ISSN: 2278-4012, Volume:8, Issue:1, JANUARY 2019

# Development of Valid and Reliable Teacher-Made Tests for Grade 10 Mathematics

**Denis Abao Tan[1], Cherly C. Cordova[2],**
**Ian Paul B. Saligumba[3] and Lester Lou B. Segumpan[4]**
Central Mindanao University, University Town, Musuan, Bukidnon, 8710 Philippines

***Abstract:*** Tests are tools utilized by the teachers to evaluate their teaching and students' learning to improve instruction, curriculum and consequently compute grades. This study hopes to develop reliable and valid teacher-made tests. Tests for Grade 10 Mathematics in the first and second grading periods were content and face validated by three (3) experts. The content validation was done via congruency with the objectives per topic indicated as well as the level of cognitive domain set for the item in the Table of Specifications (TOS) based on Bloom's Taxonomy. Then, the tests underwent item analysis and distractor analysis utilizing the Item Analyzer software. Results indicated that the tests items were congruent with the set objectives and levels of cognitive domain. These tests have KR 20 of 0.82 and 0.85 for the first and second grading periodical examinations, respectively. These were of average level of difficulty and with reasonable items. The first 38 items were retained, 14 items were revised and 8 items were rejected. Distractor analysis showed the distractors to be changed while revising an item. Also, it indicated that the stem be improved when distractors were plausible. The tests were valid and reliable hence, measure actual performance of the students.

***Key Words:*** test development, valid test, reliable test, distractor analysis, item analysis

## Introduction

Mathematics teachers in the high school are confronted with multifarious tasks, just like teachers, instructors or professors in other levels do. These tasks include lesson planning and preparation, classroom management, teaching- learning process, language proficiency, assessment of learning and reinforcement of learning which are all inherent to the teaching profession. Among these many tasks that a teacher need to perform proficiently in the classroom is the assessment of students' learning through formative or summative test that he/she made for a specific period in a term or school year.

Classroom assessment plays an important role in determining the quality of the input, process and output elements of education. Classroom assessment is a teacher's way of gathering information about what students have learned. They can then use them to reach important decisions about the students' grades, the content of future courses, and the revision of the

ISSN: 2278-4012, Volume:8, Issue:1, JANUARY 2019

structure or content of a course or curriculum (Brookhart, 1999). In addition, they might also be used to gain information about students' specific weaknesses and special instructional needs and to identify any concepts or procedures which may need to be re-taught or reviewed (Yu, 2012).

One of the many classroom assessment tools made by the teachers are either formative or summative tests. And the questions on how well these teacher-made tests measure what they intend to measure and how consistent they are remain to be unanswered if teacher does not proceed to item analysis and establishment of the validity and reliability of these tests. In most cases, item analysis, which includes determination of item difficulty and discrimination, as well as distractor analysis are not done because it is time consuming and demanding if done manually.

Development of valid and reliable tests give teachers the necessary confidence that the decisions they made as a result of the tests are indeed pictures of the realities of their students and their classroom instruction. The results of the tests eventually describe actual competencies mastered, less mastered and not mastered by the students – which provide feedback on topics to be reteach, students' performance and the like. More so, this give teachers a feeling of confidence that the grades they give to their students are reflections of their learned competencies since these are measured by valid and reliable tests, hence, this investigation.

**Review of Literature**

Through time, educators have been utilizing tests in measuring and evaluating students' understanding of any subject matter at hand. However, tests which are developed my teachers may not be able to measure what it intends to measure (Cronbach, 1971; Messick, 1989) and therefore is not valid. This scenario gives rise to efforts of educators in constructing tests which are valid and reliable so that the grades they give to their students will speak of their actual performance in class. Teachers try harder in assuring that each item constructed in their teacher-made tests ranges from easy to difficult items with high discrimination index. Also, teachers provide plausible options in every multiple choice test so that only those who mastered the competencies will be able to answer the item correctly.

Validity is never an integral attribute of a test. It refers to the type of decisions which are based on the test scores and the consequent utilization of those scores (Cronbach, 1988). A specific test provides accurate information for a specific purpose of that test but not for another.

The development of test includes three stages: framework definition, test specification, and item selection (National Research Council, 1998). In the framework definition phase, teachers use content standards to define the scope of the subject area to be assessed. The development of test specifications follow which delineate how the domain will be represented. This gives rise to the making of a table of specification (TOS). The final stage consists of item construction based on the TOS.

ISSN: 2278-4012, Volume:8, Issue:1, JANUARY 2019

Among the stages in test development, the most crucial and time-consuming is the third phase – the item construction based on the TOS. Hamilton and Koretz (2002) described different test formats. They emphasized that formats vary according to the purpose of the tests. In the United States of America, 48 states utilized uniform statewide assessment programs which consist of multiple-choice items because it is highly reliable and it is low cost test's type. However, in the recent years when outcome-based education is introduced, many had utilized other forms of assessments like short-answer item test, extended-response or essay, portfolio assessment, and the like. Formats of the test other than multiple-choice are mostly expensive, and oftentimes people scored them rather than machines. More testing time are consumed too using other test format. In which case, it generates problems like liability per unit time and higher difficulty in comparing test forms. However, many educators and policymakers consider other authentic measurements far better than the multiple choice tests. They believe that these type of assessment measures other skills of students that cannot be captured by mere multiple choice item. Since multiple-choice items are cheaper to be validated compared to other formats, the researchers decided that the tests to be developed will be of this type.

Multiple choice tests require students to select the answer from a number of some possible alternatives (Kolawole, 2005). Multiple choice items provide the reasonable opportunity to students to demonstrate their capability and testers to prove their honesty. The impartiality of multiple choice test is based on its development and scoring as items cover wider curriculum contents and objectives of instruction. It is declared as having good legitimacy since it has the tendency to cover all aspects of learning content (Alonge, 2003 and Lawal, 2001).

The quality and properties of a multiple choice tests are bases for its usefulness in achieving objectives of testing. The significance of difficulty indices and discrimination power in multiple choice items cannot be overemphasized The Classical Test Theory (CTT) as opined by Schumacker (2005) employs traditional item and sample dependent statistics i.e. item difficulty and item discrimination. In classical theory, item difficulty and item discrimination are the two statistics that form the cornerstone of the test.

Adewuyi and Oluotun (2001) defined difficulty index of an item as the degree to which an item has been responded correctly by the students. That is the proportions of the students that hand-picked the right option (Alonge, 2003). The closer to a value of one (1), the difficulty index indicates simpler item and the closer it is to zero means the item becomes more difficult.

The difficulty index of the item expresses how easy the item was for the students in that specific group. When the difficulty index is higher, it means that the questions are easier, and when the difficulty index is lower, the questions tend to be more difficult. In fact, "Easiness index" is another term for difficulty index (Zafar, 2008). Multiple choice tests with less numbers

ISSN: 2278-4012, Volume:8, Issue:1, JANUARY 2019

of options have superior difficulty indices than those with the bigger number of options (Abiri, 2006).

According to Alonge (2003), the discrimination power of multiple choice items is the ability to discriminate between the outstanding students and poor students. On the other hand, discrimination power is described by Oyejide (1991) as the strength of each item to distinguish the higher achievers (those who are more competent) from, the lower achievers (those who are less competent). The test discrimination power ranges from zero to one (0-1). The closer this value is to one (1) the better the item is (Oyedeji, 1991 and Kelly, 1989). The index of discrimination is also the extent to which a test is correctly responded to by those examinees possessing more of the traits being measured (Ebel, 1979 and Alonge, 2003).

