

Developing a Test for Testing Students' Mathematics Competency as Prospective Secondary School Mathematics Teachers

Awil¹, Syahrullah Asyari², Muhammad Darwis M³, Sahlan Sidjara⁴, Ikhbariaty Kautsar Qadry⁵

{awi.dassa@unm.ac.id¹, syahrullah_math@unm.ac.id², darwismath2011@gmail.com²}

Universitas Negeri Makassar, Jln. Dg. Tata Raya Makassar^{1,2,3,4,5}

Abstract. This was developmental research aimed at developing a test for testing students' mathematics competency as prospective secondary school mathematics teachers. The research respondents were 32 third-year students of the department of mathematics, Universitas Negeri Makassar. The test was developed through devising the test specification, writing down, reviewing, trying out, and analyzing the items of the test from both difficulty level and item discrimination. The item discrimination coefficient (c) was from -1 to 1 . The analysis results were classified into three categories: accepted without revision, if $0.30 \leq c \leq 1$, accepted with revision, if $0.10 \leq c \leq 0.29$ and rejected if $-1 \leq c \leq 0.10$. There were 45 multiple-choice items analyzed. The analysis results showed that 21 items were accepted without revision, 17 items were accepted with revision in terms of the language construction and the difficulty level of the items, whereas the rest were rejected. Thus, 38 of the 45 items were feasible to use.

Keywords: Test, Competency, Prospective Mathematics Teacher.

1 Introduction

Verily, students of mathematics education are prospective mathematics teachers either at junior or senior secondary schools. As prospective mathematics teachers, they are certainly expected to realize the importance of mathematics as a means to fulfill human needs and to develop sciences. Moreover, they are demanded to master mathematics to be able to contribute to national interest and competitiveness improvement. Unfortunately, several facts are showing that the quality of students of mathematics education in Universitas Negeri Makassar still attracts more attention from stakeholders to make them ready to contribute and compete with. Our experiences in examining students' theses in the last three years 2016 to 2018 indicated that there were some fundamental problems in terms of the mathematical knowledge of many mathematics education students. If the problem is not solved, then it will likely block graduations to be professional teachers who are ready to contribute to national interest and competitiveness improvement.

Some examples regarding the fundamental problems were the existence of many students who did not know the way to show whether two straight-line equations[6], intercept or not; their difficulty in converting repeated decimal form to the general form of rational number, and the lack of their attention to the way of writing mathematical solution correctly, since considering that the final result was the most important one, etc. Also, they had not understood the

significance of condition holding on an equation, so that they sometimes did not write it down when solving a question or problem; some were confused if they were dealt with the difference from two sets or the complement of a set; there were also students composing two functions, without paying attention to the condition holding on, etc.

According to Fennema & Franke, no doubt that teacher's knowledge is one of the most influential things in the classroom activities towards students learn [1]. There have been researches about the relationship and influence of a teacher's mathematical knowledge towards students' learning achievement[5][7][8]. These indicate that the quality of a prospective mathematics teacher gains more attention from many parties.

As an effort to solve the problem, we then design a test that could be used to measure mathematical knowledge of prospective secondary school mathematics teachers. The test is expected to be able to reveal the feasibility of mathematics education students to be graduates who master mathematics.

2 Research methods

This was developmental research. This is oriented to the development of a product whose developmental process is described thoroughly, and the product obtained is then evaluated [2]. This research was conducted in the Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Negeri Makassar by trying out the test to the third-year students. To make it easier in arranging a test instrument, there are eight phases in developing a test, that is: (1) arranging a test specification; (2) writing down a test; (3) examining the test; (4) trying out the test; (5) analyzing the question items; (6) improving the test; (7) assembling the test; (8) applying the test; (9) interpreting the test results [3]. Within this research, however, the test development was just limited to the fifth phase, i.e analyzing the question items, for some reasonable reasons. The analysis of the question items was limited to the analysis of difficulty level and item discrimination.

3 Results dan discussion

3.1 Results

Forty-five items of multiple-choice questions had been tried out to 32 third-year mathematics education students of the department of mathematics, Universitas Negeri Makassar. Further, it was analyzed question items based on the students' responses. The analysis results are presented as follows.

