International Journal of Research Studies in Education
2015 October, Volume 4 Number 4, 85-92

Construction, validation and DIF determination of a test on problem solving for pre-service mathematics teachers

Torio, Von Anthony G.
Institute of Teaching and Learning, Philippine Normal University, Philippines (torio.vag@pnu.edu.ph)
Cabrillas-Torio, Myla Zenaida
Institute of Teaching and Learning, Philippine Normal University, Philippines (cabrillas.mzc@pnu.edu.ph)

ISSN: 2243-7703
Online ISSN: 2243-7711

OPEN ACCESS

Received: 7 January 2015 **Revised**: 18 January 2015 **Accepted**: 4 February 2015
Available Online: 28 February 2015 **DOI**: 10.5861/ijrse.2015.1052

Abstract

The field of Mathematics requires a lot of critical thinking and problem solving. These skills are honed as early as the basic education years of the students. Teacher training institutions should make it a necessity that they are able to cater to pre-service teachers who will be able to meet the demands of mathematics education. This paper aimed to construct, validate and determine the Differential Item Functioning (DIF) of a problem solving test for pre-service mathematics teachers. This test will later be used as a basis for designing interventions to improve the mathematical aptitude of pre-service teachers. Two universities were chosen with 100 third year students taking Bachelor of Science in Secondary Education major in Mathematics. The tests were constructed and validated by experts before administration to the participants. The participants took the examination and the test was improved based on the results of the item analysis. Differential Item Functioning was also tested using the Standardization Method. The study led to the development of a 60-item validated test on problem solving. After subjecting to DIF analysis, it was found that the two groups tested performed with no significant difference between individuals of similar abilities. The test yielded norm values per sub-skill of the test which will help gauge pre-service mathematics teachers' problem-solving skills.

Keywords: problem solving; pre-service teachers; mathematics education; Philippines; differential item functioning

Construction, validation and DIF determination of a test on problem solving for pre-service mathematics teachers

1. Introduction

A hundred years ago, mathematics teaching is simply focused on rote memorization of the four basic operations, addition, subtraction, multiplication and division. This emphasis on rote learning is no longer viable today. Students are now facing so many challenges and therefore must be proficient with skills needed in learning mathematics both as a subject and as a tool. The present K to 12 curriculum of the Philippines emphasizes that problem solving and critical thinking are the two ultimate goals of Mathematics Instruction. To achieve this goal, a teacher must be able to identify students' strengths and weaknesses in solving word problems in terms of the five sub-skills of problem solving such as (1) comprehension; (2) analysis; (3) organization; (4) identifying the process to be used; and (5) finding an answer and verification. This can be done through testing. A valid test must be able to fairly measure what it intends to measure. This brings the next concern of the study after development: validation and determination of possible differential item functioning for the future improvement of the test.

Item analysis is an integral part of the validation process in test construction. Determination of the difficulty index and discrimination indices are essential in the refinement of the examination. Once a test is constructed and validated, it is equally important to establish that the test is fair. The degree of fairness of the test is set through a method referred to as Differential Item Functioning (DIF) Analysis. This is what the study is about, it will deal with the construction of problem solving test for pre-service teachers, validate it and determine possible DIF. The problem solving capability of pre-service teachers is hypothesized to be one of the defining factors in the success of a teacher in the future. The test developed in this study is one way to detect problem solving difficulties of pre-service mathematics teachers for possible remediation and appropriate action.

1.1 Studies about Mathematics Education

The study of Language Factor in Mathematics tests Abedi and Lord (2001) showed that impact of students' language background on their performance on Math Word problems. The study found that English language learners scored significantly lower than proficient speakers in English. It appears therefore that modifying the linguistic structures in math word problems can affect student performance. This idea of the researcher was considered in the construction of the test on problem solving in the present study.

