Journal of Education and Teaching (JET)
Vol. 7 No. 1 2026

Innovative Online Diagnostic Assessment for Evaluating Science Learning Readiness in the Merdeka Curriculum Framework

Widowati Pusporini[1], Dadan Rosana[2], Didik Setyawarno[3], Eko Widodo[4], Maryati[5]
Science Education, Universitas Negeri Yogyakarta, Indonesia[1,2,3,4,5]
Corresponding email: widowatipusporini@uny.ac.id

Draft article history
Submitted: 02-04-2025;
Revised: 09-13-2025;
Accepted: 10-01-2025;

ABSTRACT: The main objective of this study was to develop an Online Diagnostic Assessment (ODA) that can effectively measure junior high school students' readiness to learn science within the framework of the Merdeka Curriculum. Unlike previous diagnostic tools that are mostly offline or limited in scope, this ODA provides an innovative, technology-based solution that identifies student readiness and misconceptions while offering timely feedback to support differentiated learning. Using the 4-D development model, the research includes the Define, Design, Develop, and Disseminate stages. At the Define stage, the instrument grids were identified according to the curriculum. In the Design stage, the assessment prototype was developed using Google Form. The Develop stage involved validation by experts and field tests in two junior high schools, resulting in a valid and reliable instrument, with an Aiken's V value above 0.80, Cronbach's Alpha 0.736, and McDonald's Omega 0.768. The results of the EFA analysis showed all items had Measures of Sampling Adequacy (MSA) of more than 0.5. The distribution of respondents' scores shows that the instrument is effective in differentiating students' abilities. Thus, the developed ODA is valid, reliable, and ready to be used to assess students' readiness in science learning.

Keywords: 4D, merdeka curriculum, online diagnostic assessment.

ABSTRAK: Tujuan utama dari penelitian ini adalah untuk mengembangkan *Online Diagnostic Assessment* (ODA) yang dapat secara efektif mengukur kesiapan belajar sains siswa sekolah menengah pertama dalam kerangka Kurikulum Merdeka. Berbeda dengan alat diagnostik sebelumnya yang sebagian besar bersifat luring atau terbatas cakupannya, ODA ini menawarkan solusi inovatif berbasis teknologi yang mampu mengidentifikasi kesiapan dan miskonsepsi siswa sekaligus memberikan umpan balik tepat waktu untuk mendukung pembelajaran berdiferensiasi. Penelitian ini menggunakan model pengembangan 4-D yang meliputi tahap *Define*, *Design*, *Develop*, dan *Disseminate*. Pada tahap *Define*, kisi-kisi instrumen diidentifikasi berdasarkan kurikulum. Pada tahap *Design*, prototipe penilaian dikembangkan menggunakan Google Form. Tahap *Develop* melibatkan validasi oleh para ahli dan uji lapangan di dua sekolah menengah pertama, yang menghasilkan instrumen yang valid dan reliabel, dengan nilai Aiken's V di atas 0,80, Cronbach's Alpha sebesar 0,736, dan McDonald's Omega sebesar 0,768. Hasil analisis EFA menunjukkan bahwa semua butir memiliki *Measures of Sampling Adequacy* (MSA) lebih dari 0,5. Distribusi skor responden menunjukkan bahwa instrumen ini efektif dalam membedakan kemampuan siswa. Dengan demikian, ODA yang dikembangkan dinyatakan valid, reliabel, dan siap digunakan untuk menilai kesiapan siswa dalam pembelajaran sains.

Kata kunci: 4D, Kurikulum Merdeka, *online diagnostic assessment*.

INTRODUCTION

The integration of technology in education has had a significant impact on the teaching and learning process (Rapaka et al., 2025; Zhang & Xu, 2025). Technology has increased students' access to resources, made learning more engaging, and improved information transfer between teachers and students (Ghory & Ghafory, 2021). Technology has also facilitated the development of new metrics to evaluate student understanding, going beyond traditional assessment methods to provide a more comprehensive assessment (Leitão et al., 2020; Salinas-Navarro et al., 2024). Research has shown that the use of instructional technology positively impacts student learning, increases interest and satisfaction, and is now an integral part of the learning environment (Draude & Brace, 1999). Educational technology serves as a medium to solve learning problems, improve performance, and increase student engagement (Benjamin, 2024; Kalyani, 2024). Technology allows teachers to create diverse learning materials, incorporating multimedia elements and interactive components, which can increase students' desire to learn and promote active critical thinking (Sudarsana et al., 2019).

One of the key developments is the introduction and use of Online Diagnostic Assessments (ODA), which provide a dynamic platform to identify students' level of understanding in greater detail. ODA allows teachers to know students' strengths and weaknesses early on. Then, learning can be tailored to students' individual needs. In the context of a modern curriculum such as Merdeka Curriculum, ODA becomes a very important tool to improve teaching efficiency and effectiveness.

In implementing the Merdeka Curriculum, ODA plays a significant role in supporting a more personalized and adaptive approach to learning. As an initial assessment, ODA provides a comprehensive picture of students' readiness to face learning materials. By knowing the extent of students' understanding, teachers can design more relevant and targeted teaching strategies, making the learning process more responsive to individual needs. ODA also provides quick feedback, which is not only beneficial for teachers but also for students, as they can immediately recognize areas that need improvement and take corrective measures.

The advantage of ODA lies in its ability to be integrated into the learning process seamlessly, creating a more adaptive learning environment. ODA-enabled technology allows teachers to conduct continuous progress monitoring, facilitating the recognition of students' specific needs over time. For example, if a student is having difficulty with a particular concept, the teacher can immediately provide the necessary intervention before the difficulty impacts the understanding of other material. ODA also enables differentiation in learning by providing additional materials or further challenges for students who need them.

In addition, the implementation can also play a role in reducing inequalities in education. With more individualized monitoring, students who need additional assistance can be recognized immediately and receive more intensive support. This is especially important in the context of the Merdeka Curriculum, which emphasizes the development of students' full potential, both academically and

Corresponding author: Widowati Pusporini (widowatipusporini@uny.ac.id)

non-academically. ODA helps create an inclusive learning environment where every student has an equal opportunity to succeed, regardless of their background. However, while ODA offers many benefits, some challenges need to be addressed. One of the main challenges is the validity of the instruments used in ODA. Further research is needed to ensure that the diagnostic tools used in ODA actually measure student understanding accurately, especially in the context of the Merdeka Curriculum, which emphasizes competency-based learning. In addition, the adaptation of ODA to student diversity is also an important issue. Each student has a different learning style, so it is important to ensure that the learning environment can accommodate these differences.

Another possible obstacle is related to the technology skills of teachers and students. Although technology has become an integral part of education, many teachers may not feel confident in using technology-based tools, including ODA. Therefore, adequate training is needed to enable teachers to use ODA effectively as part of their teaching strategies. Support from the school and government is also crucial to ensure that the necessary infrastructure to support the use of ODA is in place.

