Formatif: Jurnal Ilmiah Pendidikan MIPA, September 2021, 11 (2), 139-150
http://dx.doi.org/10.30998/formatif.v11i2.5922
p-ISSN: 2088-351X e-ISSN: 2502-5457
Accredited (S2) by Ministry of Research and Technology of Indonesia No. 148/M/KPT/2020
Available online at https://journal.lppmunindra.ac.id/index.php/Formatif/index

Validation of Instrument Multiple Representations for Analyzing the Multiple Representation Capability of Students in Hydrocarbon Materials

Antonia Fransiska Laka (*), Hari Sutrisno
Universitas Negeri Yogyakarta, DI Yogyakarta 55281, Indonesia

Received: February 25, 2020
Revised: February 20, 2021
Accepted: July 26, 2021

Abstract

This study aims to validate multiple representational instruments to analyze the ability of multiple representations of students on hydrocarbon material. This research uses a descriptive quantitative method with a non-experimental approach. This research uses the stratified purposive sampling method with 123 students who will respond to 35 items of multiple-choice questions covering macroscopic, microscopic, symbolic, and mathematical aspects. The data analysis technique used in the research is qualitative data analysis and quantitative data analysis. The Rasch model in this research analyzed instruments such as uni-dimensionality, item fit, test reliability, and difficulty level of the item. The data analysis shows that the average Aiken index is 0.961 on the substance aspect, 0.93 on the construction aspect, and 0.950 on the language aspect for the theoretical validation results. The highest Aiken index is 1.000, and the lowest is 0.896. Uni-dimensionality was 32.4%, the result of the item fit analysis obtained 1 item that was not fit, namely item number 29, and for the reliability test results: the person reliability value was 0.65, and the item reliability was 0.97. The analysis results of the difficulty level of the items on the instrument of measuring the cognitive abilities of students with multiple representation types were nine items in the easy category, 14 items in the medium category, and 11 items in the difficult category. Therefore, based on the resulting validity and reliability categories, the compiled test instrument can be used as a tool to measure students' multiple representation abilities.

Keywords: Multiple Representation, descriptive quantitative, Rasch model and hydrocarbon material.

(*) Corresponding Author: 23antonialaka@gmail.com, Phone number: (+62)85333900377

How to Cite: Laka A.F. & Sutrisno, H. (2021). Validation of instrument multiple representations for analyzing the multiple representation capability of students in hydrocarbon materials. *Formatif: Jurnal Ilmiah Pendidikan MIPA*, 11 (2): 139-150. http://dx.doi.org/10.30998/formatif.v11i2.5922

INTRODUCTION

Chemistry is the study of the structure, properties, reactions of elements and substances. Chemistry includes the notion of macroscopic, submicroscopic, and symbolic aspects (Talanquer, 2011). The use of different representations can help students connect one aspect to another in a better way. Conceptual understanding is the key to learning chemistry. Students can have a strong understanding of chemical concepts when they can relate their insights into different representations (Hernandez et al., 2014; Wu & Shah, 2003), and how to relate each new concept or fact in three macroscopic aspects: how chemical phenomena can be observed using the five human senses such as color, smell, and others. Submicroscopic is the interaction or form of invisible molecules that include atoms, molecules, and ions. Symbolic representation consists of formulas, equations,

symbols, mathematics, and graphics (Milenkovic et al., 2014). However, current chemistry teaching rarely helps students connect multiple representations. This teaching method often causes student confusion which harms student motivation and achievement in chemistry class (Adedoyin & Mokobi, 2013). Students have more difficulties in studying microscopic and symbolic representations than macroscopic representations. This happens because the microscopic and symbolic aspects are abstract and invisible, requiring a reasoning process (Demircioglu et al., 2013; Chandrasegaran et al., 2007).

Assessment is an essential component in education. The interaction between assessment, curriculum, and instruction is essential to improve the teaching and learning process (Ghazali, 2016). Research conducted by Setiadi (2016) shows that many teachers ignore the pre-test function at the planning stage and do not perform instrument analysis before the assessment process. Teachers also experience difficulties obtaining assessment results to determine student learning progress and student learning difficulties (Maisyaroh et al., 2014).

