REiD (Research and Evaluation in Education), 6(1), 2020, 41-50
Available online at: http://journal.uny.ac.id/index.php/reid

NGSS-oriented chemistry test instruments: Validity and reliability analysis with the Rasch model

***[1]Roudloh Muna Lia; [1]Ani Rusilowati; [1]Wiwi Isnaeni**
[1]Graduate School, Universitas Negeri Semarang
Jl. Kelud Utara III, Petompon, Gajahmungkur, Kota Semarang, Jawa Tengah 50237, Indonesia
*Corresponding Author. E-mail: aliamoetz@yahoo.co.id

Submitted: 10 February 2020 | Revised: 27 April 2020 | Accepted: 1 May 2020

Abstract

The instrument of measuring test attributes must be valid and reliable. This study was carried out since the validity and reliability testing of the chemistry items used by the testee is necessary. This study aims to estimate the validity and determine the reliability of chemical test instruments oriented Next Generation Science Standards (NGSS). The research was conducted through a quantitative descriptive approach in two vocational schools of engineering program which had 130 testees. The instrument used was an NGSS-oriented chemistry test instrument containing 35 items and an expert validation questionnaire. The obtained test participant's response from the test instrument was collected through the documentation method. Item in NGSS test were presented to three subject matters experts. The validities used were the content validity and the construct validity. The reliability was tested through internal consistency and interrater consistency approaches. The results show that content validity (Aiken's V) is at a range of 0.50 to 1.00. The value of the unexplained variance is less than 10%, which means that it is well-categorized. This analysis is strengthened by CFA which has a goodness of fit and a good measurement model fit. The parameters used to test model fit are CFI, NFI, RMSEA and the value of loading factor. Some results values are over 0.90 and RMSEA is 0.00 and more than 0.3 of loading factor value on each item. All scales had alpha reliability more than the criteria of 0.70. Thus, the developed chemical test item were proven as valid and reliable instruments.

Keywords: *validity, reliability, NGSS*

How to cite: Lia, R., Rusilowati, A., & Isnaeni, W. (2020). NGSS-oriented chemistry test instruments: Validity and reliability analysis with the Rasch model. *REiD (Research and Evaluation in Education), 6*(1), 41-50. doi:https://doi.org/10.21831/reid.v6i1.30112.

Introduction

In the Government Regulation No. 32 of 2013, it is written that learning process in the education unit is carried out interactively, inspiratively, pleasently, defiantly which motivates students to participate actively, as well as providing sufficient space for initiative, creativity, and independence by following their talents, interests, physical and psychological development of students. Educators or teachers are required to carry out the mandate of government regulation. The implementation of learning will be achieved based on the goals set if it is suitable for the students' talents and interests. Students from the Engineering Program of vocational school will be less suitable if Business Economics subject is taught because it does not match with interests and expertise areas of students, likewise the Chemistry lessons that are applied at Vocational High School (VHS). The existence of chemistry subjects in the Engineering Skills Program can support the development of learners' competencies if the material is adjusted to the expertise area of students (Wena, 2009 in Banne, 2018, p. 45). If the chemistry is taught separately and it is not associated with productive subjects in the expertise area which is occupied, the chemistry subject will be irrelevant (Astuti, Sunarno, & Sudarisman, 2016).

Facts in the field from the results of questionnaire distribution in vocational students showed as many as 76 % of students stated that chemistry was a difficult subject. The reason is that students are less interested in chemistry lessons because they consider that chemistry subject is not important for them (Lia & Isnaeni, 2018, p. 403). Chemistry as an adaptive subject in VHS is expected to be in accordance with productive material needs. One way to present chemistry subjects to be in accordance with productive material in learners' expertise area is through Next Generation Science Standards (NGSS) (Lia, 2019, p. 113).

NGSS provides the opportunity to include engineering in science (National Research Council, 2013, p. xviii). One of the assessment challenges in NGSS is creating assignments that include the practical side of science and engineering (Damelin, 2017). NGSS offers a new standard combining content and practice in science and engineering (National Research Council, 2013). NGSS creates a new vision for science education based on the idea that science is a unity of knowledge and a set of practices related to developing knowledge (Penuel, Harris, & DeBarger, 2015, p. 45). This teaching and learning approach is built on decades of research that identifies problems through learning in science classes and promising strategies to make learning to be more meaningful and effective for students (Reiser, 2013).

NGSS-oriented chemistry learning had been successfully developed by Lia (2019). After the learning process has been implemented, it is followed by an assessment activity. Assessment is an activity conducted to measure and assess the curriculum achievement level (Sudrajat, 2016, p. 1). Through assessment, any lacks in learning can be identified and can be evaluated.

