



The Validity of HOTS Assessment Instrument to Measure Ability of Problem-Solving on Thermochemistry Materials

Ismi Inayati¹✉, Endang Susilaningsih², Jumaeri³

¹. Chemistry Education, Postgraduate Universitas Negeri Semarang, Indonesia

^{2,3}. Departmen of Chemistry, FMIPA, Universitas Negeri Semarang, Indonesia

Article Info

Keywords:

*HOTS, Problem-Solving,
Thermochemistry,
Validity*

Abstract

This research is motivated by the lack of variety of HOTS questions to measure ability of problem-solving on thermochemistry materials. The objective of this research is to develop a HOTS assessment instrument to measure valid problem-solving abilities on thermochemical materials. The research method used is R&D with a 4D model adapted from Thiagarajan that is converted into 3D, that are define, design, and develop. The subjects of this research were 77 students of class XI MIPA SMA Negeri 1 Pemasang. The data collection sources and methods include: interviews, questionnaires, and tests. The analysis results of the content validity of the instrument obtained that the Aiken'S coefficient in all aspects measured was above 0.75. This shows that the HOTS assessment instrument is very valid. The validity of the construct is seen from the value raw variance explained by measures 48.6% has very good criteria so that the question items are valid. The analysis results of the quality of the HOTS questions to measure problem-solving abilities are seen in terms of the validity of the items, it is known that there are 16 items that are said to be valid and 4 items are said to be invalid because they do not meet the requirements outfit MNSQ, Outfit ZSTD, and Point Measure Correlation.

✉ Correspondence Address (author1):

E-mail: isminayati@gmail.com

INTRODUCTION

The assessment of learning outcomes aims to determine the success of processes and teaching in schools which is far from effective in changing student behavior towards the expected educational goals (Gündüz et al., 2016; Tim Pusdiklat Pegawai, 2016). This assessment of learning outcomes is expected to help students

improve Higher Order Thinking Skills (HOTS), because HOTS can encourage students to think broadly and deeply about the subject matter. The HOTS assessment instrument is very rare to be found when it is compared to LOTS (Lower Order Thinking Skills). Whereas, in Bloom's taxonomy remembering, understanding, and applying are at the lower level, but they are still often used in assessment (Kusuma et al., 2017;

Widana, 2018). The HOTS includes analyzing (C4), evaluating (C5), and creating (C6) (Abidin et al., 2019; Chalkiadaki, 2018; Talmi et al., 2018). There are some aspects that show the higher-order thinking skills possessed by students, that are the ability to think critically, think creatively, and solve problems. Those aspects cannot be owned directly, but require a practice process to work on high-level questions (HOTS). The HOTS assessment in measuring the ability to (1) transfer one concept to another; (2) processing and applying information; (3) search for links of various types of information; (4) using the information to solve problems; and (5) critically reviewing ideas and information (Mujib, 2019).

HOTS questions are highly recommended for use in various forms of class assessment and school examinations. The characteristics of the HOTS questions include measuring higher-order thinking skills, using contextual problems, and using various forms of questions (Widana, 2017). The use of HOTS questions, students can practice their ability to master concepts evenly so that they can analyze, synthesize, evaluate, and create a concept well (Ichsan et al., 2019). One of the HOTS characteristic is the problem-solving skill. Problem-solving based learning is one of the innovative learning models that can provide active learning conditions for students (Nugroho et al., 2017). The Indicator of problem-solving skills used is adapted from the book *How To Solve It* by George Polya, that are understanding the problem, making plans, implementing plans, and checking (Siswanti et al., 2016). Problem-solving skills are important to be improved because they play a very important role in life in order to develop students' abilities in dealing with problems (Jayadiningrat & Ati, 2018). To increase the potential of students, they must be trained to solve HOTS problems (Harta, 2017; Setiawan et al., 2021). The students are expected to be able to overcome the problems that exist in society, especially in the field of chemistry so that it becomes a habit and the formation of good character.

Thermochemistry is a basic chemical concept that is very important in daily life such as heat, temperature, enthalpy, and energy changes (Rahmawati et al., 2021). However, students' understanding of thermochemistry material. So that, thermochemistry learning requires concept analysis and problem-solving skills (Siswanti et al., 2016). Chemistry learning can also help equip students not only with chemical concepts but experiences and facts in daily life and culture (Sudarmin et al., 2018). Based on the results of interviews conducted with chemistry teachers at SMA Negeri 1 Pematang, it was revealed that chemistry teachers still had difficulties in developing a HOTS assessment instrument that could measure problem-solving abilities, for example in thermochemistry material. This causes the HOTS assessment instrument to measure the resulting problem-solving ability is not feasible and is still very rarely applied.

The instrument used in the assessment has several conditions, one of which is to be valid (Sumintono & Widhiarso, 2015). Validity in the preparation of the assessment instrument is very important and needs to be done. Several types of validity that are measured in the assessment instrument are content validity, construct validity, and item validity. The validity of the content relates to whether the statement items arranged in the questionnaire or test have covered all the material to be measured (Budiastuti & Bandur, 2018). The validity of the content of the instrument was obtained by giving questionnaires to experts, namely assessment experts and learning experts (Bashoor & Supahar, 2018). An assessment instrument to produce accurate information requires proper analysis. One of them is by using the Item Response Theory (IRT) analysis of the Rasch Model. The analysis of the Rasch model can detect individuals whose response patterns do not match and invalid items (outliers or misfits) (Susilaningih et al., 2021). The Rasch model used in this study has several advantages, namely it can identify measurement inaccuracies, predict missing data, distinguish the ability of respondents with the same raw

score, and also identify any indications of guesswork and fraud in choosing (Marfu'i et al., 2019).

Several research results have been conducted to measure the validity of the HOTS assessment instrument, namely as follows; there are 16 questions of HOTS items were obtained that the questions reached the valid criteria secara (Saraswati et al., 2021); The item fit or item fit with the Rasch model can be said to be mathematically valid as an empirical item based on the INFIT MNSQ average value and standard deviation (Widiyawati et al., 2019); unidimensionality in the Rasch Model is used to find out whether it measures what it should measure so that it can also be called construct validity (Riswanda, 2018).

