The Development of Instruments to Identify Basic Mathematics Understanding of New Students of FMIPA Syiah Kuala University

Rini Oktavia¹, Ridha Ferdhiana², Risky Ayunanda¹, Nurmaulidar¹, Intan Syahrini¹, Hafnani¹, Tarmizi Usman¹, Taufiq Iskandar¹

¹Department of Mathematics, FMIPA, Syiah Kuala University, Banda Aceh, 23111, Indonesia
²Department of Statistics, FMIPA, Syiah Kuala University, Banda Aceh, 23111, Indonesia

rini_oktavia@unsyiah.ac.id

Abstract: This study aims to identify basic mathematics skills and understanding of new students of FMIPA Unsyiah. Prior to the beginning of August 2015, the research team developed an instrument to identify the basic skills of new students of FMIPA Unsyiah. The questions included in the instrument are developed based on some anecdotal or empirical input about students' misconceptions in some Basic Mathematics concepts. In this article, we will report the process of developing the instrument as well as the results of psychometric analysis performed to measure the quality of the developed instruments. Classical test theory analysis results reported in this article include the values of the point bi-serial indices and the difficulty indices obtained based on the student responses on the instrument. The data was obtained from the first administration participated by 325 new students. The result of item analysis shows that the point bi-serial indices of items in this instrument ranges from 0.269 up to 0.582 with difficulty indices ranging from 0.16 up to 0.92. The results indicate that items in this instrument can discriminate students' ability well and can measure students' ability from every level of ability, but revisions are still needed to improve it.

1. Introduction

This study aims to identify basic mathematics understanding of incoming students in Faculty of Mathematics and Natural Science of Syiah Kuala University (FMIPA Unsyiah). Prior to the beginning of 2015/2016 academic year, a team of lecturers in the Department of Mathematics FMIPA Unsyiah developed an instrument to measure students' understanding of several important concepts in basic mathematics. This effort was initiated to provide empirical evidence on incoming students' understanding of such concepts.

If we look at a broader perspective regarding Indonesian students' achievement in mathematics, survey results from the Program for International Assessment (PISA) organized by the Organization for Economic Co-operation and Development (OECD) showed that the achievement of Indonesian students in mathematics is still low. Based on the results of the survey aimed at measuring the ability of 15-year-old students in the field of literacy, mathematics, and science, the average score of Indonesian students' mathematics tests in PISA 2015 is 386 which is far below the average score of all
OECD countries and economies, which is 490 [1]. Although in its countries and economies participating in PISA 2015, this fast growth is far from satisfactory. The achievement of Indonesian students in mathematics is still below its counterparts in South East Asia. The average of Vietnam students in PISA 2015 was 495, meanwhile the average score of Thailand students was 415.

The comparison obtained from PISA 2015 also matches the other international assessment result, namely the Trends in International Mathematics and Science Study (TIMMS) 2015 where the average grade of Indonesian 4th graders is only 397, under 400 which becomes the low value limit in TIMMS as reported by Mullis, IVS, Martin, MO, Foy, P., and Hooper, M. in the 2015 TIMMS results report published in 2016 [2]. If PISA measures the mathematical ability of 15-year-old students in mathematics problems found in everyday life, TIMMS measures the ability of grade 4 and grade 8 students in mathematics problems listed in the school curriculum. By 2015, 8th graders from Indonesia do not participate in TIMMS, but if you look at TIMMS 2007 and 2011 results, the average grade of Indonesian Grade 8 students in Mathematics ranges from 376-388, still below the lower score limits set by TIMMS which is 400. The achievement of Indonesian students in mathematics also decreased significantly from 2007 and 2011, both in the cognitive domain and in mathematical content tested as reported in TIMSS 2011 International Results in Mathematics [3].

Therefore, it is reasonable to expect that the poor preparation in mathematics of incoming students of FMIPA Unsyiah reflects the poor performance of Indonesian students on TIMMS and PISA. We would like to investigate whether the performance of new students of FMIPA Unsyiah align with Indonesian students’ performances in international assessments. In order to provide an empirical evidence, a diagnostic test should be conducted to identify the students’ understanding of basic mathematics. In this paper, the process of developing the instrument and the results of several psychometric analyses from the students’ responses will be discussed.