Le and Klein (2002) in their book entitled, "Making Sense of Test-Based Accountability in Education", stressed out the significance of developing valid and reliable tests to determine its alignment to the content standards and to ensure that it measures actual learning of the students. Chapter three of this book focuses on the technical criteria for evaluating tests. It presented concepts like reliability, validity, and fairness. These concepts highlight the challenges that many of the test developers faces today. These includes implementation of test-based accountability systems, and utilization of the test results to make decisions about education programs and the performance of the students. Test scores being numeric often assumed to have justifiable degree of precision. This is the reason why evaluation of the technical characteristics of the test be made by those who will use the test scores to come up with decisions. The evaluation is necessary in making informed decisions about what are the actual meaning of the scores and how much assurance can be set in them.

To date, teachers have used assessments for different purposes and this has influenced teaching and learning differently. These are identified as summative assessments and formative assessments. Summative assessments are usually used to meet the requirements of graduation and academic placement, while formative assessments are mainly administered in order to garner advice on pedagogical decision-making processes and curriculum change. For instance, high-stakes standardized tests can be given as examples of summative assessment, while the tests developed by teachers or assigned homework and projects are examples of formative assessment (Liang, 2010). In this study, the purpose of assessment was to determine the performance of the students. This performance can be dependably measured by a valid and reliable teacher-made test (Cordova & Tan, 2018; Fulgencio & Tan, 2018; Saligumba & Tan, 2018; Salingay & Tan, 2018; Segumpan & Tan, 2018), which underwent review, revision and item analysis.

Classroom assessment is divided into learning and performance-focused environments (Buldur, 2014; Brookhart, 1997; İlhan, 2017; McMillan & Workman, 1998). While learning-focused assessments are dominated by an evaluation understanding aimed for students' learning;

ISSN: 2278-4012, Volume:8, Issue:1, JANUARY 2019

ultimately in performance-based assessment, their role is to understand and assess student learning (Buldur & Doğan, 2014; İlhan, 2017). In this study, the establishment of validity and reliability of a performance-based teacher-made test is the main focused.

Conceptual Framework

Valid and reliable tests are essential to the study of abilities, aptitudes, and attitudes. In this case, it is about 100 years already that classical test theory (CTT) has serve the testing field. However, major and positive changes has occurred in the psychological and education assessment during its implementation (see, for example, Hambleton & Jodoin, 2003).

Frederic Lord and other psychometrician in 1950s and 1960s initially developed Item Response Theory (IRT) (Lord, 1952; Lord & Novick, 1968) whose goal was to develop a method that is able to evaluate the respondents without necessarily depending on the items given in the test (Hambleton & Jodoin, 2003). To overcome the many limitations of the classical measurement theory, IRT evolved (Hambleton, 1994). IRT is comprised of different models in mathematics, it is considered a statistical theory with following characteristics: a) to calculate respondents' scores based on their skills or covert traits and b) to institute a relationship between students' item performance and the group of traits subordinate item performance through "item characteristic curve" (Hambleton et al., 1991). These characteristic are conceivable because IRT represents the same items used in different respondents which will keep their statistical properties (for instance, difficulty and discrimination), and respondents' scores that reflect ability or latent traits on a given subject of concern do not rest on the specific test items being administered.

The use of Item Response Theory (IRT) with test development has several advantages because it produces person parameter invariance (test scores are not dependent on the particular choice of test items) when model fit is existing, and test material tasks provide the volume of information or "measurement precision" apprehended by the test on the scale assessing the concept of interest (Embretson, 1996; Hambleton et al., 2000) and other structures too.

**Methodology**

The study was conducted at CMU Laboratory High School, College of Education, Central Mindanao University. The respondents of this study were the Grade 10 students of the CMULHS, College of Education, who were officially enrolled during the SY 2017-2018 taking Grade 10 mathematics subject. The study utilized the Grade 10 mathematics tests developed by the teacher as well as the results of the examination during the first and second grading period. It analyzed the test items constructed by the teachers versus the Table of Specifications (TOS) of the test to establish content validity of the test. Content and competency-standards plotted in the TOS were cross-examined first using the standards set by Department of Education (DepEd) in the said subject.

ISSN: 2278-4012, Volume:8, Issue:1, JANUARY 2019

The tests answered by the students were analyzed using the item analyzer software purchased by the CMULHS- Parent Teachers Association (CMULHS-PTA) developed by Dr. Cesar Bermundo in 2005. The software reported the different data that are needed in the item analysis and distractors analysis of the tests. These data were encoded and analyzed to obtain information in order to answer the questions put forward for this investigation.

The data were treated using the descriptive statistics such as frequency counts, and percentages to describe the variables under study. Item analysis to determine index of difficulty and index of discrimination were conducted. Distractor analysis to determine if options are plausible or not was done. Finally, Kuder-Richardson 20 (KR20) was reported to determine the reliability of the tests.

**Findings**

This section presents the description, analysis and interpretation of data. The presentation is arranged in accordance to the order of the research questions as reflected in the statement of the problem.

Learning Competencies Measured by the Teacher-made Tests

Table 1 presents the table of specification (TOS) which contains the learning competencies that the teacher-made tests measured for the specified grading periods. There are two (2) sets of learning competencies since there are two tests representing the first (Table 1A) and second grading (Table 1B) periodical examinations. Each topic has specified learning objectives or outcomes to be measured if learned by the students during the period specified.

For the first grading, the topics include generalizing a pattern, arithmetic sequence, arithmetic series, problems involving arithmetic sequence, geometric sequences, geometric series, problems involving geometric sequences and other types of sequences. Students need to define, describe, recognize and differentiate polynomial functions and polynomial expressions.

ISSN: 2278-4012, Volume:8, Issue:1, JANUARY 2019

Table 1A TOS for the First Grading Examination in Math 10 (SY 2017-2018)

| Topics | Objectives<br>At the end of the grading period, the students must have | Time Spent | Total | % | Remembering | Understanding | Applying | Analyzing | Evaluating |
|---|---|---|---|---|---|---|---|---|---|
| Generalizing a Pattern | 1. Observed and generalized a pattern; | | | | | | | | |
| | 1.1 arithmetic sequence | 1 | 5 | 8.33 | 1 | 2 | 3 | 4 | 5 |
| | 1.2 quadratic sequence | 1 | 5 | 8.33 | 6 | 7 | 8 | 9 | 10 |
| | 1.3 geometric sequence | 1 | 5 | 8.33 | 11 | 12 | 13 | 14 | 15 |
| Arithmetic Sequences | 2. Defined, illustrated, and graphed an arithmetic sequence;<br>3. Determined the terms of an arithmetic sequence including the general nth term of the sequence; | 2 | 10 | 16.67 | 16,17 | 18,19 | 20,21 | 22,23 | 24,25 |
| Arithmetic Series | 4. Calculated the sum of terms of a given sequence; | 1 | 5 | 8.33 | 26 | 27 | 28 | 29 | 30 |
| Problems involving Arithmetic Sequences | 5. Solved problems involving sequences and their sums; (arithmetic sequences and series) | 2 | 5 | 8.33 | | | 31,32 | 33,34 | 35 |
| Geometric Sequences | 6. Defined, illustrated, and graphed a geometric sequence;<br>7. Determined the terms of a geometric sequence including the general nth term of the sequence; | 2 | 10 | 16.67 | 36,37 | 38,39 | 40,41 | 42,43 | 44,45 |
| Geometric Series | 8. Calculated the sum of terms of a given geometric sequence, both finite and infinite; | 1 | 5 | 8.33 | 46 | 47 | 48 | 49 | 50 |
| Problems involving Geometric Sequences | 9. Solved problems involving sequences and their sums; (geometric sequences and series) | 2 | 5 | 8.33 | | | 51,52 | 53,54 | 55 |
| Other Types of Sequences | 10. Defined and illustrated a sequence and some types of sequences(e.g., harmonic, Fibonacci); | 1 | 5 | 8.33 | 56 | 57 | 58 | 59 | 60 |
| | Total | 14 | 60 | 100 | 10 | 10 | 14 | 14 | 12 |
| | | | % | 100% | 16.67 | 16.67 | 23.33 | 23.33 | 20 |

They also are expected to perform operations (synthetic division, etc.) on polynomial expressions, state theorems and proofs of the remainder theorem, and factor theorem, determine the zeroes of the polynomial functions, state and provide proof of rational zero theorem, graph and solve real world problems involving these concepts. The DepEd Curriculum Guide for Grade 10 Mathematics indicates similar competencies. These topics and the content standards as well as performance standards are alike (DepEd Curriculum Guide, 2015).

