REiD (Research and Evaluation in Education), 5(1), 2019, 10-20
Available online at: http://journal.uny.ac.id/index.php/reid

Methods used by mathematics teachers in developing parallel multiple-choice test items in school

[*1]Kartika Pramudita; [2]R. Rosnawati; [3]Socheath Mam
[1,2]Department of Educational Research and Evaluation,
Graduate School of Universitas Negeri Yogyakarta
Jl. Colombo No. 1, Karangmalang, Depok, Sleman, Yogyakarta 55281, Indonesia
[3]Faculty of Education, Royal University of Phnom Penh, Cambodia
Russian Federation Boulevard, Toul Kork, Phnom Penh, Cambodia
[*]Corresponding Author. E-mail: kartika_pramudita.2017@student.uny.ac.id

Submitted: 04 December 2018 | Revised: 12 February 2019 | Accepted: 14 February 2019

Abstract

The study was aimed at describing five methods of the development of parallel test items of the multiple-choice type in mathematics at Yogyakarta (primary education level). The study was descriptive research involving 22 mathematics teachers as the respondents. Data collection was conducted through interviews and document reviews concerning the developed test packages. A questionnaire was used to gather data about the procedure the teachers employed in developing the tests. Findings show that the teachers used five methods in developing the test item; namely (1) randomizing the item numbers; (2) randomizing the sequences of response options; (3) writing items using the same contexts but different figures; (4) using anchor items; and (5) writing different items based on the same specification table. All of the respondents stated that they developed the table of the specification before developing the test items and that most of them (77%) did the validation of the instruments in content and language.

Keywords: *parallel test items, test item development, mathematics evaluation, multiple-choice testing*

Permalink/DOI: https://doi.org/10.21831/reid.v5i1.22219

Introduction

Evaluation is one of the essential aspects of education that will contribute to the achievement of educational quality. One of the objectives of evaluation is to know students' real competence. Effective evaluation can differentiate between high- and low-achieving students. An effective evaluation gathers evidences that are valid concerning learning outcome. The process and product of evaluation are also able to give improvement to students' motivation and achievement in learning (Stiggins & Chappuis, 2012, p. 3). One type of evaluation conducted in school is cognitive evaluation. Cognitive evaluation can be performed by using tests that will show the individual or group characteristics (Rasyid & Mansur, 2008, p. 11).

Assessment for learning is integral to best practice in teaching and learning. The development of a measurable test instrument must be done through qualitative and empirical research. According to Mardapi (2008, p. 15), a test instrument, either test or non-test, must have evidence for validity and reliability so that test results can be comparable and economical. A test is said to be valid if it measures what it is supposed to measure. A test with high validity will have a low error of measurement, meaning that the scores obtained by testees are close to the original scores. A test is said to be reliable if the observed scores have a high correlation with the original scores. Sources for an instrument validity can be traced from the contents of the test, in the forms of qualitative analyses of the materials, constructs, and language of the test.

A test battery used in an evaluation can be in various test items. The test form selected must be in line with the objective of the testing. One common test form is the multiple-choice test. A multiple choice test item consists of a stem followed by several alternative responses (Kehoe, 1995b, p. 2). The multiple-choice test form is suitable for testing that involves an enormous amount of material, such as the national examination (NE) or national-standard school examination (NSSE). This is the superiority of multiple-choice testing in that it covers a high number of items, is objective, is efficient, and can be highly reliable (Reynolds, Livingston, & Willson, 2009, pp. 184–186). A multiple-choice test can measure all the thinking processes in the cognitive domain from the lowest to the highest levels. This can be highly suitable for testing in the field of mathematics (Torres, Lopes, Babo, & Azevedo, 2011, p. 11). A number of studies have been done for the evaluation of mathematics learning using the multiple-choice test mode. One study is conducted to measure the high-order thinking skills in mathematics for junior high schools students using a multiple-choice test with four options (Rosnawati, Kartowagiran, & Jailani, 2015, pp. 189–196).

