JPPIPA 10(10) (2024)

Jurnal Penelitian Pendidikan IPA

Journal of Research in Science Education

http://jppipa.unram.ac.id/index.php/jppipa/index

Authentic Assessment Instrument on Redox Reactions to Assess Students' Cognitive Skills

Qurratu 'Ainillana[1*], Isana Supiah Yosephine Louise[1]

[1] Departement of Chemistry Education, Faculty of Mathematics and Natural Science, Universitas Negeri Yogyakarta, Indonesia.

Received: April 29, 2024
Revised: August 09, 2024
Accepted: October 25, 2024
Published: October 31, 2024

Corresponding Author:
Qurratu 'Ainillana
qurratuainillana.2020@student.uny.ac.id

DOI: 10.29303/jppipa.v10i10.8791

Abstract: Authentic assessment is an assessment approach that enable students to demonstrate in-depth understanding and solve complex problem using systematic thinking skills, but teachers find it difficult to develop due to limited knowledge and time. The aim of this research is to produce an authentic assessment instrument for redox reaction materials; determine the characteristics, quality, and feasibility. The instrument development model used was a combination of Oriondo & Dallo-Antonio with Cohen Swerdlik. The development stages were assessment planning, instrument construction, and product testing. The research subjects were 155 of 10th-grade students from three public high schools in Bukittinggi, selected using random sampling technique. Data collection techniques were tests and non-test, the instruments were open ended questions and validation sheets. The authentic assessment instrument developed criteria are: (1) can assess students' higher order thinking skills; (2) integrate knowledge with skills and various scientific disciplines; (3) develop 21st century competencies; and (4) involve real or everyday life contexts for cognitive aspect. The trial results were analyzed using the PCM 1-PL approach at IRT using Winstep and Parscale program. The characteristics of the item were stated fit because it met the acceptance requirements for MNSQ, ZSTD, and PT-Measure Correlation and the difficulty is in a relatively good level. The reliability of the instrument was classified as good, with test reliability of 0.72 and item reliability of 0.96.

Keywords: Assessment Instruments; Authentic Assessment; Higher Order Thinking Skills; Redox Reactions

Introduction

Assessment in learning is essential but difficult to do, because it is impossible to know what is exactly in students' minds (Stowe & Cooper, 2019). Assessment is carried out by reasoning through the evidence shown by students on exams, assignments, or performance in various activities designed from competencies or learning indicators. Without it, it is hard to know whether students are just remembering or really understanding the material. Designing learning assessments that can provide sufficient information to evaluate students' abilities is a challenging task for teachers. The 2013 curriculum is based on Minister of Education and Culture Regulation no. 104 of 2014, requires the use of authentic assessment as the main assessment approach in assessing learning outcomes.

Developing authentic assessment instruments requires careful planning while teachers have limited time (Ambiyar et al., 2020; Ekawati, 2012; Ismiati et al., 2019; Wulandari et al., 2018). Authentic assessment can explore students' in-depth understanding to solve complex problems using critical thinking skills systematically by integrating several scientific disciplines and/or between theory and practice. It means authentic assessment involves questions with a high level of thinking skills known as HOTS (Higher Order Thinking Skills) questions (Bushkofsky, 2016).

How to Cite:
'Ainillana, Q., & Louise, I. S. Y. (2024). Authentic Assessment Instrument on Redox Reactions to Assess Students' Cognitive Skills. *Jurnal Penelitian Pendidikan IPA*, *10*(10), 7437–7446. https://doi.org/10.29303/jppipa.v10i10.8791

However, some high school chemistry teachers who are graduated from UNY developed UAS questions that are still dominated by LOTS and MOTS questions (Iskandar & Senam, 2015). Some teachers have the misperception that high-level thinking skills questions are identical to difficult questions, so they are not easy to apply to classes with students who have various cognitive abilities (Nurmawati et al., 2021).

Teachers find difficulties in involving students in the assessment process (Ekawati, 2012; Wahyuni et al., 2021; Hanifah & Irambona, 2019; Ismiati et al., 2019; Kartowagiran & Jaedun, 2016; Suastra & Ristiati, 2017). The character of students is unique, so teachers have difficulty managing time and implementing instruments such as directing students to instill good attitudes in accordance with learning objectives (Nuriana, 2018; (Sudiana et al., 2018) Sund, 2016). In fact, the integration of knowledge, skills and attitudes in assessment is one of the dimensions of authentic assessment (Gulikers et al., 2004).

Apart from the attitude aspect, similar obstacles are also experienced in assessing other observation technique assessments such as assessing practical skills or performance aspects. Research results show that direct assessment of practical skills is still limited (Hancock & Hollamby, 2020). The condition was more challenging because of the Covid-19 Pandemic, assessment in learning online must remain active and authentic. However, no instructional design has been found in online authentic assessment (Sutadji et al., 2021). Teachers still have to carry out various types of authentic assessments such as performance assessments, simple projects, portfolios and written assessments (Salirawati, 2021). This condition is a major obstacle to chemistry learning, especially laboratory activities, because apart from being encouraged to learn using new technology, educators must also quickly explore and design alternative assessments to replace direct exams and written tests (Lau et al., 2020).

Authentic assessment strategies are expected to achieve learning objectives and support a balance between learning activities, assessments and learning outcomes so that the real-world impacts produced are of truly authentic quality (Hasan & Cerimagic, 2021). Students in Taiwan have a good understanding of the concept of redox reactions based on the release and binding of oxygen, but one third of class X students cannot explain the phenomenon using redox reaction theory accurately (Chiang et al., 2014). Adu-Gyamfi & Ampiah (2019) stated that alternative conceptions of chemistry students are associated with the application of oxidation and reduction processes in real life contexts. In other research, it was found that prospective chemistry teachers were unable to connect the three levels of chemical representation (macroscopic, microscopic and symbolic) in chemistry learning, especially the redox concept (Hadinugrahaningsih et al., 2022). The assessment, teaching and learning processes are closely related to each other and are part of the pedagogical process (Villarroel et al., 2018).

