RESPONSE LATENCY EFFECTS ON CLASSICAL AND ITEM RESPONSE
THEORY PARAMETERS USING DIFFERENT SCORING PROCEDURES

A dissertation presented to
the faculty of
the College of Education of Ohio University

In partial fulfillment
of the requirements for the degree
Doctor of Philosophy

Faisal A. Abdelfattah
August 2007
This dissertation titled
RESPONSE LATENCY EFFECTS ON CLASSICAL AND ITEM RESPONSE
THEORY PARAMETERS USING DIFFERENT SCORING PROCEDURES

by

FAISAL A. ABDELFATTAH

has been approved for
the Department of Educational Studies
and the College of Education by

______________________________
George A. Johanson
Professor of Educational Studies

______________________________
Renée A. Middleton
Dean, College of Education
Abstract

ABDELFATTAH, FAISAL A., Ph.D., August 2007, Educational Research and Evaluation

RESPONSE LATENCY EFFECTS ON CLASSICAL AND ITEM RESPONSE THEORY PARAMETERS USING DIFFERENT SCORING PROCEDURES (162 pp.)

Director of Dissertation: George A. Johanson

Test performance may have little or no personal consequences for examinees in numerous testing situations. In such situations, test administrators are likely to view the effort levels of examinees as a matter for concern. Response latency or item response time was used in an attempt to detect the occurrence of a type of guessing that is expected to occur if the examinee does not put forth adequate effort. That is, responding to an item without adequately reading it and engaging in solution-oriented behavior.

The study employed response latency as an indicator of examinees rapid-guessing behavior based on response latencies lower than a particular threshold for each item. The identified responses were rescored differently using the Omitted, the Not-presented, or the Zero scoring procedures. The objective was to evaluate the use of different scoring procedures on Classical Test Theory (CTT) and Item Response Theory (IRT) parameter estimates for dichotomously scored items that were obtained from a sample of 586 ninth grade students from five high schools in Jordan who took a computer-administration of a mathematics test.
The results revealed that rapid guessing may have occurred fairly early in the test. Further, it was found that examinees showed rapid-guessing behavior on every item on the test and with greater frequency on the later items. A single group within-subjects design (repeated measure) ANOVA was used to assess changes in CTT and IRT item parameter estimates across scoring procedures.

Significant differences in classical item difficulty and discrimination indices for the Omitted and the Zero procedures contrasted with the default scoring procedure and there was no significant difference in difficulty and discrimination estimates between the Original scoring procedure and the Not-presented procedure. Further, there were no significant differences in IRT parameter estimates across scoring procedures except for persons’ parameter estimates. With regard to pass/fail decisions, it appears that identifying individual examinee rapid-guessing responses and rescoring them may well influence the scores and, therefore, the decision concerning performance may change accordingly.

Approved: _____________________________________________________________

George A. Johanson

Professor of Educational Studies
Dedication

To my parents, the endless source of giving, the meaning of all of my existence and confidence.

To my wife SUHA, for the honest love, care, understanding, and endless support.

To my lovely kids, LINA and AHMAD, for their smile and hope.

Your positive feelings and inspiration encouraged me to pursue my graduate work; you have confidence in me when I doubted myself
Acknowledgements

My gratitude and praise to Allah for granting me the strength and patience to accomplish my Ph.D. study.

I would like to express my gratitude to my advisor, Prof. George Johanson, for his unlimited support throughout my study. Without his patience, encouragement, and guidance I could not complete this work. George was always there to meet and talk about many problems I faced.

Also, I wish to thank my committee members, Dr. Gordon Brooks, he was a good teacher for statistics courses, I learned a lot from him, Dr. Tracy Leinbaugh, and Dr. Greg Kessler, for their valuable comments and insights. I have been blessed to have you all in my committee.

I will not forget my friends, who always encouraged me to do this work and gave me their trust. My special thanks to Dr. Hamzeh Dodeen, Dr. Ibrahim Al-mujaini, Dr. Ziyad Akir, and Dr. Talal Alodwan who encouraged me in different stages of my graduate work.
Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>iii</td>
</tr>
<tr>
<td>Dedication</td>
<td>v</td>
</tr>
<tr>
<td>Acknowledgements</td>
<td>vi</td>
</tr>
<tr>
<td>List of Tables</td>
<td>x</td>
</tr>
<tr>
<td>List of Figures</td>
<td>xiii</td>
</tr>
<tr>
<td>CHAPTER ONE</td>
<td>1</td>
</tr>
<tr>
<td>Introduction</td>
<td>1</td>
</tr>
<tr>
<td>Background of the Study</td>
<td>1</td>
</tr>
<tr>
<td>Purposes of Assessment in Education</td>
<td>2</td>
</tr>
<tr>
<td>The Role of Computers in Assessment</td>
<td>7</td>
</tr>
<tr>
<td>Response Latency</td>
<td>8</td>
</tr>
<tr>
<td>The Statement of the Problem</td>
<td>13</td>
</tr>
<tr>
<td>Research Questions</td>
<td>14</td>
</tr>
<tr>
<td>Rationale</td>
<td>15</td>
</tr>
<tr>
<td>Limitations and Delimitations of the Study</td>
<td>17</td>
</tr>
<tr>
<td>Definition of Terms</td>
<td>17</td>
</tr>
<tr>
<td>CHAPTER TWO</td>
<td>20</td>
</tr>
<tr>
<td>Review of the Literature</td>
<td>20</td>
</tr>
</tbody>
</table>
Analyses for Research Question Four .............................................73

Analyses for Research Question Five .............................................77

CHAPTER FIVE .........................................................................................79

Discussion..............................................................................................79

REFERENCES ..........................................................................................88

APPENDICES .........................................................................................97

APPENDIX A Mathematics Test (Arabic Translation).................................98

APPENDIX B Mathematics Test (in English).............................................118

APPENDIX C Sample of Items from the Computer Administration.............137

APPENDIX D Medians of response latency thresholds............................140

APPENDIX E Number of Rapid-guessing Responses across Examinees........141

APPENDIX F Statistics for Repeated Measures ANOVA Assumptions...........143

APPENDIX G 68% CI across Scoring Procedures....................................149
List of Tables

Table 1. Descriptive Statistics for Guessing-behavior .......................................................58

Table 2. Confidence Intervals for Proportion of Guessing Committed Based on Item Level ....................................................................................................................59

Table 3. Descriptive Statistics for CTT Difficulty and Discrimination Indices ..............64

Table 4. Univariate ANOVA of Within-Subjects for CTT Difficulty Indices ..............67

Table 5. Tests of Within-Subjects Contrasts for CTT Difficulty Indices .......................67

Table 6. Univariate ANOVA of Within-Subjects for CTT Discrimination (a) Indices ....68

Table 7. Tests of Within-subjects Contrasts for CTT Discrimination (a) Indices ............68