The assessment instrument in measuring the question attributes as students' evaluation material must be valid and reliable. Therefore, further research on the development of the NGSS learning model, namely the preparation of chemical items needs to be conducted. The NGSS-oriented chemistry items developed provide breakthroughs to give students a more meaningful assessment. Assessment becomes more meaningful because it is associated with technical material by following the field occupied by students. Before carrying out the test, some practicums were oriented towards NGSS which made the chemical side more desirable (Lia, 2019, p. 113).

The NGSS-oriented chemistry question items must have two important requirements. Those are having a good validity and reliability level. Validity and reliability will be fulfilled if the questions have been arranged. Item analysis is analyzed in order to obtain the adequate quality of the question, and data processing and interpretation of the assessment result (Kadir, 2015, p. 71). Reynolds, Livingston, and Willson (2010, p. 144) state that validity means the extent to which theoretical and empirical evidence supports the meaning and interpretation of test scores. In addition, Dewi and Sukadiyanto (2015, p. 230) explain that a valid test is a test that can measure accurately and thoroughly the symptoms which are to be measured). Reliability is test consistency (Bhakti, 2015; Khumaedi, 2012). It means that a reliable test must have consistent results even if tested repeatedly at different times. It is in accordance with the theory explained by Reynolds et al. (2010, p. 91) that reliability is the accuracy or stability of the assessment results. The measuring tools used by evaluators when carrying out evaluation activities must have accuracy, consistency, and stability so that the measurement results obtained can measure accurately (Amalia & Susilaningsih, 2014). A set of tests must have accuracy when it is used. It also should be consistent and stable in the sense that there is no change from one measurement time to another (Utami, 2018, p. 5).

This study aims to estimate the validity and determine the reliability of chemical test instruments oriented NGSS to measure the level of understanding of chemical material in engineering. Research on the validity and reliability of the test instruments has been conducted by Mohamad, Sulaiman, Sern, and Salleh (2015), Kusaeri, Sutini, Suparto, and Wardah (2019), and Iskandar (2017). The differences between previous and current research are the analysis of the validity of the construct using the confirmatory factor analysis (CFA) modification and the Rasch model. It is expected that research on validity and reliability will increase knowledge in the field of teaching, especially in the evaluation of learning.

Rasch model used in this study has several advantages which can identify the error response, predict missing data scores, distinguish the ability of respondents with the same raw score, and also identify any indications of guesses and cheaters (Sumintono & Widhiarso, 2015, pp. 44–45). These advantages make the Rasch model more accurate (Lord in Nurcahyo, 2016). Rasch modeling can produce standard error measurement values which can improve the accuracy of calculations (Ardiyanti, 2016, p. 261). Sabekti and Khoirunnisa (2018, p. 69) confirm that the Rasch model is more recommended to be used in the development of test instruments.

An assessment of the appropriateness of the item's display and/or content validity becomes the earlier steps. Assessments carried out by a panel of experts and chemistry teachers are also included in the expert panel (Ismail, Permanasari, & Setiawan, 2016, p. 239). Instruments that have been compiled and validated by experts are then validated empirically through trial instruments in small classes (Prabowo & Ristiani, 2011, p. 80).

The high of agreement among experts who assess the feasibility of an item can be estimated and quantified. Then, the statistical calculation is used as an indicator of the item content validity and the test content validity. This study used an assessment procedure in measuring validity thorough a content validity coefficient (the content validity of the test with a V index) proposed by Aiken's V. The construct validity was tested using CFA with the help of Lisrel 8.8 software. Proof of construct validity used first order confirmatory factor analysis which calculated the estimated value of the item against its latent variable. According to Sitninjak and Sugiarto in Rusilowati (2014, p. 131), the validity of an observed variable can be seen from the factor loading of the variable against latent variable. Variables are labelled as good construct validity when the goodness of fit and the measurement model fit are met.

Method

The study was conducted in two vocational high schools in Engineering Program with a total of 130 testees. The instrument used was an NGSS-oriented chemical test instrument, amounting to 35 items and validation sheet. Based on the test instrument, the result of the test participants' answers was obtained and collected through the documentation method.

Three experts were assessing to obtain three sheets of questionnaire result. The validity was estimated by content validity, validity in large class trials, and construct validity. Then, the reliability was estimated through internal consistency and interrater consistency approaches. To analysis the content validity, the Aiken's V Formula was used. The construct validity with CFA was used with the help of Lisrel 8.8 software. The internal consistency reliability used in this study is the Spearman-Brown's formula in small class trials, whereas in large class trials, the Rasch alpha Cronbach model and interrater reliability using three raters tested using two-way ANOVA with Ebel formula were used.

