Assessing Student Performance Using Test Item Analysis and its Relevance to the State Exit Final Exams of MAT0024 Classes - An Action Research Project*

Dr. Mohammad Shakil

Department of Mathematics

Miami Dade College

Hialeah, FL 33012, USA; E-mail: mshakil@mdc.edu

Abstract

The classroom assessment and action research are the two most crucial components of the teaching and learning process. These are also essential parts of the scholarship of teaching and learning. Action Research is an important, recent development in classroom assessment techniques, defined as teacher-initiated classroom research which seeks to increase the teacher's understanding of classroom teaching and learning and to bring about improvements in classroom practices. Assessing the student performance is very important when the learning goals involve the acquisition of skills that can be demonstrated through action. Many researchers have worked and developed useful theories and taxonomies on the assessment of academic skills, intellectual development, and cognitive abilities of students, both from the analytical and quantitative point of view. Different kinds of assessments are appropriate in different settings. Item analysis is one powerful technique available to instructors for the guidance and improvement of instruction. In this project, student performance using test item analysis and its relevance to the State Exit Final Exams of MAT0024 classes have been investigated.

Keywords: Action Research, Discriminators, Discrimination Index, Item Analysis, Item Difficulty, Point-Biserial, Reliability.

***Part of this article was presented on MDC Conference Day, March 6th, 2008 at MDC, Kendall Campus.**

1. Introduction

Assessing student performance is very important when the learning goals involve the acquisition of skills that can be demonstrated through action. Many researchers have worked and developed useful theories and taxonomies (for example, Bloom's taxonomy) on the assessment of academic skills, intellectual development, and cognitive abilities of students, both from the analytical and quantitative point of view. For details on Bloom's cognitive taxonomy and its applications, see, for example, Bloom (1956), Ausbel (1968), Bloom et al. (1971), Simpson (1972), Krathwohl et al. (1973), Angelo & Cross (1993), and Mertler (2003), among others. Different kinds of assessments are appropriate in different settings. One of the most important and authentic techniques of assessing and estimating student performance across the full domain of learning outcomes as targeted by the instructor is the classroom test. Each item on a test is intended to sample student performance on a particular learning outcome. Thus, creating valid and reliable classroom tests are very important to an instructor for assessing student performance, achievement and success in the class. The same principle applies to the State Exit Exams and Classroom Tests conducted by the instructors, state and other agencies. Moreover, it is important to note that, most of the time, it is not well known whether the test items (e.g., multiple-choice) accompanied with the textbooks or test-generator software or constructed by the instructors are already tested for their validity and reliability. One powerful technique available to the instructors for the guidance and improvement of instruction is the test item analysis. It appears from the literature that, in spite of the extensive work on item analysis and its applications, very little attention has been paid to this kind of quantitative study of item analysis of state exit exams or classroom tests, particularly at Miami Dade College. After thorough search of the literature, the author of the present article has been able to find two references of this kind of study, that is, Hostetter & Haky (2005), and Hotiu (2006). Accordingly, in this project, student performance using test item analysis and its relevance to the State Exit Final Exams of MAT0024 classes have been investigated. By conducting the test item analysis of the State Exit Final Exams of some of my MAT0024 classes, this project discusses how well these exams distinguish among students according to the how well they met the learning goals of these classes. The data obtained from these exit exams are presented here as an item analysis report, which, it is hoped, will be helpful in recognizing the most critical pieces of the state exit test items data, and evaluating whether or not that test item needs revision. The organization of this paper is as follows. Section 2 discusses briefly 'what action research is'. In Section 3, an overview of some important statistical aspects of test item analysis is presented. Section 4 contains the test item analysis and other statistical analyses of the State Exit Final Exams of MAT0024 classes. Some conclusions are drawn in Section 5.

2. An Overview of Action Research

This section discusses briefly 'what action research is'.

2.1 What Is Action Research?

The development of the general idea of "action research" began with the work of Kurt Lewin (1946) in his paper entitled "Action Research and Minority Problems," where he describes action research as "a comparative research on the conditions and effects of various forms of social action and research leading to social action" that uses "a spiral of steps, each of which is composed of a circle of planning, action, and fact-finding about

the result of the action". Further development continued with the contributions by many other authors later, among them Kemmis (1983), Ebbutt (1985), Hopkins (1985), Elliott (1991), Richards et al. (1992), Nunan (1992), Brown (1994), and Greenwood et al. (1998), are notable. For recent developments on the theory of action research and its applications, the interested readers are referred to Brydon-Miller et al. (2003), Gustavsen (2003), Dick (2004), Elvin (2004), Barazangi (2006), Greenwood (2007), and Taylor & Pettit (2007), and references therein. As cited in Gabel (1995), following are some of the commonly used definitions of action research:

- Action Research aims to contribute both to the practical concerns of people in an immediate problematic situation and to the goals of social science by joint collaboration within a mutually acceptable ethical framework. (Rapoport, 1970).

- Action Research is a form of self-reflective enquiry undertaken by participants in social (including educational) situations in order to improve the rationality and justice of (a) their own social or educational practices, (b) their understanding of these practices, and (c) the situations in which the practices are carried out. It is most rationally empowering when undertaken by participants collaboratively... ...sometimes in cooperation with outsiders. (Kemmis, 1983).

- Action Research is the systematic study of attempts to improve educational practice by groups of participants by means of their own practical actions and by means of their own reflection upon the effects of those actions. (Ebbutt, 1985).

In the field of education, the term action research is defined as inquiry or research in the context of focused efforts in order to improve the quality of an educational institution and its performance. Typically, in an educational institution, the action research is designed and conducted by the instructors in their classes to analyze the data to improve their own teaching. It can be done by an individual instructor or by a team of instructors as a collaborative inquiry. Action research gives an instructor opportunities to reflect on and assess his/her teaching and its effectiveness by applying and testing new ideas, methods, and educational theory for the purpose of improving teaching, or to evaluate and implement an educational plan. According to Richards et al. (1992), action research is defined as teacher-initiated classroom research, which seeks to increase the teacher's understanding of classroom teaching and learning and to bring about improvements in classroom practices. Nunan (1992) defines it as a form of self-reflective inquiry carried out by practitioners, aimed at solving problems, improving practice, or enhancing understanding. According to Brown (1994), "Action research is any action undertaken by teachers to collect data and evaluate their own teaching. It differs from formal research, therefore, in that it is usually conducted by the teacher as a researcher, in a specific classroom situation, with the aim being to improve the situation or teacher rather than to spawn generalizeable knowledge. Action research usually entails observing, reflecting, planning and acting. In its simplest sense, it is a cycle of action and critical reflection, hence the name, action research."

2.2 My Action Research Project

There are many ways in which an instructor can exploit the classroom tests for assessing student performance, achievement and success in the class. It is one of the

most important and authentic techniques of assessing and estimating student performance across the full domain of learning outcomes as targeted by the instructor. One powerful technique available to an instructor for the guidance and improvement of instruction is the test item analysis. In this project, I have investigated student performance using test item analysis and its relevance to the State Exit Final Exams of MAT0024 classes. By conducting the test item analysis of the State Exit Final Exams of some of my MAT0024 classes, that is, Fall 2006-1, Spring 2006-2 and Fall 2007-1, this project discusses how well these exams distinguish among students according to the how well they met the learning goals of these classes. The data obtained from these exit exams are presented here as an item analysis report based upon the classical test theory (CRT), which is one of the important, commonly used types of Item Analysis. It is hoped that the present study would be helpful in recognizing the most critical pieces of the state exit test items data, and evaluating whether or not that test item needs revision. The methods discussed in this project can be used to describe the relevance of test item analysis to classroom tests. These procedures can also be used or modified to measure, describe and improve tests or surveys such as college mathematics placement exams (that is, CPT), mathematics study skills, attitude survey, test anxiety, information literacy, other general education learning outcomes, etc. Further research based on Bloom's cognitive taxonomy of test items (see, for example, the references as cited above), the applicability of Beta-Binomial models and Bayesian analysis of test items (see, for example, Duncan,1974; Gross & Shulman, 1980; Wilcox, 1981; and Gelman, 2006; among others), and item response theory (IRT) using the 1-parameter logistic model (also known as Rasch model), 2- & 3- parameter logistic models, plots of the item characteristic curves (ICCs) of different test items, and other characteristics of measurement instruments of IRT are under investigation by the present author and will be reported soon at an appropriate time. For details on IRT and recent developments, see, for example, Rasch (1960/1980), Lord & Novick (1968), Lord (1980), Wright (1992), Hambleton et al. (1991), Linden & Hambleton (1997), Thissen & Steinberg (1997), and Gleason (2008), among others.

3. An Overview of Test Item Analysis

In this section, an overview of test item analysis is presented.

