JEC | Journal of Educational Chemistry
Journal of Educational Chemistry (3 (1), 2021, 11-22)
Website: https://www.journal.walisongo.ac.id/index.php/jec
ISSN 2715-3029 (p) 2685-4880 (e)
DOI: 10.21580/jec.2021.3.1.6582

Chemistry Learning Outcomes Assessment: How is The Quality of The Tests Made by The Teacher?

Rizki Nor Amelia[1], Anggi Ristiyana Puspita Sari[2], Sri Rejeki Dwi Astuti[3*]
[1]Department of Education Research and Evaluation, Universitas Negeri Yogyakarta
[2]Department of Chemistry Education, Universitas Palangka Raya
[3]Department of Chemistry Education, Universitas Negeri Yogyakarta
*E-mail Corresponding Author: srirejeki.dwiastuti@yahoo.com

Abstract

The teacher-made chemistry test must have a good quality, due to the decision taken from the tests has an impact on the students. Therefore, the purpose of the research is to explore the quality of teacher-made chemistry tests such as item fit and person fit, item difficulty, and test reliability. The sample consisted of 356 senior students from senior high schools in Yogyakarta that were selected by cluster random sampling technique. The research used the teacher-made chemistry test consisted of 40 multiple choice items which were collected using documentation technique. Data were analyzed with Rasch Model using Winsteps 3.73 version. The result showed that all items in the teacher-made chemical test were proven to have good quality (fit model, good item difficulty, and good test reliability). Moreover, 18 students were identified as misfit persons. From the findings, the test can be used to assess the students' learning outcomes, especially for the try-out of the final exam in senior high school. Besides, the students identified as person misfits should be further examined and receive teachers' guidance.

Keywords: item fit, Rasch model, person fit, teacher-made chemistry test

Abstrak

Instrumen tes kimia buatan guru haruslah memiliki kualitas yang baik, mengingat keputusan yang diambil dari hasil tes tersebut memiliki dampak terhadap siswa. Oleh sebab itu, penelitian ini bertujuan untuk mengidentifikasi kualitas tes kimia buatan guru seperti item fit, person fit, tingkat kesukaran butir soal, dan reliabilitas tes. Sampel yang digunakan pada penelitian ini sebanyak 356 siswa kelas XII SMA Negeri di Kota Yogyakarta yang dipilih menggunakan teknik cluster random sampling. Instrumen yang digunakan dalam penelitian ini adalah instrumen tes kimia buatan guru yang tersusun atas 40 butir soal yang dikumpulkan dengan teknik dokumentasi. Data dianalisis menggunakan Pemodelan Rasch dengan berbantuan Software Winsteps versi 3.73. Hasil analisis menunjukkan bahwa semua item penyusun instrumen tes kimia buatan guru terbukti memiliki kualitas yang baik (semua item terbukti fit model, memiliki tingkat kesukaran butir yang baik dan memiliki reliabilitas tes yang baik). Sementara itu, 18 siswa dari total sampel yang digunakan, teridentifikasi sebagai person misfit. Berdasarkan hasil tersebut, instrumen tes kimia buatan guru ini dapat digunakan untuk menilai hasil belajar siswa, khususnya untuk try-out ujian akhir nasional bidang kimia di SMA. Selain itu, siswa yang teridentifikasi sebagai person misfit sebaiknya diperiksa lebih lanjut untuk mendapatkan bimbingan guru.

Kata kunci: item fit, model Rasch, person fit, tes hasil belajar kimia

Introduction

Teachers play an important role in the assessment and evaluation process of students' abilities. Assessment is a method and technique used to collect information related to students' abilities, while evaluation is a process or activity of collecting information related to students' abilities to determine the success of a learning goal (Tosuncuoglu, 2018). Therefore, the assessment of learning outcomes can be applied in the learning evaluation process.

To figure out that the results of the assessment reflect the real situation, the process of evaluating learning outcomes must be carried out properly. One way to make the assessment run optimally is to use an assessment instrument that is in line with the purpose of the assessment. If the assessment runs optimally, the teacher can figure out various developments in learning outcomes and provide feedback to students (Nurhayati & Ahmad, 2018).

In addition to using the right assessment instrument, the teacher must also be careful in choosing the data analysis technique which will be used to analyze the measurement data. There are various models of analysis of measurement results, one of which is the Rasch model. Rasch is a measurement model based on modern test theory related to consideration of the respondent's ability to answer tests on the level of difficulty of the items that make up an instrument (Rasch, 1980). Mathematically, the Rasch model is equivalent to the 1-PL (1-Parameter Logistics) model in IRT (Item Response Theory), but is developed separately and is not defined as a special case of the 2-PL model (Rasch 1980).

