# Development of Grade Six Mathematics Item Bank by Applying Item Response Theory

Aye Aye Aung[1#] & San Lin[2]

[1#] Department of Educational Psychology, Yangon University of Education, Myanmar

[2] Basic Education High School (B.E.H.S) (Branch) Kyaukuk, Minbya Township, Rakhine State, Myanmar

[#]corresponding author <ayeaung777@gmail.com>

**Received** first draft 16 June 2020. Received reports from first reviewer (1 July); second reviewer (17 July). Received revised draft 4 October. **Accepted** to publish 1 November 2020.

## Abstract

*The purpose of the study was to construct a Mathematics Item Bank for Grade Six students by using Item Response Theory (IRT). The participants were selected by using stratified random sampling technique. Descriptive research design and quantitative survey method were used in this study. Lesson 1 to 5 of Grade Six Mathematics Textbook (1) and Lesson 1 to 6 of Grade Six Mathematics Textbook (2) were selected as the content area for item bank according to the monthly plan. In the preliminary test, after item analysis with 2-PLM of IRT, out of 219 multiple-choice items, 170 items were remained for three field tests (Form A, Form B and Form C) and each form contained about 56 items. These field tests were administered to 1513 Grade Six students from the selected schools in Yangon Region in which 505 students were administered for Form A, 504 students for Form B and 504 students for Form C. Then, 164 items of field test (56 items in Form A, 55 items in Form B and 55 items in Form C) had the discrimination range from 0 to +2 and the difficulty range from −3 to +3. Therefore, all items were stored in item bank. To support advantage of item bank, a new sample test was developed by using the items from item bank. It was found that this test was more appropriate for examinees whose ability (θ) range is −1.65 to +1.95 and provided the highest information for the examinees with ability level −0.05. In conclusion, it was found that the items that were biased against different groups could not be determined Hence suggestion was made that Differential Item Functioning (DIF) should be calculated in the future study for checking of items that are biased against racial or ethnic minorities.*

***Keywords:*** Item banking; Item response theory; Difficulty; Discrimination

## Introduction

### Background and Importance of the Study

A test is an evaluation device or procedure for a sample of an examinee's behavior in a specified domain (Weirsma & Jurs, 1990). Tests are used for many purposes in education: promoting students from one grade to the next, awarding the degrees or diplomas, evaluating the quality of education, and identifying the workers in the need of training. Moreover, test results are important devices to share information with the boards of education, parents, and the general public through the media.

Furthermore, it is very important to evaluate the students' achievements, comprehension levels and proficiency in the subject. Therefore, high-quality assessment systems must be required. To get the reliable evaluation or result, the tests must be constructed fairly and accurately. In other words, tests should have objective, unbiased items because each student's ability can be studied by scoring these items.

However, there are many steps in the process of constructing a good quality test. Whenever a test is constructed, the development of test specifications, item writing, pilot and field testing, as well as calibrating the items should be done first. Then, the good items are selected for the final test. Therefore, creating a good test takes a lot of time, effort and money. In order to avoid the repetition of the same process every time when a test is administered, the concept of item banking is becoming more and more popular.

Moreover, challenges about the testing, especially in basic education, are occurred. In practice, most teachers are making items in the tests easily without accountability. These items cannot show the real ability. The cause of these problems is the lack of use of systematic item bank. To overcome these challenges, the culture of widely used item bank is needed because utilizing of item bank makes the testing reliable and systematic.

The item banking plays an important role in constructing the tests. In item banks, many reliable test items are stored according to the content areas, age or grade levels, item characteristics and so on. Generally, there are two test theories which are used to develop the item bank. They are Classical Test Theory (CTT) and Item Response Theory (IRT). Because of the limitations of classical test theory, in this study, IRT measurement system is used.

Nowadays, item banking is important in education in addition to being a big business in the 21st century. Many countries conducted researches about the item banks, as well as computerized item banks and item banks are used for a variety of purposes. On the other hand, the items from the item bank are used in their testing for the quality educational assessments. For these benefits, therefore, item bank was developed for Grade Six Mathematics by applying Item Response Theory (IRT) in this study. Besides, it is expected that this study will provide some information about the concept and the importance of item banking in the role of educational testing.

## Purpose of the Study

The main purpose of the study is to develop an item bank for Grade Six Mathematics by applying the Item Response Theory (IRT). The next purpose is to develop a new sample test by using the items from the item bank.

## Research Questions

The following are Research Questions (RQs) formulated as guide of this study:

(1) What are the elements of the item bank for Grade Six Mathematics like?
(2) How the development of Grade Six Mathematics test could be implemented based on item bank prepared?

## Review of Related Literature

### Role of Item Banking in Education

Item banks are defined as the files of various suitable test items that are coded by subject area, instructional objective measured including various pertinent item characteristics like item difficulty and item discriminating power (Gronlund, 1998, as cited in Rudner & Lawrence, 1998). During the mid-1960s in England, the term "item bank" was used to describe the group of test items that were organized, classified and catalogued like books in a library (Choppin, 1985). Moreover, item banks are called by such terms as "question banks," "item pools," "items collection," "item reservoirs," and "test items libraries" (Millman & Arter, 1984).

The item banking plays the important role in the test construction. New tests or subtests can be developed, without piloting and evaluating to a large set of items, by drawing items from the item bank and then their characteristics can be predicted. Therefore, item banking provides substantial savings of time and energy over conventional test development (Rudner & Lawrence, 1998).

Besides, some advantages to item banking include flexibility, security and consistency (Umar, 1999). A test developer can construct different tests with predictable characteristics by using the items from an item bank and can compare the performance of examinees who sat for different tests on the same scale. Rudner and Lawrence (1998) indicated that items in an item bank can be edited, withdraw and populated with new items when needed. As a critical component of any high-quality assessment, item banking is the foundation for the development of valid, reliable content and defensible test forms (Millman & Arter, 1984). Therefore, item banking is really playing important role in educational assessment.

