Jurnal Pendidikan Sains Universitas Muhammadiyah Semarang.
October Vol. 10(2) pp 44-52 ISSN:2339-0786
DOI: https://doi.org/10.26714/jps.10.2.2022.44-52

JPS JOURNAL PENDIDIKAN SAINS
JURNAL PENDIDIKAN SAINS
p-ISSN : 2339-0786 o-ISSN : 2502-1443
UNIVERSITAS MUHAMMADIYAH SEMARANG

Feasibility Analysis of Four Tier Multiple Choice Diagnostic Test Instruments on Reaction Rate Material

R. Usman Rery [a,1], Abdullah [b,2], Huswatun Hasanah [c,3,*]

[a,b,c] Chemistry Education Program, University of Riau, Pekanbaru, 28293, Indonesia
[1] rery1959@gmail.com; [2] abdoel71@gmail.com; [3] huswatun.hasanah1044@student.unri.ac.id*
* corresponding author

Article history

Submission : 2022-01-26
Revised : 2022-09-09
Accepted : 2022-11-02

Keyword

Feasibility
Diagnostic Test
Four Tier Multiple Choice
Reaction Rate

Abstract

This study aimed to analyze the feasibility of the four-tier multiple-choice diagnostic test instrument on the reaction rate material. The type of research used was development research referring to R&D (Research and Development) with a 4D development model. The 4D stage includes four stages: Define, Design, Develop, and Disseminate. The research subjects at the time of the small group trial were 20 students who had studied the reaction rate material. The data collection instrument used was a validation sheet. The data analysis technique was carried out through the calculations of validity, reliability, discriminatory power, and tier of the difficulty of the questions. The results showed that the four-tier multiple choice diagnostic test was feasible to identifying students' misconceptions with an average percentage of the content validity test of 96.97% with valid criteria, the value of the construction validity test was r_{count} > 0.444 with the category of all items valid, the score the reliability test was 0.83 (reliable) with very high criteria, the value of the discriminating power of the questions was 0.5 with good criteria, and the value of the difficulty tier of the questions was 0.5 with moderate criteria.

1. INTRODUCTION

Learning is an activity designed by educators to enable students to learn new skills and values in a structured process at the design, implementation, and evaluation stages (Sagala, 2010). Referring to these, ideal learning can be carried out by emphasizing the ability to observe, classify, conclude, predict, and communicate, and can build an individual understanding in solving a problem (Winaryati, 2014). In the process of building the concept of knowledge independently, not all the concepts that are built are in accordance with the actual concept, which will produce different concepts or experience misconceptions.

One of the things that can prove that learning has been carried out well is the evaluation system used. Evaluation plays a pivotal role in determining the success of students after the learning process. This also determines whether the learning that has taken place is going well or not. In connection with what Fortuna, et al. (2013) explained that evaluation is a determinant in graduation or failure of learning outcomes. In addition, it can also affect the further learning process. Often in the learning process, students have difficulty understanding the concepts they built independently and many of them misunderstand the concepts. However, the teacher does not aware of and knows how to detect the misconceptions occurred. Therefore, a special evaluation tool is required to diagnose misunderstandings from the concepts studied.

Based on the chemistry teacher interviews, it was found that the students have difficulties understanding the abstract and complex concept of reaction rate because the basic concepts obtained from their understanding are not correct. This can result in students experiencing misconceptions. This misconception has not been identified because the teachers only measure the tier of understanding of students with ordinary multiple-choice test instruments and have never tested students' conceptual understanding using other special tests.

The misconception is an understanding or mastery of concepts that are not in accordance with the intended meaning. Misconceptions can impact poor learning outcomes and it is difficult to understand the concept related to the material. It is important to diagnose misconceptions first to find the weaknesses of students in certain parts of the material so that it can be used as a reference to determine better learning in the future. This requires a form of evaluation that can describe the misconceptions experienced by the students.

One of the evaluation instruments that can detect misconceptions is a diagnostic test. Diagnostic tests in terms of function are defined as tests that can describe the strengths and weaknesses of students when learning something so that they can be used as guidelines for follow-up. In addition, diagnostic tests can show students' thinking skills in answering the questions provided even though the answers they choose are wrong. Diagnostic tests are also very easy to perform and assess, so it is very helpful for researchers to explain students' understanding of concepts.