By using the Rasch model in learning evaluation, the teacher can perform several stages of analysis in one analysis process, that is by proving the validity of the instrument used, identifying the characteristics of the instrument such as the level of difficulty of item (b), identifying test reliability, and analyzing students' abilities (θ) in more accurate detail (Runnels, 2012). The presence of the Rasch model as a new measurement system aims to overcome the limitations of the classical measurement system or Classical Test Theory (CTT) (Ashraf & Jaseem, 2020; Yilmaz, 2019).

In general, the fit of the model to the data is a major concern when applying analysis using modern test approaches. If the data deviates greatly from the Rasch model, the cause needs to be considered, namely the person misfits or item may need to be eliminated (Boone & Noltemeyer, 2017). Therefore, specifically for the Rasch model, there are two types of fit, namely item fit and person fit, which illustrate the validity of the Rasch model measurement (Wright & Stone, 1999).

Item fit and person fit can be used to detect differences between empirical data and Rasch model data (Bond & Fox, 2015).

The research urgency of item fit and person fit is undeniable considering that both have a strong linear relationship, both of which play an important role in the construction of the test, especially concerning evaluation and item selection issues and in making decisions on test scores based on individual response results. Therefore, through the identification of fit items, errors that occur during the calibration phase of the development of an instrument can be detected. For example, if there is an item that has a poor discriminating power parameter, then item fit statistics will identify the problem. In addition, person fit will be able to show whether or not there are deviations in the response pattern (director with a score that is too high or too low) due to cheating, caring, responding, lucky guessing, responding, and random responding (Karabatsos, 2003). Respondents who are included in-person fit are only able to answer the item correctly when the item has a level of difficulty below the respondent's ability.

Historically, research on item fit was started by Andersen in 1973 and Yen in 1981; while research related to person fit was pioneered by Spearman in 1910, Thurstone in 1927, and Cronbach in 1946 (Karabatsos, 2003). In Indonesia itself, the results of research on the quality of teacher-made chemical test instruments that specifically describe fit and person fit items are relatively difficult to find. Research related to teacher-made chemical test instruments is generally limited to qualitative assessments, for example, the research of Nugraheni, Widodo, & Sugiyo (2013) which examined objectivity, practicality, and economic terms; while the research by Sakinah & Ritonga (2017) examined the aspects of the material, language, and construction. Research from Muna, Hanafi, & Rahim (2019) and Sopiah, Sidauruk, & Asi (2019) was conducted slightly differently by describing the quality of teacher-made tests in terms of the proportion of the cognitive domain of Bloom's taxonomy. In addition, the item analysis carried out was still using the classical test theory approach as research from Irmayta, Rudibyani, & Efkar (2018); Jayanti (2020); Prabayanti, Sudiyana, & Wiratini (2018); Purba, Susanti, & Rosna (2019); and Yustika, Susatyo, & Nuswowati (2014). The analysis using the classical theory approach will produce different item parameter characteristics if tested on different test takers. It is different from research by Wiyarsi, et al (2019) which has used Rasch modeling, but the analysis carried out is only limited to analysis of test validity and reliability.

Identification of item and person fit, for example, is carried out to determine and ensure that the items that make up the chemistry test are feasible so that the resulting scores will truly reflect the students' overall chemical abilities. With the information on the results of the analysis related to the chemical test instrument made by the teacher, it is hoped that the teacher can make a better chemical test instrument. This is because quality teacher-made chemistry tests are needed considering the decisions taken from the test results have an impact on students. Therefore, this study aims to identify the quality of teacher-made chemistry tests such as item fit, person fit, item difficulty level, and test reliability.

Research Method

The approach used in the research is a quantitative approach, where research data is obtained through documentation techniques. The instrument used is a summative assessment instrument in the form of a national final test try-out which is prepared by a chemistry teacher under the coordination of the Principal Work Consultation (MKKS) of the Special Region of Yogyakarta. This instrument is in the form of a multiple-choice test consisting of 40 questions to measure students' cognitive chemical abilities which include 72.5% understanding, 17.5% implementation, and 10% reasoning.