Table 8. Descriptive Statistics for IRT Item Parameter Estimates ..................................69

Table 9. Tests of Within-subjects Effects for IRT Difficulty (b) Parameter Estimates ....71

Table 10. Tests of Within-subjects Effects for IRT Discrimination (a) Parameter Estimates ..................................................................................................................72

Table 11. Tests of Within-subjects Effects for IRT Pseudo-guessing (c) Parameter Estimates ..................................................................................................................73

Table 12. Tests of Within-subjects Contrasts for IRT Pseudo-guessing Parameter Estimates ..................................................................................................................73

Table 13. Descriptive Statistics for IRT Person Parameter Estimates ............................74
Table 14. Tests of Within-subjects Effects for IRT Person ($\theta$) Parameter Estimates.......75

Table 15. Tests of Within-subjects Contrasts for IRT Person ($\theta$) Parameter Estimates....75

Table 16. Proportion of Students Succeeding Based on Different Cut-scores ..........77

Table 17. 95% CI for the Proportion of Succeeded Students across Scoring Procedures..............................................................................................................78

Table 18. Medians of Response Latency Thresholds ..............................................140

Table 19. Percentages of Examinees According to Number of Guesses.................141

Table 20. One-sample Kolmogorov-Smirnov Test for Classical Difficulty Index.......143

Table 21. One-sample Kolmogorov-Smirnov Test for Classical Discrimination Index ............................................................................................................143

Table 22. One-sample Kolmogorov-Smirnov Test for IRT Discrimination Parameter ..143

Table 23. One-sample Kolmogorov-Smirnov Test for IRT Difficulty Parameter...........144

Table 24. One-sample Kolmogorov-Smirnov Test for IRT Pseudo-Guessing Parameter .............................................................................................................144

Table 25. One-sample Kolmogorov-Smirnov Test for IRT Ability Parameter..............144

Table 26. Skewness and Kurtosis for IRT Ability Parameter........................................145

Table 27. Mauchly's Test of Sphericity for CTT Difficulty Indices................................145

Table 28. Mauchly's Test of Sphericity for CTT Discrimination Indices .....................146

Table 29. Mauchly's Test of Sphericity for IRT Difficulty Parameter Estimates.............146
Table 30. Mauchly's Test of Sphericity for IRT Discrimination Parameter Estimates ...
147

Table 31. Mauchly's Test of Sphericity for IRT Pseudo-guessing Parameter Estimates
147

Table 32. Mauchly's Test of Sphericity for IRT Person Parameter Estimates ..........148

Table 33. 68% CI for the Proportion of Succeeded Students across Scoring Procedures
149
List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 1</td>
<td>Distribution of examinee response times for an information literacy item</td>
<td>38</td>
</tr>
<tr>
<td>Figure 2</td>
<td>Frequency of solution-oriented responses across items.</td>
<td>54</td>
</tr>
<tr>
<td>Figure 3</td>
<td>Frequency of examinees exhibited rapid-guessing behavior.</td>
<td>55</td>
</tr>
<tr>
<td>Figure 4</td>
<td>Frequency of correct and incorrect responses under rapid-guessing behavior</td>
<td>56</td>
</tr>
<tr>
<td>Figure 5</td>
<td>Percentage of correct and incorrect responses under rapid-guessing behavior</td>
<td>57</td>
</tr>
<tr>
<td>Figure 6</td>
<td>The relationship between scores based on Original responses and the Not-presented scoring procedure</td>
<td>76</td>
</tr>
</tbody>
</table>
CHAPTER ONE

Introduction

Background of the Study

A major purpose of assessment in educational settings is to measure students’ achievement and to make a variety of decisions based on students’ performance. The accuracy of decisions may vary based on the educational consequences to the students or to the school they belong. If the students do not put forth a reasonable effort, the results clearly may underestimate their proficiency level and lead to invalid interpretations of the obtained scores.

Advances in technology and the growing presence of computers in assessment provide tremendous opportunity to explore new ways to improve the quality of assessment data (Klein & Hamilton, 1999). It is relatively easy with computer administered tests to collect additional information related to the interaction between each individual examinee and a single item on the test. The time required for a response can give us a good indicator of the student’s effort while testing. This is referred to as the “response latency.” Low-effort responses can be rescored in different ways to affect items’ and persons’ characteristics which may have an impact on educational decisions (Bhola, 1994; Hadadi & Luecht 1998).

This chapter begins by describing the purposes and the importance of assessment in education, then the improvements that computers can make to collect accurate data. This is followed by a description of an effort moderated model which relies on individual item response latency.
Assessment of student learning plays a significant role in the educational process. The effectiveness of instruction and instructional decisions depend largely on the ability of educators to construct and select tests and assessments that provide valid, reliable, and fair measures of learning outcomes (Linn & Gronlund, 1995). The nature and the quality of gathered information can control the educational development efforts and direct the instruction.

Education policymakers and professionals often desire to use tests for multiple purposes, such as monitoring the educational system, aiding instructional planning, motivating students to perform better, acting as a mechanism to change instructional content, and holding schools and educators accountable. In addition, they use tests for certifying students as having attained specific levels of achievement (Klein & Hamilton, 1999; Hamilton, Stecher, & Klein, 2002).

Making inferences about a student’s performance goes beyond the specific test that is used. Sometimes, teachers would like to know the degree of a student’s understanding of specific concepts based on the score that he or she obtains on achievement tests, which corresponds to the knowledge and skills the student learns in the usual school subjects. Achievement tests have commonly been used to measure students’ educational progress, but the number of purposes they expect to serve has grown substantially.

Achievement tests might be helpful in answering questions like: To what extent are students making progress beyond the minimum basics? To what extent are students
achieving the learning goals of the course? To what extent are students ready for the next learning experience? Or, how well can we predict students’ performance in the future and generalize from the test, which is a sample, to the larger and broader sample of questions? (Linn & Gronlund, 1995). Also, tests might be helpful in evaluating the effectiveness of the instruction. The quality of such evaluations depends to a large extent on the nature and quality of the collected information during the assessment setting.

Large-scale assessments use tests, often consisting of standardized multiple-choice, essay, and other types of instruments, to gather information from large numbers of students across different classrooms, schools, and educational districts as a part of a district’s or country’s testing program to monitor student achievement and for school and teacher accountability goals (Bond, 1995). Generally, large-scale tests are used for several other purposes which might include providing evidence of educational quality for public review, providing information for teachers to help them improve their instructional practices, and providing information for teachers, parents, and students to monitor student progress. Moreover, tests are used for evaluating the effectiveness of reform efforts or curricula, judging whether students should be promoted from a particular grade to the next, and determining whether students should be placed into specialized educational programs (remedial, gifted, or bilingual classes) or whether they will receive a high school diploma.

Scores on large-scale assessments can affect the students who take them and the teachers and administrators who educate those students. Also, it is clear that quality
assessment is likely to lead to improvements in student learning (Hamilton, Stecher, & Klein, 2002).

