ASIA PACIFIC JOURNAL
OF EDUCATORS AND EDUCATION

Volume 37, Number 2, 2022
https://doi.org/10.21315/apjee2022.37.2.2

Research Article:

Computerised Testlet Instrument to Assess Students' Conceptual Understanding of Chemistry on the Topic of Stoichiometry: Psychometric Analysis

Sri Yamtinah*, Shifi Syarifa Fahmina, Dimas Gilang Ramadhani, Sulistyo Saputro and Ari Syahidul Shidiq
Chemistry Education, Faculty of Teacher Training and Education, Universitas Sebelas Maret,
Jl. Ir. Sutami No.36, Kentingan, Kec. Jebres, Kota Surakarta, Jawa Tengah 57126, Indonesia

*Corresponding author: jengtina@staff.uns.ac.id

ABSTRACT

Measuring students' conceptual understanding is a challenge for teachers. Teachers usually use open-ended instruments to measure students' conceptual understanding in depth. However, this instrument is considered highly subjective in the assessment and requires a relatively long time to check the answers. Computerised Testlet instruments developed by combining the advantages of open-ended and closed-ended instruments are deemed the solution to these problems. This study, therefore, aims to analyse the quality of the Computerised Testlet instrument psychometrically. The Rasch analysis model with Winsteps software was used in this study. A total of 10 Testlet items with three supporting questions were developed on the topic of stoichiometry. A total of 413 students (N = 413: 236 female, 177 male) from three different schools took tests on stoichiometry using the Testlet instrument, which was provided through a web developed previously (computerised). Rasch analysis model helped psychometric analysis regarding reliability, linear validity of students' knowledge, and difficulty of items. This study reveals that Computerised Testlet able to identify students' thinking processes. Besides, the psychometric analysis results showed that using Testlet to measure conceptual knowledge had good unidimensionality and reliability; it could be used to measure stoichiometric conceptual knowledge. Some questions needed to be revised as they proved unable to distinguish between high and low students' conceptual knowledge; in general, the instrument could be used and provided a good function in analysing conceptual knowledge on the topic of stoichiometry.

Keywords: Computerised Testlet, conceptual understanding, stoichiometry, Rasch analysis

Published: 30 December 2022

To cite this article: Sri Yamtinah, Shifi Syarifa Fahmina, Dimas Gilang Ramadhani, Sulistyo Saputro, & Ari Syahidul Shidiq. (2022). Computerised Testlet instrument to assess students' conceptual understanding of chemistry on the topic of stoichiometry: Psychometric analysis. *Asia Pacific Journal of Educators and Education, 37*(2), 5-29. https://doi.org/10.21315/apjee2022.37.2.2

INTRODUCTION

Educational assessment has transformed in the last few decades. This transformation impacts not only the assessment on a large scale but also the assessment in the classroom. Three factors contribute to the transformation: changes in general education goals, the relationship between assessment and learning, and the limitations of assessment methods for reporting the students' conceptual understanding and skills (Brookhart, 2005; Kelting-Gibson et al., 2014; Marzano et al., 1993).

Another opinion suggests that the transformation of assessment in science education is based on the assumption of decomposability and decontextuality of science, which results in a system of judgements with simple inference; it is often referred to as "traditional assessment" (Klassen, 2006). This assessment model was replaced by an assessment that pays more attention to cognitive psychology, the development of the philosophy of science, and the development of constructivism theory. It is primarily about cognitive processes in students who emphasise the context in learning and assessment. These changes led to the development of various methods of contextual assessment in science education (Klassen, 2006; Taylor & Watson, 2000).

The recent study on educational assessment focused on the interaction between classroom assessment and the lack of use of assessment instruments associated with the experience acquired by students in the classroom. It becomes the research basis on classroom assessment to improve and develop classroom assessment that can significantly contribute to improving learning quality (Black & Wiliam, 1998). In addition, technological developments in the 21st century encourage technology integration with learning and assessment processes (Yamtinah, Saputro, et al., 2019).

Technology-based assessment is in line with the expected transformation in science education. Contextualisation of the learning process that aligns with the assessment can be done with technology. In addition, it makes it easier for classroom teachers to carry out the process of assessing students' knowledge and skills (Fahmina et al., 2019; Yamtinah et al., 2021; Yamtinah, Indriyanti, et al., 2019). Other advantages of technology-based assessment are that it can automatically provide diagnostic information from students' knowledge and skills, can determine the quality of items, and is more efficient because it can be used many times (Griffin & Care, 2015; Kapsalis, 2009; Khlaisang & Koraneekij, 2019).

Amid the transformation of science education and all the advantages provided by technology-based assessment, it turns out that many science class teachers still preferred to use paper and pencil-type tests to conduct assessments. Moreover, the form of the instrument used is inseparable from traditional multiple-choice (close-ended) and essay (open-ended) (Shidiq et al., 2016; Yamtinah et al., 2017). The discrepancy between the expectations of transformation in science education and the teacher's reliance on conventional instruments is one of the problems raised in this study.

Classroom teachers, moreover, tend to have limited information about innovative forms of assessment instruments. Therefore, they chose the traditional multiple-choice and essay forms of instruments. Either multiple-choice or essay both have its advantages and disadvantages. Multiple-choice assessment instruments have advantages, such as faster assessment times and fewer subjective responses to students' answers. However, the disadvantages are that this instrument cannot identify students' thinking processes to solve the given problem, the high factor of guessing answers students, and more likely students to cheat (Gurel et al., 2015; Shidiq et al., 2019; 2016; Yamtinah et al., 2021). Meanwhile, the essay (open-ended) assessment instruments have the advantage of discovering students' thinking processes in solving given problems. In addition, the form of an open-ended instrument also allows students to explore their answers. Nevertheless, as a disadvantage, this instrument takes longer and has a relatively large error rate when scoring. The high subjectivity factor of the teacher who scores is the main disadvantage (Fahmina et al., 2019; Hecht et al., 2017; Wainer & Kiely, 1987).

LITERATURE REVIEW

Computerised Testlet Assessment

The disadvantages of multiple-choice (close-ended) and essay (open-ended) assessment instruments, as well as the demands of technology-based assessments instrument that can accurately assess students' knowledge and skills, have then led to the need for new forms of assessment instruments that can combine the advantages of multiple-choice and essays and technology-based assessments. Therefore, a computerised form of the Testlet instrument was developed to combine all these advantages. A Testlet instrument is a group of items (questions) related to a particular topic set into a single unit and contains several predetermined steps. The Testlet is included in the type of test that produces more than one response. Furthermore, this Testlet has a relatively hierarchical answer about the knowledge (construct) to be measured (Huang & Wang, 2012; Luecht et al., 2006; Shidiq et al., 2016; Wainer et al., 2013).

The Computerised Testlet assessment instrument further provides an opportunity for the teacher not only to ask questions to students but also to describe the contextualisation of the problems to be asked to students through the "stem" or main questions in the Testlet instrument construction. Then, this stem or main question will become the problem context to be asked hierarchically in supporting questions. In this way, students' thinking processes are hierarchically identified while still simplifying the assessment process like traditional multiple choice (Chang & Wang, 2010; Lutviana et al., 2019; Wang & Wilson, 2005; Xiaohui et al., 2002). Because the Testlet has supporting questions that distinguish it from traditional multiple-choice, the graded response model is used as a guide for scoring using this Computerised Testlet instrument (Chang & Wang, 2010; Zhu & Stone, 2012).

Computerised Testlet Assessment and Stoichiometry

In this study, a Computerised Testlet assessment instrument was developed on the topic of stoichiometry. This topic was chosen because it has the characteristics of hierarchical knowledge and becomes the foundation for other topics in chemistry (Cacciatore & Sevian, 2006). In addition, this topic is considered complex by students (Mahaffy, 2006). Stoichiometry includes simple reaction equations, basic chemical laws, and chemical calculations. There have been many studies on stoichiometry topics in chemistry education, such as the use of blogs to improve students' stoichiometry knowledge (Le Maire et al., 2018) and the effect of hands-on activities on improving stoichiometry knowledge (Ajayi, 2017). However, there is no reported work on using the Computerised Testlet assessment instrument to determine students' conceptual knowledge of the topic of stoichiometry. In fact, the characteristics of the testlet which has a structured problem with several layers are hierarchically compatible with the characteristics of the stoichiometric material which is also hierarchical (Cacciatore & Sevian, 2006; Chang & Wang, 2010)

This study, therefore, administered a Computerised Testlet instrument developed previously for students to determine their conceptual knowledge of stoichiometry. The assessment instrument developed required an analysis of item quality. It was done to ensure the instrument had good construction and reliability and could measure what it was supposed to measure. Hence, this study aimed to analyse the quality of the Computerised Testlet assessment instrument's items on a psychometrically stoichiometry topic.

Psychometric Analysis

An instrument's validity and reliability analysis is part of the psychometric analysis (Finney, 2007). Teachers and researchers usually conduct psychometric analysis using a Classical Test Theory (CTT). CTT is preferred as it has a simpler mathematical equation and a more moderate sample size (Tangio, 2019). Almost all psychometric analyses of an assessment instrument on chemistry learning used CTT to determine the data validity and reliability (Adams & Callahan, 1995; Bolarinwa, 2015; Chen & Liu, 2020; Currell & Jeukendrup, 2008; Feyzíoglu et al., 2012; Garner, 2005). In CTT, the item parameter seen is only the quality of the instrument item without looking at the respondent who is working on the instrument; consequently, of course, very little information will be obtained, and the diversity of the data obtained will be minimal (Hambleton & Jones, 1993).