The domains of chemistry tested included 20% of Basic Chemistry (atomic structure, periodic system of elements, chemical bonds, nomenclature of inorganic and organic compounds, simple reaction equations, and basic laws of chemistry); 17.5% Analytical Chemistry (solution (non)-electrolyte, acid-base, solution stoichiometry, buffer solution, salt hydrolysis, solubility and solubility product); 35% Physical Chemistry material (thermochemistry, reaction rate, chemical equilibrium, colloidal chemical bonds, and colligative properties of solutions, redox and electrochemical reactions); 20% Organic Chemistry materials (carbon compounds, petroleum, and macromolecules: polymers, carbohydrates, and proteins, as well as their quantitative analysis methods, fats-oils); and 7.5% Inorganic Chemical materials (chemical bonds, chemical elements found in nature including radioactivity, properties, benefits, reactivity, and production).

The sample used in this study was 356 students of grade XII of State Senior High School in Yogyakarta which were selected by using the cluster random sampling technique. The response data obtained were then analyzed using Rasch Modeling with the help of Winsteps software version 3.73 (Linacre, 2009) to identify the presence or absence of misfit items and person misfits, and to determine psychometric characteristics in the form of difficulty level (b) items and test reliability.

Finding and Discussion

A good assessment instrument must meet two conditions, namely the instrument must be able to measure what it is supposed to measure (the test must be able to measure what the measurement objective is) and be reliable (give steady/fixed results when tested on different test takers. Furthermore, there is one additional requirement that an assessment instrument must have according to modern test theory, namely the characteristics of the test. The characteristics of the tests revealed in this study include the level of difficulty of the items (Hambleton & Swaminathan, 1985) and item fit (DeMars, 2010).

Table 1

Summary of MNSQ Outfit Statistics and Item Difficulty (b)

No	MNSQ	b (logit)	No	MNSQ	b (logit)
1	1,24	0,05	21	0,97	-0,18
2	1,05	-1,36	22	0,90	0,27
3	0,89	-1,72	23	0,76	*-2,44*
4	1,09	-0,62	24	1,13	-0,51
5	0,97	0,11	25	1,20	0,28
6	0,95	0,14	26	0,88	1,19
7	1,05	0,01	27	1,14	0,48
8	1,06	-1,27	28	1,25	-1,34
9	1,01	-0,36	29	1,54	-1,43
10	0,76	-1,23	30	1,01	-1,75
11	1,08	1,74	31	1,01	-0,54
12	0,98	1,09	32	0,78	1,84
13	0,92	-0,33	33	0,87	1,07
14	0,85	0,42	34	0,93	-0,81
15	0,81	-2,00	35	1,00	0,77
16	1,04	1,13	36	0,99	0,66
17	1,02	0,38	37	1,14	1,22
18	1,43	*2,72*	38	0,9	0,35
19	0,82	0,45	39	1,08	0,45
20	1,02	0,87	40	1,31	0,21

Before estimating the item parameters using Rasch modeling, first, an item fit and person fit analysis was performed. The analysis was conducted to determine whether the measurement data were suitable to be analyzed using the Rasch model or not. Item fit in the Rasch model was seen in each item, not all items, besides that person fit was seen in every respondent involved in the measurement process (DeMars, 2010). Linacre (2002) provided a rule of thumb to assess the implications of the fit of the model on the measurement, namely MNSQ > 2.0 which means that it destroys the measurement system; 1.5 < MNSQ 2.0 which means that it has no meaning for the measurement; 0.5 MNSQ 1.5 which means useful for measurement; and MNSQ < 0.5 which means that it is not useful for measurement even though it does not damage the measurement system.

Table 1 presents statistics on the MNSQ outfit (outlier-sensitive or information-weighted fit Mean Square) which was used to identify whether an item is fit or misfit to the Rasch model along with the characteristics of item parameters in the form of item difficulty level (b) for each item making up the instrument. teacher-made chemistry test. Based on Table 1, it can be seen that the MNSQ outfit statistics ranged from 0.76 to 1.54; while the item difficulty level ranged from -2.44 logit to 2.7 logit.

Based on Linacre's (2002) criteria, all the items of the teacher-made chemical test instrument were proven to fit the Rasch model. This states that there are no misfit items (items having a fit statistic that is too high or too low) in the analyzed instrument. In the absence of misfit items, it can be interpreted that the scores obtained by students as a response in answering the teacher-made chemistry test questions have already reflected the students' actual chemical abilities. In a measurement instrument, misfit items must be watched out for because these items are items that do not contribute much to the reliability of the test score. If an item is found to be a misfit to the Rasch model, then there is an indication if the item's construction is indeed flawed or problematic (e.g. the item has poor discriminatory power) or the item's parameters have questionable validity (the item measures other abilities).