In the context of the Merdeka Curriculum, there are also questions about how accurate ODA is in measuring students' mastery of competencies. Further research needs to find out whether ODA can really provide an accurate picture of the extent to which students master certain competencies. This is important because Merdeka Curriculum emphasizes results-oriented learning, where students are expected to master certain competencies before moving on to the next stage.

In addition, research needs to examine how ODA affects students' learning motivation. In some cases, students may feel pressured by the constant assessment, especially if the results show that they have not achieved the expected understanding. Therefore, it is important to find ways to use ODA as a tool that motivates students to learn better, rather than as a tool that creates anxiety.

In the implementation of the Merdeka Curriculum for Class VIII materials, several research gaps related to ODA still need to be bridged. One of them is the lack of in-depth research on the effectiveness of ODA in measuring students' readiness to master the competencies taught in this curriculum. ODA provide initial information about student readiness. But further research is needed to ensure how accurate ODA is in predicting student success in achieving these competencies.

Research should include an analysis of how ODA is used by teachers as part of their teaching strategies. Many questions remain unanswered, such as how teachers use ODA results to design more effective lessons or how ODA affects the way teachers provide feedback to students. Other factors that need to be considered are school support, availability of resources, and teachers' understanding of how to use technology in the teaching process. In addition, constraints in the implementation of ODA in schools need a further investigation. For example, infrastructure limitations, such as uneven internet access or a lack of

Corresponding author: Widowati Pusporini (widowatipusporini@uny.ac.id)

adequate devices, can be barriers to have an effective ODA implementation. Moreover, teacher training is an important factor to consider. Teachers need an adequate and proper skills to use ODA properly and utilize it as an effective teaching aid.

Overall, ODA has great potential to improve the quality of learning, especially in the context of the Merdeka Curriculum, which emphasizes more personalized and adaptive learning. However, to maximize this potential, further research is needed on various aspects of ODA implementation, including the validity of the instrument, adaptation to student diversity, and challenges faced in its implementation. With more in-depth research, we can optimize the role of ODA in achieving more effective, inclusive, and sustainable learning goals.

RESEARCH METHOD

The type of the research was a research and development (R & D) using the 4-D (Four D) model. The development stages included define, design, develop, and disseminate (Paidi, 2011; Subali, 2019; Thiagarajan et al., 1974). Initial research was carried out by analyzing the learning outcomes of class VIII science subjects in depth of the Merdeka curriculum, designing online diagnostic assessments for student learning readiness in implementing the Merdeka curriculum, including indicators, validation, and initial revision, implementation, or empirical testing in the field, and dissemination through workshops attended by teachers, lecturers, and observers of science education.

The research consisted of four stages: needs analysis, which emphasizes the importance of developing online diagnostic assessment instruments for learning readiness in the Merdeka Curriculum, planning on campus, developing instruments through expert revision and empirical testing, and disseminating results through science workshops. Then, they can be widely used by educational institutions.

Corresponding author: Widowati Pusporini (widowatipusporini@uny.ac.id)

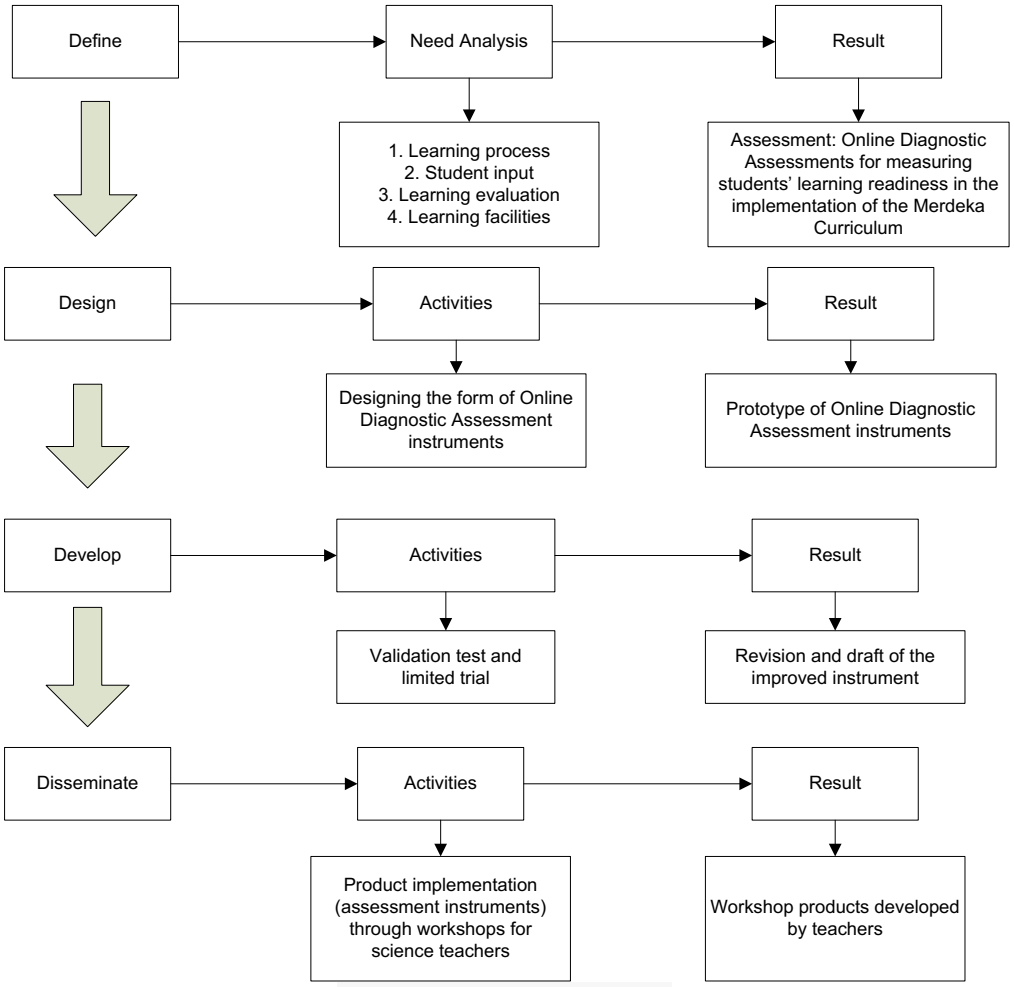

Figure 1. Research Design

The research consisted of four important stages. The first stage is identification (define), where an analysis of the curriculum and science learning outcomes aim to determine students' initial competencies and formulate online diagnostic assessment indicators. Furthermore, at the planning stage (design), the prototype of the online diagnostic assessment is prepared based on the learning outcomes set out in the curriculum. Then, at the development stage, the revised assessment instrument is produced through expert validation and limited testing with students. Finally, at the dissemination stage, the assessment instrument was widely introduced through workshops and seminars for science teachers and lecturers. This research was conducted in several junior high schools in Sleman, Yogyakarta, with the subjects being science education experts (lecturers) and students from four classes. It collected qualitative data from expert input as the basis for instrument revision and quantitative data from expert assessment sheets to assess the quality of content, construct, and language of the instrument. Field tests were conducted to empirically evaluate the quality of the instrument items.

