JRERE 11 (1) 2022 1-8

Journal of Research and Educational Research Evaluation

http://journal.unnes.ac.id/sju/index.php/jere

Validity and Reliability of Chemical Test Instruments for Acid and Base Solutions Oriented Generic Skills Science

Muhammad Riza[1]✉, Kartono Kartono[2], Endang Susilaningsih[3]

[1,2,3]Pascasarjana Universitas Negeri Semarang, Indonesia

Article Info

History Articles
Received:
15 December 2021
Accepted:
11 January 2022
Published:
30 March 2022

Keywords:
Evaluation, science generic skills, validity, reliability

Abstract

The 2013 curriculum prioritizes the generic skills needed in science learning. Generic science skills require valid and reliable measuring tools. The purpose of this research is to reveal validity and reliability Chemical test instruments oriented to generic science skills on acid and base solutions. The subjects used were 35 students for the small class test and 125 students for the large class test. The method used is the quantitative method in the form of questionnaire sheets and KGS-oriented acid-base chemistry questions. Construct validity using confirmatory first order analysis. Interterrater reliability using 3 raters and tested using two-way annova with Ebel formula. The reliability of the large class test used the cronbach alpha formula. content validity of each item in each 81.25% eligible category. Construct validity seen from the item polarity of the items has a positive Point Measure Correlation (Pt. Mea-Corr). A total of 32 items have a strong or high correlation number. Three questions, namely item number 23 (0.57), 27 (0.49), and 33 (0.67) have a moderate correlation. Reliability between expert raters was analyzed using the Ebel formula resulting in 0.79 which was in the currently category. Small class test reliability 0.97 high category. The reliability of the large class test item reliability is 0.90, personal reliability is 0.94, and the KR-20 is 0.99 which is classified as high. Thus, the level of generic science ability that has been tested for validity and Reliability can be used by educators to determine the level of students' generic science abilities.

✉Correspondence Address :
Educational Research and Evaluation,
Pascasarjana, Universitas Negeri Semarang,
Indonesia 50229
E-mail : muhammadrizakhoirullah@students.unnes.ac.id

p-ISSN 2252-6420
e-ISSN 2503-1732

INTRODUCTION

Assessment is a process of collecting and processing information with the aim of measuring the achievement of learning outcomes (Svenningsson et al., 2018). The teacher conducts assessment and evaluation activities with the aim of knowing the progress and learning outcomes of students, diagnosing learning difficulties, providing feedback for improving the teaching and learning process, and determining grade increases (Brassil & Couch, 2019). Assessment and evaluation can obtain information about the implementation of learning and the learning success of students, teachers, and the learning process (Polizzi et al., 2021). Assessment and evaluation information makes decisions about learning, student difficulties, guidance efforts if needed and the existence of a curriculum. Assessment function for students to identify the level of learning success. Assessment function for teachers to identify success in teaching (Seery et al., 2019).

Chemistry as a part of science, chemistry is a science based on experiments whose development and application demands high standards of experimental work. Chemistry experiments or practicums help students gain technical skills (Hernandez et al., 2021). Students participate actively and are trained to develop their scientific attitude in the implementation of practicum. Assessment in chemistry learning can measure all aspects of the output of the learning process. The implementation of the assessment so far is still traditional in that it only measures the knowledge of students, while in the implementation of the 2013 curriculum the assessment measures the knowledge, attitudes and skills of students (Hasan Ashari et al., 2016).

Generic science skills are basic skills contained in students and need to be developed by teachers (Pedaste et al., 2021). Generic science skills are skills that train students how to think and solve problems in chemistry that are adapted to the development of students. Generic science skills are skills used to learn various concepts and solve problems in science (Christian et al., 2021). The chemistry learning model does not only emphasize the mastery of chemical concepts, but also emphasizes thinking skills, communicating the process and results of learning chemistry in schools, as well as generic science skills to be applied in solving everyday life problems. Generic science skills are the ability to think and act that students have based on their scientific knowledge.