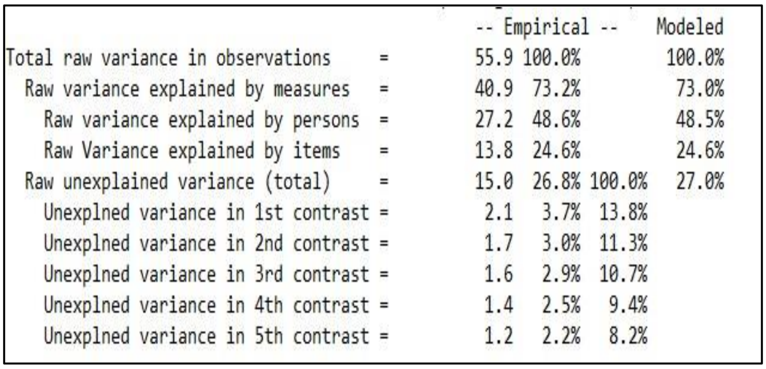
Findings and Discussion

Validity Test

Content validity was estimated with Aiken's V index. Items in NGSS test were presented to three experts to assess the compatibility of the material, construction, language and compatibility with NGSS. The experts also filled out a questionnaire containing the conclusions of the experts' assessment of chemistry-oriented items in NGSS. Quantitative data that present a summary of quantitative expert agreement coefficient data are shown in Table 1.

Table 1. Coefficient Data of Expert Agreement

Item Number	Aiken's V Index	Criterion	Conclusion
1, 5, 6, 7, 8, 9, 11, 12, 14, 16, 17, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35	1.0	valid	eligible
2, 3, 4, 13, 15, 18 19, 20	0.8	not valid enough	little revision
21	0.7	not valid enough	little revision
10	0.5	not valid enough	little revision



```
                                          -- Empirical --    Modeled
Total raw variance in observations     =   55.9 100.0%        100.0%
  Raw variance explained by measures   =   40.9  73.2%         73.0%
    Raw variance explained by persons  =   27.2  48.6%         48.5%
    Raw Variance explained by items    =   13.8  24.6%         24.6%
  Raw unexplained variance (total)     =   15.0  26.8% 100.0%  27.0%
    Unexplned variance in 1st contrast =    2.1   3.7%  13.8%
    Unexplned variance in 2nd contrast =    1.7   3.0%  11.3%
    Unexplned variance in 3rd contrast =    1.6   2.9%  10.7%
    Unexplned variance in 4th contrast =    1.4   2.5%   9.4%
    Unexplned variance in 5th contrast =    1.2   2.2%   8.2%
```

Figure 1. Unidimensionality Test

Based on the results of the data analysis in Table 1, 25 of 35 items are valid and 10 items are not valid enough, which means that there are some revisions. Comparing to previous studies, the quality classification research items analyzed are better than the result of research by Hasnah (2017) in which only nine of 40 items are categorized well.

Construct validity was proven by combining the factor analysis of the Rasch model and CFA (using Lisrel 8.8 software). The first step to see the construct validity with the Rasch model is through Output Diagnosis Item Polarity (Hayati & Lailatussaadah, 2016, p. 173). All items have a positive Point Measure Correction (Pt. Mea- Corr). A total of 14 items have strong or high correction numbers. One of the items (question number 5) has a moderate correlation number (0.57). It is in accordance with the opinion of Othman, Salleh, Hussein, and Wahid (2014, p. 117) that the high *Pt. Mea Corr* (0.68- 1.00) shows that a question item can distinguish respondents' ability.

The result of the correlation figures on *Pt. Mea Corr* is strengthened to the results of the unidimensionality test through the output table unidimensionality. The output table unidimensionality is presented in Figure 1.

The raw variance in Figure 1 shows a high number (73.2%). According to the opinion of Hakiki, Fitri, and Agung (2018, p. 42), the results of the analysis which have a unidimensionality requirement of more than 60 % show special meaning. The instrument which is developed can measure what should be measured. Variance values that cannot be explained (unexplained variance) successively are 3.7; 3.0; 2.9; 2.5; and 2.2. It shows that the variances which cannot be explained by the instruments are all less than 10%. It indicates that the unidimensionality in the instruments falls into a good category (Wibisono, 2014, p. 744).

The construct validity test on Rasch is only for the response of the tested item, whereas to find out the covariance between the test items, the CFA model with the Lisrel or Amos or SPSS programs is needed. About specifying a model for a data set, the procedures for CFA appear to be more advanced, simpler, and more user-friendly than those developed for Rasch (IRT). The CFA model can calculate an accurate estimate of the chi-square size of the fit model and related degrees (Reise, Widaman, & Pugh, 1993, pp. 554–563). Therefore, the researchers strengthened the construct validity test through the Lisrel program.