METHOD

The method used in this research is R&D with a 4D model adapted from Thiagarajan which is converted into 3D, that are defining, designing, and developing. In the defining stage, it is carried out by analyzing learning carried out by chemistry teachers in class, analysis of test instruments that are often developed and used by teachers and analysis of thermochemical material. The second stage is to design a HOTS assessment instrument product to measure problem-solving abilities by reviewing thermochemical material that is in accordance with Basic Competence. In the third stage, the development was carried out by means of instrument validation by experts, small-scale trials, and large-scale trials.

This research was conducted at SMA Negeri 1 Pematang with research subjects as many as 77 students from class XI MIPA. Subject selection based on the recommendation of the chemistry teacher at the school. The data collection sources and methods include: interviews, questionnaires, and tests. The developed instrument needs to be analyzed. The item of analysis is very important because it aims to find out the validity of the tests that have been made (Quaigrain & Arhin, 2017). The analysis in this study is the analysis of the validity of the content, constructs, and

questions. Content validity was analyzed by Aiken' V (Ortega-Toro et al., 2019). Meanwhile, construct validity and items were analyzed using the Rasch Model with the help of the Winstep program 3.73.

The first step in developing an assessment instrument is measuring content validity (Ikhsanudin & Subali, 2018). Content validity is carried out by distributing questionnaires to validator lecturers who are material experts and evaluation experts for HOTS validation and participating teachers (Andrian et al., 2018; Boateng et al., 2018; Hidayati et al., 2019; Setiawan et al., 2021). Validation test using a questionnaire by giving a score of 4, 3, 2, and 1 with the answer choices according to the content of the question, that are: "Very valid", "valid", "Quite valid", "Invalid" and "Invalid" (Prasetya et al., 2019). The analysis results are obtained from the formula:

$$V = \sum S/[n(c - 1)]$$

- s = r - lo
- lo = lowest validity assessment score (1)
- c = the highest score of validity assessment (4)
- r = numbers given by expert

The content validity criteria for each aspect are presented in Table 1.

Table 1. Content Validity Criteria

Aiken'V Coefficient	Category
0.75 < V ≤ 1.00	Very Good
0.50 < V ≤ 0.75	Good
0 < V ≤ 0.50	Very Poor

The construct validity of the Rasch Model with the Winsteps program is presented in *output tabels 23. Item: Dimensionality*. Construct validity can be determined by looking at the *Raw Variance* and *Unexplained variance* by the criteria presented in Table 2.

Table 2. Construct validity criteria

Range (%)	Criteria
60 < % ≤ 100	Excelent
40 < % ≤ 60	Very Good
20 < % ≤ 40	Good
0 < % ≤ 20	Very Poor

(Sumintono & Widhiarso, 2015)

The validity of the items in the Rasch Model is if it meets the requirements for Outfit MNSQ, Outfit ZSTD, and Pt Measure Corr (Sumintono & Widhiarso, 2015). The criteria for the validity of the items are listed in Table 3.

Table 3. Item validity criteria

Interval	Criteria
$0,5 < \text{MNSQ} < 1,5$	Accepted
$-2,0 < \text{ZSTD} < 2,0$	Accepted
$0,4 < \text{Pt. Measure Corr} < 0,85$	Accepted

(Sumintono & Widhiarso, 2015:12)

RESULTS AND DISCUSSION

This reserach aims to produce a HOTS assessment instrument product to measure problem-solving abilities on valid thermochemical materials. The item has high validity if the score on the item has a correlation with the total score (Wardany et al., 2017). This research was conducted in 3 stages, that are defining, designing, and developing. The defining stage is a study of relevant sources (literature studies and field studies) which aims to determine and define needs by analyzing material objectives and limitations (Habibah & Widodo, 2017). In addition, interviews with chemistry teachers were conducted to find out the question instruments used in the assessment of thermochemistry materials.

Based on the results of the interview, it was identified that only a few questions on thermochemistry materials used HOTS. Besides, the teacher has not been maximal in developing HOTS questions on thermochemistry material. The difficulty in developing the HOTS questions is due to the teacher's limited time. This is supported by Purwasih' research (2020) that difficulties in developing HOTS questions, one of which is the limited time to develop questions. Besides the teacher has a limit time, the lack of training makes HOTS questions also an obstacle for teachers. Even if there is training, the material provided is not specifically about HOTS assessment in chemistry learning (Nurmawati et al., 2021).

The second stage of this research is to design a HOTS assessment instrument product to measure problem-solving abilities. The researcher examines the thermochemistry material in accordance with Basic Competence 3.4, namely analyzing the concept of the enthalpy change of a reaction at constant pressure in a thermochemical equation and 3.5 analyze types of reaction enthalpy, Hess's law and the concept of bond energy. Based on the Basic Competences, the researchers designed a HOTS assessment instrument for thermochemistry material that can be used to measure students' problem-solving abilities.

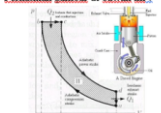
The last stage is the development of the HOTS question instrument product, which consists of 20 questions to measure problem-solving abilities. The design of the question instrument that has been developed is then validated in terms of material, construction, and language (Astuti et al., 2020; Festiana et al., 2020). The process is to determine the validity of the contents of the instrument. Content validity is a test carried out on an instrument to determine the suitability between the theory and the items of the instrument made, so that the item of the instrument is able to represent the overall content of the material being tested (Sugiharni, 2017).

The product validation process is carried out through four stages, which are expert validation relevant to the field of study, practitioner teacher validation, small-scale trials, and large group trials (Andrian et al., 2018; Pandra et al., 2021; Supriyadi, 2021). Expert assessment will be used to prove the content validity of the developed assessment instrument (Calonge-pascual et al., 2020; Pandra et al., 2021). The question instruments that have been made are validated by 2 validator lecturers and 1 participant teacher. Experts and participating teachers were asked to check the suitability of the items with the indicators of problem-solving, question writing, and the suitability of distractors in multiple choice. In addition, expert validation asked for suggestion so that the instrument developed is better. If there are still mistakes in making the instrument, then the

instrument is revised again (Arifin, 2017). The results of improvements from expert suggestions

are presented in Table 4.