Multiple-choice tests frequently studied are those of the NE and NSEE. Some of the problems related to the use of these two tests are the quality of the test and frauds frequently occur during test administrations. A study shows that, based on item response theory analyses, of the 40 items of the Mathematics NE for the junior high school, 28 are good and 12 are poor (Kartianom & Mardapi, 2017, p. 172). To look at the fraud practices during the administering of the national examination can be done from the NE integrity indexes.

In some regions, integrity indexes are found low, showing high fraud in the administration of the exams. This condition indicates that students of the primary and junior secondary schools are still fearful of the exams, although the results are not the only determinations for passing. The national exam, however, is used as a criterion for admission to the higher school level. For such, students give all kinds of efforts to get good results; one of which is by sharing answer keys. The multiple-choice system makes it possible for the test takers to exchange answers easily. This chance raises illegal cooperation among the test takers, which cause the test results to be invalid. Consequently, the exam results do not at all reflect the real competences of the students.

This problem needs a solution. One solution taken by the government is by giving out several parallel tests. Development of parallel tests takes different ways among subject matters in its method and rules. In the mathematics subject matter, item stems and options involve a lot of figures. Differences in the figures can have an impact on the levels of item difficulties. Even numbers and odd numbers give different difficulty levels. The choice of distractors also influences difficulty levels. In the development of the test packages for mathematics, therefore, must obey the rules.

In another angle, mathematics teachers are expected to prepare the students in approaching the national examination. In order to know the teachers' readiness to do it, research needs to be conducted. A study on the competence and readiness of mathematics teachers looked at the self-efficacy of mathematics teachers in Yogyakarta. The findings show that the self-efficacy of 43.07% of the teachers is at the low category, 55.47% at the medium category, and the rest 1.46% at the high category (Widdiharto, Kartowagiran, & Sugiman, 2017, pp. 69–75). These findings indicated that teachers' confidence in facing the NE was at the medium level. Probing further on the competence and readiness of teachers in approaching the NE and NSSE, it was necessary to know the teachers' competencies in developing test practices and try-outs for the NE. The purpose of the try-outs was to see each student's competence achievement to be used as a basis for improvement activities. It is, therefore, crucial that the test items developed by teachers be functional in showing the students' competences.

Another thing to be conducted is that which could minimize students' interaction in doing the test. This minimalization is done by developing several test packages. The packages should be parallel so that they would not raise a new problem. A parallel test must have

identical objective, difficulty level, and format so that the test will be the same, but the items will be different. If the packages have been able to minimize frauds but have different levels of difficulty, the results will not be valid either. It is, therefore, necessary that the development of the test packages consider the parallelism of the items that are developed by teachers through a variety of methods. Before testing the parallelism of the test packages, it is necessary to gather information concerning the methods used by the teachers to develop the test packages. This paper is to figure out how teachers develop parallel test items of the multiple-choice type in mathematics.

Method

The research employed a descriptive research approach to obtain information about the methods that the teachers used for developing the mathematics test packages in the school. The study used interviews and document reviews as the test techniques and questionnaires as the non-test technique for collecting pertinent data. Open-ended interviews were given to 22 mathematics teachers. Each teacher was given the freedom to provide information to the method he/she used in developing the test packages. Each teacher was allowed to have more than one response, depending on his/her experiences.

The research instrument used to gather data was an interview guide. It contained questions about the methods to be used by the teachers to develop the test packages and the reasons for selecting the methods. In order to obtain evidence that the teachers did use the packages, documents review was done. Besides finding that the packages were there, it was also used for finding results of the tests to the students.

The questionnaires were used to look at the procedures for developing the packages. They were used to know the steps the teachers employed in developing the packages from the formulation of the objectives, construction of the specification table, to the item validation of content and language. They were also used to obtain evidence on the consistency between the item development and the test development procedure. The questionnaires were completed by check and cross marks. A check mark was given if a teacher did the step in the test development, a cross mark when a teacher did not.