The increased reliance on assessment and testing as an educational reform tool has raised serious issues related to the nature and the quality of the tests, the possible effects of testing on students, and the fairness of tests in large-scale assessments (Linn & Gronlund, 2002). It becomes necessary for educational policymakers and practitioners to understand the strengths and the weaknesses of large-scale assessments both as measures of student and school performance and as policy tools to change practice.

Two kinds of tests that differ in their consequences appear to be responsible for some of the variation in student performance. For one, high-stakes tests are often used to refer to a set of policies which include procedures that provide rewards and/or approval as a consequence of test scores on large-scale assessments. Second, low-stakes tests are used typically for judging schools’ effectiveness, tracking the school-improvement efforts, and to provide information on student performance for teachers, policymakers, parents, and others. On the individual level, the consequences of low-stakes tests are considered to be relatively minor for both teachers and students. Conversely, on the institutional level, high performing schools receive academic achievement rewards, which might include public recognition, money to use for school improvement, or direct cash bonuses for staff. Constantly, low-performing schools face various forms of correction actions and possibly reconstitution by officials (Greene, Winters, & Forster, 2004; Hamilton, Stecher, & Klein, 2002; Thomas, 2005).
Students’ success on such tests depends not only on the teacher, but also on students’ motives, interests, energy level, and other factors related to their background. Greene, Winters, and Forster (2004) reported that increased randomness in student answers on low-stakes tests might result from their investing lower levels of effort in taking those tests.

Although the intent of assessments of student achievement is to measure what is called “maximum performance” they can do so only if students attempt to do their best when taking the assessment. If a student is not motivated to put forth a reasonable effort, the results clearly may underestimate his or her maximal proficiency and so lead to invalid interpretations of the obtained scores and a distortion of the results (Linn & Gronlund, 2002; Wise, 2006).

Accurately assessing achievement is one of the big tasks that educational professionals try to accomplish. Standardized tests provide an effective tool for assessing students and schools. For this reason, countries, states, and school districts are turning to use such tests to compare schools and evaluate their progress. Policymakers, parents, and citizens in the United States are keenly interested in students’ achievement and how U.S. students compare with other countries (Walberg, 2001).

Thomas (2005) reported results from the National Assessment of Educational Progress (NAEP), in which many students were failing to master basic skills required for entering college or gaining employment. These studies showed that American students compared unfavorably to their peers in other nations.
A number of people have argued that NAEP provides an underestimate of student achievement because the assessment has no direct consequences for the students; that is, to the individual student, the NAEP is a low-stakes test. Conversely, it is hypothesized that a high-stakes testing environment, in which a student's performance has serious consequences for students (admission, promotion or graduation), such as the American College Testing Program Assessment Exam (ACT) and the Scholastic Aptitude Test (SAT), on which the student score forms an important factor in admission to colleges and universities, would provide a more accurate estimate of student performance. Since such tests affect the future of personal educational opportunities and life chances for students, those who take the tests have a higher stake in performing well. These larger consequences are presumed to motivate performance commensurate with students' actual abilities (Kiplinger & Linn, 1996).

Scholars have expended much effort in trying to explain the seeming inconsistency between student scores and what the student knows. Measurement experts intend to obtain valid test scores from each test administration. Whenever the examinee perceives that his or her performance has little or no consequences, it is more likely that he or she will put little effort into the assessment. Consequently, their scores may not accurately reflect their true abilities (Wise & DeMars, 2003, 2005). This case represents a direct threat to test score validity, and experts have a responsibility for taking corrective action.
The Role of Computers in Assessment

The new methodologies in assessment offer opportunities for improvement in the quality of data, which would lead to greater precision and increased validity. The reduction of measurement error would be beneficial for school accountability and the accuracy of decisions (Rabinowitz & Brandt, 2001).

Advances in technology and the rapidly growing presence of computers in schools provide opportunities to explore alternative modes of testing. Using computers to administer tests is becoming increasingly common. Computers are also becoming central in the administration of many large-scale testing programs (Klein & Hamilton, 1999). Computers have made it easy to manipulate information electronically and to gather information become broader and even more accurate.

The change that has the most potential to improve both the efficiency with which information is handled and the quality of that information is the use of the computer to administer tests. Moreover, several advantages of computer-based testing arise over current paper-and-pencil tests. Ease of scheduling, the incorporation of multimedia questions, immediate scoring, the possibility of separating the test-taker’s speed from the accuracy of responses, and the administration of tailored tests are several areas where computer-based testing (CBT) offers distinct rewards. Although these advantages can be substantial in some cases, the more significant changes depend on using the computer to do things that cannot be reasonably done using paper-and-pencil tests. Computer-based tests help assess not only the product of a student’s thinking, but the process that the
student uses to answer a question, including how the question is solved and the number of hints that may be needed to reach the correct answer (Linn & Gronlund, 2002).

Contrasting performance between CBT and paper-and-pencil tests showed a nonsignificant difference due to the test modes. The results showed that neither feelings of anxiety toward the computer nor lack of experience using computers adversely affects examinee performance on CBT (Wise, Barnes, Harvey, & Plake, 1989). Results from Chin, Donn, and Conry (1991) revealed similar scores of anxiety in the computer-based group in comparison to the paper-and-pencil group.

Response Latency

Using a computer to administer tests greatly expands the range of performance tasks that can be included in standardized tests and the information gathered during the testing session. Additional data at the item level, namely “response latency,” can be gathered easily and behaviors related to the answer can be recorded. Response latency is the time used by an examinee when responding to an item.

Moreover, computer-based testing appears to be remarkably attractive for low-stakes testing programs. The capability to measure response latency at the item level appears to be easy and valuable. These response latencies can be used for both assessing the degree to which examinees give effort to the test and assessing the degree to which items receive good effort (Wise, 2006).

Different behaviors may appear during the testing session. Mislevy and Wu (1996) reported several reasons that some responses may not be observed for all examinees to all test questions. Under a typical testing setting, some examinees will not
reach the last questions on a test because of the time limit. Items of this sort are referred to as not-reached. Even when items are reached and read, some examinees may appraise the item’s content and decide for their own reasons not to respond. Items of this sort are referred to as omitted items. On the other hand, if the examinee realizes that there are no direct personal consequences based on his or her performance, the examinee may engage in random-guessing response behavior.

Because low-stakes testing continues to play a considerable role in current educational measurement practice, it is important that an understanding of the nature and the extent of the threat posed by lack of examinee effort be sought by experts. Such understanding requires reasonable measures of examinee effort. Literature mentioned in Wise (2006) shows that there are three primary choices of effort measure: examinee self-reports, person-fit statistics, and response-time based measures.