The use of modern tests such as Item Response Theory (IRT) has then begun to be used to see in more detail the instrument's ability to measure student understanding. Rasch's analysis of psychometric assessment in chemistry and science has been reported by many researchers (Lu & Bi, 2016; Syang & Dale, 1993). The Rasch model can also describe the distribution of a person's abilities and their relationship to the test items used. Thus, it is interesting for this study to use an IRT approach, such as the Rasch model analysis, to analyse the psychometric aspects of the Computerised Testlet assessment instrument in measuring students' conceptual understanding of stoichiometry.

METHODOLOGY

Research

In this quantitative research, before testing, the instrument underwent the expert validation stage and was tested on a trial scale. This study focused on the psychometric analysis of the Computerised Testlet used to measure students' conceptual understanding of chemistry on the topic of stoichiometry. As shown in Figure 1, obtained data would be analysed for item quality, reliability and validity. Validity included unidimensional and local independent tests. The first was conducted to see whether the instrument could measure something latent or, in this study, was the ability to understand students' chemical concepts on the topic of stoichiometry. The latter local independent test was to see the instrument independence, where each item did not affect other items. Meanwhile, reliability comprised item reliability to determine whether the limit was found and people reliability to know whether the sample selection used to measure the instrument's quality was correct. Then, Cronbach's Alpha measures the reliability of a construct, whereas composite reliability measures the reliability of the value of a construct. In this regard, composite reliability is considered to be better in estimating the internal consistency of a construct. Moreover, item quality includes an item fit test to determine the quality of the items, Wright's map to see the distribution of item and person data, and distractor analysis to see how well the distractors work and analyse the log distribution of each item.

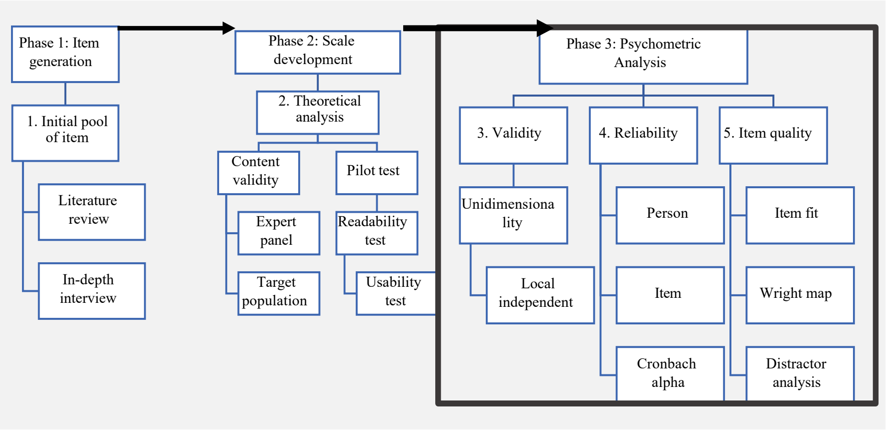


Figure 1. Research design

Instrument

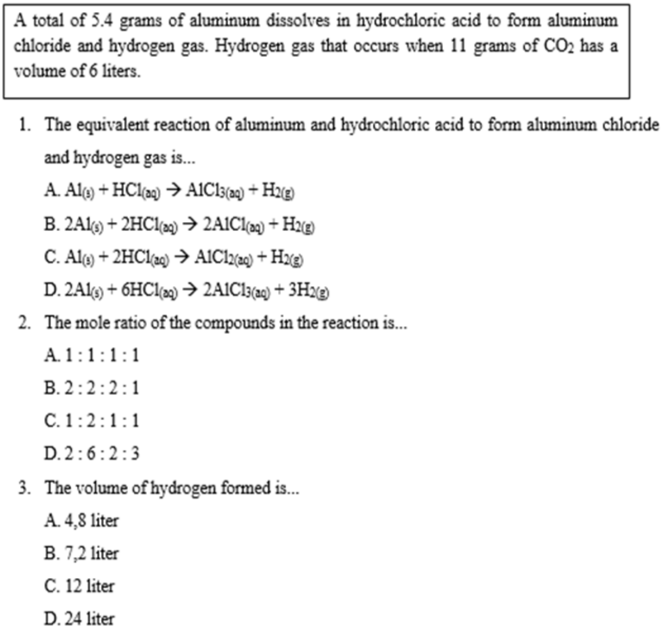
Developing an instrument requires several phases and many steps that need methodological and theoretical decisions. Classical (Nunnally & Bernstein, 1994) and contemporary (Boateng et al., 2018; DeVellis, 2017) instrument development illustrates that instrument development involves several phases of preparation and testing to obtain a good instrument.

Almost all research and development carried out more than five steps, the most common of which are initial product development, product validation and testing. In this research, the researchers developed the instrument described by Boateng et al. (2018) and followed a procedure consisting of three phases and a total of five steps, as shown in Figure 1.

In the first phase, the construction of the Computerised Testlet development began by looking for topics on the chemical concepts that caused the most problems for students. The stoichiometry was chosen based on an in-depth interview and study literature, which provided data that this concept has a high difficulty level. The construction of basic competencies related to stoichiometric topics was carried out to determine competency and question indicators. The Testlet was also developed based on the needs obtained from the in-depth interview process and the literature review carried out. The second phase, the development of instrument and application software, was carried out. The panel of experts consisted of three professors in the chemistry teaching field, two experts in chemistry learning assessment, two experts in chemistry learning content, and seven chemistry teachers with more than 15 years of teaching experience. The expert panel checked the content validity of the proposed items, leading to a test of mastery of chemical concepts in the topic of stoichiometry. Next was a pilot test consisting of a readability test and a Testlet application usability test. The readability of items included language and graphics displayed in the Testlet instrument. Meanwhile, the usability test comprised the ease of operating the application and the problems that arose during the use of the application. In the third and final phase, the items were tested on high school students, and then the tests' psychometric results were analysed.

Two general concepts often used in factor Testlet models are independent of each item and multidimensional. Independent Testlet means that each item developed is not related to other items. On the other hand, by using the multidimensional concept, each developed item can be linked to other items. This multidimensional concept makes more sense when applied to several items related to the context but not made directly from each other (DeMars, 2012). For example, it is when presented with phenomena or events related to chemical concepts. For instance, in the dissolution of aluminium in hydrochloric acid, producing aluminum chloride and hydrogen gas, writing down the equation for the occurring reaction is asked in the first level. In the next level, the mole ratio of each compound in the reaction is determined. At the third level, the volume of hydrogen gas formed is determined. Levels 2 and 3 cannot be answered when the items in the first level are not solved correctly because the equivalent reaction determines the mole ratio of each compound in the reaction. Thus, it is impossible to determine if one cannot solve the equivalent reaction and the mole ratio of each compound in the reaction in the third level. Figure 2 is an example of a Testlet item.

In this study, all data were collected using a Computerised Testlet instrument. The Computerised Testlet instrument contained 10 Testlet items, and the participants were gathered at the computer centre of each school and took turns working on the items given. Participants were also given 90 minutes to work on the questions. The instrument Testlet was distributed randomly in ten rods, each consisting of three graded items. The first item was a basic item to investigate students' basic knowledge, a bridge in working on the second item, which was the basis for the third item. Therefore, the items in the Computerised Testlet instrument are called graded items.



A total of 5.4 grams of aluminum dissolves in hydrochloric acid to form aluminum chloride and hydrogen gas. Hydrogen gas that occurs when 11 grams of CO_2 has a volume of 6 liters.

1. The equivalent reaction of aluminum and hydrochloric acid to form aluminum chloride and hydrogen gas is...
   A. $Al_{(s)} + HCl_{(aq)} \rightarrow AlCl_{3(aq)} + H_{2(g)}$
   B. $2Al_{(s)} + 2HCl_{(aq)} \rightarrow 2AlCl_{(aq)} + H_{2(g)}$
   C. $Al_{(s)} + 2HCl_{(aq)} \rightarrow AlCl_{2(aq)} + H_{2(g)}$
   D. $2Al_{(s)} + 6HCl_{(aq)} \rightarrow 2AlCl_{3(aq)} + 3H_{2(g)}$
2. The mole ratio of the compounds in the reaction is...
   A. 1 : 1 : 1 : 1
   B. 2 : 2 : 2 : 1
   C. 1 : 2 : 1 : 1
   D. 2 : 6 : 2 : 3
3. The volume of hydrogen formed is...
   A. 4,8 liter
   B. 7,2 liter
   C. 12 liter
   D. 24 liter

Figure 2. A Testlet with one stem consisting of three layers of questions

Participants

The participants of this study were 413 high school students in Surakarta City, consisting of 236 female students and 177 male students with an age range of 15–17 years old. These students were from three different schools. The Testlet computerised assessment instrument was given to students after they had completed four lesson meetings related to the topic of stoichiometry. Students also took this test in their school's computer laboratory, done online using a web-based Computerised Testlet assessment instrument system developed previously. All participants have also agreed to participate in this research and were willing if their test results, which were part of this research, to be published.