Findings and Discussion

Findings

The key findings of the study are that in developing mathematics test packages, teachers had applied five methods including (1) randomizing the item numbers; (2) randomizing the sequences of response options; (3) writing items using the same contexts but different figures; (4) using anchor items; and (5) writing different items based on the same specification table.

The majority of teachers up to 37 % (of 22 teachers) used the same contexts with different figures to construct test items (as seen in Figure 1). It was followed by 21 % that developed different test items from the same table of specification. Meanwhile, other proportions developed the same items in different item numbers, developed the same items with different orders for the options, and used anchor items.

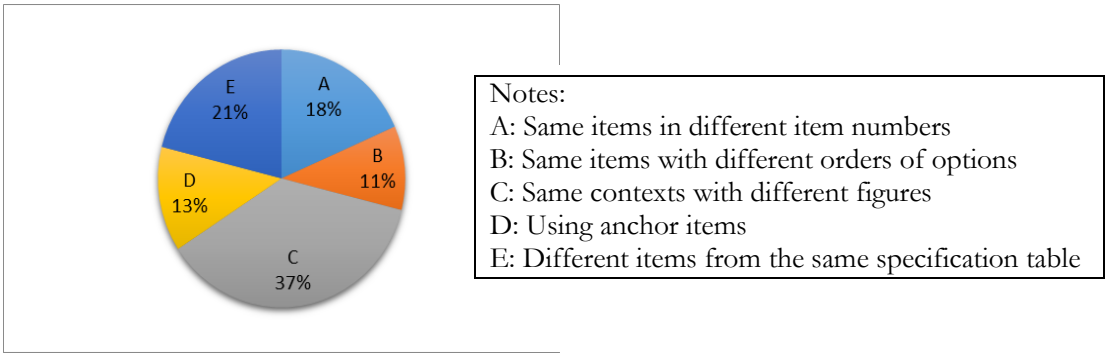

Figure 1. Methods of test package development

ISSN 2460-6995

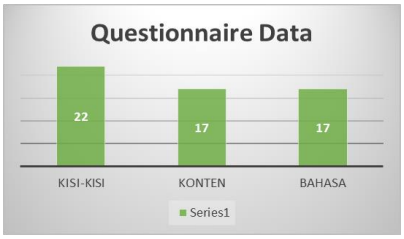

Figure 2. Data of instrument validation

Table 1. Randomization of the orders of response options

Package	Sample Item
Package 1 Sequencing from small to large numbers	Line gradient passing through point A(n, 3) and B(6, 2n) is 7. Value of n is … . A. 2 B. 3 C. 4 D. 5
Package 2 Sequencing from large to small numbers	Line gradient passing through point A(n, 3) and B(6, 2n) is 7. Value of n is … . A. 5 B. 4 C. 3 D. 2

Figure 2 presents a diagram of the results of questionnaires completed by 22 respondent teachers. It shows that all the teachers constructed the specification table before beginning to write the test items. Next, 17 teachers had their items validated in content and language by peer teachers. The rest five teachers did not have their items validated.

In developing test items, one should follow all steps set up in the procedure. After writing the items, teachers should have subjected them to peer validation by their colleagues as experts (Torres et al., 2011, p. 7).

Randomizing Item Numbers

From the interview, 18% of the teacher state they randomized items numbers to produce parallel items. Thus, the same test items were developed but were sequenced in different numbers. The difficulty levels and differentiating powers of the items were the same. The distractor functioning was the same too because identical distractors were used.

The method of randomizing item numbers is easy to use, does not take much time, and produces many test packages, as many as the test items. The interview reveals that some respondents commented that developing the items by changing the options order gave advantage to the students who got an item order that is the same with the content order. However, those who got items orders that are different from the content orders were put to a disadvantage because mathematics is built of axiomatic and deductive systems such that content sequences are highly compact.