ISSN: 2278-4012, Volume:8, Issue:1, JANUARY 2019

Table 1B TOS for the Second Grading Examination in Math 10

(SY 2017-2018)

| Topics | Objectives<br>At the end of the grading period, the students must have | Time Spent | Total | % | Remembering | Understanding | Applying | Analyzing | Evaluating |
|---|---|---|---|---|---|---|---|---|---|
| Basic Concepts of Polynomial Functions | 1. Defined and described polynomial functions;<br>2. Recognized examples of polynomial functions;<br>3. Differentiated between polynomial functions and polynomial expressions; | **1** | **5** | **7.14** | **1** | **2** | **3** | **4** | **5** |
| Operations on Polynomial Functions | 4. Performed operations on polynomial functions;<br>4.1 addition/subtraction | 1 | 5 | 7.14 | 6 | 7 | 8 | 9 | 10 |
| | 4.2 multiplication | 1 | 5 | 7.14 | 11 | 12 | 13 | 14 | 15 |
| | 4.3 division<br>5. Described, illustrated, and performed the synthetic division process for dividing polynomial expressions by a binomial; | **1** | **5** | **7.14** | **16** | **17** | **18** | **19** | **20** |
| The Remainder Theorem | 6. Stated, provided a proof of the theorem and applied the remainder theorem; | 2 | 10 | 14.28 | 21,22 | 23,24 | 25,26 | 27,28 | 29,30 |
| The Factor Theorem | 7. Stated, provided a proof of the theorem and applied the factor theorem; | 1 | 5 | 7.14 | 31 | 32 | 33 | 34 | 35 |
| Roots and Zeroes | 8. Determined the zeroes of polynomial functions; | 1 | 5 | 7.14 | 36 | 37 | 38 | 39 | 40 |
| Rational Zero Theorem | 9. Stated, provided a proof of the theorem and applied the rational zero theorem; | 2 | 10 | 14.28 | 41,42 | 43,44 | 45,46 | 47,48 | 49,50 |
| Composite Functions | 10. Described the nature and solved problems involving composite functions | 2 | 10 | 14.28 | 51,52 | 53,54 | 55,56 | 57,58 | 59,60 |
| Graphs of Polynomial Functions | 11. Sketched graphs of polynomial functions; | 2 | 10 | 14.28 | 61,62 | 63,64 | 65,66 | 67,68 | 69,70 |
| | Total | 14 | 70 | 100 | 14 | 14 | 14 | 14 | 14 |
| | | | | 100% | 20 | 20 | 20 | 20 | 20 |

For the second grading period, the topics include basic concepts of polynomial functions, operations of polynomial functions, the remainder theorem, the factor theorem, roots and zero theorem, composite functions, and graphs of polynomial functions. Competencies to be developed by students include generalize pattern of varied series and sequences, define and determine the different series and sequences, calculate the sums of the given series, solve problems involving series and the like (DepEd Curriculum Guide, 2015).

The two tables indicated the number of hours each topic was discussed in the classroom which become bases for the number of questions to be given on that specific topic per percentage. This distribution ensures that the topics discussed in a longer period will have more items in the test and vice versa. This TOS is in line with the assessment standards set for

ISSN: 2278-4012, Volume:8, Issue:1, JANUARY 2019

classroom use, which is indicative of the content validity of the test developed from these blueprints.

These assessment practices of the mathematics teacher is supported by National Research Council (1998). The utilization of TOS as blueprint in the construction of the test ensures validity of the instrument.

Moreover, Husna and others (2018) believed that cognitive competence plays an important role in enhancing mathematical literacy in the teaching and learning mathematics in secondary schools. Hence, for them, the quality of test items requires to be aligned with Bloom's cognitive taxonomy. This notion supported the findings of this study.

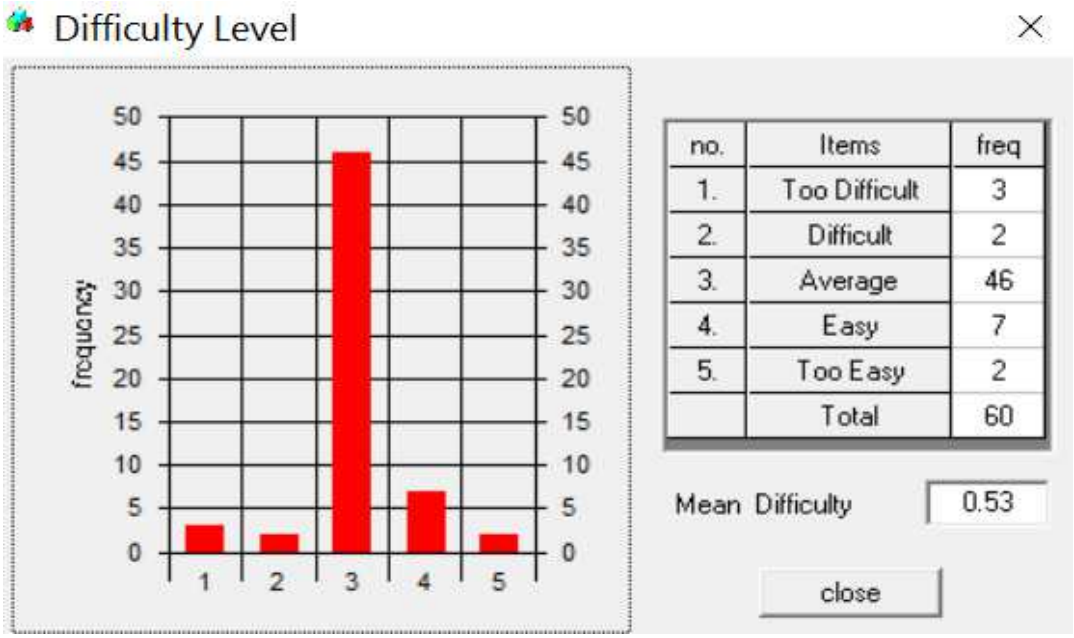



Figure 1A. Distribution and Mean Difficulty Index Level of Grade 10 Mathematics for the First Grading Period

Figure 1A shows the frequency of items with their difficulty level. As reflected, majority of the items in the Grade 10 Mathematics for the first grading falls in an average level of difficulty (46 0ut of 60). There are 3 and 2 items which are too difficult and too easy, respectively. Only 2 items are considered difficult while 7 are easy. The mean difficulty level of the test is 0.53 which is considered average difficulty. Adewuyi and Oluotun (2001) described difficulty index as the percentage of students who got the answer correctly. This indicates that approximately 53% of the Grade 10 students answered the test items correctly.