Data Analysis

Student responses were collected using a web-based Computerised Testlet instrument—the data were obtained in the form of answers to each number of questions provided. The score was based on graded scoring as revealed by the Graded Response Model (GRM) (Zhu & Stone, 2012) scoring guidelines with details in Table 1.

Table 1. Computerised testlet instrument scoring guidelines

No	Assessment aspect	Score
1.	The answer to the first step is wrong.	0
2.	The answer in the first step is correct but wrong, or no answer in the second and third steps.	1
3.	The answer is correct in the first and second steps but wrong in the third.	2
4.	The answer in all steps is correct.	3

The results of student responses were then analysed using Rasch modelling in polytomy data analysis through Winsteps software. Rasch modelling has benefits in the psychometric evaluation of assessment data. The Rasch model also can convert raw scores into a measure of ability based on equidistant logit. This model can estimate the difficulty of items on the same interval scale as the estimated ability of students so that items can be evaluated. That way, the rater can see how well the item's difficulty level matches the student's ability being measured (Ahmad & Siew, 2021; Pentecost & Barbera, 2013). The data obtained from the testing process of the Rasch model included the value of reliability, construct validity, student abilities, item difficulty and item fit.

In addition, the instrument testing with the Rasch model must meet the requirements of unidimensionality and local independence. Factors also influence how this test works to estimate the person's level of understanding. This test explains how the instruments developed can measure what should be measured. Unidimensionality in this measurement could determine other components with latent properties (conceptual understanding of chemistry on stoichiometry). Unidimensionality also analysed the conformity level of the main features with local independence. In the figure showing the unidimensionality test results, the researchers could find data related to 69.4% of raw variance explained by measures and leaving 30.4% of raw variance unexplained. Overall, it is considered that the data met unidimensionality and local independence requirements, providing evidence of the instrument's construct validity (Linacre, 2011). Therefore, the researchers could use this instrument to measure the conceptual understanding of stoichiometry (Sumintono & Widhiarso, 2014).

Moreover, this test's reliability is divided into two index indicators: items and persons. It contains reliability according to classical theory through Cronbach's alpha of 0–1. This coefficient measures the consistency of a person's answer to the items (questions) of the assessment instrument. Then, the Testlet instrument consisting of ten bars (questions) can be categorised as having an "acceptable" internal consistency if it has a range of 0.70 and 0.80, "good" consistency between 0.80 and 0.90, and "very good" consistency if it is above 0.90. Therefore, this Testlet instrument has a value compared to these criteria to measure the consistency level that can measure items and people.

The item difficulty and separation analysis are shown on the Wright map. The Rasch model relates to the response of the person, both of which are on the log-odds (logit) scale in the same place. Logit placement and scale allow for direct analysis of person and item capabilities. The placement of zero (0) logit is the basis that the ability of the person and item difficulty is at an average level and will be higher in accordance with the increase in the logit value. Besides, the logit value with a negative value (−) indicates that the difficulty level gets lower; thus, the person's ability is lower too. The symbol M indicates the placement of the mean or average, (S) represents one standard deviation, and (T) means two standard deviations (Jin et al., 2020; Rachmatullah et al., 2017).

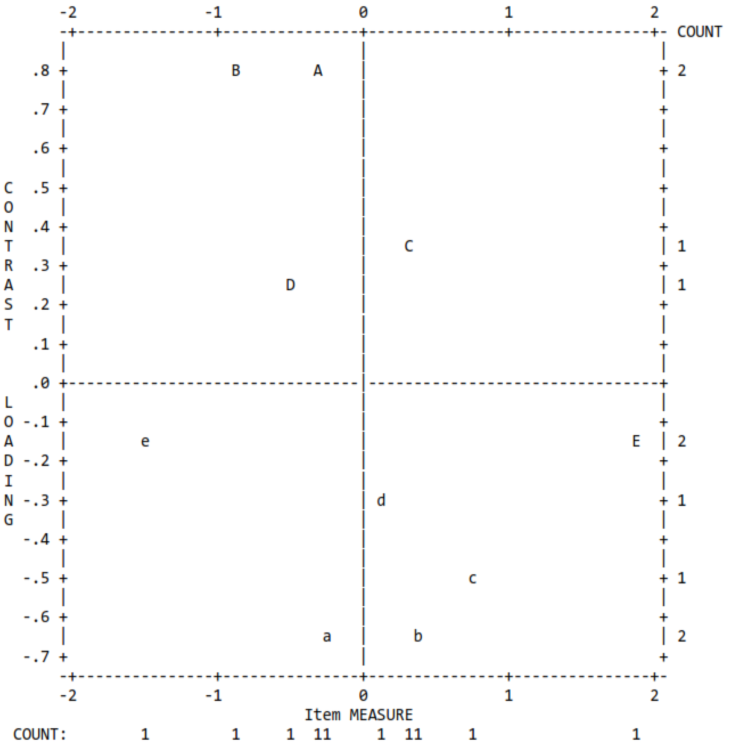
Furthermore, item fit indicates the expected level between the actual item characteristics and the characteristics of the Rasch model. Items are categorised as fit with the model when the MNSQ outfit value is 0.5 to 1.5, the ZSTD outfit value is in the -2 to +2 range, and the correlation measurement point value is in the 0.4 to 0.85 range (Wei et al., 2012).

RESULTS

Validity and Reliability of Measures

The Rasch model analysis is based on unidimensional and locally independent. Unidimensional has the assumption that items measure something latent; this study measured conceptual understanding of chemistry on the topic of stoichiometry. Besides, being locally independent means that one item's success or failure is unrelated to other items. In this research, the Testlet consists of several stems in one stem, comprising three interrelated multiple-choice questions. This analysis could not use a dichotomy but should use the measurement of the polytomies of the three questions in one stem. Here, the relationship between each question in the stem, of course, has a connection, therefore in local independent analysis, it cannot be used with a dichotomy with the intention that this analysis does not violate the assumptions of the Rasch model about local independence (Deng & Wang, 2017).

Local independence was then evaluated, and the data found no significant correlation between items. The most considerable correlation was found to have a value of 0.14; this value can indicate that the Testlet theoretically did not violate the assumptions of the Rasch model, while according to the standard, items with a correlation >0.7 are categorised as having a high local dependent. Local independence shows that its success does not influence the subject's success in overcoming other problems. Local independence also occurs when the subject's success in overcoming a problem is due to his ability, not caused by other factors (Taskin et al., 2015).



```
     -2             -1              0              1              2
     -+--------------+--------------+--------------+--------------+- COUNT
      |                             |                             |
   .8 +                B       A    |                             + 2
      |                             |                             |
   .7 +                             |                             +
      |                             |                             |
   .6 +                             |                             +
      |                             |                             |
C  .5 +                             |                             +
O     |                             |                             |
N  .4 +                             |                             +
T     |                             |    C                        | 1
R  .3 +                             |                             +
A     |                      D      |                             | 1
S  .2 +                             |                             +
T     |                             |                             |
   .1 +                             |                             +
      |                             |                             |
   .0 +-----------------------------|-----------------------------+
L     |                             |                             |
O -.1 +                             |                             +
A     |       e                     |                           E | 2
D -.2 +                             |                             +
I     |                             |                             |
N -.3 +                             | d                           + 1
G     |                             |                             |
  -.4 +                             |                             +
      |                             |                             |
  -.5 +                             |          c                  + 1
      |                             |                             |
  -.6 +                             |                             +
      |                         a   |     b                       | 2
  -.7 +                             |                             +
     -+--------------+--------------+--------------+--------------+-
     -2             -1              0              1              2
                              Item MEASURE
COUNT:        1        1     1 11     1  11    1                1
```

Figure 3. The plot of item loading

The reliability value can be seen in Table 2. The test results obtained an item reliability value of 0.98, personal reliability of 0.89, and Cronbach Alpha reliability of 0.95. These values indicate that the items in the Computerised Testlet instrument had good consistency when used from time to time (Korayem et al., 2017; Fahmina et al., 2019) content validity was applied to the computerised testlet instrument. Computerised testlet is a group of multiple-choice items to reveal the same information developed in a computerised system. The computerised testlet instrument in this study consists of sevens stem (the subject matter). In other words, the Computerised Testlet instrument will produce the same information when carried out by several researchers with the same object (Nedungadi et al., 2019). The second reliability value was the person reliability of 0.89, included in the good category. This value denotes that the respondents used in this study had good answer consistency. The third reliability value was the Cronbach Alpha reliability of 0.95, showing that the interaction between a person and an item was very good overall (Sabah & Hammouri, 2009).