Research credibility refers to how accurate the answers to research questions are or the strength of research conclusions. The indicators of success in the measurement instrument are the reliability and validity of the measurements. Validity refers to the accuracy of the measurement. Validity refers to how well the assessment tool can measure the desired result (Kimberlin & Winterstein, 2008; Sullivan, 2011). Reliability shows that multiple-choice measures something consistently. Reliability is not the type of knowledge, ability, and or skill that is measured. Therefore, evidence of validity becomes an essential aspect before concluding that multiple-choice sequences are valid (Peeters et al., 2013). A field trial process determines the quality of multiple-choice to evaluate the characteristics of each item. Many teachers do not perform instrument analysis before the assessment process. Many teachers never do a pre-test and multiple-choice analysis because they do not have the competence to analyze tests (Sanoya et al., 2017). The teacher arranges the assessment test instrument only according to the material that has been taught to students. The arranged tests only focus on setting numbers and do not stimulate how students should solve problems. As a result, students' abilities are not too prominent in high thinking, critical, creative, and problem-solving (Baharudin, 2013).

Learning about the preparation of multiple representation-based test instruments is essential because it helps measure and analyze students' multiple representations of hydrocarbon material. This helps teachers in using the right strategy, approach, or learning model. This research provides many benefits for students, teachers, schools, and stakeholders. An excellent and credible instrument with validation and reliability is needed to measure the multiple representation abilities of students. This study aims to validate the test instrument to measure multiple representations of high school students in Yogyakarta on hydrocarbon material. The research problem discussed relates to hydrocarbon material only.

METHODS

This research was a quantitative descriptive study using research samples from the same class XI on hydrocarbon material. This research used a non-experimental design where the researcher did not give special treatment to students. The population of this study was all eleventh-grade high school students in Yogyakarta. The sample was taken using the stratified purposive sampling method from 3 high schools consisting of 123 students as the research sample.

The data collection instrument used in this study was multiple-choice questions based on multiple representations consisting of 35 items to determine the student's ability to answer multiple representation type questions. The test instrument was given when the hydrocarbon material has been taught. The instrument was equipped with an answer sheet, instructions for solving the questions, and answer keys. The instrument items were arranged systematically by considering the breadth and depth of the material.

Table 1. Grid of multiple representation instruments for hydrocarbon materials

Item Number	Theory	Item Indicator	Cognitive Level	Aspect of Representation
1	Hydrocarbon compounds in everyday life	Students can analyze the symbolic representation of a substance by analyzing the macroscopic representation of the questions presented.	C4	Macroscopic Symbolic
2		Students can analyze hydrocarbon compounds from the given chemical formula.	C4	Macroscopic Symbolic
3		Students can recognize a material by analyzing the symbolic representation of the questions presented.	C4	Macroscopic Symbolic
4	Specificity of Carbon Atom	Students can analyze the elements C, H, and O in carbon compounds through experiments.	C4	Macroscopic Symbolic
5		Students can analyze the peculiarities of carbon atoms in carbon compounds.	C4	Microscopic Symbolic
6	Primary, secondary, tertiary C atom	Students can distinguish primary, secondary, tertiary, and quaternary C atoms.	C2	Macroscopic Symbolic
7,8,9		Students can distinguish primary, secondary, tertiary, and quaternary C atoms.	C2	Symbolic
10	The names of alkanes, alkenes and alkyne compounds are in accordance with IUPAC regulations	According to IUPAC, students can analyze hydrocarbon names given the symbolic structure of a hydrocarbon compound and its name.	C4	Symbolic
11		Given a symbolic representation of alkane compounds, students can analyze these alkane compounds.	C4	Macroscopic Symbolic
12		Students can analyze the names of alkene compounds according to IUPAC rules.	C4	Symbolic
13		Students can interpret the symbolic representation of an alkene compound and analyze the name of the compound.	C4	Macroscopic Symbolic
14	The names of alkanes, alkenes and alkyne compounds are in accordance with IUPAC regulations	Students can interpret the symbolic representation of a compound and analyze the IUPAC name of the compound.	C4	Symbolic