Therefore, it is important for teachers and prospective educators have complete knowledge to develop authentic assessment instruments on redox reactions material. This research aims to develop an authentic assessment instrument on class X high school redox reaction material which is still rarely found in previous research results. This instrument includes assessment instruments on cognitive aspect with indicators and scoring guidelines that can be used in both online and offline learning. The question instrument items are also designed to develop students' high-level thinking skills, and are linked to daily life, and encourage students to be able to integrate various scientific disciplines.

Method

Research Design

This research is development research that followed the specific procedure for compiling and developing instrument by Oriondo & Dallo-Antonio (1984) combined with the instrument development model of Cohen & Swerdlik (2018).

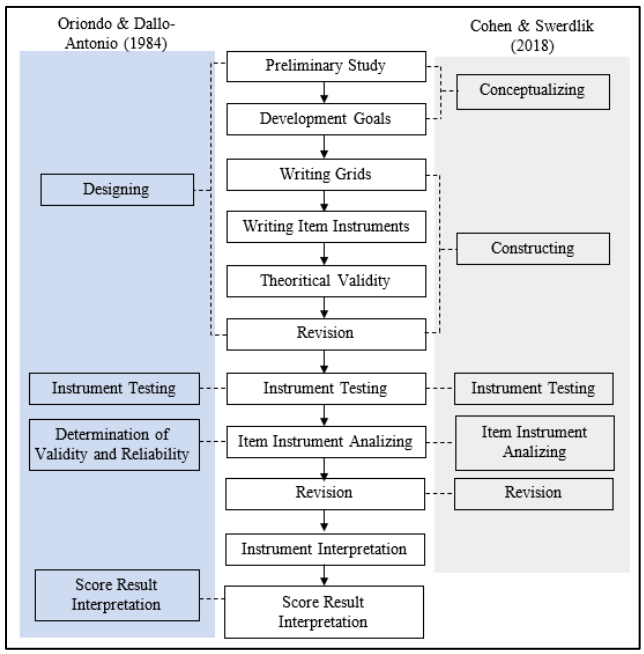

Figure 1. Research Flow

Time and Place

The empirical test of the instrument was carried out in the fourth week of April 2022 to the second week of

May 2022 at SMAN 1 Bukittinggi, SMAN 3 Bukittinggi, and SMAN 5 Bukittinggi.

Subject

The research subjects were 155 students in the empirical testing stage. The research subjects were 10th grade of science students from three public high schools in Bukittinggi and the population was the 10th grade students of public senior high school in Bukittinggi. The sample selection technique uses random sampling.

Procedure

The instrument development stages include preliminary studies, determining development objectives, writing grids, writing instrument items, determining theoretical validity, revisions, instrument testing, instrument item analysis, interpretation of instrument implementation score results.

Data, Instruments, and Data Collection Techniques

The data obtained is quantitative data (student scores resulting from empirical tests) and qualitative data (results from content validation). Data collection techniques used test and non-test techniques. The data collection instruments were an instrument validation sheet, two packages of question instrument consist of 12 items open ended question each.

Data Analysis Technique

The results of content validation were analyzed qualitatively. The empirical test results were analyzed using an approach Partial Credit Model 1-Parameter Logistic (PCM 1-PL) pada Item Response Theory (IRT) using the program Winstep and Parscale.

Result and Discussion

Initial Product Result

At the initial stage, assessment planning is carried out, consist of a preliminary study and determining the objectives of instrument development. Based on the literature review, in general the criteria for authentic assessment in the cognitive aspect are (1) solving complex problems, conceptual thinking and inquiry skills, ensuring levels of critical thinking (C4 – C6) are engaged, (2) coherence of knowledge and skills and resources used, (3) integration of various scientific disciplines, connecting with prior knowledge, (4) developing 21st century competencies, (5) real context or everyday life.

Table 1. Authentic Assessment Indicators

Cognitive Level	Authentic Assessment Indicators
Analyzing (C4)	Apply the knowledge to analyze phenomena in everyday life.
	Investigate an event and relate it to the knowledge gained
	Compare situations based on observations and knowledge gained
Evaluating (C5)	Proving that a phenomenon is in accordance/not in accordance with the knowledge held.
	Formulate the conclusion of a discourse related to phenomena and organize it to answer questions.
	Critical thinking in considering a term/statement based on examples in everyday life.
	Evaluate alternative problem solving strategies and solutions
Creating (C6)	Combining the data or information obtained to create statements that represent more general relationships and broader terms that apply (resuming).

Instrument construction is carried out based on the results of assessment planning, including writing instrument grids and writing instrument items. Indicators of authentic assessment instruments in the cognitive aspect can be seen in Table 1. The instrument design was then validated in content with material experts, instrument development experts, and practitioners or teachers in schools. Validation was carried out qualitatively, then the results of the validator's corrections were revised for empirical testing.

Empirical Product Test Result

Analysis of Prerequisite Assumptions

Unidimensional Test

To test the adequacy of the sample size, the Kaiser-Mayer-Olkin Measure of Sampling Adequacy (KMO-MSA) test was carried out. A KMO test value > 0.6 indicates a sufficient sample size (Shrestha, 2021). The closer the KMO test value is to 1, the more ideal the sample size is. Bartlett's test is indicated by a small significance value (usually smaller than 0.050) which shows that the correlation matrix is identical to the identity matrix (Bartlett, 1951). The KMO and Bartlett tests using the SPSS 16.0 program are shown in Table 2.

Table 2. KMO-MSA and Bartlett Test Result

Kaiser Mayer-Olkin Measure of Sampling Adequacy		0.688
Bartlett's Test of Sphericity	Approx. Chi-Square	360.894
	df	66
	Sig.	.000

Based on the results of the KMO and Bartlett tests in Table 2, the significance value is 0.000. It indicates that the correlation between variables is zero. The results of the KMO and Bartlett tests shows that the sample size of 155 with 12 items used is appropriate so it can be continued.