Table 2. Summary statistics of person and item

Parameter	Separation	Reliability	Cronbach Alpha
Item (10)	2.31	0.98	0.95
Person (413)	2.88	0.89	

Item Difficulty and Person Ability

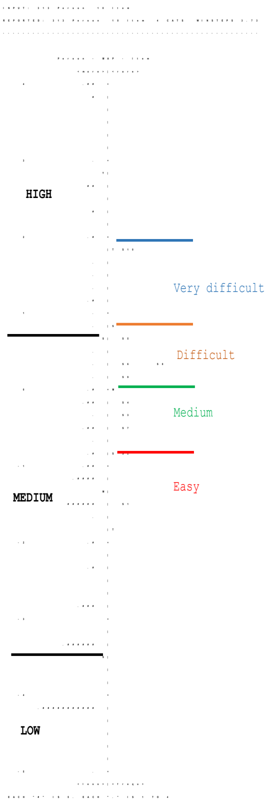


Figure 4. Wright map Testlet, containing logit distribution data for the person and item difficulty level

Item separation shows the grouping of questions, while person separation demonstrates the group of students' abilities. As shown in Figure 4, the left side of the vertical line is the distribution of students' abilities, and the right side is the distribution of item difficulty levels. For example, the Computerised Testlet instrument tested resulted in a person separation value of 2.80. This value can be rounded up to 3, meaning that the Computerised Testlet instrument can classify students' conceptual understanding of chemistry skills into three levels: high ability, medium ability and low ability.

In Figure 4, it can also be seen that students' conceptual understanding of chemistry abilities varied but was dominated by moderate abilities, clustering in the middle, from −3.5 logs to 0.5 logs. Thus, while students with high abilities are at the top logit, students with low abilities are at the bottom. In addition, the value of stem separation generated through Rasch modelling was 2.31, which was then simplified to 3. This value means that the Computerised Testlet instrument consisted of 10 items with high, medium, and low difficulty levels. The easiest stem questions were at the bottom logit; the higher individuals go, the higher the difficulty level. Therefore, the difficult and medium levels dominated the stem of this question. Students with high abilities (at the top logit) could do all stems correctly, while students with low abilities could only solve basic questions (Nedungadi et al., 2019).

Figure 4 presents the approximate person and item maps of the Computerised Testlet instrument. Each "#" sign represents five students, and a "." signifies 1−4 students. On the left side of the vertical line is the distribution of students' conceptual understanding of chemistry (person) from high ability (top) to low ability (bottom). Meanwhile, on the right side of the line is the distribution of each item's difficulty level, which stems from easy to very difficult. The figure depicts that the distribution of students' conceptual understanding of chemistry ability spread from −5.0 logs to 4.0 logs, while the distribution of items for each stem ranged from −1.50 logs to 1.80. Items were also arranged by logit from lowest to highest logit. Items S1−S10 were estimated to be close to each other, and all items were distributed to match the level of students' conceptual understanding of chemistry.

Further, the most challenging question was in question 10; this number was in logit 1.9 and was in the area (T) or twice the standard deviation. The difficulty level of question number 10 could still be categorised as could be done by most people with high understanding abilities, and it can be seen from the symbols "#" and "." above (T). It is also necessary to look at the item suitability section to determine the next step regarding whether this type of question needs to be maintained or not suitable for measuring student understanding. On the other hand, the most simple question was in question number 1, looking at it at the lowest level with a logit value of −1.53. The researchers determined that this question had a low difficulty level, and many students could do well on this question. However, this analysis was insufficient to explain whether the item was feasible to use; thus, with item fit, there will be more information about how the item can measure stoichiometry conceptual understanding (Lu & Bi, 2016; Wei et al., 2012).

Item Fit

According to Table 3, the total fit index could be accepted according to the MNSQ Criteria, ZSTD Criteria, or point correlation criteria. Overall, according to the Rasch model and the Computerised Testlet instrument, the data were reliable and acceptable for measuring students' conceptual understanding. However, further analysis of each item separately was required for instrument optimisation. This analysis assumes that students/persons with abilities in the upper group have a higher chance of answering items correctly than those with low abilities. This analysis was used to find the relationship between the student's person's ability to use the item. Items also function to separate problematic items so they can be improved or even not used. Item fit analysis was with residual analysis, unweighted (outfit), and weighted (infit) indices (Nedungadi et al., 2019; Paek et al., 2009; Taskin et al., 2015).

Table 3. Item fit statistics

Item	Difficulty value	S.E.	Infit		Outfit		Pt. Mea Corr.
			MNSQ	ZSTD	MNSQ	ZSTD	
S1	−1.53	0.09	1.36	3.9	2.04	6.3	0.72
S9	0.10	0.10	1.54	4.8	1.62	4.1	0.78
S3	−0.32	0.10	1.36	3.5	1.01	0.1	0.81
S2	−0.85	0.09	1.17	1.8	0.97	−0.2	0.82
S10	1.90	0.13	1.03	0.3	0.49	−2.2	0.81
S4	−0.24	0.10	0.89	−1.1	0.73	−2.5	0.85
S5	0.73	0.11	0.89	−1.1	0.65	−2.4	0.84
S6	0.39	0.10	0.86	−1.5	0.67	−2.6	0.85
S8	0.31	0.10	0.70	−3.4	0.55	−3.9	0.88
S7	−0.50	0.10	0.48	−7.0	0.50	−5.3	0.90

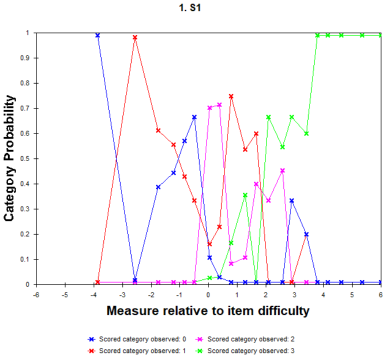
Based on the analysis of item number 1, having an MNSQ value that exceeded 1.5 beyond the specified range, the researchers could quickly conclude that this item was problematic for most students. The fit analysis on the MNSQ infit and outfit also found that this question was relatively easy and could be done by all students with low or high abilities. Further analysis was then carried out to see if this article could be used in the distractor analysis as in the figure. It shows that this question could not distinguish; students with high or low abilities could all do the questions well.

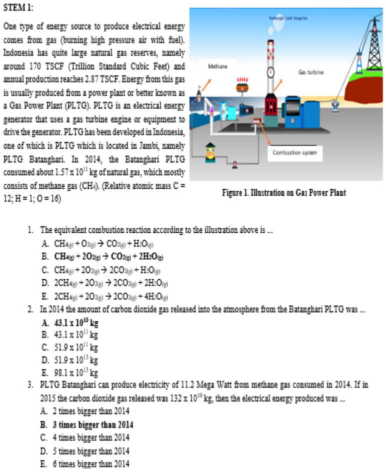
In addition, item number 9 had almost the same problem as number 1; the researchers judged that all questions in the medium category could be used, but not all questions in this category could distinguish. Item 9 also had an MNSQ value of both infit and outfit above 1.5, showing that this question was only considered easy for those with high abilities and was too difficult for those with low abilities.

Item 10, as the most challenging question, had an MNSQ infit value close to 1 and an outfit of 0.5, which was still within the allowable range. The researchers saw how the persons responded, and their understanding level would improve this analysis. It implies that each item (regardless of difficulty score) performed well enough for most students.

Moreover, ZSTD values are often used to test the hypothesis that the data fit a "perfect" model. It is recommended that MNSQ scores be evaluated before considering ZSTD scores and that "ZSTD is only useful for rescuing insignificant MNSQ [scores] when the sample size is small, or the test length is short." Related to that, items 1 and 9 no longer met the MNSQ and the ZSTD standards; this item needed to be revised (Jin et al., 2020; Lu & Bi, 2016).

Distractor Analysis




Pembangkit Listrik Tenaga Gas
Methane
Gas turbine
Combustion system

STEM 1:

One type of energy source to produce electrical energy comes from gas (burning high pressure air with fuel). Indonesia has quite large natural gas reserves, namely around 170 TSCF (Trillion Standard Cubic Feet) and annual production reaches 2.87 TSCF. Energy from this gas is usually produced from a power plant or better known as a Gas Power Plant (PLTG). PLTG is an electrical energy generator that uses a gas turbine engine or equipment to drive the generator. PLTG has been developed in Indonesia, one of which is PLTG which is located in Jambi, namely PLTG Batanghari. In 2014, the Batanghari PLTG consumed about 1.57 x 10^{11} kg of natural gas, which mostly consists of methane gas (CH_4). (Relative atomic mass C = 12; H = 1; O = 16)

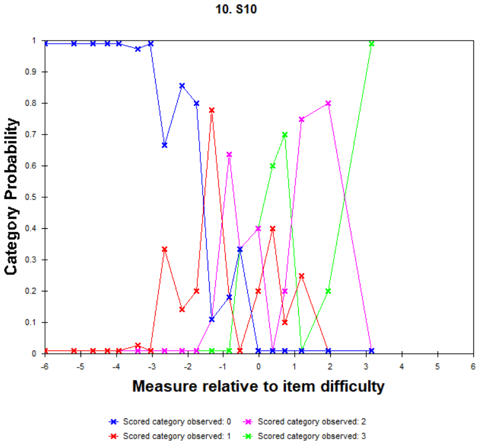
Figure 1. Illustration on Gas Power Plant

