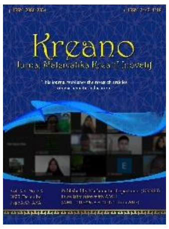
UNNES JOURNALS

Kreano, Jurnal Matematika Kreatif-Inovatif
Volume 14, number 2, 2023, pp. 373-385



Development of Four-tier Multiple-choices Test to Diagnose Student's Misconception

Anggit Prabowo[1,2], Tatang Herman[1], Siti Fatimah[1], Arum Dwi Ahniah[2]

[1]Universitas Pendidikan Indonesia, Indonesia
[2]Universitas Ahmad Dahlan, Indonesia

Corresponding Author: anggit.prabowo@pmat.uad.ac.id*

Abstract

Studies in Indonesia delineate that junior high school students in Indonesia have low mastery of circle material. Students with low mastery of the material tend to have misconceptions. The goal of this study is to develop a four-tier diagnostic test for circle material to identify the level of their understanding, including misconception in circle material. This development used the ADDIE model. It comprises four phases: Analysis, Design, Development, Implementation, and Evaluation. The test developed consists of 20 items which four experts in learning mathematics have validated. They are two lecturers and two teachers of mathematics who have teaching experience of more than ten years. The validated test was implemented on 34 grade 8 students in Yogyakarta. From the test results obtained information, of the 20 test items, on average, have an ideal difficulty level. All items have a good discriminant index. Test reliability was estimated using the Cronbach Alpha formula. The estimation results show a test reliability coefficient of 0.72. Students as test respondents stated that the tests developed contained easy-to-understand instructions, easy-to-understand language, sufficient test time, clear pictures, and enough items. This test can help junior high school mathematics teachers in Indonesia to identify their students' misconceptions and level of understanding, especially regarding circle material.

Keywords: *circle*, diagnosis, *four-tier*, *test*

Information of Article

Subject classification	*97D60 Student assessment, achievement control, and rating*
Submitted	*22 June 2023*
Review Start	*25 June 2023*
Initial Review Finish	*21 August 2023*
Round 1 Finish	*21 October 2023*
Accepted	*23 October 2023*
Published	*1 Desember 2023*
Similarity Check	*10%*

Abstrak

Hasil-hasil studi *menunjukkan siswa sekolah menengah pertama di Indonesia memiliki penguasaan yang rendah pada materi lingkaran. Siswa dengan penguasaan materi yang rendah terindikasi mengalami miskonsepsi. Penelitian ini bertujuan untuk mengembangkan tes diagnostik tipe four-tier pada materi lingkaran untuk mendiagnosis level pemahaman siswa pada materi lingkaran, termasuk miskonsepsi yang dialami siswa. Pengembangan ini menggunakan model ADDIE (Analysis, Design, Devalopment, Implementation, Evaluation). Tes yang dikembangkan terdiri atas 20 butir yang telah divalidasi oleh 4 orang ahli dalam pembelajaran matematika yang terdiri atas 2 orang dosen pembelajaran matematika dan 2 orang guru matematika yang memiliki pengalaman mengajar lebih dari 10 tahun. Tes yang sudah divalidasi diujicobakan kepada 34 siswa kelas 8 di Yogyakarta. Dari hasil uji coba didapat informasi bahwa dari 20 butir soal tes rata-rata memiliki tingkat kesukaran yang ideal. Keseluruhan butir memiliki daya pembeda yang baik. Keandalan tes diestimasi dengan menggunakan formula Cronbach Alpha. Hasil estimasi menunjukkan koefisien reliabilitas tes sebesar 0,72. Siswa sebagai responden uji coba menyatakan bahwa tes yang dikembangkan memuat petunjuk yang mudah dimengerti, bahasa yang mudah dipahami, waktu tes yang cukup, gambar yang jelas, dan jumlah butir soal yang cukup. Tes yang dikembangkan ini dapat membantu guru matematika sekolah menengah pertama di Indonesia untuk mengidentifikasi miskonsepsi dan level pemahaman siswanya, khususnya pada materi lingkaran.*

INTRODUCTION

Recently, STEM was being integrated into teaching. "STEM" refers to a multidisciplinary educational perspective combining mathematics with science, technology, and engineering (Chesky & Wolfmeyer, 2015). Mathematics has a more vital role in those fields because it supports students in mastering other fields (Shim, Shakawi, & Azizan, 2017). There are a lot of occupations, notably in science, technology, and engineering, that depends on mathematics (Li & Schoenfeld, 2019), even daily activities. In science, for example, mathematics supports the development of formulas to find unknown geometric and parameter values related to inheritance, measurements, and relationships of points, numbers, angles, and lines in space (Swaranjit, 2015).

The important role of mathematics makes it necessary for Indonesian students to learn mathematics from elementary education to higher education. To evaluate students' performance in mathematics, several types of assessments are conducted in Indonesia education system, such as assessments by teachers, schools, and the government. The type of assessment used to evaluate is formative and summative. Summative evaluations are used to assess students' learning, skill development, and academic achievement, whereas formative assessments are utilized by teachers to modify their teaching and learning practices and increase student progress (Bhat, 2019).