Conceptually, to make a test across NGSS, three components should be recked, namely DCIs, SEs, and also CCs. DCIs are very dependent on the material that will be made from the instrument. Then, SEPs and CCs are the characteristics of NGSS-oriented statistics. SEPs consist of six aspects with 15 indicators. CCs consist of three aspects with 14 indicators. The results of the NGSS instrument construct validity with CFA prove that the dimensions of CCs which consist of three aspects with 14 indicators are evidenced by the factor loading value and item compatibility parameters. The analysis of CCs components consisting of three aspects and 14 indicators is generated in a diagram presented in Figure 2.

Analysis through CFA proved that CCs dimensions which consisted of three aspects with 14 indicators are evidenced by the value of loading factor and items that are compatible with the parameters. All factor loading's value shows that there are more than 0.3. Factor loadings which are less than 0.5 are removed (Arifin, Yusoff, & Naing, 2012). The parameters that are used to test model fit are CFI, NFI, and RMSEA. CFI and NFI are over 0.90 (CFI=0.92; NFI=0.90) and RMSEA is 0.00. It is compatible with the theory that the expected CFI and NFI values are above 0.90 (Zehir, Akyuz, Eren, & Turhan, 2013, p. 9). RMSEA is recommended to be under 0.05 though acceptable up to 0.08 (Sohail & Jang, 2017). In Rusilowati (2014, p. 134), it is stated that the compatibility of the model that is developed by empirical data at a minimum can be seen from three match sizes that represent the three categories of match test different models. When two of the three categories are significant, the model developed is compatible with the data. All model fits were acceptable and according to the literature, the validity of the measurements in the current study met the criteria.

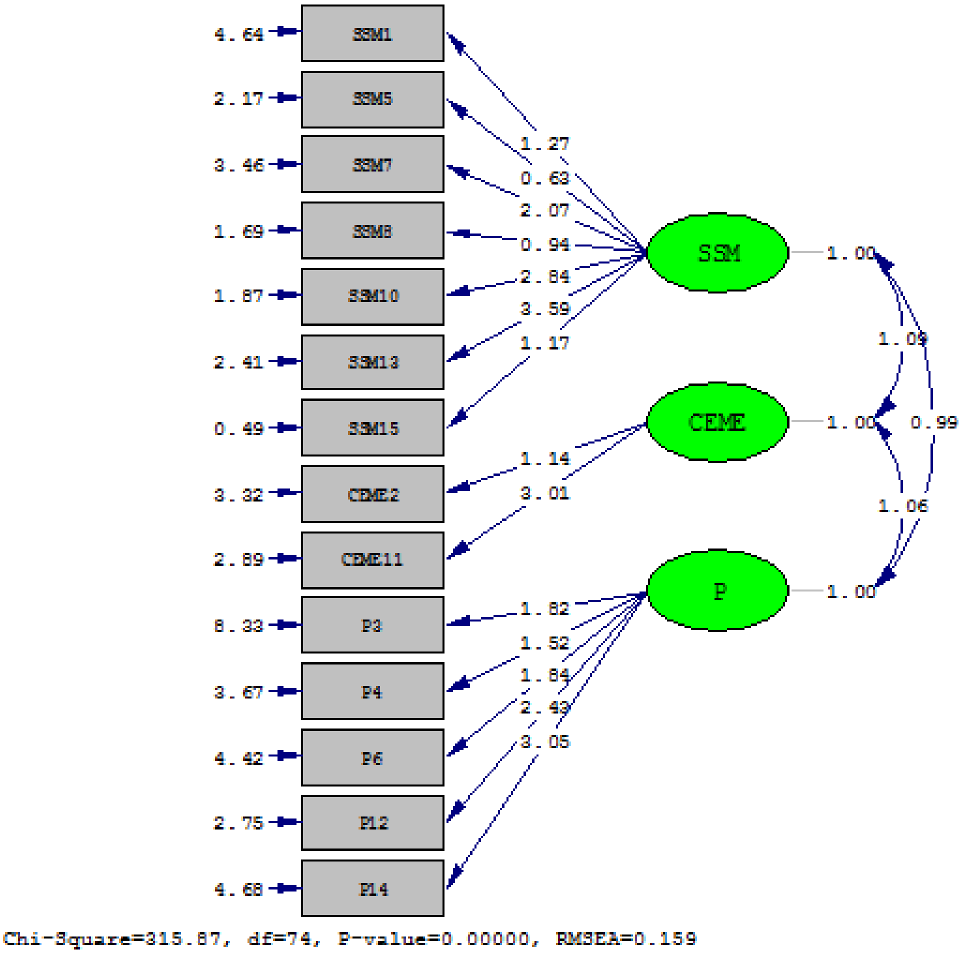


Figure 2. CCs Path Diagram

The validity of the large class trial phase was analyzed using the Rasch through the Output model, item fit order. The output is presented in Table 2.