The type of diagnostic test that can be used is the four-tier multiple-choice diagnostic test. The 4-tier diagnostic test is an update from the previous test tier, namely the 3-tier diagnostic test because the 4-tier type has two degrees of certainty located at the first and third tiers, which makes it superior in identifying misconceptions compared to the 3-tier diagnostic test type (Ismail et al., 2015). The advantages of the 4-tier diagnostic test type are: (1) being able to compare the tier of certainty of students in selecting answers and reasons (2) detecting student misunderstandings in more depth, (3) determining parts of the material that required more attention (4) compiling appropriate learning better (Fariyani et al., 2015). Mubarak et al. (2016) have developed a three-tier multiple-choice diagnostic test to identify the misconceptions of class XI students. The results show that the instrument developed is good and valid with a CVR (content validity ratio) value of 0.99 and a mean of 1.52 with an instrument reliability value of 0.90.This study aims to analyze the feasibility of the four-tier multiple-choice diagnostic test instrument on the reaction rate material.

2. METHOD

This research was carried out in grade XII Mathematics and Science at SMAN 2 *Pekanbaru*. The duration of data collection started from June 2021 to August 2021. This research was development research that referred to R and D (Research and Development) with a 4D development model. The 4D stage has four stages: define, design, develop, and disseminate or be adapted in the 4P: definition, design, development, and deployment. The research subjects during the small group trial were 20 students who had previously studied the subject of reaction rates. The data collection tool was in the form of a validation sheet to determine the quality or feasibility of a test instrument before being tested.

The data analysis technique was carried out by calculating:

1) Validity

a. Content Validity

Content validity was analyzed by calculating the score of each assessment on the validation sheet carried out by three material expert validators. The assessment of the instrument was based on aspects of the material, construction, language, and appearance using a Likert scale of 1-4. Table 1 shows the categories of the Likert scale rating.

Table 1. Category of Likert scale assessment 1-4 (Sugiyono, 2017)

Rating Scale	Criteria
4	SS: Very Suitable
3	S: Suitable
2	KS: Less Suitable
1	TS: Not Suitable

To calculate the percentage of each validator assessment can be formulated as follows. Content validity criteria can be seen as follows:

$$\mathrm{P} = \frac{n}{\mathrm{N}} \times 100\% \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(1)$$

Content validity criteria can be seen as follows:

Table 2. Content validity criteria (Riduwan, 2015)

Number	(%)	Criteria
1	80.00-100	Valid
2	60.00-79.99	Quite Valid
3	50.00-59.99	Less Valid
4	0.00-49.99	Invalid

b. Construction Validity

Construction validity was calculated using the Pearson Product Moment correlation formula (Arikunto, 2010):

$$r_{XY} = \frac{N\sum XY - (\sum X)(\sum Y)}{\sqrt{\{N\sum X^2 - (\sum X)^2\}\{N\sum Y^2 - (\sum Y)^2\}}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(2)$$

Description:

r_{xy} = Correlation coefficient between correlated variables
X = Item score
Y = Total score
N = Number of Subjects

The value of r_{xy} is compared with the value of $r_{xytable}$ and is associated with the following criteria.

Table 3. Construction validity test criteria (Arikunto, 2016)

r_{xy}	Criteria
$r_{xycount} > r_{xytable}$	Valid
$r_{xycount} < r_{xytable}$	Invalid

2) Reliability

The reliability test was analysed using the halving technique and calculated using the Spearman-Brown correlation formula as follows (Sukardi, 2003).

$$r_{1\,1} = \frac{2.\,r_b}{1 + r_b} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(3)$$

Description:

r_{11} = Instrument reliability coefficient

r_b = Product moment correlation between hemispheres (odd-even) or (beginning-end)

The value of the reliable coefficient (r_{11}) can be compared with the r_{table} coefficient and adjusted according to the following criteria:

Table 4. Terms of reliability test (Riduwan, 2015)

r_{11}	Description
$r_{11} > r_{table}$	Reliable
$r_{11} < r_{table}$	Unreliable

Table 5. Reliability criteria (Ratnawulan and Rusdiana, 2017)

Reliability Index	Criteria
0.800-1.000	Very High
0.600-0.799	High
0.400-0.599	Enough
0.200-0.399	Low
0.000-0.199	Very Low