To complete the information about the characteristics of the items that make up the teacher-made chemical test instrument, the item parameter was also identified, namely the level of difficulty. In the IRT approach, the level of difficulty (parameter b) is defined as a point or location where the S-shaped curve has the steepest slope on an ability scale (Al-khadher & Albursan, 2017; Adedoyin & Mokobi, 2013), whose magnitude ranges from logit – to logit +. Although generally only -2 logit to 2 logit, it is neither too easy nor too difficult for the intended test subject (DeMars, 2010; Hambleton & Swaminathan, 1985; Hambleton, Swaminathan, & Rogers, 1991). Therefore, in this study, the item is said to have a low level of difficulty (easy item) if b < 2.0 logit; moderate difficulty level (medium item) if -2.0 logit b 2.0 logit; and high difficulty level (difficult item) if b > 2.0 logit.

Based on these guidelines, only item number 18 was classified as a difficult item (2.72 logit) and only item number 23 was classified as an easy item (-2.44 logit), while the remaining 38 items were classified as medium item (-2.0 logit b 1.84 logit). In general, it can be concluded that the teacher-made chemistry test instrument has good quality items because it has met the requirements as written by Hambleton & Swaminathan (1985) that an item is said to be "good" if it has a good level of difficulty, namely -2 logit b + 2 logs.

Using the same criteria, the person fit statistics in Table 2 are interpreted. Table 2 presents statistics on MNSQ outfits and special abilities (θ) for respondents who were included in the person misfit. Based on Table 2, it can be seen that the MNSQ outfit statistics ranged from 1.53 to the maximum (infinity); while the students' abilities ranged from -1.53 logit to 5.47 logit. The results are 18 students who were included in the person misfit, namely students with serial numbers 7, 16, 26, 34, 36, 60, 63, 70, 77, 96, 98, 100, 125, 136, 145, 232, 349, and 353. This indicates that the ability of the 18 students has a response pattern that cannot be predicted by the model (Smith, 2001). Whereas through the response pattern, the accuracy of each student's response to each item can be described (Sumintono & Widhiarso, 2015). One way to identify the cause of a person misfit is through the Guttman matrix or scalograms. The Guttman matrix can provide valuable information because the item questions have been ordered from the easiest item (number 23) to the hardest item (number 18). This matrix can also show the unidimensionality of the data (Hambleton & Swaminathan, 1985). The following presents the identification of five examples of students who were classified as person misfit based on the Guttman matrix in Figure 1.

Table 2

Summary of MNSQ Outfit Statistics and Ability (θ) for Person Misfit

Subject Code	MNSQ	θ	Subject Code	MNSQ	θ
7	1.62	2.11	96	1.88	1.37
16	2.26	2.11	98	1.53	-0.11
26	1.53	0.14	100	max	5.47
34	1.55	1.53	125	2.82	2.99
36	1.69	2.64	136	2.35	-1.72
60	3.54	2.99	145	1.54	0.01
63	3.70	3.46	232	1.65	0.39
70	1.60	3.46	349	2.06	-1.05
77	1.62	-1.53	353	1.55	-0.24

Based on Figure 1, it can be stated that students with serial numbers 63, 70, 60, and 16 were classified as person misfit with the Rasch model. This is because these students had unusual response patterns, which were able to answer correctly on difficult items (item number 18) but answered incorrectly on relatively easy items (previous items). If it is based on the definition of the Rasch model, students with lower abilities will not have the opportunity to solve questions that have a higher level of difficulty, then it can be stated that the answers given by these students are possible guesses that happen to be right (lucky guessing) or even the result of cheating. This is because students were only able to give correct responses to items that have a level of difficulty below their ability level ($\theta > b$).

The identification results are in line with Meijer (1996) and Karabatsos (2003) who stated that there are at least five things that cause a person misfit, namely cheating (for example, copying answers from other test-takers) referring to unfair behavior for correct answers to items that he actually cannot answer correctly; careless responding occurs when the test-taker correctly answers the difficult items but in an unclear way answers incorrectly on the easy items; lucky guessing occurs if the test-taker correctly guesses an item that does not know the correct answer; creative responding only occurs when the test-taker has a high ability to respond incorrectly to an item which is actually easy because the test-taker interprets the item uniquely and creatively, and random responding refers to a situation where the test-taker chooses a multiple-choice option at random in responding to a question item.