Data were analysed quantitatively and qualitatively. Content, construct, and language analysis were conducted using Aiken's formula, with a minimum validity score of 0.80 for five validators (Retnawati, 2014; Setyawarno, 2020). Construct

Corresponding author: Widowati Pusporini (widowatipusporini@uny.ac.id)

quality was analysed using Exploratory Factor Analysis (EFA) to test the reliability and consistency of the instrument, measured by Cronbach's Alpha and McDonald's Omega (Retnawati, 2014). Instrument retested on junior high school students with EFA and Item Response Theory (IRT) (Retnawati, 2014) using Jamovi software. The requirement for the size of the KMO is > 0,5 (Sutopo, Y dan Slamet, A, (2017). Quantitative to qualitative scale conversion analysis was conducted to assess the feasibility of the instrument from the aspects of content, construct, and language, based on the criteria set (Suparwoto, 2003). The fit of the Online Diagnostic Assessment instrument items was analysed using the Rasch model for dichotomous data with the Jamovi application, based on INFIT and OUTFIT values to determine whether the item fits the model (Adams & Kho, 1996).

RESULT AND DISCUSSION

The research was a research and development (R&D) project that aims to produce an online diagnostic assessment (ODA) product to measure student readiness in science learning in junior high schools in implementing the Merdeka Curriculum. The development used the 4-D model (Four D), which consisted of four stages: define, design, develop, and disseminate (Paidi, 2011; Thiagarajan et al., 1974). In the first stage, the instrument grids were identified in accordance with the junior high school science curriculum. The second stage involved online assessment planning. Next, the third stage was the preparation and field testing with experts. And, the fourth stage was the dissemination of results through science teacher training in Yogyakarta.

The Define Stage

This stage aims to determine and define an online diagnostic assessment (ODA) instrument following the junior high school science curriculum through literature studies and previous research. The analysis includes learning outcomes and science materials of the Merdeka curriculum. The result is a diagnostic assessment grid that measures students' prerequisite abilities before entering learning materials. The ODA includes three assessment components aligned with science literacy, namely: content, cognitive process, and context, which are relevant as the basis for developing this instrument (OECD, 2018; Suprayitno, 2019).

The Design Stage

This stage involved developing a prototype of an online diagnostic assessment tool using Google Forms to measure junior high school students' readiness for the Merdeka Curriculum. The format chosen was multiple-choice and true-false questions, designed to provide automatic feedback on student readiness or suggestions for additional material. The assessment instrument is oriented towards science literacy, covering aspects of content, context, knowledge, and cognitive processes. The products of this stage include assessment indicators and draft questions to measure students' readiness for AKM and PISA.

Corresponding author: Widowati Pusporini (widowatipusporini@uny.ac.id)

Journal of Education and Teaching (JET) Volume 7 No. 1, 2026
DOI: 10.51454/jet.v7i1.546

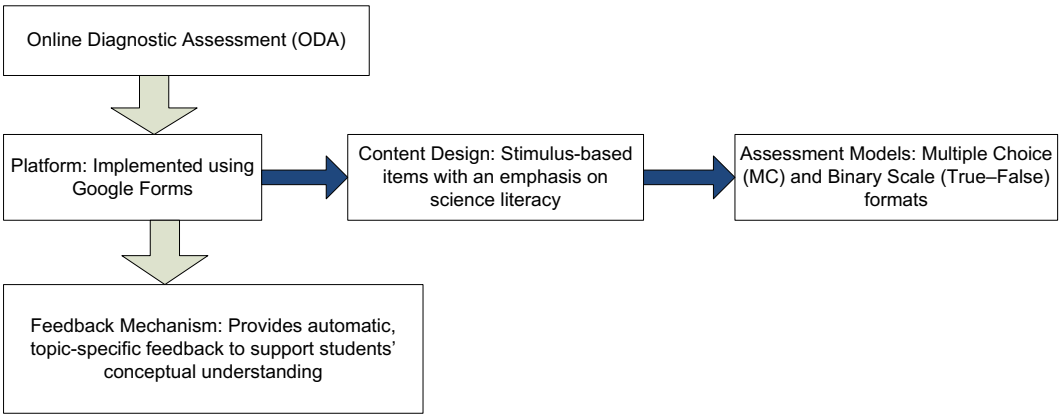

Figure 2. Online Diagnostic Assessment Application Design

The Develop Stage

This stage aims to produce an online diagnostic assessment (ODA) product that can measure students' readiness in learning science in junior high school following the implementation of the Merdeka Curriculum. This stage of development includes several steps, namely Google Form-based questions equipped with automatic feedback, product validation by experts related to construct, content, and language, and field tests conducted at SMP N 2 Mlati and SMP N 7 Muhammadiyah Yogyakarta. The results of the validation and field test were used to revise the product. Data were analyzed quantitatively and qualitatively, including validity analysis using Aiken's formula (Aiken, 1985; Setyawarno, 2020). Validity is determined through expert judgment, with valid results if Aiken ≥ 0,80.

Corresponding author: Widowati Pusporini (widowatipusporini@uny.ac.id)

