Jurnal Nalar Pendidikan
Volume 13, Number 1, 2025 pp. 10-17
P-ISSN: 2339-0794 E-ISSN : 2477-0515
Open Access: https://dx.doi.org/10.26858/jnp.v13i1.72706

Development of a Two-Tier Multiple-Choice Diagnostic Test to Identify Students' Misconceptions in Chemistry Using the Rasch Model

Rafsanjani Supardi[1*], Muhammad Hasim[2], Wirawan Setialaksana[3], Sriwidayani Syam[4], Andi Ratu Ayuashari Anwar[5]

[1,4,5] Academic Administration, State University of Makassar, Indonesia
[2] Mechanical Engineering Education, State University of Makassar, Indonesia
[3] Informatics and Computer Engineering Education, State University of Makassar, Indonesia

ARTICLE INFO

Article history:
Accepted May 15, 2025
Revised May 31, 2025
Published June 14, 2025
Available online June 15, 2025

Kata Kunci:
tes diagnostik, miskonsepsi kimia, two-tier multiple choice

Keywords:
diagnostic test, chemistry misconceptions, two-tier multiple

E-ISSN: 2477-0515

How to Cite (APA Style):
Supardi, R., Hasim, M., Setialaksana, W., Syam, S., Anwar, A,R,A. (2025). Development of a Two-Tier Multiple-Choice Diagnostic Test to Identify Students' Misconceptions in Chemistry Using the Rasch Model. *Jurnal Nalar Pendidikan. 13 (1). 10-17.*

ABSTRAK

Studi ini dirancang untuk menciptakan alat evaluasi diagnostik berbentuk pilihan ganda dua tingkat guna mengidentifikasi kesalahan pemahaman peserta didik dalam mata pelajaran kimia, terutama pada bahasan larutan penghantar dan non-penghantar listrik serta proses reduksi dan oksidasi. Metode pengembangan alat ini mengacu pada kerangka kerja Wilson, Orindo, dan Antonio yang telah disesuaikan menjadi tiga fase inti: perencanaan, penilaian kelayakan, dan pengujian lapangan. Proses validasi dilakukan melalui penilaian pakar serta penerapan pendekatan statistik Rasch. Alat evaluasi ini terbukti memiliki kualitas isi yang sangat baik dengan skor Aiken mencapai 0,86, menunjukkan kecocokannya dengan materi pembelajaran dan tujuan pengembangan keterampilan analitis siswa. Pengujian di lapangan membuktikan bahwa alat ini memenuhi seluruh kriteria pengukuran Rasch, termasuk kesatuan dimensi, kebebasan jawaban, konsistensi parameter, kecocokan model, dan variasi tingkat kesulitan butir soal. Alat ini juga memiliki konsistensi pengukuran yang baik, dengan performa terbaik dalam menilai peserta yang memiliki kemampuan menengah. Dengan demikian, instrumen ini layak digunakan sebagai alat identifikasi kesalahan konsep dalam pembelajaran kimia sekaligus pedoman dalam menyusun strategi pembelajaran yang tepat.

ABSTRACT

This study is designed to develop a two-tier multiple-choice diagnostic assessment tool to identify students' misconceptions in chemistry, particularly on the topics of conductive and non-conductive solutions, as well as reduction and oxidation processes. The development method follows the adapted framework of Wilson, Orindo, and Antonio, streamlined into three core phases: planning, feasibility assessment, and field testing. Validation was conducted through expert evaluation and the application of Rasch statistical analysis. The assessment tool demonstrated excellent content quality, with an Aiken score of 0.86, confirming its alignment with learning materials and the objective of enhancing students' analytical skills. Field testing confirmed that the tool meets all Rasch measurement criteria, including unidimensionality, answer independence, parameter consistency, model fit, and item difficulty variation. The tool also exhibited strong measurement consistency, performing optimally in assessing students with intermediate ability levels. Thus, this instrument is deemed suitable for identifying conceptual errors in chemistry learning and serves as a guide for formulating effective instructional strategies.

*Corresponding author
E-mail adderesses: rafsanjani.supardi@unm.ac.id

INTRODUCTION

Education plays a strategic role in national development, including in Indonesia. In facing the challenges of globalization and the Fourth Industrial Revolution, the national education system is expected to produce graduates who not only possess technical skills but are also capable of critical thinking and deep understanding of scientific concepts (Fonna, 2019; Waruwu et al., 2022). These competencies are essential for solving complex problems (Astuti, Sugandi, & Pertiwi, 2023). Nevertheless, various challenges remain, such as disparities in educational quality across regions, limited supporting infrastructure, and instructional approaches that are yet to foster conceptual understanding effectively.

Chemistry, as a branch of science, plays a crucial role in developing competent human resources in science and technology. It examines matter and its changes, encompassing scientific products, processes, attitudes, and applications. However, many students in Indonesia face difficulties in learning chemistry due to the abstract nature of its content and the complex interrelation among concepts (Jariyah & Efendi, 2023).