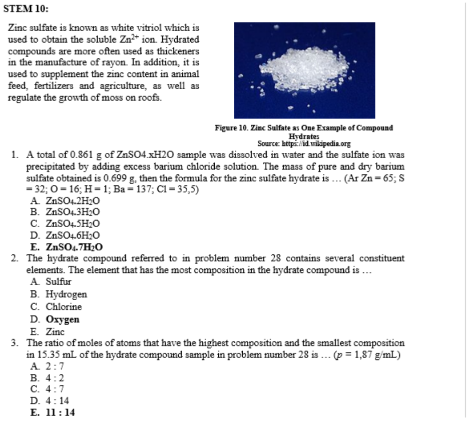
1. The equivalent combustion reaction according to the illustration above is ...
   A. $CH_{4(g)} + O_{2(g)} \rightarrow CO_{2(g)} + H_2O_{(g)}$
   B. $CH_{4(g)} + 2O_{2(g)} \rightarrow CO_{2(g)} + 2H_2O_{(g)}$
   C. $CH_{4(g)} + 2O_{2(g)} \rightarrow 2CO_{2(g)} + H_2O_{(g)}$
   D. $2CH_{4(g)} + 2O_{2(g)} \rightarrow 2CO_{2(g)} + 2H_2O_{(g)}$
   E. $2CH_{4(g)} + 2O_{2(g)} \rightarrow 2CO_{2(g)} + 4H_2O_{(g)}$
2. In 2014 the amount of carbon dioxide gas released into the atmosphere from the Batanghari PLTG was ...
   A. 43.1 x 10^{10} kg
   B. 43.1 x 10^{11} kg
   C. 51.9 x 10^{11} kg
   D. 51.9 x 10^{13} kg
   E. 98.1 x 10^{13} kg
3. PLTG Batanghari can produce electricity of 11.2 Mega Watt from methane gas consumed in 2014. If in 2015 the carbon dioxide gas released was 132 x 10^{10} kg, then the electrical energy produced was ...
   A. 2 times bigger than 2014
   B. 3 times bigger than 2014
   C. 4 times bigger than 2014
   D. 5 times bigger than 2014
   E. 6 times bigger than 2014

Figure 5. Distractor analysis stem number 1

Items in stem 1 discuss the reactions that occur in the combustion process of hydrocarbon compounds. Stem 1 had the problem with the lowest level of difficulty. A stem set consisted of three questions related to one another, where students would work on question number 2 if they could work on question number 1. It is impossible to predict the amount of carbon dioxide gas if they cannot master the concept of an equivalent reaction. Items on this stem started by balancing the reaction; if it fails to make the reaction equal, it is impossible to determine the mass ratio on the respondent with a score of 0 with a logit of −2 to 0 logit. Score 1 is at logit −3 to 2; score 2 is at logit 0 to 3 logit. In addition, item 1 had a low level of difficulty, in which low-category people should easily answer this question while high-category people could not. In the Testlet, the questions in stem 1 consisted of three questions, and all three were related. This question was mainly responded to correctly by students with high abilities, while those in the low category could only answer up to point 1. Alternatively, low comprehension abilities could only answer the reaction equation. Besides, it will be impossible to determine the amount of carbon dioxide gas formed, such as the questions in item 1 in layers. Students with high and low abilities could only answer 2 and 3. Thus, this question could not distinguish between high and low comprehension abilities (He et al., 2016; Luo et al., 2018).

Further investigation at the distractor analysis stage provides an overview of how this item should be improved. Item stem 1 needed total improvement because a person with a high, medium, or low ability level could work on this type of question. Suggestions for making improvements are ways to improve the quality of the questions to be better.




STEM 10:

Zinc sulfate is known as white vitriol which is used to obtain the soluble Zn^{2+} ion. Hydrated compounds are more often used as thickeners in the manufacture of rayon. In addition, it is used to supplement the zinc content in animal feed, fertilizers and agriculture, as well as regulate the growth of moss on roofs.

Figure 10. Zinc Sulfate as One Example of Compound Hydrates
Source: https://id.wikipedia.org

1. A total of 0.861 g of ZnSO4.xH2O sample was dissolved in water and the sulfate ion was precipitated by adding excess barium chloride solution. The mass of pure and dry barium sulfate obtained is 0.699 g, then the formula for the zinc sulfate hydrate is … (Ar Zn = 65; S = 32; O = 16; H = 1; Ba = 137; Cl = 35,5)
   A. $ZnSO_4.2H_2O$
   B. $ZnSO_4.3H_2O$
   C. $ZnSO_4.5H_2O$
   D. $ZnSO_4.6H_2O$
   E. $ZnSO_4.7H_2O$
2. The hydrate compound referred to in problem number 28 contains several constituent elements. The element that has the most composition in the hydrate compound is …
   A. Sulfur
   B. Hydrogen
   C. Chlorine
   D. **Oxygen**
   E. Zinc
3. The ratio of moles of atoms that have the highest composition and the smallest composition in 15.35 mL of the hydrate compound sample in problem number 28 is … (p = 1,87 g/mL)
   A. 2 : 7
   B. 4 : 2
   C. 4 : 7
   D. 4 : 14
   E. 11 : 14

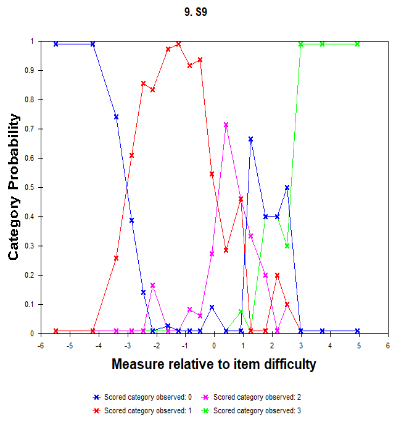
Figure 6. Distractor analysis stem number 10

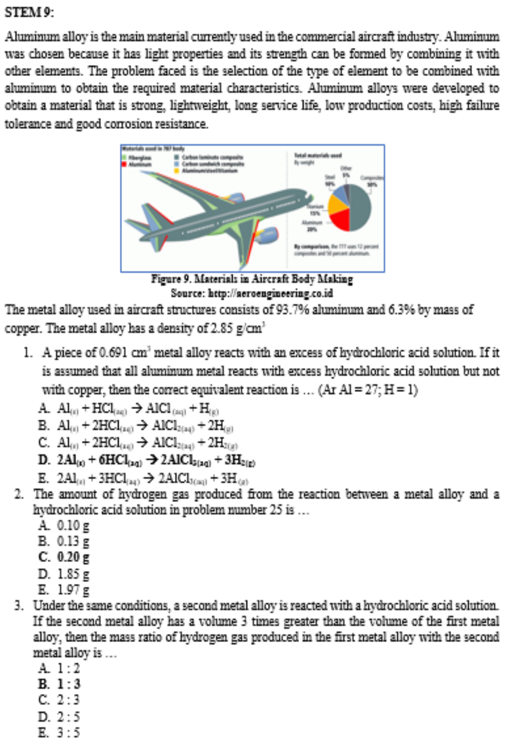
Stem 10 was the most difficult question and had the highest logit value. The analysis was carried out to determine how difficult the questions were and whether they were still feasible. Stem 10 is about hydrate compounds and their functions. In this problem, the first layer determined the chemical formula of the hydrated compound, and the next layer determined the element with the most components of a hydrated compound. The problem in this second layer had a relationship with the first layer. It will not determine the most elemental composition if the hydrated compound cannot be determined. Then, the last layer determines the mole ratio of the element with the most and the smallest composition. It has a relationship with the previous two problems, making students have to determine the ratio correctly from the existing hydrate reaction (Van Bramer et al., 2001).

Questions with a score of 0 or students who could not understand conceptual understanding on this topic were in the −4 logit to 0 logits, and it is a long-range where almost all students with low abilities would fail on this stem. Problems with a value of 1 or could balance the hydrate reaction had a range between −3 logit and −2 logit; almost all students with medium and low ability could complete this well. The value of 2 was in the range of −1 logit to 2 logits; one should have strong analysis abilities on this question. Therefore, with low conceptual understanding skills, it will be challenging to understand this type of question. Meanwhile, a perfect score of 3 on this stem was from 0 logits to 4 logits; this question was proven to distinguish between low and high abilities because only high abilities could do it correctly. This question could still be used because, according to the fit and distractor analysis (Jin et al., 2020; Trail & Howe, 1994), this question was still within the limits specified for questions in the high category.

In contrast, many people failed at the second layer, and some students in high ability persons failed to work on the questions at this layer. It indicates a problem in the second layer that made this question ineffective. This investigation focused on how the person's answer patterns and improvements were made to improve the instrument used.

Moreover, stem 10 was the most difficult question, and the question in the difficult category does not mean that it is not used. The distractor analysis found that people could answer this question with high abilities, and those with low categories could not. Persons with low categories would also stop at the second layer, and most of them failed on the first layer. Then, the medium and high categories had a greater chance of reaching layers 2 and 3. Thus, stem 10 can be used to measure the stoichiometry of conceptual understanding (Taskin et al., 2015).





STEM 9:

Aluminum alloy is the main material currently used in the commercial aircraft industry. Aluminum was chosen because it has light properties and its strength can be formed by combining it with other elements. The problem faced is the selection of the type of element to be combined with aluminum to obtain the required material characteristics. Aluminum alloys were developed to obtain a material that is strong, lightweight, long service life, low production costs, high failure tolerance and good corrosion resistance.