The national examination was the formative assessment conducted by Indonesia government. The subjects tested, especially at the junior high school (JHS) level, consist of mathematics, Indonesian, English, and sciences. The results of the national exam on several occasions showed that the average mathematics score of students taking national exams in Indonesia was the lowest among other subjects tested in the national examination (Prabowo, Rahmawati, & Anggoro, 2019). Only 46.19 out of a possible 100 points were earned on average by Indonesian students in mathematics in the last period. Mastery of Indonesian, English, and Sciences material was 66.12, 50.96, and 49.43, respectively. It indicates that mathematics is the most difficult for JHS students in Indonesia. Studies also reveal that mathematics is complicated (Setiana, Ili, Rumasoreng, & Prabowo, 2020) and commonly perceived to be difficult (Fritz, Haase, & Rasanen, 2019).

The content of mathematics in Indonesia JHS students contains numbers, algebra, geometry and measurement, statistics, and probability. In more detail, of the various mathematics materials tested, geometry and measurement materials are classified as materials with a small average percentage of being answered correctly by students taking the national examination (Prabowo, Anggoro, Adiyanto, & Rahmawati, 2018; Retnawati, Arlinwibowo, & Sulistyaningsih, 2017).

One of the subjects tested on geometry and measurement is circle. At the last national exam, the average score of students' mastery of circle material was only 35.77. The low mastery of students in Indonesia on circle material also occurred in previous years. Indonesian students' mastery of circle material ranged from 40 to 50. The low student mastery of circle material is also described in various study results (Rejeki & Putri, 2018). The low mastery of the circle concept is related to determining the circle elements such as the center point and radius (Sudihartinih & Purniati, 2019), determining the length of the circular (Lestari, Mardiyana, & Slamet, 2020).

Low mastery of the material is closely related to misconceptions about students’ comprehension of the material related (Kusmaryono, Basir, & Saputro, 2020). Misconceptions are misunderstandings and misinterpretations based on wrong meanings (Ojose, 2015). It harms students' cognitive development because they build their own concepts (Aydin, Keles, & Hasiloglu, 2012), which are far from the correct concept.

Misconceptions detected early will be easy to be corrected. To identify the students' misconceptions, the following tools might be used: interviews, open-ended questions, and multiple-choice with two or three tiers (Ojose, 2015). Using interview, students can set their own schedules and can learn more in-depth information. However, this technique needed a lot of time, and there were few respondents. For open-ended tests, which provide respondents the chance to create answers based on their own word choices, it is possible to test with a greater number of participants than in interviews. But the open-ended test has a drawback in that it takes a long time to evaluate the data and concluded. It enables ease of administration and analysis for multiple-choice examinations and can be given to many respondents. Multiple-choice exams, however, cannot distinguish between students' accurate and incorrect responses, making it impossible to undertake in-depth study on them (Kaltakci-Gurel, D., Eryilmaz & McDermott, 2017).

Development of a diagnostic test that not only has a simple administration method but can also be used with a greater number of test takers and provides detailed information on students' level of knowledge is required to address the shortcomings of previous approaches. Two-tier and three-tier multiple choice are two of the test formats that are feasible (Ojose, 2015). However, it still has limitations, one of which is that it cannot identify the reason why students actually encounter misunderstandings (Gurel, Eryilmaz, & McDermott, 2015).

The four-tier model is the diagnostic test model that can provide the most complete information in making a diagnosis. In this model, an item is equipped with answer choices, reasons for choosing answers, and the degree of confidence for each response and the reason (Caleon & Subramaniam, 2010a). This test model was originally developed on physics materials, such as optics (Caleon & Subramaniam, 2010b), optical tools, and waves (Zaleha, Samsudin, & Nugraha, 2017). This model has not been developed in the field of mathematics yet. The models developed in the field of mathematics are only two-tier (Lin, Yang, & Li, 2016). This model has a lack because it only provides alternative answers and underlying reasons for answering questions, regardless of the level of confidence students choose answers (Kutluay, 2005). Hence, it is essential to develop a four-tier diagnostic test to identify students' misconceptions about circles material.

METHOD

The diagnostic test is a kind of assessment that is part of instruction. Therefore, the ADDIE model was used in developing the diagnostic test in this study. It is one of the most widely used models for instructional design, which serves as a manual for creating successful designs, systematic, and easy to apply so that the resulting product is well-tested (Aldoobie, 2015). The steps of the ADDIE model are Analysis, Design, Development, Implementation, and Evaluation (Lu & Sides, 2022). The tasks and output of each step are presented in Table 1.