After the KMO and Bartlett tests were carried out and it was determined that the sample size was appropriate, the anti-image value was also checked. Anti-image correlation is a concept used in Exploratory Factor Analysis to measure the unique contribution of each variable to a common factor. The anti-image correlation test results table shows the numbers that form a diagonal marked "*a*", which denotes the MSA figure (Measure of Sampling Adequacy) a variable. The anti-image correlation value on the diagonal number must be greater than 0.5 (Shrestha, 2021).

Based on the data in Table 3, the anti-image correlation value for items 2 – 12 has a value greater than 0.5, indicating that this item has variables that contribute significantly to the main factor. However, in other research it was found that the anti-image correlation value is also acceptable if it is greater than 0.4 (Suyanta et al., 2020).

Table 3. Anti-Image Correlation Test Result

Item	Anti-Image Correlation Value	Decision
1	0.466	Use
2	0.676	Use
3	0.607	Use
4	0.822	Use
5	0.747	Use
6	0.681	Use
7	0.659	Use
8	0.801	Use
9	0.646	Use
10	0.624	Use
11	0.698	Use
12	0.720	Use

Because the results of the KMO and Bartlett tests as well as anti-image correlation showed suitable results, next factor analysis was carried out for the unidimensional assumption. Factor analysis is used to identify underlying factors or traits that explain patterns or correlations among observed variables (Shrestha, 2021). The unidimensional assumption can be analyzed using the SPSS 16.0 program by looking at the Eigen values. Table 4 below shows the Eigen values for items 1 – 12.

Table 4. Exploratory Factor Analysis Test Result

Component	Total	% Variance	Cumulative %
1	2.994	24.947	24.947
2	1.881	15.673	40.620
3	1.181	9.845	50.465
4	1.049	8.743	59.208
5	0.911	7.591	66.799
6	0.783	6.521	73.320
7	0.751	6.262	79.582
8	0.727	6.059	85.641
9	0.523	4.356	89.997
10	0.470	3.918	93.915
11	0.375	3.123	97.037
12	0.356	2.963	100.000

The results of factor analysis can also be seen more clearly in the scree plot which shows the Eigen value graph to identify bent points or breaking points where the curve becomes flat. The number of points above the curve that is flatter indicates the number of factors that need to be maintained (Costello & Osborne, 2005). The graph in Figure 1 shows that there are 4 points before the flat curve (Eigen value > 1), meaning that there are 4 factors formed.

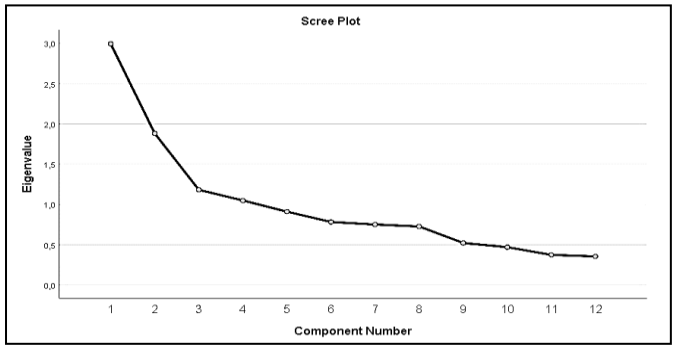

Figure 2. Scree Plot

Unidimensional test can also obtain from the results of the analysis of the Winstep 3.73 program in Table 5 which is shown from the raw variance value (Eigen Value) with a minimum requirement of 20% (Brentari & Golia, 2007); Sumintono & Widhiarso, 2013). Based on the following analysis results table, the Eigen value of the instrument during the trial was 46.7%. This means that this instrument meets the unidimensional assumption. The instrument contains only one dominant component. If an item does not contain one dominant component, then the test cannot accurately measure the component being measured (Meijer & Tendeiro, 2017).

Table 5. Eigen Value from Winstep Program

Standardize Residual variance (in Eigenvalue units)	Empirical		Modeled
Total raw variance in observations	43.2	100%	100.0%
Raw variance explained by measures	20.2	46.7%	46.8%
Raw variance explained by persons	9.6	22.3%	22.3%
Raw variance explained by items	10.6	24.5%	24.5%
Raw unexplained variance (total)	23.0	53.3%	100% 53.2%
Unexplained variance in 1st contrast	2.6	6.1%	11.4%
Unexplained variance in 2nd contrast	2.1	4.9%	9.2%
Unexplained variance in 3rd contrast	2.0	4.5%	8.5%
Unexplained variance in 4th contrast	1.5	3.6%	6.7%
Unexplained variance in 5th contrast	1.5	3.5%	6.5%

Local Independence Assumption

The local independence assumption test states that the possibility of a correct answer to a question item does not depend on the responses to other question items, depending on the participant's ability (Monseur et al., 2011). Based on Table 6, every value below the diagonal line in the matrix is zero. This shows that the assumption of local independence in this instrument is met. This means that the responses given by participants are considered independent compared to others or the students' skills in answering questions do not influence their ability to answer other questions.

Table 6. Local Independence Assumption Test Result

	k1	k2	k3	k4	k5	k6	k7	k8	k9	k10
k1	0,06826									
k2	0,02775	0,01221								
k3	0,01526	0,00669	0,00409							
k4	0,01376	0,00599	0,00335	0,00341						
k5	0,01241	0,00547	0,00326	0,00289	0,00385					
k6	0,01462	0,00640	0,00381	0,00344	0,00394	0,00471				
k7	0,01332	0,00553	0,00347	0,00293	0,00366	0,00378	0,00410			
k8	0,01845	0,00817	0,00454	0,00405	0,00408	0,00442	0,00415	0,00684		
k9	0,02607	0,01081	0,00647	0,00554	0,00515	0,00607	0,00577	0,00693	0,01146	
k10	0,11173	0,03793	0,02084	0,02309	0,02494	0,02863	0,02801	0,01932	0,03833	0,48536

Invariance Assumption of Item Parameter

To test the assumption of parameter invariance, researchers can compare items using different groups (Nguyen et al., 2014). In this case the groups are divided into odd and even groups. Ideally, the equation of the line formed is a straight line with a slope (slope) is equal to 1 and the intercept is at 0. This indicates that the item parameters (difficulty level, discrimination power, and guessing) remain constant across different groups or conditions. The linear equation shows the relationship between item parameters and latent traits (ability) from participants (Nguyen et al., 2014). The results of the test for the invariance assumption of item parameters can be seen in Figure 2 below.