Figure 9. Materials in Aircraft Body Making
Source: http://aeroengineering.co.id

The metal alloy used in aircraft structures consists of 93.7% aluminum and 6.3% by mass of copper. The metal alloy has a density of 2.85 g/cm^3

1. A piece of 0.691 cm^3 metal alloy reacts with an excess of hydrochloric acid solution. If it is assumed that all aluminum metal reacts with excess hydrochloric acid solution but not with copper, then the correct equivalent reaction is … (Ar Al = 27; H = 1)
   A. $Al_{(s)} + HCl_{(aq)} \rightarrow AlCl_{(aq)} + H_{(g)}$
   B. $Al_{(s)} + 2HCl_{(aq)} \rightarrow AlCl_{2(aq)} + 2H_{(g)}$
   C. $Al_{(s)} + 2HCl_{(aq)} \rightarrow AlCl_{2(aq)} + 2H_{2(g)}$
   D. $2Al_{(s)} + 6HCl_{(aq)} \rightarrow 2AlCl_{3(aq)} + 3H_{2(g)}$
   E. $2Al_{(s)} + 3HCl_{(aq)} \rightarrow 2AlCl_{3(aq)} + 3H_{(g)}$
2. The amount of hydrogen gas produced from the reaction between a metal alloy and a hydrochloric acid solution in problem number 25 is …
   A. 0.10 g
   B. 0.13 g
   C. 0.20 g
   D. 1.85 g
   E. 1.97 g
3. Under the same conditions, a second metal alloy is reacted with a hydrochloric acid solution. If the second metal alloy has a volume 3 times greater than the volume of the first metal alloy, then the mass ratio of hydrogen gas produced in the first metal alloy with the second metal alloy is …
   A. 1 : 2
   B. 1 : 3
   C. 2 : 3
   D. 2 : 5
   E. 3 : 5

Figure 7. Distractor analysis stem number 9

Stem 9 discusses metal alloys and their composition. In this problem, individuals will look for reactions that occur in the reaction of metals with acids, determine the mass of products produced from a substance, and determine the ratio of the masses of the products of the first and second reactions. A value of 0 in the range of −4 logit to 3 logits indicates that this question could not determine the difference between one ability, with a vast range, making low or high abilities have the same potential for not answering these bars correctly. At a value of 1 or could work on the first layer, i.e., questions related to the reactions that occur in the reaction of metals with acids, it was in logit in the range of −4 to 2 logit. In addition, in the highest proportion with answers worth 1, almost all respondents could only answer at this layer. It shows that this question could only be used for respondents with moderate ability and could not distinguish between high and low-ability concept understanding.

The value of 2, which was only in the logit −1 to 2 range, was concise and showed that very few students of high ability reached this stage. Determining the mass of the product produced from a substance was a problem for most students. It indicates the need for revision at this stage. Then, at a value of 3 in the range of 1 logit to 5 logits, the third layer discusses determining the mass ratio of the first and second reaction products. Although the third layer had an easier difficulty than the second one, it had a problem in the second layer, resulting in many respondents failing at the next layer (Hecht et al., 2017).

Stem 9 is also a unique question because it is not a difficult or easy question category and is in the standard deviation range, and 0 logit indicates a medium-category question. The analysis on stem 9 revealed that most people had problems with the questions on the second layer. People could work on this problem only at the first layer (Heredia & Lewis, 2012; Tarhan & Acar-Sesen, 2013)

DISCUSSION

Construction validity determines the extent to which the test structure matches the theoretical structure of the construct that has been designed or defined (Aydin & Uzuntiryaki, 2009). The test instrument used must be designed to be unidimensional (measure one particular construct); it must be ensured and determined that only one construct is measured (Finney, 2007). In this research, the Testlet used assessed the conceptual understanding of stoichiometry. In accordance with that situation, the test instrument should be unidimensional. In other words, each item served as a measure of stoichiometric conceptual understanding. As indicated during the evaluation of the assumptions of the Rasch model, the TCI is unidimensional and therefore fits into the designed theoretical structure. The data in Figure 2 show the degree of unidimensionality according to the criteria, 69.4% of the raw variance explained by measure, leaving 30.4% of the standard variance unexplained. It is, therefore, essential to demonstrate how this instrument can assess unidimensionality in this study, assessing the conceptual understanding of stoichiometry (Taskin et al., 2015).

Like validity, reliability cannot be determined based on a single coefficient; evidence of reliability can be seen from various sources and forms (Bolarinwa, 2015). In the Testlet instrument, reliability was divided into person reliability and item reliability. The first looked at how appropriately the participants were involved in this research. The person's reliability also showed whether the participant represented the ideal situation or whether the selected participant had the too-low or too-high ability. Ideally, the distribution of the ability of a person or participant should be divided into high, medium, and low abilities. Meanwhile, the item reliability indicates whether the question items used have the reliability to be used now and in the future. The test results obtained item reliability values of 0.98, a person's reliability of 0.89, and Cronbach Alpha reliability of 0.95. These values suggest that the items in the Computerised Testlet instrument had good consistency when used over time (Scantlebury et al., 2001; Fahmina et al., 2019). In other words, a Computerised Testlet instrument will produce the same information when carried out by several researchers with the same object (Nedungadi et al., 2019). Then, the second reliability value was the person's reliability of 0.89, included in the good category. In addition, this value denotes that the respondents used in this study had good answer consistency. The third reliability value was the Cronbach Alpha reliability of 0.95, meaning that the overall interaction between people and items was very good.

Moreover, a Testlet is needed to measure conceptual understanding of stoichiometry, and accurate data is required about the level of items outside the standard items to determine conclusions about the quality of the items used. Items are at different difficulty levels; how to make this difficulty level the information needed to determine the item's quality, as stem 1 shows the question with the easiest difficulty level with a logit value of –1.53. Stem 10 was the most difficult question with a 1.90 logit item, and this stem was not directly categorised into questions to be revised or unused. The item map only provides information on how ones compare the item's difficulty level and the person's ability. The item map is also not absolute to determine whether the item is working correctly. Item fit is the next step of explaining how the Computerised Testlet can work in conceptual understanding. Two questions were of concern in item fit on stems 1 and 9, and it can be seen that it is outside the standard set for a fit item (Jin et al., 2020; Lu & Bi, 2016; Rachmatullah et al., 2018).

The Computerised Testlet, according to the distractor analysis results, has advantages in distinguishing students' abilities and can map students' conceptual abilities well. The developed Computerised Testlet had three layers for each item. The first to third questions in the same STEM had a related hierarchy. Students should work on the first question to work on the next question, which aims to reduce cheating in the process of moving the questions. The Computerised Testlet also has a function like an open-ended question, where students are indirectly required to work on each question in one item coherently and describe the concepts they understand. Students will write them down on an answer sheet only in open-ended questions, while the Testlet helps by providing answer options. Like some previous research, the test instrument in a Testlet combines the advantages of multiple-choice and description questions. Huang and Wang (2012) provided a Testlet instrument design as a set of items giving a stimulus. This instrument can be used in solving problems, especially in terms of time efficiency, when giving students assignments or exercises to know various stimuli. The basic idea is that students process stimuli to fulfill several items to reveal the same information. These items make it possible to measure the same achievement outside of the traits measured using the overall test (Wainer et al., 2007). Further, the Computerised Testlet can have the same role as the open-ended test in that it gives freedom to each test taker to express his reasoning power so that the answers given by each test taker will show complex thinking skills. Distractor analysis data on items 1, 9 and 10, as shown in Figures 5, 6 and 7, revealed that students were required to do the same as when working on open-ended questions; only each step of working on the questions was assisted with answer choices such as only objective test questions. Based on distractor analysis, the distribution of students' scores for each ability, high, medium and high, had differences, as shown in Figure 9. Students had difficulty in layer 2, where most students with low abilities and some with moderate experienced it, while high abilities could easily work. This analysis uncovered that students were given convenience in expressing answers; at layer 1, item 9, students would express the correct chemical reaction. It would relate to the amount of gas produced in layer 2. Then, students should express the correct reaction before determining the gas produced. Students' answers also would reflect how they could understand chemical concepts and distinguish between students with high and low abilities without reducing their right to demonstrate their ability to understand hierarchical chemical concepts (Liu & Hannig, 2017).

The objective test has the opportunity to answer correctly by guessing quite highly, as indicated by the amount of blind guessing and pseudo-level chance. The scoring on the objective test is dichotomous, so it is not optimal to determine the ability of the test taker (Kieftenbeld & Natesan, 2012). Meanwhile, its advantages are that it is easier and faster to analyse the results of scores than other types of tests (Yamtinah et al., 2018). Even though the Testlet is the same as using multiple choices, it will be more difficult for blind guessing and the pseudo-level chance to occur. In the data on the distribution of student answers, such as in distractor analysis, it can be seen from Items 9 and 10, showing that the Computerised Testlet was difficult for blind guessing and pseudo-level chance as students would experience problems in the next layer if the previous layer could not be answered correctly. The distribution of student scores also showed that difficult questions, such as in question number 10, could only be answered well by students with high abilities, while students with low abilities would only be on a score of 1, even if many were 0.

CONCLUSIONS, LIMITATIONS AND RECOMMENDATIONS

The reliability of this test was divided into two index indicators: item and person. The item reliability value of 0.98 indicates that the items in the Computerised Testlet instrument had good consistency when used from time to time. People's reliability value of 0.89, included in the good category, means that the respondents involved in this study had good answer consistency. In addition, Cronbach's Alpha reliability of 0.95 suggests that the overall interaction between the person and the item was very good. Then, the unidimensionality test resulted in 69.4% of the raw variance explained by measure and left 30.4% of the raw variance unexplained. Overall, it is considered that the data met the requirements of unidimensionality so that this instrument could be used to measure the conceptual understanding of stoichiometry.