Table 1. Student Response Criteria

Steps	Tasks	Output
Analysis	Needs assessment	Problem Statement
Design	Write objectives	Measurable objective
	Create test blueprint	Test blueprint
Development	Develop diagnostic items test	Four-tier diagnostic item test
Implementation	Try out	Items characteristic
Evaluation	Revise the product	Revised the product

In the analysis phase, a need assessment was conducted to identify the problem that occurred. The problem is needing a four-tier diagnostic test to identify students' misconceptions in circles. In the design phase, objectives were determined. In this phase, the blueprint for the test was set (*see Table 2*).

Table 2. Test Blueprint

Competency	Indicator	Number of item
Explain and solve problems related to central angles, inscribed angles, arc lengths, and the sector of a circle, and their relationships	Identify the elements of a circle	1, 2, 3, 4
	Determine the circumference and area of a circle	7, 12
	Determine the relationship between the central angle and the inscribed angle	5, 6, 13
	Determine the relationship between arc length and the area of sector	8, 9, 10, 11, 14, 15

In the development phase, diagnostic items were developed based on the blueprint. At this stage, the test was validated by four experts. They are two mathematics lecturers and two mathematics teachers with more than ten years of teaching experience. In the implementation phase, the developed diagnostic item test was implemented with 34 junior high school students in Yogyakarta, Indonesia. In the evaluation phase, evaluation was carried out in all the stages (analysis, design, development, and implementation). In this phase, the test was revised based on implementation, including the characteristics of the difficulty index (Dif-I), discriminating index (Dis-I), and reliability.

Instruments and Data Collection

The instruments used in this study were the validation sheet and student response questionnaire. The type of validity analysis for the test was content validity. It represents the evidence of the degree to which the assessment components of the instrument are pertinent and represent a specific construct for a given assessment purpose, known as content validity. In contrast to other types of validity, this validity relates to test-based validity rather than score-based validity (Almanasreh, Moles, & Chen, 2019). It is determined by using expert agreement (Retnawati, 2016). The validated aspects include material, construction, language, and appearance. Aspects of student responses that were asked in the questionnaire included: instructions for clarity, material studied and tested, ease of language selection, time effectiveness, picture clarity, and the adequacy of the number of items.

Analysing of Data

To establish the test's validity, content validity analysis of expert judgment data was performed. The test is declared valid if it meets the criteria: in accordance with the indicators to be measured, the items are formulated clearly, use clear and informative language, and contain clear instructions. The validity was analyzed by using Aiken's V index formula. The following is the formulation of Aiken's item validity index.

$$V = \frac{\sum s}{n(c-1)}$$

V represents the validity index of the item. s means the scores the rater gave minus the lowest score in the category used.

$$s = r - l_o$$

r is the score given by the rater and l_o is the lowest score in the category used.

Item Difficulty Index (Dif-I) is the proportion of test takers who answered an item correctly (Sayyah et al., 2012). The following is the formulation to identify Dif-I.

$$P_i = \frac{n_i}{N}$$

P_i is the difficulty Index of item-i. n_i is the number of students who answer the item-i correctly. N is the total of students who answer the item-i. Items with Dif-I <0.3 are classified as too difficult, Dif-I > 0.7 are classified as too easy, and if Dif-I ranges from 0.3 to 0.7, it is recommended (Musa, Shaheen, & Elmardi, 2018).

Item discriminating index (Dis-I) is the difference between the percentage of examinees with high ability and those with low ability who get the items correctly. Dis-I describes how well the items differentiate student abilities (Dhakne-Palwe, Gujarathi, & Almale, 2015). Before identifying the Dis-I, students were divided to two groups (high and low group) based on their total score. The following is the formulation to identify Dis-I.

$$D_i = \frac{nU_i}{NU_i} - \frac{nL_i}{NL_i}$$

D_i is the discriminating index of item-i. nU_i is the number of students in the high group who answer the item-i correctly. NU_i is the total of students in the high group. nL_i is the number of students in the low group who answer the item-i correctly. NL_i is the total of students in the low group. The acceptable Dis-I is between 0.2 to 0.29, while more than 0.29 is good and excellent (Shete, Kausar, Lakhkar, & Khan, 2015).

The reliability of the diagnostic test was estimated using Cronbach's Alpha Formula.

$$\alpha = \frac{n}{n-1}\left(1 - \frac{\sum s^2(X_i)}{s^2(Y)}\right)$$

α is the coefficient of reliability. n refers to the number of items. $s^2(X_i)$ is the variance of item- i and $s^2(Y)$ is the variance of total scores. The test is said to be reliable if it has a reliability coefficient of more than 0.5 (Gugiu & Gugiu, 2018).

This study also identifies the student’s responses to the test. A questionnaire with 8 items/statements was used to explore the quality of instructions provided, language used, time allotted, pictures presented, and number of items provided. Each item contains four alternative responses: very good (4), good (3), poor (2), and very poor (1) were used. The average of students’ responses ($\bar{x}$) is categorized based on the following criteria (Table 3).