3.1 Item Analysis

Item analysis is a process which examines student responses to individual test items (questions) in order to assess the quality of those items and of the test as a whole. It is a valuable, powerful technique available to teaching professionals and instructors for the guidance and improvement of instructions. It enables instructors to increase their test construction skills, identify specific areas of course content which need greater emphasis or clarity, and improve other classroom practices. According to Thompson & Levitov, (1985, p. 163), "Item analysis investigates the performance of items considered individually either in relation to some external criterion or in relation to the remaining items on the test." For example, when norm-referenced tests (NRTs) are developed for instructional purposes, such as placement test, or to assess the effects of educational programs, or for educational research purposes, it can be very important to conduct item and test analyses. Similarly, criterion-referenced tests (CRTs) compare students' performance to some preesablished criteria or objectives (such as classroom tests designed by the instructors). These analyses evaluate the quality of items and of the test as a whole. Such analyses can also be employed to revise and improve both items and

the test as a whole. Many researchers have contributed to the theory of test item analysis, among them Galton, Pearson, Spearman, and Thorndike are notable. For details on these pioneers of test item analysis theories and their contributions, see, for example, Gulliksen (1987), among others. For recent developments on the test item analysis practices, see Crocker & Algina (1986), Gronlund & Linn (1990), Pedhazur & Schemlkin (1991), Sax (1989), Thorndike, et al. (1991), Elvin (2003), and references therein.

3.2 Classical Test Theory (CTT)

An item analysis involves many statistics that can provide useful information for improving the quality and accuracy of multiple-choice or true/false items (questions). It describes the statistical analyses which allow measurement of the effectiveness of individual test items. An understanding of the factors which govern effectiveness (and a means of measuring them) can enable us to create more effective test questions and also regulate and standardize existing tests. The item analysis is an important phase in the development of an exam program. For example, a test or exam consisting of multiple-choice or true-false items is used to determine the proficiency (or ability) level of an examinee in a particular discipline or subject. Most of the times, the test or exam score obtained contributes a considerable weight in determining whether or not an examinee has passed or failed the subject. That is, the proficiency (or ability) level of an examinee is estimated using the total test score obtained from the number of correct responses to the test items. If the test score is equal to a cut-off score or greater than a cut-off score, then the examinee is considered to pass the subject, otherwise, it is considered a failure. This approach of using the test score as proficiency (or ability) estimate is called as the true score model (TSM) or classical test theory (CTT) approach. Classical Item Analysis, based on traditional classical theory models, forms the foundation for looking at the performance of each item in a test. The development of the CTT began with the work of Charles Spearman (1904) in his paper entitled “General intelligence: Objectively determined and measured”. Further development continued with the contributions by many researchers later, among them Francis Galton (1822 – 1911), Karl Pearson (1857 – 1936), and Edward Thorndike (1874 – 1949) are notable, (for details, see, for example, Nunnally, 1967; Gulliksen 1987; among others). For recent developments on the theory of CTT and its applications, the interested readers are referred to Chase (1999), Haladyna (1999), Nitko (2001), Tanner (2001), Oosterhof (2001), Mertler (2003), and references therein. The TSM equation is given by

$$X = T + \varepsilon \,,$$

where $X = observed\ score$, $T = true\ score$, $\varepsilon = random\ error$, and $E(X) = T$. Note that, in the above TSM equation, the true score reflects the exact value of the examinee’s ability or proficiency. Also, the TSM assumes that abilities (or traits) are constant and the variation in observed scores are caused by random errors, which may result from factors such as guessing, lack of preparation, or stress. Thus, in CTT, all test items and statistics are test-dependent. The trait (or ability) of an examinee is defined in terms of a test, whereas the difficulty of a test item is defined in terms of the group of examinees. According to Hambleton, et. al (1991, p. 3), “Examinee characteristics and test item characteristics cannot be separated: each can be interpreted only in the context

of the other." Some important criterias which are employed in the determination of the validity of a multiple-choice exam are following:

- Whether the test items were too difficult or too easy.
- Whether the test items discriminated between those examinees who really knew the material and those who did not.
- Whether the incorrect responses to a test item were distractors or non-distractors.

3.3 Item Analysis Statistics

An item analysis involves many statistics that can provide useful information for determining the validity and improving the quality and accuracy of multiple-choice or true/false items. These statistics are used to measure the ability levels of examinees from their responses to each item. The ParSCORE™ item analysis generated by Miami Dade College – Hialeah Campus Reading Lab when a multiple-choice MAT0024 State Exit Final Exam is machine scored consists of three types of reports, that is, a summary of test statistics, a test frequency table, and item statistics. The test statistics summary and frequency table describe the distribution of test scores, (for details on these, see, for example, Agresti and Finlay, 1997; Tamhane and Dunlop, 2000; among others). The item analysis statistics evaluate class-wide performance on each test item. The ParSCORE™ report on item analysis statistics gives an overall view of the test results and evaluates each test item, which are also useful in comparing the item analysis for different test forms. In what follows, descriptions of some useful, common item analysis statistics, that is, item difficulty, item discrimination, distractor analysis, and reliability, are presented below, (for details on these, see, for example, Wood, 1960; Lord & Novick, 1968; Henrysson, 1971; Nunally, 1978; Thompson & Levitov, 1985; Crocker & Algina, 1986; Ebel & Frisbie, 1986; Suen, 1990; Thorndike et al., 1991; DeVellis, 1991; Millman & Greene, 1993; Haladyna, 1999; Tanner, 2001; Haladyna et al., 2002; Mertler, 2003; among others). For the sake of completeness, definitions of some test statistics as reported in the ParSCORE™ analysis are also provided.

(I) Item Difficulty: Item difficulty is a measure of the difficulty of an item. For items (that is, multiple-choice questions) with one correct alternative worth a single point, the item difficulty (also known as the item difficulty index, or the difficulty level index, or the difficulty factor, or the item facility index, or the item easiness index, or the p-value) is defined as the proportion of respondents (examinees) selecting the answer to the item correctly, and is given by

$$p = \frac{c}{n}$$

where $p =$ the difficulty factor, $c =$ the number of respondents selecting the correct answer to an item, and $n =$ total number of respondents. Item difficulty is relevant for determining whether students have learned the concept being tested. It also plays an important role in the ability of an item to discriminate between students who know the tested material and those who do not. Note that

(i) $0 \le p \le 1$.

(ii) A higher value of p indicate low difficulty level index, that is, the item is easy. A lower value of p indicate high difficulty level index, that is, the item is difficult. In general, an ideal test should have an overall item difficulty of around 0.5; however it is acceptable for individual items to have higher or lower facility (ranging from 0.2 to 0.8). In a criterion-referenced test (CRT), with emphasis on mastery-testing of the topics covered, the optimal value of p for many items is expected to be 0.90 or above. On the other hand, in a norm-referenced test (NRT), with emphasis on discriminating between different levels of achievement, it is given by $p \approx 0.50$. For details on these, see, for example, Chase(1999), among others.

(iii) To maximize item discrimination, ideal (or moderate or desirable) item difficulty level, denoted as p_M, is defined as a point midway between the probability of success, denoted as p_S, of answering the multiple - choice item correctly (that is, 1.00 divided by the number of choices) and a perfect score (that is, 1.00) for the item, and is given by

$$p_M = p_S + \frac{1 - p_S}{2}.$$

(iv) Thus, using the above formula in (iv), ideal (or moderate or desirable) item difficulty levels for multiple-choice items can be easily calculated, which are provided in the following table, (for details, see, for example, Lord, 1952; among others).

Number of Alternatives	Probability of Success (p_S)	Ideal Item Difficulty Level (p_M)
2	0.50	0.75
3	0.33	0.67
4	0.25	0.63
5	0.20	0.60

(Ia) Mean Item Difficulty (or Mean Item Easiness): Mean item difficulty is the average of difficulty easiness of all test items. It is an overall measure of the test difficulty and ideally ranges between 60 % and 80 % (that is, $0.60 \le p \le 0.80$) for classroom achievement tests. Lower numbers indicate a difficult test while higher numbers indicate an easy test.

(II) Item Discrimination: The item discrimination (or the item discrimination index) is a basic measure of the validity of an item. It is defined as the discriminating power or the degree of an item's ability to discriminate (or differentiate) between high achievers (that is, those who scored high on the total test) and low achievers (that is, those who scored low), which are determined on the same criterion, that is, (1) internal criterion, for example, test itself; and (2) external criterion, for example, intelligence test or other achievement test. Further, the computation of the item discrimination index assumes that the distribution of test scores is normal and that there is a normal distribution underlying the right or wrong dichotomy of a student's performance on an item. For details on the item discrimination index, see, for example, Kelly (1939), Wood (1960), Henrysson (1971), Nunally (1972), Ebel (1979), Popham (1981), Ebel & Frisbie (1986), Weirsma & Jurs (1990), Glass & Hopkins (1995), Brown (1996), Chase (1999), Haladyna (1999), Nitko (2001), Tanner (2001), Oosterhof (2001), Haladyna et al. (2002), and Mertler (2003), among others. There are several ways to compute the item discrimination, but, as shown on the ParSCORE™ item analysis report and also as reported in the literature, the following formulas are most commonly used indicators of item's discrimination effectiveness.

(a) Item Discrimination Index (or Item Discriminating Power, or D-Statistics), D: Let the students' test scores be rank-ordered from lowest to highest. Let

$$p_U = \frac{No.\ of\ \ students\ in\ upper\ 25\% - 30\%\ group\ answering\ the\ item\ correctly}{Total\ \ Number\ of\ \ students\ in\ upper\ 25\% - 30\%\ group},$$

and

$$p_L = \frac{No.\ of\ \ students\ in\ lower\ 25\% - 30\%\ group\ answering\ the\ item\ correctly}{Total\ \ Number\ of\ \ students\ in\ lower\ 25\% - 30\%\ group}$$

The ParSCORE^{TM} item analysis report considers the upper 27% and the lower 27% as the analysis groups. The item discrimination index, D, is given by

$$D = p_U - p_L .$$

Note that

(i) $-1 \le D \le +1$.