Then, item analysis provides information that in this instrument, the difficulty level of the questions was divided into easy, medium, difficult, and very difficult, where most questions were in the medium and difficult distribution. The distribution of persons and items has shown good results, where the questions were well distributed. In item fit, the researchers found several notes showing the peculiarities of the Testlet, where students should understand the previous question to answer the question in one stem. Otherwise, students would be known to make guesses in working on the questions. Moreover, the statistical fit analysis uncovered that the items on the Computerised Testlet test instrument could distinguish low and high-ability students. The Computerised Teslet has also proven to be used and worked effectively to measure students' chemical understanding of the topic of stoichiometry.

Further, computerised Testlets have advantages in terms of speed of analysis and have many similarities with open-ended tests, especially regarding the hierarchy of doing tests. The Computerised Testlet can be used as a practical and precise instrument to reduce relatively high guessing, as indicated by the amount of blind guessing and pseudo-level chance. The use of Computerised Testlets from Indonesia or even Asia Pacific needs to be done to assist teachers in conducting fair, objective, and fast tests with open-ended questions. The computerised test in Indonesia or Asia Pacific has also been running and has a significant growth speed; this makes this application an opportunity to be used because of the speed of analysis and can help students think hierarchically as in open-ended questions.

REFERENCES

Adams, C. M., & Callahan, C. M. (1995). The reliability and validity of a performance task for evaluating science process skills. *Gifted Child Quarterly*, *39*(1), 14–20. https://doi.org/10.1177/001698629503900103

Ahmad, J., & Siew, N. M. (2021). Curiosity towards stem education: A questionnaire for primary school students. *Journal of Baltic Science Education*, *20*(2), 289–304. https://doi.org/10.33225/jbse/21.20.289

Ajayi, O. V. (2017). Effect of hands-on activities on senior secondary chemistry students achievement and retention in stoichiometry in zone C of benue state. *SSRN Electronic Journal*, *5*(8), 839–842. https://doi.org/10.2139/ssrn.2992803

Aydin, Y. Ç., & Uzuntiryaki, E. (2009). Development and psychometric evaluation of the high school chemistry self-efficacy scale. *Educational and Psychological Measurement*, *69*(5), 868–880. https://doi.org/10.1177/0013164409332213

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, *5*(1), 7–74. https://doi.org/10.1080/0969595980050102

Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quiñonez, H. R., & Young, S. R. (2018). Best practices for developing and validating scales for health, social, and behavioral research: A primer. *Frontiers in Public Health.* https://doi.org/10.3389/fpubh.2018.00149

Bolarinwa, O. (2015). Principles and methods of validity and reliability testing of questionnaires used in social and health science researches. *Nigerian Postgraduate Medical Journal*, *22*(4), 195. https://doi.org/10.4103/1117-1936.173959

Brookhart, S. M. (2005). Developing measurement theory for classroom assessment purposes and uses. *Educational Measurement: Issues and Practice*, *22*(4), 5–12. https://doi.org/10.1111/j.1745-3992.2003.tb00139.x

Cacciatore, K. L., & Sevian, H. (2006). Teaching lab report writing through inquiry: A green chemistry stoichiometry experiment for general chemistry. *Journal of Chemical Education*, *83*(7), 1039–1041. https://doi.org/10.1021/ed083p1039

Chang, Y., & Wang, J. (2010). *Examining testlet effects on the PIRLS 2006 assessment.* Paper presented at the 4th IEA International Research Conference.

Chen, S. Y., & Liu, S. Y. (2020). Using augmented reality to experiment with elements in a chemistry course. *Computers in Human Behavior*, *111*(October 2019), 106418. https://doi.org/10.1016/j.chb.2020.106418

Currell, K., & Jeukendrup, A. (2008). Validity, reliability and sensitivity of measures of sporting performance LK-. *Sports Medicine TA - TT -*, *38*(4), 297–316. https://rug.on.worldcat.org/oclc/367041412

Deng, Y., & Wang, H. (2017). Research on evaluation of Chinese students' competence in written scientific argumentation in the context of chemistry. *Chemistry Education Research and Practice*, *18*(1), 127–150. https://doi.org/10.1039/c6rp00076b

DeMars, C. E. (2012). Confirming testlet effects. *Applied Psychological Measurement, 36*(2), 104–121.

DeVellis, R. F. (2017). *Scale development: Theory and applications* (4th ed.). Thousand Oaks, CA: Sage.

Fahmina, S. S., Masykuri, M., Ramadhani, D. G., & Yamtinah, S. (2019). Content validity uses Rasch model on computerised testlet instrument to measure chemical literacy capabilities. *AIP Conference Proceedings*, *2194*(December). https://doi.org/10.1063/1.5139755

Feyzíoglu, B., Demirdag, B., Akyildiz, M., & Altun, E. (2012). Developing a Science process skills test for secondary students: Validity and reliability study. *Educational Sciences: Theory & Practice, 12*(3), 1899-1906.

Finney, S. J. (2007). Book review: Exploratory and confirmatory factor analysis: understanding concepts and applications. *Applied Psychological Measurement*, *31*(3), 245–248. https://doi.org/10.1177/0146621606290168

Garner, J. P. (2005). Stereotypies and other abnormal repetitive behaviors: Potential impact on validity, reliability, and replicability of scientific outcomes. *ILAR Journal*, *46*(2), 106–117. https://doi.org/10.1093/ilar.46.2.106

Griffin, P., & Care, E. (Eds.) (2015). *Assessment and teaching of 21st century skills: Methods and approach.* Springer. https://doi.org/10.1007/978-94-017-9395-7

Gurel, D. K., Eryilmaz, A., & McDermott, L. C. (2015). A review and comparison of diagnostic instruments to identify students' misconceptions in science. *Eurasia Journal of Mathematics, Science and Technology Education*, *11*(5), 989–1008. https://doi.org/10.12973/eurasia.2015.1369a

Hambleton, R. K., & Jones, R. W. (1993). An NCME instructional module on comparison of classical test theory and item response theory and their applications to test development items: Instructional topics in educational measurement. *Educational Measurement: Issues and Practice*, *12*(3), 38–47.

He, P., Liu, X., Zheng, C., & Jia, M. (2016). Using Rasch measurement to validate an instrument for measuring the quality of classroom teaching in secondary chemistry lessons. *Chemistry Education Research and Practice*, *17*(2), 381–393. https://doi.org/10.1039/c6rp00004e

Hecht, M. Siegle, T. Weirich, S., Hecht, M., & Siegle, T. (2017). A model for the estimation of testlet response time to optimize test assembly in paper-and-pencil large-scale assessments. *Journal for Educational Research Online Journal for Educational Research Online 9*(1), 32–51. https://doi.org/10.25656/01:12965

Heredia, K., & Lewis, J. E. (2012). A psychometric evaluation of the Colorado learning attitudes about science survey for use in chemistry. *Journal of Chemical Education*, *89*(4), 436–441. https://doi.org/10.1021/ed100590t

Huang, H. Y., & Wang, W. C. (2012). Higher order testlet response models for hierarchical latent traits and testlet-based items. *Educational and Psychological Measurement*, *73*(3), 491–511. https://doi.org/10.1177/0013164412454431

Jin, Y., Rodriguez, C. A., Shah, L., & Rushton, G. T. (2020). Examining the psychometric [roperties of the redox concept inventory: A Rasch approach. *Journal of Chemical Education*, *97*(12), 4235–4244. https://doi.org/10.1021/acs.jchemed.0c00479

Kapsalis, V. (2009). Implementation of an assessment system incorporating web-based parameterized questions. *International Journal of Emerging Technologies in Learning*, *4*(3), 20–28. https://doi.org/10.3991/ijet.v4i3.884

Kelting-Gibson, L., Gallavan, N. P., St. Arnauld, E., Black, G., Cayson, A., Davis, J., Evans, K. D., Johnson, P. P., Levandowski, B., Mosley, K., Rickey, D., Shulsky, D. D., Thomas, D., Williamson, A. M., & Wolfgang, J. I. (2014). Four facets of classroom assessments: obstacles, obligations, outcomes, and opportunities. *Action in Teacher Education*, *36*(5–6), 363–376. https://doi.org/10.1080/01626620.2014.977688

Khlaisang, J., & Koraneekij, P. (2019). Open online assessment management system platform and instrument to enhance the information, media, and ICT literacy skills of 21st century learners. *International Journal of Emerging Technologies in Learning*, *14*(7), 111–127. https://doi.org/10.3991/ijet.v14i07.9953

Kieftenbeld, V., & Natesan, P. (2012). Recovery of graded response model parameters: A comparison of marginal maximum likelihood and Markov chain Monte Carlo estimation. *Applied Psychological Measurement, 36*(5), 399–419. https://doi.org/10.1177/0146621612446170

Klassen, S. (2006). Contextual assessment in science education: Background, issues, and policy. *Science Education*, *90*(5), 820–851. https://doi.org/10.1002/sce.20150

Korayem, M. H., Hoshiar, A. K., & Ghofrani, M. (2017). Comprehensive modelling and simulation of cylindrical nanoparticles manipulation by using a virtual reality environment. *Journal of Molecular Graphics and Modelling*, *75*, 266–276. https://doi.org/10.1016/j.jmgm.2017.06.006

Le Maire, N. V., Verpoorten, D. P., Fauconnier, M. L. S., & Colaux-Castillo, C. G. (2018). Clash of chemists: A gamified blog to master the concept of limiting reagent stoichiometry. *Journal of Chemical Education*, *95*(3), 410–415. https://doi.org/10.1021/acs.jchemed.7b00256

Linacre, J. M. (2011). *A user's guide to Winsteps & Ministep Rasch Model computer program.* Winsteps.