The first method to detect examinee effort is to use self-report instruments, which ask for example, “I am not concerned about the score I receive on this test.” Self-report instruments are easy to administer and score, but problems may arise when inferences are drawn about student responses. Examinees may be giving socially acceptable responses that do not match with what they actually believe. The information gathered through self-reports may not correspond to the student’s actual behavior during the examination. Younger children may not understand the self-report questions and be incapable of giving accurate responses. In some cases, students who did not try to do well on the test might falsely report giving good effort because of fear of disapproval or punishment from the test giver. Other students who believe they did not perform well on the test might
underreport their effort because of a predisposition to attribute failure on the test to lack of effort over lack of ability (Pintrich & Schunk, 2002; Wise, 2006). And, of course, students who choose not to expend the necessary energy on the test for a valid score may also choose not to expend the necessary energy on the self-report instrument for a valid score.

The second method to detect examinee-effort is to examine person-fit statistics which compare the examinee’s response pattern with a theoretical measurement model. This choice has a clear advantage of being based on an observed item response pattern rather than self-report questions. The use of person-fit indices, as an exploratory technique, applied typically in situations in which the kind of aberrant responses can be expected, is unknown, leads to numerous interpretations. It may become difficult to conclude unambiguously that a particular instance of misfit is due to lack of effort (Wise & DeMars, 2006). Misfit may indicate that the examinee engaged in cheating, creative responses, careless responses, guessing, or other unusual response behaviors. Researchers cannot always be confident of the kind of aberrant responses underlying test performance because different forms of aberrant behavior can have multiple explanations and may result in the same kind of item response pattern (Meijer, 1996). Thus, it becomes difficult to rely solely on person-fit statistics to measure examinee effort.

A third method is to use measures based on response latency. In 2005, Wise and Kong developed an index named Response Time Effort (RTE) to measure the examinee’s test-taking effort in computer-based settings. RTE is based on prior research done by Schnipke (1995, 1996, 1999) and Schnipke and Scrams (1997a, 2002) that distinguishes
appropriate solution-oriented behavior (where the examinees actively seek to determine the correct answer to test items) from rapid-guessing behavior (where the examinees determine the answer by rapid-guessing; perhaps because they do not have enough time to fully think about the item). Wise and Kong discovered that rapid-guessing behavior appears to be obvious in untimed data in low-stakes computer-based tests. This behavior could be observed throughout the test and not just toward the end of speeded tests as proposed earlier by Schnipke and Scrams (Wise & Kong, 2005; Wise, 2006).

The increased use of standardized tests which have no impact on grades may have discouraged students so that they do not try very hard, and do not perform at their actual ability levels. Students are not always motivated to give their best effort (O’Neil, Abedi, Miyoshi, & Mastergeorge, 2005). The motivational level of students may affect their performance, no matter their achievement level. Students may seriously underperform, make random-guessing responses, omit answers, or not finish the test, which are all signals of low motivation and noncompliance (Haladyna & Downing, 2004). The resulting scores may thus underrepresent what students know and influence the validity of decisions made based on these results.

Sundre and Wise (2003) and Wise and DeMars (2005) introduced motivation filtering as an attempt to increase the validity and credibility of low-stakes assessment. The logic underlying motivation filtering is that the data from those students giving low effort are untrustworthy, and by deleting these data, the remaining data are assumed to better represent the ability levels of the target group of students (Wise, Wise, & Bhola,
Motivation filtering is thus assumed to be appropriate when inferences are based on a group of students at the national, institutional, or class level.

Response latency information has an important advantage over self-reports in that it represents a direct observation of examinee behavior and it does not rely on examinee judgments. Moreover, data for response latency can be collected in an easy and non-reactive way. Examinees may have little awareness that response time data are even being recorded. Response latency information is at the item level and therefore allows tracking the changes in the level of effort during a testing session. Unlike person-fit statistics, response latency is a direct way to examine the effort that the examinee puts forth on test items and may occur along with other aberrant behaviors during the testing session.

Recording response latency might raise some ethical concerns; this study was not employed to modify examinees’ scores while taking the test, nor changing their scores according to the response latency. The primary interest will be in comparing item parameter estimates after considering response latency per item.

In the current study, response latency will be used in an attempt to detect the occurrence of a type of guessing that is expected to occur if the examinee does not put forth adequate effort; perhaps even responding to an item without reading it. If examinees spend less than a predetermined threshold of time on a particular item, then that item will be flagged and rescored in a variety of ways including: not reached, omitted, zero, or the response will remain as the examinee answered it, and scored 1 if correct, otherwise 0. The rescored data sets will be used to determine if these different rescaling procedures
effectively change the relevant parameter estimates and decisions based on examinees’ scores compared with the original data. Differences will be based on both classical test theory (CTT) and the three-parameter logistic item response theory model (3PL-IRT).

The Statement of the Problem

An understanding of the nature of rapid-guessing behavior and the development of indices to detect this behavior can help in the process of test construction, testing of response validity, and the verification of decisions made using these test results.

While testing, some examinees may neither engage entirely in the test nor give enough effort answering the test questions. The resulting test scores may be biased in representing what they know and can do (Wise & Demars, 2003). Wise and his colleagues (Sundre & Wise, 2003; Wise & DeMars, 2005; Wise & Kong, 2005) deleted the entire test record for any unmotivated examinee. Their justification was, in part, that the interest of researchers in low-stakes assessment testing is the average proficiency of a group of examinees, not the individual examinee. Of course, deleting records from the obtained data set will reduce the number of students and the sample may be less representative of the targeted population.

In the proposed study, the interest will be in using three different scoring methods to replace the guessed responses instead of deleting them, and to see how these scoring procedures impact different parameter estimations for persons and items in terms of CTT and IRT models. Specifically, the purpose of this study will be: (a) to examine the stability of classical parameter estimates (item difficulty, discrimination, reliability coefficient), (b) to examine the stability of IRT parameter estimates for persons and
items, (c) to explore the impact of using different scoring procedures on pass/fail
decisions. All analyses will be based on an effort-moderated model across different
scoring procedures for a set of responses obtained from low-stake achievement tests and
where person decisions are pass/fail.

Based on the findings of this proposed research, measurement experts might
effectively use the effort-moderated model to help resolve the rapid-guessing problem.
The more effective scoring procedures could be incorporated into the item and person
parameter estimation processes.

Research Questions

The overall purpose of this study is to examine the effects of using different
scoring procedures on CTT and IRT person and item parameter estimates considering
response latency. In particular, the current study will be focused on answering the
following questions:

1. Is there a significant proportion of students who exhibit rapid-guessing
   behavior?

   The RTE index will be used to indicate the percentage of items with solution-
   oriented behavior. Rapid-guessing responses will be detected and rescoring using different
   scoring procedures. The resulting estimates for item and person may be different from
   those obtained from the whole sample which includes the rapid-guessed responses. One
   of the current research goals is to examine the extent of rapid-guessing behavior that
   would seriously influence classical item statistics, such as item means (difficulty), item-
   total correlations (discrimination), and test score reliability. More specifically:
2. Are there significant differences in classical item indices (item means, item-total correlations, and reliability) across the different scoring procedures?