These learning difficulties often lead to misconceptions understandings that deviate from scientifically accepted explanations (Septiyani, 2019). Students frequently struggle with concepts such as atomic structure, chemical bonding, and redox reactions, which involve phenomena that are not directly observable (Izza, Nurhamidah, & Elvinawati, 2021). Misconceptions tend to persist and can continue into higher education levels if not properly identified and corrected (Izza et al., 2021), thereby hindering the acquisition of more advanced concepts and diminishing the quality of chemistry education.

Evaluating misconceptions in chemistry is fundamental for identifying students' cognitive barriers. Such misconceptions often stem from literal interpretations of chemical phenomena, inaccurate symbolic representations, or flawed prior learning experiences (Elvia, Rohiat, & Ginting, 2020). Misconceptions are systematic and form coherent alternative frameworks (Rokhim, Rahayu, & Dasna, 2023), requiring diagnostic instruments that assess not just the accuracy of students' answers but also the reasoning behind them.

Literature indicates that chemistry learning challenges extend beyond symbolic misrepresentations to include the failure to integrate macroscopic, microscopic, and symbolic representations (Lusyana & Aini, 2025). These complexities underscore the significance of diagnosing misconceptions through well-constructed instruments. Over the past five years, researchers have explored diverse formats of diagnostic tests, ranging from two-tier to five-tier multiple-choice instruments, to capture better students' conceptual errors (Astuti, Bhakti, & Prasetya, 2021; Erlangga & Susanti, 2022; Inggit, Liliawati, & Suryana, 2021; Sitorus & Dalimunthe, 2024). However, many of these studies still rely on classical test analysis approaches.

The emergence of Item Response Theory (IRT) provides a more advanced framework for test analysis, overcoming the limitations of Classical Test Theory (CTT). One key advantage of IRT is its ability to produce item parameters that are invariant across different test-taker populations (Mulyani, Efendi, & Ramalis, 2021; Sarea & Ruslan, 2019; Antara, 2020). In contrast to CTT, where item difficulty is sample-dependent, IRT maintains item parameter stability regardless of students' ability levels. This makes IRT a more robust approach for analyzing diagnostic instruments, particularly in identifying student misconceptions (Safitri et al., 2024; Sari et al., 2024).

Accordingly, mapping misconceptions in chemistry through IRT-based diagnostic tools not only strengthens measurement validity but also provides empirical foundations for designing targeted instructional interventions. Developing diagnostic instruments that integrate both conceptual depth and psychometric rigor is essential for advancing chemistry education toward deeper conceptual understanding.

RESEARCH METHOD

Participants

The diagnostic tool was tested on junior and senior high school students in Parepare City. The sample size was determined based on the minimum requirement for item analysis using the Rasch model. Linacre (1994) recommends a sample range between 108 and 243 participants for stable item calibration. Therefore, this study involved 230 students from two educational institutions, meeting the criteria for an ideal sample size. The instrument focused on two main topics, namely electrolyte and non-electrolyte solutions, as well as redox processes. These were developed in accordance with the national curriculum's Standard Competencies, Basic Competencies, and Learning Indicators. The analysis was carried out using the Quest software based on Item Response Theory.

Research Instrument

The study adopted a modified development framework based on the model proposed by Wilson, Orindo, and Antonio. The instrument development process consisted of three primary stages: formulation of the test design, pilot testing, and analysis of measurement results. A matrix format was used to organize test specifications and item indicators. Each item was constructed according to the targeted learning

outcomes, covering concepts related to electrolyte solutions and electron transfer in redox reactions for Grade 10 students in the second semester. Cognitive levels were aligned with the revised Bloom's taxonomy (Azwar, 2019).

Content and Empirical Validation

Content validation focused on scientific accuracy and alignment with educational objectives. The validation team included three university academics and five secondary school chemistry teachers. Their assessments informed a limited field trial. Aiken's V coefficient was used to determine the level of content validity. Items were rated using a five-point Likert scale, where a score of 1 indicated very low relevance, and 5 indicated ideal alignment between the test item and the measured indicator (Azwar, 2019). The mathematical formula for Aiken's V is expressed as:

$$V = \sum \frac{s}{n(c-1)} \quad \text{.....(2)}$$

Description :

s : r-lo

lo : lowest possible rating

c : highest possible rating

r : the rating given by a rater

n : number of raters

Field testing was then conducted to evaluate the empirical properties of the instrument. Item difficulty parameters were analyzed using the Rasch model. According to Hambleton, Swaminathan, and Rogers (1991), acceptable difficulty values range from -2.0 to +2.0. Items outside this range are considered either too easy or too difficult.