Item Number	Theory	Item Indicator	Cognitive Level	Aspect of Representation
15		Students can analyze the names of hydrocarbons from the chemical formula of these compounds.	C4	Symbolic
16	The structure of alkanes, alkenes, and alkyne compounds	Students can analyze the microscopic representation of an alkane group of compounds from a given compound name.	C4	Macroscopic Microscopic
17		Given a visible picture of a material containing an alkene compound, students can analyze the symbolic representation of the structure of the compound.	C4	Microscopic Symbolic
18		Students can analyze the symbolic representation of alkyne compounds from given compound names.	C4	Symbolic
19,20, 21,22	Boiling point of hydrocarbons	Students can categorize the order of the boiling points of several hydrocarbon compounds based on their relative molecular masses and structures.	C6	Symbolic
23,24, 25,26	Structural isomers (framework, position, function) and geometric isomers (cis, trans)	Students can predict structural isomers (frame, position, function) and geometric isomers (cis, trans).	C2	Symbolic
27		Students can predict structural isomers (frame, position, function) and geometric isomers (cis, trans).	C2	Symbolic Mathematics
28	Types of reactions of alkanes, alkenes, and alkyne	Students can analyze the types of reactions of alkanes, alkenes, and alkyne.	C4	Symbolic Mathematics
29,30, 31		Students can analyze the types of reactions of alkanes, alkenes, and alkyne.	C4	Symbolic
32	Types of reactions of alkanes, alkenes, and alkyne	Students can analyze the types of reactions of alkanes, alkenes, and alkyne.	C4	Microscopic Symbolic
33		Students can analyze the types of reactions of alkanes, alkenes, and alkyne.	C4	Symbolic Mathematics
34,35		Students can analyze the types of reactions of alkanes, alkenes, and alkyne.	C4	Symbolic

The purpose of this study was to validate the instrument test. Therefore, validation was carried out in two ways, namely theoretical validation, and empirical validation, for theoretical validation carried out by expert judgment and education practitioners. The prepared instrument was then validated by the validator using V-Aiken—for empirical validation, done by testing the instrument to 123 students from 3 schools in Yogyakarta city. The empirical validation of the instruments was analyzed using the Rasch Model to determine the validity, reliability, fit items, item difficulty level, and function of information.

RESULTS & DISCUSSION

Result

The application of the Rasch Model in this study is essential in determining the validity and reliability of the instrument to define valid and reliable item constructs and provide a clear definition of construct that can be measured consistently with theory. This model can be used effectively for measured items and good response patterns (Milenkovic, Segedinac & Hrin, 2014).

The instrument test consists of 35 multiple choice items equipped with a grid and an answer key. The judgment of experts and education practitioners acts as a theoretical validation for questions that have been arranged in the aspects of substance, construction, and language. Then the validation results were analyzed using V-Aiken. Table 2 shows the results of the Aiken index analysis.

The results of expert judgment and education practitioners (4 raters in total) and a scale of 5 indicate that the Aiken average index is 0.973 on the substance aspect, 0.973 on the construction aspect, and 0.966 on the language aspect. The highest Aiken index is 1,000, and the lowest is 0.938. Items can be accepted if the index obtained is more significant than 0.88. The analysis results showed that the total Aiken index of 35 items was more than 0.88. It can be concluded that the multiple representational instruments arranged have met theoretical validity and can be used for testing empirical validity (Aiken, L. R., 1980).

Table 2. Aiken Index for Substance, Construction and Language Aspects

Item Number	Aiken Index of Content Aspect	Aiken Index of Construction Aspect	Aiken Index of Language Aspect	Item Number	Aiken Index of Content Aspect	Aiken Index of Construction Aspect	Aiken Index of Language Aspect
1	1,000	0,938	0,938	9	1,000	1,000	0,938
2	1,000	1,000	0,938	10	0,938	0,938	1,000
3	1,000	1,000	1,000	11	1,000	1,000	0,938
4	1,000	0,938	0,938	12	0,938	0,938	0,938
5	1,000	1,000	0,938	13	1,000	1,000	0,938
6	1,000	0,938	0,938	14	0,938	1,000	0,938
7	0,938	0,938	1,000	15	1,000	0,938	1,000
8	1,000	1,000	1,000	16	0,938	0,938	1,000
17	0,938	1,000	0,938	27	0,938	1,000	1,000
18	1,000	0,938	0,938	28	1,000	0,938	1,000
19	0,938	1,000	0,938	29	0,938	1,000	1,000
20	0,938	0,938	1,000	30	0,938	1,000	0,938
21	1,000	1,000	1,000	31	0,938	0,938	0,938
22	1,000	0,938	0,938	32	1,000	1,000	1,000
23	1,000	1,000	0,938	33	1,000	1,000	0,938
24	0,938	1,000	1,000	34	1,000	1,000	0,938
25	1,000	1,000	1,000	35	0,938	0,938	1,000
26	0,938	0,938	1,000				
				Avg	0,973	0,973	0,966