The assessment of generic science skills can be done through tests and through practicum. In the test, questions are given with multiple choice and essay types with integrated science generic skills. The assessment model with practicum can be taken from some of the generic science skills in the applied practicum. Basic chemistry practicum on redox and electrochemistry practicum, chemical kinetics practicum and introduction of functional groups takes generic science skills in the form of direct observation skills, symbolic language, and logical inference applied at the student level. Organic chemistry practicums in distillation, solubility and recrystallization tests, compositions and chains of hydrocarbons, alcohols-phenols, aldehydes and ketones take generic science skills in the form of observation and logical inference applied at the student level.

Evaluation tools in the form of tests can be used to determine the improvement of learning outcomes so that the quality can be known (Werno Sujito et al., 2015). Teachers as educators must be able to make evaluation tools so that they can give instructions that the task is successful or not (Kang et al., 2019). Evaluation can improve students' understanding. The role of evaluation is important to provide information about the learning outcomes that have been owned by students (Barlow et al., 2020). The information obtained by the teacher determines whether the goals that have been set have been achieved or not (Polizzi et al., 2021).

The teacher arranges evaluation tools according to good and correct rules so that student learning outcomes truly reflect the actual results (Seery et al., 2019). The low value of student learning outcomes is not always caused by the unpreparedness of students in answering questions, but can also be caused by question items that do not measure the material that they actually want to measure (Tiruneh et al., 2017). The evaluation tool compiled is in the form of an instrument in the form of a test. The material mastery test is important to determine the ability of students and to measure the success of students' learning and to measure the success of students' learning (Tzivinikou et al., 2021). The test questions used are in the form of objective and essay tests.

The test items should be arranged according to the standards of the test compiler (Pedaste et al., 2021). Compiling tests requires knowledge, skills and high accuracy (Yang et al., 2021). The test used by the testee needs to know whether the test is of good or poor quality so it is necessary to do a test analysis. Aims to help improve the test through revision in finding out whether students have mastered the material taught by the teacher (Jescovitch et al., 2021). The method developed to analyze items can use the classical test theory approach and modern test theory (Brassil & Couch, 2019). The classical approach uses classical test theory, while the modern approach uses item response theory (IRT).

The purpose of this research is to reveal validity and reliability Chemical test instruments oriented to generic science skills on acid and base solutions. Test analysis in this study used the Rasch/Item Response Item 1 parameter model. The one-parameter model or Rasch model has several advantages, namely being able to identify response errors, being able to predict missing data scores, being able to distinguish the ability of respondents with the same number of raw scores, and being able to identify indications of guesswork and cheats (Rachmatullah & Ha, 2019). This advantage makes the Rasch model more accurate. Rasch modeling is able to produce standard error measurement values so that it can increase the accuracy of calculations (Hofer et al., 2017).

METHODS

The research was conducted in two places, the first at SMK Al Furqon and the second at SMK Ky Ageng Giri Kusuma. The small class test at SMK Al Furqon used 35 X TKRO subjects and the large class test at SMK Ky Ageng Giri Kusuma used 125 X TKRO subjects. The method used is the quantitative method in the form of questionnaire sheets and KGS-oriented acid-base chemistry questions. Students are given material on acid and base solutions that are oriented with KGS. Content validity was assessed by 1 KGS expert, 1 chemistry expert and 1 chemistry teacher at SMK Ky Ageng Giri Kusuma. Content validity uses Arikunto criteria. Construct validity using confirmatory first order analysis. Interterrater reliability using 3 raters and tested using two way annova with Ebel formula. The reliability of the small class trial used the Spearman Brown formula and the large class test used the Rasch alpha cronbach formula for reliability as well as item and person characteristics (Taber, 2018).

RESULTS AND DISCUSSION

Content Validity

Content validity was assessed by several experts. The content validity test was carried out by 3 experts to see the suitability of the material, construction, language and conformity with KGS. The experts also filled out a questionnaire containing the conclusions of the experts' assessment of the KGS-oriented chemistry questions. Quantitative data that presents the validation criteria of the experts are presented in Table 1.

Table 1. Expert Validation Criteria

No Item	Percentage	Criteria	Conclusion
1. 4. 7. 9. 11. 13. 15. 16. 17. 18. 19. 20. 21. 22. 23. 24. 25. 26. 27. 28. 30. 31. 32. 33. 34. 35. 36. 38. 40	83 - 89	Very good	Worthy
2. 3. 5. 6. 8. 10. 12. 14. 29. 37. 39.	90 - 96	Very good	Worthy

Based on the results of data analysis in Table 1, information can be obtained that a number of 40 items are valid and there are no revisions.