Table 4. Suggestions and improvements to the question instruments based on experts

Suggestion (Before Repair)				After Repair					
There are no Problem-Solving Indicators (IPM) on the question grid yet				Addition of Problem-Solving Indicators (IPM) to the question grid					
No.	Indikator Pencapaian Kompetensi	Indikator Soal	Tingkat Kognitif	No.	Indikator Pencapaian Kompetensi	Indikator Pemecahan Masalah	Indikator Soal	Tingkat Kognitif	Soal
1.	Menganalisis perbedaan sistem dan lingkungan	Menganalisis perbedaan sistem dan lingkungan pada mesin kendaraan bermotor	C4	1.	Menganalisis perbedaan sistem dan lingkungan	a. Menjelaskan masalah dengan kalimat sendiri (tidak memahuri masalah) b. Melakukan langkah/ strategi penyelesaian soal (tidak menyalah rancangan) c. Melaksanakan	Menganalisis perbedaan sistem dan lingkungan pada mesin kendaraan bermotor	C4	Perhatikan gambar di bawah ini !  Gambar di atas merupakan gambar m

The question about tempeh fermentation is replaced with tape fermentation

The replacement of the question of tempeh fermentation is replaced with tape fermentation

Tempe merupakan makanan yang kaya akan protein. Proses pembuatan tempe, cukup mudah, yaitu :

1. Kedelai disortir dan dicuci bersih kemudian direbus
2. Kupas kulit ari lalu rendam selama 12-16 jam pada suhu kamar
3. Kedelai tersebut dimasak lalu tiriskan dan dinginkan
4. Tambahkan inokulum
5. Simpan kedelai dalam wadah terbuka

Pernyataan yang sesuai untuk deskripsi prosedural di atas yaitu ...

Tape merupakan makanan cemilan yang populer. Proses pembuatan tape cukup mudah, yaitu :

- a. Kupas singkong, potong, kemudian cuci hingga bersih
- b. Kukus singkong di atas air mendidih hingga ¾ matang, lalu dinginkan
- c. Masukkan singkong ke dalam wadah lalu taburi dengan ragi yang telah dihaluskan dengan menggunakan saringan
- d. Singkong yang telah diberi ragi ini kemudian ditutup kembali dengan daun pisang.
- e. Setelah singkong ditutup dengan daun pisang, diamlkan selama 1-2 hari hingga sudah terasa lunak dan manis. Saat itulah singkong telah menjadi tape.
- f. Reaksi yang terjadi yaitu $C_6H_{12}O_6 \rightarrow 2C_2H_5OH + 2CO_2 + 2 ATP$

Pernyataan yang sesuai untuk deskripsi prosedural di atas yaitu ...

It is better to add answer choices about the energy required to form 1 mole of ATP into ADP

Added answer choices about the energy required to make 1 mole of ATP into ADP

Pernyataan yang sesuai untuk deskripsi prosedural di atas yaitu ...

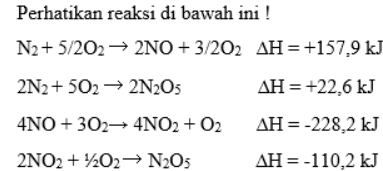
- A. Fermentasi tempe merupakan proses reaksi anaerob. Pada langkah di atas, proses fermentasi dilakukan dalam wadah terbuka sehingga tempe tidak berhasil dibuat. Reaksi yang terjadi yaitu reaksi eksoterm $C_6H_{12}O_6 \rightarrow 2C_2H_5OH + 2CO_2 + 2 ATP$ (Energi yang dilepaskan: 118 kJ per mol)
- B. Fermentasi tempe merupakan proses reaksi aerob. Pada langkah di atas, proses fermentasi dilakukan dalam wadah terbuka sehingga tempe berhasil dibuat. Reaksi yang terjadi yaitu reaksi eksoterm $C_6H_{12}O_6 \rightarrow 2C_2H_5OH + 2CO_2 + 2 ATP$ (Energi yang dilepaskan: 118 kJ per mol)
- C. Setelah kedelai direbus kemudian ditiriskan. Hal ini berguna untuk menghilangkan air yang dapat mengganggu proses fermentasi.
- D. Tempe tidak berhasil dibuat karena terlalu lama pada proses perendaman.
- E. Tempe berhasil dibuat karena langkah-langkah di atas sudah sesuai.

Pernyataan yang sesuai untuk deskripsi prosedural di atas yaitu ...

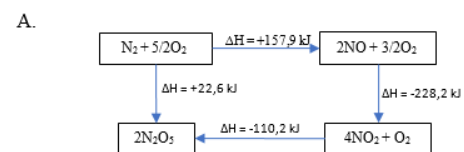
- A. Fermentasi tape merupakan proses reaksi anaerob. Pada langkah di atas, proses fermentasi dilakukan dalam wadah terbuka sehingga tape tidak berhasil dibuat. Reaksi yang terjadi yaitu reaksi endoterm yang menyerap energi sebesar 61 kJ per mol glukosa
- B. Fermentasi tape merupakan proses reaksi aerob. Pada langkah di atas, proses fermentasi dilakukan dalam wadah terbuka sehingga tempe berhasil dibuat. Reaksi yang terjadi yaitu reaksi eksoterm yang menghasilkan energi sebesar 30,5 kJ per mol glukosa
- C. Tape tidak berhasil dibuat pada wadah terbuka. Reaksi fermentasi singkong menjadi tape merupakan reaksi eksoterm yang menghasilkan energi 2 ATP atau sebesar 61 kJ per mol glukosa
- D. Tape berhasil dibuat karena enzim pada ragi tidak pecah apabila terdapat udara yang mengganggu proses pemecahan enzim tersebut.
- E. Tape berhasil dibuat karena langkah-langkah di atas sudah sesuai. Proses pembuatan tape merupakan reaksi eksoterm yang menghasilkan energi sebesar 61 kJ per mol glukosa

There are no multiple-choice questions in the form of graphs/diagrams

Added answer choices in the form of graphs/diagrams



Tentukan grafik yang sesuai dengan reaksi di atas!