3. Do the different scoring procedures yield significantly different IRT item parameter estimates?

4. Do the different scoring procedures yield significantly different IRT person parameter estimates?

5. Do pass/fail decisions for each examinee differ across the different scoring procedures?

**Rationale**

Examinees often differ in the degree of effort they put forth taking a low-stakes test. Some of them will give full effort; other examinees might try, but not as much as if they were taking the test for a course grade or other individual consequence. Still others might try to answer some questions but on the remaining questions they will possibly just guess an answer and move to the next one. It becomes crucial to identify the guessers and develop ways to deal with this behavior that does not detract from the accuracy of the test results.

Wise (2006) discussed the responsibility of measurement experts within low-stakes test settings, to either develop methods for identifying and managing data from examinees who do not give good effort or to implement testing practices that elevate examinee’s effort. Using computer-based testing and employing the effort-moderated model may represent one strategy. A second strategy may be developing relatively short
tests, with limited amounts of reading that are basically motivating to examinees and obtain a maximum effort.

Sometimes high-stakes testing programs administer low-stakes tests in the early stages of the program. The purpose may be to obtain data that later can be used in item calibration, test form construction, or linking and equating (DeMars, 2000; Wise, 2006). It appears that there is a great deal of measurement research conducted in low-stakes settings. For example, in large-scale assessments, measurement professionals and test developers frequently use students or volunteers for which there are no consequences associated with test results for test development and calibration purposes. It is an important issue because measurement professionals and test developers want item parameters calibrated from pilot tests to reflect relatively properties of the items they will have on the final form of the test. If some item properties change more than others, the test developers will not get the items they intended.

Scores from low-stakes assessments may underestimate what students know and thus threaten the validity of decisions made based on these scores. If performance standards are based on underestimated values, they may result in cut scores for the associated high-stakes tests being too low. Calibration on high-stakes tests is similarly problematic (DeMars, 2000). Accurate estimates from these tests are important because results from them may be used in making essential policy decisions or establishing norms for the published standardized tests.
Limitations and Delimitations of the Study

1. All data were obtained via a computer-administered test. Students’ attitudes towards computers and proficiency of using a computer may impact the results.

2. The mathematics test consisted of 35 multiple-choice items; other item formats may result in different outcomes.

3. Items will be calibrated using 3PL-IRT to get the item and person parameters. Other IRT models (1PL-IRT and 2PL-IRT) may reveal different results.

4. Thresholds were determined during the piloting stage of the test on 20 students from ninth grade. Decisions about randomly guessed responses depend on the reasonableness of the predetermined latency threshold. To the best of the researcher’s knowledge, this method has not been introduced in the literature. In this study, threshold is based on the real time that an examinee needs to read the entire item stem and the options.

5. Only students from the ninth grade from Jordan participated in this study.

6. Results obtained from the study are based on a low-stakes, computer-administered test and may not apply to high-stakes tests and computer-adaptive tests.

Definition of Terms

For the purpose of this study, the following terms will be defined accordingly:

Response latency: defined as the time elapsed between presenting the question on the computer screen and the response to that question.
Threshold: the minimum time that an examinee needs to read and respond to the presented question. Wise and Kong (2005) assigned a threshold to each item according to both the total number of characters in the stem and options and whether or not ancillary materials (such as a table or graph) were included with the item. Wise (2006) used a visual inspection of response latency presented by line graphs to choose thresholds.

Solution-oriented behavior: the examinee reads each item carefully and fully considers the solution (Schnipke & Scrams, 1997).

Rapid-guessing behavior: the examinee skims the item briefly for key words without reading it thoroughly (and this guessing behavior may not be affected by item content). The concept, rapid-guessing, is used to describe the examinee’s behavior when he/she runs out of time while taking tests (Schnipke & Scrams, 1997; Thomas, 2006).

Random-guessing behavior: This concept is used interchangeably with rapid-guessing behavior in timed or untimed tests (Thomas, 2006).

Effort-moderated model: a suggested model that incorporates response latency within the standard 3PL-IRT model (Wise & DeMars, 2006).

Response time effort (RTE) index: the amount of overall response effort/time that the examinee puts forth in the test (Wise & Kong, 2005). RTE is a proportion established for each examinee based on the number of solution-oriented behavior items out of the total number of items in the test.
Motivation filtering: collecting data regarding student’s effort immediately after administering the test using a motivation scale. Filtering can be accomplished by rank ordering motivation scores from low to high and removing successively proportions of examinees with the lowest motivation scores from the data (Wise, Wise, & Bhola, 2006).

Large-scale assessment: gathering information from large numbers of students across different classrooms, schools, and educational districts using standardized tests, which typically include multiple-choice and essay questions (Hamilton, Stecher, & Klein, 2002).

High-stakes tests: tests that reflect a set of policies and procedures that provide rewards and/or approval as a consequence of test scores for examinees (Thomas, 2005).

Low-stakes tests: tests that include no or minor consequences of test scores for examinees; these are commonly used to compare a group of examinees with national norms (Greene, Winters, & Forster, 2004; Thomas, 2005).
CHAPTER TWO

Review of the Literature

Introduction

This study will investigate the consequences of using different scoring procedures on CTT and IRT estimates based on response latency, using data collected via computer-based administration. This chapter will compare test performance on large-scale assessments that have no direct consequences for individual examinees with test performance that has high consequences. A review of the relevant literature in the area of response latency as an indicator of rapid-guessing behavior in large-scale assessment settings will be presented. In addition, this chapter will discuss methods that have been used in previous research dealing with rapid-guessing responses. The current study will also address an alternative approach for rescoring rapid-guessing responses to avoid deleting these responses.

A particular area of concentration involves assessing the consequences of employing different scoring procedures on (PASS, FAIL) decisions based on a student’s performance. The literature review section will conclude with a short summary of the most relevant aspects of the topics presented.

Test Performance on Low-stakes Assessment

A number of studies have investigated the relations between examinee effort and test performance within different stakes settings. The results have been fairly consistent, where less-motivated examinees perform less well than their highly-motivated counterparts.
Kiplinger and Linn (1993) examined the effect of the testing environment as related to varying the stakes of the test on student performance for standardized mathematics tests. A set of 17 items from NAEP released items were embedded in Georgia Curriculum-Based Assessments (CBA), part of the state assessment program, that is considered having higher stakes for teachers and school systems; on the other side, NAEP current administration considered having low-stakes. Kiplinger and Linn assumed that teachers devote great efforts to encouraging students to put forth their best efforts on the assessment. Comparative analyses of student performance on the same set of items were conducted to determine whether performance is enhanced when such items are administered in a higher stakes environment than does NAEP. Effect size was calculated using the difference between the means of conditions divided by the pooled estimate of the standard deviation; a small effect size of 0.20 was found for the first nine items while an effect size of 0.04 was found for the last eight items on the two administrations. Kiplinger and Linn concluded that students’ achievement is related to testing stakes and no one can conclude that the stakes of assessment would not be accompanied by a significant increase in achievement scores.