Construct Validity

A total of 32 items have a strong or high correlation number. Three questions, namely item number 23 (0.57), 27 (0.49), and 33 (0.67) have a moderate correlation. This is in accordance with the opinion of Norasmah, Salleh & Hussein (2014, p.117) that Pt. A high Mea Corr (0.68-1.00) indicates an item can distinguish between respondents' abilities.

Table 2. Unidimensionality Test

Unidimensionality Test	Empirical
Total raw variance is observation	100
Raw variance explained by measures	77.9
Raw variance explained by persons	65.3
Raw variance explained by items	12.6
Raw explaned variance (total)	22.1
Unexplained variance in 1st contrast	5.2
Unexplained variance in 2nd contrast	2.6
Unexplained variance in 3nd contrast	2.0
Unexplained variance in 4nd contrast	1.8
Unexplained variance in 5nd contrast	1.2

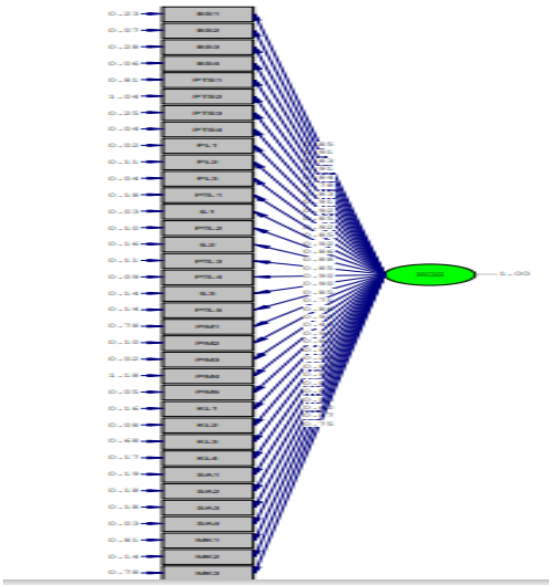
The results of the correlation figures at Pt. The Mea Corr is strengthened on the results of the unidimensionality test through the unidimensionality output table. The unidimensionality table output is presented in Table 2.

Looking at the raw variance in Table 1 shows a figure that is close to high, namely 77.9%. The results of the analysis have a requirement for unidimensionality of more than 60%, indicating that it is special, meaning that the developed instrument is able to measure what it should measure (Ludwig et al., 2021). The unexplained variance values are 5.2, respectively; 2.6; 2.0; 1.8; 1.2. this shows that the variances that cannot be explained by the instrument are all below 10%. Unexplained variance below 10% indicates that the unidimensionality in the instrument is in the good category.

The construct validity test at Rasch is only for the responses of the tested items, while to determine the covariance between the test items, a CFA model is needed with the Lisrel or Amos or SPSS programs (Eser et al., 2019). With regard to defining a model for a data set, the procedure for CFA appears to be more advanced and simpler than that developed for Rasch (IRT). The CFA model can calculate an accurate estimate of the chi-square measure of model fit and related degrees. The construct validity test was strengthened by the Lisrel program (Foster et al., 2019).

Conceptually, the chemical questions developed were constructed on 9 KGS indicators, namely Symbolic Language (BS), Understanding of Scale (PTS), Direct Observation (PL), Indirect Observation (PTL), Logical Inference (IL), Mathematical Modeling (PM), Logical Framework (KL), Cause and Effect (SA), and Building Concepts (MK). One set of questions consisting of 35 questions. Nine indicators were tested at once because the CFA logic was theorized that each item only measures one factor (Pumptow & Brahm, 2021). A variable is said to have good construct validity when the goodness of fit and

the measurement model fit are met. The goodness of fit and measurement model fit values shown in the loading factor are presented on a path diagram. The path diagram for this CFA test is presented in Figure 1.