Table 3. Student Response Criteria

Interval	**Criteria**
$M_i + 1.8\ SB_i < \bar{x}$	Very good
$M_i + 1.8\ SB_i \geq \bar{x} > M_i + 0.6SB_i$	Good
$M_i + 0.6\ SB_i \geq \bar{x} > M_i - 0.6\ SB_i$	Moderate
$M_i - 0.6\ SB_i \geq \bar{x} > M_i - 1,.\ SB_i$	Poor
$M_i - 1.8\ SB_i > \bar{x}$	Very poor

$\bar{x}$= average score, $M_i = \frac{1}{2}$(Ideal maximum score + ideal minimum score), and $SB_i = \frac{1}{6}$(Ideal maximum score - ideal minimum score).

To identify the level of students’ understanding, Table 4 presents the criteria.

Table 4. Analysis of Four-tier Diagnostic Test Items

Tier 1	Tier 2	Tier 3	Tier 4	Conclusion
F	S	F	S	Misconceptions (M)
F	S	F	NS	Does not understand the concept (NU)
F	NS	F	S	
F	NS	F	NS	
R	S	R	S	Understands the concept (UC)
R	S	R	NS	
R	NS	R	S	
R	NS	R	NS	
R	S	F	S	Partial understanding (PU)
R	S	F	NS	
R	NS	F	S	
R	NS	F	NS	
F	S	R	S	
F	S	R	NS	
F	NS	R	S	
F	NS	R	NS	

(Rawh, Samsudin, & Nugraha, 2020)
F = False; R = Right; S = Sure; NS = Not Sure

Students are categorized as having misconceptions if they are wrong and sure in answering questions and giving reasons. Students are categorized as not understanding the concept if they are wrong and sure about answering questions and wrong and not sure in giving reasons, wrong and not sure about answering questions and wrong and sure in giving reasons, and wrong and not sure about answering questions and giving reasons.

Students are categorized as understanding the concept if they are correct and confident in answering questions and giving reasons, correct and confident in answering questions and correct but not confident in giving reasons, correct but unsure in answering the question and correct and confident in giving reasons, correct but not sure in answering the question and giving reasons, correct and confident in answering questions but incorrect and confident in giving reasons, and correct and confident in answering questions but incorrect and unsure in giving reasons.

Students are categorized as having partial understanding of the concept if they are correct and confident in answering questions and giving reasons, correct and confident in answering questions and correct but not confident in giving reasons, correct but unsure in answering the question and correct and confident in giving reasons, correct but not sure in answering the question and giving reasons, correct and confident in answering questions but incorrect and confident in giving reasons, and correct and confident in answering questions but incorrect and unsure in giving reasons.

RESULTS AND DISCUSSION

Results

The Stages of Four-Tier Diagnostic Test Development

The development of a four-tier diagnostic test for 8th students in Indonesia at circle material was started with the analysis phase. In this phase, problems in instruction were identified. First, junior high school students in Indonesia had difficulties in mastering circle material. Second, this difficulty needs to be diagnosed to identify students' strengths and weaknesses in circle material. The diagnosis can also give a preview of the category of students' misconceptions in circle.

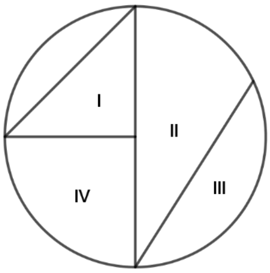
Based on the analysis, the objective of the design phase of the study was to develop a four-tier type diagnostic test instrument for circle material. The blueprint of the test was created. The content of the material referred to the curriculum in Indonesia regarding the basic competencies that students in Indonesia want to achieve, especially in the basic competency points for mathematics of 8th grade (competency numbers 3.7 and 4.7). The blueprint consists of the measured basic competency, indicators, type, and number of items.

The test was developed based on the blueprint. Four-tier items with four options are the type of test. The level of confidence is "sure" and "not sure". The test consists of 20 items. It had been validated by two experienced mathematics lecturers and two experienced mathematics teachers. Their judgment was analyzed by using Aiken's V formula. Its Aiken's V index was 0.95. The minimum index to state the valid instrument (4 raters with 4 options) is a minimum of 0.92 (Aiken, 1985). It means the test is valid.

Table 5 presents one of the sample questions. This item measures the student's ability to show the elements of a circle, in this case, the sector of the circle.

The valid test was implemented on 34 junior high school students in Yogyakarta. They did 20 items test and gave the response of the implementation of the test. It was conducted by using a paper-based test. The responses were analyzed to identify the Dif-I, Dis-I, coefficient of reliability, students' responses, and their level of concept understanding in circle.