Randomizing Sequences of Options

A total of 11% of respondents experienced randomizing the order of the response options. In developing multiple-choice test, randomizing the response options orders can minimize illegal interaction among the testees. The interview result reveals that randomizing the order of the options may result in two possibilities. First, if students find out that the options are different only in the orders, they can work out a way to interact with each other. In other words, this method still makes it possible for them to interact although they get different test packages. Second, if the students do not realize that the tests are different only in the options orders, they will not get advantage from their interaction. Thus, in this case,

the method functions well in minimizing frauds.

At least two test packages is needed in using this method, since sequencing can be done in two ways; from small to great or from great to small. An example of test package development by altering the options size is shown in Table 1. In Table 1, the stem in the two packages is the same, but the options order is different, although the options are the same. Test packages that have all the options in figures can only be developed in two different versions. If the response options are not in the form of figure, more packages can be obtained (Table 3). In this version, the stem and options are the same, but the options order is different. The number of packages that can be developed depends on the number of options. For example, a three-option item can be sequenced in several versions (Table 2).

Table 2. Randomization of response options

Package	Option Order
Package 1	A. P1
	B. P2
	C. P3
Package 2	A. P1
	B. P3
	C. P2
Package 3	A. P2
	B. P3
	C. P1
Package 4	A. P2
	B. P1
	C. P3
Package 5	A. P3
	B. P1
	C. P2
Package 6	A. P3
	B. P2
	C. P1

Table 3. Randomization of response options

Package	Item
Package 1	Line equation that passes the point (0. -2) and point (4. 1) is … . A. $y = \frac{3}{4}x - 2$ B. $y = \frac{4}{3}x - 2$ C. $y = \frac{3}{4}x + 2$ D. $y = \frac{4}{3}x + 2$
Package 2 Option A is exchanged with D and B with C.	Line equity that passes the point (0. -2) and point (4. 1) is … . A. $y = \frac{4}{3}x + 2$ B. $y = \frac{3}{4}x + 2$ C. $y = \frac{4}{3}x - 2$ D. $y = \frac{3}{4}x - 2$
Package 3 Option A is exchanged with C and B with D.	Line equation that passes the point (0. -2) and point (4. 1) is … . A. $y = \frac{3}{4}x + 2$ B. $y = \frac{4}{3}x + 2$ C. $y = \frac{3}{4}x - 2$ D. $y = \frac{4}{3}x - 2$
Package 4 Option A is exchanged with B and C with D.	Line equation that passes the point (0. -2) and point (4. 1) is … . A. $y = \frac{4}{3}x - 2$ B. $y = \frac{3}{4}x - 2$ C. $y = \frac{4}{3}x + 2$ D. $y = \frac{3}{4}x + 2$
Etc.	

ISSN 2460-6995

From Table 2, it can be seen that six test packages can be developed from changing the orders of three response options. The number of test packages obtained by randomizing the orders of the options is n!, where n is a number of options. The number of packages can increase if a number of ways of item combination are impacted by certain items. Suppose Package 1 is an initial item; Packages 2 – 6 can be constructed in the way shown in Table 3. Package 7 can be constructed by exchanging options A and B on an even item and B and C on an odd item. This way of combining items will give a larger number of packages.

Table 3 presents an original item of two test packages with no order of options. Package 2 is obtained by changing options A and B in the initial item. The number of options influences the number of packages. Generally, the mathematics items for primary and junior secondary schools have four options, while senior secondary schools have five options.

The method of constructing test items by changing the orders of the options is intended to maintain item characteristics. Also, the distractors are also expected to function effectively. Numerous studies have been conducted that are related to the quality of multiple-choice tests. The studies commonly look into the quality of items in terms of levels of difficulty, differentiating powers, and distractor effectiveness. In addition to revealing information about test qualities, these studies also look into aspects that need to be improved to increase the quality of tests to be able to measure well.