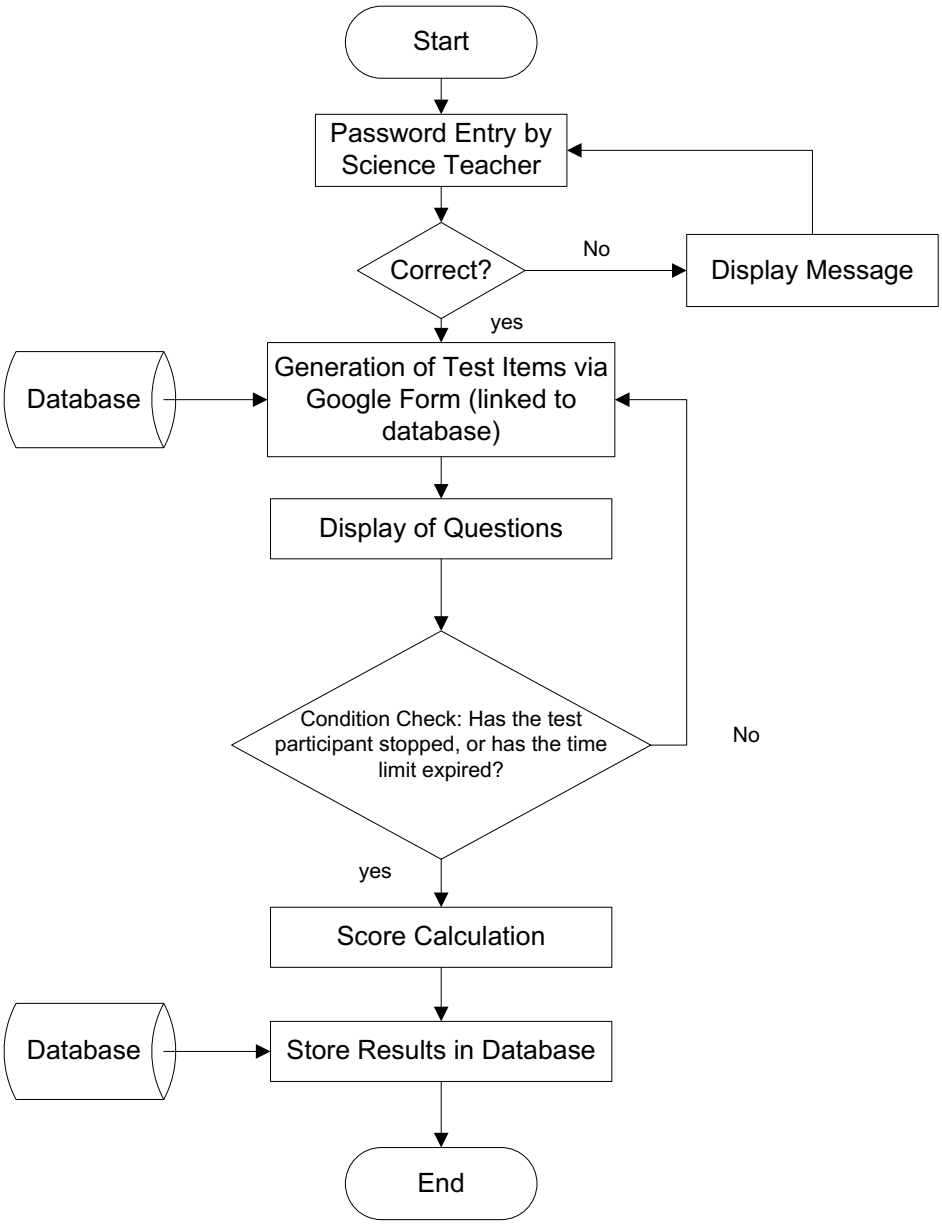

Figure 3. Online Diagnostic Assessment Application Algorithm

The Desseminate Stage

This stage introduces the developed online diagnostic assessment instrument on a wider scale, which is delivered in a science teacher workshop, national seminar, or international seminar in the field of education, attended by various groups, including teachers, lecturers, and science education students. The number of respondents was 67. Then, the results were analyzed for validity and reliability. The reliability value shows strong, namely the reliability value with Cronbach's alpha of 0.736 and reliability with McDonald's of 0.768. Meanwhile, the results of the construct validity analysis with EFA show the MSA value > 0.5. The study tests the validity of measurement instruments using factor analysis techniques. Exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) were used to assess construct validity (Ardi Waluyo & dan Sulhadi, 2020; Sarip et al., 2022).

Corresponding author: Widowati Pusporini (widowatipusporini@uny.ac.id)

The results of the construct validity analysis with EFA show that the 30 items are valid since all KMO MSA values are > 0.5 (Sutopo, Y., and Slamet, A., 2017). This result follows the study Suharsono & Istiqomah (2014). The self-efficacy scale was adapted and validated using qualitative and quantitative item analysis, with an item-total correlation > 0.50 considered acceptable. In addition to validity and reliability, discriminant analysis and IRT analysis with the Rusc Model were also conducted. And, the results are presented in Table 1.

Table 1. Difficulty and Discrimination Index

Item	Difficulty	ULI	RIT	RIR
1	0.687	0.3636	0.2226	0.1254
2	0.91	0.2273	0.3881	0.3343
3	0.582	0.5455	0.3937	0.2991
4	0.701	0.5	0.402	0.315
5	0.299	0.7727	0.6462	0.5834
6	0.284	0.0909	0.1437	0.0476
7	0.164	0.2273	0.3061	0.2317
8	0.119	0.3182	0.5502	0.4989
9	0.552	-0.0455	0.0726	-0.034
10	0.806	0.3636	0.3646	0.2878
11	0.537	0.4545	0.398	0.3025
12	0.821	0.4091	0.4363	0.3662
13	0.388	0.0455	0.1957	0.0927
14	0.493	0.7273	0.5189	0.4785
15	0.358	0.6818	0.596	0.5235
16	0.478	0.2727	0.366	0.2681
17	0.791	0.5	0.518	0.4499
18	0.284	0.2727	0.3281	0.2381
19	0.701	0.2273	0.1923	0.0956
20	0.642	0.5909	0.4577	0.3708
21	0.403	0.6818	0.5349	0.4534
22	0.657	0.2727	0.2347	0.1355
23	0.388	-0.6364	-0.5393	-0.6073
24	0.403	0.3636	0.3849	0.2902
25	0.672	0.5	0.3954	0.3056
26	0.433	0.5	0.384	0.2881
27	0.776	0.3636	0.3081	0.22
28	0.478	0.5	0.4236	0.3299
29	0.373	0.2273	0.2221	0.1207
30	0.597	0.5909	0.5017	0.4168

The table 1 presents the results of the Item Difficulty and Discrimination Index analysis for a number of items in the test. Item Difficulty measures the difficulty of the item, with values from 0 to 1. Higher values indicate easy items, such as item 2 with a value of 0.910, which means that more than 90% of respondents

Corresponding author: Widowati Pusporini (widowatipusporini@uny.ac.id)

answered correctly. In contrast, item 7 has a difficulty of 0.164, indicating that this item is difficult, as only a few respondents answered correctly. Upper-Lower Index (ULI) measures the ability of an item to differentiate respondents based on ability level. For example, item 14 with a ULI of 0.7273 shows high effectiveness in distinguishing between high and low ability participants, while item 9 with a negative ULI (-0.0455) does not distinguish effectively. Item-Total Correlation (RIT) assesses the consistency of the items with the overall performance, where item 13 has a value of 0.5960, indicating high consistency, while item 10, with 0.3466, is less consistent. Item-Rest Correlation (RIR) measures an item's unique contribution to the overall test, with item 14 having an RIR of 0.5235, indicating a significant contribution. Meanwhile item 9, with a negative value (-0.0340), is less effective and may need to be improved or removed. Overall, item 14 was rated effective, item 9 was problematic, and item 2 was too easy for respondents. This analysis provides guidance to improve or replace less effective items.