### How to Construct an Item Bank?

Although item bank has an enormous potential to ease and improve test construction process, it demands skills and professional expertise. The following steps are involved in constructing an item bank (Tshering, 2006):

(1) The goals and objectives of item bank have to be identified.
(2) Appropriate people will have to be identified for developing items and performing item content matching.
(3) The items have to be field-tested through different tests across wide range of abilities.
(4) The items from different tests have to be calibrated on a common scale by using suitable Item Response Theory (IRT) models.
(5) Item bank data base has to be developed.
(6) Item bank has to be replenished with new items.

But only the procedure of step 1 to 5 can be carried out for constructing item bank since all the above steps cannot be applied in this study.

### Item Response Theory

IRT is a test theory that expresses the relationship between observable test performance (responses) and unobservable traits underlying the test performance (Tshering, 2006). In IRT, the construct or unobservable mental attribute measured by the items may be an academic proficiency or aptitude, or it may be an attitude or belief. Thus, the IRT score is often called an ability or person parameter, latent trait, or proficiency, referred to as "θ" that is an ability parameter. According to DeMars (2010), values for this ability parameter (θ) theoretically range from negative infinity (−∞) to positive infinity (+∞) but most examinees will have values for the ability parameter (θ) between −3 and +3 in practical.

Moreover, IRT item parameters include item difficulty, *b*-parameter (location), item discrimination, *a*-parameter (slope) as well as pseudo-guessing, *c*-parameter (asymptote) and they are estimated directly using logistic models instead of proportions. The item difficulty, *b*-parameter tells how difficult the item is. The item discrimination, *a*-parameter describes how well an item can differentiate between the examinees having abilities below the item location and those having abilities above the item location (Nu Nu Khaing, 2011, as cited in Htet Htet Lin, 2014). The item discrimination, *a*-parameter, is called the *slope*, tells how steeply the probability of correct response changes at the steepest point on the Item Characteristics Curves (ICC) as the proficiency or trait increases.

Next, the lower asymptote parameter or *c*-parameter, sometimes called guessing parameter or pseudo-chance-level parameter, provides the probability that an examinee with a very low level of θ will answer the item correctly only by guessing (DeMars, 2010).

There are three popular models of IRT;

(1) The one-parameter logistic model (1PLM) or, Rasch model, only uses item difficulty (*b*) as a parameter for calculating a person's ability.
(2) The two-parameter logistic model (2PLM) uses both item difficulty (*b*) and item discrimination (*a*) as parameters, and
(3) The three-parameter logistic model (3PLM) uses item difficulty (*b*), item discrimination (*a*) and the guessing parameter (*c*).

### How to Develop a New Test from the Item Bank?

In order to develop a new test, the test developer should determine the purpose of the test. After that a blueprint or a table of specifications should be drawn to outline this test. It is a two-fold table on which the learning outcomes are listed along one side of a table and the subject matter topics along the other. Then, the items are selected from the item bank according to the content, grade level and on the basis of their item parameters to meet the particular testing goals (Baker, 2001). Without pilot testing, the characteristics of this new test can be predicted (Rudner & Lawrence, 1998). With reasonable accuracy, how much skill an examinee should possess to answer this new test can also be predicted.

## Method

### Sample of the Study

The preliminary test was administered to 167 samples of Grade Six students form Basic Education High School (BEHS) (1) Thingankyun Township in Yangon Region. For field testing, sample of 1513 Grade Six students (male = 764 and female = 749) were selected by using stratified random sampling technique from eight schools of four districts in Yangon Region.

## Research Method

In this study, the participants were selected by using stratified random sampling method. Descriptive research design and quantitative survey method were used.

## Content Area for Item Banking

Firstly, Grade Six Mathematics Text Book and Teacher Manual of Grade Six Mathematics, together with the instructional objectives and aims of teaching Mathematics, were studied. There are nine chapters in Grade Six Mathematics Textbook (1) and ten chapters in Grade Six Mathematics Textbook (2). According to the monthly lesson plan, the items were selected from the content area of the chapters 1 to 5 of Grade Six Mathematics Textbook (1) and the chapters 1 to 6 of Grade Six Mathematic Textbook (2) because the students had learnt only those selected chapters during preliminary testing time.

## Procedure

**Administering the preliminary test.** The preliminary test was administered to a sample of 167 Grade Six students from No.1, Basic Education High School (BEHS), Thingankyun, Thingankyun Township in Yangon Region. Since the preliminary test had four parts of questions and each part involved about 50 items, students were administered with test items of two parts per day. If one part was administered in the morning, then another part was administered in the afternoon after break time. It took two days for the whole test that included four parts. The time duration of each part was 60 minutes.

Since the items were objective types, the scoring keys were 1 for correct answer and 0 for incorrect answer. Then, the data were analyzed by using Two-Parameter Logistic Model (2-PLM) of Item Response Theory (IRT) through BILOG-MG software. As the items were calibrated with Two-Parameter Logistic IRT Model, the characteristics of the items were described by the item discrimination (a) which ranges from 0 to +2, and item difficulty (*b*) which ranges within $-3$ and $+3$. The variability of *a*-values ranged from $+0.171$ to $+1.713$ and *b*-values ranged from –3.187 to +5.771. As 49 items were not in the range of difficulty $-3$ and $+3$, they were removed and so 170 items remained for field test. Since the test took time of the participants to respond, all items were not administered at the same time for field testing. Thus, three groups of items with the nearly same difficulty range and same content (but not identical) were left for field testing.