ISSN: 2278-4012, Volume:8, Issue:1, JANUARY 2019

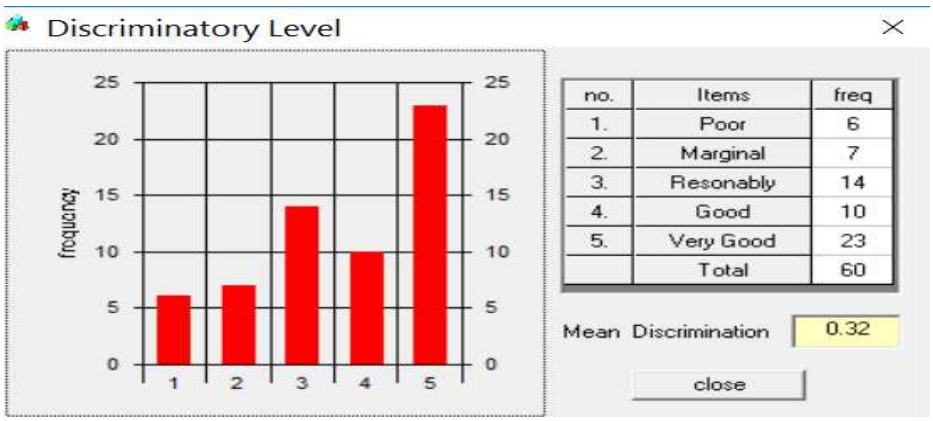



Figure 1B. Distribution and Mean Discrimination Index Level of Grade 10 Mathematics for the First Grading Period

Figure 1B displays the frequency of items with its level of discrimination index. As reflected, most of the items in the Grade 10 Mathematics for the first grading are very good items. These items discriminate performing from non-performing students. Six out of 60 are poor items, seven (7) are marginal, 14 are reasonably good, and 10 are good items. This indicates that each item of the test has varied level of discrimination index (Please see Appendix A). This also illustrates that the test itself has very good discriminatory index in identifying the performing from not. The mean discrimination index level of the test is 0.32 which means that it contains good items. Having a test with good items specifies that these items are well-written and well-thought of by the subject teacher. This further support the giving of grades of the teacher to his/her students because their scores indicates that they actually possess the competence the teacher tried to measure in the test. This is supported by Alonge (2003) when he emphasized that higher discrimination index of a test indicates the extent to which the test is correctly responded to by the students who possess the traits being measured.

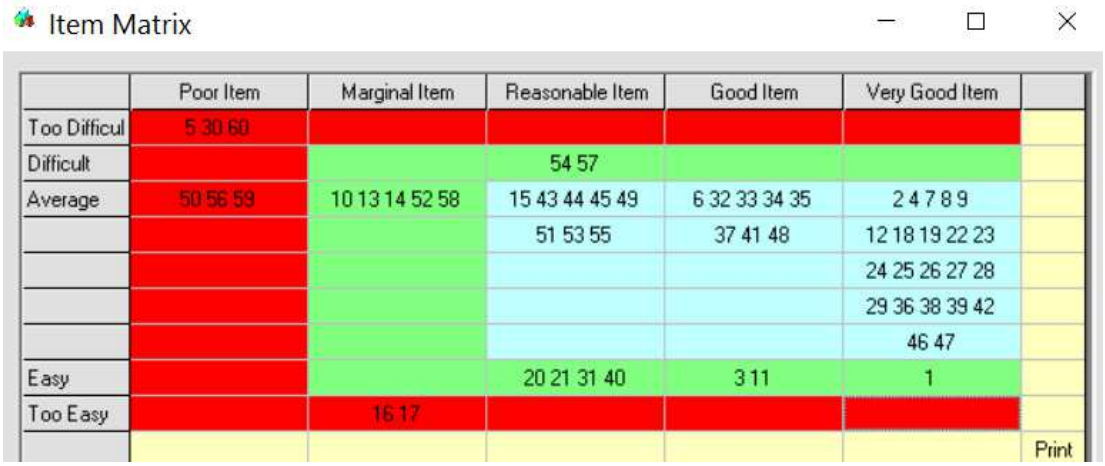



Figure 1C. Distribution of Test Items Based on each Difficulty and Discrimination Indices

ISSN: 2278-4012, Volume:8, Issue:1, JANUARY 2019

Figure 1C presents the distribution of the items based on their difficulty and discrimination indices. As illustrated, items number 5, 30, and 60 are poor and too difficult items. Items number 50, 56 and 59 are poor and average items. Items number 16 and 17 are marginal item yet too easy ones. These items are actually beyond the acceptable standards for item difficulty and discrimination indices. These further specify that these items will be rejected during revision of the test because these are poor and marginal items which are too difficult and too easy ones for the students. These items cannot identify students who are performing and non-performing in the class.

The same figure shows that there are 14 items that are needed to be revised. These are items number 10, 13, 14, 52, and 58 which are marginal and average items, items number 45 and 57 which are reasonable and difficult items, items number 20, 21, 31 and 40 which are reasonable but easy items, items number 3, and 11 which are good but easy items, and item number 1 which is very good but easy items. These fourteen items will be reconstructed according to how well the stem was constructed or how well the distractors function as such in an item. Per item distractor analysis would help the teacher decides on what to do on each of these to be revised test items (Please see Appendix A). Distractors with N indicates that it does not function as a distractor and needs to be changed.

The rest of the items not mentioned are considered items to be retained because these are reasonably good, good and very good items ranging from difficult to easy level of difficulty indices. These items are neither too easy nor too difficult to answer by the students. More so, these items can separate students who are performing and not performing ones. These items now form part of the questions in the test bank for Math 10 examinations.

The reliability index of the test is 0.85 which indicates that the test is a very good one for a classroom test. This exhibits the consistency of the test in the competencies it is designed to measure regardless of time and who will take the test.

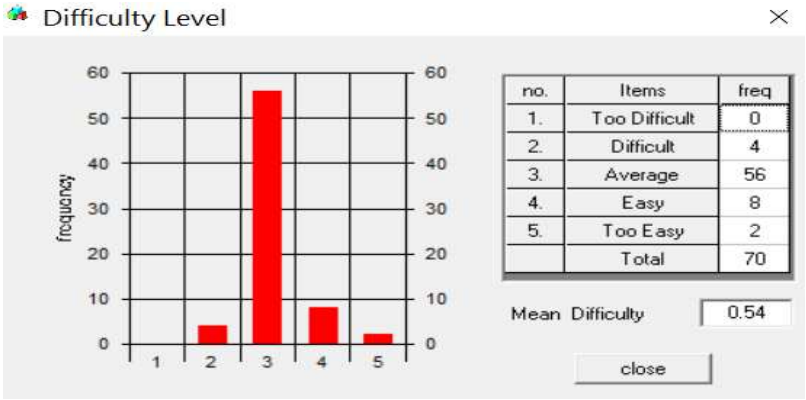



Figure 2A. Distribution of Difficulty Indices of Items in the Grade 10 Mathematics for Second Grading Period

ISSN: 2278-4012, Volume:8, Issue:1, JANUARY 2019

Figure 2A illustrates the difficulty indices of the test items as well as its mean difficulty level. As shown, there are two (2) too easy items, eight (8) easy items, 56 average items and four (4) difficult items. This indicates that majority of the test items are in the average level of difficulty which is favorable for a teacher-made test. There are only to items which are too easy, which indicates further that the test is not too difficult nor too easy for the students to answer. The mean difficulty index of the test is 0.54 which means it has average level of difficulty.