Item Characteristics

To address the limitations of classical test theory, the study applied Item Response Theory. This approach allows for a more sensitive analysis of item difficulty, sample independence, item invariance, unidimensionality, and examinee ability (Disnawati et al., 2024). Accordingly, several psychometric assumptions were evaluated, including unidimensionality, local independence, and parameter invariance, which are required for valid IRT analysis (DeMars, 2010).

Unidimensionality was assessed through factor analysis to determine whether the items measured a single underlying construct (Astuti et al., 2024). Local independence refers to the condition where a student's response to one item is not influenced by their response to another item after controlling for ability (Wulansari & Kirana, 2023).

Parameter invariance implies that item difficulty remains stable across groups with varying abilities, and test-taker ability is consistent across items with differing levels of difficulty (Disnawati et al., 2024). An item is considered to fit the Rasch model if it satisfies the criteria for Outfit MNSQ (0.5 to 1.5), ZSTD (-2 to +2), and Infit MNSQ (0.77 to 1.33), as proposed by Muntazhimah (2023). Meeting these conditions ensures that the instrument aligns with the requirements of the Rasch measurement model.

RESULTS AND DISCUSSION

The final product of this research is a two-tier multiple-choice diagnostic evaluation tool designed to identify conceptual errors in chemistry learning. This evaluation tool was constructed through the adaptation of the development framework proposed by Wilson, Oriondo, and Antonio, encompassing three main phases: (1) evaluation design planning, (2) product trial implementation, and (3) measurement result analysis. The final design of the instrument comprises 35 items distributed proportionally across various subject topics and higher-order thinking skill levels.

Content validation by experts indicated that all components of the instrument, including learning materials, evaluation framework, achievement indicators, and test items, met essential validity criteria. This finding is supported by Aiken's V validity coefficient score of 0.86, indicating that 86% of the items accurately represent the subject content and higher-order thinking dimensions being measured, as well as aligning with the applicable curriculum. In the empirical validation stage, the instrument was tested against five main psychometric parameters: (1) unidimensionality, (2) local independence, (3) parameter invariance, (4) model fit, and (5) item difficulty index. The unidimensionality test using confirmatory factor analysis revealed the presence of ten principal components dominating the instrument structure, with the test results visualized in Figure 1 in the form of a dimensional scree plot.

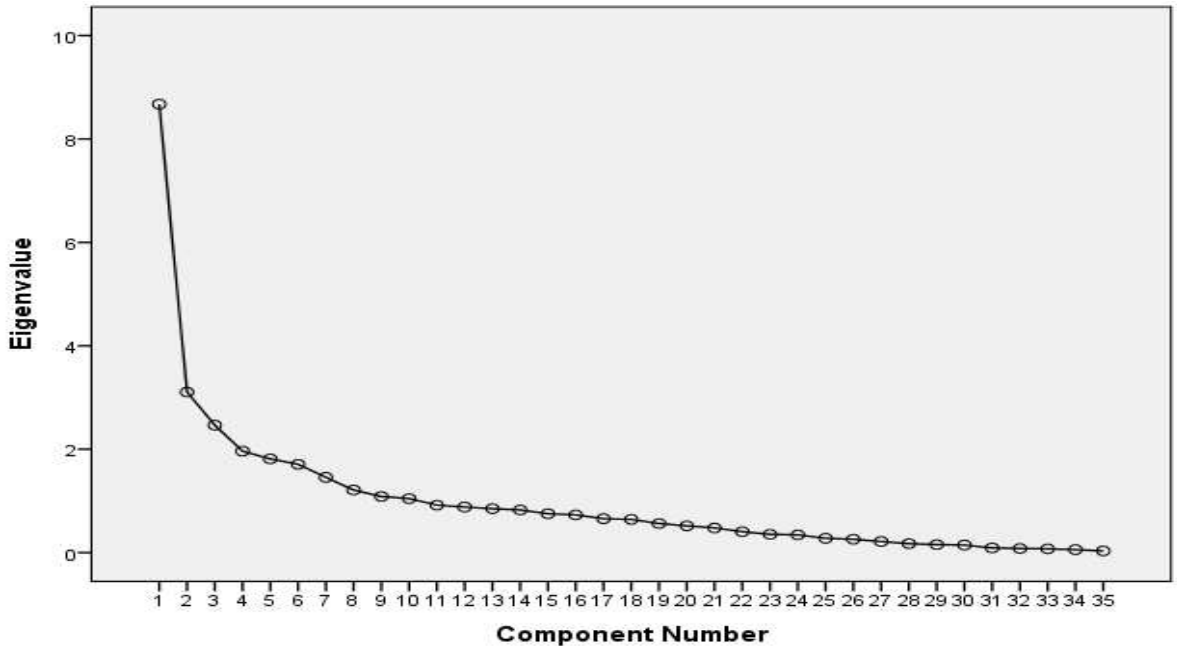

Figure 1. Scree Plot of Instrument Eigenvalues

The interpretation of the unidimensionality test results revealed one principal component that stood out significantly from the others. This main component had an eigenvalue of 8.672 with a cumulative variance contribution of 24.778%. In the context of the local independence test, the condition is fulfilled when the variables influencing participants' performance remain fixed so that responses to one item are not statistically related to responses to other items (Saepuzaman et al., 2021). This concept implies that there should be no significant correlation between items, or inter-item correlation values should approach zero. The data from the local independence assumption test are presented in the following analysis.