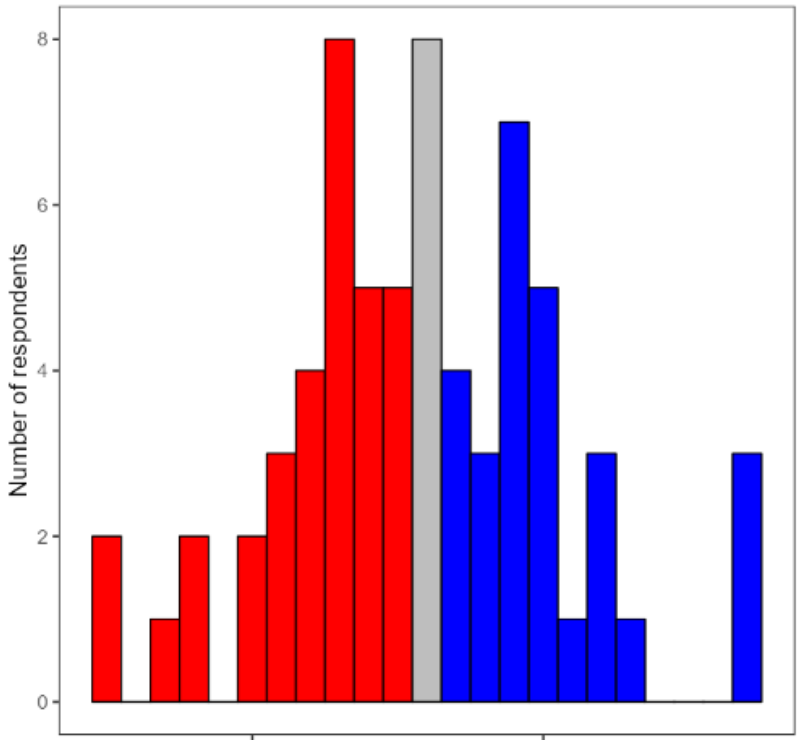

Figure 4. Histogram of Total Score

The histogram shows the distribution of respondents' total scores, with the horizontal axis representing the range of scores and the vertical axis the number of respondents. The red color depicts low-scoring respondents, while the blue color indicates high-scoring respondents, with the gray color around scores 14-15 possibly indicating the median point. Most of the low-scoring respondents (5-13) are concentrated at scores 9-11, with a peak of about 8 respondents. Respondents with high scores (15-24) are more spread out, with peaks around scores 18 and 20. Also, this histogram shows that the majority of respondents are in the middle of the distribution, with a clear distinction between the more concentrated low score group and the more spread out high score group.

Corresponding author: Widowati Pusporini (widowatipusporini@uny.ac.id)

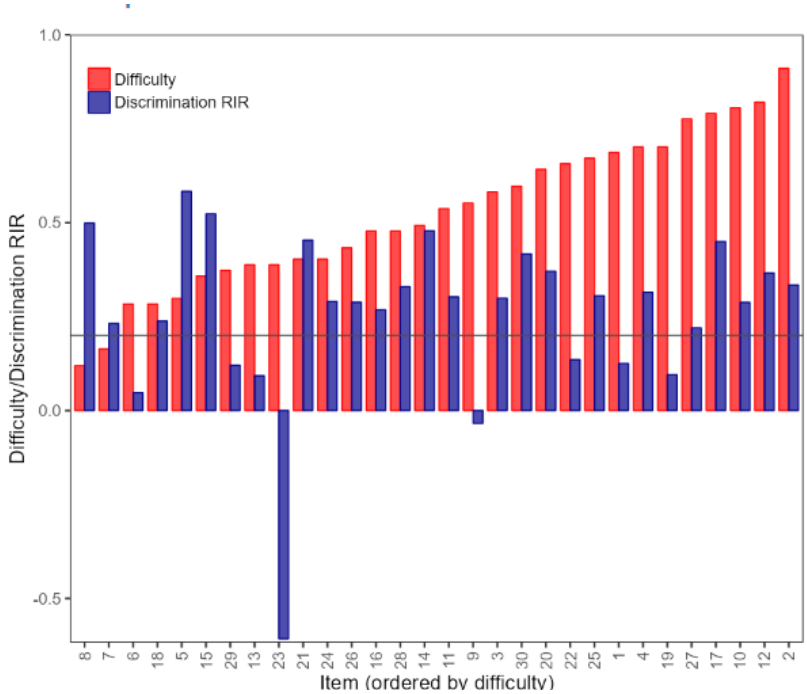

Figure 5. Discriminan Plot

This graph shows the analysis results of the two aspects of the items, namely difficulty and discrimination RIR, each item ranked by difficulty level. The difficulty level is indicated by a red bar, which represents how challenging an item is for respondents to answer correctly. The higher the difficulty value, the more difficult the item is to answer. The graph shows that the further to the right, or the larger the item number, the difficulty level tends to increase, indicating a pattern of increasing difficulty according to item number.

Meanwhile, the discrimination RIR is indicated by a blue bar, which measures the ability an item to differentiate between high and low of participants' ability. A positive value indicates that the item can discriminate well. In contrast, a negative value or close to zero indicates that the item is less effective in discriminating between participants. Some items, such as item 23, have a sizable negative power difference, around -0.5. It means that the item discriminates participants in reverse, where lower ability participants are more likely to answer correctly than high ability participants.

Based on the analysis, item 23 stands out as a misfit, as its discriminating power is negative, which indicates that this item may not be functioning as intended and needs to be considered for revision or deletion. Most of the other items have positive discriminating power, especially the items in the beginning and middle, such as items 8, 15, and 29, indicating that they are functioning well. There are some exceptions, but there is not always a direct relationship between difficulty and distinctiveness. Some of the more difficult items, such as item 2, still have good distinguishing power, while easier items, such as item 7, also show high distinguishing power. However, some items are difficult but low in discriminating power, such as item 23, which suggests that although they are difficult, they do not provide good information in discriminating between participants.

Corresponding author: Widowati Pusporini (widowatipusporini@uny.ac.id)

Overall, items with higher difficulty tend to have lower power, although there are some exceptions. Items with negative power, such as item 23, need to be re-evaluated as they may lead to less accurate measurement results. In contrast, items with high positive power, such as items 7, 8, and 15, can be considered as quality items in the measurement of this test.

Table 2. Item Statistics

Item	Proportion	Measure	S.E.Measure	Infit	Outfit
1	0.687	-0.8707	0.276	1.051	1.252
2	0.91	-2.515	0.437	0.898	0.903
3	0.582	-0.3711	0.261	0.982	0.968
4	0.701	-0.9478	0.279	0.969	0.924
5	0.299	0.944	0.28	0.819	0.744
6	0.284	1.0237	0.284	1.103	1.204
7	0.164	1.7867	0.342	0.955	1.197
8	0.119	2.1834	0.389	0.832	0.621
9	0.552	-0.2362	0.259	1.187	1.189
10	0.806	-1.5665	0.32	0.958	0.965
11	0.537	-0.1694	0.258	0.98	0.979
12	0.821	-1.6722	0.33	0.92	0.821
13	0.388	0.5028	0.264	1.094	1.114
14	0.493	0.0299	0.258	0.876	0.865
15	0.358	0.6443	0.268	0.848	0.838
16	0.478	0.0963	0.258	1.01	0.981
17	0.791	-1.4665	0.312	0.873	0.786
18	0.284	1.0237	0.284	1.0	1.01
19	0.701	-0.9478	0.279	1.076	1.217
20	0.642	-0.6496	0.268	0.934	0.907
21	0.403	0.4335	0.262	0.902	0.864
22	0.657	-0.7218	0.27	1.078	1.087
23	0.388	0.5028	0.264	1.546	1.763
24	0.403	0.4335	0.262	0.984	0.957
25	0.672	-0.7955	0.273	0.978	0.962
26	0.433	0.2972	0.26	0.99	0.999
27	0.776	-1.3713	0.305	1.008	1.022
28	0.478	0.0963	0.258	0.961	0.937
29	0.373	0.573	0.266	1.074	1.132
30	0.597	-0.4395	0.262	0.91	0.894