Later, Kiplinger, and Linn (1996) reported results for students’ responses to motivation items for the mathematics assessment of the 1992 NAEP; they noted that 40% of grade 8 students said that it was "not very important" or only "somewhat important" for them to do well on the mathematics test. Sixty-nine percent of these students reported that they tried "not at all hard" or only "about hard" on the mathematics test. The results for grade 12 students showed higher percentages on the aforementioned items. These
results raise skepticism about the validity of such test scores and the importance of performing well on these assessments.

DeMars (2000) conducted a study to examine how scores on the Michigan High School Proficiency Test (HSPT) changed when the stakes of the test changed. She considered the final pilot administration as the low-stakes administration. The high-stakes administration of the test was the operational administration, in which the results were used for state diploma approval. Findings showed higher scores under consequential conditions. Also, the results revealed bigger effects on the constructed-response scores compared with multiple-choice items when increasing the stakes. Regarding item characteristics, the study found that the estimated difficulty is different under both test administrations. Therefore, estimating item difficulty from pilot administrations could lead to inaccurate estimates if it is used to equate test forms to be administered under higher stakes.

Zenisky and Baldwin (2006) found that item properties such as difficulty, cognitive area assessed, and complexity impacted response time latency. They provided evidence that time latency differences are due more to group differences, such as English as a First Language or English as a Second Language for students, than item differences.

Wise and DeMars (2005) synthesized 15 empirical studies that investigated the relationship between test performance and examinee effort in low-stake assessment tests. They found an average effect size exceeding 0.5 standard deviations between the two groups of high and low effort. Test givers would view effect sizes of this magnitude as meaningful.
In many low-stakes test settings, it is known that many examinees put forth a good effort while taking tests even when there are no personal benefits from the assessment experience. The problem comes from the difficulty in assessing how many examinees try hard and how many do not, which makes it difficult to determine the degree to which low examinee effort has biased the test data. Although it may be reasonable for these examinees not to try hard on test items, this change of motivation clearly has implications for the accuracy and quality of the resultant person and item statistics (Bracy, 2007; Eklof, 2006; Parshall, 2002).

**Measuring Test Taking Effort**

Given that insufficient effort may cause a threat to the validity of inferences made regarding assessment results, researchers attempt to measure the degree of effort that examinees expend toward their assessment tests to explore the presence and impact of this probable source of score bias. Most research studies on test-taking motivation have used some type of post-test self-report instruments. For example, in some studies, one or two questions could be asked about the effort that the student put forth in answering the questions. Additionally, in other studies, a longer scale has been used (Halkitis, Jones, & Pradhan, 1996; Wise & DeMars, 2003; Wise, Bhola, & Yang, 2006).

Many studies focused on developing approaches to increase students’ effort during examination by raising the consequence attached to student performance. Wolf and Smith (1995) developed a brief motivational scale, an 8-item Likert-type scale in a single-factor structure that was administered after the college students responded to two parallel tests under two experimental conditions. For one of the conditions, the score was
counted as part of the course grade and for the other condition, it did not count. Wolf and Smith reported that the test condition with credit consequences resulted in significantly higher reported motivation and test score performance.

Sundre (1999) modified the Wolf and Smith (1995) motivational scale into a 10-item scale. The revised scale, called the Student Opinion Scale (SOS), yields a total motivation score and two 5-item subscale scores (named Importance scale and Effort scale). Sundre replicated Wolf and Smith’s design in an attempt to demonstrate differential effect sizes for examinees reporting different motivation levels in both consequences and no-consequences settings. A total of 90 college students participated in the study; each of them received five points for their participation. The results showed significantly higher means of motivation for the consequences condition ($M = 43.11$) compared to the no-consequences condition ($M = 37.05$). This study also investigated the relationship between motivation and performance; the results presented lower and non-significant correlation coefficients between motivation and the consequences of the test performance in contrast to the Wolf and Smith study (Sundre, 1999; Sundre & Moore, 2002).

Smith and Smith (2002) conducted a study to test a hypothesis that the consequences of a test to an examinee influences motivation, performance, and test anxiety. Wolf and Smith’s (1995) survey was used to measure test motivation. Students were encouraged to participate to receive four points toward their final average in the class. Results obtained from 112 undergraduate college students showed greater motivation and performance under the consequential condition, where they had been told
“this test counts for your grade,” as compared to the non-consequential condition, were they had been told “this test does not count for you grade,” with an effect size of 1.58 standard deviation units.

Recently, a brief vignette was used to illustrate how differences in motivation and other psychological processes may contribute to performance on high-stakes mathematics assessment. The differences have been found to be important to achievement. The study presented quotes from interviews with grade eight examinees (Ryan, Ryan, Arbuthnot, & Samuels, 2007).

As an attempt to increase effort and, therefore, performance on mathematics, a subtest of the Iowa Tests of Basic Skills (ITBS) was used by Brown and Walberg (1993). They changed the test instructions to encourage students to try harder. The experimental condition of the test instruction told students the importance of doing their best for themselves, for their teachers, and for their parents. Conversely, a control group of students received the standardized test instructions. The experimental group mean was significantly higher than the control group mean with an effect size of .30 standard deviation units. These results were consistent at the school level, but not when comparing between groups across teachers and schools, suggesting that test instructions may interact with other variables on their effects on student performance. Following the administration of the assessment, students’ efforts were measured by their reactions to the script that was read to them through the post-test interview.

Findings from Smith and Smith (2002), Sundre (1999), and Sundre and Moore (2002) are limited to college courses and testing settings that accompanied these courses.
In addition, the focus of these studies was on performance on course examinations because these are the available settings as reported by Smith and Smith. It is worthy to mention that these examinations involve substantial consequences to students in which they received some extra points for participation in both consequential and non-consequential testing settings.

Another approach found to increase students’ effort during examination and thus increase student performance was by offering financial rewards. O’Neil, Abedi, Miyoshi, and Mastergeorge (2005) provided a monetary incentive to maximize student effort, specified by a self-assessment questionnaire, and, therefore, to increase performance. Twenty items from the released TIMSS 2000 mathematics literacy scale were used to examine the assumption that offering $10 per correct item (high-stakes administration) will motivate students to perform significantly higher in mathematics and to exhibit more effort than those who do not receive a monetary incentive (low-stakes administration). Results showed that students in the incentive group performed no better than the students in the control group; interestingly, student effort was not related to performance on the test.

