JPPIPA 9(3) (2023)

Jurnal Penelitian Pendidikan IPA

Journal of Research in Science Education

http://jppipa.unram.ac.id/index.php/jppipa/index

Five-Tier Diagnostic Test Instrument Validation on Reaction Rate Materials: To Identify the Causes of Misconception and Student Representation

Deni Ainur Rokhim[1,2], Hayuni Retno Widarti[1*], Sutrisno[1]

[1]Chemistry Department, Faculty of Mathematics and Natural Science, Universitas Negeri Malang, Jalan Semarang No 5, Sumbersari, Lowokwaru, Kota Malang, 65145, Indonesia
[2]Chemistry, SMAN 3 Sidoarjo, Indonesia

Received: January 18, 2023
Revised: March 13, 2023
Accepted: March 25, 2023
Published: March 31, 2023

Corresponding Author:
Hayuni Retno Widarti
hayuni.retno.fmipa@um.ac.id

DOI: 10.29303/jppipa.v9i3.2952

Abstract: This study aims to determine the validity of the five-tier diagnostic instrument. Empirical validation tests were carried out to identify the items' difficulty level, the different power of the items, the percentage of distractor effectiveness, and test the validity and reliability of the items. The results show that the five-tier diagnostic test instrument is feasible and valid to use by looking at the test difficulty level, discriminating power, distractor effectiveness, item validity, empirical validity, and reliability test. The difficulty level of the questions is included in the average level of easy difficulty at Tier A and Tier R. In contrast, and the Multiple representatives are included in the medium level of difficulty. The average results at Tier A, Tier R, and Multiple Representative in the differential power test have sufficient differential power.

Keywords: Five-Tier Diagnostic; Instrument; Misconceptions; Validation.

Introduction

The reaction rate is part of an abstract chemical concept, so it often makes it difficult for students to understand this concept (Nazar et al., 2013). These difficulties can cause students to experience misconceptions about the reaction rate material. In Sinaga's research (2006), it was stated that almost half of the students experienced difficulty understanding the influence of catalysts and temperature on reaction rates (Sinaga, 2006).

A misconception is a condition where there is a misunderstanding of the correct concept in a scientific study. An understanding that is inconsistent with or erroneous with scientific concepts but is believed to be true by students indicates misconceptions among students (Habiddin & Page, 2019a; Jusniar et al., 2020; Widarti et al., 2017). This misconception often occurs in chemistry lessons because students consider chemistry difficult and abstract.

Misconceptions can also cause students to be left behind in learning in class. The lag is because chemicals are closely related to other chemicals. Misconceptions in chemistry lessons are fatal because concepts in chemistry are interrelated with one another. If there is a misconception or concept error at the beginning of learning, that will affect subsequent learning. In addition, misconceptions will result in low student abilities and failure to achieve mastery (Nazar et al., 2013).

Identifying misconceptions can be used in various ways. A diagnostic test can determine students' understanding and knowledge in understanding the reaction rate material (Anam et al., 2019). Diagnostic test instruments commonly used in science education include concept maps, open-ended questionnaires, pictures, word associations, interviews, multiple-choice

How to Cite:
Rokhim, D.A., Widarti, H.R., & Sutrisno, S. (2023). Five-Tier Diagnostic Test Instrument Validation on Reaction Rate Materials: To Identify the Causes of Misconception and Student Representation. *Jurnal Penelitian Pendidikan IPA*, *9*(3), 1380–1385. https://doi.org/10.29303/jppipa.v9i3.2952

tests, and two-tiered, three-tiered, and four-tiered multiple-choice tests. four-tier (Habiddin & Page, 2019; Hakimah et al., 2021; Qodriyah et al., 2020). These weaknesses will be given appropriate treatment and appropriate follow-up. In research from Qodriyah et al. (2020) regarding student misconceptions, a multilevel diagnostic test was carried out, and their answers were analyzed and categorized based on their level of understanding. The results show a misconception of 29.8%, or it can be said that misconceptions in this class fall into the low category (Qodriyah et al., 2020).