Table 2. Item Fit

Item's	Outfit		
Number	MNSQ	ZSTD	PT. Mea Corr
1	2.11	3.9	0.68
3	1.99	3.2	0.75
6	1.47	1.7	0.75
2	1.33	1.4	0.73
7	1.33	1.7	0.75
9	1.16	0.8	0.74
12	1.08	0.5	0.80
5	1.03	0.2	0.57
8	0.94	-0.2	0.69
10	0.80	-1.0	0.87
4	0.81	-0.8	0.82
14	0.77	-1.0	0.83
13	0.56	-2.0	0.90
15	0.59	-1.9	0.78
11	0.48	-1.6	0.86

The item fit information is useful for identifying the indications of misconception (Sumintono & Widhiarso, 2015, p. 77). In Table 2, based on MNSQ, ZSTD, and Pt. Mea Corr, it can be concluded that 15 items were classified as valid, but there is one item namely question number 1 which is indicated as a misconception. The MNSQ value is 2.11 and the ZSTD is 3.9 which represents unexpected data. The cause of outlier MNSQ and ZSTD values is from some testee's answers. Those are reversed between "the oxidation-reduction reaction and the reason", but Pt. Mea Corr is still within the limit of more than 0.4 and less than 0.85. Therefore, 15 items have been used to measure the quality of education because these questions have been analyzed. It is in accordance with the opinion of Pancoro (2011, p. 94) that test questions need to be first analyzed to have the same characteristics so that they can be used to measure the quality of education.

Reliability Test

The reliability test consists of (a) inter-rater reliability, (b) small-scale trial reliability, and (c) large-class trial reliability. Based on Table 3, the values of the reliability of the tests are 0.17, 0.82, and 0.94. Inter-rater reliability (among experts) is very low, the reliability of small class trials is very high, and the reliability of large classes is special. A discussion of the three reliability tests is elaborated as follows.

Inter-rater Reliability

Inter-rater reliability is a preliminary part of a study (Dockrell et al., 2012, p. 633). Interrater reliability was calculated after calculating the content validity among three validators. Level agreement between three validators can be explained through the reliability coefficient between rater (assessors) using two-way ANOVA-analysis with the Ebel formula. Two-way ANOVA analysis through SPSS 16.0 is presented in Table 4.

In Table 4, it can be explained that Rater is the assessor and Item is a matter of Items. The mean square value of Rater is 0.495, the value of the item is 0159 and the interaction between Rater and Item (Rater * Item) is 0.132. These values are entered in the Ebel formula and produce a reliability coefficient of 0.17. The reliability coefficient of r value is less than 0.2. The reliability

Table 3. Reliability Data Analysis

Trial Phase	Reliability	N of Items
Expert (Expert Judgment)	0.17	35
Small Class	0.82	25
Big Class	0.94	15

Table 4. Output Reliability of Two-Way ANOVA

Source	Mean Square
Rater	0.495
Item	0.159
Rater*Item	0.132

among the assessors in assessing the contents of the instrument is still not consistent (Rusilowati, 2014, p. 29). When the reliability coefficient obtained is not high enough, there are inconsistencies among raters (Pinilih, Budiharti, & Ekawati, 2013, p. 25). The reason for this inconsistency in this research is the difference in viewpoints in evaluating chemical test instruments. For example, expert 1 puts more emphasis on its chemical content while expert 3 is more inclined in evaluating the appearance and suitability of the answers.

Small Class Trial Reliability

Reliability using the Spearman-Brown formula was applied to small classes and searched using the Anastes Description application. The reliability coefficient of small class tests based on Table 3 shows that the coefficient number is 0.82. Figures for reliability coefficient is 0.8 r < 1.0, which indicates very high reliability.

Big Class Trial Reliability

In the big class stage, the reliability is seen with the help of Winstep 3.73 program. Reliability in the Rasch model is illustrated by the presence of a separation index. The separation indexes reported are the item reliability and the person reliability which are supplemented by Cronbach Alpha KR-20 of reliability coefficient figures. Those are three successive coefficient numbers (0.91, 0.98 and 0.94). All three of these figures indicate very high reliability. Separation reliability (item or person reliability) is categorized as high value because the study sample and grain difficulty level have a wide range and produce a small measurement error. Broad grain means that the item has a difficulty level from the easiest to the most difficult. Similarly, in the study sample, a broad sample means that the sample can spread from the smartest to the least clever (Linacre, 2016, p. 256). The output reliability can be seen in Table 5. In Table 5, in addition to the reliability coefficient, there is also important information related to the statistical summary of the test participant's overall response patterns, namely (a) INFIT MNSQ ZSTD, and OUTFIT MNSQ ZSTD, and (b) Separation.