It is better to add the mass of glucose to be fermented, the volume of CO₂ gas to be decomposed, and the moles of alcohol formation

Glukosa merupakan jenis karbohidrat yang banyak dijumpai pada berbagai makanan. Glukosa dapat diubah menjadi alkohol melalui reaksi fermentasi. Pada proses termokimia, alkohol dapat dibuat dari fermentasi glukosa dengan perubahan entalpi sebesar $-x$ kJ/mol. Reaksi pembentukan alkohol memiliki perubahan entalpi $-y$ kJ/mol. Jika reaksi penguraian gas karbon dioksida mempunyai perubahan entalpi $+z$ kJ/mol, berapakah perubahan entalpi pembentukan glukosa?

- A. $z + (x - y)$ kJ/mol
- B. $z + 2(x - y)$ kJ/mol
- C. $x - 2x + y$ kJ/mol
- D. $x - 2(x + y)$ kJ/mol
- E. $x - (x + y)$ kJ/mol

Add the mass of glucose to be fermented, the volume of CO₂ gas that is decomposed, and the moles of alcohol formation

Glukosa merupakan jenis karbohidrat yang banyak dijumpai pada berbagai makanan. Glukosa dapat diubah menjadi alkohol melalui reaksi fermentasi. Pada proses termokimia, alkohol dapat dibuat dari fermentasi 540 gram glukosa dengan perubahan entalpi sebesar $-3x$ kJ/mol. Reaksi pembentukan 1 mol alkohol memiliki perubahan entalpi $-y$ kJ/mol. Jika reaksi penguraian 5,6 liter gas karbon dioksida (keadaan STP) mempunyai perubahan entalpi $+z$ kJ/mol, berapakah perubahan entalpi pembentukan glukosa?

- A. $z + (x - 4y)$ kJ
- B. $z + 2(x - 4y)$ kJ
- C. $x - 2x + 4y$ kJ
- D. $x - 2(y + 4z)$ kJ
- E. $x - (y + 4z)$ kJ

Based on Table 4, the items that are considered less good are corrected according to suggestions from experts so that the items can still be used (Riyani et al., 2017).

Researchers make a grid of assessment instruments based only on indicators of competency achievement and have not adjusted to problem solving indicators in accordance with the research. So that in repair number 1, the question instrument grid is given a problem-solving indicator so that the question instrument can measure the problem-solving abilities of students. The second improvement is to replace the stimulus problem which was originally in the form of the process of making tempeh, replaced by fermenting cassava into tape because according to the validator, the process of making tempeh is more complex. The third improvement is still related to the previous question, only in the answer choices it is recommended to add choices about the energy needed to form 1 mole of ATP into ADP. This is intended to adjust to the characteristics of the HOTS items. From the questions made, it turns out that there are no answer choices in the form of graphs/diagrams, so that item number 14 is corrected by adding answer choices in the form of diagrams. The last improvement is that the problem regarding Hess's law is still considered too simple, so it is recommended to add

question variables in the form of the mass of glucose to be fermented, the volume of CO₂ gas that is described, and the moles of alcohol formation. The improvements that have been made are expected to improve the quality of the assessment instruments developed for the better and can be used to measure the problem solving abilities of students.

The developed instrument can measure accurately or provide measurement results in accordance with the purpose of the measurement if the instrument has high validity so that the results of the measurement are quantities that accurately describe what is being measured. Content validity is used to distinguish between appropriate items and items that do not match the research objectives. The content validity of the expert judgment was analyzed using the Aiken's V formula. Aiken's V is used to determine the content validity coefficient based on the assessment of a number of experts (n) on a number of construct items measured (Pandra et al., 2021). In this study, the researcher asked 3 raters ($n = 3$), 20 items to be measured ($m = 20$), and 4 categories ($c = 4$). The number of smallest rating categories formulated by Aiken is 2 and the most is 7 (Bashooir & Supahar, 2018). The results of the content validity analysis of the instrument are presented in Table 5.

Table 5. Instrument content validity

Aspect	Aiken'V	Category
Content		
The questions are in accordance with the basic competencies	1.00	Very Good
The questions are in accordance with the indicators of competency achievement	0.89	Very Good
Items in accordance with the measurement objectives	0.78	Very Good
Problem-Solving		
The question points can develop the ability to identify facts related to the problem.	0.89	Very Good
The question items can develop the ability to determine concepts or categories	0.78	Very Good
The questions can develop the ability to determine information/data related to the given problem.	0.89	Very Good
The question items can develop the ability to determine the details of the problem (time, place, actor).	0.78	Very Good
The question points can develop the ability to map sub-problems and sub-solutions.	0.78	Very Good
The questions can develop the ability to map sub-problems and sub-solutions	0.78	Very Good
The question can develop the ability to choose theories, principles and approaches to solving related problems.	0.78	Very Good
The questions can develop the ability to design a list of problems to be solved.	0.78	Very Good
The questions can develop the ability to design work steps related to solutions that have been made	0.89	Very Good
The questions can develop the ability to check the feasibility of the solution made.	0.78	Very Good
The questions can develop the ability to make assumptions regarding the solutions made	0.78	Very Good
The questions can develop the ability to predict the results that will be obtained through the solutions that have been made	0.89	Very Good
The questions can develop the ability to choose the right media, convey and communicate the solutions that have been made	0.78	Very Good
Language		
Communicative sentence formulation	1.00	Very Good
Using good sentences and correct language, according to the type of language	0.89	Very Good
The formulation of the sentence does not cause double interpretation or misunderstanding	1.00	Very Good
Using standard language	1.00	Very Good

The assessment instrument has good criteria and can be used if the Aiken'V coefficient is above 0.5. Based on the Aiken table, if there are 3 raters ($n = 3$), there are 20 items ($m = 20$), and 4 categories, then the minimum acceptable limit is 0.65 (Pandora et al., 2021). Based on Table 5, all the aspects measured were obtained by Aiken'V coefficients above 0.75 with very good criteria. This means that the HOTS assessment instrument to measure problem-solving abilities is proven valid. After the assessment instrument was valid, a small-scale trial was conducted. This small-scale trial aims to find out and identify various problems such as weaknesses or product deficiencies when used by students. The data obtained from this trial is used as a basis for revising the product before being used in large-scale trials.