A diagnostic test instrument is a type of instrument that detect student errors to be used as material for improvement in learning on that material. Diagnostic instruments use tests and non-tests (Ardiansah et al., 2017). Weaknesses and strengths possessed by students can be known through the results of diagnostic tests that have been carried out. These weaknesses will be given appropriate treatment and appropriate follow-up.

This research develops a four-tier diagnostic test instrument on the material of reaction rate by adding one-level questions, which is then called a five-tier diagnostic test instrument. This test instrument can analyze the profiles of multiple representations of students and sources of information used to answer questions. The profile of the multiple representations of the students is obtained from the results of the student's answers in the form of an overview related to the multiple representations. Drawing instruments need to be added to diagnostic instruments because drawing is a powerful way to think and communicate in any field of science. Besides that, drawing is a processing ability that is a part of science to make hypotheses, design experiments, visualize and interpret data, and present results (Ainsworth et al., 2011; Quillin & Thomas, 2015). Moreover, drawing is a constructive and motivating activity because drawing is a combination of the use of hands and mind (Anam et al., 2019). This drawing instrument was modified with a four-tier diagnostic test instrument to produce a five-tier diagnostic test instrument.

Method

This study is a quantitative analysis that aims to determine the feasibility of the five-tier diagnostic test items. The population in this study were students of class XI and XII Science public high schools who had completed learning chemistry on the subject of reaction rates.

The instrument compiled and developed is a five-tier diagnostic test instrument. The number of questions compiled is 15 questions. The instrument was then validated before being used to collect research data. In addition to the five-tier diagnostic test, the instrument used interviews, which can be conducted to support research data.

Empirical validation tests were carried out to identify the items' difficulty level, the different power of the items, the percentage of distractor effectiveness, and test the validity and reliability of the items. An empirical validation test was carried out using SPSS 16.0 software. The results of the empirical validation test become the basis for revising the questions in the five-tier diagnostic instrument prototype. The revisions will produce the final product of a five-tier diagnostic instrument on reaction rate.

Result and Discussion

Quantitative analysis was carried out using the SPSS 16.0 program. The program automatically analyzes the difficulty level, discriminating power, distractor effectiveness, item reliability, and other statistical data. The results of the analysis of the test instrument at the difficulty level of the items can be seen in Table 1.

Based on the table 1, a total of 15 questions were analyzed to determine the difficulty level of the items. These questions are included in the average easy difficulty level at Tier A and Tier R, while the Multiple representations are included in the medium difficulty level. Magdalena et al. (2021) said that the difficulty level of the questions needs to be seen from the student's ability to answer the questions given, not from the perspective of the teacher who created the questions (Magdalena et al., 2021). The average value of the difficulty level can show that, overall, the questions on the instrument have fulfilled the difficulty requirements, and students can answer them well. According to Arikunto (2009), a good question is relatively easy (Arikunto, 2009).

Table 1. Item Difficulty Level

Question Number	Tier A	Tier R	Representative
1	0.94	0.72	0.98
2	0.94	0.94	0.92
3	0.92	0.94	0.82
4	0.90	0.80	0.34
5	0.56	0.66	0.48
6	0.90	0.86	0.20
7	0.92	0.96	0.76
8	0.90	0.88	0.34
9	0.92	0.72	0.04
10	0.54	0.96	0.05
11	0.90	0.94	0.56
12	0.94	0.88	0.38
13	0.66	0.76	0.24
14	0.54	0.80	0.88
15	0.90	0.90	0.32
Average	0.83	0.85	0.49

Based on the table 1, it can be seen that multiple representations have lower scores compared to Tier A and Tier R. This is because students know more about the answers to each question and can give a reason. However, they cannot explain it in multiple representations. Habidin & Page (2019) research said that students answering Tier A could only use their content knowledge, while to answer Tier R, they need good conceptual knowledge (Habiddin & Page, 2019b). line with Atikah's research (2020), those with good content and conceptual understanding can answer the questions correctly (Atikah, 2020).