Table 1. Local Independence of the Instrument

	K1	K2	K3	K4	K5	K6	K7	K8	K9	K10
K1	0,002	0,001	0,001	0,001	0,001	0,001	0,001	0,002	0,003	0,004
K2		0,001	0,001	0,000	0,000	0,001	0,001	0,001	0,001	0,002
K3			0,001	0,000	0,000	0,001	0,001	0,001	0,002	0,003
K4				0,000	0,000	0,000	0,000	0,001	0,001	0,001
K5					0,000	0,000	0,000	0,001	0,001	0,001
K6						0,001	0,001	0,001	0,001	0,002
K7							0,001	0,001	0,002	0,002
K8								0,002	0,002	0,004
K9									0,004	0,005
K10										0,013

The test results in Table 1 indicate that the correlation values of the participants' theta groups (matrix diagonals) tend to be low or near zero. This proves the absence of significant item interdependence, thus satisfying the requirement for local independence. The minimal residual correlation level (< 0.01) indicates that each item independently measures the intended construct without systematic influence from responses to other items. Furthermore, the consistent stability of the diagonal values supports the validity of the measurement model in Item Response Theory (IRT). To test parameter invariance, the respondents or test items were divided into two groups to determine the correlation magnitude between the two (Antara, 2020; Kurniawan & Munadi, 2019). The visualization results in Figure 2 show that the distribution of the estimation points is very close to the reference line (y = x), with a correlation coefficient of 0.938. The closeness of the points to the straight line (y = x) and the high correlation value (>0.90) serve as strong evidence that the item parameter invariance assumption is met. Thus, the developed instrument demonstrates good invariance properties.

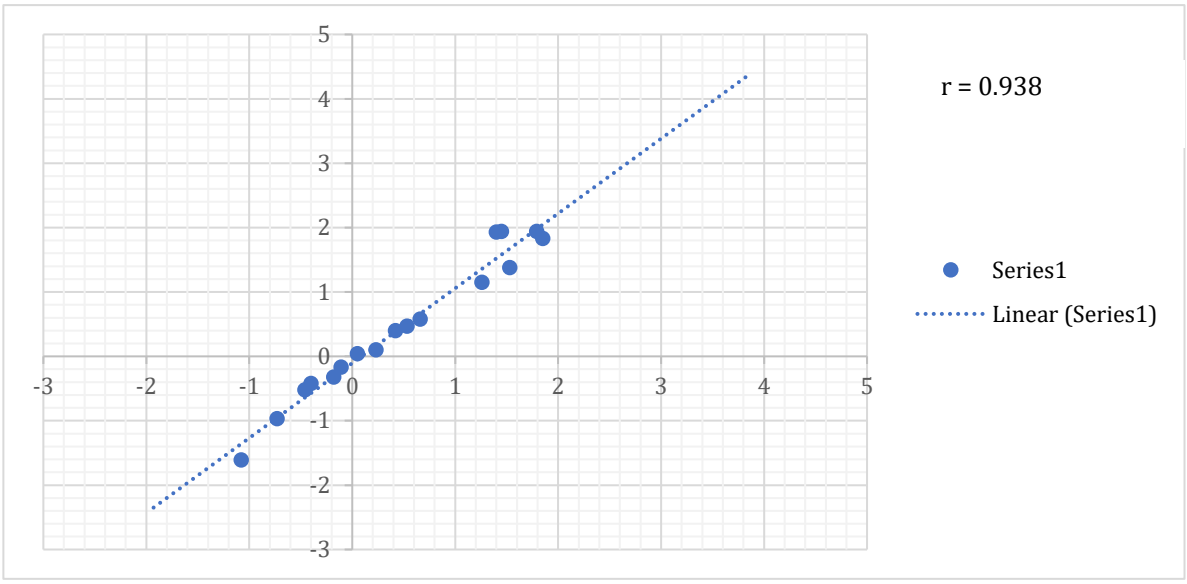

Figure 2. Invariance of Item Parameters

Further testing results confirmed the invariance of participants' ability parameters relative to item characteristics. The correlation value of 0.997 obtained from grouping participant abilities shows a very high level of stability. This finding reinforces the evidence that participants' ability scores remain consistent regardless of the varying difficulty levels of the items they face. The analysis of item and ability parameter invariance shows that the plots in both graphs are relatively close together. This indicates that the assumption of invariance for both item and ability parameters has been fulfilled.