A small-scale trial was carried out to 20 students of class XI. The small-scale trial aims to determine the readability of the item and the results obtained can be used as item development (Pandora et al., 2021). Students are asked for suggestions and input about the tests or questions they are working on. These suggestions or inputs are used as material to make improvements to the assessment instruments developed before large-scale trials.

The results of the small-scale trial showed that all the items (20 items) tested were valid because they met the requirements *outfit MNSQ*, *outfit ZSTD*, and *Point Measure Correlation* so that all the questions can be used for large-scale trials. The uniformity of the question instruments is categorized as good, as indicated by the results of the construct validity that meet the requirements. This is supported by research (Purba, 2018) which obtained 16 items of misfit questions, 32 items of fit questions, so that the final instrument was 32 items.

The large-scale trial aims to determine the effectiveness of the changes that have been made to the results of expert validation and small-scale trials whether the HOTS assessment instrument to measure problem-solving abilities can be used. Large-scale trials are carried out as in small-scale trials. The difference is that the number of

participants who are subjects in large-scale trials is more than in small-scale trials. Large-scale trials were carried out on 77 students of class XI. The analysis on the large-scale trial includes the analysis of construct validity and item validity. The analysis of construct validity using the Rasch model in the Winstep program was tested on *output tables 23 unidimensionalitas*. The unidimensionality of the instrument can indicate whether the assessment instrument developed is able to measure what it is supposed to measure. The results of the construct validity analysis with the Winstep program can be seen in Figure 1.

Table of STANDARDIZED RESIDUAL variance (in Eigenvalue units)			
	-- Empirical	--	Modeled
Total raw variance in observations =	38.9	100.0%	100.0%
Raw variance explained by measures =	18.9	48.6%	48.0%
Raw variance explained by persons =	6.3	16.2%	16.0%
Raw variance explained by items =	12.6	32.4%	32.0%
Raw unexplained variance (total) =	20.0	51.4%	52.0%
Unexplained variance in 1st contrast =	2.5	6.5%	12.7%
Unexplained variance in 2nd contrast =	1.9	4.9%	9.5%
Unexplained variance in 3rd contrast =	1.8	4.6%	8.9%
Unexplained variance in 4th contrast =	1.6	4.2%	8.2%
Unexplained variance in 5th contrast =	1.5	3.8%	7.5%

Figure 1. Construct Validity Test Results

Based on Figure 1. the score of *raw variance explained* by empirical measures is 48.6% while the Rasch model predicts 48.0%. In this case, the empirical construct validity has almost the same value as the predictions of the Rasch model. The results of the construct validity have good criteria because it meets the minimum unidimensionality criteria of 20%. The score of the first to fifth unexplained variance is below 15%, which means the instrument uniformity is in the good category. This indicates that the questions used in this study are related to the content of the material (Musa et al., 2017). This is supported by Saidi & Siew's research (2019) that the *Raw variance explained by measures* higher than 20% is acceptable, higher than 40% is good, while higher than 60% is excellent. Besides, *unexplained variance* for 1 to 5 contrast less than 10%, which falls within the ideal range value of less than 15%. The results of the analysis show that the construct validity in the study shows the uniformity of the instruments which are in the good category. This shows that the questions used in this study are related to the content of the material.

In the Rasch model, to see the quality of the items from the validity aspect, that is, if they

meet the requirements, *Outfit MNSQ*, *Outfit ZSTD*, and *Pt Measure Corr* (Sumintono & Widhiarso, 2015). After testing the validity aspects of each item using the Rasch model, the validity of the items developed so as to obtain a item suitability between student responses and the developed assessment instrument. The results of the analysis of the validity of the HOTS items to measure problem-solving abilities are presented in Figure 2.

ITEM STATISTICS: MISFIT ORDER														
ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	S.E.	MODEL MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PT-MEASURE CORR.	EXP.	EXACT OBS%	MATCH EXP%	ITEM	
4	63	77	-1.80	.32	1.38	2.0	2.12	2.3	A	.05	.37	76.6	82.6	14
17	13	77	2.10	.33	1.27	1.4	2.11	2.2	D	.15	.38	83.1	84.1	117
7	61	77	-1.69	.31	1.05	-4.1	1.68	1.7	D	.30	.39	79.2	80.7	17
6	91	77	-2.90	.44	.98	-11.31	.6	0	D	.23	.26	92.2	92.2	16
5	33	77	.46	.26	1.06	.6	1.04	.3	C	.44	.47	70.1	71.1	15
13	35	77	1.03	.25	1.03	-.4	.94	-.2	C	.46	.46	70.1	74.3	113
12	58	77	-1.33	.29	.94	-4.1	1.04	.3	C	.44	.41	79.2	78.2	112
1	69	77	-2.55	.40	1.03	-21.00	.2	0	D	.23	.30	89.6	89.6	11
19	25	77	1.03	.28	1.02	-2.1	1.02	.2	C	.44	.46	75.3	74.5	119
8	51	77	-.77	.27	.99	-11.00	.1	0	D	.46	.45	74.0	73.5	118
9	13	77	-1.10	.33	.90	-.4	.98	-.1	C	.43	.38	85.7	84.1	19
3	59	77	-1.41	.30	.98	-1.1	.80	-.6	C	.45	.41	75.3	79.0	13
10	61	77	-1.60	.31	.98	-1.1	.83	-.4	C	.42	.39	79.2	80.7	110
14	22	77	1.27	.28	.96	-2.2	.92	-.3	C	.47	.45	79.2	76.7	114
16	9	77	2.61	.38	.86	-5.5	.94	-.1	C	.42	.34	90.9	88.4	116
2	68	77	-2.40	.38	.87	-4.5	.95	-.8	C	.43	.32	92.2	89.6	120
18	13	77	2.10	.33	.86	-7.7	.77	-.7	D	.48	.38	88.3	84.1	118
20	8	77	2.76	.40	.86	-5.5	.49	-.9	C	.45	.32	92.2	89.6	120
11	37	77	-1.24	.29	.82	-11.3	.67	-1.7	D	.56	.42	83.1	77.3	11
15	13	77	2.10	.33	.81	-9.9	.64	-9.9	A	.52	.38	88.3	84.1	115
MEAN	39.6	77.0	.00	.33	.98	.0	1.04	.1			82.0	81.7		
S.D.	23.3	.0	1.88	.05	.14	-.7	.44	1.0			7.0	5.9		