(ii) Items with positive values of D are known as positively discriminating items, and those with negative values of D are known as negatively discriminating items.

(iii) If $D = 0$, that is, $p_U = p_L$, there is no discrimination between the upper and lower groups.

(iv) If $D = +1.00$, that is, $p_U = 1.00 \; and \; p_L = 0$, there is a perfect discrimination between the two groups.

(v) If $D = -1.00$, that is, $p_U = 0 \; and \; p_L = 1.00$, it means that all members of the lower group answered the item correctly and all members of the upper group answered the item incorrectly. This indicates the invalidity of the item, that is, the item has been miskeyed and needs to be rewritten or eliminated.

(vi) A guideline for the value of an item discrimination index is provided in the following table, see, for example, Chase(1999), and **Mertler(2003),** among others.

Item Discrimination Index, D	**Quality of an Item**
$D \ge 0.50$	Very Good Item; Definitely Retain
$0.40 \le D \le 0.49$	Good Item; Very Usable
$0.30 \le D \le 0.39$	Fair Quality; Usable Item
$0.20 \le D \le 0.29$	Potentially Poor Item; Consider Revising
$D < 0.20$	Potentially Very Poor; Possibly Revise Substantially, or Discard

(b) Mean Item Discrimination Index, D:

This is the average discrimination index for all test items combined. A large positive value (above 0.30) indicates good discrimination between the upper and lower scoring students. Tests that do not discriminate well are generally not very reliable and should be reviewed.

(c) Point-Biserial Correlation (or Item-Total Correlation or Item Discrimination) Coefficient, r_{pbis} : The point-biserial correlation coefficient is another item discrimination index of assessing the usefulness (or validity) of an item as a measure of individual differences in knowledge, skill, ability, attitude, or personality characteristic. It is defined as the correlation between the student performance on an item (correct or incorrect) and overall test score, and is given by either of the following two equations (which are mathematically equivalent).

(i) Suen (1990); DeVellis (1991); **Haladyna (1999)**

$$r_{pbis} = \left[\frac{\bar{X}_C - \bar{X}_T}{s} \right] \sqrt{\frac{p}{q}} ,$$

where $r_{pbis} =$ the point-biserial correlation coefficient; $\bar{X}_C =$ the mean total score for examinees who have answered the item correctly; $\bar{X}_T =$ the mean total score for all examines; $p =$ the difficulty value of the item; $q = 1 - p$; and $s =$ the standard deviation of total exam scores.

(ii) Brown (1996)

$$r_{pbis} = \left[\frac{m_p - m_q}{s} \right] \sqrt{p\,q} ,$$

where $r_{pbis} =$ the point-biserial correlation coefficient; $m_p =$ the mean total score for examinees who have answered the item correctly; $m_q =$ the mean total score for examinees who have answered the item incorrectly; $p =$ the difficulty value of the item; $q = 1 - p$; and $s =$ the standard deviation of total exam scores.

Note that

(i) The interpretation of the point-biserial correlation coefficient, r_{pbis} , is same as that of the D -statistic.

(ii) It assumes that the distribution of test scores is normal and that there is a normal distribution underlying the right or wrong dichotomy of a student performance on an item.

(iii) It is mathematically equivalent to the Pearson (product moment) correlation coefficient, which can be shown by assigning two distinct numerical values to the dichotomous variable (test item), that is, incorrect = 0 and correct = 1.

(iv) $-1 \le r_{pbis} \le +1$.

(v) $r_{pbis} \approx 0$ means little correlation between the score on the item and the score on the test.

(vi) A high positive value of r_{pbis} indicates that the examinees who answered the item correctly also received higher scores on the test than those examinees who answered the item incorrectly.

(vii) A negative value indicates that the examinees who answered the item correctly received low scores on the test and those examinees who answered the item incorrectly did better on the test. It is advisable that an item with $r_{pbis} \approx 0$ or with large negative value of r_{pbis} should be eliminated or revised. Also, an item with low positive value of r_{pbis} should be revised for improvement.

(viii) Generally, the value of r_{pbis} for an item may be put into two categories as provided in the following table.

Point-Biserial Correlation Coefficient, r_{pbis}	**Quality**
$r_{pbis} \ge 0.30$	Acceptable Range
$r_{pbis} \approx 1$	Ideal Value

(ix) The statistical significance of the point-biserial correlation coefficient, r_{pbis}, may be determined by applying the Student's t test, (for details, see, for example, Triola, 2007, among others).

Remark: It should be noted that the use of point-biserial correlation coefficient, r_{pbis}, is more advantageous than that of item discrimination index statistics, D, because every student taking the test is taken into consideration in the computation of r_{pbis}, whereas only 54 % of test-takers passing each item in both groups (that is, the upper 27 % + the lower 27 %) are used to compute D.

(d) Mean Item-Total Correlation Coefficient, r_{pbis} **:** It is defined as the average correlation of all the test items with the total score. It is a measure of overall test discrimination. A large positive value indicates good discrimination between students.

(III) Internal Consistency Reliability Coefficient (Kuder-Richardson 20, KR_{20}, Reliability Estimate): The statistic that measures the test reliability of inter-item consistency, that is, how well the test items are correlated with one another, is called the internal consistency reliability coefficient of the test. For a test, having multiple-choice items that are scored correct or incorrect, and that is administered only once, the Kuder-Richardson formula 20 (also known as KR-20) is used to measure the internal consistency reliability of the test scores (see, for example, Nunally, 1972; and Haladyna, 1999, among others). The KR-20 is also reported in the ParSCORE™ item analysis. It is given by the following formula:

$$KR_{20} = \frac{n\left(s^2 - \sum_{i=1}^{n} p_i\, q_i\right)}{s^2\,(n-1)}$$

where KR_{20} = the reliability index for the total test; n = the number of items in the test; s^2 = the variance of test scores; p_i = the difficulty value of the item; and $q_i = 1 - p_i$. Note that

(i) $0.0 \le KR_{20} \le 1.0$.

(ii) $KR_{20} \approx 0$ indicates a weaker relationship between test items, that is, the overall test score is less reliable. A large value of KR_{20} indicates high reliability.

(iii) Generally, the value of KR_{20} for an item may be put into the following categories as provided in the table below.

KR_{20}	**Quality**
$KR_{20} \ge 0.60$	Acceptable Range
$KR_{20} \ge 0.75$	Desirable
$0.80 \le KR_{20} \le 0.85$	Better t
$KR_{20} \approx 1$	Ideal Value

(iv) **Remarks:** The reliability of a test can be improved as follows:

a) By increasing the number of items in the test for which the following Spearman-Brown prophecy formula is used (Mertler, 2003).

$$r_{est} = \frac{n\, r}{1 + (n-1)r}$$

where r_{est} = the estimated new reliability coefficient; r = the original KR_{20} reliability coefficient; n = the number of times the test is lengthened.

b) Or, using the items that have high discrimination values in the test.

c) Or, performing an item-total statistic analysis as described above.

(IV) Standard Error of Measurement (SE_m): It is another important component of test item analysis to measure the internal consistency reliability of a test see, for example, Nunally, 1972; and Mertler, 2003, among others). It is given by the following formula:

$$SE_m = s\sqrt{1 - KR_{20}}\ ,\ 0.0 \le KR_{20} \le 1.0\,,$$

where SE_m = the standard error of measurement; s = the standard deviation of test scores; and KR_{20} = the reliability coefficient for the total test.

Note that

(i) $SE_m = 0$, when $KR_{20} = 1$.

(ii) $SE_m = 1$, when $KR_{20} = 0$.

(iii) A small value of SE_m (e.g., < 3) indicates high reliability; whereas a large value of SE_m indicates low reliability.

(iv) **Remark:** Higher reliability coefficient (i.e., $KR_{20} \approx 1$) and smaller standard deviation for a test indicate smaller standard error of measurement. This is considered to be more desirable situation for classroom tests.

(v) Test Item Distractor Analysis: It is an important and useful component of test item analysis. A test item distractor is defined as the incorrect response options in a multiple-choice test item. According to the research, there is a relationship between the quality of the distractors in a test item and the student performance on the test item, which also affect the student performance on his/her total test score. The performance of these incorrect item response options can be determined through the test item distractor analysis frequency table which contains the frequency, or number of students, that

selected each incorrect option. The test item distractor analysis is also provided in the ParSCORE^{TM} item analysis report. For details on test item distractor analysis, see, for example, Thompson & Levitov (1985), DeVellis (1991), Milman & Greene (1993), Haladyna (1999), and Mertler (2003), among others. A general guideline for the item distractor analysis is provided in the following table:

Item Response Options	**Item Difficulty** p	**Item Discrimination Index** D or r_{pbis}
Correct Response	$0.35 \le p \le 0.85$ (Better)	$D \ge 0.30$ or $r_{pbis} \ge 0.30$ (Better)
Distractors	$p \ge 0.02$ (Better)	$D \le 0$ or $r_{pbis} \le 0$ (Better)

(v) Mean: The mean is a measure of central tendency and gives the average test score of a sample of respondents (examinees), and is given by

$$\bar{x} = \frac{\sum_{i=1}^{n}(x_i)}{n},$$

where $x_i = individual\ test\ score$, $x_i = individual\ test\ score$, $n = no.\ of\ respondents$.