Table 5. Four-Tier Diagnostic Test Item

Tier		Question
Tier 1	:	1. Look at the following picture!  The sector of a circle is indicated by the number.... A. I B. II C. III D. IV
Tier 2	:	My level of confidence in choosing the answer: ○ Sure ○ Not sure
Tier 3	:	The reason I chose the answer: A. because the sector is the area bounded by arc and chord B. because the sector is the area bounded by diameter and arc C. because the sector is the area bounded by an arc and two radiuses D. because the sector is the area bounded by a chord and two radiuses
Tier 4	:	My level of confidence in choosing the reason for the answer: ○ Sure ○ Not sure

The Difficulty Index Items of Four-tier Diagnostic Test

The results of the student's answers were analyzed quantitatively to determine the DIf-I, Dis-I, and reliability of the test. Table 6 presents the level of difficulty of each item and the percentage of each level.

Table 6. The Level of Difficulty of the Items

Item number	Dif-I	Criteria
1	0.45	Moderate
2	0.55	Moderate
3	0.75	Too easy
4	0.55	Moderate
5	0.45	Moderate
6	0.65	Moderate
7	0.60	Moderate
8	0.60	Moderate
9	0.65	Moderate
10	0.35	Moderate
11	0.60	Moderate
12	0.65	Moderate
13	0.75	Too easy
14	0.55	Moderate
15	0.80	Too easy
16	0.40	Moderate
17	0.75	Too easy
18	0.45	Moderate
19	0.60	Moderate
20	0.60	Moderate

Table 7. Percentage of Each Level of Difficulty

Criteria	Items Number	Quantity	Percentage
Too easy	3, 13, 15, 17	4	20%
Moderate	1, 2, 4, 5, 6, 7, 8, 9, 10, 12, 14, 15, 16, 18, 19, 20	16	80%
Too difficult	-	0	0%

Tables 6 and 7 show that of the 20 items, there are four items (20%) that are too easy (items numbers 3, 11, 13, 17). The other items are moderate.

The Discriminant Index Items of Four-tier Diagnostic Test

The Dis-I is of each item presented in Table 8.

Table 8. The Discrimination Index of the Items

Item number	Dis-I	Criteria
1	0.3	Excellent
2	0.5	Excellent
3	0.3	Excellent
4	0.5	Excellent
5	0.3	Excellent
6	0.3	Excellent
7	0.4	Excellent
8	0.2	Acceptable
9	0.3	Excellent
10	0.3	Excellent
11	0.4	Excellent
12	0.4	Excellent
13	0.3	Excellent
14	0.3	Excellent
15	0.2	Acceptable
16	0.2	Acceptable
17	0.3	Excellent
18	0.3	Excellent
19	0.4	Excellent
20	0.4	Excellent

Table 9. Percentage of Each Criterion of the Discrimination Index

Criteria	Items Number	Quantity	%
Acceptable	8, 15, 16	3	15%
Excellent	1, 2, 3, 4, 5, 6, 7, 9, 10, 11, 12, 13, 14, 17, 18, 19, 20	17	85%

Table 8 and 9 present that of the 20 items, all of them (100%) have acceptable and excellent Dis-I.

The Reliability of Four-tier Diagnostic Test

The reliability of the test is indicated by the coefficient of reliability. It was estimated by using Cronbach Alpha formula. Measured by using the SPSS package program, obtained the reliability coefficient of 0.72.

The Response of the Students toward the Four-tier Diagnostic Test

Students' responses to the tests developed are shown in table 10.

Table 10. Student Responses

Statement	Score	Category
The instructions provided are clear and easy to understand	3.44	Very good
The language used in the questions is easy to understand	3.50	Very good
The time allotted is sufficient to complete the test	2.94	Good
The pictures in the questions are clear and easy to understand	3.26	Good
The number of items is sufficient	3.09	Good

To get the test runs well, the instructions were provided. It consists of the instructions to guidance students during the test. Their responses toward the instructions are clear and easy to understand. This makes them focused on the process of test. The language used in the test is Indonesia, their daily language. The structure of the sentences was constructed based on the General Guidelines for Indonesian Spelling. It makes students familiar and easy to understand the meaning of each statement. To answer 20 items, they have 80 minutes. In means in average, they have 4 minutes for each item. Student stated that it was enough to answer 20 items with four-tiers. From 20 items of, 15 of them provide the picture as the additional information of the item. Students stated that the pictures were clear and informative in supporting the information of the item. This type of test is four-tier. Each item contains four questions. This is a consideration in determining the number of questions. For the students, 20 items are enough and ideal for them to do it in their best performance.

The Level of Student's Understanding

The level of student's understanding reported based on their answer in each item and the criteria at Table 4. Table 11 presents the samples of responses of 3 students in answer item number 1.

Table 11. Sample of Students' Responses

Student	Tier 1	Tier 2	Tier 3	Tier 4	Conclusion
1	R	S	R	S	Understands the concept
2	F	S	F	S	Misconceptions
3	R	NS	F	NS	Partial understanding

The indicator of item number 1 is the student can identify the element of a circle. Student number 1 understands the concept of the element of a circle because he/she was sure and correct in answering tier-1 and tier-3. Student number 2 has a misconception in identifying the element of a circle because he/she was sure and wrong in answering tier-1 and tier-3. Student number 3 has a partial understanding of identifying the element of a circle because he/she was sure and right in answer tier-1, but he/she was unsure with his/her wrong answer in tier-3.