Empirical validation was carried out to prove construct validity and determine whether the sample measurement was carried out using the SPSS 16 Program. Subjects in

the empirical validation were 123 eleventh grade students of 3 senior high schools in Yogyakarta. The first stage is to test the adequacy of the sample. The measure (size) or several samples is categorized as sufficient and fulfills the requirements for analysis of the results of the KMO-MSA test analysis are more than 0.05 (>0.50), and the significance of the Bartlett test is less than 0.01 (< 0.01). Table 3 shows the results of the sample adequacy test.

Table 3. The results of KMO-MSA and Bartlett tests

Test Analysis	Test Result	Criteria	Conclusion
KMO-MSA	0,796	>0,50	Sample measure (size)
Bartlett Significance Test	0,00	<0,01	is eligible for further analysis.

Table 3 shows that the results of the KMO-MSA test analysis are 0.796 and the Bartlett Significance Test is 0.00. It can be concluded that the test sample size qualifies for further analysis.

The uni-dimensionality of the instrument is an important measure to evaluate whether the instrument can measure what it should be measured, namely, to measure the ability of multiple student representations that can be obtained from dimensionality items in the Winstep program. Uni-dimensionality can be analyzed by looking at the measurement of natural variance data, which shows the minimum uni-dimensionality requirement of 20%. If the value is more than 40%, it means better, and if the value is more than 60%, it means special. Another thing that can be informed is the variance that the instrument cannot explain, ideally not exceeding 15% (Sumintono, B. & Widhiarso, W., 2014).

Table of STANDARDIZED RESIDUAL variance (in Eigenvalue units)

		-- Empirical --			Modeled
Total raw variance in observations	=	51.8	100.0%		100.0%
Raw variance explained by measures	=	16.8	32.4%		32.3%
Raw variance explained by persons	=	4.5	8.6%		8.6%
Raw Variance explained by items	=	12.3	23.8%		23.7%
Raw unexplained variance (total)	=	35.0	67.6%	100.0%	67.7%
Unexplned variance in 1st contrast	=	2.4	4.6%	6.9%	
Unexplned variance in 2nd contrast	=	2.1	4.1%	6.1%	
Unexplned variance in 3rd contrast	=	2.1	4.1%	6.0%	
Unexplned variance in 4th contrast	=	1.9	3.7%	5.5%	
Unexplned variance in 5th contrast	=	1.9	3.6%	5.3%	

Figure 1. The results of unidimensional analysis on winstep program

Based on Figure 1, the results of the measurement of natural variance explained by measures obtained were 32.4%. This shows that the test instrument compiled can only measure 32.4% of the desired ability, while there are 67.7% that cannot be explained. However, the minimum dimension of 20% has been fulfilled, and it can be concluded that the instrument compiled can only measure one multiple representation capability (Sumintono, B. & Widhiarso, W., 2015).

Item fit describes whether the item is functioning normally to take measurements or not. Analysis of fit items was carried out using the help of the Winstep program. Table 4 shows the criteria for the level of item suitability following Sumintono & Widhiarso (2014).

Table 4. The criteria of fit item

Item Fit	Criteria
Outfit Mean Square (MNSQ)	0,5 < MNSQ < 1,5
Outfit Z-Standard (ZSTD)	-2 < ZSTD < +2
Point Measure Correlation (Pt Mean Corr)	0,4 < Pt. Measure Corr < 0,85

If the test instrument meets one criterion, it can be said to be fit. If it does not meet 1 criterion, for example, the MNSQ outfit does not meet the criteria, it can be matched again with ZSTD or Pt. Mean Corr outfit (Sumintono & Widhiarso, 2014). Table 5 shows the results of the item fit analysis of this study.