(vi) Median: If all scores are ranked from lowest to highest, the median is the middle score. Half of the scores will be lower than the median. The median is also known as the 50th percentile or the 2nd quartile.

(vii) Range of Scores: It is defined as the difference of the highest and lowest test scores. The range is a basic measure of variability.

(viii) Standard Deviation: For a sample of n examinees, the standard deviation, denoted by s, of test scores is given by the following equation

$$s = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n-1}},$$

where $x_i = individual\ test\ score$ and $\bar{x} = average\ test\ score$. The standard deviation is a measure of variability or the spread of the score distribution. It measures how far the scores deviate from the mean. If the scores are grouped closely together, the test will have a small standard deviation. A test with a large value of the standard deviation is considered better in discriminating the student performance levels.

(ix) Variance: For a sample of n examinees, the variance, denoted by s^2, of test scores is defined as the square of the standard deviation, and is given by the following equation

$$s^2 = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n-1}.$$

(x) Skewness: For a sample of n examinees, the skewness, denoted by β_3, of the distribution of the test scores is given by the following equation

$$\beta_3 = \frac{n}{(n-1)(n-2)}\left[\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s}\right)^3\right],$$

where $x_i = individual\ test\ score$, $\bar{x} = average\ test\ score$ and $s = s\tan dard\ deviation\ of\ test\ scores$. It measures the lack of symmetry of the distribution. The skewness is 0 for symmetric distribution and is negative or positive depending on whether the distribution is negatively skewed (has a longer left tail) or positively skewed (has a longer right tail).

(xi) Kurtosis: For a sample of n examinees, the kurtosis, denoted by β_4, of the distribution of the test scores is given by the following equation

$$\beta_4 = \left\{\frac{n(n+1)}{(n-1)(n-2)(n-3)}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s}\right)^4\right\} - \frac{3(n-1)^2}{(n-2)(n-3)},$$

where $x_i = individual\ test\ score$, $\bar{x} = average\ test\ score$, and $s = s\tan dard\ deviation\ of\ test\ scores$. It measures the tail-heaviness (the amount of probability in the tails). For the normal distribution, $\beta_4 = 3$. Thus, depending on whether $\beta_4 > 3\ or < 3$, a distribution is heavier tailed or lighter tailed than the normal distribution.

4. Results of the Research

This section consists of four parts, which are described below.

4.1 Test Item Analysis of 20071 MAT0024 Versions A and B State Exit Final Exams

An item analysis of the data obtained from my Fall 2007-1 MAT0024 class State Exit Final Exam Items (Versions A and B) is presented here based upon the classical test theory (CRT). Various test item statistics and relevant statistical graphs (for both test forms, Versions A and B) using the ParSCORE^{TM} item analysis report and the Minitab software are computed and summarized in the Tables 1 – 5 below. Each version consisted of 30 items. There were two different groups of 7 students for each version.

- It appears from these statistical analyses that a large value of $KR_{20} = 0.90\ (\approx 1)$ for Version B indicates its high reliability in comparison to Version A, which is also substantiated by large positive values of $Mean\ DI = 0.450 > 0.3$ and $Mean\ Pt.\ Bisr. = 0.4223$, small value of standard error of measurement (that is, $SEM = 1.82$), and an ideal value of mean (that is, $\mu = 19.57 > 18$, the passing score) for Version B. These analyses are also evident by the bar charts and scatter plots drawn for various test item statistics using Minitab, that is, item difficulty (p), item discrimination index (D) and point-biserial correlation coefficient (r_{pbis}), which are presented below in Figures 1 and 2.

- The results indicate a definite correlation between item difficulty level and item discrimination index. For example, as the item difficulty level increases, the item discrimination index (D or r) also increases. However, there is an optimum level of item difficulty level, that is, 40 % - 70 % in Version A and 40 % - 50 % in Version B, after which the item discrimination index (D or r) starts decreasing. This means that the test items were too difficult in these ranges for both the high scorers and the low scorers, and did not have a good and effective discriminating power.

- **Filter for Selecting, Rejecting and Modifying Test Items:** The analysis also indicated two extremes, that is, the test items which were too easy (with item difficulty level as 100 %) and too difficult (with item difficulty level as 0 %). This implies that these test items did not have the effective discriminating power between students of different abilities (that is, between high achievers and low achievers). This process may be used for the selection, rejection and modification of test items (Figures 1 and 2).

Table 1

A Comparison of 20071 MAT0024 Ver. A and B State Exit Test Items

Exam. Version	Reliability $KR-20$	Mean	SD	SEM	$p < 0.3$	$0.3 \le p \le 0.7$	$p > 0.7$	$D > 0.2$
A	0.53	17.14	2.80	1.92	8	10	12	14
B	0.90	19.57	5.75	1.82	1	15	14	20

Exam. Version	*Mean DI*	*Mean Pt. Bisr.*
A	0.233	0.2060
B	0.450	0.4223

Table 2

MAT0024_2007_1_Ver_A

Data Display

Row	PU	PL	Disc. Ind. (D)	Difficulty (p)	Difficulty (p) %	Pt-Bis (r)
1	1.0	0.0	1.0	0.4286	42.86	0.78
2	1.0	1.0	0.0	0.8571	85.71	0.02
3	1.0	0.5	0.5	0.8571	85.71	0.46
4	1.0	0.0	1.0	0.5714	57.14	0.66
5	1.0	0.0	1.0	0.5714	57.14	0.77
6	1.0	0.0	1.0	0.7143	71.43	0.82
7	0.5	0.0	0.5	0.5714	57.14	0.56
8	1.0	1.0	0.0	1.0000	100.00	0.00
9	0.0	0.5	-0.5	0.1429	14.29	-0.46
10	0.5	0.5	0.0	0.4286	42.86	0.27
11	0.5	0.5	0.0	0.4286	42.86	-0.15
12	1.0	1.0	0.0	1.0000	100.00	0.00
13	1.0	1.0	0.0	1.0000	100.00	0.00
14	0.0	0.0	0.0	0.0000	0.00	0.00
15	1.0	0.5	0.5	0.5714	57.14	0.25
16	1.0	0.5	0.5	0.7143	71.43	0.37
17	1.0	0.5	0.5	0.8571	85.71	0.60
18	1.0	1.0	0.0	1.0000	100.00	0.00
19	1.0	1.0	0.0	1.0000	100.00	0.00
20	1.0	0.5	0.5	0.8571	85.71	0.46
21	1.0	0.5	0.5	0.8571	85.71	0.46
22	0.5	0.5	0.0	0.5714	57.14	-0.16
23	0.0	0.5	-0.5	0.1429	14.29	-0.46
24	0.5	1.0	-0.5	0.5714	57.14	-0.27
25	0.0	0.0	0.0	0.2857	28.57	0.08
26	0.0	0.0	0.0	0.1429	14.29	-0.02
27	1.0	0.5	0.5	0.4286	42.86	0.37
28	0.5	0.0	0.5	0.1429	14.29	0.71
29	0.5	0.0	0.5	0.2857	28.57	0.53
30	0.0	0.5	-0.5	0.1429	14.29	-0.46

Table 3

Descriptive Statistics: MAT0024_2007_1_Ver_A

Variable	Mean	SE Mean	StDev	Variance	Minimum	Q1
Disc. Ind. (D)	0.2333	0.0821	0.4498	0.2023	-0.5000	0.000000000
Difficulty (p)	0.5714	0.0573	0.3139	0.0985	0.000000000	0.2857
Difficulty (p) %	57.14	5.73	31.39	985.11	0.000000000	28.57
Pt-Bis (r)	0.2063	0.0703	0.3850	0.1482	-0.4600	-0.00500

Variable	Median	Q3	Maximum
Disc. Ind. (D)	0.000000000	0.5000	1.0000
Difficulty (p)	0.5714	0.8571	1.0000
Difficulty (p) %	57.14	85.71	100.00
Pt-Bis (r)	0.1650	0.5375	0.8200

❖ Filter for Selecting, Rejecting and Modifying Test Items (Figure 1)