Discussion

In instructional, diagnostic testing is one type of test. It is also called the analytical test. Teachers use this test to obtain evidence that details a learner's progress on a given subject (Adom, Mensah, & Dake, 2020). In this study, 20 items of a four-tier diagnostic test were developed. After experts judged its validity, it was tested on 34 students. 80% of items are moderate, and 20% of items are too easy. The Dif-I of these (too easy) items is close to the ideal category. Their Dif-I spread from 0.75 to 0.80. The ideal Dif-I is between 0.3 to 0.7 (Musa et al., 2018).

Theoretically, too-easy items are less informative. This is because all students, both those with high and low abilities, answered the questions correctly. As a result, these items do not discriminate well between students' abilities (small difference). Items that are too easy or difficult tend to have poor discriminant power (Musa et al., 2018; Quaigrain & Arhin, 2017), although it may not always be like that (Hingorjo & Jaleel, 2012). In this test, item numbers 3, 11, 13, and 17 tend to have low discriminating power (0.20, 0.43, 0.38, and 0.32, respectively). Because the tests developed are diagnostic tests to identify student strengths and weaknesses and diagnose student misconceptions, they can still be used because they provide information on student strengths.

The consistency of the measurement results of these items is shown by the reliability coefficient obtained from estimation using the Cronbach Alpha formula. In the social, behavioral, and educational sciences, it stands for the most widely used indicator of internal consistency. It is usually interpreted as the average of all possible split-half coefficients (Mohajan, 2017). The test reliability coefficient is 0.72. This exceeds the minimum criteria for the reliability coefficient set by experts. The expected reliability coefficient is at least 0.8 (Quaigrain & Arhin, 2017), 0.5 (Gugiu & Gugiu, 2018), or 0.6 (Rudner & Schafer, 2002). Studies on the meta-analysis of the reliability coefficients reveal that the average reliability of published studies is only about 0.75, with a reporting coefficient of 75% greater than 0.70, a reporting coefficient of 49% greater than 0.80, and only 14% reporting coefficient greater than 0.90 (Peterson, 1994). It indicates that the diagnostic test developed gives consistent measurement. It means that an observed score for a measure actually matches its true score (Mohajan, 2017).

The test was also responded either by students as test users. They stated that the test instructions are easy to understand. The language used in the items is appropriate with their age and daily language. It is important because test questions must use language that is complies with the language used in learning (Clay, 2001) so that it is easy to understand so that they can understand the meaning of the problem.

Regarding the time to do the test, Haladyna's study in 2002 reviewed 27 psychology textbooks and 27 research on the taxonomy of multiple-choice texts, showing that there were no references that mentioned the ideal time to take the test (Brothen, 2012). Because it was developed in Indonesia, the time is adjusted to the time that is often used in national mathematics exams, 4 minutes for the one-tier type. Because this test is four-tier where students still must determine the level of confidence in answering, choosing reasons, and determining the level of confidence, the test developers added 2 minutes. Thus, the time to work on one test item is an average of 6 minutes. For 20 items, tester provides 120 minutes. Most of the students state that the time allotted sufficient for them to complete the test.

The pictures on the test are also presented clearly, and the number of test items is enough for them so that the concentration of students to do the test is stable. Based on this information, the diagnostic test developed is ideal for wider implementation to diagnose students' misconceptions of circle material. Identification of student misconceptions and their sources will assist teachers in overcoming and planning appropriate learning for their students (Ojose, 2015).

Using this test, the level of student's understanding of circle material can be identified based on the student's responses. Because each item represents the indicator of the competency measured, the analysis reports the level of their understanding of the indicator related.

Implication

This four-tier diagnostic test will help mathematics teachers at junior high schools in Indonesia to identify their students' level understanding at circle material. Whether they understand the concept, do not understand the concept, have partial understanding, or have misconception. For the students, it can be used to evaluate their comprehension of circle material. For the parents, its result will give information about their child's strengths or weaknesses in circle material. This report can be used as a consideration for following up by providing programs, for example by giving enrichment or remedial program. For the following research, this type of test can be developed for the other material in mathematics or other subjects.

Limitation

This study was only tried on 34 students from one school in Yogyakarta. If the participants came from more schools, the data would be more representative to get information on the test. From the aspect of the validity of the test, although it has been analyzed by three experts, it would be more valid when the experts are more.

CONCLUSION

The product of this development research is a four-tier diagnostic test for the circle material of junior high school students in Indonesia. The test consists of 20 items. It has been validated by three experienced experts in the field of mathematics learning. Of the 20 test items, on average, they have an ideal level of difficulty and have a good discrimination index. The consistency of the test is indicated by the reliability coefficient of 0.72. It was estimated using the Cronbach Alpha formula. Students as test users stated that the tests developed contained easy-to-understand instructions, easy-to-understand language, sufficient test time, clear pictures, and enough items.