Constructing Items Using the Same Context but Different Figures

In the study, 37% of the respondents constructed the test items using the same contexts but different figures. This method (see Table 4) results in two test packages that will be able to minimize the testees' interaction.

The teachers revealed that this method of test construction decreases the students' chance to cooperate. However, item construction using this method should be done carefully by paying full attention to the figures being used in each package. Even though the figures in each package are different, care must be taken in terms of even and odd figures since there are different perceptions of these figures between boys and girls (Wilkie & Bodenhausen, 2015, pp. 3–9). Besides, the size of the figures must also be taken into great account to make sure that the item difficulties are equal. Item difficulty levels influence discriminating powers; good items will we correctly answered by 30% to 80% of the testees (Kehoe, 1995a, p. 1). These percentages must be taken care of so that the test administration is minimized from frauds, and the results are fair to all the testees.

Table 4. Items with the same context but different figures

Package	Item
Package 1	A room with an air-conditioning of 3^{o}C. After the device is activated, the room temperature reduces 2^{o}C every 4 minutes. When the air-conditioner has been activated for 28 minutes, the room temperature will presently be … ^{o}C. A. –20^{o}C B. –15^{o}C C. –12^{o}C D. –11^{o}C
Package 2	An air-conditioning set is 5^{o}C. After it is activated, the temperature of the device reduces 4^{o}C every 8 minutes. When the air-conditioner has been activated for 32 minutes, its temperature will presently be … ^{o}C. A. 21 B. 16 C. –11 D. –59

Table 4 presents two test items into different test packages developed with the same context but different figures. Test item 1 uses figures that are relatively smaller than those in Package 2. The combination of even and odd numbers, however, is equal. Item 1 uses 3, and item 2 uses 5. These are two odd numbers with a small difference. Later on, Package 1 uses 4 while package 2 uses 8. Meanwhile, 32 and 28 are not far apart; both are two-digit and even figures.

Using Anchor Items

From the interviews, 13% of respondents developed the test packages using anchor items. Some studies have been done to obtain evidence for the functioning of anchor items. Studies show that the more anchor items used, the better the results are for the test equalization (Kartono, 2008, pp. 317–318). It means that anchor items function to equalize tests. One study increased the anchor items of a physics test up to 40%; the results show that items at the low, mid, and high difficulty levels are not yet equal (Abdullah, Mansyur, & Rosdiyanah, 2016, pp. 217–218). This inequality may be due to the fact that physics tests involve items with figures in them. The use of different figures in items will have an impact on the item difficulty levels. Even and odd figures also influence difficulty levels. Mathematics subject matter involves a lot of figures in its tests; and, thus, in using this method, developers must be accurate and careful to produce parallel tests.

Developing Items Using the Same Specification Table

Based on the results of the interviews, 21% of respondents constructed a test specification table and developed from it some different test packages. This mode of instrument development can be done in several ways, such as using various figures in the test items, making the same problem with different contexts, etc. This method of test development is effective in reducing frauds when the test is based on the teacher's narratives.

The two test items presented in Table 5 are developed from the same indicator, problem-solving in daily life using line arithmetic. The contexts and figures used in the items are different. In package 1, what is known is the first leg and amount of increase per year; while in package 2, what is known is the line from leg 1 to leg 3. The figures used in the two items are also different. The teacher needs to pay attention to these differences. The case is feared in which students can complete package 1 but not package 2 because of the different contexts. This condition may cause invalid testing so that the objective of the evaluation is not achieved. In order to prevent this from happening, it is suggested that teachers know and have information about parallel testing and the ways to develop parallel tests.