Discriminating power analysis is used to examine test questions from the aspect of the ability of the test to distinguish students who fall into the low and high categories (Magdalena et al., 2021). The higher the value of the discriminating power of the items, the better it will be in distinguishing the abilities or achievements of students. A high discriminating power value indicates better item quality in identifying students between high achievers and achievers (Jusniar et al., 2020). The average Tier A, R, and Multiple Representative results have sufficient discriminating power. It shows that half of the items on the five-tier diagnostic instrument are good enough to differentiate students' abilities or achievements. There are items with very different power values compared to the others, namely at number 15. Tuckman & Harper (2012) suggested that a value of DI above 0.20 is useful (Tuckman & Harper, 2012).

The effectiveness of the distractor is a parameter to determine whether the distractor or the wrong answer is functioning effectively or not. The distractor must be chosen by at least one student so that the distractor used in the question item can be seen properly. A good distractor is a distractor that at least 5% of students choose; otherwise, it is considered a bad distractor (DiBattista & Kurzawa, 2011). This parameter also determines students' low conceptual understanding due to choosing the wrong answer or reason (Habiddin & Page, 2019b).

Table 2. Index Driscrimination

Question Number	Tier A	Tier R	Representative
1	0.10	0.50	0.05
2	0.10	0.10	0.18
3	0.20	0.10	0.18
4	0.20	0.40	0.23
5	0.50	0.30	0.41
6	0.20	0.10	0.27
7	0.20	0.00	0.27
8	0.20	0.30	0.14
9	0.20	0.50	0.09
10	0.60	0.00	0.41
11	0.20	0.10	0.32
12	0.10	0.30	0.55
13	0.20	0.20	0.09
14	0.60	0.40	0.18
15	0.20	0.00	0.18
Average	0.24	0.22	0.24

Based on the table above, the effectiveness of distractors is known at Tier A; on average, they only choose two answer options for each number. The distractor or the wrong answer cannot function properly in the two options not selected because less than 5% or none of the students chose that answer. However, in numbers 13 and 14, only one distractor needs to be fixed. Whereas in number 11, number 12, and number 15, the wrong answer option can function properly because the value of the effectiveness of the distractor is more than 5%, meaning that it qualifies as a distractor. The distractors not selected by the Testee will be reconsidered. When deciding whether a question should be revised or replaced, the values of all parameters must be considered. In some circumstances, even questions with bad distractors can be defended because the main purpose of this instrument is to identify students' understanding, not discriminate between low-achieving students (Suruchi, S., and Rana, 2014).

Table 3. Distractor Effectiveness

Tier	Options	1	2	3	4	5
A	A	-	6.00%	-	-	-
	B	6.00%	-	8.00%	10.00%	56.00%
	C	94.00%	-	92.00%	90.00%	44.00%
	D	-	94.00%	-	-	-
Tier	Options	6	7	8	9	10
A	A	90.00%	-	10.00%	-	-
	B	10.00%	92.00%	90.00%	8.00%	54.00%
	C	-	-	-	92.00%	46.00%
	D	-	8.00%	-	-	-
Tier	Options	11	12	13	14	15
A	A	90.00%	94.00%	12.00%	18.00%	90.00%
	B	10.00%	6.00%	22.00%	-	10.00%
	C			66.00%	28.00%	
	D			-	54.00%	

After validating the contents of the items obtained from the validator's assessment, 3 experts/validators consisted of 2 lecturers majoring in chemistry at Malang State University and 1 teacher from SMA Negeri 3 Sidoarjo, as can be seen in the following Figure 1.