Table 2 shows statistical information related to the performance of some items in a test, including the proportion of correct answers (proportion), difficulty level (measure), standard error of measurement (S.E. Measure), and Infit and Outfit values to evaluate the fit of items to the measurement model. The Proportion column shows the percentage of respondents who answered correctly on each item, where higher values mean more participants answered correctly. For example, Item 2 has a proportion of 0.910, which means 91% of respondents answered correctly. The score indicates that this item is very easy. In contrast, Item 8 has a proportion of 0.119, indicating that only about 12% of participants answered correctly, signaling that this item is difficult. The Measure column

Corresponding author: Widowati Pusporini (widowatipusporini@uny.ac.id)

describes the level of difficulty in logit units. A positive value indicates a more difficult item, while a negative value indicates an easier item. For example, Item 2 with a Measure value of -2.5150 is very easy. Meanwhile, Item 8 with a value of 2.1834 is very difficult. In general, the higher the Measure value, the more difficult the item is for participants to answer correctly.

The S.E. Measure column indicates the uncertainty in the item difficulty estimate. The smaller the value, the more accurate the difficulty estimate. Item 1 has an S.E. Measure of 0.276, which means the difficulty estimate is fairly accurate, while Item 8 has an S.E. Measure of 0.389, indicating more uncertainty. The Infit value illustrates how well the item fits the measurement model, particularly against responses that provide important information. The expected value is 1.0. If this value is too high (e.g., above 1.5), the item is considered too noisy. In contrast, a too low value (below 0.7) indicates that the item is too perfect. In Item 23, the Infit value is 1.546, indicating inappropriate variation, while Item 10 has an Infit of 0.958, close to the expected value.

Outfit values, similar to Infit, are more sensitive to outliers or unusual responses. The expected Outfit value is also 1.0. Item 23 has an Outfit of 1.763, indicating many outliers, while Item 4, with a value of 0.744, indicates a more appropriate level of variation. From this analysis, Item 2 and Item 23 show different characteristics. Item 2 is an easy item with a high proportion of correct answers and good Infit and Outfit. Meanwhile, Item 23 has high Infit and Outfit values, indicating that this item needs further evaluation. Most of the other items, such as Item 4, Item 10, and Item 15, have Infit and Outfit values close to 1.0, indicating that these items work according to the model. Overall, this table provides an indication that the majority of the items in this test are of good quality. Although there are a few exceptions, such as Item 23, which needs further evaluation to ensure its quality and relevance in measurement.

Corresponding author: Widowati Pusporini (widowatipusporini@uny.ac.id)

Journal of Education and Teaching (JET) Volume 7 No. 1, 2026
DOI: 10.51454/jet.v7i1.546

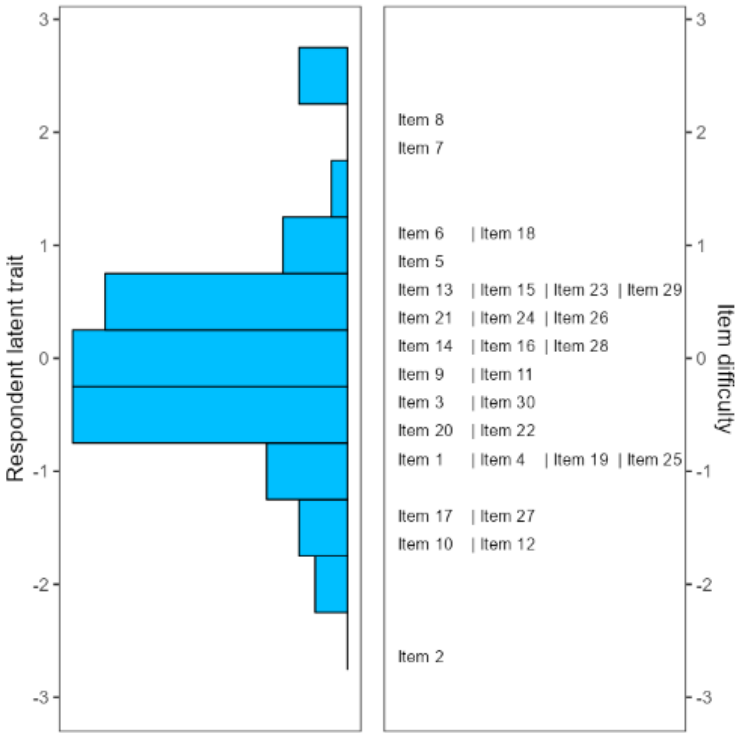

Figure 6. Wright Map Diagram

The Wright Map diagram in Figure 6 shows two main components in the analysis of test items, namely the distribution of respondent ability and the level of item difficulty in logit units. On the left side, the distribution of participants' ability is seen with most respondents having average ability around logit 0. Meanwhile, only a few have very high ability (+2 logit and above) or very low ability (-1 logit and below). On the right side, the level of item difficulty is shown, with more difficult items, such as Items 8 and 7, at the top (around +2 logit), and easier items, such as Item 2, at the bottom (-2.5 logit). Most items fall in the logit range of 0 to +1, which corresponds to the majority of participants' ability. This suggests that the test is suitable for measuring participants with average to slightly above average ability. However, participants with very high or very low ability may find the test less challenging or too difficult due to the limited number of items that match their ability. Item 2 is the easiest, while Items 8 and 7 are the most difficult.

Corresponding author: Widowati Pusporini (widowatipusporini@uny.ac.id)