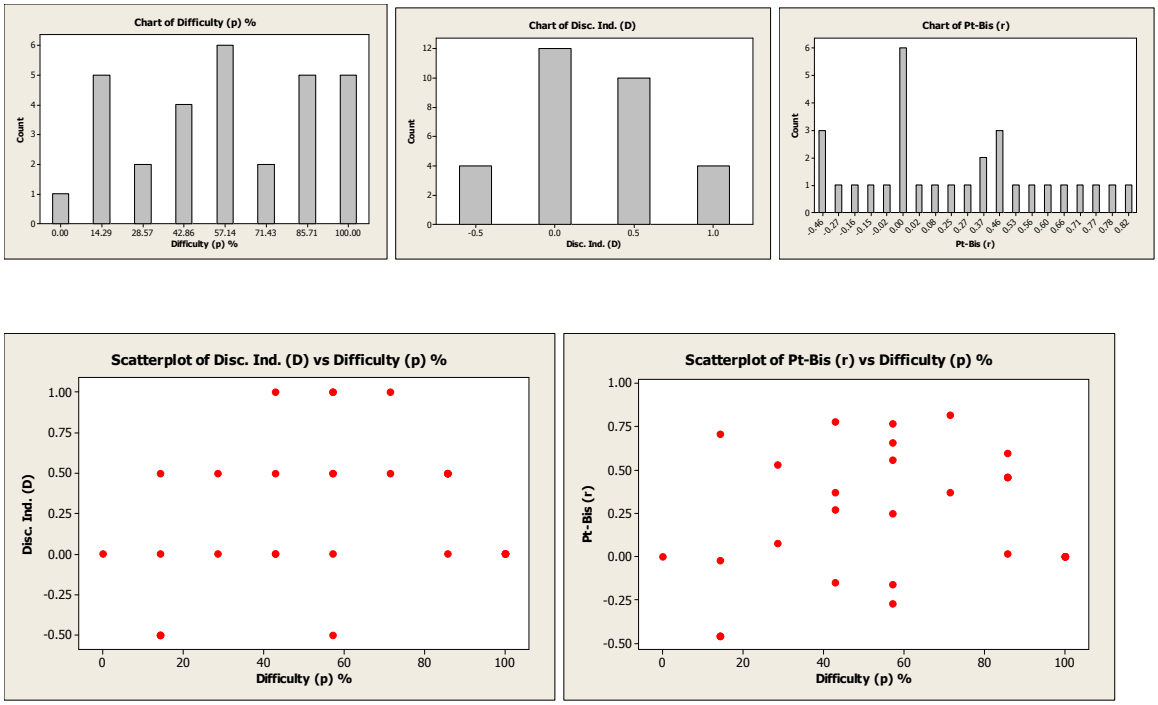


Figure 1

(Bar Charts and Scatter Plots for p, D, and r_{pbis}, Version A)

Table 4

MAT0024_2007_1_Ver_B

Data Display

Row	PU	PL	Disc. Ind. (D)	Difficulty (p)	Difficulty (p) %	Pt-Bis (r)
1	1.0	1.0	0.0	1.0000	100.00	0.00
2	1.0	1.0	0.0	0.7143	71.43	0.06
3	1.0	1.0	0.0	1.0000	100.00	0.00
4	1.0	1.0	0.0	0.8571	85.71	0.11
5	1.0	0.5	0.5	0.8571	85.71	0.54
6	1.0	0.5	0.5	0.7143	71.43	0.67
7	1.0	0.0	1.0	0.4286	42.86	0.92
8	1.0	0.5	0.5	0.4286	42.86	0.37
9	0.5	0.5	0.0	0.4286	42.86	0.42
10	1.0	0.0	1.0	0.4286	42.86	0.92
11	1.0	0.5	0.5	0.5714	57.14	0.69
12	1.0	1.0	0.0	1.0000	100.00	0.00
13	1.0	0.5	0.5	0.8571	85.71	0.32
14	0.5	0.0	0.5	0.4286	42.86	0.37
15	1.0	0.5	0.5	0.5714	57.14	0.54
16	0.5	0.0	0.5	0.5714	57.14	0.34
17	1.0	0.0	1.0	0.5714	57.14	0.69
18	1.0	1.0	0.0	1.0000	100.00	0.00
19	1.0	1.0	0.0	1.0000	100.00	0.00
20	1.0	0.5	0.5	0.8571	85.71	0.54
21	0.5	1.0	-0.5	0.8571	85.71	-0.39
22	1.0	0.5	0.5	0.7143	71.43	0.67
23	0.5	0.0	0.5	0.1429	14.29	0.67
24	1.0	0.0	1.0	0.4286	42.86	0.92
25	1.0	0.0	1.0	0.5714	57.14	0.44
26	1.0	0.0	1.0	0.4286	42.86	0.67
27	1.0	0.5	0.5	0.7143	71.43	0.06
28	0.5	0.0	0.5	0.1429	14.29	0.67
29	1.0	0.5	0.5	0.8571	85.71	0.54
30	1.0	0.0	1.0	0.4286	42.86	0.92

Table 5

Descriptive Statistics: MAT0024_2007_1_Ver_B

Variable	Mean	SE Mean	StDev	Variance	Minimum	Q1
Disc. Ind. (D)	0.4500	0.0733	0.4015	0.1612	-0.5000	0.000000000
Difficulty (p)	0.6524	0.0458	0.2508	0.0629	0.1429	0.4286
Difficulty (p) %	65.24	4.58	25.08	628.81	14.29	42.86
Pt-Bis (r)	0.4223	0.0628	0.3440	0.1183	-0.3900	0.0600

Variable	Median	Q3	Maximum
Disc. Ind. (D)	0.5000	0.6250	1.0000
Difficulty (p)	0.6429	0.8571	1.0000
Difficulty (p) %	64.29	85.71	100.00
Pt-Bis (r)	0.4900	0.6700	0.9200

❖ **Filter for Selecting, Rejecting and Modifying Test Items (Figure 2)**

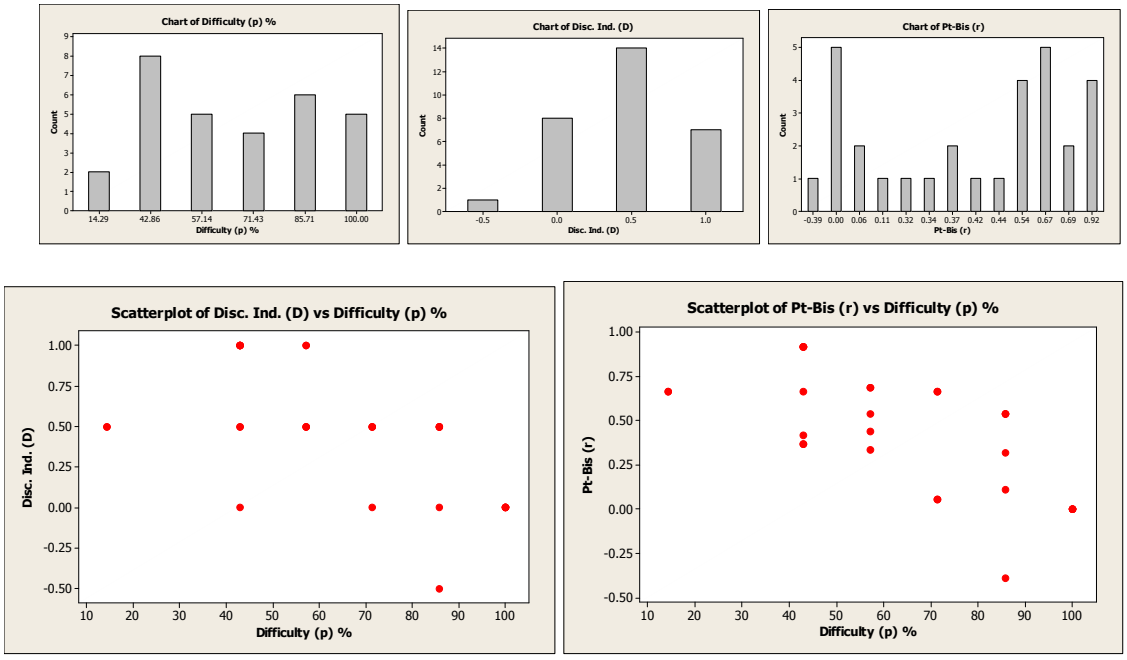


Figure 2

(**Bar Charts and Scatter Plots for** p **,** D **, and** r_{pbis} **, Version B)**

4.2 A Comparison of 2007-1 MAT0024 Ver. A and B State Exit Exams Performance

A Two-Sample T-Test: To identify if there is a significant difference between the 2007-1 MAT0024 Versions A and B state exit exams performance of the students, a two-sample T-test was conducted using the Minitab and Statdisk software. For this, first the assumption of normality was checked using the histograms and Anderson-Darling Test for both groups. The results are provided in the Tables 6 -7 and Figures 3-4 below. It is evident that the normality tests are easily met. Moreover, at the significance level of $\alpha = 0.05$, the two-sample T-test conducted fails to reject the claim that $\mu_A = \mu_B$, that is, the sample does not provide enough evidence to reject the claim.

Figure 3

(Anderson-Darling Normality Tests for 2007-1 MAT0024 A & B Exit Exam Scores)

Figure 4

(Two-Sample T-Test for 2007-1 MAT0024 A & B Exit Exam Scores)

Table 6

Descriptive Statistics: 2007-1A, 2007-1B

Variable	Total Count	N	Mean	SE Mean	StDev	Variance	Minimum	Q1	Median
2007-1A	7	7	17.14	1.14	3.02	9.14	13.00	14.00	17.00
2007-1B	7	7	19.57	2.35	6.21	38.62	12.00	15.00	18.00

Variable	Q3	Maximum	Skewness	Kurtosis
2007-1A	19.00	22.00	0.16	-0.03
2007-1B	25.00	29.00	0.40	-1.31

Table 7

Two-Sample T-Test and CI: 2007-1A, 2007-1B (Assume Unequal Variances)

```
Two-sample T for 2007-1A vs 2007-1B

          N   Mean  StDev  SE Mean
2007-1A   7  17.14   3.02      1.1
2007-1B   7  19.57   6.21      2.3


Difference = mu (2007-1A) - mu (2007-1B)
Estimate for difference:  -2.42857
95% CI for difference:  (-8.45211, 3.59497)
T-Test of difference = 0 (vs not =): T-Value = -0.93  P-Value = 0.380  DF = 8
```

