris, M., Loka, I. N., Hayati, H., & Yanti, E. (2021). Optimization of achievement in chemistry learning outcomes by implementing discovery learning-based strategy. *Proceeding of the 2nd Annual Conference on Education and Social Science (ACCESS 2020)*, 274-278. https://doi.org/10.2991/assehr.k.210525.089

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

Muraki, E. (1997). A generalized partial credit model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 153–164). Springer. https://doi.org/10.1007/978-1-4757-2691-6

Nunes, S., Oliveira, T., & Oliveira, A. (2017). Item response theory – a first approach. *AIP Conference Proceeding*, *1863*(1). https://doi.org/10.1063/1.4992683

Nye, C. D., Joo, S. H., Zhang, B., & Stark, S. (2020). Advancing and evaluating IRT model data fit indices in organizational research. *Organizational Research Methods*, *23*(3), 457–486. https://doi.org/10.1177/1094428119833158

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, *24*(1), 50–64. https://doi.org/10.1177/01466216000241003

Pada, A. U. T., Mustakim, S. S., & Subali, B. (2018). Construct validity of creative thinking skills instrument for biology student teachers in the subject of human physiology. *Jurnal Penelitian dan Evaluasi Pendidikan*, *22*(2), 119–129. https://doi.org/10.21831/pep.v22i2.22369

Pardede, T., Santoso, A., Diki, D., Retnawati, H., Rafi, I., Apino, E., & Rosyada, M. N. (2023). Gaining a deeper understanding of the meaning of the carelessness parameter in the 4PL IRT model and strategies for estimating it. *REID (Research and Evaluation in Education)*, *9*(1), 86–117. https://doi.org/10.21831/reid.v9i1.63230

Penfield, R. D. (2010). Explaining crossing DIF in polytomous items using differential step functioning effects. *Applied Psychological Measurement*, *34*(8), 563–579. https://doi.org/10.1177/0146621610377083

Penfield, R. D. (2014). An NCME instructional module on polytomous item response theory models. *Educational Measurement Issues and Practice*, *33*(1), 36–48. https://doi.org/10.1111/emip.12023

Phakiti, A. (2018). Assessing higher-order thinking skills in language learning. In J. I. Liontas (Ed.), *The TESOL encyclopedia of English language teaching*. John Wiley & Son. https://doi.org/10.1002/9781118784235.eelt0380

Pratama, G. S., & Retnawati, H. (2018). Urgency of higher order thinking skills (HOTS) content analysis in mathematics textbook. *Journal of Physics: Conference Series*, *1097*(1). https://doi.org/10.1088/1742-6596/1097/1/012147

Pratomo, H., Fitriyana, N., Wiyarsi, A., & Marfuatun. (2025). Mapping chemistry learning difficulties of secondary school students: A cross-grade study. *Journal of Education and Learning*, *19*(2), 909–920. https://doi.org/10.11591/edulearn.v19i2.21826

Preston, K. S. J., & Reise, S. P. (2013). Estimating the nominal response model under nonnormal conditions. *Educational and Psychological Measurement*, *74*(3), 377–399. https://doi.org/10.1177/0013164413507063

Qiu, X.-L., & Torre, J. d. l. (2023). A dual process item response theory model for polytomous multidimensional forced-choice items. *British Journal of Mathematical and Statistical Psychology*, *76*(3), 491–512. https://doi.org/10.1111/bmsp.12303

Rahman, T. (2019). Peningkatan kemampuan berfikir kritis matematis siswa melalui model pembelajaran berbasis masalah. *Symmetry: Pasundan Journal of Research in Mathematics Learning and Education*, *4*(2), 72–80. https://doi.org/10.23969/symmetry.v4i2.2085

Rampean, B. A. O., & Rohaeti, E. (2025). The development of an integrated instrument to measure higher order thinking skills and scientific attitudes. *Journal of Turkish Science Education*, *22*(1), 48–62. https://doi.org/10.36681/tused.2025.004

Reise, S. P. (2014). Item response theory. In R.L. Cautin & S.O. Lilienfeld (Eds.), *The encyclopedia of clinical psychology* (pp.1–10). John Wiley & Son. https://doi.org/10.1002/9781118625392.wbecp357

Reise, S. P. (2016). The emergence of item response theory models and the patient reported outcomes measurement information systems. *Austrian Journal of Statistics*, *38*(4), 211-220. https://doi.org/10.17713/ajs.v38i4.274

Retnawati, H. (2014). *Teori respons butir dan penerapannya: Untuk peneliti, praktisi pengukuran dan pengajian mahasiswa dan pascasarjana.* Nuha Medika.