Table 5. Items constructed out of the same indicator

Type	Item
Package 1	Amount of sugar consumption by people in a village is 1,000 kg in 2013 and is always doubled each year. The total sugar consumption from 2013 to 2018 is … . A. 66,000 kg B. 65,000 kg C. 64,000 kg D. 63,000 kg E. 62,000 kg
Package 2	A scavenger collects trash plastic bottles. On the first day, he gets 2.5 kg, on the second day 3 kg, and on the third day 3.5 kg, and so forth following an arithmetic line system. If the plastic bottles are sold to a collector at Rp10,000.00/kg, in 15 days the scavenger earns …. A. Rp800,000.00 B. Rp900,000.00 C. Rp1,000,000.00 D. Rp1,200,000.00 E. Rp1,500,000.00

ISSN 2460-6995

Discussions

The research findings show that 18% of the respondents stated that they developed test packages by randomizing the order of the item numbers believed to be able to produce parallel sets of items. This method had also been done in the entrance testing at Muhammadiyah University of Bengkulu. The randomization of item numbers used the Linear Congruent Method (LCM) computer software. This selection system ran effectively (Gunawan & Prabowo, 2017, pp. 144–151). The test consisted of 100 items scheduled for 90 minutes. One of the test items is numerical. This test item has identical characteristics as numerical items tested in the school mathematics so that the method of randomizing the item numbers is effective. One advantage of this method of developing parallel tests of the multiple-choice type is that it can produce test packages in a large number. The number of test packages will be the same as the number of test items. It is the combination of all items in the test. A simple illustration of a test with three items can be seen in Table 6.

Table 6. Randomization of test item numbers

Package	Item Number
Package 1	1, 2, 3
Package 2	1, 3, 2
Package 3	2, 3, 1
Package 4	2, 1, 3
Package 5	3, 2, 1
Package 6	3, 1, 2

A test with three items can be developed into six test packages. The number of the packages is the combination of all the test items; so, if a test has an n item, the number of the packages that can be developed is n. A test consisting of 40 items can be developed by randomization of the item numbers into 40! packages.

Findings show that 11% of the respondents developed the packages by reordering the response options. In 2016, a study investigated the influence of distractor revision upon item validity and reliability. The study found that it did (Ali, Carr, & Ruit, 2016, pp. 6–9). Some other studies reveal that the quality of an item is influenced by the quality of distractors. Another study found that the quality of distractors has an impact on the item's difficulty level (Tarrant & Ware, 2010, pp. 539–543). The number of distractors, on the other hand, does not impact the item quality (Royal & Dorman, 2018, pp. 3–5). In conclusion, by maintaining the parallelism of the distractors, parallel instrument packages can be obtained.

In the interviews with the teachers, it was found that they randomized the response option by using google doc. It was a computer application for on-line testing. In the process, the teacher input a test set through the application. Google doc. would automatically shuffle the response options of each item. When the students open the application to do the test, they will get items with different orders of the options. This application helped teachers in providing test packages by using one initial test set. This computer application can be used with, of course, the backing of the school facilities for on-line testing. One weakness, however, lies in the fact that the computer application did not sequence figures from small to large or from large of small. It becomes a violation of the rules for randomizing response options. The use of google doc application must consider the form of the options. It would be best used for options that do not use series orders such as sizes of figures.

The method of constructing test packages from the same table of the specification was claimed by 37% of the respondents. Conditions and considerations must be taken into account when developing test packages using this method. However, not all the rules were followed. The teachers merely considered the contexts to get parallel levels of difficulty. As can be seen in Package 1 and Package 2, the options consist of one correct answer and three distractors.

Determining the correct answer within the options was almost not a problem. The problem lies, however, on providing distractors that can function well. Instrument development must also consider the parallel functioning of the distractors because they also contribute to the quality of the item. Distractors were made to lead low students to select them so that the item can distinguish between

low-achieving and high-achieving students. Worse, it should not happen that low-achieving students choose the correct answer while high-achieving students choose the wrong options. In this case, distractors do not function well. Table 7 presents some possibilities to help distractors functioning.