Table 5 shows the results of the item fit analysis to measure the multiple representation abilities of students. According to Table 5, there is one unfit item from 35 items after being analyzed using the Rasch model PCN 1-PL. The results of the item fit analysis obtained one unfit item, namely item number 29. The unfit items were reduced due to an indication that students had misconceptions about these items, so there were 34 items left to measure the ability of multiple representations of students.

Reliability is part of the validity assessment (Sullivan, 2011). Reliability refers to whether a rating instrument provides the same results whenever used in the same setting with the same type of subject. Reliability means consistent or reliable results. The test reliability analysis is seen from the Cronbach alpha value obtained from the Winstep program's analysis, namely from summary statistics. According to Bhatnagar et al. (2014), using criteria to see reliability based on the Cronbach Alpha value can be seen in Table 6.

Table 5. The result of item fit analysis

Item Number	MNSQ Outfit	ZSTD Outfit	Pt. Mean Corr Outfit	Result	Item Number	MNSQ Outfit	ZSTD Outfit	Pt. Mean Corr Outfit	Result
1	1.11	0.6	0,19	fit	19	0.86	- 1.4	0.47	fit
2	1.03	0.2	0.25	fit	20	1.14	1.6	0.12	fit
3	1.09	0.7	0.17	fit	21	0.91	- 0.5	0.37	fit
4	0.99	- 0.1	0.30	fit	22	0.86	- 0.5	0.38	fit
5	1.01	0.1	0.30	fit	23	0.85	- 1.3	0.44	fit
6	0.74	- 2.1	0.57	fit	24	0.89	- 0.9	0.38	fit
7	0.78	- 0.4	0.23	fit	25	1.09	0.9	0.19	fit
8	0.88	- 0.6	0.37	fit	26	1.17	1.1	0.17	fit
9	0.45	- 0.5	0.20	fit	27	1.18	1.7	0.09	fit
10	0.97	- 0.3	0.34	fit	28	1.25	1.9	0.05	fit
11	0.98	0.0	0.26	fit	29	2.27	4.3	0.05	Does not fit
12	1.12	0.6	0.25	fit	30	1.02	0.3	0.27	fit
13	0.97	- 0.2	0.37	fit	31	1.22	1.1	0.10	fit
14	0.86	- 1.1	0.48	fit	32	0.99	- 0.1	0.33	fit
15	0.84	- 1.2	0.47	fit	33	1.34	1.1	0.07	fit
16	0.93	- 0.4	0.42	fit	34	0.93	- 0.6	0.37	fit
17	0.90	- 0.4	0.30	fit	35	0.97	0.0	0.30	fit
18	0.78	- 1.0	0.41	fit					

Table 6. The criteria of reliability

Alpha Cronbach Value	Category
$\alpha < 0.5$	Unacceptable
$0.5 < \alpha < 0.6$	Poor
$0.6 < \alpha < 0.7$	Acceptable
$0.7 < \alpha < 0.9$	Good
$\alpha > 0.9$	Excellent

The Alpha Cronbach value obtained in this study was 0.67. This shows that the reliability category of the instrument test to measure the multiple representation ability of students is included in the accepted category (Bhatnagar et al., 2014). Item consistency can be used for other samples with the same or nearly identical characteristics (Shah et al., 2017).