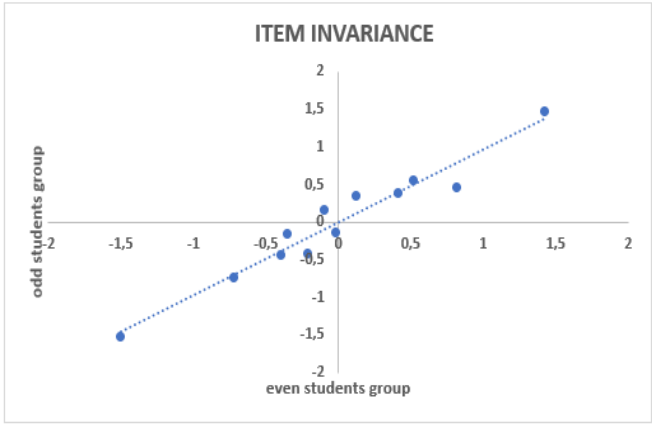

Figure 3. Item Invariance Assumption Test Result

Based on the picture, the two groups are correlated, as evidenced by the number of points that approach the linear line, so it can be said that the assumption of the question item parameters is met. This means that even though the instrument product is tested on different students, the characteristics of the instrument items will not change.

Test of the assumption of invariance of ability parameters is carried out by estimating ability parameters for different groups of questions. On the graph, a straight-line equation will be formed with a slope equal to 1, which shows that the probability of a correct response increases linearly along with the ability parameter (Galdin & Laurencelle, 2010; DeMars, 2010). The results of the ability parameter invariance assumption test can be seen in Figure 3.

Based on the data, it appears that the two groups of items are correlated. This can be seen from the trendline which forms a straight-line equation with a slope of 1.

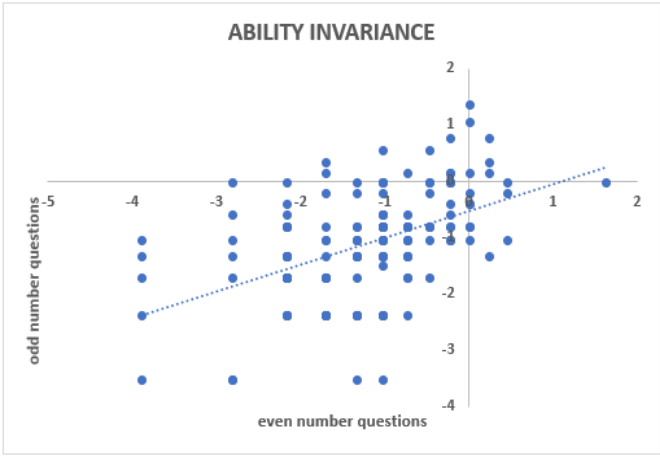

Figure 4. Ability Invariance Assumption Test Result

The trendline in the ability invariance graph shows the relationship between an individual's ability and the possibility of the individual's correct response to a particular item. Thus, the three assumption tests have been met, so that the estimation of the item parameters can be carried out using Rasch modeling.

Analysis of Question Item Parameters
Item Fit

This analysis is seen from the value outfit mean square (MNSQ), outfit z-standard (ZSTD), and PT-measure correlation (Fuady et al., 2023).

Based on the empirical test results (Table 7), the average MNSQ value was obtained he enters 1.01 and a standard deviation of 0.28 during testing, which is within the accepted criteria. Based on these values, items 1 (question number 1 package A), 3 (question number 3 package A), and 22 (question number 10 package B) do not meet the criteria outfit MNSQ and outfit ZSTD. Item 22 also does not meet all criteria. Therefore, items 1, 3, and 22 should be deleted from the question instrument.

Table 7. Item Fit Result

Item	Question Number	Criteria			Decision
		Outfit MNSQ	Outfit ZSTD	PT-Measur Cor	
B1	1 paket A	1.68	2.9	0.43	Rejected
B2	2 paket A	0.74	-1.5	0.62	Accepted
B3	3 paket A	1.63	3.0	0.47	Rejected
B4	4 paket A	0.52	-2.3	0.73	Accepted
B5	5 paket A	0.92	-0.3	0.61	Accepted
B6	6 paket A	0.93	-0.2	0.67	Accepted
B7	7 paket A	0.97	0.0	0.50	Accepted
B8	8 paket A	0.65	-1.9	0.73	Accepted
B9	9 paket A	1.00	0.1	0.57	Accepted
B10	10 paket A	0.98	-0.1	0.64	Accepted
B11	11 paket A	0.91	-0.1	0.52	Accepted
B12	12*	1.12	0.7	0.53	Accepted
B13	1 paket B	1.35	1.8	0.37	Accepted
B14	2 paket B	0.77	-1.7	0.41	Accepted
B15	3 paket B	1.09	0.5	0.65	Accepted
B16	4 paket B	0.90	-0.6	0.45	Accepted
B17	5 paket B	0.66	-1.8	0.55	Accepted
B18	6 paket B	0.67	-2.5	0.57	Accepted
B19	7 paket B	0.99	0.0	0.48	Accepted
B20	8 paket B	1.17	1.2	0.51	Accepted
B21	9 paket B	1.09	0.5	0.39	Accepted
B22	10 paket B	1.83	3.7	0.12	Rejected
B23	11 paket B	1.25	0.6	0.19	Accepted

Item Validity

The validity of each item can be determined based on the results of program analysis Parscale based on the graph on Item Information Curve (IIC). A higher information value indicates higher psychometric quality which contributes to the validity of the assessment instrument (Kalkbrenner, 2021; Nguyen et al., 2014; (Yang & Kao, 2014).