In the study on Teacher's Mathematical Beliefs by Handal (2003), it was argued that despite many educational reforms, a large number of teachers still perceive mathematics in traditional rather than progressive terms. As such, students have to learn mathematics by rote and removed from the application to real life situations. Teachers cannot deviate from this practice due to so many mediating factors such as pressure of examination, parent's traditional expectations, demands for covering the syllabus and supervisory style. Most of these factors also exist in our educational atmosphere. The present study will serve as starting point where teachers can start deviating from this traditional practice. Result of the test will describe the present characteristics of students in terms of problem solving.

Khairavi and Nordin (2011), in their study entitled "The Development and Construct Validation of the Mathematics Proficiency Test for 14-year-old students" aimed to assess the Mathematics proficiency of students in terms of conceptual understanding, procedural fluency and strategic competence. Result showed that the students were most proficient in conceptual understanding. The present study showed the same result as the previous study. While, Neil (2002) in his paper entitled "Issues in Constructing Formative Tests in Mathematics" aimed to give detailed information about student performance in mathematics. The test constructed assessed five

strands of the Mathematics Curriculum: Number, Measurement, Geometry, Algebra and Statistics. The present study aimed to assess the problem solving skill of students in terms of comprehension, analysis and organization, process to be used, finding an answer and verification.

1.2 Studies about Differential Item Functioning

Differential Item Functioning is a term that is confused with terms item bias sand item impact. Zumbo (2007) in an article about three generations of DIF Analyses made a clarification about DIF. He said in the article that in broad terms, asking the question "Is the test performing in the same manner for each group of examinees?" He proposed to have three generations of DIF praxis and theorizing. He named the first generation: Motivations for the Problem and Concept formation. In this generation, item bias is a commonly used term. In this generation, comparison of only two groups of examinees leading to the use of the terminologies: focal and reference groups denoting minority and majority groups. In this generation as well, the distinction between item impact and item bias is clarified. Item impact describes the situation in which DIF exists, due to true differences between groups in the underlying ability interest being measured by the item.

On the other hand, item bias is described the situations in which there is DIF because of some characteristic of the test item that is not relevant to the underlying ability of interest. Second Generation is entitled: Embodying the new terms and building frameworks for empirically investigating DIF. In this generation, the term DIF is widely accepted and impact and bias are considered separate terms. At least three frameworks for thinking about DIF were conceived by Zumbo (2007) in this generation: (1) Modeling item responses via contingency tables and/or regression models; (2) item response theory (IRT), and (3) Multidimensional models. The third generation of DIF is signaled by the "wanting to know why DIF occurs". The third generation is most clearly characterized as conceiving of DIF as occurring because of some characteristic of the test item and/or testing situation that is not relevant to the underlying ability of interest.

Schmitt and Dorans (1988) in their study entitled, Differential Item Functioning for Minority Examinees on the SAT used the Standardization approach to assess differential item functioning. Results of studies conducted on Asian-Americans, Hispanics, and Blacks on the Scholastic Aptitude Test (SAT) are described and then synthesized across studies. They limited the groups to include examinees who speak English as their best language. The results revealed that very few items across forms and ethnic groups exhibited large DIF. Poncheri, Meade, and Surface (2007) studied the DIF between Native and Non-native English speakers on the International Personality Item pool. Their study examined DIF using the Five Factor Model (FFM) of personality, across native and non-native English speaking workers living within the United States. Results indicated that many items of the FFM scale did not exhibit measurement invariance across native and non-native English speakers.

Abedi, Leon, and Kao (2007) examined the DIF in reading assessments for students with disabilities. They specifically looked at group differences between students with disabilities and students without disabilities using DIF analyses in a high-stakes reading assessment. A multi-step logistic regression procedure was employed in the study. The results revealed that for grade 9, many items exhibited DIF. However, several limitations were reported about the data. There was no access to information about the testing accommodations that students with disabilities might have received, and no information about the type of disabilities. Elousa and Jauregui (2007) classified DIF sources that affects the adaptation of tests. Their basis for classification is on linguistic and cultural criteria. There are four general DIF sources: (1) cultural relevance, (2) translation problems, (3) morph syntactical differences, and (4) semantic differences. It was found that the influence of the sources on the adaptation of tests is greater among those languages belonging to different linguistic families as the cultural distance among group increases. The DIF analysis is made through Mantel-Haenzel procedure. It was further found that the level of agreement between the two procedures, judgmental and statistical, stands at 69.8%.