**Constructing field tests.** For field study, three forms were administered, i.e. Form A contained 58 items, Form B contained 56 items and Form C contained 56 items respectively. The three practical or field tests were administered to a sample of 1513 Grade Six students (505 participants for Form A, 504 participants for Form B, and 504 participants for Form C) from the selected schools in Yangon Region. The data obtained from the field testing were analyzed by using 2-PLM of IRT with the application of BILOG-MG software. Only one test was required to be taken by each participant. The time limitation of each test form was 60 minutes.

## Findings and Discussion

### Checking the Assumption of Unidimensionality

The assumption of unidimensionality means that only one trait or ability is measured by the items. It is also a common one for the test constructors since they usually desire to construct unidimensional tests to enhance the interpretability of the test scores (Kay Zune Aung & Nu

Nu Khaing, 2017). To assess the unidimensionality of the data, the scree plots of the eigenvalues of the inter-item correlation matrix for three forms were studied. The scree plots of the eigenvalues for all items of three forms were shown in Figure 1.

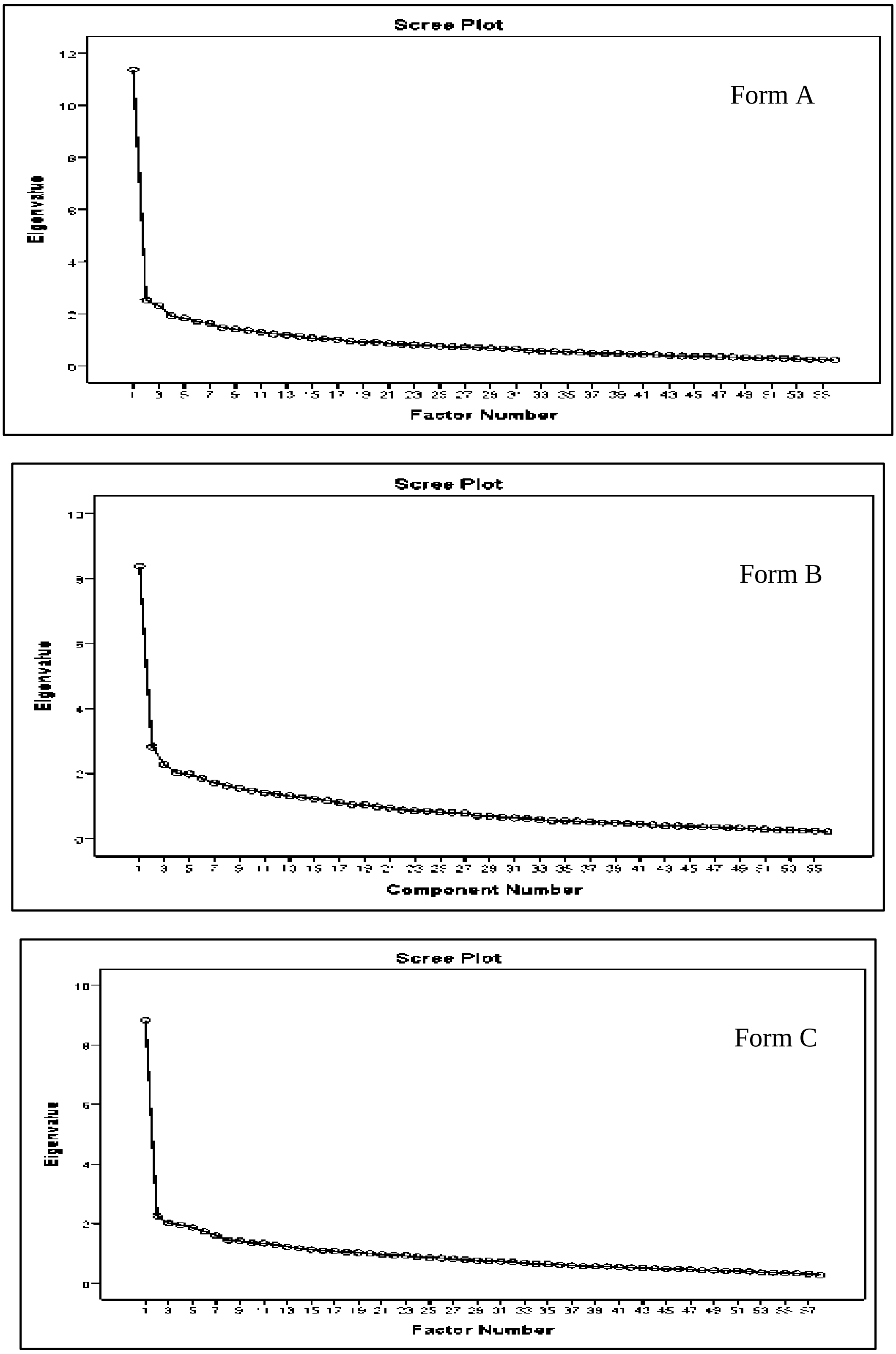


*Figure 1.* Scree Plots of the Eigenvalues for all Items of Form A, Form B and Form C.

As shown in the Figure 1,  the dominance of the first factor can be seen obviously. The largest eigenvalue of the correlation matrix for all items in Form A, Form B and Form C is about three times larger than the second largest eigenvalue. Therefore, it is sure that the three forms had the unidimensionality. It means that all items in the three forms were unidimensional and not dependence on each other (Hambleton, Swaminathan, & Rogers, 1991).

## Analyzing Item Parameters by IRT

After checking the assumption of IRT, the item parameters for all items were estimated with 2-PLM by applying BILOG-MG software. In this study, items with discrimination ($a$) values within 0 to +2 and those with difficulty ($b$) values within −3 to +3 were expected to be selected for item bank.