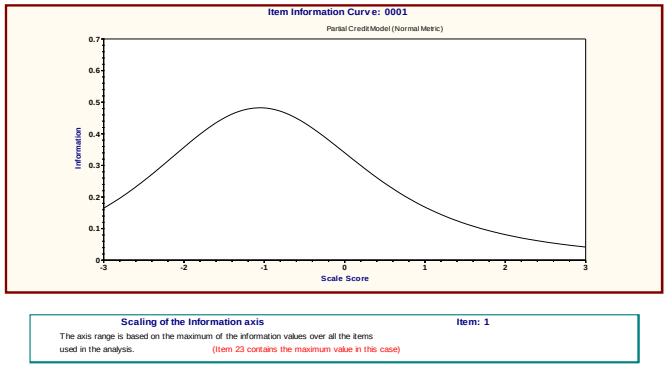

Figure 5. IIC Item 1

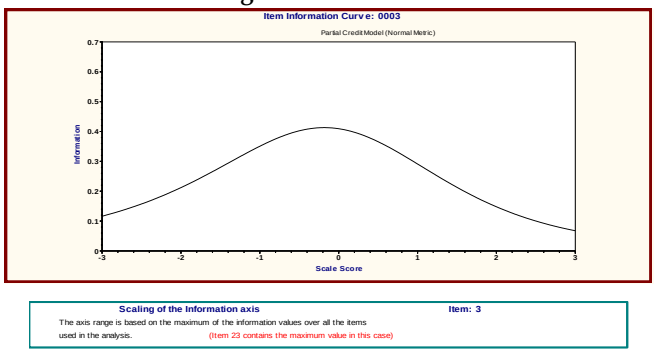

Figure 6. IIC Item 3

Figure 4 and Figure 5 are IIC for items 1 and 3 (questions number 1 and 3, package A). The highest peak is in the area of participants with abilities around -1 (below average). A steeper curve indicates that the item provides more information across the range of the latent trait (Wood, 2017; Baker & Kim, 2017). More informative and precise items better measure underlying constructs and contribute to the validity of assessment instruments (Yang & Kao, 2014; (Thorpe & Favia, 2012). Therefore, the researcher still maintains these two items because these questions are said to be valid for measuring students' abilities with these abilities.

Point 22 (question number 10 package b) shown in Figure 6, has a Gaussian IIC graph which is also good, namely it can explore the abilities of participants in the medium ability category (ability around +1). Thus, this point is also retained.

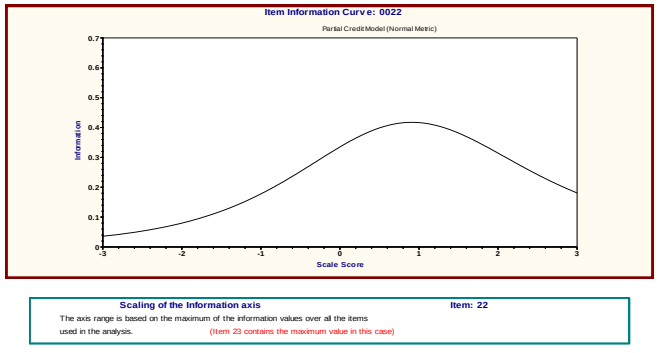

Figure 7. ICC Item 6

There are several reasons why the results of the Winstep program analysis are different from the results of the Parscale program analysis, such as differences in the IRT model, item parameters, data and analysis methods, evaluation criteria, operating system, software version, and others (McCowan & McCowan, 1999). IIC on Parscale program analysis results can provide a more comprehensive view of item performance even if it is not fit against the Rasch model.

Point 11 (question number 11 package a) and point 23 (question number 11 package b) are not good to use. The peak of the graph in Figure 8 is when students' abilities are relatively high. This is in accordance with the level of difficulty of the questions which can be seen from output phase 2 (PH2) with value location respectively +2.784 and +2.795 which indicates this question is too difficult.

Item 23 has the same question indicators as item 11, thus item 11 and item 23 were decided to be eliminated. The validator teachers also stated that material about redox reaction experiments is never touched on in learning because it is also included in the chemistry learning syllabus in 10th grade. Based on the results of this analysis, it is too difficult for students to gain understanding through the discourse and situations provided related to the knowledge possessed to prepare an experimental design.

Item 23 (question number 11 package B) has the highest peak among the other items, but the peak is in the area where the participant's ability is close to 3. This means that this item can provide more information, but only for students with high ability. Likewise, for point 11. Thus, there are total of 21 items that are acceptable and included in this instrument.

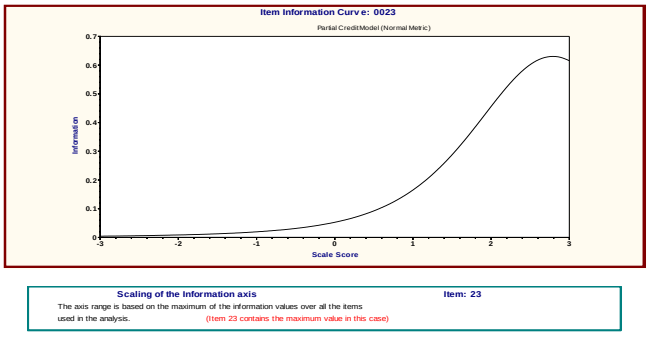

Figure 8. IIC Item 23

Reliability

Based on analysis using the program Winstep, the item reliability value is obtained, reliability person, and Cronbach alpha as stated in Table 8.

Table 8. Reliability Test Result

Reliability	Value	Category
Item Reliability	0.96	High
Person Reliability	0.72	Medium

On the results of program analysis Parscale, the reliability of the instrument in instrument trials can be seen in the menu Total Information. In IRT analysis, the value of information is reliability while standard error Also known as measurement error. Reliability is inversely proportional to standard error, the greater the reliability, standard error getting smaller (Pratama et al., 2020).