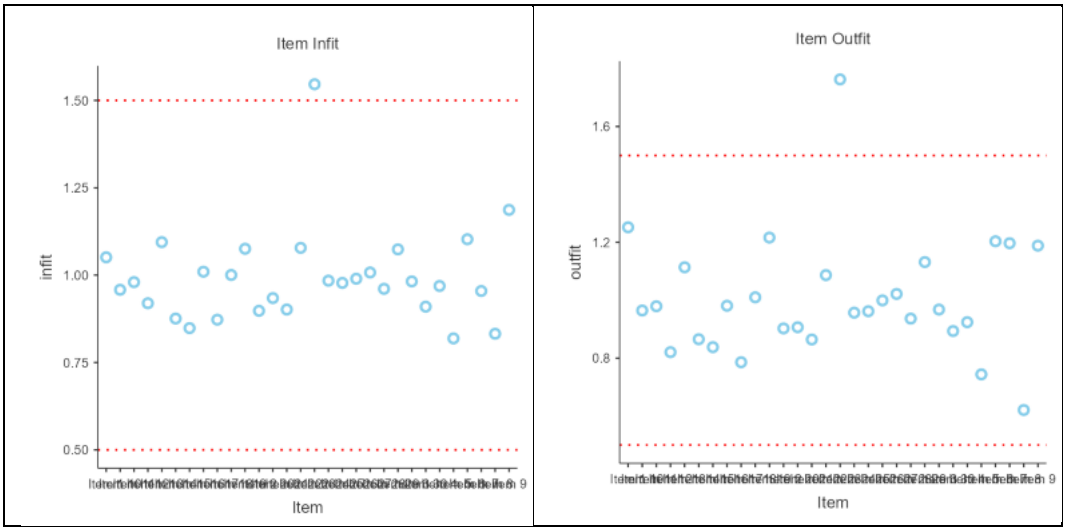

Figure 7. Infit Items and Outfit Items

The figure shows two graphs, namely " Infit Item" and " Outfit Item," which are used in the Item Response Theory (IRT) model to evaluate the extent of the test items function in measuring individual abilities. The Infit graph shows the sensitivity of the items to individual abilities with a value range of 0.50 to 1.50. Most items have infit values between 0.75 and 1.25, indicating a good fit with the model. However, some items outside the range, especially those approaching 1.50, require further attention because they may not fit the model well. The Outfit graph shows the sensitivity of the items to outliers, with most items falling within a reasonable range of values. Several items with outfit values above 1.2 indicate minor misfits, while one item approaches 1.6, indicating a greater outlier effect. Overall, both graphs show that most items fall within reasonable limits of fit. But items that fall outside the limits need further examination to ensure their function in measuring the intended ability.

Discussion

Instrument validity is an important aspect in ensuring that the developed product is truly capable of measuring what is intended, namely, students' learning readiness in the implementation of the Independent Curriculum at the junior high school level. The instrument was validated through three aspects: content, construct, and language, using the Aiken formula. The validity value calculated using the Aiken formula (Aiken's V). It shows that all items assessed by five experts have a V value higher than 0.80. According to the Aiken table, this value indicates that all items are considered valid (Aiken's, 1980). The results of the expert assessment include aspects of content (such as the suitability of questions to indicators and homogeneity of answers), constructs (neatness of questions, accuracy of question formulation, and structure of answer choices), and language (use of good and communicative Indonesian). All of these aspects received a V value ≥ 0.85, which means valid.

The reliability value of the instrument was measured using two methods, namely Cronbach's Alpha and McDonald's Omega. The results showed that the

Corresponding author: Widowati Pusporini (widowatipusporini@uny.ac.id)

reliability value of Cronbach's Alpha was 0.736 and the reliability of McDonald's Omega was 0.768. These values indicate that the instrument has good reliability. In general, reliability higher than 0.70 is considered adequate, so this instrument is consistent in measuring (Ramly et al., 2022; Sumin et al., 2022; Wladis & Samuels, 2016).

In terms of construct validity, the results of exploratory factor analysis (EFA) show that all items have an MSA (Measures of Sampling Adequacy) value higher than 0.5, indicating that each item has an adequate correlation with the construct being measured (Rani et al., 2021). This shows that the instrument is feasible to be used to measure students' learning readiness. The results of the quantitative to qualitative scale conversion for the test items show that the total score is 75.00, which places it in the "Very Feasible" category, according to the scale used. This means that the developed product is considered very feasible by the experts. This score is obtained from an assessment of three aspects (content, construct, and language).

Item analysis is essential to develop high-quality multiple-choice tests. Effective item distractors should appeal to low- and middle-ability groups while correct answers differentiate between high-ability students (Asril & Marais, 2011). Differential Diversion Function Analysis can examine the interaction between population subgroups and option choices while controlling for ability (Green et al., 1989). Rasch model item distractor analysis can detect whether item distractors provide diagnostic information, especially for low-ability groups, by meeting content and statistical criteria (Asril & Marais, 2011). Malfunctioning item distractors can make items too easy and fail to differentiate between top and bottom groups, requiring revision (Asril & Marais, 2011). Comparing item response distributions across groups can reveal inconsistencies with one-dimensional latent variable differences (Rosenbaum, 1985). This analysis contributes to improving test quality, providing valuable insights for teaching and learning, and ensuring fair assessment across ability groups.

Item analysis is very important to evaluate the quality of multiple-choice questions (MCQs) in an examination. Difficulty index (P) and discrimination index (D) are the main parameters to assess the quality of MCQs (Pande et al., 2013; Singh Rana, 2014). Generally, items with P values between 30% and 70% are considered acceptable, while those below 30% are difficult and those above 70% are easy (Pande et al., 2013; Singh Rana, 2014). Higher D values indicate better discrimination between high and low achieving students. Studies have shown that most MCQs fall within the acceptable range for P and D (Pande et al., 2013; Singh Rana, 2014). The relationship between P and D is not completely linear; questions that are classified as medium difficulty tend to have the highest discriminating power (Pande et al., 2013; Singh Rana, 2014). Liu, (2014) states that the discrimination index can be determined mathematically based on the item difficulty level and the correlation between the item's performance and the total test score. For example, Item 14 has a ULI (Upper-Lower Index) of 0.7273, indicating that the item is very good at differentiating the abilities of test takers.

Corresponding author: Widowati Pusporini (widowatipusporini@uny.ac.id)

The discrimination graph and histogram of total scores provide a more in-depth picture of the distribution of difficulty and effectiveness of the test items in differentiating student abilities. The histogram shows the distribution of total scores of respondents, with the majority in the middle, indicating a balanced distribution between low, medium, and high-ability students. This study discusses the use of infit and outfit statistics in Item Response Theory (IRT) to assess the fit of items and people (Walker et al., 2018). The Infit graph shows that most items have values between 0.75 and 1.25, indicating a good fit, although some items approaching 1.50 need attention. Meanwhile, the Outfit graph shows that most items are within a reasonable range of values. But there are items with values above 1.2, indicating a misfit, and one item approaching 1.6, indicating the influence of an outlier. Although most items are in good fit, items that are out of bounds need further evaluation to ensure measurement accuracy.

This instrument is valid and reliable based on the validity and reliability test. The items in the test can differentiate students based on their abilities, with some items needing minor improvements for distractors or effectiveness.

CONCLUSION

The developed instrument showed high validity with an Aiken's V value above 0.80, indicating that all tested items were valid and good reliability with Cronbach's Alpha 0.736 and McDonald's Omega 0.768. It reflects adequate consistency in measurement. The results of exploratory factor analysis (EFA) showed that all items had Measures of Sampling Adequacy (MSA) of more than 0.5, proving the instrument's feasibility in measuring student readiness. Item analysis also indicated that most of the questions met the criteria for difficulty and good discrimination, although some items need revision to improve their effectiveness. In addition, the balanced distribution of scores among students with different abilities, as seen from the histogram and discrimination plot, indicated that this instrument was effective in differentiating student abilities. Overall, the developed ODA proved to be valid and reliable, ready to be used in assessing student readiness in science learning, and made a significant contribution to the development of diagnostic assessments in a broader educational context, and supported the implementation of the Merdeka Curriculum.