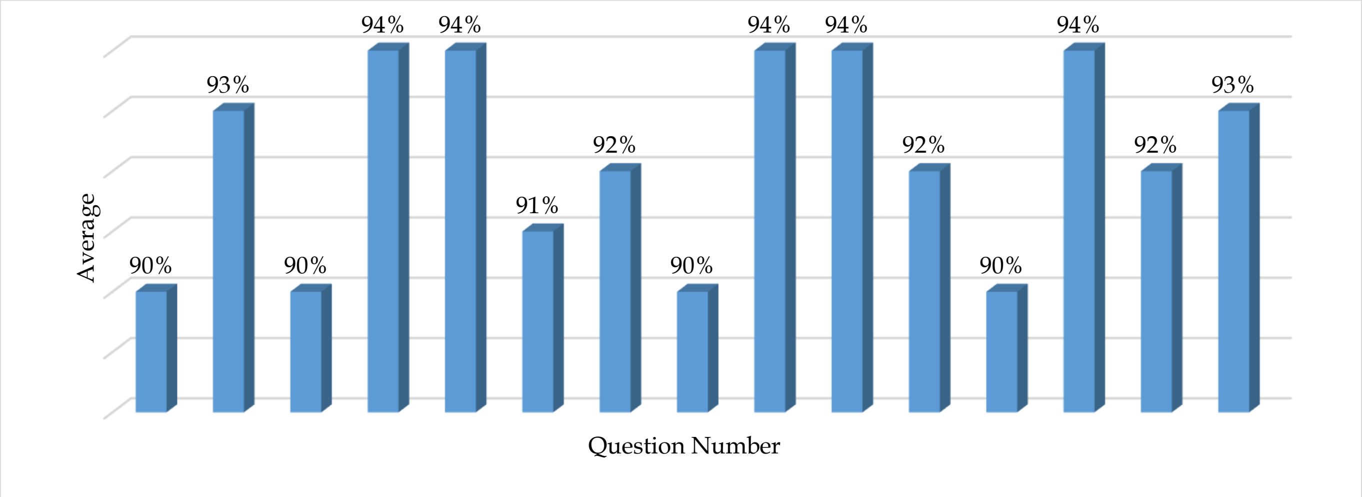

Figure 1. The Average Result of the Validation Percentage of the Item Content

Based on the figure 1 shows that the results of validating the contents of the item items carried out by the validator, as many as 15 questions can be said to be valid. Borich (1994) argues that R ≥ 75% can be classified as a good percentage of agreement by the validator (Borich, 1994). This event occurs because the lowest average percentage is 89% and can still be said to be valid.

The validity test was performed using Product Moment Pearson Correlations (Bivariate Pearson) analysis. According to Kimberlin & Winterstein (2008), validity refers to whether the information obtained from a test represents the true understanding of the examinee. The item's validity is indicated by the value of the Pearson correlation index (r count) (Kimberlin & Winterstein, 2008). Based on the table above, most items can be said to be valid. The item's validity is closely related to the index of discrimination or discriminating power because if the item can distinguish high and low-achieving students, it means that the item has been trusted to measure conceptions (Jusniar et al., 2020) .

Another possible reason for the invalidity of question number 15 is that even though it is invalid, the item remains positive, so it can still be used by revising the language of question number 15 because it is likely caused by the language of the question, which is difficult for students to understand.

Relevant research on empirical validation tests written by Putri & Ernawati (2021) shows valid results because Rxy, compared to Rtable, produces Rxy > Rtable at a significance of 5%. By comparing the rxy and rtable values, it is found that the sixteen items in the Final Draft are valid, considering that the rxy > rtable values (Putri & Ermawati, 2021).

Table 4. Empirical Validation

Question Number	Tier A	Tier R	Representative
1	0.000	0.000	0.769
2	0.003	0.011	0.234
3	0.001	0.002	0.000
4	0.000	0.000	0.000
5	0.007	0.000	0.000
6	0.000	0.033	0.000
7	0.008	0.014	0.000
8	0.013	0.641	0.000
9	0.001	0.000	0.000
10	0.001	0.000	0.000
11	0.000	0.641	0.000
12	0.000	0.002	0.000
13	0.044	0.000	0.000
14	0.001	0.079	0.000
15	0.024	0.899	0.000

Table 5. Reliability Test (Cronbach Alpha)

N	Tier A	Tier R	Representatives
15	0.707	0.664	0.957

The reliability test refers to the scoring of the items. The calculation to determine the instrument's reliability uses Cronbach's Alpha with a significance value of 5%. To get the correct data with conclusions that follow the actual situation, we need an instrument that is valid and consistent, and precise in providing (reliable) research data (Yusup, 2018). At Tier A, Cronbach's alpha coefficient is 0.707; at Tier R is 0.664; at Multiple representations is 0.957. The reliability results show that the five-tier diagnostic instrument on multiple representatives has very high criteria or is very reliable. Meanwhile, Tier A and Tier R have sufficient criteria. The item items have been tested for reliability based on the data above. According to Creswell (2012), the test instrument results have internal consistency or high regularity (Creswell, 2012). This is supported by research from Utari & Ermawati 2018 regarding the development of a four-tier misconception diagnostic test instrument with a reliability test result of 1.067, which is in the very reliable category (Utari & Ermawati, 2018).