REFERENCES

Adom, D., Mensah, J. A., & Dake, D. A. (2020). Test, measurement, and evaluation: Understanding and use of the concepts in education. *International Journal of Evaluation and Research in Education*, *9*(1), 109–119. https://doi.org/10.11591/ijere.v9i1.20457

Aldoobie, N. (2015). ADDIE Model. *American International Journal of Contemporary Research*, *5*(6), 68–72.

Almanasreh, E., Moles, R., & Chen, T. F. (2019). Evaluation of methods used for estimating content validity. *Research in Social and Administrative Pharmacy*, *15*(2), 214–221. https://doi.org/10.1016/J.SAPHARM.2018.03.066

Aydin, S., Keles, U., & Hasiloglu, A. (2012). Establishment for misconceptions that science teacher candidates have about geometric optics. *The Online Journal of New Horizons in Education*, *2*(3), 7–15.

Bhat, B. A. (2019). Formative and Summative Evaluation Techniques for Improvement of Learning Process. *European Journal of Business and Social Sciences*, *7*(5), 776–785.

Brothen, T. (2012). Time Limits on Tests: Updating the 1-Minute Rule. *Teaching of Psychology*, *39*(4), 288–292. https://doi.org/10.1177/0098628312456630

Caleon, I., & Subramaniam, R. (2010a). Development and application of a three-tier diagnostic test to assess secondary students' understanding of waves. *International Journal of Science Education*, *32*(7), 939–961. https://doi.org/10.1080/09500690902890130

Caleon, & Subramaniam, R. (2010b). Do students know What they know and what they don't know? Using a four-tier diagnostic test to assess the nature of students' alternative conceptions. *Research in Science Education*, *40*(3), 313–337.

https://doi.org/10.1007/s11165-009-9122-4

Chesky, N. Z., & Wolfmeyer, M. R. (2015). *Philosophy of STEM Education*. *Philosophy of STEM Education*. https://doi.org/10.1057/9781137535467

Clay, B. (2001). *Is This a Trick Question? Is This a Trick*. Kansas: Kansas Curriculum Center.

Dhakne-Palwe, S., Gujarathi, A., & Almale, B. (2015). Item Analysis of MCQs and Correlation between Difficulty Index, Discrimination Index and Distractor Efficiency in a Formative Examination in Community Medicine. *Journal of Research in Medical Education & Ethics*, *5*(3), 254–259. https://doi.org/10.5958/2231-6728.2015.00052.9

Fritz, A., Haase, V. G., & Rasanen, P. (2019). *International handbook of mathematical learning difficulties*. Cham, Switzerland: Springer.

Gugiu, C., & Gugiu, M. (2018). Determining the Minimum Reliability Standard Based on a Decision Criterion. *Journal of Experimental Education*, *86*(3), 458–472. https://doi.org/10.1080/00220973.2017.1315712

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

Hingorjo, M. R., & Jaleel, F. (2012). Analysis of one-best MCQs: The difficulty index, discrimination index and distractor efficiency. *Journal of the Pakistan Medical Association*, *62*(2), 142–147.

Kaltakci-Gurel, D., Eryilmaz, A., & McDermott, L. C. (2017). Development and application of a four-tier test to assess pre-service physics teachers' misconceptions about geometrical optics. *ReseaRch in science & Technological educaTion*, *35*(2), 238-260.

Kusmaryono, I., Basir, M. A., & Saputro, B. A. (2020). Ontological Misconception in Mathematics. *Infinity*, *9*(1), 15–30.

Kutluay, Y. (2005). *Diagnosis of Eleventh Grade Students' Misconceptions About Geometric Optic By a Three-Tier Test*. Ankara: Middle East Technical University.

Lestari, P., Mardiyana, M., & Slamet, I. (2020). Practicality Analysis of PBL-Based Mathematics in Circle Material. *Indonesian Journal of Science and Mathematics Education*, *03*(3), 282–291. https://doi.org/10.24042/ijsme.v3i2.7271

Li, Y., & Schoenfeld, A. H. (2019). Problematizing teaching and learning mathematics as "given" in STEM education. *International Journal of STEM Education*, *6*(1), 1-13. https://doi.org/10.1186/s40594-019-0197-9

Lin, Y. C., Yang, D. C., & Li, M. N. (2016). Diagnosing students' misconceptions in number sense via a web-based two-tier test. *Eurasia Journal of Mathematics, Science and Technology Education*, *12*(1), 41–55. https://doi.org/10.12973/eurasia.2016.1420a

Lu, L., & Sides, M. L. C. (2022). Instructional Design for Effective Teaching: The Application of ADDIE Model in College Reading Lesson. *NOSS Practitioner to Practitioner SPRING 2022*, *Spring*, 4–12.