Difficulty Level. In general, according to the classic theory, the difficulty level could be stated in some ways, that is (1) the proportion of test-takers responding correctly, (2) scale of linear difficulty, (3) Davis index, and (4) bivariate scale. The proportion of correct answers i.e the number of test-takers responding correctly on the question items analyzed, and then compared to all test-takers constitutes the most common difficulty level to use. The equation (1) used to determine the difficulty level with the proportion of correct answers is:

$$p = \frac{\sum x}{S_m N} \quad (1)$$

Where:

p is the proportion of test-takers responding correctly or difficulty level

$\sum x$ is the number of test-takers responding correctly

S_m is the maximum score

N is the number of students as test-takers

where the difficulty level is categorized as follows. Difficult (if $p < 0.3$), moderate (if $0.3 \leq p \leq 0.7$), and easy (if $p > 0.7$).

Item Discrimination. One of the quantitative analysis objectives of a question is to determine whether a question could differ groups from aspects measured by the difference existing in the group. The index used in differing test-takers with high ability from those with low ability in mathematics is so-called item discrimination. The item discrimination of the questions is the difference in the proportion of test-takers responding correctly in each group. The index shows the conformity between the function of questions and that of test wholly. Thus, the validity of the question is equal to the item discrimination, which is the ability to differ the test-takers with high ability from those with low ability in mathematics.

The number which shows the measure of item discrimination is from -1 to 1. The negative sign shows that the test-takers with low ability can answer a question correctly, whereas those with high ability cannot answer it correctly. Thus, the question whose index of item discrimination is negative shows the contrary to the quality of the test-takers.

Index of item discrimination is determined based on the division of the group into two parts, which is the upper group constituting the group of test-takers with high ability in mathematics, and lower group constituting the group of test-takers with low ability in mathematics. The high ability is indicated by the acquisition of a high score, and the low ability is indicated by the acquisition of low score. Crocker & Algina state that the index of item discrimination is defined as the difference in the proportion of correct answers in the upper group and that of correct answers in the lower group. The groups can be divided employing various methods which depend on our needs. According to Kelley, Crocker, and Algina, the most stable and sensitive method along with used widely is by determining 27% of the upper group and 27% of the lower group [4]. For many purposes, the group division can be 50% of the upper group and 50% of the lower group or about 25% to 35% of the upper group and lower group. In this research, it is used the group division as 50% - 50% for at least two reasons, i.e. the lack number of students as the test-takers and the possibility to compute the index of item discrimination of all test-takers.

To ease the computation, the score acquisition is put in the right order that is from the test-takers with high scores to those with low scores in mathematics. After ordering the scores, then subsequently dividing the class into two equal parts, that is 50% of the upper group and 50% of the lower group.

The item discrimination according to the index of item discrimination can be determined by using the following equation (2).

$$D = \frac{\sum A}{n_A} - \frac{\sum B}{n_B} \quad (2)$$

Where:

D is the index of item discrimination

$\sum A$ is the number of test-takers responding correctly in the upper group

$\sum B$ is the number of test-takers responding correctly in the lower group

n_A is the number of test-takers of the upper group

n_B is the number of test-takers of the lower group

But, because in this research, the number of test-takers of the upper group is equal to the number of those of the lower group, that is: $n_A = n_B = n$, so that the equation (3) of the item discrimination becomes:

$$D = \frac{\sum A - \sum B}{n} \quad (4)$$

The item discrimination of each question item is determined by using the following equation (4).

$$D = DLU - DLL \quad (5)$$

Where:

D is the index of item discrimination

DLU is the difficulty level of the upper group

DLL is the difficulty level of the lower group

The following table (Table 2) shows the results of computation of the item discrimination.

The difficulty level influences directly to the item discrimination index. If each person chooses an option as a correct answer ($p = 1$), or if each person chooses an option as an incorrect answer ($p = 0$), then the question cannot be used to differ the ability of students from one another. In other words, the question cannot show the existence of the difference between students with high scores and those with low scores in mathematics.