INFIT MNSQ ZSTD and OUTFIT MNSQ ZSTD

The MNSQ INFIT and MNSQ OUTFIT values are 0.99 and 1.21, respectively for persons as well as 0.98 and 1.10 for MNSQ INFIT values and MNSQ OUTFIT items. It is categorized as having a good value because the ideal value is 1 (the closer to 1 the better). The value of INFIT ZSTD and OUTFIT values are 0.99 and 1.21, respectively for persons as well as 0.98 and 1.10 for MNSQ INFIT values and MNSQ OUTFIT items. It is also categorized as having a good value because the ideal value is 1 (the closer to 1 the better). The value of INFIT ZSTD and OUTFIT ZSTD in sequence person and item are 0.0, 0.2, -0.1, 0.3. The ZSTD value is ideally 0.0, so that the ZSTD value including ideal except for the value of INFIT ZSTD in the item shows a negative value (not good).

Table 5. Output Reliability of Rasch Model

	Measured Person			
	Infit		**Outfit**	
	MNSQ	**ZSTD**	**MNSQ**	**ZSTD**
Mean	0.99	0.0	1.21	0.2
Separation			3.11	
Person Reliability			0.88	
	Measured Item			
	Infit		**Outfit**	
	MNSQ	**ZSTD**	**MNSQ**	**ZSTD**
Mean	0.98	-0.1	1.18	0.3
Separation			6.37	
Item Reliability			0.97	
KR-20 Test Reliability			0.94	

Separation

The greater the value of separation, the quality of the instrument in terms of overall respondents and grain is getting better. The separation value on the items developed is 8.45 by entering the formula H that has been explained. Score 8.45 rounded up to 8, which means that eight groups of items can be interpreted as groups of varied items.

Conclusion

This test instrument has been proven for content validity, construct validity, inter-rater reliability, and reliability with the Rasch model. The test instrument has fulfilled the content validity with expert judgment as evidenced by the acquisition of agreement index (Aiken index) ranging from 0.50 to 1.00. The lowest score (0.5) is caused by each value's interconsistence. The raw variance value in the analysis of the Rasch model's construct validity is 73.2% with a special category. Variance values that cannot be explained are less than 10%, consecutively 3.7; 3.0; 2.9; 2.5; 2.2 indicating that unidimensionality in the instrument is in a good category. The parameters used to test model fit are CFI, NFI, RMSEA, and the loading factor value. Some results values are over 0.90 (CFI=0.92; NFI=0.90) and RMSEA is 0.00, and more than 0.3 of loading factor value on each item which indicates that the variable has good validity to the construct. The test instrument increases the number of reliability coefficients at each step of the trial, i.e. 0.17, 0.82, and 0.94. The characteristics of the Rasch model items analyzed can reveal interpretations in terms of items, personnel, and instruments. Thus, the chemistry test items developed are tested to be valid, reliable and have adequate characteristics.

References

Amalia, N. F., & Susilaningsih, E. (2014). Pengembangan instrumen penilaian keterampilan berpikir kritis siswa SMA pada materi asam basa. *Jurnal Inovasi Pendidikan Kimia*, *8*(2), 1280–1389. Retrieved from https://journal.unnes.ac.id/nju/index.php/JIPK/article/view/4443

Ardiyanti, D. (2016). Aplikasi model Rasch pada pengembangan skala efikasi diri dalam pengambilan keputusan karir siswa. *Jurnal Psikologi*, *43*(3), 248–263. https://doi.org/10.22146/jpsi.17801

Arifin, W. N., Yusoff, M. S. B., & Naing, N. N. (2012). Confirmatory factor analysis (CFA) of USM Emotional Quotient Inventory (USMEQ-i) among medical degree program applicants in Universiti Sains Malaysia (USM). *Education in Medicine Journal*, *4*(2), 1–22. https://doi.org/10.5959/eimj.v4i2.33

Astuti, R., Sunarno, W., & Sudarisman, S. (2016). Pembelajaran IPA dengan pendekatan ketrampilan proses sains menggunakan metode Eksperimen Bebas Termodifikasi dan Eksperimen Terbimbing ditinjau dari sikap ilmiah dan motivasi belajar siswa. *Proceeding Biology Education Conference*, *13*(1), 338–345. Retrieved from https://jurnal.uns.ac.id/prosbi/article/view/5742