Revelle, W. (2023). *psych: Procedures for psychological, psychometric, and personality research*. https://cran.r-project.org/package=psych

Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and item response theory analyses. *Journal of Statistical Software*, *17*(5), 1–25. https://doi.org/10.18637/jss.v017.i05

Robitzsch, A., Kiefer, T., & Wu, M. (2023). *TAM: Test analysis modules*. https://cran.r-project.org/package=TAM

Rohim, A. A., Suharti, S., & Herunata. (2025). Enhancing higher order thinking skills (HOTS) through a scientific approach oriented in teaching salt hydrolysis. *Journal of Physics: Conference Series*, *3148*(1). https://doi.org/10.1088/1742-6596/3148/1/012027

Saepuzaman, D., Istiyono, E., & Haryanto. (2022). Characteristics of fundamental physics higher-order thinking skills test using item response theory analysis. *Pegem Egitim ve Ogretim Dergisi*, *12*(4), 269–279. https://doi.org/10.47750/pegegog.12.04.28

Salgado, F. J. A. (2017). The role of context and context familiarity on mathematics problems. *Revista Latinoamericana de Investigacion En Matematica Educativa*, *20*(3), 265–292. https://doi.org/10.12802/relime.17.2031

Samejima, F. (1997). Graded response model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85–100). Springer.

Schauberger, P., & Walker, A. (2025). *openxlsx: Read, write and edit xlsx files* (4.2.8.1). CRAN. https://doi.org/10.32614/CRAN.package.openxlsx

Shanmugam, S. K. S., & Rajoo, M. (2020). Examining the quality of English test items using psychometric and linguistic characteristics among grade six pupils. *Malaysian Journal of Learning and Instruction*, *17*(2), 63–101. https://doi.org/10.32890/mjli2020.17.2.3

Stanke, L., & Bulut, O. (2019). Explanatory item response models for polytomous item responses. *International Journal of Assessment Tools in Education*, *6*(2), 259–278. https://doi.org/10.21449/ijate.515085

Stojanov, A. (2018). Application of item response theory to psychological test construction. *Psychology Science and Practice*, *4*, 67–82. https://doi.org/10.37509/psyscip184067s

Su, K. (2021). A new assessment of HOCS-oriented learning for students' higher order thinking abilities by Marzano's Taxonomy. *Journal of Baltic Science Education*, *20*(2), 305–315. https://doi.org/10.33225/jbse/21.20.305

Sumintono, B. (2015). *Aplikasi pemodelan Rasch: Pada assessment pendidikan.* Penerbit Trim Komunikata.

Sun, J., Xin, T., Zhang, S., & Torre, J. d. l. (2013). A polytomous extension of the generalized distance discriminating method. *Applied Psychological Measurement*, *37*(7), 503–521. https://doi.org/10.1177/0146621613487254

Suswanti, N., Solikhin, F., & Ginting, S. M. (2024). Development of cognitive assessment instruments on buffer solution material based on HOTS (higher order thinking skills). *Journal of Chemistry Education Research*, *8*(2), 97–103. https://doi.org/10.26740/jcer.v8n2.p97-103

Swaminathan, H., Hambleton, R. K., & Rogers, H. J. (2006). Assessing the fit of item response theory models. In C.R. Rao & S. Sinharay (Eds.), *Handbook of statistics* (Vol. 26, pp. 683–718). Elsevier. https://doi.org/10.1016/S0169-7161(06)26021-8

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction*, *4*(4), 295–312. https://doi.org/10.1016/0959-4752(94)90003-5

Tan, T. K. (2024). Evaluating assessment via item response theory utilizing information function with R. *The Quantitative Methods for Psychology*, *20*(1), 33–49. https://doi.org/10.20982/tqmp.20.1.p033

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, *2*, 53–55. https://doi.org/10.5116/ijme.4dfb.8dfd

Triastuti, T., & Laksono, E. W. (2021). Validation of critical thinking test to senior high school students. *Journal of Physics Conference Series*, *1806*(1), 12201. https://doi.org/10.1088/1742-6596/1806/1/012201

Vidergor, H. E. (2018). Effectiveness of the multidimensional curriculum model in developing higher-order thinking skills in elementary and secondary students. *Curriculum Journal*, *29*(1), 95–115. https://doi.org/10.1080/09585176.2017.1318771

Wallmark, J., Ramsay, J. O., Li, J., & Wiberg, M. (2024). Analyzing polytomous test data: A comparison between an information-based IRT model and the generalized partial credit model. *Journal of Educational and Behavioral Statistics*, *49*(5), 753–779. https://doi.org/10.3102/10769986231207879