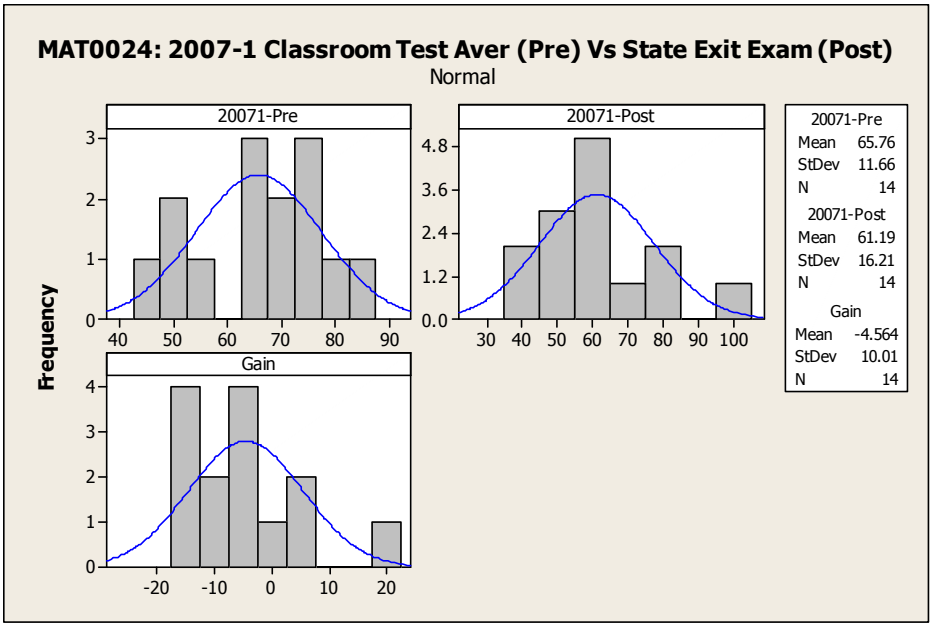
Two-Sample T-Test and CI: 2007-1A, 2007-1B (Assume Equal Variances)

```
Two-sample T for 2007-1A vs 2007-1B

          N   Mean  StDev  SE Mean
2007-1A   7  17.14   3.02      1.1
2007-1B   7  19.57   6.21      2.3


Difference = mu (2007-1A) - mu (2007-1B)
Estimate for difference:  -2.42857
95% CI for difference:  (-8.11987, 3.26273)
T-Test of difference = 0 (vs not =): T-Value = -0.93  P-Value = 0.371  DF = 12
Both use Pooled StDev = 4.8868
```

4.3 A Comparison of 2007-1 MAT0024 Classroom Test Aver (Pre) Vs State Exit Exam (Post) Performance

A Paired Samples T-Test: To identify if there is a significant gain in the 2007-1 MAT0024 posttest (state exit exam) compared to the pretest (classroom test Average) performance of the students, a paired samples T-test was conducted using the Minitab and Statdisk software. For this, first the assumption of normal distribution of the post, pre, and gain (post – pre) scores was checked using the histograms (see Figure 5). The histograms suggest that the distributions are close to normal. To check whether normality assumption for a paired samples t-test is met, the Kolmogorov-Smirnov and Shapiro-Wilk tests for the gain scores were conducted using Minitab. The results are provided in the Tables 8 -10 and Figure 5 below. It is evident that the normality tests are easily met. Moreover, at the significance level of $\alpha = 0.05$, the paired samples T-test conducted fails to reject the claim that $\mu_A = \mu_B$, that is, the sample does not provide enough evidence to reject the claim.

HISTOGRAMS



KOLMOGOROV-SMIRNOV TEST

SHAPIRO-WILK TEST

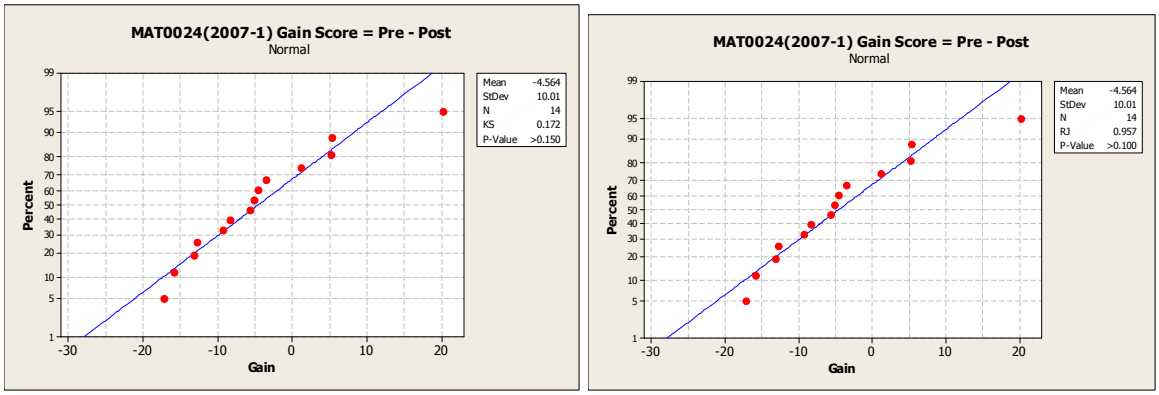


Figure 5

TESTS FOR NORMALITY
(MAT0024: 2007-1 Classroom Test Aver (Pre) Vs State Exit Exam (Post))

MAT0024 (2007-1)

Paired T-Test and CI: 20071-Post, 20071-Pre (Gain Score = Post – Pre)

Hypothesis Test for the Mean Difference: Matched Pairs

Figure 7

(Paired Samples T-Test: MAT0024 2007-1 Pre Vs Post (State Exit Exam)

Table 8

Data Display: MAT0024 (2007-1)
20071-Post, 20071-Pre (Gain Score = Post – Pre)

Row	20071-Pre	20071-Post	Gain
1	69.4	56.7	-12.7
2	63.2	50.0	-13.2
3	54.8	60.0	5.2
4	78.0	83.3	5.3
5	75.6	76.7	1.1
6	66.8	63.3	-3.5
7	51.8	46.7	-5.1
8	44.6	40.0	-4.6
9	72.6	56.7	-15.9
10	68.4	60.0	-8.4
11	67.2	50.0	-17.2
12	76.6	96.7	20.1
13	82.6	73.3	-9.3
14	49.0	43.3	-5.7

Table 9

MAT0024 (2007-1)

Descriptive Statistics: 20071-Post, 20071-Pre (Gain Score = Post – Pre)

Variable	Total Count	N	Mean	SE Mean	StDev	Variance	Minimum	Q1	Median
20071-Post	14	14	61.19	4.33	16.21	262.62	40.00	49.18	58.35
20071-Pre	14	14	65.76	3.12	11.66	136.01	44.60	54.05	67.80
Gain	14	14	-4.56	2.67	10.01	100.14	-17.20	-12.83	-5.40

Variable	Q3	Maximum	Range	IQR	Skewness	Kurtosis
20071-Post	74.15	96.70	56.70	24.98	0.84	0.22
20071-Pre	75.85	82.60	38.00	21.80	-0.51	-0.80
Gain	2.13	20.10	37.30	14.95	1.10	1.56

Table 10

MAT0024 (2007-1)

Paired T-Test and CI: 20071-Post, 20071-Pre (Gain Score = Post – Pre)

Paired T for 20071-Post - 20071-Pre

	N	Mean	StDev	SE Mean
20071-Post	14	61.1929	16.2056	4.3311
20071-Pre	14	65.7571	11.6622	3.1169
Difference	14	-4.56429	10.00704	2.67450

95% CI for mean difference: (-10.34218, 1.21361)
T-Test of mean difference = 0 (vs not = 0): T-Value = -1.71 P-Value = 0.112

4.4 A Comparison of MAT0024: 2006-1, 2006-2, 2007-1 State Exit Exams

To identify if there is a significant difference in the MAT0024: 2006-1, 2006-2, 2007-1 State Exit Exams performance of the students, one-way analysis of variance was conducted using the Minitab and Statdisk software. For this, first the assumption of normality was checked using the histograms and Anderson-Darling Test for the three groups. The results are provided in the Tables 11 -12 and Figures 7-9 below. It is evident that the normality tests are easily met. Moreover, at the significance level of $\alpha = 0.05$, the data does not provide enough evidence to indicate the claim that the sample means are unequal.