In a prior study, O’Neil, Sugrue, and Baker (1996) found an increase of 0.41 standard deviation units in test performance. This finding was based on a sub-sample from grade 8 students who received $1 as a monetary incentive compared with students who received standard NAEP instructions. In contrast, a sample of students from grade 12 showed no treatment effects (incentive or no incentive) on either test performance or self-reported effort. In addition, neither instruction that elicited a task-involved goal
orientation, for example, provided the opportunity for a personal accomplishment, nor instructions that elicited an ego-involved goal orientation, for example, comparing each student's mathematical ability with that of other students, resulted in increased effort or achievement on the test.

The previous findings, that motivation varied according to the experimental manipulations practiced in these studies, do not support either the inference that self-report instruments can be used as good indicators of student test-taking effort or that financial rewards improve student test performance or reported effort compared to that obtained from the standardized test instructions. Generally, self-report measures provide information about the general effort that examinees expended during the test setting, but this information failed to provide real effort indices down to the item level. Wise (2006) and Wise and Kong (2005) noted to a large extent that a dynamic change in the level of an examinee’s effort occurred during a testing session.

*Effort based model filtering.* Researchers have attempted to alleviate the problem of low effort responses, such as offering incentives (O’Neil, Sugrue et al., 1996; O’Neil, Abedi et al., 2005), raising the stakes of testing, choosing tests that are not too mentally taxing, and providing feedback (Wise & DeMars, 2005).

Wise, Bhola, and Yang (2006) developed an *effort-monitoring CBT model* which utilizes computer response time latency to classify responses as solution-oriented or rapid-guessing while the test is being administered. In addition, the computer displayed a warning message to examinees committing three consecutive rapid-guessing responses according to short response latencies. The purpose of the warning messages was to
motivate examinees to increase their effort while taking the test. Compared with the no-warning condition, findings showed the warning message condition resulted in increased test performance as a consequence of examinees’ higher effort.

As a follow-up to the previous study, Kong, Wise, Harmes, and Yang (2006) used praise feedback for students who exhibited solution-oriented behavior in an effort to encourage continued solution-oriented behavior and thus influence test performance. This condition was accompanied with the warning messages and control condition employed by Wise, Bhola, and Yang (2006). The difference between the praise feedback condition and each warning message as well as control conditions was not significant; the effect size was rather small. Consistent with their preceding findings, comparisons between the warning messages and the control group revealed that the effects of warnings messages were significant in both test performance and effort.

Another approach to deal with students’ effort would be to incorporate examinee effort indices into the test scoring process. Therefore, examinees’ proficiency estimates were adjusted for their effort during the examination setting. Sundre and Wise stressed “There are currently no psychometric models available for doing this” (Sundre & Wise, 2003, p. 3). Following attempts to deal with low effort filtering, solutions appeared to report and interpret aggregated results only for responses that could be identified as having adequate effort from the examinee. In regard to filtering out responses produced from low effort examinees, the authors argue that these responses are not trustworthy, as they likely underestimate true proficiencies for those examinees.
Sundre and Wise (2003) studied the effects of using various degrees of motivation filtering, based on SOS self-report, descriptive statistics, reliability, and validity for two assessment tests. By applying more rigorous motivation filters, the results showed: (a) the average of student proficiency estimate increased; (b) test reliability remained quite similar; and (c) convergent validity between examination scores and SAT showed an increase from \( r = .41 \) to \( r = .54 \). Sundre and Wise concluded that removing data for less motivated examinees resulted in a reduction of distortions in the assessment of the proficiency levels and rendered results to be more trustworthy. Another way to deal with examinees’ effort in low-stakes assessments could be achieved by altering data analyses to handle the effects of lower motivation.

*Response latency as indicator of examinee effort.* The study of response latency, also known as response time, has been recently investigated. Relatively little attention has been given to response latency in several types of tests (Schnipke & Scrams, 2002; Sundre & Wise, 2003). Because of the attractiveness of using computers in administering tests, it is feasible to collect response latency from a large number of examinees in an unobstructive manner and make use of the collected information to examine item characteristics.

Schnipke (1995) used response time latencies to detect two response strategies on a speeded computer-based test (CBT). The study classified examinees into solution-oriented behavior and rapid-guessing behavior. Solution-oriented behavior indicates examinees tried to determine the correct answer to every item; the response accuracy is determined by item and examinee characteristics. In contrast, in rapid-guessing behavior,
examinees respond rapidly to items as time expires; response accuracy is at near chance level because examinees are not fully considering the items. After removing the random-guessed responses from the data, Schnipke noted a considerable increase in the accuracy to measure ability and an increase in the explained variation for item characteristics to predict response times especially on the last half of the speeded tests.

Ferrando (2006) hypothesized that response time latency increases as the item and person position become closer on the continuum of the trait that is measured. Findings showed weak relationships to support his hypothesis, with effect sizes of 0.11 and 0.18 standard deviation units for both subscales administered.

Schnipke and Scrams (1997a) developed a mixture model to distinguish between solution-oriented behavior and rapid-guessing response strategies on individual items. The model employed response time latencies only, but response accuracy supported the classifications. Accuracy was at or below chance for responses classified as rapid guesses. Their model could be applied to other strategy usage issues as long as the strategies are associated with different response time demands. For example, cheating on items might have fast correct responses under high-stakes settings.

Commonly known, the goal of standardized testing is to assess examinee performance, usually based on the accuracy with which examinees respond to items. More able examinees are assumed to respond more accurately. Wise (2006) reported that response latency patterns might be affected by item position in the test, and some other characteristics that have been investigated. Moreover, it was found that response latency
might be affected by examinee characteristics (e.g., gender, minority groups, speededness in responding, etc.) but not their ability on the domain being measured.

Hadadi and Luecht (1998) analyzed data from two CBTs, speeded and non-speeded tests. They separated data into blocks of 30 items (positions 1-30, 31-60, etc.) and combined response times from these blocks to create stable response time distributions to detect rapid-guessing behavior. The speeded test showed evidence of both omissions which are unanswered items and rapid-guessing behavior (identified as response times of less than 8 seconds), especially on the last 30 items (of 180 items).

Wise, Kingsbury, Thomason, and Kong (2004) compared consequences for two filtering procedures on mathematics scores: self-report measure and Response Time Effort (RTE) index. Findings from 2,382 students from grades 6 through 10 indicated that motivation filtering did not show consistent effects in a K-12 testing context compared with that of higher education students found in Sundre and Wise (2003). The authors attributed these findings to the low number of students who did not give enough effort on their mathematics computer adaptive test (CAT).

Schnipke (1995) studied gender differences relative to effort; she found that rapid-guessing behavior was common among male examinees on an analytical test. In addition, rapid guessing was more common among female examinees on a quantitative test, and equally common on a verbal test.

Halkitis, Jones, and Pradhan (1996) investigated the relationship between response latency and item characteristics such as length, difficulty, and discrimination. The predicted logarithm of response latency from the mentioned item characteristics
found 50.18% of the variance in the logs of time latency could be predicted from these variables. This model could be used as an initial estimate of the total time required to finish the test.