SUMMARY OF 124 MEASURED Person

	TOTAL SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	21.1	35.0	.60	.42	.99	.0	1.01	.1
S.D.	4.2	.0	.74	.02	.21	1.1	.47	.9
MAX.	29.0	35.0	2.14	.59	1.59	2.7	4.41	2.9
MIN.	4.0	35.0	-2.69	.40	.55	-2.5	.39	-1.7
REAL RMSE	.44	TRUE SD	.59	SEPARATION	1.35	Person	RELIABILITY	.65
MODEL RMSE	.42	TRUE SD	.60	SEPARATION	1.43	Person	RELIABILITY	.67
S.E. OF Person MEAN = .07								

Figur

e 2. The results of person reliability analysis

SUMMARY OF 35 MEASURED Item

	TOTAL SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	74.7	124.0	.00	.24	1.00	.0	1.01	.1
S.D.	30.7	.0	1.42	.09	.08	1.0	.27	1.2
MAX.	122.0	124.0	2.84	.72	1.15	2.5	2.26	4.3
MIN.	14.0	124.0	-3.84	.19	.81	-1.9	.45	-2.1
REAL RMSE	.26	TRUE SD	1.40	SEPARATION	5.36	Item	RELIABILITY	.97
MODEL RMSE	.26	TRUE SD	1.40	SEPARATION	5.44	Item	RELIABILITY	.97
S.E. OF Item MEAN = .24								

Figure 3. The results of item reliability analysis

Table 7. The criteria of item difficulty

Criteria	Category
< -1	Easy
$-1 < b < +1$	Moderate
$> +1$	Difficult

Table 8. Shows the result of difficulty item in this study

Item Number	Difficulty Level	Category	Item Number	Difficulty Level	Category
1	0.25	Moderate	21	-1.85	Easy
2	-0.74	Moderate	22	-1.78	Easy
3	2.46	Difficult	23	0.77	Moderate
4	2.18	Difficult	24	1.76	Difficult
5	-0.30	Moderate	25	1.59	Difficult
6	-0.87	Moderate	26	1.94	Difficult
7	-1.78	Easy	27	1.50	Difficult
8	-1.40	Easy	28	1.65	Difficult
9	-3.55	Easy	29	Does not fit	
10	0.64	Moderate	30	1.10	Difficult
11	-1.16	Easy	31	1.35	Difficult
12	-1.61	Easy	32	-0.06	Moderate
13	-0.43	Moderate	33	2.19	Difficult
14	-0.75	Moderate	34	1.59	Difficult
15	-0.88	Moderate	35	-0.97	Moderate
16	-1.08	Easy			
17	-0.91	Moderate			
18	-1.42	Easy			
19	-0.40	Moderate			
20	0.98	Moderate			

The value of person reliability obtained from the analysis is 0.65, and the item reliability obtained from the analysis is 0.97. The value of person reliability is 0.65, and item reliability is 0.97. It can be concluded that the consistency of the answers from students is weak (accepted), and the quality of the items in the instrument has a remarkable reliability aspect (Sumintono & Widhiarso, 2015).

Based on classical theory, item difficulty is the percentage of students who answered an item correctly. The greater of test-takers who work on the item questions correctly, the easier it will be. If the exam questions are tough, then most of the test scores will be very low. If the item test is straightforward, then the test score will be very satisfying (Afolabi et al., 2016). The item difficulty level was analyzed using the Winstep program obtained from the item measure information. Table 7 shows the level of problem criteria according to Adedoyin & Mokobi (2013).

Table 8 provides information about the difficulty level of the items from the arranged instruments. The analysis results of the difficulty level of the items on the instrument of measuring the cognitive abilities of students with multiple representations types were nine items in the easy category, 14 items in the medium category, and 11 items in the difficult category.

Discussion

The result of this study is theoretical validity analyzed by expert judgment and education practitioner judgment (4 rater in total) and 5 rating scale category shows that the average index of Aiken is 0.973 for the substance and construction aspects, and 0.966 for the language aspect. The highest Aiken index is 1,000, and the lowest is 0.938. an item with an index of more than 0.88 is acceptable. The analysis results indicate that the total Aiken index from 35 item questions is more than 0.88. It can be concluded that the

multiple representation instrument fulfills the theoretical validity and can be used for empirical validity testing (Aiken, L. R., 1980).