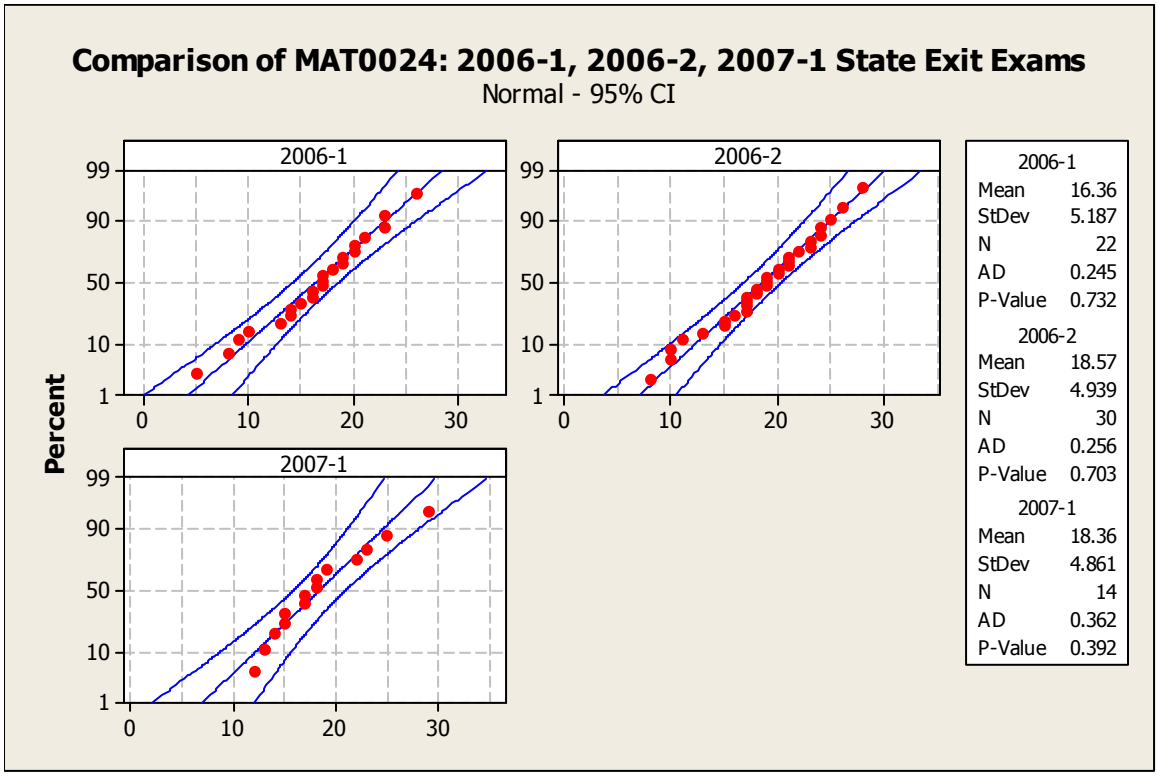


Figure 7

(Normality Tests: MAT0024 2006-1, 2006-2, 2007-1 State Exit Exams)

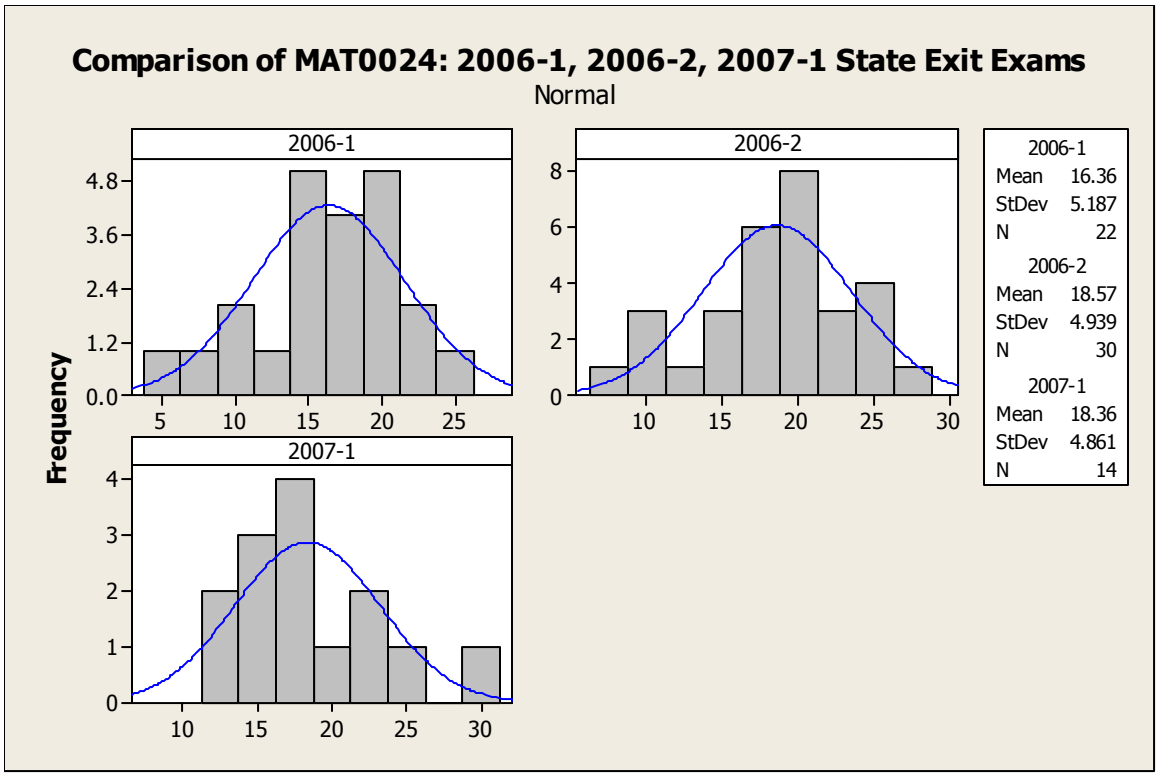


Figure 8

(Normality Tests: MAT0024 2006-1, 2006-2, 2007-1 State Exit Exams)

Table 11

One-way ANOVA
MAT0024: 2006-1, 2006-2, 2007-1 State Exit Exams

```
Source  DF      SS     MS     F      P
Factor   2    67.4   33.7  1.34  0.268
Error   63  1579.7   25.1
Total   65  1647.0

S = 5.007   R-Sq = 4.09%   R-Sq(adj) = 1.04%

                          Individual 95% CIs For Mean Based on
                          Pooled StDev
Level    N    Mean  StDev  ---------+---------+---------+---------+
2006-1  22  16.364  5.187  (----------*---------)
2006-2  30  18.567  4.939            (--------*--------)
2007-1  14  18.357  4.861       (-------------*------------)
                          ---------+---------+---------+---------+
                                16.0      18.0      20.0      22.0

Pooled StDev = 5.007
```

Table 12

One-way ANOVA (Analysis of Variance)
MAT0024: 2006-1, 2006-2, 2007-1 State Exit Exams

One-way Analysis of Variance: Hypothesis Test

Figure 9

(One-way ANOVA: MAT0024 2006-1, 2006-2, 2007-1 State Exit Exams)

5. Concluding Remarks

- This paper discusses the classroom assessment and action research, which are the two most crucial components of the teaching and learning process. Student performance using test item analysis and its relevance to the State Exit Final Exams of MAT0024 classes have been investigated. By conducting the test item analysis of the State Exit Final Exams of some of my MAT0024 classes, this project discusses how well these exams distinguish among students according to the how well they met the learning goals of these classes.
- It is hoped that the present study would be helpful in recognizing the most critical pieces of the state exit test items data, and evaluating whether or not that test item needs revision. The methods discussed in this project can be used to describe the relevance of test item analysis to classroom tests. These procedures can also be used or modified to measure, describe and improve tests or surveys such as college mathematics placement exams (that is, CPT), mathematics study skills, attitude survey, test anxiety, information literacy, other general education learning outcomes, etc.

- Further research based on Bloom's cognitive taxonomy of test items, the applicability of Beta-Binomial models and Bayesian analysis of test items and item response theory (IRT) using the 1-parameter logistic model (also known as Rasch model), 2- & 3- parameter logistic models, plots of the item characteristic curves (ICCs) of different test items, and other characteristics of measurement instruments of IRT are under investigation by the present author and will be reported soon at an appropriate time.

- Finally, this action research project has given me new directions about the needs of my students in MAT0024 and other mathematics classes. It has helped me to know about their learning styles, individual differences and ability. It has also given me insight to construct valid and reliable tests & exams for more student successes and achievements in my math classes. This action research project has provided me inputs to coordinate with my colleagues in mathematics and other disciplines at the Hialeah Campus as well as college wide to identify methods to improve classroom practices through test item analysis and action research in order to enhance the student success and achievement in the class and, later, in their lives, which are also the MDC QEP and General Education Learning Outcomes.

Acknowledgments

I would like to express my sincere gratitude and thanks to the LAS Chair, Dr. Cary Castro, the Academic Dean, Dr. Ana Maria Bradley-Hess, and the President, Dr. Cindy Miles, of Miami-Dade College, Hialeah Campus, for their continued encouragement, support and patronage. I would like to thank the Hialeah Campus College Prep Lab Coordinator, Professor Javier Duenas and the Lab Instructor, Mr. Cesar Rueda, for their kind support and cooperation in providing me with the ParSCORE™ item analysis reports on the MAT0024 State Exit Final Exams. I'm also thankful to Dr. Hanadi Saleh, MDC CT & D Instructional Designer/Trainer, Hialeah Campus, for her valuable and useful comments, suggestions, and contributions to the Power Point which considerably improved the quality of this presentation. I would also like to acknowledge my sincere indebtedness to the works of various authors and resources on the subject which I have consulted during the preparation of this research project. Last but not the least I am thankful to the authorities of Miami-Dade College for allowing and giving me an opportunity to present this paper on MDC Conference Day.

References

Angelo, T. A. and Cross, K. P. (1993). *Classroom Assessment Techniques – A Handbook for College Teachers.* Jossey-Bass, San Francisco.

Agresti, A. and Finlay, B. (1997). *Statistical Methods for the Social Sciences*. Prentice Hall, Upper Saddle River, NJ.

Ausubel, D. P. (1968). *Educational Psychology: A Cognitive View*. Holt, Reinhart & Winston, Troy, Mo.

Barazangi, N. H. (2006). An ethical theory of action research pedagogy. *Action Research*, **4(1)**, 97-116.