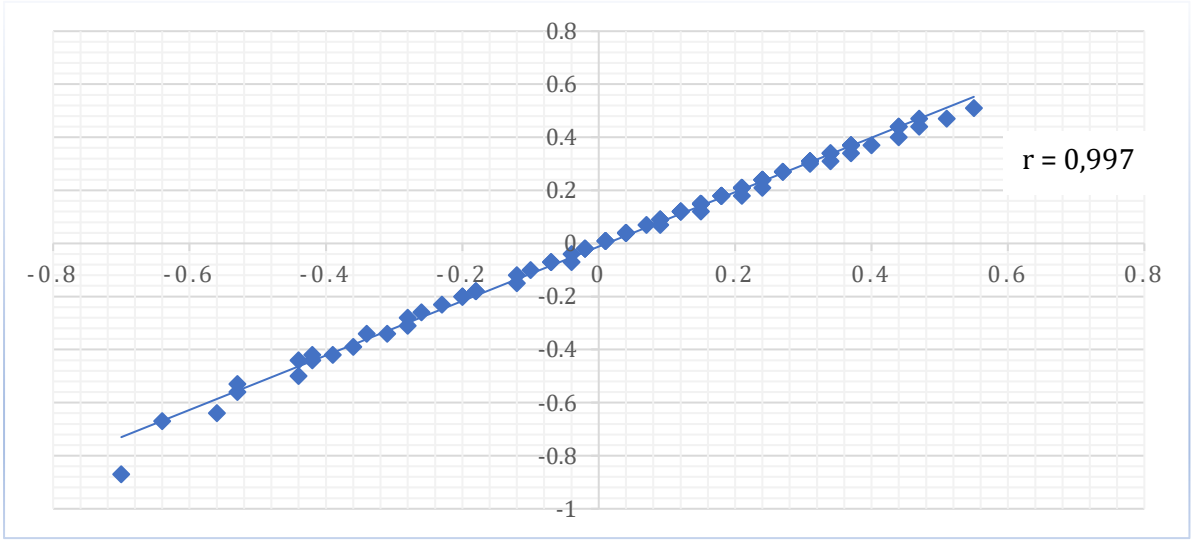

Figure 3. Invariance of Ability Parameters

An item is considered to fit the model if it meets the criteria of outfit MNSQ within the range of 0.5 to 1.5 and outfit ZSTD values between -2 and +2. In addition to the outfit indicators, item fit can also be analyzed through the infit MNSQ, ideally ranging from 0.77 to 1.33. The distribution of items based on infit MNSQ is visualized in the item fit map for the model used.