The next step is empirical validation. The first phase is to do a sample adequacy test. The result of the KMO-MSA test analysis is 0.796, and Bartlett Significance Test is 0.00. It concludes that the sample size of the test is eligible for further analysis. The second phase is unidimensional analysis. Unidimensional is defined as the presence of a latent underlying data trait. The result shows that the instrument can only measure one representation of multiple capabilities (Sumintono & Widhiarso, 2015). The third phase is checking item fit. One unfit item is number 29 from 35 items after the analysis process using the Rasch model PCM 1-PL. The unfit items are removed for students' misconception on the items, leaving 34 items to measure the ability of multiple representations of students. The fourth phase is to measure item and person reliability. The Alpha Cronbach value of this study is 0.67. It shows that the reliability category in the test instrument to measure the multiple representation capabilities of students is categorized as received. Item and person reliability showed that It could be concluded that both consistencies of respondents' answers and quality of items in special instruments are good. The fifth phase is the analysis difficulty level of items. The analysis results of the difficulty level of the items on the instrument of measuring the cognitive abilities of students with multiple representations types were nine items in the easy category, 14 items in the medium category, and 11 items in the difficult category. The most challenging item is number 3, and the most straightforward item is number 9.

The instrument of assessment must be reliable and valid to get credible measurement results. Thus, the reliability and validity of each assessment instrument used to measure the study results should be analyzed. Validity refers to how accurate a study answers research questions or the strength of research conclusions (Sullivan, 2011). The reliability of test instruments is categorized as good (Bhatnagar & Many, 2014). Based on the value of person reliability and item reliability, it can be concluded that the consistency of answers from respondents and quality items in a particular instrument is suitable. It indicates that the arranged instruments can be categorized as good questions. The test instrument can be used to diagnose or measure students' ability of multiple representations of hydrocarbon. The arranged items determine the students' ability moderate level only (Sumintono & Widhiarso, 2015).

CONCLUSION

The test instrument has fulfilled the theoretical validity performed by the expert judgment and educational practitioners who have been analyzed using the V-Aiken index. The analysis results show that the test instruments have fulfilled theoretical validity and can be used for further analysis. The arranged test instruments have fulfilled the empirical validity with empirical evidence of 34 qualified items from a total of 35 items that have been analyzed using PCM 1-PL. The following research suggests that chemistry teachers can apply the instruments prepared to determine students' ability of multiple representations to help teachers decide the best strategy, approach, or appropriate learning models then students understand multiple representations in learning hydrocarbon. This instrument can be used as an example to prepare test instruments for other chemical contents.

ACKNOWLEDGEMENT

The authors are thankful to all teachers and students at state senior high school in Yogyakarta, many thanks for their supports on this study.

REFERENCES

Adedoyin, O. O., & Mokobi, T. (2013). Using IRT psychometric analysis in examining the quality of junior certificate mathematics multiple-choice examination test items. *International Journal of Asian Social Science, 3*(4), 992-1011.

Aiken, L. R. (1980). Content validity and reliability of single items or questionnaires. *Educational and Psychological Measurement, 40*(4), 955-959. DOI: 10.1177/001316448004000419

Ainsworth, S. (1999). The functions of multiple representations. *Computers & Education, 33*(2-3), 131-152. DOI: https://doi.org/10.1016/S0360-1315(99)00029-9

Baharudin. (2013). Menganalisis instrumen penilaian pembelajaran matematika pada materi segi empat sekolah menengah pertama negeri 1 Dompu. *Jurnal Kependidikan, 15*(1), 1-10.

Bhatnagar, R., Kim, J., & Many J. E. (2014). Candidate surveys on program evaluation: examining instrument reliability, validity, and program effectiveness. *American Journal of Educational Research, 2*(8), 683-690. DOI: 10.12691/education-2-8-18

Chandrasegaran, A. L., Treagust, D. F., & Mocerino, M. (2007). The development of a two-tier multiple-choice diagnostic instrument for evaluating secondary school students' ability to describe and explain chemical reactions using multiple levels of representation. *Chemistry Education Research and Practice, 8*(3), 293-307. DOI: 10.1039/B7RP90006F

Gabel, D. (1999). Improving teaching and learning through chemistry education research: A look to the future. *Journal of Chemical Education, 76*(4), 548-554. DOI: 10.1021/ed076p548

Ghazali, N. H. M. (2016). A reliability and validity of an instrument to evaluate the school-based assessment system: A pilot study. *International Journal of Evaluation and Research in Education, 5*(2), 148-157