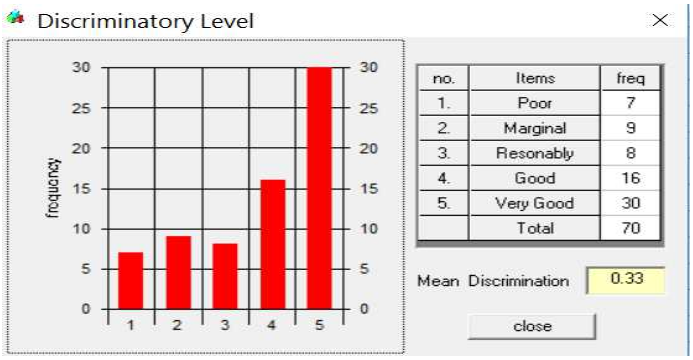
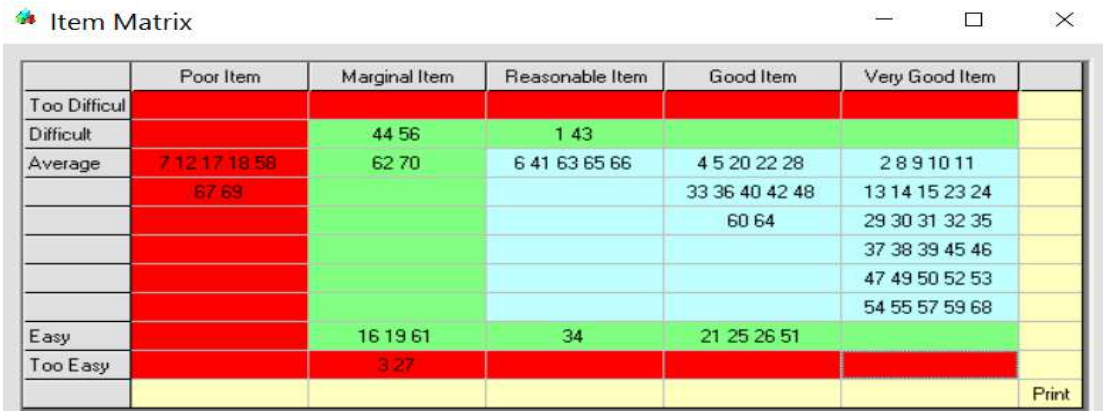



Figure 2B. Distribution of Discrimination Indices of Items in the Grade 10 Mathematics for Second Grading Period

The discriminatory indices of each test item as well as the mean discriminatory index of the test is shown in Figure 2B. This illustrates that items in the second periodical examination for Grade 10 Mathematics vary in their discriminatory levels. There are seven (7) items which are poor, nine (9) are marginal, eight (8) are reasonable, 16 are good and 30 are very good items. These show that majority (30 out of 70 items) are very good items indicating that these items are able to distinguish performing from non-performing students (Alonge, 2003). Also, only 7 (7 out of 70) are considered poor. This displays that the formulation of the test was made with utmost care and thorough thinking by the subject teacher. The mean discrimination index of the test is 0.33 which means that the test contains reasonably good items. This further exhibits the test development skill of the math teacher in the unit. For a more detailed item discrimination index per item is presented in Appendix B.



ISSN: 2278-4012, Volume:8, Issue:1, JANUARY 2019

Figure 2C. Distribution of Test Items Based on each Difficulty and Discrimination Indices of the Grade 10 Mathematics for the Second Grading Period

The distribution of the items based on their difficulty and discrimination indices are presented in Figure 2C. As shown, items number 7, 12, 17, 18, 58, 67, and 69 are poor and average items. Items number 3 and 27 are marginal and too easy items. These items are actually beyond the acceptable standards for item difficulty and discrimination indices. These further specify that these items were rejected during revision of the test because these are poor and marginal items which are average and too easy ones for the students. These items cannot identify students who are performing and non-performing in the class. These findings are supported by Abiri (2016) and Zafar (2018), items with high easiness index and low discrimination index must be excluded in the test revision.

The same figure shows that there are 14 items that are needed to be revised. These are items number 44 and 56, 62 and 70, and 16, 19 and 61 which are marginal items but difficult, average and easy items, respectively. Items number 1, and 43 and 34 are reasonable and difficulty and easy items, respectively. While items number 21, 25, 26 and 51 are good but easy items. These fourteen items were reconstructed according to how well the stem was constructed or how well the distractors function as such in an item. Per item distractor analysis would help the teacher decides on what to do on each of these to be revised test items (Please see Appendix B).

The rest of the items not mentioned are considered items to be retained because these are reasonably good, good and very good items ranging from difficult to easy level of difficulty indices. These items are neither too easy nor too difficult to answer by the students. More so, these items can separate students who are performing and not performing ones. These items now form part of the questions in the test bank for Math 10 examinations.

The reliability index of the test is 0.85 which means that the test is a very good classroom test. This is supported by Macmillan (2000) when he opined that the level of reliability and validity of measurement tools plays a role in determining students' actual performance in class which is related to success. He added that when assessments by teachers were found to be inappropriate or that did not serve their purpose (not valid), it is important for teachers to question the decisions made with regard to these tools rather than simply questioning their application frequency. This implies that validity and reliability of the teacher-made tests are very essential in the teaching-learning process. Hence, establishment of these assessment characteristics is a must.

## Conclusion

Based on the results and findings of the study, the following conclusions were drawn:

ISSN: 2278-4012, Volume:8, Issue:1, JANUARY 2019

The tests measure the competencies developed by the students while learning generalizing a pattern, arithmetic sequence, arithmetic series, problems involving arithmetic sequence, geometric sequences, geometric series, problems involving geometric sequences and other types of sequences for the first grading period; and basic concepts of polynomial functions, operations of polynomial functions, the remainder theorem, the factor theorem, roots and zero theorem, composite functions, and graphs of polynomial functions for the second grading period.

The tests are of average difficulty with good and reasonable items. Items are neither too easy nor too difficult for the students and these can distinguish high performing students to low performing ones.

The first grading test has 38 items which are retained, 14 items are revised and 8 items are rejected. The revised test with good items contains 52 items with difficulty level ranging from easy to difficult items. On the other hand, the second grading test has 47 items which are retained, 14 items are revised and 9 items are rejected. These leave a 61-item test with average level of difficulty and high discrimination index.

The teacher-made tests are valid and reliable with 52 good items for the first grading and 61 reasonably good items for the second grading.

**Suggestions and Recommendations**

Teachers are encouraged to develop their formative and summative test based on the topics intended for the sessions. Term or periodic examinations will be made with TOS as its blueprint to ensure content validity of the tests.

Construction of items in a test will be made with care such that it will be not too easy nor too difficult for the students to answer. Also, options to be given in a multiple choice test be made plausible to identify those who are studying and guessing during the test.

Identifying test items which are retained, revised and rejected is very helpful for teachers to come up with a test bank. These test items may be modified or used for the next school year because these are good items. Hence, teachers are enjoined to do item analysis of their tests.

Teachers are challenged to establish validity and reliability of their tests. These ensures valid grades and good inputs for instructional innovations, curriculum revision and intervention development to enhance student outcomes and sustain quality programs and graduates.

**Acknowledgement**

The researchers greatly acknowledged Central Mindanao University headed by Dr. Maria Luisa R. Soliven for the research grant that enabled them to do this work. Also, they would like to extend their gratitude to the Vice President for Research, Development and Extension, Dr.

Luzviminda T. Simborio, the Director of Research, Dr. Angela Grace T. Bruno, the research coordinator, Prof. Elhrich Ray J. Magday, the Dean of the College of Education, Dr. Raul C. Orongan, and the students for the support extended during the planning up to the finalization of this paper.