There are several approaches to calculate for the DIF of the test develop. The proponent is choosing

standardization approaches in determining DIF because of its simplicity in arriving at values necessary for calculation. Other DIF methodologies require more sophisticated software because of the complexity of the calculation process.

2. Methodology

This study used the descriptive method of research. It dealt with the development of a test on problem solving, validated the test and determined the test's DIF through standardization approaches. The study involved four phases; Phase 1 is the construction and validation of the problem solving test; Phase 2 is the Item analysis phase; Phase 3 is the DIF determination; and Phase 4 is the final editing.

2.1 Phases of the Study

Phase 1: Construction and validation of the problem solving test for pre-service mathematics teachers

The aim was to develop an instrument on problem solving in Mathematics for pre-service teachers. The test constructed was designed to compose items/problems that will assess five sub-skills of problem solving. The five sub-skills are: (1) Comprehension, (2) Analysis and Organization, (3) Identifying Processes to be used, (4) Finding an Answer, and (5) Verification. The preparation of the test started with the preparation of table of specifications (TOS) that will ensure equal distribution of the items on the five sub skills initially identified. A pool of 75 items were constructed and subjected to content analysis, validation and review by experts. Three experts with a minimum qualification of a doctoral degree in Mathematics were considered in the validation process.

Phase 2: Item Analysis

After the content and face validation of the instrument by the three experts, the instrument was subjected to Item Analysis to decide on items requiring revision, rejection or acceptance. The items that remained after the item analysis was further subjected to final editing to determine flaws in grammar, item format, level of readability and appropriateness and accuracy of illustrations used. Considering the suggestions of content experts, a final copy was made and reproduced.

Phase 3: DIF determination

The problem solving test that remained after the item analysis was subjected to field testing to determine its DIF. The standardization approach was used to determine the DIF of the items. The calculated DIF served as a means to decide on the further review that will be made on the items that remained.

Phase 4: Final Editing

The qualified items identified were subjected to final editing by content experts to look at other flaws in the content, format, readability, appropriateness and accuracy of the illustrations used in the study.

2.2 Samples of the Study

The researcher identified two universities to be part of the study. The presence of resource persons in both universities made the entire processes of the study to flow efficiently. The samples consisted of one hundred (100) third year students taking Bachelor of Science in Secondary Education specializing in Mathematics during the school year 2013-2014 in the two universities. The universities considered in the study are identified as University A and University B. The respondents consisted of forty-eight (48) students from University A and fifty-two (52) students from University B. The two universities are also identified as reference group and focus group respectively. Both Universities are state universities offering teacher education programs.

3. Results and Discussion

3.1 Construction and Validation of the Problem Solving Test

The preparation of the test took almost a month to complete. The succeeding sections will help show the details of the results. Out of the Seventy-five (75) items constructed, only sixty items made it through the validation process and item analysis. Fifteen items were removed from the test The following table reveals the result of the item analysis done on the sixty-item test:

Table 1

Distribution of difficulty level per Sub-Skill

Item No.	Sub-Skill	Easy	Average	Hard	Total
1 – 12	Comprehension	5	6	1	12
13 – 24	Analysis & Organization	3	9	0	12
25– 36	Process	3	7	2	12
37 – 48	Finding an Answer	2	9	1	12
49 – 60	Verification	0	10	2	12
	TOTAL	13	41	6	60

Table 1 shows the distribution of the difficulty level of items per sub-skill. It will be noticed that majority of the items, 41 out of 60 are concentrated on the average level, a few, 6 out of 60 or 10% is hard and 13 out of 60 are easy items. This is a good indication on the level of difficulty of the test developed. The results of the test will also serve as the norm for future reference.