Chi-Square=5173.68, df=560, P-value=0.00000, RMSEA=0.258

Goodness of Fit Statistics

Degrees of Freedom = 560
Minimum Fit Function Chi-Square = 4309.08 (P = 0.0)
Normal Theory Weighted Least Squares Chi-Square = 5173.68 (P = 0.0)
Estimated Non-centrality Parameter (NCP) = 4613.68
90 Percent Confidence Interval for NCP = (4387.10 ; 4847.55)

Minimum Fit Function Value = 34.75
Population Discrepancy Function Value (F0) = 37.21
90 Percent Confidence Interval for F0 = (35.38 ; 39.09)
Root Mean Square Error of Approximation (RMSEA) = 0.26
90 Percent Confidence Interval for RMSEA = (0.25 ; 0.26)
P-Value for Test of Close Fit (RMSEA < 0.05) = 0.00

Expected Cross-Validation Index (ECVI) = 42.85
90 Percent Confidence Interval for ECVI = (41.02 ; 44.74)
ECVI for Saturated Model = 10.16
ECVI for Independence Model = 350.57

Chi-Square for Independence Model with 595 Degrees of Freedom = 43401.06
Independence AIC = 43471.06
Model AIC = 5313.68
Saturated AIC = 1260.00
Independence CAIC = 43605.05
Model CAIC = 5581.66
Saturated CAIC = 3671.84

Normed Fit Index (NFI) = 0.90
Non-Normed Fit Index (NNFI) = 0.91
Parsimony Normed Fit Index (PNFI) = 0.85
Comparative Fit Index (CFI) = 0.91
Incremental Fit Index (IFI) = 0.91
Relative Fit Index (RFI) = 0.89

Critical N (CN) = 19.44

Figure 1. Path Diagram and Goodness of Fit Results for KGS Oriented Questions

Based on Figure 1, it can be seen that the construct instrument used to form a research model, in the confirmatory factor analysis process, has met the goodness of fit criteria that have been determined. The goodness of fit test probability value shows the CFI. value 0.91 (>0.9) and NFI 0.9 (>0.9). RMSEA value 0.258 (cut off value RMSEA < 0.08). Two of the three categories of compatibility tests meet the good-fit test or are in accordance with stating that the support for the fit of the model developed by empirical data is at least seen from the three fit measures that represent the three categories are significant, then the model developed is suitable or in accordance with the data.

Figure 1 shows that the measurement model fit is fulfilled, as evidenced by looking at the factor loading value of each item > 0.3 which indicates that the variable is said to have good validity to the construct.

Reliability Estimation of KGS Oriented Chemistry Test Instruments

The tested reliability is (a) inter rater reliability, (b) small class trial reliability and (c) large class trial reliability. Based on the analysis, the results are obtained in Table 2.

Table 3. Estimated Reliability

Trial Stage	Reliability	N of items
Expert	0.79	40
Small class	0.97	40
Large Class	0.99	35

Based on Table 2 the successive test reliability values are 0.79; 0.97; 0.99. Inter rater reliability (between experts) is categorized as very low because the experts have different backgrounds. Expert 1 is more critical about generic science skills. Expert 2 is more critical of chemistry content and Expert 3 is more critical of writing. This causes the interrater reliability is currently. The reliability of the small class and large class trials is categorized as special (Küçükerdönmez et al., 2021).

Inter rater reliability

Inter rater reliability is calculated after calculating the content validity between 3 validators. The level of agreement between the 3 validators can be explained through the reliability coefficient between raters (appraisers) using a two way annova analysis with the ebel formula. Alpha reliability 0.78 is in the medium category. The results of the ICC reliability of 0.542 are in the medium category.

Small Grade Test Reliability

Reliability using the Spearman-Brown formula is applied to small classes and is searched using the Anates Description application. Coefficient value 0.97. The reliability coefficient value of 0.97 indicates high reliability because it is > 0.7 (Taber, 2018).

Large Grade Test Reliability

The reliability of the large class stage was seen with the help of the Winstep 3.73 program. The reliability of the Rasch model is described by the presence of a separation index. The reported separation indices are item reliability and person reliability, plus the reliability coefficient of Cronbach Alpha KR-20, the three coefficients are 0.90, 0.94 and 0.99, respectively. These three figures show very high reliability. The separation reliability is of high value because the research sample and the grain difficulty level have a wide range and produce small measurement errors. A broad item means that the item has a difficulty level from the easiest to the most difficult. Likewise for the research sample, a wide sample means that the sample has abilities that are spread from the most intelligent to the least intelligent (Jescovitch et al., 2021). Output reliability can be seen in Table 4.