## References

Abiri, J. O. O. (2006). Elements of Evaluation Measurement and Statistical Techniques in Education. Ilorin: Library and Publication committee, University of Ilorin, Nigeria.

Adewumi, J. O. and Oluokun, O. (2001). Introduction to Test and Measurement in Education. Oyo: Odumatt Press Publishers.

Brookhart, S. M. (1999). *The art and science of classroom assessment: The missing part of pedagogy. ASHE-ERIC Higher Education Report (Vol: 27, No: 1).* Washington, DC: The George Washington University, Graduate School of Education and Human Development.

Brookhart, S. M., & DeVoge, J. G. (1999). Testing a theory about the role of classroom assessment in student motivation and achievement. *Applied Measurement in Education, 12*(4), 409-425.

Brookhart, S. M., & Durkin, T. D. (2003). Classroom assessment, student motivation, and achievement in high school social study classes. *Applied Measurement in Education, 16*(1), 27-54.

Buldur, S. (2014). The investigation of the relationship between the students' perceptions about the classroom assessment environment and their achievement-goal orientations: Gender Perspective. *Education and Science, 39*(4), 213-225.

Buldur, S., & Doğan, A. (2014). Adaptation of the students' perceptions of the science and technology course classroom assessment environment scale (SPCAES) into Turkish. *Education and Science, 39*(176), 199-211.

Cangelosi, J.S. (1988). Development and Validation of the Underprepared Mathematics Teacher Assessment, Journal for Research in Mathematics Education, Vol. 19, No. 3 (May, 1988), pp. 233-24.

Cordova, C., & Tan, DA. (2018). Mathematics Proficiency, Attitude and Performance of Grade 9 Students in Private High School in Bukidnon, Philippines‖. *Asian Academic Research Journal of Social Sciences and Humanities*, vol. 5, issue 2, pp. 103-116, February 2018.

Department of Education Curriculum Guide (2015). Mathematics Curriculum Guide for K to 12. Retrieved from www.deped.gov.ph on October 10, 2018.

Embretson, S.E. (1996). The new rules of measurement. Psychol Assess. 8(4):341–9. doi:10.1037/1040-3590.8.4.341.

Ciubal-Fulgencio, N., & Tan, D. (2018). Effects of mathematics communication strategies on attitude and performance of grade 8 students, *Asian Academic Research Journal of Multi-disciplinary*, Volume 5, Issue 2, February 2018.

Hambleton, R.K. (1994). Item response theory: A broad psychometric framework for measurement advances. Psicothema. 1994; 6(3) :535–56.

Hambleton, R.K. Robin F, and Xing D. (2000). Item response models for the analysis of educational and psychological test data. In: Tinsley HEA, Brown SD, editors. Handbook of Applied Multivariate Statistics and Mathematical Modeling. San Diego: Academic; p. 553–85.

Hambleton, R.K., and Jodoin M. (2003). Item response theory: models and features. In: Fernández-Ballesteros R, editor. Encyclopedia of Psychological Assessment. London: Sage; p. 509–14.

Hambleton, R.K., (2005). Applications of item response theory to improve health outcomes assessment: Developing item banks, linking instruments, and computer-adaptive testing. In: Lipscomb J, Gotay CC, Snyder C, editors. Outcomes Assessment in Cancer: Measures, Methods and Applications. Cambridge: Cambridge University Press; p. 445–64.

Hambleton, R.K, and Jodoin, M. (2003). Item response theory: models and features. In: Fernández-Ballesteros R, editor. Encyclopedia of Psychological Assessment. London: Sage; p. 509–14.

Hamilton, L.S. and Koretz, D. M. (2002). Tests and their Use in Test-Based Accountability Systems, Making Sense of Test-Based Accountability in Education. RAND Publication. (Eds) Laura S. Hamilton, Brian M. Stecher, Stephen P. Klein.

M Husna *et al. (*2018). *J. Phys.: Conf. Ser*. **1088** 012043.

Kolawole, E. B. (2005). Test and Measurement. Lagos: Bolabay Publications.

Kolawole, E.B. (2007). A Comparative Analysis of Psychometric Properties of Two Nigerian Examining Bodies for Senior Secondary Schools Mathematics. Research Journal of Applied Sciences, 2(8): 913-915.

Lawal, A. (2001). Evaluation of Students' Learning Outcomes 1: Types and Uses of Tests.

Le, V.N. and Klein, S. P. (2002). Technical Criteria for Evaluating Tests, Making Sense of Test-Based Accountability in Education. RAND Publication. (Eds) Laura S. Hamilton, Brian M. Stecher, Stephen P. Klein.

Lord, F. M. (1952). A theory of test scores. Psychometrika; Monograph 7.Google Scholar.

Lord F.M., and Novick, M. R. (1968). Statistical Theories of Mental Test Scores. Reading: Addison-Wesley; Google Scholar.

İlhan, M. (2017). Öğrencilerin sınıf değerlendirme atmosferine ilişkin algılarının kula yönelik tutumları üzerindeki yordayıcı rolü [The predictive role of students' perceptions of classroom assessment environment on their attitudes towards school].*Kastamonu Eğitim Dergisi*, *25*(1), 111-128.

McMillan, J. H. (2000). Fundamental assessment principles for teachers and school administrators. *Practical Assessment, Research &Evaluation, 7*(8).

Messick, S. (1989). Validity, In R.L. Linn (Ed) Educational Measurement (3rd ed, pp. 13-103), New York: Mcmillan.

Saligumba, I.P., & Tan, D. (2018). Gradual Release of Responsibility Instructional Model: Its Effects On Students' Mathematics Performance And Self-Efficacy. *International Journal of Scientific & Technology Research.* Volume 7, Issue 8, August 2018.

Salingay, N., & Tan, D. (2018). Concrete-Pictorial-Abstract Approach On Students' Attitude And Performance In Mathematics, *International Journal of Scientific & Technology Research*, Volume 7, Issue 5, May 2018.

Segumpan, L., & Tan, D. (2018). Mathematics performance and anxiety of junior high school students in a flipped classroom, *European Journal of Education Studies*, Volume 4 ,Issue 12.

Schumacker, R.E. (2005). Class Test Analysis. London, Applied Measurement Associates.

Villarreiz, A. (1982). Occupational Interest of Fourth Year Students. Unpublished Master's Thesis, Bukidnon State College, Malaybalay, Bukidnon.

Yao, L., (2010) Fall. Reporting Valid and Reliable Overall Scores and Domain Scores, Journal of Educational Measurement, Vol. 47, No. 3 (Fall 2010), pp. 339-360.