2. Literature Review
In developing the instrument, a thorough discussion regarding diagnostic assessment in education were commenced. In education, diagnosis is defined differently and is frequently viewed from different perspectives and conducted based on the perspectives. Rupp, Templin, and Hanson in [4] explained that a diagnosis is basically an action to analyze a problem with precision and to identify the cause of the problem in order to make a decision. This agrees with Ketterlin-Geller and Yovanoff [5] who define diagnosis from a clinical perspective as a clinical definition in which results of an assessment are used to determine whether a specific condition exists or not. Ketterlin-Geller and Yovanoff [5] also give an alternative definition of diagnosis in education as definition in education which results of an assessment provide information about whether students have certain knowledge and skills on the content domain as well as preconceptions or misconceptions about the material. In addition to suggesting that diagnostic assessment is needed when there is a problem that has been previously identified; this definition also suggests that a diagnostic assessment is needed for teachers to prepare for appropriate instructional strategies that are based on students’ needs.

Cognitive diagnostic assessment (CDA) is a relatively new diagnostic assessment in psychometrics that aims to provide formative detailed diagnostic detailed feedback that inform what examinees’ know and are able to do [6, 7, 8]. DiBello, Roussos, and Stout in [6], also recognize two types of practices in constructing a diagnostic assessment: (1) designing a test by using an existing assessment and applying more appropriate data analysis to gather richer information to identify the problems faced by students in learning and (2) designing a test from the beginning for a diagnostic purpose. In designing an effective diagnostic assessment, the primary goal is to satisfy the assessment purpose, so the assessment purpose should be carefully described.

Since assessments are tools to examine learners’ cognitive capabilities by gathering evidence from a measurement instrument administered to examinees, consisting of several tasks that need to be executed. Therefore, choosing the tasks should be based on the type and amount of evidence needed to support desired inferences about examinees’ attributes. Test developers should consider a wide variety of possible tasks, choosing feasible ones that best match the purpose of the assessment; task
developers should not avoid tasks that require combinations of multiple skills per task [6]. In cases where the assessment instrument already exists, the problem becomes how much information about the set of desired attributes the test can provide. In the case when new tasks are explicitly designed for a new diagnostic assessment, DiBello, Roussos, and Stout in [6] suggest that in developing an assessment, test developers should provide an analysis of the tasks that explain the skills or attributes that are involved in the assessment, the difficulty level of the instrument, and the form of interactions that are needed to gather the data from the instrument.

In this study, we will design a test instrument from the beginning for several concepts in basic mathematics that include material in arithmetic, algebra, and geometry. In the following section we will discuss the process of developing the instrument that follows the suggestion from DiBello, Roussos, and Stout [6] mentioned above. In the following sections, we will also discuss the results that we gain from the administration of the instrument based on classical test theory (CTT) analysis and the conclusions that were inferred from the results.

3. Methodology
In order to develop a diagnostic assessment to identify students’ understanding of important basic mathematics concepts, the research team collected information regarding students’ common misconceptions in basic mathematics. The misconceptions which are either scientifically proven or anecdotal based on instructors’ experiences in teaching basic mathematics were recorded and the matrix of cognitive domains and mathematical concepts that are targeted to be measured were developed. The research team also referred to the national curriculum of middle and high school mathematics in determining the contents that will be included in the instrument. After the cognitive domains and the contents were determined then the research team develop the items.

To ensure the content validity and the face validity of the instrument, we requested all faculty members of the Department of Mathematics of Syiah Kuala University to evaluate and provide critiques and suggestions to improve the instrument. The final version of the instrument includes 15 items consisting of five arithmetic items, five algebra items, and five geometry items. An example of the items that is translated to English from Bahasa Indonesia is given in Fig. 1.

![Figure 1. An example of items in the Instrument](image)

The participants of this study were 325 new students of FMIPA Unsyyiah enrolling in 2015/2016 academic year. The survey was administered in August 2015. There are 15 multiple choice basic mathematics items and one short answer item, and several demographic questions to gain students’ characteristics included in the questionnaire. The administration took place in the matriculation session in the beginning of the semester. To measure the reliability of students’ responses on the instrument, another identical administration were conducted two weeks after the first administration that involved 85 new students from the same population. The second administration was conducted to investigate the test-retest reliability of the students’ responses. Analysis reliability and validity of our instrument will be discussed in a separate article.