Bloom, B. S. (1956). *Taxonomy of Educational Objectives, Handbook I: The Cognitive Domain*. David McKay Co., Inc., New York.

Bloom, B. S., Hastings, J. T. and Madaus, G. F. (1971). *Handbook on Formative and Summative Evaluation of Student Learning*. McGraw-Hill, New York.

Brown, H. D. (1994). *Teaching by Principles: An Interactive Approach to Language Pedagogy*. Prentice Hall, Englewood Cliffs, NJ.

Brown, J. D. (1996). *Testing in language programs*. Prentice Hall, Upper Saddle River, NJ.

Brydon-Miller, M., Greenwood, D. and Maguire, P. (2003). Why Action Research?. *Action Research*, **1(1**), 9-28.

Chase, C. I. (1999). *Contemporary assessment for educators*. Longman, New York.

Crocker, L. and Algina, J. (1986). *Introduction to classical and modern test theory*. Holt, Rinehart and Winston, New York.

DeVellis, R. F. (1991). *Scale development: Theory and applications*. Sage Publications, Newbury Park.

Dick, B. (2004). Action research literature: Themes and trends. *Action Research*, **2(4)**, 425-444.

Duncan, G. T. (1974). An empirical Bayes approach to scoring multiple-choice tests in the misinformation model. *Journal of the American Statistical Association*, **69(345)**, 50-57.

Ebel, R.L. (1979). *Essentials of educational measurement* (3rd ed). Prentice Hall, Englewood Cliffs, NJ.

Ebel, R. L. and Frisbie, D. A. (1986). *Essentials of educational measurement*. Prentice-Hall, Inc, Englewood Cliffs, NJ.

Ebbutt (1985). Educational action research: Some general concerns and specific quibbles. *In Burgess R (ed.) 'Issues in educational research: Qualitative methods'*. Falmer Press, Lewes.

Elliott, J. (1991). *Action research for educational change*. Open University Press, Philadelphia.

Elvin, C. (2003). Test Item Analysis Using Microsoft Excel Spreadsheet Program. *The Language Teacher*, **27 (11)**, 13-18

Elvin, C. (2004). My Students' DVD Audio and Subtitle Preferences for Aural English Study: An Action Research Project. *Explorations in Teacher Education*, **12 (4)**, 3-17.

Gabel, D (1995). An Introduction to Action Research.
http://physicsed.buffalostate.edu/danowner/actionrsch.html

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, **1(3)**, 515-533.

Glass, G. V. and Hopkins, K. D. (1995). *Statistical Methods in Education and Psychology*, *3rd edition*, Allyn & Bacon, Boston.

Gleason, J. (2008). An evaluation of mathematics competitions using item response theory. *Notices of the AMS*, 55(1), 8-15.

Greenwood, D. J. and Lewin, M. (1998), *Introduction to Action Research*, Sage, London.

Greenwood, D. J (2007). Teaching/learning action research requires fundamental reforms in public higher education**.** *Action Research*, **5(3)**, 249-264.

Gronlund, N.E., & Linn, R.L. (1990). *Measurement and evaluation in teaching (6th ed*). MacMillan, New York.

Gross, A. L. and Shulman, V. (1980). The applicability of the beta binomial model for criterion referenced testing. Journal of Educational Measurement, **17(3)**, 195-201.

Gulliksen, H. (1987). *Theory of mental tests*. Erlbaum, Hillsdale, NJ.

Gustavsen, B. (2003). New Forms of Knowledge Production and the Role of Action Research. *Action Research*, **1(2)**, 153-164.

Haladyna. T. M. (1999). *Developing and validating multiple-choice exam items, 2nd ed*. Lawrence Erlbaum Associates, Mahwah, NJ.

Haladyna, T. M., Downing, S.M. and Rodriguez, M.C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, **15(3)**, 309-334.

Hambleton, R. K., Swaminathan, H. and Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. Sage Press, Newbury Park, CA.

Henrysson, S. (1971). Gathering, analyzing, and using data on test items. *In R.L. Thorndike (Ed.), Educational Measurement (p. 141)*. American Council on Education, Washington DC.

Hopkins, D. (1985). *A teacher's guide to classroom research.* Open University Press, Philadelphia.

Hostetter, L. and Haky, J. E. (2005). A classification scheme for preparing effective multiple-choice questions based on item response theory. *Florida Academy of Sciences, Annual Meeting*. University of South Florida, March, 2005 (cited in Hotiu, 2006).

Hotiu, A. (2006). The relationship between item difficulty and discrimination indices in multiple-choice tests in a physical science course. *Master in Science Thesis, Charles Schmidt College of Science*. Florida Atlantic University, Boca Raton, Florida.

Kelley, T. L. (1939). The selection of upper and lower groups for the validation of test items. *J. Ed. Psych.,* **30**, 17-24.

Kemmis, S. (1983). Action Research. *In DS Anderson & C Blakers (eds), Youth, Transition and Social Research*. Australian National University, Canberra.

Krathwohl, D. R., Bloom, B. S. and Bertram, B. M. (1973). *Taxonomy of Educational Objectives, the Classification of Educational Goals. Handbook II: Affective Domain*. David McKay Co., Inc., New York.

Lewin, K. (1946). Action research and minority problems. *Journal of Social Issues*, 2,

34-46.

Lord, F. M. (1952). The Relationship of the Reliability of Multiple-Choice Test to the Distribution of Item Difficulties. *Psychometrika*, 18, 181-194.

Lord, F. M. and Novick, M. R. (1968). *Statistical Theories of Mental Test Scores*. Addison-Wesley, Reading, MA.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Lawrence Erlbaum Associates, Inc, New Jersey.

Mertler, C. A. (2003). *Classroom Assessment – A Practical Guide for Educators*. Pyrczak Publishing, Los Angeles, CA.

Millman, J. and Greene, J. (1993). The specification and development of tests of achievement and ability. *In R.L. Linn (Ed.), Educational measurement (pp. 335-366)*. Oryx Press, Phoenix, AZ.

Nitko, A. J. (2001). *Educational assessment of students (3rd edition)*. Prentice Hall, Upper Saddle River, NJ

Nunan, D. (1992). *Research Methods in Language Learning*. Cambridge University Press, Cambridge.

Nunnally, J. C. (1972). *Educational measurement and evaluation (2nd ed)*. McGraw-Hill, New York.

Nunnally, J. C. (1978). *Psychometrics Theory, Second Edition.* : McGraw Hill, New York.

Oosterhof, A. (2001). *Classroom applications for educational measurement*. Merrill Prentice Hall, Upper Saddle River, NJ.

Pedhazur, E. J. and Schmelkin, L. P. (1991). *Measurement, design, and analysis: An integrated approach*. Erlbaum , Hillsdale, NJ.

Popham, W. J. (1981). Modern educational measurement. Prentice-Hall, Englewood Cliff, NJ.

Rapoport, R. (1970). Three dilemmas in action research. *Human Relations*, **23(6)**, 499-513.

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests*. (Copenhagen, Danish Institute for Educational Research), expanded edition (1980) with foreword and afterword by B.D. Wright. The University of Chicago Press, Chicago.

Richards, J. C., Platt, J. and Platt, H. (1992). *Dictionary of Language Teaching and Applied Linguistics*, Second Edition, Longman, London.

Sax, G. (1989). *Principles of educational and psychological measurement and evaluation (3rd ed)*. Wadsworth, Belmont, CA.

Simpson, E.J. (1972). *The Classification of Educational Objectives in the Psychomotor Domain.* Gryphon House, Washington, DC.

Spearman, C. (1904). “General intelligence,” objectively determined and measured. *American Journal of Psychology*, **15**, 201-293.

Suen, H. K. (1990). *Principles of exam theories*. Lawrence Erlbaum Associates, Hillsdale, NJ.

Tamhane, A. C. and Dunlop, D. D. (2000). *Statistics and Data Analysis from Elementary to Intermediate*. Prentice Hall, Upper Saddle River, NJ.

Tanner, D. E. (2001). *Assessing academic achievement*. Allyn & Bacon, Boston.

Taylor, P. and Pettit, J (2007). Learning and teaching participation through action research: Experiences from an innovative masters programme. *Action Research*, **5(3)**, 231-247.

Thompson, B. and Levitov, J. E. (1985). Using microcomputers to score and evaluate test items. *Collegiate Microcomputer*, **3**, 163-168.

Thorndike, R. M., Cunningham, G. K., Thorndike, R. L. and Hagen, E.P. (1991). *Measurement and evaluation in psychology and education (5th ed)*. MacMillan, New York.

Triola, M. F. (2006). *Elementary Statistics*. Pearson Addison-Wesley, New York.

Van der Linden, W. J. and Hambleton, R. K. (Eds.) (1997). *Handbook of modern item response theory.* Springer, New York.

Wiersma, W. and Jurs, S. G. (1990). *Educational measurement and testing (2nd ed)*. Allyn and Bacon, Boston, MA.

Wilcox, R. R. (1981). A review of the beta-binomial model and its extensions. *Journal of Educational Statistics*, **6(1)**, 3-32.

Wood, D. A. (1960). *Test construction: Development and interpretation of achievement tests*. Charles E. Merrill Books, Inc, Columbus, OH.

Wright, B. D. (1992). IRT in the 1990s: Which Models Work Best?. *Rasch measurement transactions*, **6(1)**, 196-200