Table 2

Performance of Samples per Sub-skill

Sub skill	Mean	
	Reference Group N = 48	Focal Group N = 52
Comprehension (12)	9.27	7.21
Analysis & Organization (12)	5.17	4.90
Process (12)	3.90	4.38
Finding an Answer (12)	2.56	2.94
Verification (12)	2.83	3.58

There are 12 items on each sub-skill. Using the mean of the reference group and the focal group in each sub-skill, it could be seen from Table 2 that both the reference group and the focal group performed best in comprehension. It is followed by the sub-skill, analysis and organization. However, students were found to perform least in verification for the reference group while for the focal group, they performed least in Finding an answer, followed closely by verification. It may be interpreted that participants are good in understanding the problem but are weak in finding an answer. The pre-service teachers may know what the problem is about but are not capable of determining the correct answer and verifying if the answers that they got are correct. These skills are considered to be of higher order. This may imply that if pre-service teachers lack the ability to verify the correctness of their response, they lack the confidence in the answer that they got.

From Table 3, it could be seen that the test in every sub-skill has an average difficulty and the test as a whole also has an average difficulty. Good tests are neither very difficult nor very easy. This may mean that the problem solving test is showing one good characteristic of tests. The good characteristic is associated with the average level of difficulty of the items in the sub-skills and the problem solving test as a whole.

Table 3

Difficulty Index of the Test by Sub-skill

Sub skill	Total Score		Difficulty Index	Description
	Upper 27%	Lower 27%		
Comprehension	262	173	0.67	Average
Analysis & Organization	236	159	0.61	Average
Process	219	167	0.60	Average
Finding an Answer	178	151	0.51	Average
Verification	199	94	0.45	Average
WHOLE TEST	217.8	148.8	0.57	Average

3.2 DIF Detection

The second part of the data analysis is DIF Detection. The study made use of the Standardization Methodology. In general sense, an item is said to have DIF if the probability of correctly answering an item is lower in one group as compared to another group with similar abilities. The probability that an examinee with a given score will be able to correctly answer a particular item may be estimated. In order to make an accurate estimate, the observed proportion correct among those with the same score may be used.

The developed Problem solving test has sixty (60) validated items. Based on the result of the test, thirty-four (34) pairs of scores qualified for close examination. These pairs of scores are those that are similar for the two groups University A, which is designated as the base or the reference group and University B which is the focal group. The standardization approach to determining DIF got its name from the standardization group. This standardization group supplies a weighting function. Each of the members of the identified pairs will have a specific weight before accumulating the weighted differences across score values to arrive at a calculated index.

In order to calculate the index, the standardized p-difference was used. It is defined using the following equation:

$$\text{DSTD} = \sum_{s=1}^{S} K_s \left[P_{fs} - P_{bs}\right] / \sum_{s=1}^{S} K_s$$

Where:

$K_s / \sum_{s=1}^{S} K_s$ is the weighting factor at a given score value supplied by the standardization group to weight differences in performance between the focal group and the base group.

DSTD uses common weights or standards for both P_{fs} and P_{bs}. This is the very essence of the idea of standardization. This standardization factor is $K_s / \sum_{s=1}^{S} K_s$. For the purpose of the study, the following DSTD was calculated:

$$DSTD = \sum_{s=1}^{S} K_s \left[P_{fs} - P_{bs}\right] / \sum_{s=1}^{S} K_s = (-8.7\text{x}10^{-15})/40$$

$$DSTD = -2.17\text{x}10^{-16}$$

In interpreting this DSTD, we consider cutoffs. The cutoffs will set the boundary between an item requing close examination and items not requiring close consideration. There are two cutoffs for DSTD values that can be considered. One is $|DSTD| \geq 0.05$. When this happens, It can be interpreted that a significant number of items have to be reviewed. In most cases, the items considered for review will still be found to be acceptable. The second cutoff is $|DSTD| \geq 0.10$ which will identify relatively few items. The items with falling to this cutoff will require closer examination. For the purpose of the study both cutoffs were used.