In recording the students’ responses, a correct response is scored 1 and a false response is scored 0, that gives 0 is the possible lowest total score and 15 is the possible maximum total score. The CTT analyses conducted include the calculations of difficulty and point bi-serial indices of each item as well as the validity and reliability indices of the students’ responses. The choice of using CTT
analyses was based on the fact that our data have three content constructs or dimensions that are arithmetic, algebra, and geometry. This multi dimensions characteristic make it inappropriate to use other item analysis such as Item Response Theory (IRT) or Rasch measurement.

CTT has been the foundation for measurement theory for more than eight decades. In CTT, it is assumed that the raw score \( X \) obtained by an examinee by adding the true score \( T \) of the examinee and a random error \( E \) or \( X = T + E \). The true score of a person \( T \) is a hypothetical score that is assumed as the mean total score of an examinees if the examinee takes the same test again and again for infinitely many times. CTT concentrates on two main statistics: item difficulty and item discrimination. Item difficulty is calculated by taking the ratio of the number of correct responses with the number of participants responded to the item [9]. The higher the item difficulty index, the easier the item is. The desired average of item difficulty indices in an instrument is 0.5.

Item discrimination index describes the capacity of the item to discriminate students’ ability. The higher the discrimination index of an item, the more powerful the item in discriminating students’ abilities. In this study we the point bi-serial index as the discrimination index of an item. Point bi-serial index of an item is actually the value of Pearson Correlation between students’ responses on the item with their total scores [10]. According to Penn [11] dan McGahee & Ball [12] the value of revpoint biserial index in interval 0.2-0.29 is moderate, in interval 0.3-0.39 is considered good, and in interval 0.4-0.7 is considered very good. An item with point bi-serial index that is less than 0.2 is considered bad item and need to be revised or excluded from the instrument. The formula to calculate the point bi-serial correlation of an item is given below:

\[
r_{\text{bis}} = \frac{\bar{Y}_1 - \bar{Y}_0}{\sigma_Y} \sqrt{p \times q}
\]

where:
- \( r_{\text{bis}} \) = point bi-serial correlation
- \( \bar{Y}_1 \) = the average of total scores of students who responded correctly to the item
- \( \bar{Y}_0 \) = the average of total scores of students who responded falsely to the item
- \( \sigma_Y \) = standard deviation of the total scores
- \( p \) = proportion of correct responses
- \( q \) = proportion of false responses  [13]

In the following sections the results of the CTT analyses are provided. The values of the difficulty index and point bi-serial correlation of each item will be analyzed. Then the conclusions about the quality of the instrument will be discussed.

4. Results and Analysis

From the distribution of the total scores, we found that the average total score is 7.6 with standard deviation of 2.973. It was also found that a little bit more than half of the students (50.2 %) scored between 0-7. This means less than half of the students are not capable on responding correctly 8 or more items in the instrument. This finding is quite disturbing since the items in the instruments are measured students’ understanding of important basic mathematics concepts that they should have mastered before entering college. This finding informed the instructors of basic mathematics courses about materials that need to be addressed more thoroughly in class. Due to the limited space in this article, more qualitative analyses of the mastery that students have in each item will be reported separately.

An item is considered to have medium difficulty is its difficulty index falls in the interval of 0.3 - 0.8 [14]. An item with difficulty index higher than 0.8 is considered an easy item, meanwhile an item with difficulty index less than 0.3 is considered a hard item. It was found that the average difficulty indices of the items in the instrument is 0.51 inferring that the instrument has a desired overall difficulty. Only three items fall in the easy category (item 2, 3, and 15). Meanwhile, 5 items are categorized as hard (item 6, 7, 11, 12, and 13). The hard items measure students’ understanding on graphic of functions, exponential and rational expressions, line equations, and factoring polynomials.
When we observed the point biserial index values of each item, we found that only one item that falls in the moderate category, that is item 11. Considering the difficulty index of this item is also very small, we can argue that the small point bi-serial index might be caused by the guessing factor. The small correlation between the students’ responses on item 11 and their total scores might indicate that students who have higher total scores did not answer correctly because their guesses were false, meanwhile students who have low total scores might respond to this item correctly due to the correct guessing. Overall, the point bi-serial index of each item in the instrument can be considered very good. Only item 14 that can be categorized as a good item in discriminating students’ ability in basic mathematics. All point bi-serial correlations are statistically significant with level of significances that are below 0.001. The values of the difficulty indices and point bi-serial indices of the items in our instrument are given in Table 1 below.