Figure 2 described the matrix plot of item characteristics curves (ICCs) for all items on Form A with 58 items. It was found that most items in Form A had discrimination ($a$) values within 0 to +2 and those with difficulty ($b$) values within −3 to +3. But item 11 and item 13 were very difficult items because their difficulty ($b$) values were greater than +3. In addition those items had discrimination less than 0.4 and thus they should not be included in the operational test (DeMars, 2010). Therefore, except these two items, the other items of Form A (56 items) were regarded to be kept in the item bank.

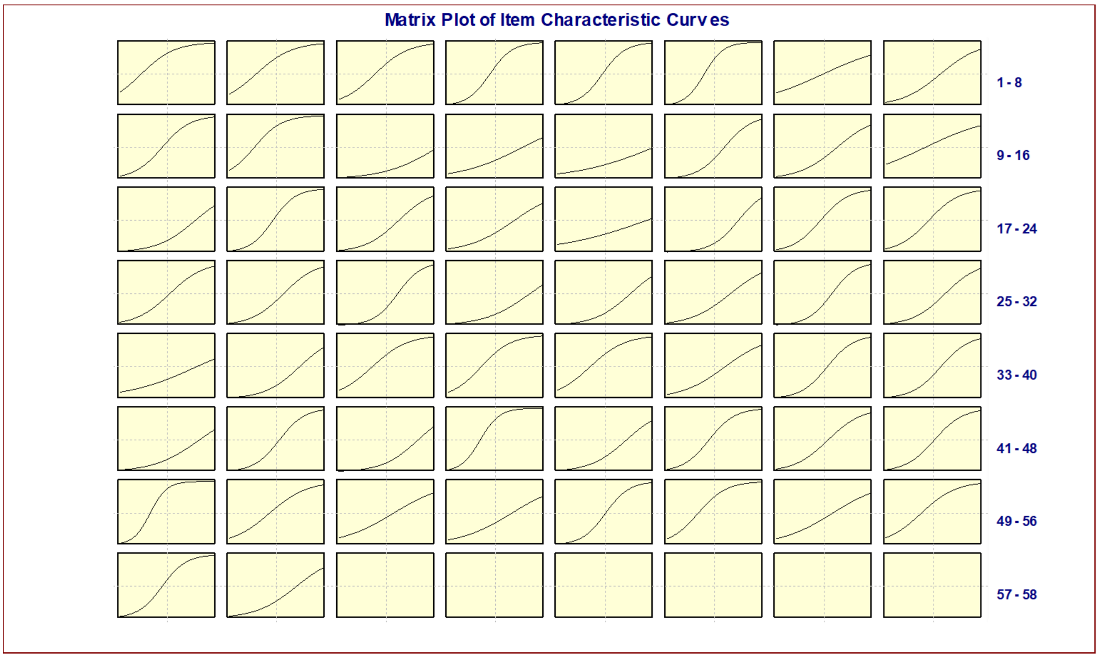


*Figure 2.* Matrix Plot of Item Characteristic Curves for Form A.

In Figure 3, it could be seen that only item 21 ($a = 0.303$, $b = 3.051$) had difficulty value greater than $+3$, and discrimination less than 0.4. Other items in Form B had the difficulty range between $-3$ and $+3$ and discrimination range between 0 and $+2$. Therefore, out of 56 items in Form B, item 21 was eliminated and the remaining items (55 items) of the Form B were regarded as the acceptable items to be kept in the item bank.

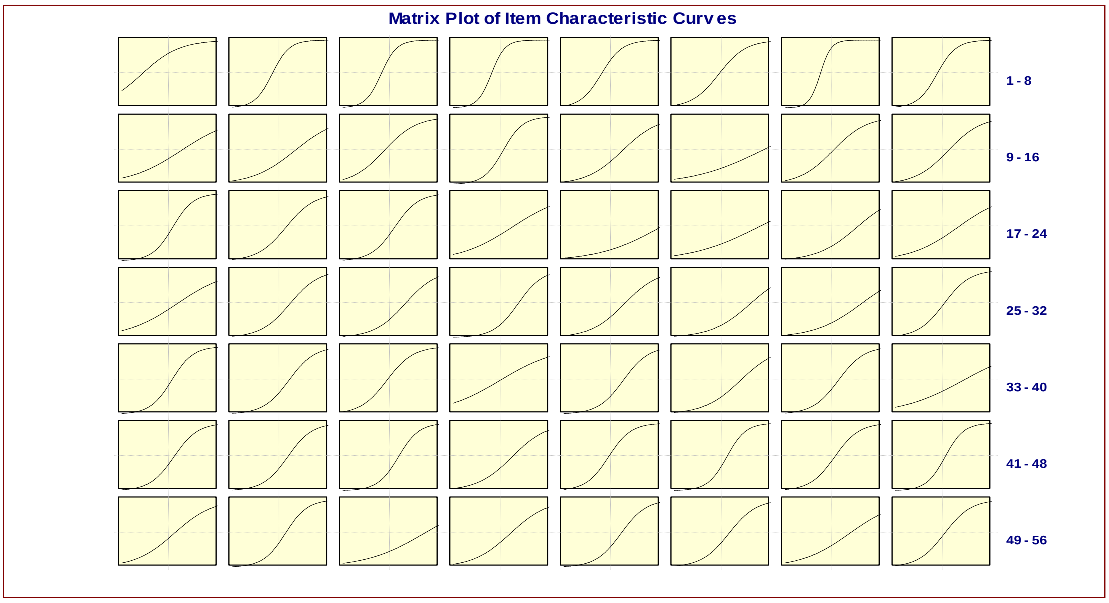


*Figure 3.* Matrix Plot of Item Characteristic Curves for Form B.

The matrix plot of item characteristic curves (ICCs) for all items in Form C with 56 items was shown in Figure 4. Although the difficulty range of most items in Form C had between −3 and +3, three items (item 34, item 50 and item 56) had the difficulty values (5.076 ~ 5.658) greater than +3, and they had discrimination (0.179 ~ 0.213) less than 0.4. Therefore, they were removed and the remaining items (53 items) of the Form C were assigned as the good items and were added to the item bank.