The computed DSTD = $-2.17\text{x}10^{-16}$ is much less than the 0.05 critical value, meaning that the test is found to have a very small DIF. This means that the constructed test has no significant DIF. Furhter, this means

that there is no need to closely examine the items for further validation. The very small DIF calculated in this study may be associated to the process undergone in constructing, selecting, revising and validating items before they are formally tested for DIF.

4. Conclusion

This study entitled, "Construction, Validation & DIF Determination of Test on Problem Solving for Pre-Service Mathematics Teachers" aimed to develop a test on problem solving. Part of the development of the test on problem solving is two significant tasks of validation and determination of DIF. Problem solving is an integral part of not only of mathematics pre-service teacher's macro skill but also of pre-service teachers in general. The pre-service teachers' ability to solve problems faced is a determining factor of his success in the teaching profession. A test is a means to measure student's problem solving skill. In this study, there are four phases undergone. These phases lead to the development of a sixty-item test to test the problem solving capability of mathematics pre-service teachers. The items were divided into five sub-skills. The five sub-skills are: Comprehension, Analysis and Organization, Identifying Processes to be used, Finding an Answer, Verification. Each sub-skill has twelve items each.

Another important phase in the study is the validation and DIF determination of the developed test. The test was administered to two universities, A and B. The results of the data analysis led to the following findings: (1) Both the focal group and the reference group did best in comprehension, the reference group performed least in verification while the focal group performed least in finding an answer; (2) The mean of the scores (23.67) of all the subjects revealed a lower than half score (30 points); (3) The results of the item analysis revealed the following: 13 easy items, 41 average items and 6 difficult items; this means that only 10% of the items will be considered for rejection. The other part of the validation involves DIF determination. The main purpose of the DIF determination is for the construction of a fair test. The DIF method used in this study is the standardization approach. With this approach, an index (DSTD) is to be calculated. The results of the DIF analysis through the standardization approach led to a DSTD = $-2.17x10^{-16}$ which is much less than the 0.05 or the 0.10 critical values. This means that the constructed test has DIF that is not significant.

5. References

Abedi, J., & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education, 14*(3), 219–234. http://dx.doi.org/10.1207/S15324818AME1403_2

Abedi, J., Leon, S., & Kao, J. (2007). *Examining differential item functioning in reading assessments for students with disabilities*. Minneapolis, MN: University of Minnesota, Partnership for accessible reading assessment.

Elousa, P., & Jauregui, A. L. (2007). Potential of sources of differential item functioning in the adaptation of tests. *International Journal of Testing, 7*(1), 39-52. http://dx.doi.org/10.1080/15305050709336857

Handal, B. (2003). Teacher's mathematical beliefs: A review. *The Mathematics Educator, 13*(2), 47-57.

Khairavi, A. Z., & Nordin, M. S. (2011). The development and construct validation of the mathematics proficiency test for 14-year-old students. *Asia Pacific Journal of Educators and Education, 26*(1), 33-50.

Neil, W. A. (2002). *Issues in constructing formative tests in mathematics.* Paper presented at the International Association for Educational Assessment Conference, Hong Kong.

Poncheri, R. M., Meade, A. W., & Surface, E. A. (2007). *Differential item functioning and personality; Comparing native and non-native speakers*. Paper presented at the 22nd Annual Conference of the Society for Industrial and Organizational Psychology, New York, New York.

Schmitt, A. P., & Dorans, N. J. (1988). *Differential item functioning for minority examinees on the SAT.* Educational Testing Services. Princeton, New Jersey.

Zumbo, B. D. (2007). Three generations of dif analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly, 42*(2), 223-233. http://dx.doi.org/10.1080/15434300701375832