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org

Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2026). *dplyr: A grammar of data manipulation* (1.2.0). CRAN. https://doi.org/10.32614/CRAN.package.dplyr

Widiantie, R., Handayani, H., Setiawati, I., & Roqiqulqolby, A. (2022). Development of higher order thinking skills assessment based on scientific article reviews. *Proceeding of the 2nd Universitas Kuningan International Conference on System, Engineering, and Technology (UNISET 2021)*. EAI. https://doi.org/10.4108/eai.2-12-2021.2320320

Wildan, W., Supriadi, S., Laksmiwati, D., & Analita, R. N. (2021). Environmental chemistry course assisted problem-based learning in developing students' higher-order thinking skills and characters. *Acta Chimica Asiana*, *4*(2), 141–146. https://doi.org/10.29303/aca.v4i2.54

Willse, J. T. (2026). *CTT: Classical test theory functions* (2.3.4). CRAN. https://cran.r-project.org/package=CTT

Xia, Y., & Yang, Y. (2019). RMSEA, CFI, and TLI in structural equation modeling with ordered categorical data: The story they tell depends on the estimation methods. *Behavior Research Methods*, *51*(1), 409–428. https://doi.org/10.3758/s13428-018-1055-2

Yosepha, A., Ali, M., Wahyudin, D., & Rusman. (2023). The role of multi-dimensional curriculum design in improving higher-order thinking skills. *International Journal of Learning, Teaching and Educational Research*, *22*(7), 219–239. https://doi.org/10.26803/ijlter.22.7.12

APPENDIX

Indicator	Designing the procedure to prepare a buffer solution with a specific target pH by selecting the correct chemical combination and determining the appropriate reactants ratios	
Item Code	Item3	
Cognitive Level	C6 – Creating	
Item	A student wants to make a buffer solution by mixing an acid and a base. a. Below are several acid and base solutions, which combination is the best for making a buffer solution? 1) Hydrocyanic Acid (HCN) 2) Ammonium Hydroxide (NH_4OH) 3) Acetic Acid (CH_3COOH) 4) Sodium Hydroxide (NaOH) 5) Potassium Hydroxide (KOH) 6) Hydrochloric Acid (HCl) 7) Nitrous Acid (HNO_2) 8) Nitric Acid (HNO_3) b. From the previously selected combination, how can the student obtain an acid buffer solution with a pH of 5? Write and explain one the reactions! (known: Ka $CH_3COOH = 2x10^{-5}$; Ka HCN = 6.2×10^{-10}; Kb $NH_4OH = 10^{-5}$; Ka $HNO_2 = 5 \times 10^{-4}$)	
Rubric	0	No answer / wrong answer
	1	Able to choose the right acid-base combinations to make a buffer solution for part (a), but does not write any calculation or explanation for part (b)
	2	Able to choose the right combination and identify the most suitable weak acid based on its ionization constant value, but unable to calculate the ratio/proportion of components to achieve the target pH
	3	Able to choose the right combination and calculate the exact reactant mole ratio to achieve the final pH, but does not make the reaction with 'initial', 'reacted', 'remainder' or give a complete explanation
	4	Able to choose the right combination, calculate the exact reactant mole ratio, make reactions accompanied by the correct amount of 'initial', 'reacted', 'remainder', and can explain the components of buffer solution correctly

Indicator	Evaluating the validity of the statement by constructing accurate stoichiometric calculations	
Item Code	Item8	
Cognitive Level	C5 – Evaluating	
Item	In a beaker, there is a buffer solution made by mixing 50 mL of 0.2 M HCl and 100 mL of 0.4 M NH_3. A total 30 mL of 0.2 M NaOH is then added to buffer solution, changing the pH of mixture to 9.95. Is the statement correct? Prove it with calculations! (Given that Kb of NH_3 = 10-5, log 3 = 0.47, log 6 = 0.78, log 9 = 0.95)	
Rubric	0	No answer / wrong answer
	1	Able to state whether the statement is right or wrong based on a guess, but without any chemical reaction or calculation
	2	Able to make the initial reaction correctly and determine the initial moles, but there are still errors in calculating the 'initial', 'reacted', 'remainder' of the second stage reaction after the addition of the strong base
	3	Able to calculate the amount of 'initial', 'reacted', 'remainder' correctly after the strong base addition and calculate the final pH accurately, but does not state the conclusion whether the statement is right or wrong.
	4	Able to calculate the amount of 'initial', 'reacted', 'remainder' correctly, calculate the final pH accurately, and can conclude the statement in the prompt is right or wrong.