Banne, K. (2018). Meningkatkan aktivitas belajar kimia (Redoks) siswa kelas XII TKR SMK Negeri 1 Sumarorong melalui penerapan model pembelajaran kooperatif tipe NHT dengan materi berbasis kontekstual. *Jurnal MEKOM (Media Komunikasi Pendidikan Kejuruan)*, *5*(1), 45–50. https://doi.org/10.26858/mekom.v5i1.8223

Bhakti, Y. B. (2015). Pengaruh jumlah alternatif jawaban dan teknik penskoran terhadap reliabilitas tes. *Formatif: Jurnal Ilmiah Pendidikan MIPA*, *5*(1), 1–13. https://doi.org/10.30998/formatif.v5i1.168

Damelin, D. (2017). Using technology to enhance NGSS-aligned assessment tasks for classroom formative use. Retrieved from The Concord Consortium website: https://concord.org/newsletter/2017-spring/using-technology-enhance-ngss-aligned-assessment-tasks/

Dewi, P. C. P., & Sukadiyanto, S. (2015). Pengembangan tes keterampilan olahraga woodball untuk pemula. *Jurnal*

Keolahragaan, *3*(2), 228–240. https://doi.org/10.21831/jk.v3i2.6254

Dockrell, S., O'Grady, E., Bennett, K., Mullarkey, C., Mc Connell, R., Ruddy, R., … Flannery, C. (2012). An investigation of the reliability of Rapid Upper Limb Assessment (RULA) as a method of assessment of children's computing posture. *Applied Ergonomics*, *43*(3), 632–636. https://doi.org/10.1016/j.apergo.2011.09.009

Government Regulation No. 32 of 2013, on National Education Standard. , (2013).

Hakiki, A. W., Fitri, A. R., & Agung, I. M. (2018). Analisis properti psikometri subtes Merkaufgaben (ME) dengan Rasch model. *Jurnal Psikologi*, *14*(1), 40–49. https://doi.org/10.24014/jp.v14i1.4900

Hasnah, H. (2017). Analisis kualitas soal matematika Ujian Sekolah kelas XII IPA SMA Negeri di Watansoppeng berdasarkan Teori Respon Butir. *PEP Educational Assessment*, *1*(1), 27–33. Retrieved from https://ojs.unm.ac.id/UEA/article/view/3776

Hayati, S., & Lailatussaadah, L. (2016). Validitas dan reliabilitas instrumen pengetahuan pembelajaran aktif, kreatif dan menyenangkan (PAKEM) menggunakan model Rasch. *Jurnal Ilmiah Didaktika*, *16*(2), 169–179. https://doi.org/10.22373/jid.v16i2.593

Iskandar, A. (2017). *Teknik analisis validitas konstruk dan reliabilitas instrument test dan non test dengan software LISREL*. https://doi.org/10.31227/osf.io/nbhxq

Ismail, I., Permanasari, A., & Setiawan, W. (2016). STEM virtual lab: An alternative practical media to enhance student's scientific literacy. *Jurnal Pendidikan IPA Indonesia*, *5*(2), 239–246. https://doi.org/10.15294/jpii.v5i2.5492

Kadir, A. (2015). Menyusun dan menganalisisi tes hasil belajar. *AL-TA'DIB : Jurnal Kajian Ilmu Kependidikan*, *8*(2), 70–81. https://doi.org/10.31332/atdb.v8i2.411

Khumaedi, M. (2012). Reliabilitas instrumen penelitian pendidikan. *Jurnal Pendidikan Teknik Mesin*, *12*(1), 25–30. Retrieved from https://journal.unnes.ac.id/nju/index.php/JPTM/article/view/5273

Kusaeri, K., Sutini, S., Suparto, S., & Wardah, F. (2019). The validity and inter-rater reliability of project assessment in mathematics learning. *Beta: Jurnal Tadris Matematika*, *12*(1), 1–13. https://doi.org/10.20414/betajtm.v12i1.266

Lia, R. M. (2019). *Pengembangan butir soal Kimia berorientasi NGSS dan analisisnya menggunakan model Rasch*. Master thesis, Universitas negeri Semarang, Semarang.

Lia, R. M., & Isnaeni, I. (2018). Evaluation of Chemistry learning programs at vocational high school Semarang on Vehicle Engineering field. *Proceedings of the International Conference on Science and Education and Technology 2018 (ISET 2018)*, 403–407. https://doi.org/10.2991/iset-18.2018.82

Linacre, J. M. (2016). *A user's guide to WINSTEPS MINISTEP Rasch-model computer programs*. Chicago, IL: Winsteps.com.