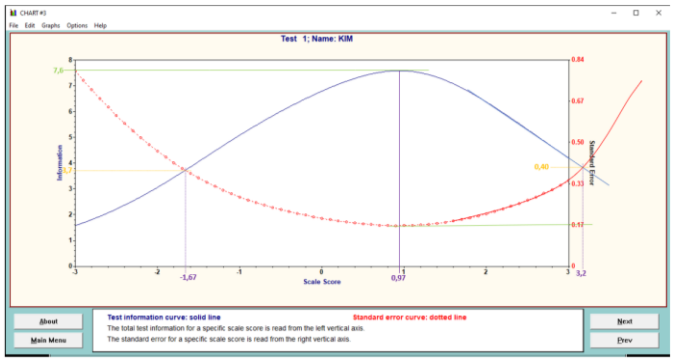

Figure 9. Test Information

The information value on the graph based on the curve intersection point is 3.7 with standard error 0.40. Based on this graph, it can be concluded that overall, the items in this question are suitable or reliable for measuring students' cognitive abilities with abilities ranging from -1.67 to +3.2. Apart from that, the question is not reliable because the error is high. The highest information was obtained with a value of 7.6 at an ability of 0.97 or close to 1 and with standards error 0.17, meaning that the questions most reliably measure the abilities of students with moderate abilities.

Difficulty Level

The item difficulty level parameter value has a range between -2.0 to +2.0 as a standard for determining easy to difficult items (Hambleton et al., 1991). The most difficult items based on analysis using the program Parscale is item 11 with value location +2,784 and item 23 with value location +2,795. This is in accordance with the results of the program analysis Winstep, the most difficult question is item 23 with marks measure +2.08 and item 11, but with value measure +1.44. The easiest item is item 13 with value location -1.118 and item 15 with value location -0.914. However, both are still within the criteria values. These results are also in accordance with measure order on Winstep where item 13 is the easiest item with value measure -1.35 and item 15 has value measure -1.25. But point 1 also has value measure low (classified as easy questions) with value measure -1.26 while on program analysis results Parscale mark location item 1 is -0.864. Table 21 shows a summary of item difficulty levels.

Table 9. Item Difficulty Level

Item	Location Value (Parscale)	Difficulty Level	Measure Value (Winstep)	Difficulty Level
1	-0.864	Easy	-1.26	Easy
2	-0.833	Easy	-1.02	Easy
3	-0.427	Medium	-0.83	Medium
4	1.991	Difficult	0.39	Medium
5	1.369	Medium	0.33	Medium
6	0.954	Medium	0.20	Medium
7	1.878	Difficult	0.83	Medium
8	0.886	Medium	-0.07	Medium
9	1.469	Medium	0.35	Medium
10	0.751	Medium	-0.20	Medium
11	2.784	Difficult	1.44	Difficult
12	1.251	Medium	0.59	Medium
13	-1.118	Easy	-1.35	Easy
14	1.044	Medium	-0.45	Medium
15	-0.914	Easy	-1.25	Easy
16	0.782	Medium	-0.17	Medium
17	1.660	Medium	0.39	Medium
18	0.465	Medium	-0.35	Medium
19	1.181	Medium	0.38	Medium
20	0.243	Medium	-0.39	Medium
21	1.192	Medium	0.19	Medium
22	1.203	Medium	0.16	Medium
23	2.795	Difficult	2.08	Difficult

Thus, questions in the medium category have more composition. This proves that HOTS questions are not always identical with difficult questions. HOTS questions are questions that can assess students' abilities in analyzing, evaluating and creating based on contextual problems and non-routine matters (Widana, 2017).

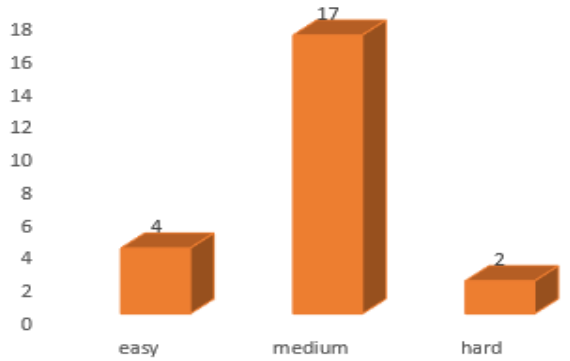

Figure 10. Distribution of Number of Items Based on Difficulty Level

Conclusion

Based on the discussion, it can be concluded that (1) the authentic instrument developed criteria are can assess higher order thinking skills; have coherence of knowledge with skills and various scientific disciplines; developing 21st century competencies; and involve real or everyday life contexts; (2) the characteristic are outfit MNSQ is 0.52 – 1.83; outfit ZSTD -2.3 – 3.7; PT value-Measure Corr is 0.12 – 0.73 and the difficulty level of the questions is relatively good with a range of -1.35 to +2.08 and is dominated by medium category questions; (3) the instrument is declared theoretically and empirically valid; (4) item reliability is in the high category with a value of 0.96 and person reliability is in the medium category with a value of 0.72.

Acknowledgments

Place acknowledgments, including information on grants received, before the references, in a separate section, and not as a footnote on the title page

Author Contributions

Conceptualization, Q.A. and I.S.Y.L.; methodology, Q.A. and I.S.Y.L.; validation, I.S.Y.L.; formal analysis, Q.A.; investigation, I.S.Y.L.; resources, Q.A.; data curation, Q.A.; writing—original draft preparation, Q.A.; writing—review and editing, Q.A. and I.S.Y.L.; visualization, Q.A.; supervision, I.S.Y.L.; project administration, Q.A. All authors have read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

References

Adu-Gyamfi, K., & Ampiah, J. G. (2019). Students' alternative conceptions associated with application of redox reactions in everyday life. *Asian Education Studies*, *4*(1), 29. https://doi.org/10.20849/aes.v4i1.590

Ambiyar, Efendi, R., Irawati, Y., Waskito, & Suryadimal. (2020). Effectiveness e-authentic assessment in computer network course. *Journal of Physics: Conference Series*, *1481*(1). https://doi.org/10.1088/1742-6596/1481/1/012131

Baker, F. B., & Kim, S.-H. (2017). *The Basics of Item Response Theory Using R*. Springer. https://doi.org/10.1007/978-3-319-54205-8

Bartlett, M. S. (1951). The effect of standardization on a χ2 approximation in factor analysis. In *Source: Biometrika* (Vol. 38, Issue 3). https://doi.org/10.2307/2332580

Brentari, E., & Golia, S. (2007). Unidimensionality in the Rasch Model: How to detect and interpret. *Statistica*, 253–261. https://doi.org/https://doi.org/10.6092/issn.1973-2201/3508

Bushkofsky, N. (2016). Developing authentic summative assessments that correlate to the Next Generation Science Standards for a middle school science classroom. *Graduate Research Papers*. 110. https://scholarworks.uni.edu/grp/11.