3) Distinguishing Power

Distinguishing power was analyzed using the following formula (Ratnawulan and Rusdiana, 2017):

$$\text{DP} = \frac{B_A - B_B}{0{,}5\ \text{N}} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(4)$$

Description:

DP = Dissimilarity of questions

B_A = Number of correct in the group with high scores

B_B = Number of correct in the group with low scores

N = Number of respondents

Distinguishing power is classified according to the following criteria:

Table 6. Criteria for distinguishing power (Arikunto, 2016)

Distinguishing Power Range	Criteria
0.00-0.20	Poor
0.21-0.40	Satisfactory
0.41-0.70	Good
0.71-1.00	Excellent

4) Difficulty Tier

The tier of difficulty per item can be calculated by the following equation (Bagiyono, 2017):

$$P = \frac{N_p}{N} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots..(5)$$

with:
P = Proportion
N_p = Number of subjects who answered the question correctly
N = Number of subjects

The criteria for the item difficulty index are determined as follows:

Table 7. Difficulty index criteria (Arikunto, 2016)

Difficulty Index Range	Criteria
0.00-0.30	Hard
0.31-0.70	Medium
0.71-1.00	Easy

3. RESULTS AND DISCUSSION

Designed Instruments were then tested for quality or suitability. The following is a description of the results and discussion of the instrument quality test:

1) Validity

a. Content Validity

The Four Tier Diagnostic Test Instrument (ITDET) prepared was then consulted with the supervisors. Then the product or ITDET that has received reviews and suggestions from the supervisor is continued with a content validity test by three validators (material experts). After the content validity was carried out by the three validators, several suggestions for improvement were obtained. Reviews and inputs from material expert validators were followed up by revising the items. Then the percentage of ITDET eligibility can be calculated based on the value given by the validator. The instrument tested is declared valid with the acquisition of the percentage value of each validator as follows:

Table 8. Percentage value of content validity of the 4-tier diagnostic test instrument

Validator	Total score	Percentage (%)	Criteria
V1	44	100	Valid
V2	43	97.73	Valid
V3	41	93.18	Valid
Average percentage		96.97	Valid

Based on Table 8, the value of the instrument content validity obtained was 96.97%. According to the criteria by Riduwan (2015), the percentage is in the range of 80.00%-100%, so the ITDET used has valid criteria.

b. Construction Validity

Based on the results of calculations using SPSS 25 analysis, the correlation value between 16 items and the total score is obtained in the following display:

Table 9. The results of the construct validity test

Number	r_{count}	r_{table}	category
r_1y	.594	0.444	Valid
r_2y	.556		
r_3y	.465		
r_4y	.553		
r_5y	.449		
r_6y	.469		
r_7y	.463		
r_8y	.512		
r_9y	.653		
$r_{10}y$	.462		
$r_{11}y$	.560		
$r_{12}y$	.517		
$r_{13}y$	.487		
$r_{14}y$	.632		
$r_{15}y$	.608		
$r_{16}y$	.627		

The results of the calculation of construct validity on 20 students using SPSS 25 analysis yielded 16 questions in the valid category with $r_{count} > r_{table}$ or $r_{count} > 0.444$, with a minimum r_{count} of 0.449 and a maximum r_{count} of 0.653. This refers to the opinion of Arikunto (2016), if $r_{count} > r_{table}$, the item can be said to be valid. The r_{table} value was obtained from the product-moment correlation table. The r_{table} value used was adjusted to the product-moment correlation table with a sample of 20 people and the significance tier used was 5%, to obtain the r_{table} of 0.444.

2) Reliability

The results of the reliability test calculation produced rcount or r_{11} of 0.830. This means that $r_{11} > r_{table}$ or $r_{11} > 0.468$ so that the instrument is reliable and can be used. This refers to the opinion of Riduwan (2015), if $r_{11} > r_{table}$, the instrument can be declared as reliable so that the instrument can be used. The r_{table} used had a significance of 5% with degrees of freedom (dk) = 18, so that the r_{table} = 0.468 was obtained. The r_{count} obtained is in the range of 0.800-1.000 which makes it has very high-reliability criteria as described in Table 5.