Figure 2. Results of Item Validity Test

Based on Figure 2, the results of the analysis of the quality of the HOTS questions to measure problem-solving ability in terms of the validity of the items, information was obtained that there were 16 items that were said to be valid and 4 items were said to be invalid because they did not meet the requirements *outfit MNSQ*, *Outfit ZSTD*, and *Point Measure Correlation (Pt Measure Corr)* that are the questions at number 4, 7, 17 dan 20. However, there are 2 items that do not meet the requirements *Point Measure Correlation (Pt Measure Corr)* that are the questions number 6 and 1. However, both of these questions are still suitable for use because they still meet the requirements *Outfit ZSTD*. The score of *outfit ZSTD* there is no negative value, a negative value indicates a defective test item because students with lower abilities can get high scores on difficult items (Andriani et al., 2021). If the items its been declared valid (fit), it means that the items it meets the criteria and can guarantee that the level of understanding of students is indeed tested through appropriate and quality items (Palimbong et al., 2018). Meanwhile, items that is declared invalid (misfit) can be corrected at a later date.

The research of Darmana et al. (2021) There are 6 items (15%) that are not fit, namely items 40, 9, 28, 27, 36, and 37, and the results of the analysis show that 34 items (85%) are fit. Item fit analysis is used to determine whether the item functions normally or not in the measurement. The analysis shows that the item fits the model, so it can be concluded as a valid item. This item fits the model when at least two matching item criteria are accepted.

The validity of the items in this reserach is in the very strong category so that it can be concluded to have a very strong relationship. The level of suitability of the items obtained that from the 20 questions that were tested on a large scale, there were 4 items that were said to be invalid because they did not meet the requirements *outfit MNSQ*, *Outfit ZSTD*. This is supported by research by Litna et al. (2021) it is known that from 23 test items that were tested in large groups, there were three test items that were rejected, so that 20 items of mathematics test met the quality of the HOTS-based mathematical test instrument.

The results of the research by Palimbong et al. (2018), obtained from 30 EBAS questions, 26 questions were declared fit and 4 questions were not fit because they did not meet the criteria *Outfit MNSQ* and *Outfit ZSTD*. If the question has been declared fit, it means that it meets the criteria and can guarantee that the level of understanding of students is indeed tested through appropriate and quality items.

Based on these data, it can be said that a good test instrument is an instrument that can be understood by respondents well so that the test instrument is feasible to use. Achieving the criteria for content validity, construct validity, and item validity in the developed product, a final product is obtained in the form of a HOTS assessment instrument to measure valid problem solving abilities.

CONCLUSION

Based on the research that has been done, all the aspects measured are obtained by the Aiken'V coefficient above 0.75 with very good

criteria. It can be concluded that the HOTS assessment instrument to measure problem-solving ability is proven to be valid. Construct validity met the requirements with a value below 10%, and there were 16 items met the requirements *Outfit MNSQ*, *Outfit ZSTD*, and *Pt Measure Corr* and also 4 questions are said to be invalid because they do not meet the requirements *outfit MNSQ*, *Outfit ZSTD*, and *Point Measure Correlation (Pt Measure Corr)* that are the questions number 4, 7, 17 and 20.

ACKNOWLEDGMENTS

The authors would like to acknowledge the support from Chemistry Education Study Program, Postgraduate of Universitas Negeri Semarang.

REFERENCES

- Abidin, A. Z., Istiyono, E., Fadilah, N., & Dwandaru, W. S. B. (2019). A computerized adaptive test for measuring the physics critical thinking skills. *International Journal of Evaluation and Research in Education*, 8(3), 376–383. <https://doi.org/10.11591/ijere.v8i3.19642>
- Aisah, S. (2020). Pengembangan Instrumen Penilaian Higher Order Thinking Skills (HOTS) Pada Mata Pelajaran Korespondensi Kelas X OTP di SMK Negeri 1 Jombang. *Jurnal Pendidikan Administrasi Perkantoran (JPAP)*, 8(2015), 194–204. <https://journal.unesa.ac.id/index.php/jpap>
- Andrian, D., Kartowagiran, B., & Hadi, S. (2018). The instrument development to evaluate local curriculum in Indonesia. *International Journal of Instruction*, 11(4), 921–934. <https://doi.org/10.12973/iji.2018.11458a>
- Andriani, F., Indrowati, M., & Sugiharto, B. (2021). Biologi Analysis items of the four-tier immune system multiple choice test instrument using rasch analysis. *Biosfer: Jurnal Pendidikan Biologi*, 14(1), 99–119. <https://doi.org/https://doi.org/10.21009/biosferjpb.18020>
- Arifin, Z. (2017). Kriteria Instrumen dalam Suatu Penelitian. *Jurnal THEOREMS (The Original Research of Mathematics)*, 2(1), 28–36.
- Astuti, T. N., Sugiyarto, K. H., & Ikhsan, J. (2020). Effect of 3D visualization on students' critical thinking skills and scientific attitude in chemistry. *International Journal of Instruction*, 13(1), 151–164. <https://doi.org/10.29333/iji.2020.13110a>
- Bashooir, K., & Supahar. (2018). Validitas dan reliabilitas instrumen asesmen kinerja literasi sains pelajaran Fisika berbasis STEM. *Jurnal Penelitian Dan Evaluasi Pendidikan*, 22(2), 168–181. <https://doi.org/10.21831/pep.v22i2.20270>
- Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quinonez, H. R., & Young, S. L. (2018). Best Practices for Developing and Validating Scales for Health, Social, and Behavioral Research: A Primer. *Frontiers in Public Health*, 6(June), 1–18. <https://doi.org/10.3389/fpubh.2018.00149>
- Budiastuti, D., & Bandur, A. (2018). *Validitas dan reliabilitas Penelitian dengan Analisis “dengan NVIVO, SPSS dan AMOS.”* Mitra Wacana Media. <https://doi.org/10.31219/osf.io/tr4m7>
- Calonge-pascual, S., Fuentes-jim, F., & Casaj, A. (2020). Design and Validity of a Choice-Modeling Questionnaire to Analyze the Feasibility of Implementing Physical Activity on Prescription at Primary Health-Care Settings. *Int. J. Environ. Res. Public Health*, 17(6627), 1–12. <https://doi.org/doi:10.3390/ijerph17186627>
- Chalkiadaki, A. (2018). A Systematic Literature Review of 21st Century Skills and Competencies in Primary Education. *International Journal of Instruction*, 11(3), 1–16.
- Darmana, A., Sutiani, A., Nasution, H. A., Ismanisa, I., & Nurhaswinda, N. (2021). Analysis of Rasch Model for the Validation of Chemistry National Exam Instruments. *Jurnal Pendidikan Sains Indonesia*, 9(3), 329–345. <https://doi.org/10.24815/jpsi.v9i3.19618>
- Festiana, I., Firman, H., Setiawan, A., & Muslim. (2020). Development and validation of concept mastery physics test on the electricity topics. *International Journal of Scientific and Technology Research*, 9(1), 772–776.
- Gündüz, A. Y., Alemdağ, E., Yaşar, S., & Erdem, M. (2016). Design of a problem-based online learning environment and evaluation of its effectiveness. *Turkish Online Journal of Educational Technology*, 15(3), 49–57.
- Habibah, F. N., & Widodo, A. T. (2017). Pengembangan Perangkat Pembelajaran Kontekstual Berpendekatan Inkuiri Terbimbing Materi KSP. *Journal of Innovative*