Figure 1
Guttman Matrix

```
GUTTMAN SCALOGRAM OF RESPONSES:
Person |Item
       |213 2 2 13 32 12   42231113233231123131
       |3503928804414931715602587499765032667128
       |----------------------------------------
100    +1111111111111111111111111111111111111111  100
63     +1111011111111111111111111111111011111111   63
70     +1111111111110111111111111111111111110111   70
60     +1111101111011110111111111111111111111111   60
16     +1101111110111111111111110110111011110111   16
```

Meanwhile, the student with serial number 100 was also identified by the Rasch model as a person misfit because he or she managed to answer 40 items correctly (getting an extreme score) so that the fit statistic could not be measured (overfit). According to Meijer & Sitsma (2001), the measurement of person fit not only identifies an unlikely response pattern but also a response pattern that is too likely. The Rasch model predicts that uncertainty and too much certainty actually show a constraint on the response.

The information related to the reliability of the estimated test scores using the Rasch model shows that reliability is included in the good category. This is based on two reliability coefficients that are special characteristics of the results of the analysis using the Rasch model, namely person reliability and item reliability which are interpreted as the same as the reliability coefficients on the classical measurement (e.g. KR-20 or Cronbach Alpha) (Linacre, 2009). A value closer to 1 indicates a more internally consistent measure (Boone, Staver, & Yale, 2014).

There were two statistics of person reliability and item reliability that were raised by the results of the analysis of the Rasch model, namely the person reliability model and the item reliability model which provided an upper limit on the reliability of persons and items, and real person reliability and item reliability which provided a lower limit on the reliability of persons and items (Boone, Staver, & Yale, 2014). In this study, the coefficients of the model person reliability and real person reliability are 0.86 and 0.85, respectively; while the model coefficient of item reliability and real item reliability are both 0.99. The coefficient has a very high value ($\alpha > 0.900$), so it has a very good internal consistency (Tavakol & Dennick, 2011). A test that has good internal consistency will give relatively the same measurement results even though it is tested on different subjects. This is because the test is not influenced by the characteristics of the test-taker.

Conclusion

Based on the results of research and discussion, it can be concluded that all the constituent items of the teacher-made chemical test instrument are proven to be fit with the Rasch model. In addition, the test has good reliability in terms of person and item. The item difficulty level analysis also supports that the items in the teacher-made chemistry test have good psychometric characteristics. Meanwhile, 18 students were identified to be included into the person misfits. Based on these results, this teacher-made chemical test instrument can be used to assess students' learning outcomes, especially for the try-out of the national final exam in chemistry in high school. In addition, students who were identified as person misfits should be further examined for teacher guidance.

References

Adedoyin, O.O., & Mokobi, T. (2013). Using IRT Psychometric Analysis in Examining the Quality of Junior Certificate Mathematics Multiple Choice Examination Test Items. *International Journal of Asian Social Science*, 3(4). 992-1011.

Al-khadher, M.M.A., & Albursan, I.S. (2017). Accuracy of Measurement in the Classical and the Modern Test Theory: An Empirical Study on a Children Intelligence Test. International *Journal of Psychological Studies*, 9(1), 71-80.

Ashraf, Z.A., & Jaseem, K. (2020). Classical and Modern Methods in Item Analysis of Test Tools. *International Journal of Research and Review*, 7(5), 397-403.

Bond, T.G., & Fox, C.M. (2015). *Applying the Rasch Model Fundamental Measurement in the Human Sciences (3rd ed.)*. New Jersey: Lawrence Erlbaum Associatcs.

Boone, W.J., & Noltemeyer, A. (2017). Rasch Analysis: A primer for School Psychology Researchers and Practitioners. *Cogent Education*, 4(1), 1-13.

Boone, W.J., Staver, J.R., & Yale, M.S. (2014). *Rasch Analysis in the Human Sciences*. Dordrecht: Springer.

DeMars, C. (2010). *Item Response Theory: Understanding Statistics Measurement*. New York: Oxford University Press, Inc.

Hambleton, R.K., & Swaminathan, H. (1985). *Items Response Theory: Principles and Application*. Boston: Kluwer-Nijjhoff Publish.

Karabatsos, G. (2003). Comparing the Abberant Response Detection Performance of Thirty-Six Person-Fit Statistics. *Applied Measurement in Education*, 16(4), 277-298.