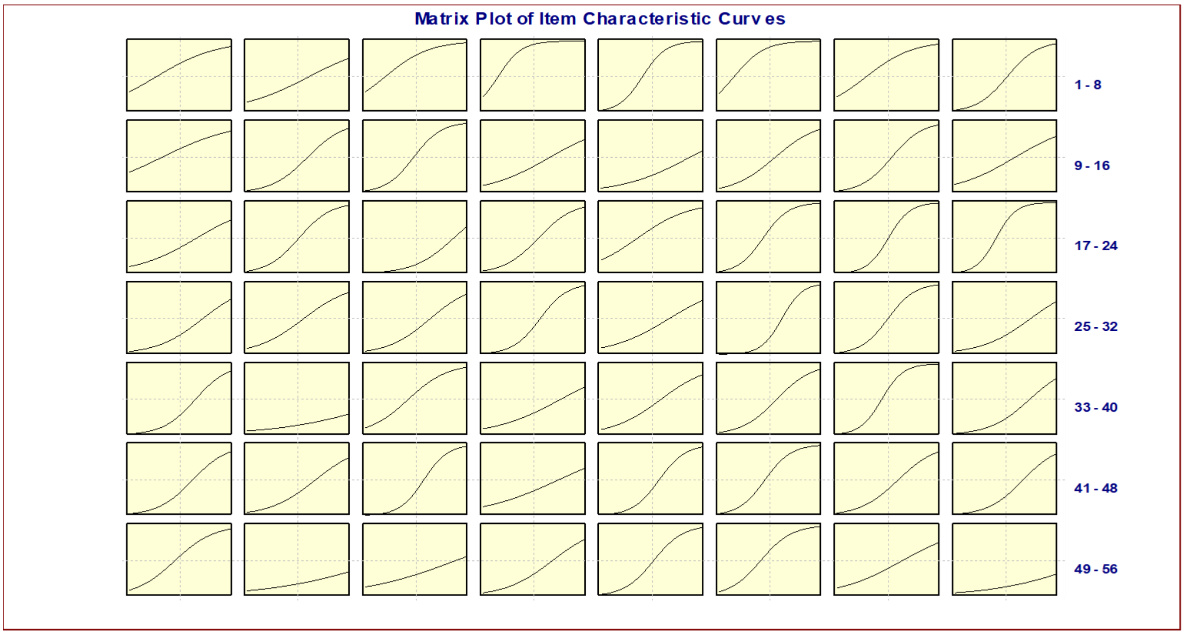
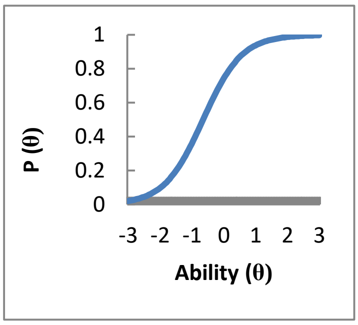
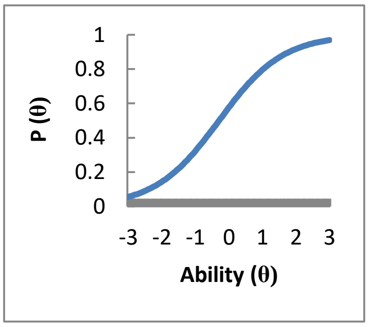


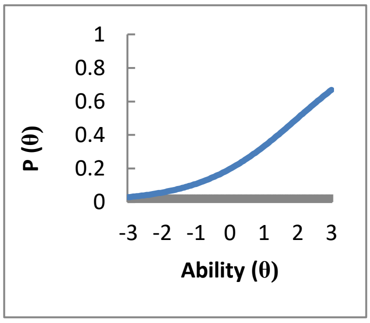
*Figure 4.* Matrix Plot of Item Characteristic Curves for Form C.

After analyzing the item parameters of the three forms, 164 items (56 items from Form A, 55 items from Form B, and 53 items from Form C) with their respective item codes, lessons, item parameters and item characteristic curves were installed in the item bank. Table 1 illustrated some sample items from the item bank with their respective item parameters.

Table 1

*Sample Items with Their Parameters*

| **No** | **Items** | **Learning Outcomes** | **Item Parameters** | | **Item Characteristic Curves (ICCs)** |
|---|---|---|---|---|---|
| | | | **a** | **b** | |
| **1** | 45 x [ 36 ÷ {8 - (16 ÷ 4)}] =<br>A. 105 B. 205<br>C. 305 D. 405<br>E. 505<br>**ANS: D** | Comprehension | 0.969 | −0.647 |  <br>Good discrimination |
| **2** | 9.8 + 10.035 =<br>A. 1.9835<br>B. 19.835<br>C. 198.35 | Comprehension | 0.617 | −0.292 |  |

| | | | | | |
|---|---|---|---|---|---|
| | D.0.19835<br>E. 1983.5<br>**ANS: B** | | | | Fair discrimination |
| **3** | If p = 3 and q = 2, then<br>(p + q) $^{2}$ =<br>A. 5 B. 25 C. 10<br>D. 9 E. 4<br>**ANS: B** | Comprehension | 0.413 | 2.001 |  <br>Poor discrimination |

## Test Information Functions of Three Field Tests

Next, the test information functions were calculated in order to know the maximum amount of information obtained from the tests. The steeper the slope (*a*-parameter) is, the greater the test information is (Baker, 2001). The test information function curve for three tests was also illustrated in Figure 5. Comparing the tests, it was found that Form C could give less test information than the other two forms because Form C had a little lower value of *a*-parameter than the other two forms. Besides, it can be also seen that Form B with the highest discrimination could provide more information.