- Science Education*, 6(1), 66–74.
<https://doi.org/10.15294/jise.v6i1.17066>
- Harta, J. (2017). Pengembangan Soal Esai Berbasis HOTS untuk Menyelidiki Keterampilan Pemecahan Masalah Siswa SMA. *Jurnal Penelitian*, 21(1), 62–69.
- Hidayati, K., Budiyono, & Sugiman. (2019). Using alignment index and polytomous item response theory on statistics essay test. *Eurasian Journal of Educational Research*, 2019(79), 115–132.
<https://doi.org/10.14689/ejer.2019.79.6>
- Ichsan, I. Z., Sigit, D. V., Miarsyah, M., Ali, A., Arif, W. P., & Prayitno, T. A. (2019). HOTS-AEP: Higher order thinking skills from elementary to master students in environmental learning. *European Journal of Educational Research*, 8(4), 935–942.
<https://doi.org/10.12973/eujer.8.4.935>
- Ikhsanudin, & Subali, B. (2018). Content validity analysis of first semester formative test on biology subject for senior high school. *Journal of Physics: Conference Series*, 1097(1).
<https://doi.org/10.1088/1742-6596/1097/1/012039>
- Jayadiningrat, M. G., & Ati, E. K. (2018). Peningkatan Keterampilan Memecahkan Masalah Melalui Model Pembelajaran Problem Based Learning (Pbl) Pada Mata Pelajaran Kimia. *Jurnal Pendidikan Kimia Indonesia*, 2(1), 1.
<https://doi.org/10.23887/jpk.v2i1.14133>
- Kusuma, M. D., Rosidin, U., Abdurrahman, A., & Suyatna, A. (2017). The Development of Higher Order Thinking Skill (Hots) Instrument Assessment In Physics Study. *IOSR Journal of Research & Method in Education (IOSRJRME)*, 07(01), 26–32. <https://doi.org/10.9790/7388-0701052632>
- Marfu'i, L. N. R., Ilfiandra, & Nurhuda. (2019). The analysis of critical thinking skills test in social-problems for physics education students with Rasch Model. *Journal of Physics: Conference Series*, 1280(5). <https://doi.org/10.1088/1742-6596/1280/5/052012>
- Mujib, M. F. R. (2019). *Modul Penyusunan Soal Keterampilan Berpikir Tingkat Tinggi (HOTS)*. Direktorat Pembinaan Sekolah Menengah Atas.
- Musa, N. A. C., Mahmud, Z., & Baharun, N. (2017). Exploring students' perceived and actual ability in solving statistical problems based on Rasch measurement tools. *Journal of Physics: Conference Series*, 890(1).
<https://doi.org/10.1088/1742-6596/890/1/012096>
- Nugroho, K. M., Raharjo, S. B., & Masykuri, M. (2017). Pengembangan E-modul Kimia Berbasis Problem Solving dengan Menggunakan Moodle pada Materi Hidrolisis Garam untuk Kelas XI SMA/MA Semester II. *Jurnal Inkuiri*, 6(1), 175–180.
- Nurmawati, N., Driana, E., & Ernawati, E. (2021). Pemahaman Guru Kimia Sma Tentang Penilaian Kemampuan Berpikir Tingkat Tinggi Dan Implementasinya. *Edusains*, 12(2), 233–242.
<https://doi.org/10.15408/es.v12i2.13613>
- Ortega-Toro, E., García-Angulo, A., Giménez-Egido, J. M., García-Angulo, F. J., & Palao, J. M. (2019). Design, validation, and reliability of an observation instrument for technical and tactical actions of the offense phase in soccer. *Frontiers in Psychology*, 10(JAN), 1–9.
<https://doi.org/10.3389/fpsyg.2019.00022>
- Palimbong, J., Mujasam, & Allo, A. Y. T. (2018). Item Analysis Using Rasch Model in Semester Final Exam Evaluation Study Subject in Physics Class X TKJ SMK Negeri 2 Manokwari. *Physics Education Journal*, 1(1), 43–51. i.yusuf@unipa.ac.id
- Pandra, V., Kartowagiran, B., & Sugiman. (2021). Mathematics Test Development By Item Response Theory Approach And Its Measurement On Elementary School Students. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, 12(5), 464–483.
<https://doi.org/10.17762/turcomat.v12i5.994>
- Prasetya, A., Rosidin, U., & Herlina, K. (2019). Development of Instrument Assessment for Learning the Polytomous Response Models to Train Higher Order Thinking Skills (HOTS). *Journal of Physics: Conference Series*, 1155(1).
<https://doi.org/10.1088/1742-6596/1155/1/012032>
- Purba, S. E. D. (2018). Analisis model Rasch instrumen tes prestasi pada mata pelajaran dasar dan pengukuran listrik [Rasch model analysis of achievement test instruments on basic subjects and electrical measurements]. *Wiyata Dharma: Jurnal Penelitian Dan Evaluasi Pendidikan*, 6(2), 142.
- Purwasih, J. H. G. (2020). Kendala Calon Pendidik Dalam Membuat Soal Pilihan Ganda Higher Order Thinking (Hot). *Jurnal Sosial Humaniora*, 13(1), 12.
<https://doi.org/10.12962/j24433527.v13i1.6746>