Chiang, W.-W., Chiu, M.-H., Chung, S.-L., & Liu, C.-K. (2014). Survey of High School Students' Understanding of Oxidation-Reduction Reaction.

Journal of Baltic Science Education. http://oaji.net/articles/2015/987-1450980905.pdf

Cohen, R. Jay., & Swerdlik, M. E. . (2018). *Psychological Testing and Assessment : an introduction to tests and measurement* (9th ed.). McGraw-Hill Education.

Costello, A. B., & Osborne, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research & Evaluation*, *10*. https://doi.org/10.7275/jyj1-4868

DeMars, C. (2010). *Item Response Theory: Understanding Statistic Measurement* (N. Beretvas & P. Leavy, Eds.). Oxford University Press.

Ekawati, D. (2012). The implementation of authentic assessment in Vocational High School 1 Kuala Cenaku. *1st English Language and Literature International Conference*. https://jurnal.unimus.ac.id/index.php/ELLIC/article/view/2415

Fuady, I., Priyansyah, R. N., Ernawati, E., & Prasanti, D. (2023). Validity and Reliability Tests on the Nomophobia Instrument with the Rasch Model. *Indigenous: Jurnal Ilmiah Psikologi*, *7*(3), 276–287. https://doi.org/10.23917/indigenous.v7i3.19152

Galdin, M., & Laurencelle, L. (2010). Assessing parameter invariance in item response theory's logistic two item parameter model: A Monte Carlo investigation. *Tutorials in Quantitative Methods for Psychology*, *6*(2), 39–51. https://doi.org/10.20982/tqmp.06.2.p039

Gulikers, J. T. M., Bastiaens, T. J., & Kirschner, P. A. (2004). A five-dimensional framework for authentic assessment. *Educational Technology Research and Development*, *52*(3), 67–87. https://doi.org/10.1007/BF02504676

Hadinugrahaningsih, T., Rahmawat, Y., & Suryani, E. (2022). An analysis of preservice chemistry teachers' misconceptions of reduction-oxidation reaction concepts. *Journal of Technology and Science Education*, *12*(2), 448–465. https://doi.org/10.3926/jotse.1566

Hambleton, R. K. ., Swaminathan, Hariharan., & Rogers, H. Jane. (1991). *Fundamentals of Item Response Theory* (Vol. 2). SAGE Publications, Inc.

Hancock, L. M., & Hollamby, M. J. (2020). Assessing the practical skills of undergraduates: The evolution of a station-based practical exam. *Journal of Chemical Education*, *97*(4), 972–979. https://doi.org/10.1021/acs.jchemed.9b00733

Hanifah, M., & Irambona, A. (2019). Authentic assessment: Evaluation and its application in science learning. *Psychology, Evaluation, and Technology in Educational Research*, *1*(2), 81. https://doi.org/10.33292/petier.v1i2.4

Hasan, R., & Cerimagic, S. (2021). Authentic assessment during Covid-19: an Australian postgraduate computing degree program example. *Journal of Learning Development in Higher Education*, *22*. https://doi.org/10.47408/jldhe.vi22.690

Iskandar, D., & Senam, S. (2015). Studi kemampuan guru kimia SMA lulusan UNY dalam mengembangkan soal UAS berbasis HOTS. *Jurnal Inovasi Pendidikan IPA*, *1*(1), 65–72. https://doi.org/10.21831/jipi.v1i1.4533

Ismiati, I., Nahadi, N., & Halimatul, H. S. (2019). Analysis of the need to development an authentic assessment instrument on buffer material. *Journal of Physics: Conference Series*, *1157*(4). https://doi.org/10.1088/1742-6596/1157/4/042044

Kalkbrenner, M. T. (2021). A practical guide to instrument development and score validation in the social science: The MEASURE approach. *Practical Assessment, Research, and Evaluation*, *26*, 1–18. https://doi.org/10.7275/svg4-e671

Kartowagiran, B., & Jaedun, A. (2016). Model asesmen autentik untuk menilai hasil belajar siswa Sekolah Menengah Pertama (SMP): Implementasi asesmen autentik di SMP. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *20*(2), 131–141. https://doi.org/10.21831/pep.v20i2.10063

Lau, P. N., Chua, Y. T., Teow, Y., & Xue, X. (2020). Implementing alternative assessment strategies in chemistry amidst COVID-19: Tensions and reflections. *Education Sciences*, *10*(11), 1–15. https://doi.org/10.3390/educsci10110323

McCowan, R. J., & McCowan, S. C. (1999). *Item Analysis for Criterion-Referenced Tests*. Center for Development of Human Services. http://www.bsc-cdhs.org

Meijer, R. R., & Tendeiro, J. N. (2017). Unidimensional item response theory. In *The Wiley Handbook of Psychometric Testing: A Multidisciplinary Reference on Survey, Scale and Test Development* (Vols. 1–2, pp. 413–443). Wiley Blackwell. https://doi.org/10.1002/9781118489772.ch15

Monseur, C., Baye, A., Lafontaine, D., & Quittre, V. (2011). PISA test format assessment and the local independence assumption. *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments*, *4*. https://orbi.uliege.be/bitstream/2268/103137/1/IERI_Monograph_Volume04_Chapter_6.pdf