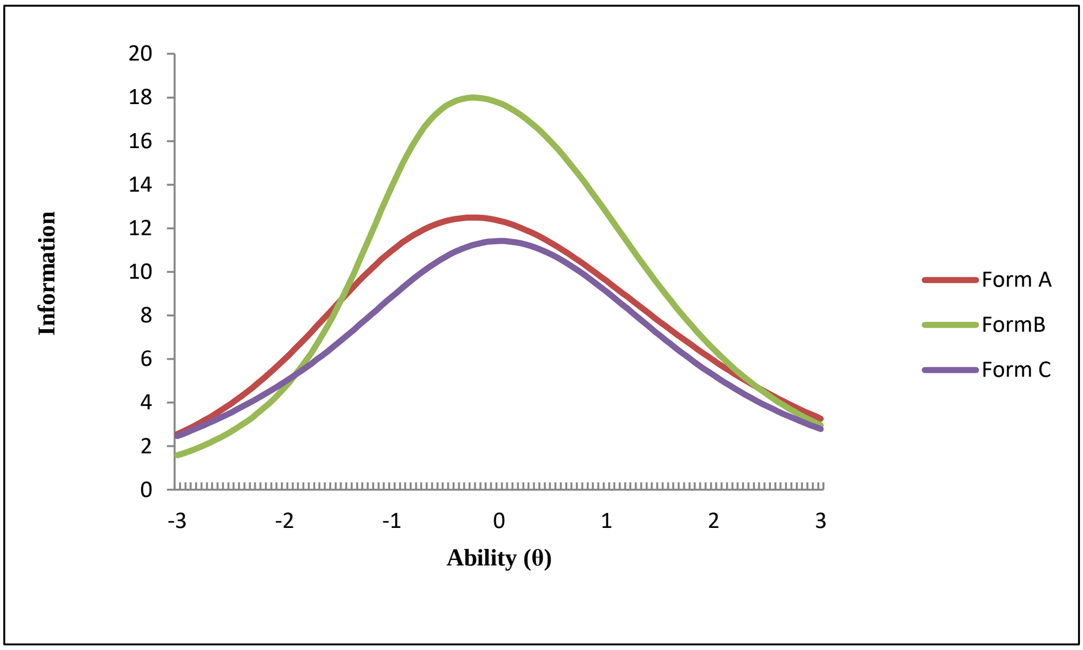


*Figure 5*. Test Information Curves for three tests.

## Developing a New Test from the Item Bank

As a next step, a new sample test was developed by using the items from the bank. First, the purpose of the test was defined. It was intended to measure average ability of students (i.e., θ = −3 to +3). Then, a table of specifications in which the learning outcomes are listed along one side of a table and the subject matter topics are described along the other was prepared for the new test (See Table 2). 60 items from the bank were selected (20 items from the Form A, 21 items from the Form B and 19 items from the Form C) to develop the new test.

Table 2
*Table of Specifications for the New Test*

| Chapter | | Content | Learning Outcomes | | | Total Items | Total Percent | Content Weight |
|---|---|---|---|---|---|---|---|---|
| | | | Knowledge | Compre-hension | Application | | | |
| **Maths-1** | 1 | Natural Numbers, Whole Numbers and Their Operations | 2 | 3 | - | 5 | 8.3% | 8 |
| | 2 | Multiple Numbers, Factors, Prime Factorization, Highest Common Factor and Least Common Multiple | 3 | 5 | 2 | 10 | 16.7% | 16 |
| | 3 | The Fractions and Decimal Numbers | 4 | 5 | 3 | 12 | 20.0% | 20 |
| | 4 | Ratio, Percentage and Average | 2 | 4 | 1 | 7 | 11.7% | 12 |
| | 5 | Introduction to Algebra | 1 | 6 | - | 7 | 11.7% | 12 |
| **Maths-2** | 1 | The Geometric Figures in Environment | 1 | 2 | 1 | 4 | 6.7% | 6 |
| | 2 | Points, Lines, Rays and Segments | - | 1 | 1 | 2 | 3.3% | 4 |
| | 3 | The Angles | 2 | 3 | - | 5 | 8.3% | 8 |
| | 4 | The Basic Drawings by Using Set Squares | 2 | - | - | 2 | 3.3% | 4 |
| | 5 | Triangles | - | 1 | 1 | 2 | 3.3% | 4 |
| | 6 | Circles | 2 | 1 | 1 | 4 | 6.7% | 6 |
| | | Total | 19 | 31 | 10 | 60 | | 100 |
| | | Percentage | 31.6% | 51.7% | 16.7% | | 100% | |

According to Baker (2001), the items would be selected from the item bank on the basis of their contents and their item parameters to meet a particular testing goal. Therefore, a greater number of items with average difficulty values were selected to measure the most of the examinees more precisely (i.e., θ = −3 to +3). The items with fair discrimination values then were selected to measure the average examinees.

To be known precisely the maximum amount of information obtained by the new test, the test information curve (TIC) of the new test was also computed. Figure 6 illustrated the test information curve of the new test with the standard error (SE).