- Quaigrain, K., & Arhin, A. K. (2017). Using reliability and item analysis to evaluate a teacher-developed test in educational measurement and evaluation. *Cogent Education*, 4(1). <https://doi.org/10.1080/2331186X.2017.1301013>
- Rahmawati, Y., Ramadhani, S. F., Afrizal, A., Puspitasari, M., & Mardiah, A. (2021). Development of Students' Conceptual Understanding through STEAM Project Integration in Thermochemistry. *JTK (Jurnal Tadris Kimiya)*, 6(1), 62–74. <https://doi.org/10.15575/jtk.v6i1.5498>
- Riswanda, J. (2018). Pengembangan Soal Berbasis Higher Order Thinking Skill (Hots) Serta Implementasinya di SMA Negeri 8 Palembang. *Jurnal Penelitian Pendidikan Biologi*, 2(1), 49–58.
- Riyani, R., Maizora, S., & Hanifah, H. (2017). Uji Validitas Pengembangan Tes Untuk Mengukur Kemampuan Pemahaman Relasional Pada Materi Persamaan Kuadrat Siswa Kelas Viii Smp. *Jurnal Penelitian Pembelajaran Matematika Sekolah (JP2MS)*, 1(1), 60–65. <https://doi.org/10.33369/jp2ms.1.1.60-65>
- Saidi, S. S., & Siew, N. M. (2019). Reliability and Validity Analysis of Statistical Reasoning Test Survey Instrument using the Rasch Measurement Model. *International Electronic Journal of Mathematics Education*, 14(3), 535–546. <https://doi.org/10.29333/iejme/5755>
- Saraswati, S., Rodliyah, I., & Rahmawati, N. D. (2021). Analisis Instrumen Penilaian Berbasis Higher Order Thinking Skills pada Mata Kuliah Matematika Lanjut. *Inomatika*, 3(2), 138–151. <https://doi.org/10.35438/inomatika.v3i2.275>
- Setiawan, J., Sudrajat, A., Aman, & Kumalasari, D. (2021). Development of higher order thinking skill assessment instruments in learning Indonesian history. *International Journal of Evaluation and Research in Education*, 10(2), 545–552. <https://doi.org/10.11591/ijere.v10i2.20796>
- Siswanti, S., Saputro, S., & Utomo, S. B. (2016). Pengembangan Modul Termokimia Berbasis Problem Solving Untuk Siswa Sma/Ma Kelas Xi Semester 1 Kurikulum 2013. *INKUIRI: Jurnal Pendidikan IPA*, 5(1), 28–36. <https://doi.org/10.20961/inkui.v5i1.9500>
- Sudarmin, S., Mursiti, S., & Asih, A. G. (2018). The use of scientific direct instruction model with video learning of ethnoscience to improve students' critical thinking skills. *Journal of Physics: Conference Series*, 1006(1). <https://doi.org/10.1088/1742-6596/1006/1/012011>
- Sugiharni, G. A. D. (2017). Validitas Isi Instrumen Pengujian Modul Digital Matematika Diskrit Berbasis Open Source di STIKOM Bali. *E-Proceedings KNS&I STIKOM Bali*, 678–684. <http://knsi.stikom-bali.ac.id/index.php/e proceedings/article/view/123/118%0Ahttp://knsi.stikom-bali.ac.id/index.php/e proceedings/article/view/123>
- Sumintono, B., & Widhiarso, W. (2015). *Aplikasi Pemodelan Rasch: pada Assessment Pendidikan* (Issue October). Trim Komunikata.
- Supriyadi. (2021). Evaluation Instrument Development for Scientific Writing Instruction with A Constructivism Approach. *Technium Social Sciences Journal*, 21, 345–363.
- Susilaningsih, E., Nuswawati, M., Aprilia, N., & Luthfiah, A. (2021). Dissemination of test instruments as product of the development research to measure the problem-solving ability of class X students by online in the pandemic period. *Journal of Physics: Conference Series*, 1918(3). <https://doi.org/10.1088/1742-6596/1918/3/032024>
- Talmi, I., Hazzan, O., & Katz, R. (2018). Intrinsic Motivation and 21st-Century Skills in an Undergraduate Engineering Project: The Formula Student Project. *Higher Education Studies*, 8(4), 46. <https://doi.org/10.5539/hes.v8n4p46>
- Tim Pusklat Pegawai. (2016). *PENILAIAN HASIL BELAJAR*. Pusklat Pegawai Kemendikbud.
- Wardany, K., Sajidan, & Ramli, M. (2017). Pengembangan Penilaian Untuk Mengukur Higher Order Thinking Skills Siswa. *Jurnal Inkuiri*, 6(2), 1–16. <http://jurnal.uns.ac.id/inkui>
- Widana, I. W. (2017). Modul Penyusunan Soal Higher Ordher Thinking Skill (HOTS). In *Books* (Vol. 53, Issue 9). Direktorat Pembinaan Sekolah Menengah Atas.
- Widana, I. W. (2018). Higher Order Thinking Skills Assessment towards Critical Thinking on Mathematics Lesson. *International Journal of Social Sciences and Humanities (IJSSH)*, 2(1), 24–32. <https://doi.org/10.29332/ijssh.v2n1.74>
- Widiyawati, Y., Nurwahidah, I., & Sari, D. S. (2019). Pengembangan Instrumen Integrated Science Test Tipe Pilihan Ganda Beralasan Untuk