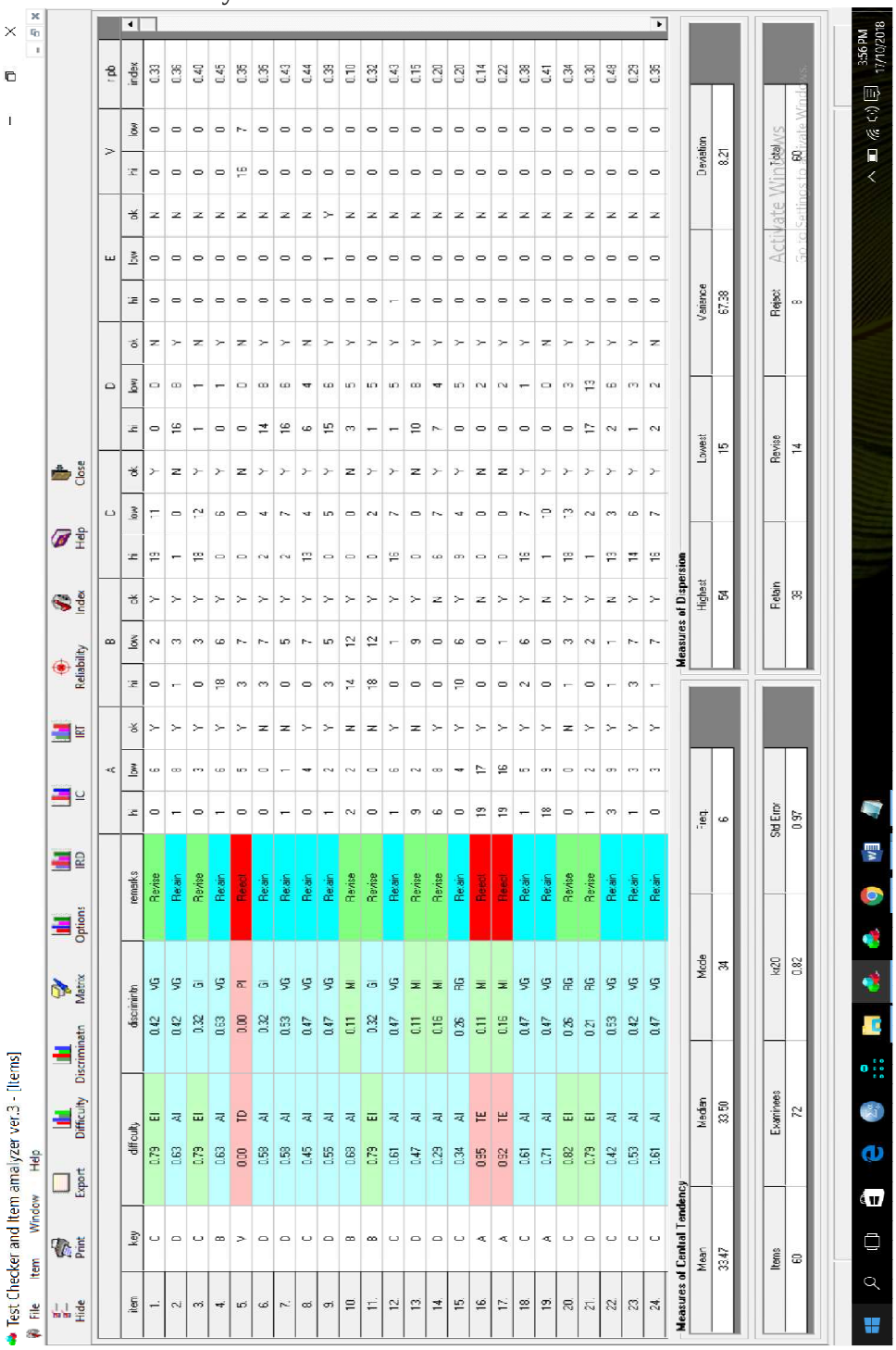
Yu, V. (2012). *Six ways continuous assessment can help students to learn*. New York. Knewton.

Zafar, M. (2008). Item Analysis Assumptions (Difficulty & Discrimination Indexes). Assessment Unit Dept. of Medical Education Ext. 47142.

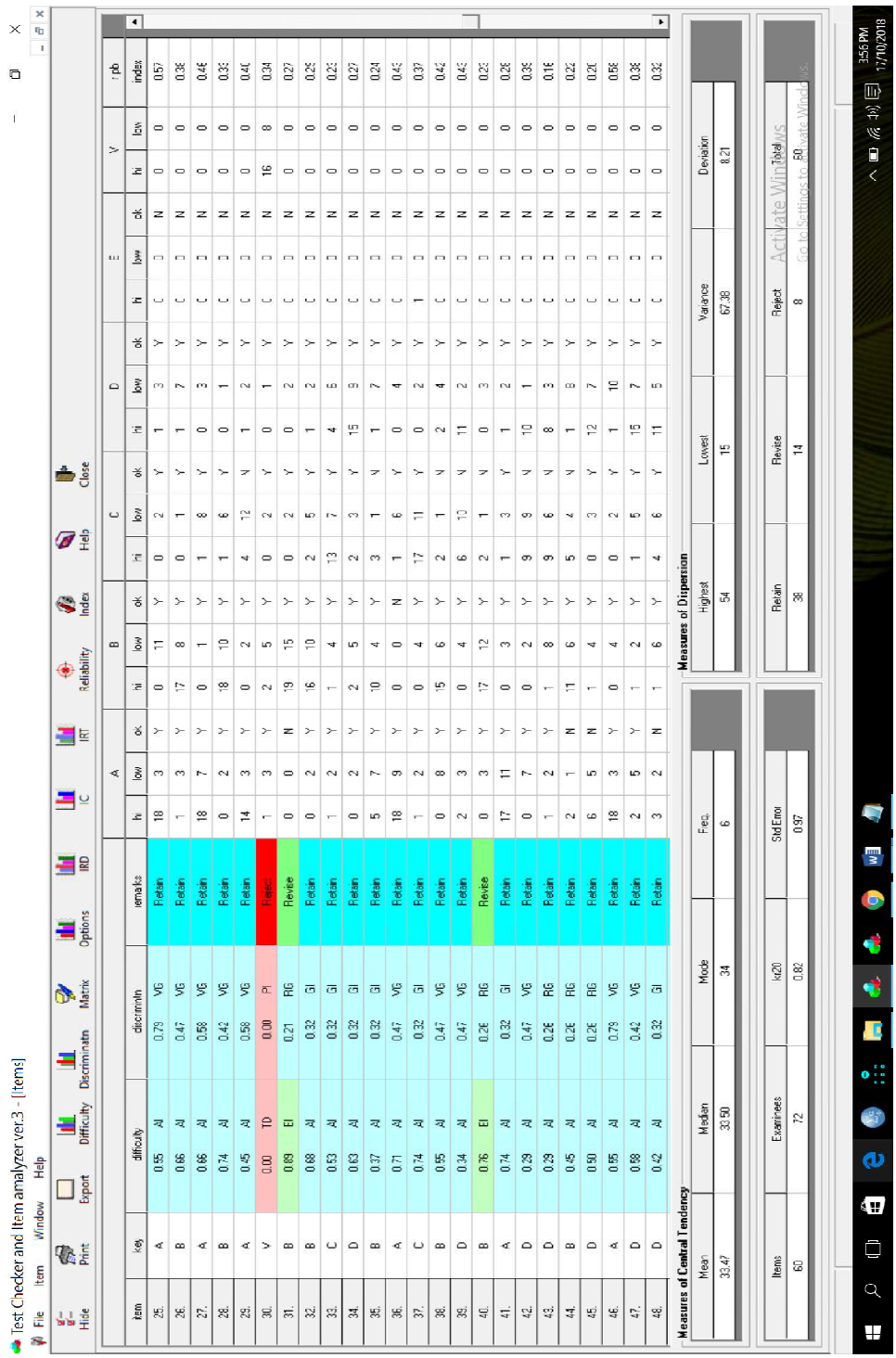
ISSN: 2278-4012, Volume:8, Issue:1, JANUARY 2019