Mohajan, H. K. (2017). Two Criteria for Good Measurements in Research: Validity and Reliability. *Annals of Spiru Haret University. Economic Series*, *17*(4), 59–82. https://doi.org/10.26458/1746

Musa, A., Shaheen, S., & Elmardi, A. (2018). Item difficulty & item discrimination as quality indicators of physiology MCQ examinations at the Faculty of Medicine Khartoum University. *Khartoum Medical Journal*, *11*(2), 1477–1468.

Ojose, B. (2015). Students' Misconceptions in Mathematics: Analysis of Remedies and What Research Says. *Ohio Journal of School Mathematics*, *72*(Fall), 30–34.

Peterson, R. A. (1994). A Meta-Analysis of Cronbach's Coefficient Alpha. *Journal of Consumer Research*, *21*(2), 381-391. https://doi.org/10.1086/209405

Prabowo, A., Anggoro, R. P., Adiyanto, R., & Rahmawati, U. (2018). Interactive Multimedia-based Teaching Material for Trigonometry. In *ICRIEMS 5 Journal of Physics: Conference Series* (Vol. 1097(1), p. 012138). IOP Publishing https://doi.org/10.1088/1742-6596/1097/1/012138

Prabowo, A., Rahmawati, U., & Anggoro, R. P. (2019). Android-based Teaching Material for Statistics Integrated with Social Media WhatsApp. *International Journal on Emerging Mathematics Education*, *3*(1), 93–104. https://doi.org/10.12928/ijeme.v3i1.11961

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), 1–11. https://doi.org/10.1080/2331186X.2017.1301013

Rawh, P., Samsudin, A., & Nugraha, M. G. (2020). Pengembangan Four-Tier Diagnostic Test untuk Mengidentifikasi Profil Konsepsi Siswa pada Materi Alat-Alat Optik. *WaPFi (Wahana Pendidikan Fisika)*, *5*(1), 84–89.

Rejeki, S., & Putri, R. I. I. (2018). Models to support students' understanding of measuring area of circles. *Journal of Physics: Conference Series*

(Vol. 948, No. 1, p. 012058). IOP Publishing. https://doi.org/10.1088/1742-6596/948/1/012058

Retnawati, H. (2016). Proving Content Validity of Self-Regulated Learning Scale (The Comparison of Aiken Index and Expanded Gregory Index). *Research and Evaluation in Education*, *2*(2), 155–164.

Retnawati, H., Arlinwibowo, J., & Sulistyaningsih, E. (2017). The Students' Difficulties in Completing Geometry Items of National Examination. *International Journal on New Trends in Education and Their Implications*, *8*(4), 28–41.

Rudner, L., & Schafer, W. (2002). *What Teachers Need to Know about Assessment*. *Student Assessment series*. Washington: National Education Association.

Sayyah, M., Vakili, Z., Masoudi Alavi, N., Bigdeli, M., Soleymani, A., Assarian, M., & Azarbad, Z. (2012). An Item Analysis of Written Multiple-Choice Questions: Kashan University of Medical Sciences. *Nursing and Midwifery Studies*, *1*(2), 83–87. https://doi.org/10.5812/nms.8738

Setiana, D. S., Ili, L., Rumasoreng, M. I., & Prabowo, A. (2020). Relationship between Cooperative learning method and Students' Mathematics Learning Achievement: A Meta-Analysis Correlation. *Al-Jabar : Jurnal Pendidikan Matematika*, *11*(1), 145–158. https://doi.org/10.24042/ajpm.v11i1.6620

Shete, A., Kausar, A., Lakhkar, K., & Khan, S. (2015). Item analysis: An evaluation of multiple choice questions in Physiology examination. *Journal of Contemporary Medical Education*, *3*(3), 106-109. https://doi.org/10.5455/jcme.20151011041414

Shim, G. T. G., Shakawi, A. M. H. A., & Azizan, F. L. (2017). Relationship between Students' Diagnostic Assessment and Achievement in a Pre-University Mathematics Course. *Journal of Education and Learning*, *6*(4), 364–371. https://doi.org/10.5539/jel.v6n4p364

Sudihartinih, E., & Purniati, T. (2019). Using geogebra to develop students understanding on circle concept. *Journal of Physics: Conference Series* (Vol. 1157, 4, p. 042090). IOP Publishing. https://doi.org/10.1088/1742-6596/1157/4/042090

Swaranjit, K. (2015). Application of mathematics in sciences. *International Journal of IT, Engineering and Applied Sciences Research*, *4*(6), 83–85.

Zaleha, Z., Samsudin, A., & Nugraha, M. G. (2017). Pengembangan Instrumen Tes Diagnostik VCCI Bentuk Four-Tier Test pada Konsep Getaran. *Jurnal Pendidikan Fisika Dan Keilmuan (JPFK)*, *3*(1), 36–42. https://doi.org/10.25273/jpfk.v3i1.980