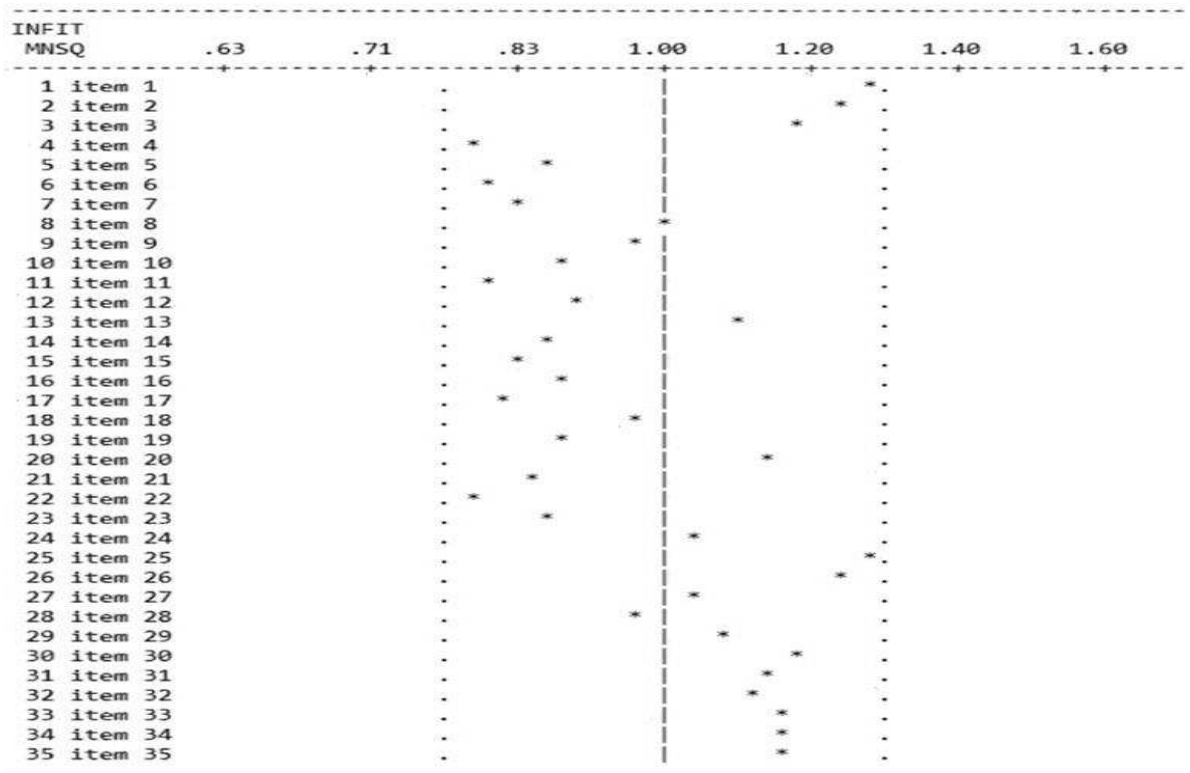

Figure 4. Item Fit Map to the Model

The development of the instrument in this study used the Rasch model of Item Response Theory (IRT), which relies on a single parameter, namely item difficulty level. Negative difficulty values indicate easy items, while positive values indicate difficult items. An item is considered high quality if its difficulty value is within the range of -2 to +2 (Hambleton et al., 1991). The analysis results showed difficulty levels ranging from -1.93 to 1.94, indicating that all items met the criteria. The Item Characteristic Curve (ICC) illustrates the relationship between the probability of answering correctly and participants' ability levels. Curves that shift to the right indicate more difficult items, while curves shifting to the left indicate easier items. The overall ICC visualization is presented in Figure 5.

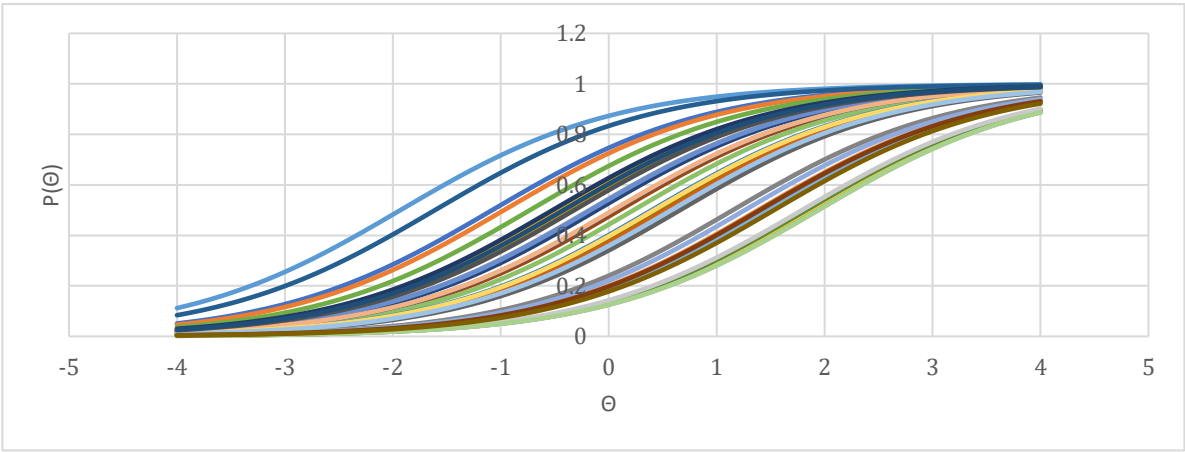

Figure 5. Test Item ICC

The Item Characteristic Curve (ICC) indicates the difficulty characteristics of test items in the form of a curve representing the relationship between the probability of a 50% correct response and the participant's ability level. Items with curves that shift further to the right have higher difficulty levels. Items with curves shifting to the left indicate that the items are easier. The overall item characteristic curves are presented in Figure 5.

The item information function measures an item's contribution in revealing the test taker's ability and is used to evaluate item quality, select items, and compare test sets (Septiyani, 2019). The plot output

shows that the test provides a maximum information value of 15.57 with a standard error of measurement of 0.253 at a participant ability level of approximately 0.25, which represents a moderate ability level. The lower and upper bounds of the interval are determined at the intersection point between the information function curve and the measurement error curve. Beyond this interval, the information provided decreases, and measurement error increases. Thus, this test instrument is best suited for participants with moderate ability levels.