It has been stated that the number showing the measure of item discrimination is from -1 to 1 or the minimum index of the item discrimination is -1 , whereas the maximum index of that is 1 . It means that the closer the item discrimination index to 1 , the better the question to differ students with high scores from those with low scores in mathematics. The negative sign shows that the students with low ability in mathematics can answer the question correctly, whereas those with high ability in mathematics answer it incorrectly. Meanwhile, the positive sign shows that the students with low ability in mathematics can answer a question incorrectly, whereas those with high ability in mathematics answer it correctly.

Criteria for choosing questions. After computing the difficulty level and item discrimination statistically, then the subsequent step is choosing the question items. The results of item analysis are in general divided into three categories that are accepted without revision,

accepted with revision and rejected. Actually until nowadays there has not been a standard as a certain reference in determining the validity and the difficulty level of a good question item. According to Nitko, the criteria for choosing question items depend on the purpose of use, which is a general or particular purpose. If the test purpose is the appropriateness in ranking the test-takers ability in a certain discipline, then the measure of the difficulty level and question validity is not able to determine. But, if the thing measured is only one competency aspect, then the difficulty level should be between 0.16 and 0.84. if the thing measured is a collection of competency aspects, then the difficulty level should be between 0.30 and 0.70 and the validity of question items is suggested to be more than 0.30. Table 8 shows the criteria for accepting or rejecting a question item [4].

In this research, question item choice is based on Table 1. The question item accepted is that whose coefficient of the difficulty level is from 0.30 through 0.70. The difficulty level indicates that the items are homogenous and can produce a vast score distribution. The item with the coefficient of validity greater than 0.3 constitutes a good item, as Nunnally stated. If the coefficient of the item is greater than 0.30, then it is categorized to be that which can differ students with high ability and those with low ability in mathematics.

In this research, the determination of item acceptance status refers to the analysis results based on the item discrimination. This is because the question items developed will be applied for the students as prospective secondary school mathematics teachers in the department of mathematics, Universitas Negeri Makassar as the prerequisite which determine the feasibility of a student to step forward to the research proposal seminar to arrange his or her thesis.

Table 1. The Criteria for Choosing Multiple-Choice Questions.

Criteria	Coefficient	Decision
	0.30 to 0.70	Accepted
Difficulty Level	0.10 to 0.29 or 0.70 to 0.90	Revised
	< 0.10 and > 0.90	Rejected
Item	> 0.30	Accepted
Discrimination	0.10 to 0.29	Revised
	< 0.10	Rejected
Answer Proportion	> 0.05	

3.2 Discussion

The test development and its item analysis were finished. At last, based on Tables 2 and 3, we know that if it is viewed either from the aspects of the difficulty level or the item discrimination, there is an intersection between the questions accepted, revised or rejected. The question items accepted or considered as good according to the two aspects are 9 (nine), that is question number 3, 5, 8, 10, 16, 23, 33, 36 and 40. Meanwhile, the good enough questions and need revision according to the two aspects are also 9 (nine), that is question number 9, 14, 15, 17, 21, 22, 31, 32, and 41. Whereas the rejected questions according to the two aspects are 5 (five) which is question number 6, 18, 26, 37, and 42.

Table 2. The Status of Acceptance of 45 Multiple-Choice Questions Based on the Results of the Difficulty Level Analysis.

No	Question Status		
	Accepted	Revised	Rejected
1	3	1	4
2	5	2	6
3	8	7	11
4	10	9	18
5	16	12	20
6	23	13	24
7	28	14	26
8	33	15	30
9	36	17	37
10	39	19	38
11	40	21	42
12	44	22	
13	45	25	
14		27	
15		29	
16		31	
17		32	
18		34	
19		35	
20		41	
21		43	
Total	13	21	11

We learn from the result of the difficulty level and item discrimination analysis that questions that have good difficulty level might have poor item discrimination analysis. Or, a question might be rejected based on the difficulty level, but if it is viewed from the item discrimination, the questions might be accepted with revision. Or, a question might be revised, if it is viewed from the difficulty level, but if it is viewed from the item discrimination, the question is good. Therefore, a test designer should be in the scrutiny of the stages of arranging test by attempting to minimize mistakes or bias on a test.