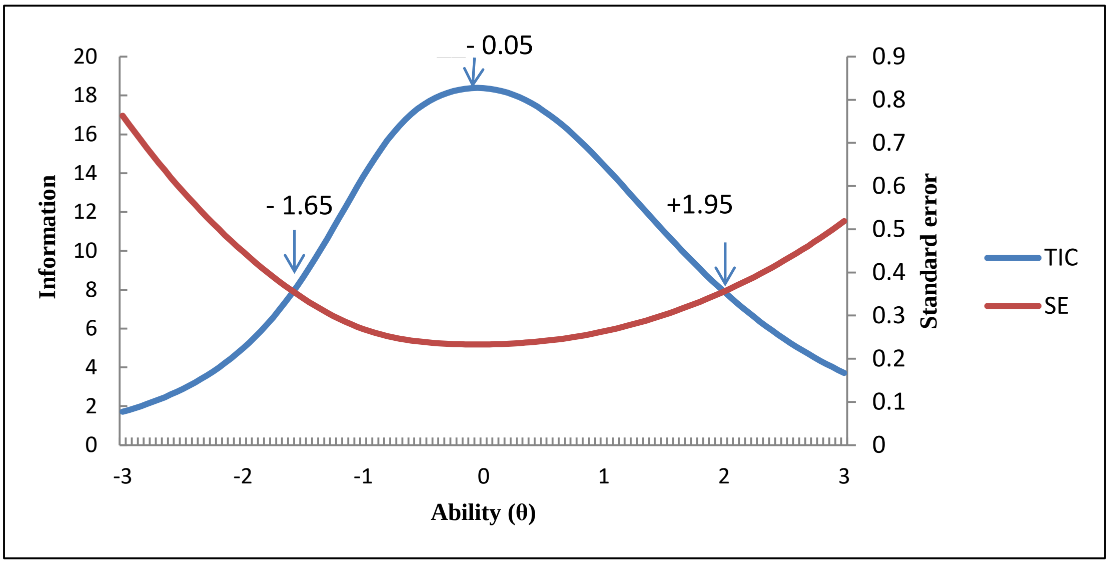


*Figure 6.* Test Information Curve for the new test.

In Figure 6, it was found that the test had the smaller standard errors across the ability scale from −1.65 to +1.95, and larger standard errors at the low and high ends of the ability scale. Thus, the test was discriminating well the examinees with the ability (θ) level −1.65 ~ +1.95, but was discriminating poorly among examinees with extremely low ability (θ) level (θ < −1.65) and who with extremely high ability (θ) level (θ > +1.95). Therefore, this test would be the most suitable for examinees whose ability (θ) range is −1.65 to +1.95. Moreover, the maximum amount of information I(θ) were 18.39 at θ = −0.05 in this scale. This meant that this test would provide the highest information for the examinees whose ability level (θ) at −0.05.

## Item Characteristic Curves and Test Characteristic Curve for the New Test

The item characteristic curves of all items in the new test were graphed to present the probability of choosing the correct answer to an item as a function of the level of the ability being measured by the test. In Figure 7, it was clearly found that the higher the examinees' ability level, the greater the probability of the examinees would correct the item. Besides, observing the item characteristics curves of this test, it contained a few items that had low discrimination values with high difficulty values in order to cover the content according to table of specifications.

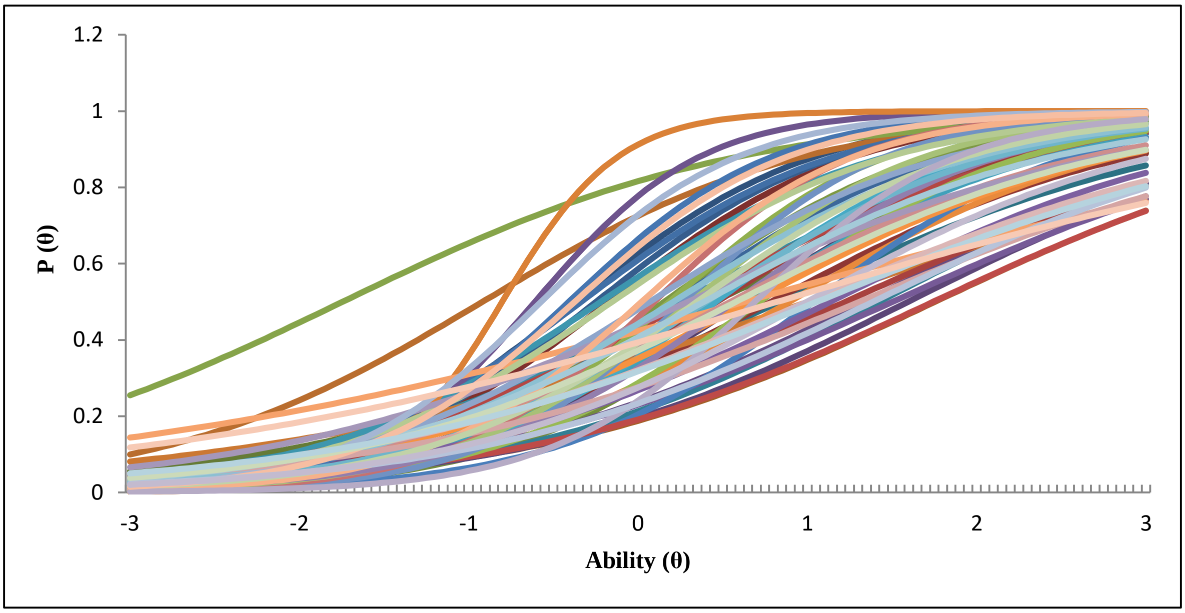


*Figure 7.* Item Characteristics Curves (ICCs) of all items in the new test.

Next, the test characteristic curve for the new test was presented in order to predict the relationship between the true score (τ) and the ability (θ) of the examinee. According to Figure 8, it could be predicted that if the examinees possess the ability level of $\theta = +1.5$, of $\theta = +1$ and of $\theta = -0.5$, they would probably get a true score of about 45, of about 39 and of about 18 respectively. Thus, it could be said that the higher the examinees' ability level (θ), the higher the scores would be obtained. Therefore, this test may be more appropriate to separate the master and non-master of students within the ability range $-1.65 \sim +1.95$.