Liu, Y., & Hannig, J. (2017). Generalized fiducial inference for logistic graded response models. *Psychometrika, 82*(4), 1097-1125. https://doi.org/10.1007/s11336-017-9554-0

Lu, S., & Bi, H. (2016). Development of a measurement instrument to assess students' electrolyte conceptual understanding. *Chemistry Education Research and Practice*, *17*(4), 1030–1040. https://doi.org/10.1039/c6rp00137h

Luecht, R., Brumfield, T., & Breithaupt, K. (2006). A testlet assembly design for adaptive multistage tests. *Applied Measurement in Education*, *19*(3), 189–202. https://doi.org/10.1207/s15324818ame1903_2

Luo, M., Wang, Z., Sun, D., & Wan, Z. H. (2018). Evaluating scientific reasoning ability : the design and validation of an assessment with a focus on reasoning and the use of evidence. *Journal of Baltic Science Education*, *19*(2), 261–275. https://files.eric.ed.gov/fulltext/EJ1271010.pdf

Lutviana, E., Rahardjo, S. B., Susanti, E., Yamtinah, S., Mulyani, S., & Saputro, S. (2019). The computer-assisted testlet assessment instrument to measure students ' learning difficulties in chemical bonding. *Journal of Physics: Conference Series*, *1156*(012019), 1–5. https://doi.org/10.1088/1742-6596/1156/1/012019

Mahaffy, P. (2006). Moving Chemistry education into 3D: A tetrahedral metaphor for understanding chemistry union carbide award for chemical education 1. *Journal of Chemical Education*, *83*(1), 49–55.

Marzano, R. J., Pickering, D., & McTighe, J. (1993). *Assessing student outcomes: performance assessment using the dimensions of learning model*. Association for Supervision and Curriculum Develpment. https://doi.org/ED419696

Nedungadi, S., Paek, S. H., & Brown, C. E. (2019). Utilizing Rasch analysis to establish the psychometric properties of a concept inventory on concepts important for developing proficiency in organic reaction mechanisms. *Chemistry Teacher International*, *0*(0), 1–10. https://doi.org/10.1515/cti-2019-0004

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Paek, I., Assessment, H., & Wilson, M. (2009). Random parameter structure and the testlet model : Extension of the Rasch testlet model. *Journal of Applied Measurement*, *10*(4), 394–407.

Pentecost, T. C., & Barbera, J. (2013). Measuring learning gains in chemical education: A comparison of two methods. *Journal of Chemical Education*, *90*(7), 839–845. https://doi.org/10.1021/ed400018v

Rachmatullah, A., Diana, S., & Ha, M. (2017). The effects of curriculum, gender and students' favorite science subject on Indonesian high-school students' conceptions of learning science. *Journal of Baltic Science Education*, *16*(5), 797–812.

Rachmatullah, A., Diana, S., & Ha, M. (2018). Identifying Indonesian upper-secondary school students' orientations to learn science and gender effect through the use of structural equation modeling. *Journal of Baltic Science Education*, *17*(4), 633–648. https://doi.org/10.33225/jbse/18.17.633

Sabah, S., & Hammouri, H. (2009). *V*alidation of a scale of attitudes toward science across countries using Rasch model: Findings from TIMSS. *Journal of Baltic Science Education*, 12(5), 692–702. https://doi.org/10.33225/jbse/13.12.692

Scantlebury K., Boone W., Kahle J. B., & Fraser B. J. (2001). Design, validation, and use of an evaluation instrument for monitoring systemic reform. *Journal of Research in Science Teaching, 38,* 646–662. https://doi.org/10.1002/tea.1024

Shidiq, A. S., Yamtinah, S., & Masykuri, M. (2019). Identifying and addressing students' learning difficulties in hydrolysis using testlet instrument. *AIP Conference Proceedings*, *2194*(020117), 1–8. https://doi.org/10.1063/1.5139849

Shidiq, A. S., Yamtinah, S., & Masykuri, M. (2016). Assessing science process skills using testlet instrument. *Assessment for Improving Students' Performance*, 231–234.

Sumintono, B., & Widhiarso, W. (2014). *Aplikasi model rasch untuk penelitian ilmu-ilmu sosial.* Trim Komunikata Publishing House.

Syang, A., & Dale, N. B. (1993). Computerised adaptive testing in computer science: Assessing Student programming abilities. *ACM SIGCSE Bulletin*, *25*(1), 53–56. https://doi.org/10.1145/169073.169109

Tangio, J. S. (2019). *Issn 1648-3898 Issn 2538-7138* Analytic approach of response pattern of diagnostic test items in evaluating students ' conceptual understanding of characteristics of particle of matter. *Journal of Baltic Science Education, 19*(5) 824–841. http://www.scientiasocialis.lt/jbse/?q=node/905

Tarhan, L., & Acar-Sesen, B. (2013). Problem based learning in acids and bases: learning achievements and students' beliefs. *Journal of Baltic Science Education*, *12*(5), 565-578. https://.doi.org/10.33225/jbse/13.12.565

Taskin, V., Bernholt, S., & Parchmann, I. (2015). An inventory for measuring student teachers' knowledge of chemical representations: design, validation, and psychometric analysis. *Chemistry Education Research and Practice*, *16*(3), 460–477. https://doi.org/10.1039/c4rp00214h

Taylor, A., & Watson, S. B. (2000). The effects of traditional classroom assessment on science learning and understanding of the processes of science. *Journal of Elementary Science Education*, *12*(1), 19–32.

Trail, C., & Howe, A. C. (1994). The mole concept: Developing an instrument to assess conceptual understanding, *Journal of Chemical Education, 71*(8), 653–655.

Van Bramer, S., Fisher, M., Cooper, M. A., Elzerman, A. W., Lee, C. M., Zeile, J. V., Jones, L. L., Dorhout, P. K., Kingsbury, C., Schelble, S., Zielinski, T. J., & Schwenz, R. W. (2001). Symposium Report: Teaching chemistry in the new century. *Journal of Chemical Education*, *78*(9), 1167–1174. https://doi.org/10.1021/ed078p1171

Wainer, H., Bradlow, E. T., & Wang, X. (2013). Testlet response theory and its applications. In *Testlet Response Theory and Its Applications*. Cambridge University Press. https://doi.org/10.1017/CBO9780511618765.002

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications.* Cambridge University Press. https://doi.org/10.1017/CBO9780511618765

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24*(3), 185–201.

Wang, W. C., & Wilson, M. (2005). The Rasch testlet model. *Applied Psychological Measurement*, *29*(2), 126–149. https://doi.org/10.1177/0146621604271053

Wei, S., Liu, X., Wang, Z., & Wang, X. (2012). Using rasch measurement to develop a computer modeling-based instrument to assess students' conceptual understanding of matter. *Journal of Chemical Education*, *89*(3), 335–345. https://doi.org/10.1021/ed100852t

Xiaohui, W., Bradlow, E. T., & Wainer, H. (2002). *Model for Testlets : A general bayesian model for testlets: theory and applications* (Issue 98). Educational Testing Service.

Yamtinah, S., Indriyanti, N. Y. Y., Saputro, S., Mulyani, S., Ulfa, M., Mahardiani, L., Satriana, T., & Shidiq, A. S. (2019). The identification and analysis of students' misconception in chemical equilibrium using computerised two-tier multiple-choice instrument. *Journal of Physics: Conference Series*, *1157(4),* 042015. https://doi.org/10.1088/1742-6596/1157/4/042015

Yamtinah, S., Masykuri, M., Ashadi, & Shidiq, A. S. (2017). Gender differences in students' attitudes toward science: An analysis of students' science process skill using testlet instrument. *AIP Conference Proceedings*, *1868030003*. https://doi.org/10.1063/1.4995102

Yamtinah, S., Saputro, S., Mulyani, S., & Shidiq, A. S. (2021). Computerized testlet instrument for assessing students' chemical literacy in high school. *Journal of Hunan University Natural Sciences, 48*(2), 156–164.

Yamtinah, S., Saputro, S., Mulyani, S., Ulfa, M., Lutviana, E., & Shidiq, A. S. (2019). Do students have enough scientific literacy? A computerised testlet instrument for measuring students' scientific literacy. *AIP Conference Proceedings*, *2194,* 020143. https://doi.org/10.1063/1.5139875

Zhu, X., & Stone, C. A. (2012). Bayesian comparison of alternative graded response models for performance assessment applications. *Educational and Psychologica*, *72*(5), 774–799. https://doi.org/10.1177/0013164411434638