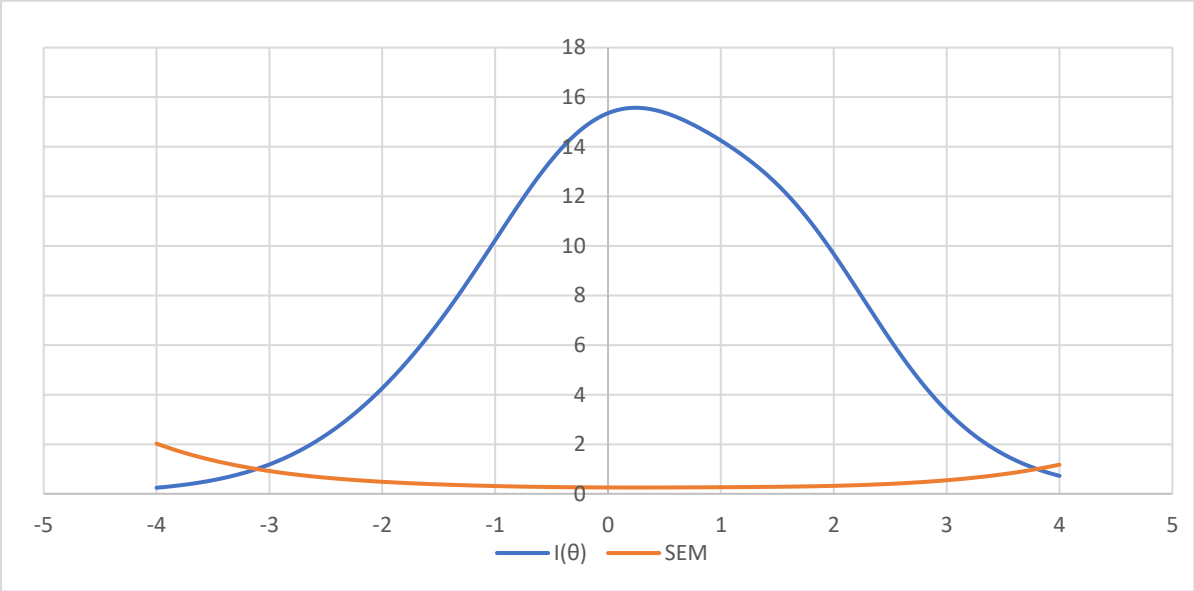

Figure 6. Information Function and Standard Error of Measurement

Figure 6 shows that the effective interval of the information function lies within the ability (theta) range of -3.1 to 3.8. Within this range, measurement error remains controlled, making the test suitable for participants with low to high ability. The information function and standard error reflect the reliability of the instrument, which, in this case, demonstrates that the test is reliable for measuring ability within the specified theta range (Wulansari & Kirana, 2023).

CONCLUSION

This study successfully developed a two-tier multiple-choice diagnostic test instrument to identify students' misconceptions about the subject matter of electrolyte and non-electrolyte solutions and redox reactions through a modified framework by Wilson, Oriondo, and Antonio in three stages: design, trial, and measurement. Expert content validation demonstrated a high level of validity (Aiken's V index), with a coefficient confirming the alignment of material and higher-order thinking skills coverage with the curriculum.

Item analysis using the Rasch model confirmed the fulfillment of five key psychometric assumptions, namely unidimensionality (with the highest eigenvalue dominant factor), local independence, parameter invariance (strong correlation between groups), model fit, and item difficulty variation that is ideal for obtaining maximum information. The test information function indicated that the instrument is most effective for students with moderate abilities, offering minimal measurement error and high reliability. Therefore, the instrument is suitable for comprehensively detecting misconceptions across varying student ability levels.

Overall, this research provides a significant contribution to the field of chemistry education by offering a valid and reliable instrument for identifying misconceptions. These findings serve as a foundation for educators to design more effective instructional strategies for improving students' conceptual understanding. The instrument also supports the transformation of chemistry learning toward deeper conceptual comprehension and adds value to the development of future valid and reliable diagnostic tools for teachers.

REFERENCES

Antara, A. A. P. (2020). *Penyetaraan vertikal dengan pendekatan klasik dan item response theory (teori dan aplikasi)*. Deepublish.

Astuti, I. A. D., Bhakti, Y. B., & Prasetya, R. (2021). Four Tier-Magnetic Diagnostic Test (4T-MDT): Instrumen

Evaluasi Medan Magnet Untuk Mengidentifikasi Miskonsepsi Siswa. *JIPFRI (Jurnal Inovasi Pendidikan Fisika Dan Riset Ilmiah)*, *5*(2), 110–115.

Astuti, N. D., Hapsan, A., Herianto, M., Warsyidah, A. A., Riskawati, N. M., Febriana, B. W., & Toron, V. B. (2024). *Prinsip-prinsip Pengukuran dan Evaluasi Pendidikan: Disertai dengan Contoh Kasus*. CV. Ruang Tentor.

Astuti, N. L. P., Sugandi, A. I., & Pertiwi, C. M. (2023). Eksplorasi kesulitan siswa SMP dalam menjawab soal kemampuan pemecahan masalah matematis pada materi fungsi kuadrat. *JPMI (Jurnal Pembelajaran Matematika Inovatif)*, *6*(4), 1441–1448.

Azwar, S. (2019). *Reliabilitas dan validitas*.

DeMars, C. (2010). *Item response theory*. Oxford University Press.

Disnawati, H., Wahyudi, E., Ismail, I. H., Dos Santos, M., Jaya, P. R. P., Jusmiana, A., … Susilowati, Y. (2024). *Esensi Pengukuran Dan Evaluasi Pendidikan: Teori dan Praktik*. CV. Ruang Tentor.