## Appendices

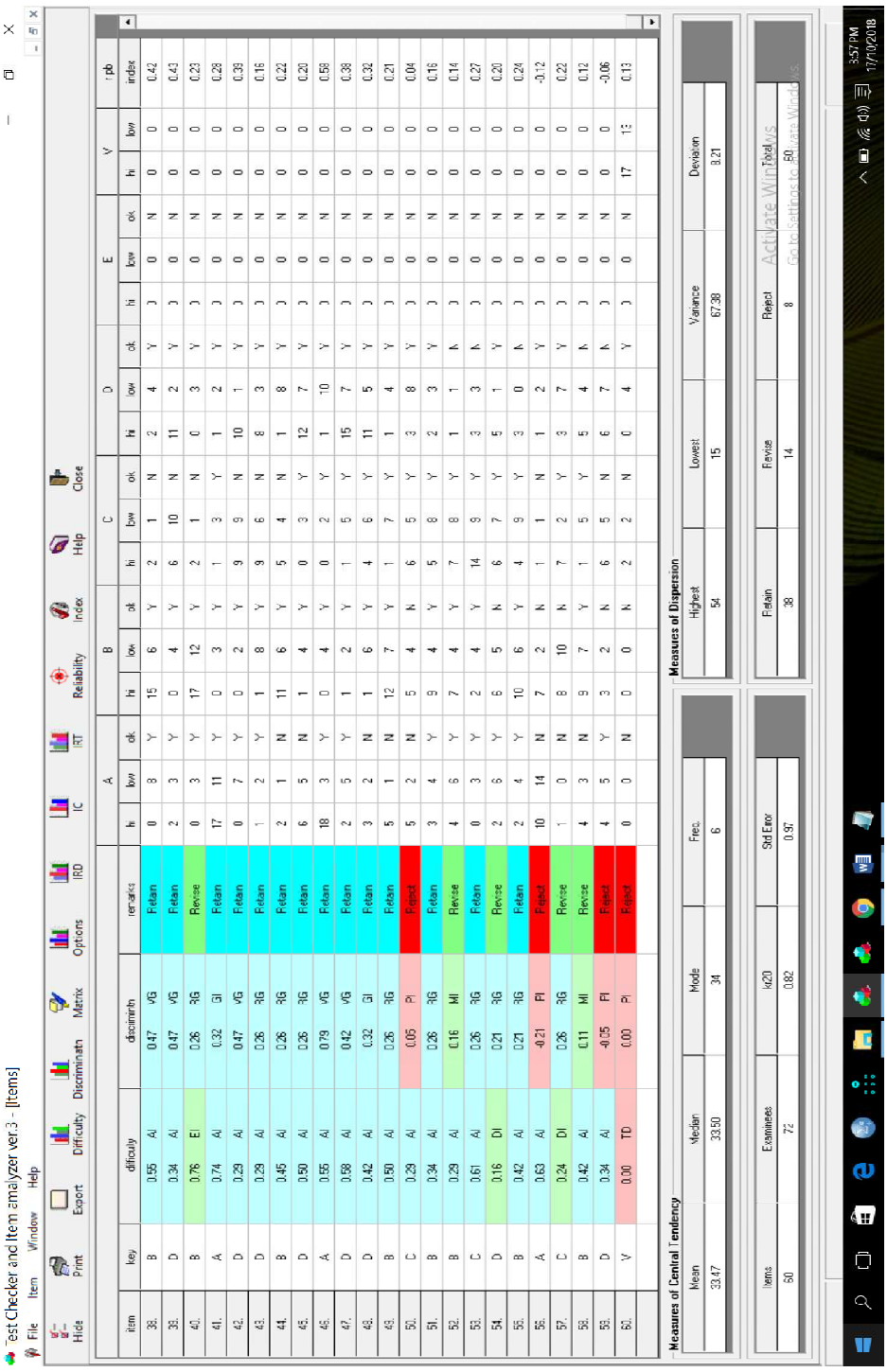
Appendix A. Distractor Analysis of the First Periodical Examination in Math 10



ISSN: 2278-4012, Volume:8, Issue:1, JANUARY 2019

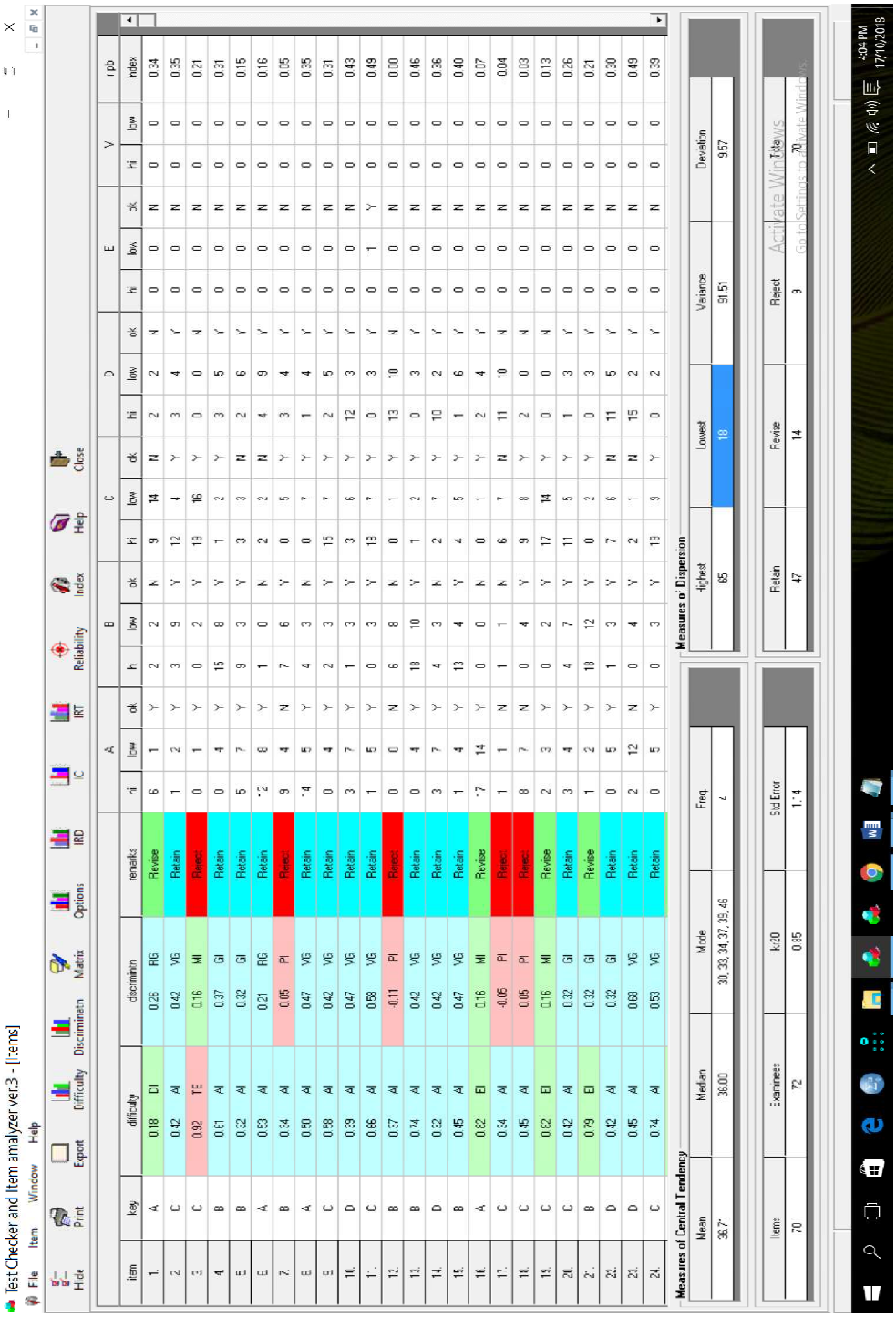


ISSN: 2278-4012, Volume:8, Issue:1, JANUARY 2019




Appendix B. Distractor Analysis of the Second Periodical Examination in Math 10

ISSN: 2278-4012, Volume:8, Issue:1, JANUARY 2019



ISSN: 2278-4012, Volume:8, Issue:1, JANUARY 2019