Mohamad, M. M., Sulaiman, N. L., Sern, L. C., & Salleh, K. M. (2015). Measuring the validity and reliability of research instruments. *Procedia - Social and Behavioral Sciences*, *204*, 164–171. https://doi.org/10.1016/j.sbspro.2015.08.129

National Research Council. (2013). *Next Generation Science Standards: For states, by states*. https://doi.org/10.17226/18290

Nurcahyo, F. A. (2016). Aplikasi IRT dalam analisis aitem tes kognitif. *Buletin Psikologi*, *24*(2), 64–75. https://doi.org/10.22146/buletinpsikologi.25218

Othman, N. B., Salleh, S. M., Hussein, H., & Wahid, H. B. A. (2014). Assessing construct validity and reliability of competitiveness scale using Rasch model approach. *The 2014 WEI International Academic Conference Proceedings*, 113–120. Retrieved from https://www.westeastinstitute.com/wp-content/uploads/2014/06/Suria-Mohd-Salleh.pdf

Pancoro, N. H. (2011). Karakteristik butir soal ulangan kenaikan kelas sebagai persiapan bank soal Bahasa Inggris. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *15*(1), 92–114. https://doi.org/10.21831/pep.v15i1.1089

Penuel, W. R., Harris, C. J., & DeBarger, A. H. (2015). Implementing the Next Generation Science Standards. *Phi Delta Kappan*, *96*(6), 45–49. https://doi.org/10.1177/0031721715575299

Pinilih, F. W., Budiharti, R., & Ekawati, E. Y. (2013). Pengembangan instrumen penilaian produk pada pembelajaran IPA untuk siswa SMP. *Jurnal Pendidikan Fisika*, *1*(2), 23–27. Retrieved from https://jurnal.fkip.uns.ac.id/index.php/pfisika/article/view/2798

Prabowo, A., & Ristiani, E. (2011). Rancang bangun instrumen tes kemampuan keruangan pengembangan tes kemampuan keruangan Hubert Maier dan identifikasi penskoran berdasar teori Van Hielle. *Kreano, Jurnal Matematika Kreatif-Inovatif*, *2*(2), 72–87. https://doi.org/10.15294/kreano.v2i2.2618

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin*, *114*(3), 552–566. https://doi.org/10.1037/0033-2909.114.3.552

Reiser, B. J. (2013). What professional development strategies are needed for successful implementation of the Next Generation Science Standards. *The Invitational Research Symposium on Science Assessment*, 1–23. Retrieved from http://www.ets.org/Media/Research/pdf/reiser.pdf

Reynolds, C. R., Livingston, R. B., & Willson, V. L. (2010). *Measurement and assessment in education* (2nd ed.). Upper Saddle River, NJ: Pearson Education.

Rusilowati, A. (2014). *Pengembangan instrumen penilaian*. Semarang: Unnes Press.

Sabekti, A. W., & Khoirunnisa, F. (2018). Penggunaan Rasch model untuk mengembangkan instrumen pengukuran kemampuan berpikir kritis siswa pada topik ikatan kimia. *Jurnal Zarah*, *6*(2), 68–75. https://doi.org/10.31629/zarah.v6i2.724

Sohail, M. S., & Jang, J. (2017). Understanding the relationships among internal marketing practices, job satisfaction, service quality and customer satisfaction: An empirical investigation of Saudi Arabia's service employees. *International Journal of Tourism Sciences*, *17*(2), 67–85. https://doi.org/10.1080/15980634.2017.1294343

Sudrajat, D. (2016). Portofolio: Sebuah model penilaian dalam Kurikulum Berbasis Kompetensi. *Intelegensia*, *1*(2), 1–9. Retrieved from http://ejurnal.unikarta.ac.id/index.php/intelegensia/article/view/257

Sumintono, B., & Widhiarso, W. (2015). *Aplikasi pemodelan Rasch pada assessment pendidikan*. Cimahi: Trim Komunikata.

Utami, B. N. (2018). *Praktik evaluasi penyuluhan pertanian*. Malang.

Wibisono, S. (2014). Aplikasi model Rasch untuk validasi instrumen pengukuran fundamentalisme agama bagi responden muslim. *JP3I (Jurnal Pengukuran Psikologi Dan Pendidikan Indonesia)*, *3*(3), 729–750. https://doi.org/10.15408/jp3i.v3i3.10731

Zehir, C., Akyuz, B., Eren, M. S., & Turhan, G. (2013). The indirect effects of servant leadership behavior on organizational citizenship behavior and job performance: Organizational justice as a mediator. *International Journal of Research in Business and Social Science (2147-4478)*, *2*(3), 1–13. https://doi.org/10.20525/ijrbs.v2i3.68