3) Distinguishing Power

The results of the analysis of the discriminatory power of 4-tier diagnostic test items can be seen in the following table:

Table 10. The results of the calculation of distinguishing power

Question number	**Value**	**Criteria**
1	0.6	Good
2	0.6	
3	0.5	
4	0.5	
5	0.4	Enough
6	0.4	
7	0.3	
8	0.5	Good
9	0.5	
10	0.6	
11	0.5	
12	0.4	Enough
13	0.5	Good
14	0.5	
15	0.5	
16	0.6	
Average	0.5	Enough

Based on the results of the calculation of the discriminatory power of items in Table 10, the mean discriminating power of the questions was 0.5. According to the discriminatory power criteria proposed by Arikunto (2016) in Table 6, the average value of discriminating power produced was in the range of 0.41-0.70. Thus, the average value of discriminating power is included in the good criteria.

4. Difficulty Tier

The results of the analysis of the difficulty tier of the 4-tier diagnostic test items are presented as follows:

Table 11. The results of the calculation of the difficulty tier

Question number	Value	Criteria
1	0.4	Medium
2	0.5	
3	0.45	
4	0.75	Easy
5	0.4	Medium
6	0.6	
7	0.55	
8	0.35	
9	0.55	
10	0.5	

11	0.45	
12	0.6	
13	0.55	
14	0.45	
15	0.45	
16	0.5	
Average	0.5	Medium

In general, the items had a moderate tier of difficulty with an average value obtained of 0.5 or ranging from 0.31 to 0.70. This is in accordance with the criteria proposed by Arikunto (2016) in Table 7.

4. CONCLUSION

It can be concluded that the 4-tier multiple choice diagnostic test was feasible to identifying students' misconceptions with an average content validity test percentage of 96.97% with valid criteria, the value of the construction validity test was $r_{count} > 0.444$ with the category of all valid items, the reliability test value was 0.83 (reliable) with very high criteria, the discriminatory power value of the questions was 0.5 with good criteria, and the difficulty tier of the questions was 0.5 with moderate criteria. This study only tested the feasibility of the four-tier multiple-choice diagnostic test instrument. Therefore, further research is needed to test the feasibility of other diagnostic tests, such as the five-tier diagnostic test.

ACKNOWLEDGEMENT

The researchers would like to thank all related parties who always provided constructive suggestions for the smoothness and benefits of this research.

REFERENCE

Arikunto, Suharsimi. (2010). Research Procedure: A Practical Approach. Jakarta: Rineka Cipta.

Arikunto, Suharsimi. (2016). Educational Evaluation Basics. Jakarta: Earth Literacy.

Bagiyono. (2017). Analysis of Tier of Difficulty and Distinguishing Power of Items for Tier 1 Radiography Training Exams. Widyanuclides, 16(1), 1-12.

Fariyani, Q., Rusilowati, A., & Sugianto. (2015). Development of Four Tier Diagnostic Tests to uncover the misconceptions of Physics for Class X students. Journal of Innovative Science education. 4(2), 41-49.

Fortuna, Dewi, Edy Chandra, and Ria Yulia Gloria. (2014). Development of a Diagnostic Test to Measure Students' Misconceptions on the Subject of the Human Regulatory System for Class XI Semester II High School Students. Journal of Scientiae Education. 2(2), 1-16.

Ismail, Ismiara Indah, A. Samsudin, E. Suhendi, and I. Kaniawati. (2015). Diagnostic Misconceptions Through Electric Dynamic Four Tier Test. Proceedings of the National Symposium on Science Innovation and Learning (SNIPS 2015) 2015.

Mubarak, Syarifatul, Endang Susilaningsih and Edy Cahyono. (2016). Development of Three Tier Multiple Choice Diagnostic Tests to Identify Misconceptions of Class XI Students. JISE : Journal of Innovative Science Education, 5(2), 101-110.

Ratnawulan, Elis and Rusdiana. (2017). Learning Evaluation. Bandung: Faithful Library.

Riduwan. (2015). Easy Learning Research for Teachers-Employees and Beginner Researchers. Bandung: Alphabeta.

Sagala, Syaiful. (2010). Concept and Meaning of Learning. Bandung: Alphabeta.

Sugiyono. (2017). Quantitative, Qualitative, and R&D Research Methods. Bandung: Alphabeta.

Sukardi. (2003). Competency Education Research Methodology and Practice. Jakarta: Earth Literacy.

Winaryati, Eny. (2014). Evaluation of Learning Supervision. Yogyakarta: Graha Ilmu.