Table 4. Output Reliability Model Raschm

Measured Person			
Infit		Outfit	
0.90	-0.5	1.16	-0.6
Mean			
3.90			
Separation			
0.94			
Person			
Reliability			
Measured Item			
Infit		Outfit	
0.84	-0.9	1.16	-0.5
Mean			
3.07			
Separation			
0.90			
Item			
Reliability			
0.99			
KR-20 Test			
Reliability			

CONCLUSION

The validity of the chemical test instrument oriented to generic science skills which was developed based on an assessment by 3 experts on 40 multiple choice questions with reasoned on four aspects, namely language, structure, material and generic science skills where the four aspects consist of 21 rules, validity based on an assessment 3 Experts analyzed using Arikunto's formula obtained the validity value of each item in each 81.25% category feasible. Construct validity seen from the item polarity of the items has a positive Point Measure Correlation (Pt. Mea-Corr). A total of 32 items have a strong or high correlation number. Three questions, namely item number 23 (0.57), 27 (0.49), and 33 (0.67) have a moderate correlation. The unidimensionality test resulted in the raw variance by measure, namely 77.9%, and the unexplained variance value, respectively, 5.2%;2.6%;2.0%;1.8%; 1.2%. This shows that the variances that cannot be explained by the instrument are all below 10%. Unexplained variance below 10% indicates that the unidimensionality in the instrument is in the good category. Construct validity through the lisrel 8.8 application produces a factor loading value of each item >

0.3 which indicates that the variable is said to have good validity to the construct.

Reliability between expert raters was analyzed using the Ebel formula resulting in 0.125 which was in the low category. Small class test reliability 0.97 high category. The reliability of the large class test item reliability is 0.90, personal reliability is 0.94 and the KR-20 is 0.99 which is classified as high.

REFERENCES

Barlow, A., Brown, S., Lutz, B., Pitterson, N., Hunsu, N., & Adesope, O. (2020). Development of the student course cognitive engagement instrument (SCCEI) for college engineering courses. *International Journal of STEM Education*, *7*(1), 1–20.

Brassil, C. E., & Couch, B. A. (2019). Multiple-true-false questions reveal more thoroughly the complexity of student thinking than multiple-choice questions: a Bayesian item response model comparison. *International Journal of STEM Education*, *6*(1), 1–17.

Christian, K. B., Kelly, A. M., & Bugallo, M. F. (2021). NGSS-based teacher professional development to implement engineering practices in STEM instruction. *International Journal of STEM Education*, *8*(1), 1–18.

Eser, E., Çevik, C., Baydur, H., Güne, S., Esgin, T. A., Eker, E., Gümü, U., & Eser, G. B. (2019). Reliability and validity of the Turkish version of the WHO-5 , in adults and older adults for its use in primary care settings. *Primary Health Care Research & Development*, 0–6.

Foster, E., Lee, C., Imamura, F., Hollidge, S. E., Westgate, K. L., Venables, M. C., Poliakov, I., Rowland, M. K., Osadchiy, T., Bradley, J. C., Simpson, E. L., Adamson, A. J., Olivier, P., Wareham, N., Forouhi, N. G., & Brage, S. (2019).Validity and reliability of an online self-report 24-h dietary recall method (intake24): a doubly labelled water study and repeated-measures analysis. *Juornal of Nutritional Science*, *8*(29), 1–9.

Hasan Ashari, L., Lestari, W., & Hidayah. (2016). Instrumen penilaian unjuk kerja siswa smp kelas viii dengan model peer asssessment berbasis android pada pembelajaran penjasorkes dalam permainan bola voli. *Journal of research and Educational Research Evaluation*

Hernandez, D., Jacomino, G., Swamy, U., Donis, K., & Eddy, S. L. (2021). Measuring supports from learning assistants that promote engagement in active learning: evaluating a novel social support instrument. *International Journal of STEM Education*, *8*(1), 1–17.

Hofer, S. I., Schumacher, R., & Rubin, H. (2017). The test of basic Mechanics Conceptual Understanding (bMCU): using Rasch analysis to develop and evaluate an efficient multiple choice test on Newton's mechanics. *International Journal of STEM Education*, *4*(1), 1–20.