Irmayta, E., Rudibyani, R. B., & Efkar, T. (2018). Pengembangan Instrumen Asesmen Pengetahuan pada Materi Asam Basa Arrhenius. *Jurnal Pendidikan dan Pembelajaran Kimia*, 7(1), 63-76

Jayanti, E. (2020). Instrumen Tes Higher Order Thinking Skills pada Materi Kimia SMA. *Orbital: Jurnal Pendidikan Kimia*, 4(2), 135-149.

Linacre, J.M. (2002). What do Infit and Outfit Mean-Square and Standardized Mean?. *Rasch Measurement Transaction*, 16, 878.

Linacre, J.M. (2009). *A User's Guide to WINSTEPS*. Chicago: Winsteps.

Meijer, R.R., & Sitsma, K. (2001). Person Fit Statistic: What is Their Purpose?. *Rasch Measurement Transactions*, 15(2), 823.

Muna, W., Hanafi, H., & Rahim, A. (2019). Analisis Kualitas Tes Buatan Guru Mata Pelajaran Bahasa Indonesia pada Siswa SMP Kelas IX Berbasis HOTS. *Jurnal Pendidikan Bahasa*, 8(2), 29-40.

Nugraheni, Y.P., Widodo, A.T., & Sugiyo, W. (2013). Kualitas Soal Buatan Guru dan Dinas Pendidikan Bidang Studi Kimia. *Chemistry in Education*, 2(1), 1-8.

Nurhayati, E., & Ahmad, T. A. (2018). Implementasi Penilaian Autentik dalam Pembelajaran Sejarah di SMA Negeri 1 Semarang. *Indonesian Journal of History Education*, 6(1), 21–30.

Prabayanti, N.M.D., Sudiana, I.K., & Wiratini, N.M. (2018). Analisis Tes Ulangan Kenaikan Kelas Buatan Guru Mata Pelajaran Kimia. *Jurnal Pendidikan Kimia Indonesia*, 2(1), 1-10.

Purba, R. A. Br., Susanti, N., & Rosna (2019). Analisis Butir Soal Ujian Semester Ganjil Kimia Kelas X SMA Negeri1 Perbaungan. *Jurnal*

Tavakol, M., & Dennick, R. (2011). Making Sense of Cronbach's alpha. *International Journal of Medical Education*, 2, 53-33.

Tosuncuoglu, I. (2018). Importance of Assessment in ELT. *Journal of Education and Training Studies*, 6(9), 163-167.

Wiyarsi, A., Fachriyah, A. R., Supriadi, D., & Ibrahim, M. (2019). A Test of Analytical Thinking and Chemical

Inovasi Pendidikan Kimia, 1(1), 38-43,

Rasch, G. (1980). *Probabilistic Models for Some Intelligence and Attainment Test*. Chicago, IL: University of Chicago Press.

Runnels, J. (2012). Using the Rasch Model to Validate a Multiple Choice English Achievement Test. *International Journal of Language Studies*, 6(4), 141-153.

Sakinah, P., & Ritonga, P.S. (2017). Analisis Butir Soal Ujian Semester Mata Pelajaran Kimia Kelas X Madrasah Aliyah di Kecamatan Pasir Penyu. *Konfigurasi: Jurnal Pendidikan Kimia dan Terapan*, 1(1), 129-137.

Sopiah, A., Sidauruk, S., & Asi, N.B. (2019). Kualitas Soal Penilaian Akhir Semester (PAS) Buatan Guru Mata Pelajaran Kimia Kelas XI IPA SMA Negeri di Kabupaten Seruyan pada Semester Ganjil Tahun Ajaran 2018/2019. *Jurnal Ilmiah Kanderang Tingang*, 10(2), 110-126.

Sumintono, B., & Widhiarso, W. (2015). *Aplikasi Permodelan Rasch Pada Assessment Pendidikan*. Cimahi: Trim Komunikata.

Representation Ability on 'Rate of Reaction' Topic. *Cakrawala Pendidikan*, 38(2), 228-242.

Wright, B., & Stone, M. (1999). *Measurement Essentials (2nd ed.)*. Wilmington: Wide Range, Inc.

Yilmaz, H.B. (2019). A Comparison of IRT Model Combinations for Assessing Fit in a Mixed Format Elementary School Science Test. *International*

Electronic Journal of Elementary Education, 11(5), 539-545.

Yustika, A., Susatyo, E. B., & Nuswowati, M. (2014). Uji Kriteria Instrumen Penilaian Hasil Belajar Kimia. *Jurnal Inovasi Pendidikan Kimia*, 8(2),1330-1339.