Based on Table 7, the possibilities of students' errors can be used as a basis for selecting distractors effectively. The distractors in Package 2 are 21, 16, and -59. For the item in Package 1, if distractors are calculated in the same way as they are in Package 2, the values 19, 14, and -53 are obtained. The item sample of Package 1 in Table 4 shows that the distractors are -20, -15, and -12. It shows that there is no parallelism in selecting distractors so that the item parallelism is doubted. Students' inaccuracy in doing Package 2 makes them choose the wrong options or distractors. Students' error in doing Package 1, if there are no good distractors, will induce them to try to find the correct answers. It may produce unfairness among testees.

From the interviews results, it is known that 13% of the respondents used anchor items to develop the packages. Development of test packages using anchor items has been done for NSSE for primary, junior secondary, and senior secondary schools, in addition to the NE. For the school examination (NSSE), the teachers were involved in developing the test items. Some items are standardized by the government, and the other is developed by the teachers. This is the anchor-based development. The anchor items function to equalize one item among the others. It is expected that the test will be able to reveal students' competencies across regions using tests that are different but equal.

Based on the results of the interviews, 21% of respondents developed different test items from the same specification table. This method requires extra time when many packages are expected to be produced. Besides, the characteristics of the items produced may not be the same so that it needs the difficulty levels testing of the items in each package. In the practice of developing different items from the same specification table of the national level, it is never achieved to produce different items having the same difficulty level albeit being developed from the same table of specification (Herkusumo, 2011). This thought must be considered when developing different test packages based on the same specification table.

Table 7. Possibilities of errors made by testees

Type	Possibility of errors of Package 2	Possibility of errors of Package 1
Error 1	Option 21 is obtained from: $\frac{32\ \text{min}\ utes}{8\ \text{min}\ utes} = 4$, so decreasing 4 times. The decrease in temp. in 32 min: $4 \times 4^{o}C = 16°C$ Room temp after 32 min: $16°C + 5°C = 21°C$ (Room temp after)	 $\frac{28\ \text{min}\ utes}{4\ \text{min}\ utes} = 7$, so decreasing 7 times. The decrease in temp. in 28 min: $7 \times 2^{o}C = 14°C$ Room temp after 28 min: $14°C + 5°C = 19°C$ (Room temp after)
Error 2	Option 16 is obtained from: $\frac{32\ \text{min}\ utes}{8\ \text{min}\ utes} = 4$, so decreasing 4 times. The decrease in temp. in 32 min: $4 \times 4^{o}C = 16°C$ (Testee stops at temp drop)	 $\frac{28\ \text{min}\ utes}{4\ \text{min}\ utes} = 7$, so decreasing 7 times. The decrease in temp. in 28 min: $7 \times 2^{o}C = 14°C$ (Testee stops at temp drop)
Error 3	Option -59 is obtained from: $\frac{32}{4} = 8$, (error in selecting a number to calculate temp.). The decrease in temp. in 32 min: $8 \times 8 = 64°C$ $5°C - 64°C = -59°C$	 $\frac{28}{2} = 14$, (error in selecting a number to calculate temp.). The decrease in temp. in 28 min: $14 \times 4 = 56°C$ $3°C - 56°C = -53°C$

ISSN 2460-6995

Conclusion and Suggestions

Conclusion

Teachers can use various methods of developing mathematics test packages by randomizing the item number, reordering the response options, using the same context with a different figure, using anchor items, and using the same table of specification. These methods are applied based on the respondents' logical thinking supported by analyses proposing that the test packages being developed are parallel. However, no theoretical bases have been used by the teachers in developing the tests. All the teachers used a specification table to develop tests while most of them had validated content and language.

Suggestions

Further research is needed to look at how the parallelism of the test packages can be developed among those five methods. Such research will be useful for the teachers to improve their theories and knowledge in developing parallel multiple-choice test items so that their evaluation of students is valid and reflect the real students' competences.