Jescovitch, L. N., Scott, E. E., Cerchiara, J. A., Merrill, J., Urban-Lurain, M., Doherty, J. H., & Haudek, K. C. (2021). Comparison of Machine Learning Performance Using Analytic and Holistic Coding Approaches Across Constructed Response Assessments Aligned to a Science Learning Progression. *Journal of Science Education and Technology*, *30*(2), 150–167.

Kang, J., Keinonen, T., Simon, S., Rannikmäe, M., Soobard, R., & Direito, I. (2019). Scenario Evaluation with Relevance and Interest (SERI): Development and Validation of a Scenario Measurement Tool for Context-Based Learning. *International Journal of Science and Mathematics Education*, *17*(7), 1317–1338.

Küçükerdönmez, Ö., Akder, R. N., Seçkiner, S., & Oksel, E. (2021). Turkish version of the ' Three-Factor Eating Questionnaire-51 ' for obese individuals : a validity and reliability study. *Public Health Nutrition*, *24*(11), 3269–3275.

Ludwig, T., Priemer, B., & Lewalter, D. (2021). Assessing Secondary School Students' Justifications for Supporting or Rejecting a Scientific Hypothesis in the Physics Lab. *Research in Science Education*, *51*(3), 819–844.

Pedaste, M., Baucal, A., & Reisenbuk, E. (2021). Towards a science inquiry test in primary education: development of items and scales. *International Journal of STEM Education*, *8*(1), 1–19.

Polizzi, S. J., Zhu, Y., Reid, J. W., Ofem, B., Salisbury, S., Beeth, M., Roehrig, G., Mohr-Schroeder, M., Sheppard, K., & Rushton, G. T. (2021). Science and mathematics

teacher communities of practice: social influences on discipline-based identity and self-efficacy beliefs. *International Journal of STEM Education*, *8*(1), 1–18.

Pumptow, M., & Brahm, T. (2021). Students' Digital Media Self-Efficacy and Its Importance for Higher Education Institutions: Development and Validation of a Survey Instrument. *Technology, Knowledge and Learning*, *26*(3), 555–575.

Rachmatullah, A., & Ha, M. (2019). Indonesian and Korean high school student's disparities in science learning orientations: an approach to multi-group structural equation modeling. *Asia-Pacific Science Education*, *5*(1), 1–17.

Seery, N., Buckley, J., Delahunty, T., & Canty, D. (2019). Integrating learners into the assessment process using adaptive comparative judgement with an ipsative approach to identifying competence based gains relative to student ability levels. *International Journal of Technology and Design Education*, *29*(4), 701–715.

Svenningsson, J., Hultén, M., & Hallström, J. (2018). Understanding attitude measurement: exploring meaning and use of the PATT short questionnaire. *International Journal of Technology and Design Education*, *28*(1), 67–83.

Taber, K. S. (2018). The Use of Cronbach's Alpha When Developing and Reporting Research Instruments in Science Education. *Research in Science Education*, *48*(6), 1273–1296.

Tiruneh, D. T., De Cock, M., Weldeslassie, A. G., Elen, J., & Janssen, R. (2017). Measuring Critical Thinking in Physics: Development and Validation of a Critical Thinking Test in Electricity and Magnetism. *International Journal of Science and Mathematics Education*, *15*(4), 663–682.

Tzivinikou, S., Charitaki, G., & Kagkara, D. (2021). Distance Education Attitudes (DEAS) During Covid-19 Crisis: Factor Structure, Reliability and Construct Validity of the Brief DEA Scale in Greek-Speaking SEND Teachers. *Technology, Knowledge and Learning*, *26*(3), 461–479.

Werno Sujito, S., Hardyanto, W., & Lestari. (2015). Pengembangan Model Pembelajaran Seni Lukis Berbantuan Aplikasi Tux Paint Guna Meningkatkan Kemampuan Menggambar Alam Di Sekolah Dasar *Journal of Educational Research and Evaluation*. *4*(1).

Yang, X., Zhang, M., Kong, L., Wang, Q., & Hong, J. C. (2021). The Effects of Scientific Self-efficacy and Cognitive Anxiety on Science Engagement with the "Question-Observation-Doing-Explanation" Model during School Disruption in COVID-19 Pandemic. *Journal of Science Education and Technology*, *30*(3), 380–393.