Gilbert, J. K., & Treagust, D. (2009). *Models and modeling in science education: Multiple representations in chemical education.* Perth: Springer. DOI: 10.1007/978-1-4020-8872-8

Hernandez, G. E., Criswell, B. A., Kirk, N. J., Sauder, D. G., & Rushton, G. T. (2014). Pushing for particulate level models of adiabatic and isothermal processes in upper-level chemistry courses: a qualitative study. *Chemistry Education Research and Practice, 15,* 354-365. DOI: 10.1039/c4rp00008k

Johnstone, A. H. (1993). The development of chemistry teaching a changing response to changing demand. *The Forum Symposium on Fievolution and Evolution in Chemical Education, 70*(9), 701-705. DOI: 10.1021/ed070p701

Johnstone, A. H. (2000). Teaching of chemistry-logical or psychological? *Chemistry Education: Research and Practice in Europe, 1*(1), 9-15. DOI: 10.1039/A9RP90001B

Kimberlin, C. L., & Winterstein, A. G. (2008). Validity and reliability of measurement instruments used in research. *American Society of Health-System Pharmacists*, 65, 2276-2284

Maisyaroh, Zulkarnain W., Setyowati, A. J., & Mahanal, S. (2014). Masalah guru dalam implementasi kurikulum 2013 dan kerangka model supervisi pengajaran. *Manajemen Pendidikan, 24*(3), 213-220.

Milenkovic, D. D., Segedinac, M. D., & Hrin, T. N. (2014). Increasing high school students' chemistry performance and reducing cognitive load through an instructional strategy based on the interaction of multiple levels of knowledge representation. *Journal of Chemical Education*, A-H. DOI: 10.1021/ed400805p

Nakhleh, M. B., & Krajcik, J. S. (1994). Influence of levels of information as presented by different technologies on students' understanding of acid, base, and ph concepts. *Journal of Research in Science Teaching, 31*(10), 1077-1096. DOI: https://doi.org/10.1002/tea.3660311004

Peeters, M. J., Beltyukova, S. A., & Martin, B. A. (2013). Educational testing and validity of conclusions in the scholarship of teaching and learning. *American Journal of Pharmaceutical Education, 77*(9), 1-9

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor test: results and implications. *Journal of Educational Statistics, 4*(3), 207-230. DOI: 10.2307/1164671

Rosengrant, D., Etkina, E., & Heuvelen, A. V. (Juli 2006). An overview of recent research on multiple representations. *Physics Education Research Conference, New York, 883*, 149-152. DOI: http://dx.doi.org/10.1063/1.2508714

Sanova, A., Bakar, A., & Afrida. (2017). Standarisasi instrumen penilaian hasil belajar dengan program anates v4 bagi-guru SMPN 17 Kota Jambi, *Jurnal Pengabdian Masyarakat, 2*(1), 1-10. Setiadi, H. (2016). Pelaksanaan penilaian pada kurikulum 2013. *Jurnal Penelitian dan Evaluasi Pendidikan, 20*(2), 166-178. DOI: http://dx.doi.org/10.21831/pep.v20i2.7173

Shah, R. L. Z. R. M., Samad, M. H. A., Shah, R. N. F. A. R. M., Adenan, N. H. (2017). Validity and reliability of graphing calculator skills test items for circles topic (CGCST) using Rasch measurement model analysis: a pilot study. *International Journal of Education and Research, 5*(8), 189-200

Sullivan, G. M. (2011). A primer on the validity of assessment instruments. *Journal of Graduate Medical Education*, 119-120

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

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

Thomas, G. P. (2017). "Triangulation": An expression for stimulating metacognitive reflection regarding the use of 'triplet' representations for chemistry learning. *Chemistry Education Research and Practice, 18*(4), 1-48. DOI: 10.1039/C6RP00227G

Wu, H. K., Krajcik, J. S., & Soloway, E. (2001). Promoting understanding of chemical representations: students' use of a visualization tool in the classroom. *Journal of Research in Science Teaching, 38*(7), 821-842. DOI: https://doi.org/10.1002/tea.1033

Wu, H. K., & Shah, P. (2003). Exploring visuospatial thinking in chemistry learning. *Science Education, 88*(3), 465-492. DOI: 10.1002/sce.10126/full