Elvia, R., Rohiat, S., & Ginting, S. M. (2020). Identifikasi miskonsepsi mahasiswa pada pembelajaran daring matematika kimia melalui tes diagnostik three tier multiple choice. *Hydrogen: Jurnal Kependidikan Kimia*, *9*(2), 84–96.

Erlangga, S. Y., & Susanti, S. (2022). Identifikasi Miskonsepsi Peserta Didik Menggunakan Instrumen Diagnostik Three Tier Pada Materi Gerak Lurus. *Jurnal Sosial Humaniora Sigli*, *5*(2), 312–316.

Fonna, N. (2019). *Pengembangan revolusi industri 4.0 dalam berbagai bidang*. Guepedia.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory* (Vol. 2). Sage.

Inggit, S. M., Liliawati, W., & Suryana, I. (2021). Identifikasi miskonsepsi dan penyebabnya menggunakan instrumen five-tier fluid static test (5tfst) pada peserta didik kelas xi sekolah menengah atas. *Journal of Teaching and Learning Physics*, *6*(1), 49–68.

Izza, R. I., Nurhamidah, N., & Elvinawati, E. (2021). Analisis miskonsepsi siswa menggunakan tes diagnostik esai berbantuan cri (certainty of response index) pada pokok bahasan asam basa. *Alotrop*, *5*(1), 55–63.

Jariyah, A., & Efendi, N. (2023). Pengaruh Penerapan Model Pembelajaran Kooperatif Tipe Jigsaw Terhadap Hasil Belajar Siswa Kelas V SDN Candipari I. *Pendas: Jurnal Ilmiah Pendidikan Dasar*, *8*(2), 3878–3896.

Kurniawan, N. I. A., & Munadi, S. (2019). Analysis of the quality of test instrument and students' accounting learning competencies at vocational school. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *23*(1), 68–75.

Linacre, J. M. (1994). Sample size and item calibration stability. *Rasch Measurement Transactions*, *7*, 328.

Lusyana, L., & Aini, F. Q. (2025). Evaluasi Penyajian Multirepresentasi pada Buku Teks Kimia Kurikulum Merdeka pada Materi Kesetimbangan Kimia. *JagoMIPA: Jurnal Pendidikan Matematika Dan IPA*, *5*(1), 223–235.

Mulyani, S., Efendi, R., & Ramalis, T. R. (2021). Karakterisasi tes keterampilan pemecahan masalah fisika berdasarkan teori respon butir. *JURNAL Pendidikan Dan*, *1*(01).

Muntazhimah, M. P. (2023). *Model Rasch: Pengembangan Instrumen Penelitian Pendidikan*. Deepublish.

Rokhim, D. A., Rahayu, S., & Dasna, I. W. (2023). Analisis miskonsepsi kimia dan instrumen diagnosisnya: literatur review. *Jurnal Inovasi Pendidikan Kimia*, *17*(1), 17–28.

Safitri, I., Lestarani, D., Imtikhanah, R. D. N. W., Akbarini, N. R., Sari, M. W., Fitrah, M., & Hapsan, A. (2024). *Teori Pengukuran dan Evaluasi*. CV. Ruang Tentor.

Sarea, M. S., & Ruslan, R. (2019). Karakteristik Butir Soal: Classical Test Theory vs Item Response Theory? *Didaktika: Jurnal Kependidikan*, *13*(1), 1–16.

Sari, N. A. A., Antia, V., Daimah, U. S., Muhakimah, I., & Dewanti, S. S. (2024). Konstruksi Instrumen Tes Kemampuan Pemecahan Masalah Menggunakan Teori Respon Butir. *Teorema: Teori Dan Riset Matematika*, *9*(2), 193–206.

Septiyani, E. (2019). *Identifikasi miskonsepsi siswa menggunakan Tes Diagnostik Four-Tier Digital Test (4TDT) berbasis website pada konsep suhu dan kalor*. Fakultas Ilmu Tarbiyah dan Keguruan UIN Syarif Hidayatullah Jakarta.

Sitorus, D. M., & Dalimunthe, M. (2024). Pengembangan Instrumen Tes Diagnostik Five-Tier Multiple Choice untuk Mengidentifikasi Miskonsepsi Siswa pada Materi Kesetimbangan Kimia. *Jurnal Pendidikan Kimia FKIP Universitas Halu Oleo*, *9*(1), 55–72.

Waruwu, E., Ndraha, A. B., & Lase, D. (2022). Peluang dan tantangan G20 dalam transformasi manajemen pendidikan di era revolusi industri 4.0 dan civil society 5.0 pasca pandemi COVID-19. *Jurnal Ilmiah Maksitek*, *7*(3), 26–32.

Wulansari, A. D., & Kirana, D. P. (2023). *Pengukuran English Vocabulary Size dengan Computerized Adaptive Testing*. Thalibul Ilmi Publishing & Education.