Conclusion

Based on the research results above, the five-tier diagnostic test instrument is feasible and valid to use by looking at the test results of difficulty level, discriminating power, distractor effectiveness, item validity, empirical validity, and test reliability. The difficulty level of the questions is included in the average level of easy difficulty at Tier A and Tier R. In contrast; the Multiple representatives are included in the medium level of difficulty. The average results at Tier A, Tier R, and Multiple Representative in the differential power test have sufficient differential power. The effectiveness of the distractor at Tier A, on average, only chooses two answer options for each number because less than 5% of students choose those answer options, and no one even chooses them. In the Tier A reliability test, Cronbach's alpha coefficient is 0.707; at Tier R is 0.664 and at Multiple representatives is 0.957, so it can be said that the five-tier diagnostic test has been tested for its reliability.

References

Ainsworth, S., Prain, V., & Tytler, R. (2011). Drawing to Learn in Science. *Science*, *333*(6046), 1096–1097. https://doi.org/10.1126/science.1204153

Anam, R. S., Widodo, A., Sopandi, W., & Wu, H.-K. (2019). Developing a Five-Tier Diagnostic Test to Identify Students' Misconceptions in Science: An Example of the Heat Transfer Concepts. *İlköğretim Online,* 1014-1029. https://doi.org/10.17051/ilkonline.2019.609690

Ardiansah, Masykuri, M., & Rahardjo, S. B. (2017). Kelayakan Instrumen Diagnostik Pada Materi Asam-. *Seminar Nasional Pendidikan Sains*, *21*, 104–111. Retrieved from https://jurnal.fkip.uns.ac.id/index.php/snps/article/view/11399

Arikunto, S. (2009). *Dasar-Dasar Evaluasi Pendidikan*. Bumi Aksara.

Atikah. (2020). *Identifikasi Pemahaman Konsep Dan Miskonsepsi Mahasiswa Jurusan Kimia Pada Materi Sifat Koligatif Larutan Dengan Menggunakan Instrumen Diagnostik Four-Tier*. Universitas Negeri Malang.

Borich, G. D. (1994). *Observation Skills for Effective Teaching* (2th Editio). Macmillan Publishing Company.

Creswell, J. W. (2012). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research.* Pearson.

DiBattista, D., & Kurzawa, L. (2011). Examination of the quality of multiple-choice items on classroom tests. *Canadian Journal for the Scholarship of Teaching and Learning, 2*(2), 4. https://doi.org/10.5206/cjsotl-rcacea.2011.2.4

Habiddin, H., & Page, E. M. (2019a). Development and Validation of a Four-Tier Diagnostic Instrument for Chemical Kinetics (FTDICK). *Indonesian Journal of Chemistry*, *19*(3), 720–736. https://doi.org/10.22146/ijc.39218

Habiddin, & Page, E. M. (2019b). Development and Validation of a Four-Tier Diagnostic Instrument for Chemical Kinetics (FTDICK). *Indonesian Journal of Chemistry*, *19*, 720–736. https://doi.org/10.22146/ijc.39218

Hakimah, N., Muchson, M., Herunata, H., Permatasari, M. B., & Santoso, A. (2021). *Identification student misconceptions on reaction rate using a Google forms three-tier tests.* https://doi.org/10.1063/5.0043114

Jusniar, J., Effendy, E., Budiasih, E., & Sutrisno, S. (2020). Developing a three-tier diagnostic instrument on chemical equilibrium (TT-DICE). *Educacion Quimica*, *31*(3), 84–102. https://doi.org/10.22201/fq.18708404e.2020.3.72133