ACKNOWLEDGMENT

We would like to express the gratitude to Yogyakarta State University (UNY) for the funding support provided through the university research grant scheme. This research was conducted as part of a research program managed by the Directorate of Research and Community Service at Yogyakarta State University. The author would also like to thank all those who have provided assistance and contributed to the implementation of this research.

REFERENCES

Corresponding author: Widowati Pusporini (widowatipusporini@uny.ac.id)

Aiken's, L. R. (1980). Content Validity and Reliability of Single Items or Questionnaires. *Educational and Psychological Measurement*, *40*(4), 955–959. https://doi.org/10.1177/001316448004000419

Ardi Waluyo, A., & dan Sulhadi, H. (2020). Analisis Faktor: Validitas Konstruk Instrumen Penilaian Keterampilan Berpikir Kritis. In *Jurnal Pendidikan Fisika Tadulako Online (JPFT)* (Vol. 8, Issue 1).

Asril, A., & Marais, I. (2011). *Applying a Rasch Model Distractor Analysis Implications For Teaching And Learning*. Sense Publishers.

Benjamin, A. (2024). Studying the student's perceptions of engagement and problem-solving skills for academic achievement in chemistry at the higher secondary level. *Education and Information Technologies*, *29*(7), 8347–8368.

bin Rani, N., Baharudin, H., & Isa Hamzah, M. (2021). Validation of Existing Arabic Language Reading Knowledge Instrument (SPSAMBA): Exploratory Factor Analysis (EFA). In *PSYCHOLOGY AND EDUCATION* (Vol. 58, Issue 1). www.psychologyandeducation.net

Draude, B., & Brace, S. (1999). *Assessing the Impact of Technology on Teaching and Learning: Student Perspectives Introduction Method Survey Results and Major Findings Conclusion Contact*. Eric. http://www.mtsu.eduk-itconf/proceed99/brace.html

Ghory, S., & Ghafory, H. (2021). The impact of modern technology in the teaching and learning process. *International Journal of Innovative Research and Scientific Studies*, *4*(3), 168–173. https://doi.org/10.53894/ijirss.v4i3.73

Green, B. F., Crone, C. R., & Folk, V. G. (1989). A Method for Studying Differential Distractor Functioning. In *Journal of Educational Measurement Summer* (Vol. 26, Issue 2).

Kalyani, L. K. (2024). The role of technology in education: Enhancing learning outcomes and 21st century skills. *International Journal of Scientific Research in Modern Science and Technology*, *3*(4), 5–10.

Leitão, G., Colonna, J., Monteiro, E., Oliveira, E., & Barreto, R. (2020). *New Metrics for Learning Evaluation in Digital Education Platforms*. http://arxiv.org/abs/2006.14711

Liu, X. S. (2014). A note on statistical power in multi-site randomized trials with multiple treatments at each site. *British Journal of Mathematical and Statistical Psychology*, *67*(2), 231–247. https://doi.org/10.1111/bmsp.12016

Pande, S. S., Pande, S. R., Parate, V. R., Nikam, A. P., & Agrekar, S. H. (2013). Correlation between difficulty & discrimination indices of MCQs in formative exam in Physiology. In *Asian Journal of Medical Education* (Vol. 7, Issue 1).

Ramly, S. N. F., Ahmad, N. J., & Yakob, N. (2022). Development, validity, and reliability of chemistry scientific creativity test for pre-university students. *International Journal of Science Education*, *44*(14), 1–16. https://doi.org/10.1080/09500693.2022.2116298

Rapaka, A., Dharmadhikari, S. C., Kasat, K., Mohan, C. R., Chouhan, K., & Gupta, M. (2025). Revolutionizing learning− A journey into educational games with immersive and AI technologies. *Entertainment Computing*, *52*, 100809.

Corresponding author: Widowati Pusporini (widowatipusporini@uny.ac.id)

Retnawati, H. (2014). Teori Respon Butir dan Penerapanya. In *Yogyakarta: Nuha Medika*. Nuha Medika.

Rosenbaum, P. R. (1985). Comparing distributions of item responses for two groups. *British Journal of Mathematical and Statistical Psychology*, *38*(2), 206–215. https://doi.org/10.1111/j.2044-8317.1985.tb00836.x

Salinas-Navarro, D. E., Vilalta-Perdomo, E., Michel-Villarreal, R., & Montesinos, L. (2024). Designing experiential learning activities with generative artificial intelligence tools for authentic assessment. *Interactive Technology and Smart Education*.

Sarip, M., Amintarti, S., Utami, N. H., Biologi, S. P., Lambung, U., Brigjen, M. J., Brigjend, J., Basri, H., Utara, K. B., Banjarmasin, K., & Selatan, K. (2022). *Validitas Dan Keterbacaan Media Ajar E-Booklet Untuk Siswa SMA/MA Materi Keanekaragaman Hayati* (Vol. 1, Issue 1).

Singh Rana, S. (2014). *Test Item Analysis and Relationship Between Difficulty Level and Discrimination Index of Test Items in an Achievement Test in Biology Education*.

Sudarsana, I. K., Nakayanti, A. R., Sapta, A., Haimah, Satria, E., Saddhono, K., Achmad Daengs, G. S., Putut, E., Helda, T., & Mursalin, M. (2019). Technology Application in Education and Learning Process. *Journal of Physics: Conference Series*, *1363*(1). https://doi.org/10.1088/1742-6596/1363/1/012061

Suharsono, Y., & Istiqomah, I. (2014). Validitas Dan Reliabilitas Skala Self-Efficacy. *Jurnal Ilmiah Psikologi Terapan*, *02*(01), 144–151.

Sumin, Sukmawati, F., Setiawati, F. A., & Asmawi, S. (2022). The Impact of Z-Score Transformation Scaling on the Validity, Reliability, and Measurement Error of Instrument SATS-36. *Jurnal Pengukuran Psikologi Dan Pendidikan Indonesia*, *11*(2), 166–180. https://doi.org/10.15408/jp3i.v11i2.26591

Sutopo, Y dan Slamet, A. (2017). *Statistika Inferensial*. ANDI.

Walker, A. A., Jennings, J. K., & Engelhard, G. (2018). Using person response functions to investigate areas of person misfit related to item characteristics. *Educational Assessment*, *23*(1), 47–68. https://doi.org/10.1080/10627197.2017.1415143

Wladis, C., & Samuels, J. (2016). Do online readiness surveys do what they claim? Validity, reliability, and subsequent student enrollment decisions. *Computers and Education*, *98*, 39–56. https://doi.org/10.1016/j.compedu.2016.03.001

Zhang, L., & Xu, J. (2025). The paradox of self-efficacy and technological dependence: Unraveling generative AI's impact on university students' task completion. *The Internet and Higher Education*, *65*, 100978.

Corresponding author: Widowati Pusporini (widowatipusporini@uny.ac.id)