The next stage is fixing or improving the questions which have not been suitable with expectation based on the item analysis. There are 45 multiple-choice items analyzed. The analysis results show that 21 items are accepted without revision, 17 items are accepted with revision in terms of the language construction and the difficulty level of the items, whereas the rest are rejected. Thus, after revision there are 38 of the 45 items are feasible to use. Subsequently, all question items which have been analyzed and revised are then unified to be one whole test. When unifying the test, one should also look at the things that can influence the validity of the questions like the consecutive number of the questions, the grouping of question items, layout, etc.

At last, we acknowledge the shortcoming of this research. It is because of the limitation of time and the difficulty of adjusting the course schedule and the classroom organization as the place for trying out the test designed so that the number of test-takers participating in this research is only one group consisting of 32 students. But actually, the case is, to get good analysis results, the number of question items should be at least 40 until 50 items and the number of test-takers is not less than 400 people [4].

Table 3. The Status of Acceptance of 45 Multiple-choice Questions Based on The Result of Item Discrimination Analysis.

No	Question Status		
	Accepted	Revised	Rejected
1	1	9	6
2	2	11	18
3	3	14	26
4	4	15	34
5	5	17	35
6	7	21	37
7	8	22	42
8	10	24	
9	12	28	
10	13	30	
11	16	31	
12	19	32	
13	20	38	
14	23	39	
15	25	41	
16	27	44	
17	29	45	
18	33		
19	36		
20	40		
21	43		
Total	21	17	7

4 Conclusions

Based on the analysis results on the aforementioned research results, it can be concluded that of 45 multiple-choice question items, there are 38 of them which are feasible to use, but some of them still need revision firstly. The revision is in terms of language construction, the difficulty level, and item discrimination. On the basis of the conclusion, it can be suggested as follows. First, reviewing the test developed, particularly the items accepted with revision, and then following it up by rewriting the questions and paying attention to the results of the difficulty level, item discrimination, and distractor function analysis. Second, this article should be completed with an analysis of the effectiveness of the function of multiple-choice item distractors. Third, the number of the try out test-takers should not be less than 400 people as recommended by experts in psychometry or not less than 10 different groups where there are about 40 people in each class.

References

- [1]Linsel, C & Anakin, M.: Diagnostic Assessment of Pre-Service Teachers' Mathematical Content Knowledge. Dalam Jenni Way (Eds.), *Mathematics Teacher Education and Development*. Melbourne: Mathematics Education Research Group of Australasia, Inc. pp. 4-27 (2012)

- [2]Richey, R.C. & Nelson, W.A.: Developmental research, In D. Jonassen (Ed.), Handbook of research for educational communications and technology. Macmillan, London. pp. 1213-1245 (1996)
- [3]Mardapi, D.: Teknik penyusunan instrumen tes dan nontes. Yogyakarta: Mitra Cendekia Press (2007)
- [4]Surapranata, S.: Analisis, Validitas, Reliabilitas dan Interpretasi Hasil Tes: Implementasi Kurikulum 2004 (Cet III). Bandung: PT Remaja Rosda Karya (2006)
- [5] Hill, Heather C., Brian Rowan, and Deborah Loewenberg Ball.: "Effects of teachers' mathematical knowledge for teaching on student achievement." *American educational research journal* 42.2. pp. 371-406. (2005)
- [6]Orton, Anthony.: "Students' understanding of differentiation." *Educational studies in mathematics* 14.3. pp.235-250. (1983)
- [7]Tella, Adedeji.: "The impact of motivation on student's academic achievement and learning outcomes in mathematics among secondary school students in Nigeria." *Eurasia Journal of Mathematics, Science & Technology Education* 3.2. pp. 149-156. (2007)
- [8] Polly, Drew, et al.: "Linking professional development, teacher outcomes, and student achievement: The case of a learner-centered mathematics program for elementary school teachers." *International Journal of Educational Research* 72. pp.26-37. (2005)