Nguyen, T. H., Han, H. R., Kim, M. T., & Chan, K. S. (2014). An introduction to item response theory for patient-reported outcome measurement. In *Patient* (Vol. 7, Issue 1, pp. 23–35). Adis International Ltd. https://doi.org/10.1007/s40271-013-0041-0

Nuriana, D. (2018). Kendala guru dalam memberikan penilaian sikap siswa pada proses pembelajaran Berdasarkan Kurikulum 2013. *Madrosatuna: Journal of Islamic Elementary School*, *2*(2), 51–62. https://doi.org/10.21070/madrosatuna.v2i2.1970

Nurmawati, N., Driana, E., & Ernawati, E. (2021). Pemahaman guru kimia sekolah menengah atas tentang penilaian kemampuan berpikir tingkat tinggi dan implementasinya. *Edusains*, *12*(2), 233–242. https://doi.org/10.15408/es.v12i2.13613

Oriondo, L. L., & Dallo-Antonio, E. M. (1984). *Evaluating Educational Outcomes (Tests, Measurement and Evaluation)*. Rex Book Store.

Pratama, M. A., Supahar, Lestari, D. P., Sari, W. K., Putri, T. S. Y., & Adiatmah, V. A. K. (2020). Data literacy assessment instrument for preparing 21 Cs literacy: Preliminary study. *Journal of Physics: Conference Series*, *1440*(1). https://doi.org/10.1088/1742-6596/1440/1/012085

Salirawati, D. (2021). Authentic assessment in the pandemic period. *Journal of The Indonesian Society of Integrated Chemistry*, *13*(1), 21–31. https://doi.org/10.22437/jisic.v13i1.11716

Shrestha, N. (2021). Factor analysis as a tool for survey analysis. *American Journal of Applied Mathematics and Statistics*, *9*(1), 4–11. https://doi.org/10.12691/ajams-9-1-2

Stowe, R. L., & Cooper, M. M. (2019). Assessment in chemistry education. *Israel Journal of Chemistry*, *59*(6), 598–607. https://doi.org/10.1002/ijch.201900024

Suastra, W., & Ristiati, N. P. (2017). Problems faced by teachers in designing and implementing authentic Assessment in science teaching. *International Research Journal of Engineering*, *3*, 27–36. https://doi.org/10.21744/irjeis.v3i4.496

Sudiana, I. K., Sastrawidana, I. D. K., & Antari, N. P. S. (2018). Kendala guru dalam penyelenggaraan penilaian sikap. *Jurnal Pendidikan Kimia Undiksha*, *2*(2), 9–74. https://doi.org/10.23887/jjpk.v2i2.21169

Sumintono, B., & Widhiarso, W. (2013). *Aplikasi Model Rasch untuk Penelitian Ilmu-Ilmu Sosial* (B. Trim, Ed.; Revisi). Trim Komunikata Publishing House.

Sund, P. (2016). Science teachers' mission impossible?: a qualitative study of obstacles in assessing students' practical abilities. *International Journal of Science Education*, *38*(14), 2220–2238. https://doi.org/10.1080/09500693.2016.1232500

Sutadji, E., Susilo, H., Wibawa, A. P., Jabari, N. A. M., & Rohmad, S. N. (2021). Adaptation strategy of authentic assessment in online learning during the covid-19 pandemic. *Journal of Physics: Conference Series*, *1810*(1). https://doi.org/10.1088/1742-6596/1810/1/012059

Suyanta, S., Muharram, M., Mulbar, U., Rauf, B., Agung, M., Ganefri, G., Ponto, H., Sila, I. N., Wahid, A., Parenreng, J. M., Yasdin, Y., Astuti, S. R. D., Puspita Sari, A. R., & Tyas, R. A. (2020). Educational LPTK, non-educational LPTK, and non-LPTK students' intention to become teacher. *Universal Journal of Educational Research*, *8*(12), 6676–6683. https://doi.org/10.13189/ujer.2020.081232

Thorpe, G. L., & Favia, A. (2012). Data Analysis Using Item Response Theory Methodology: An Introduction to Selected Programs and Applications. *Psychology Faculty Scholarship, 20,* 1-33 http://digitalcommons.library.umaine.edu/psy_facpub/20

Villarroel, V., Bloxham, S., Bruna, D., Bruna, C., & Herrera-Seda, C. (2018). Authentic assessment: creating a blueprint for course design. *Assessment and Evaluation in Higher Education*, *43*(5), 840–854. https://doi.org/10.1080/02602938.2017.1412396

Wahyuni, L. G. E., Dewi, N. L. P. E. S., & Paramartha, A. A. G. Y. (2021). Authentic Assessment Practice Teachers' Perceived Knowledge. *Advances in Social Science, Education and Humanities Research*, *540*, 316–323. https://doi.org/10.2991/assehr.k.210407.258

Widana, I. W. (2017). Higher order thinking skills assessment (HOTS). *Journal of Indonesian Student Assessment and Evaluation*, *3*(1), 32–44. https://journal.unj.ac.id/unj/index.php/jisae/article/view/4859/3601

Wood, J. (2017, November 12). *Logistic IRT Models*. QuantDev Penn State University. https://quantdev.ssri.psu.edu/sites/qdev/files/IRT_tutorial_FA17_2.html

Wulandari, A. D., Pramana Situmorang, R., & Dewi, L. (2018). Evaluasi pelaksanaan penilaian autentik pada pembelajaran IPA terhadap hasil belajar peserta didik kelas VIII SMP Negeri 3 Salatiga. *Jurnal Pendidikan Sains (JPS)*, *06*(01), 34–46. http://jurnal.unimus.ac.id/index.php/JPKIMIA

Yang, F. M., & Kao, S. T. (2014). Item response theory for measurement validity. *Shanghai Arvhives of Psychiatry*, *26*(3). https://doi.org/10.3969/j.issn.1002-0829.2014.03.010