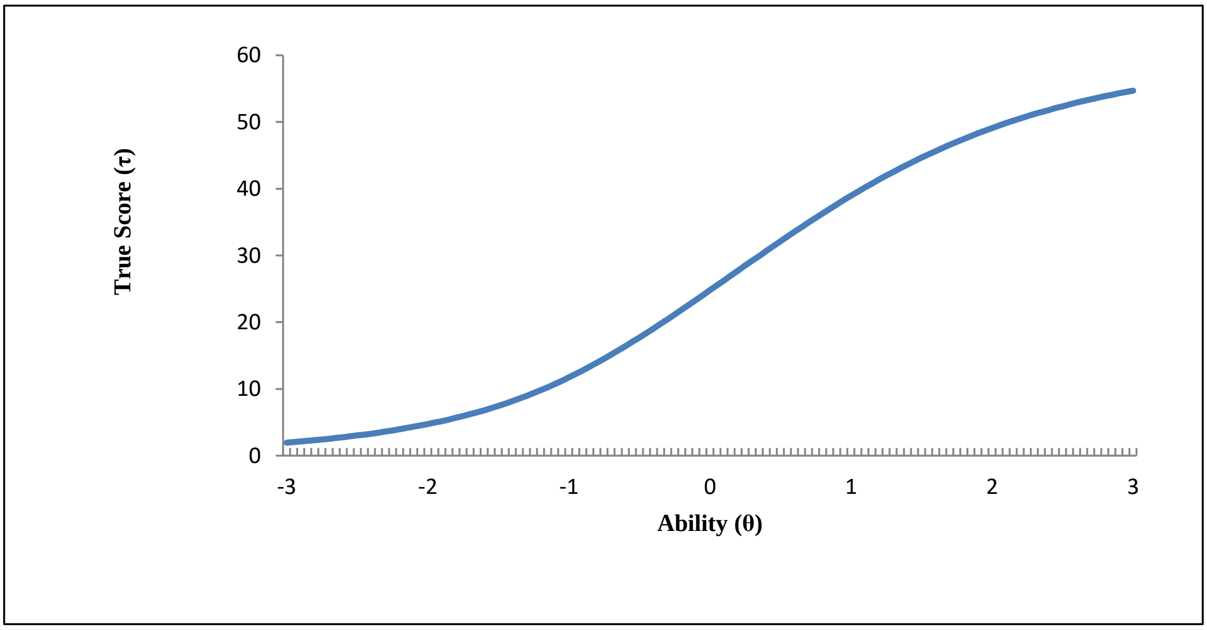


*Figure 8.* Test Characteristics Curve (TCC) of the new test.

## Conclusion

In conclusion, it is hoped that this study would help the test developers to get knowledge about how to construct an item bank and how to develop a new test or tests which are relevant to the particular testing goals with the help of item banking.

### Significance and Implications

This study has focused on the item banking for Grade Six Mathematics Achievement. Tests are the key factors in education in order to develop both teaching and learning progress. Moreover, various types of tests are usually used by the educators for the purpose of assessing students' achievement. Thus, it is essential to develop the high quality tests in order to estimate the information about students accurately. By constructing the tests systematically, the tests are more likely to be highly reliable and valid as well as are really useful in education process to some extent. But in order to develop a good quality test, many steps such as constructing table of specifications, item writing, pilot and field testing, calibrating the items, to name a few, should be taken carefully into account. But it will take a lot of time, effort and money. The best way to solve this problem is to develop the item bank. It is known that item banks are widely used in many countries for the development of tests. Therefore, tests developers in Myanmar should use the item banks because of many advantages whenever paper and pencil tests or computer-adaptive tests are created in the assessment practice.

### Limitations and Suggestions for Further Study

In this item bank, the items which were biased against different groups could not be determined. Thus, the differential item functioning (DIF) should be calculated in the future study in order to check whether the items in this item bank are biased against racial or ethnic minorities. As mentioned above on the facts that the content area was restricted, as well as there were only multiple-choice items and less number of items in this item bank, the new items need to be continuously developed and calibrated because this item bank was not sufficient enough to examine the Mathematics Achievement of Grade Six students yet.

Finally, the item bank in this study was developed for only Grade Six Mathematics. Therefore, the item banks for other grades and other subjects should be constructed in further study. If the item banks for all grade levels and subjects are developed in a systematic way, many different tests for different grade levels in different subjects will be able to be used readily in every school.

## References

Baker, F. B. (1986). Item banking in computer-based instructional systems. *Applied Psychological Measurement, 10(4),* 405-414.

Baker, F. B. (2001). *The basic item response theory* (2nd ed.). United States: ERIC Clearinghouse on Assessment and Evaluation.

Choppin, B. (1985). Principles of item banking. *Evaluation in Education, 9,* 87-90.

DeMars, C. (2010). *Item response theory*. Oxford University Press, Inc.

Hambleton, R. K., Swaminathan, H., & Rogers, J. H. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage Publications, Inc.

Htet Htet Lin. (2014). *Constructing an item bank for Grade-11 Mathematics with item response theory in Myanmar*. Unpublished master's thesis, Sagaing University of Education, Myanmar.

Kay Zune Aung & Nu Nu Khaing. (2017). Constructing Grade 5 English item bank. *Journal of the Myanmar Academy of Arts and Science, 10 A*, 287-312.

Millman, J., & Arter, J. A. (1984). Issues in item banking. *Journal of Educational Measurement, 21(4),* 315-330.

Rudner & Lawrence. (1998). *Item banking.* ERIC Clearinghouse on Assessment and Evaluation Washington DC, ED423310.

Tshering, G. (2006). *IRT in item banking, study of DIF items and test construction.* Unpublished master’s thesis, University of Twente.

Umar, J. (1999). Item banking, *Advances in measurement in educational research and assessment*, Pergamon Press, pp.207-219.

Wiersma, W., & Jurs, S. G. (1990). *Educational measurement and testing* (2nd ed.). USA: A Division of Simon & Schuster, Inc.