### TABLE 1. point bi-serial and difficulty Index

| Item | Point Bi-serial Index | Difficulty Index | Item Category |
|------|-----------------------|------------------|---------------|
| 1    | .467**                | 0.61             | Medium        |
| 2    | .422**                | 0.92             | Easy          |
| 3    | .429**                | 0.90             | Easy          |
| 4    | .529**                | 0.71             | Medium        |
| 5    | .501**                | 0.66             | Medium        |
| 6    | .485**                | 0.24             | Hard          |
| 7    | .582**                | 0.26             | Hard          |
| 8    | .537**                | 0.71             | Medium        |
| 9    | .529**                | 0.31             | Medium        |
| 10   | .530**                | 0.47             | Medium        |
| 11   | .269**                | 0.19             | Hard          |
| 12   | .463**                | 0.26             | Hard          |
| 13   | .499**                | 0.16             | Hard          |
| 14   | .313**                | 0.38             | Medium        |
| 15   | .452**                | 0.86             | Easy          |

5. Conclusions
More than half of students participated in this study did not have solid background in basic mathematics that is needed for taking quantitative-based courses in the university. The items in the instrument have the power to measure students’ understanding of basic mathematics from every level of difficulties. In predicting basic mathematics understanding of incoming students of FMIPA Unsyiah in 2015, the instrument to measure students’ understanding of basic mathematics was found to have overall moderate difficulty and only two items that are categorized as moderate or good in discriminating students’ abilities in basic mathematics. Other items are very good in their discriminating powers. This indicates that that from the CTT perspectives, the developed instrument has a very good quality. There are several minor errors found in the representations of the items in the instrument. With a minor revision, this instrument could be used in the future diagnostic tasks to identify students’ level of understanding of basic mathematics.

REFERENCES
[1] OECD. 2016. Country Note – Results from PISA 2015, Indonesia. OECD 1-8.
[2] I. V. Mullis, M. O. Martin, P. Foy, P., M. Hooper. TIMSS 2015 International Results in Mathematics. International Association for the Evaluation of Educational Achievement IEA & Boston College
[3] I. V. Mullis, M. O. Martin, P. Foy, M. Hooper, A. Arora. TIMSS 2011 International Results in Mathematics. Chesnut Hill, MA, USA & Amsterdam, The Netherland: TIMSS & PIRLS
International Study Center, Lynch School of Education, Boston College & International Association for the Evaluation of Educational Achievement IEA.

A. Rupp., J. Templin, R. A. Henson. 2010 Diagnostic measurement: Theory, methods, and applications. New York, NY: Guilford Press.

I. R. Ketterlin-Geller, P. Yovanoff. 2009 Diagnostic assessments in mathematics to support instructional decision making. Practical Assessment, Research & Evaluation,

L. DiBello, L. A. Roussos, W. Stout. 2007 Review of cognitively diagnostic assessment and a summary of psychometric models. In C. V. Rao & S. Sinharay Eds., Handbook of Statistics, 26 Psychometrics, Amsterdam, The Netherlands: Elsevier.

S. E. Embretson. 1991 A multidimensional latent trait model for measuring learning and change. Psychometrika.

J. P. Leighton, M. J. Gierl. 2007 Defining and evaluating models of cognition used in educational measurement to make inferences about examinees’ thinking processes. Educational Measurement: Issues and Practice,

M. McAlphine. 2002 A Summary of Method of Item Analysis. CAA centre, England.

L. Crocker, J. Algina. 2008 Introduction to Classical and Modern Test Theory. Cengage, United States of America.

B. Penn. 2009 Test item development and analysis. Presented at Creighton University School of Nursing Faculty Retreat. NE, Omaha. W. McGahee, J. Ball. 2008 How to Read and Really Use an Item Analysis. US National library of Madicine National Institutes of Health:

T. J. Kline. 2005 Psychological Testing : A Practical Approach to Design and Evaluation. Sage, Canada.

R. J. Cohen, M. Swerdlik. 2009 Psychological Testing and Assessmen : An Introduction to Tests and Measurement 7th edition. McGraw-Hill, United State of America