Kimberlin, C. L., & Winterstein, A. G. (2008). Validity and reliability of measurement instruments used in research. *American journal of health-system pharmacy, 65*(23), 2276–2284. https://doi.org/10.2146/ajhp070364

Magdalena, I., Fauziah, S. N., Faziah, S. N., & Nupus, F. S. (2021). Analisis Validitas, Reliabilitas, Tingkat Kesulitan dan Daya Beda Butir Soal Ujian Akhir Semester Tema 7 Kelas III SDN Karet 1 Sepatan. *BINTANG: Jurnal Pendidikan Dan Sains*, *3*(2), 198–214. Retrieved from https://www.ejournal.stitpn.ac.id/index.php/bintang/article/view/1291

Nazar, M., Winarni, S., Fitriana, R., Prodi Pendidikan Kimia Unsyiah Banda Aceh, D., & Prodi Pendidikan Kimia Unsyiah Banda Aceh, M. (2013). Identifikasi Miskonsepsi Siswa SMA Pada Konsep Faktor-faktor yang Mempengaruhi Laju Reaksi. *Biologi Edukasi: Jurnal Ilmiah Pendidikan Biologi*, *2*(3), 49–53. Retrieved from https://jurnal.unsyiah.ac.id/JBE/article/view/448

Putri, W. K., & Ermawati, F. U. (2021). Pengembangan, Uji Validitas dan Reliabilitas Tes Diagnostik Five-Tier untuk Materi Getaran Harmonis Sederhana beserta Hasil Uji Coba. *PENDIPA Journal of Science Education*, *5*(1), 92–101. https://doi.org/10.33369/pendipa.5.1.92-101

Qodriyah, N. R. L., Rokhim, D. A., Widarti, H. R., & Habiddin. (2020). Identifikasi Miskonsepsi Siswa Kelas XI SMA Negeri 4 Malang pada Materi Hidrokarbon Menggunakan Instrumen Diagnostik Three Tier. *Jurnal Inovasi Pendidikan Kimia*, *14*(2), 2642–2651.

Quillin, K., & Thomas, S. (2015). Drawing-to-Learn: A Framework for Using Drawings to Promote Model-Based Reasoning in Biology. *CBE—Life Sciences Education*, *14*. https://doi.org/10.1187/cbe.14-08-0128

Sinaga, M. S. (2006). *Analisis Kesulitan Siswa Dalam Memahami Materi Sub Pokok Bahasan Faktor-Faktor Yang Mempengaruhi Laju Reaksi Yang Diolah Dengan Reduksi Didaktik*. UPI Bandung.

Suruchi, S., and Rana, S. S. (2014). Test item analysis and the relationship between difficulty level and discrimination index of test items in an achievement test in biology. *Paripex India Journal of Research*, *3*(6), 56–68. Retrieved from https://www.worldwidejournals.com/paripex/recent_issues_pdf/2014/June/June_2014_1403953039__18.pdf

Tuckman, B. W., & Harper, B. E. (2012). *Conducting educational research* (6th ed). Rowman & Littlefield Publishers, INC.

Utari, J. I., & Ermawati, F. U. (2018). Pengembangan Instrumen Tes Diagnostik Miskonsepsi Berformat Four-Tier untuk Materi Suhu, Kalor, dan Perpindahannya. *Inovasi Pendidikan Fisika*, *7*(3), 434–439. Retrieved from https://ejournal.unesa.ac.id/index.php/inovasi-pendidikan-fisika/article/view/25537

Widarti, H. R., Permanasari, A., & Mulyani, S. (2017). Undergraduate Students' Misconception On Acid-Base And Argentometric Titrations: A Challenge To Implement Multiple Representation Learning Model With Cognitive Dissonance Strategy. *International Journal of Education*, *9*(2), 105–112. Retrieved from https://www.learntechlib.org/p/208918/

Yusup, F. (2018). Uji Validitas Dan Reliabilitas Instrumen Penelitian Kuantitatif. *Jurnal Tarbiyah: Jurnal Ilmiah Kependidikan*, *7*(1), 17–23. https://dx.doi.org/10.18592/tarbiyah.v7i1.2100