References

Abdullah, S., Mansyur, M., & Rosdiyanah, R. (2016). Pengaruh jumlah butir anchor terhadap hasil penyetaraan tes berdasarkan teori respon butir. *Jurnal Kependidikan: Penelitian Inovasi Pembelajaran*, *46*(2), 207–218. https://doi.org/10.21831/JK.V46I2.10935

Ali, S. H., Carr, P. A., & Ruit, K. G. (2016). Validity and reliability of scores obtained on multiple-choice questions: Why functioning distractors matter. *Journal of the Scholarship of Teaching and Learning*, *16*(1), 1–14. https://doi.org/10.14434/josotl.v16i1.19106

Gunawan, G., & Prabowo, D. A. (2017). Sistem ujian online seleksi penerimaan mahasiswa baru dengan pengacakan soal menggunakan Linear Congruent Method (Studi kasus di Universitas Muhammadiyah Bengkulu). *Jurnal Informatika Upgris*, *3*(2), 143–151. https://doi.org/10.26877/jiu.v3i2.1872

Herkusumo, A. P. (2011). Penyetaraan (equating) ujian akhir sekolah berstandar nasional (UASBN) dengan teori tes klasik. *Jurnal Pendidikan Dan Kebudayaan*, *17*(4), 455–471. https://doi.org/10.24832/jpnk.v17i4.41

Kartianom, K., & Mardapi, D. (2017). The utilization of junior high school mathematics national examination data: A conceptual error diagnosis. *REiD (Research and Evaluation in Education)*, *3*(2), 163–173. https://doi.org/10.21831/reid.v3i2.18120

Kartono, K. (2008). Penyetaraan tes model campuran butir dikotomus dan politomus pada tes prestasi belajar. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *12*(2), 302–320. https://doi.org/10.21831/pep.v12i2.1433

Kehoe, J. (1995a). Basic item analysis for multiple-choice tests. *Practical Assessment, Research & Evaluation*, *4*(10), 1–3.

Kehoe, J. (1995b). Writing multiple-choice test items. *ERIC/AE Digest Series EDO-TM-95-3*, *3*, 1–6.

Mardapi, D. (2008). *Teknik penyusunan instrumen tes dan nontes*. Yogyakarta: Mitra Cendekia.

Rasyid, H., & Mansur. (2008). *Penilaian hasil belajar*. Bandung: CV Wacana Prima.

Reynolds, C. R., Livingston, R. B., & Willson, V. L. (2009). *Measurement and assessment in education* (2nd ed.). Upper Saddle River, NJ: Pearson.

Rosnawati, R., Kartowagiran, B., & Jailani, J. (2015). A formative assessment model of critical thinking in mathematics learning in junior high school. *REiD (Research and Evaluation in Education)*, *1*(2), 186–198. https://doi.org/10.21831/reid.v1i2.6472

Royal, K., & Dorman, D. (2018). Comparing item performance on three- versus four-option multiple choice questions in a veterinary toxicology course.

Veterinary Sciences, *5*(2), 55. https://doi.org/10.3390/vetsci5020055

Stiggins, R. J., & Chappuis, J. (2012). *An introduction to student-involved assessment for learning*. Boston, MA: Pearson.

Tarrant, M., & Ware, J. (2010). A comparison of the psychometric properties of three- and four-option multiple-choice questions in nursing assessments. *Nurse Education Today*, *30*(6), 539–543. https://doi.org/10.1016/j.nedt.2009.11.002

Torres, C., Lopes, A. P., Babo, L., & Azevedo, J. (2011). Improving multiple-choice questions. *US-China Education Review*, *B*(1), 1–11.

Widdiharto, R., Kartowagiran, B., & Sugiman, S. (2017). A construct of the instrument for measuring junior high school mathematics teacher's self-efficacy. *REiD (Research and Evaluation in Education)*, *3*(1), 64–76. https://doi.org/10.21831/reid.v3i1.13559

Wilkie, J. E. B., & Bodenhausen, G. V. (2015). The numerology of gender: Gendered perceptions of even and odd numbers. *Frontiers in Psychology*, *6*, 810. https://doi.org/10.3389/fpsyg.2015.00810

ISSN 2460-6995