Wolf, Smith, and Birnbaum (1995) conducted a study to examine the interaction of examinee and item characteristics to explain item by item differences in performance. Results from grade 10 and grade 11 students taking a 30-item test were investigated under consequential and non-consequential conditions. Test scores for grade 10 students were used to establish placement in remedial programs for the following year. Findings revealed 10th grade students were significantly more motivated than 11th grade students who had no consequences. In addition, very few items were omitted by grade 10 students: 0.2% versus 1% for grade 11 students. The authors did not notice an increase in omissions of later items on the test for either group. Findings from the mental taxation index and item difficulty were significantly related to the difference in performance between the two groups. The mental taxation index was created by experienced mathematics educators to rate the degree of the mental energy required to solve the question. Research findings supported the notion that the test condition influences student performance and this influence may be substantially different for different kinds of items.

Recent studies have produced promising results such that indices can detect random-guessing responses in many forms leading to targeted action and more effective methods of preventing the test results to be misused. Research showed that item difficulty parameters obtained under low-stakes testing conditions tend to be higher than those obtained under high-stakes testing conditions. Consequently, when IRT parameter
estimates obtained from low-stakes testing conditions are used under higher stakes conditions, the item difficulties are biased, producing inaccurate and high estimates for student ability. Thus, the consequential decisions were made on the basis of inflated student test performance scores (Wise, 2006; Wise & DeMars, 2005; Wise, Wise, & Bhola, 2006). Therefore, precautions should be taken as the low-stakes estimates are important since they may alter the calibration of high-stakes parameter estimates.

Pilot tests share similarities with low-stakes tests, such as the National Assessment of Educational Progress (NAEP), Trends in International Mathematics and Science Study (TIMSS), and various exams conducted by the International Education Association. For these examinations, students are informed that their scores will be anonymous, and will not be counted as part of their individual school progress. Also, they receive no individual feedback. Findings about pilot tests could generalize to these other important low-stakes situations (DeMars, 2000).

Adequate effort put forth by the examinee while answering test questions can be determined by measuring item response latency. The use of computers for test administration provides a unique opportunity to measure response latency based on item level and make it quantifiable. The availability of response latency information suggests a number of useful alternatives for research and allows a detailed investigation of the relationship between the item and examinee characteristics. This relationship tends to be complicated in some testing settings and the role of experts is to resolve these complications in order to ensure measurement accuracy.
Mills (2002) argued the importance of incorporating item response latency with item selection:

For many measurement professionals, response-time analysis is a new area of interest…timing data could be incorporated into item-selection algorithms to lessen the effects of speededness, also it could be used before conducting item analysis to remove data from examinees whose response is so rapid as to indicate they did not attempt the item. (p. 213)

*Remediation for Rapid-guessed Responses*

Maximizing the validity of the test scores involves removing the effects of factors which have nothing in common with the construct being measured by the test. Any new feature added to a test that is not essential to measure the target construct is a potential threat to validity. The effects of such features must be adjusted for using more refined response models (Van Der Linden, 2002). Plake (2002) suggested removing random-guessing responses because these responses represent noise in the item data. The removal of these rapid-guessing responses from data would probably provide more accurate information from test items. In contrast, Wise, Wise, and Bhola (2006) retained only solution-oriented behavior data from 154 examinees out of 488 examinees who took the information literacy test. Findings showed a decrease in both internal consistency and standard errors of measurement, which could be explained by the decrease in the observed score standard deviations.
Rapid-guessing behavior provides little if any information about examinee ability (Schnipke & Scrams, 1997). If some examinees are engaging in rapid-guessing responses, their response latency will be unusually fast and short compared to their colleagues who engaged in solution-oriented behavior.

Schnipke and Scrams (1997b) found that rapid-guessing response was relatively uncommon on verbal and quantitative tests; only a few items were affected. In addition, the authors noticed that rapid-guessing was found on one type on the verbal test (11 out of 38 total items) and for the last five of the 30 items on the quantitative test. Rapid-guessing was more widespread on the reasoning test affecting over half of the 25 items. All examinee responses which were identified as rapid-guessing responses were removed for the analysis.

Schnipke (1995, 1999) and Schnipke and Scrams (1997a) were interested in removing individual random-guessed responses from parameter estimation. It seems that the practice of removing all affected responses would have severely reduced the number of records analyzed.

In classical test theory, the effects of all error variables are assumed to be matched across examinees by following rigorous standardization. As a result, the only systematic difference among examinees is their true scores on the variable measured. All remaining variation among the observed scores is random (measurement error). IRT models were developed to adjust for the effects of item properties: difficulty, discrimination, and guessing. These properties are systematic errors interfering with the variable to be measured. Because IRT models have precise parameters for these item properties,
estimates of the examinees’ abilities are adjusted for their effects. Consequently, these abilities are measured on a common scale (Van Der Linden, 2002).

The literature offers basic techniques to control the response latency and incorporate item latency with a particular IRT model (McCall & Bontempo, 2006; Morch & Bolt, 2006; Schnipke & Scrams, 1997). Utilizing raw data rather than relying on item response alone has been more challenging. Thissen’s model in 1983 or Schnipke and Scrams’ mixture model in 1997 seems to be promising for scaling examinees or making decisions regarding response latency assessment. Within CBT administrations, omissions may be problematic but could perhaps be allowed. Omissions would have to be scored as wrong responses, but limitations would be necessary since it is undesirable to have an excessive proportion of omissions (Folk & Smith, 2002).

Early in 1994, Bhola developed an algorithm to identify examinees who spent much less time than expected on particular items of the computer-administered Graduate Record Examination (GRE) General Test. Number of words and response latency were used to compute an examinee's mean working rate on items from a particular content area of the test. Working rate was used in combination with the number of words of later items of similar format from the same content area to compute the expected response latency on these items. If the examinee spent less than a fraction of this response latency on a particular item, the response was identified and rescored as a not reached item. The algorithm was developed using a sample of 3,000 examinees, and evaluated using another sample of 3,000 examinees drawn from a population of examinees who took Form J of the GRE. IRT calibrations, using PC-BILOG 3, were completed on both original and
rescored data. Analyses revealed that, generally, across the three content areas of the test, rescoreing using the proposed algorithm did not improve the precision of item and person parameter estimates (Bhola, 1994).

Recently, Wise and Kong (2005) noticed that rapid-guessing behaviors are also present in the data from un-speeded, low-stakes CBTs. Moreover, they found that rapid-guessing responses can occur throughout a test and not just toward the end as Bhola (1994) and Schnipke (1995) had observed with speeded high-stakes tests. Wise and Kong developed the RTE index for measuring the overall test-taking effort. RTE scores were found to: (a) show significant correlations with other measures of examinee effort, (b) be uncorrelated with measures of academic ability, and (c) reveal motivation filtering effects similar to those found with other measures of examinee effort (Sundre & Wise, 2003; Wise & DeMars, 2005).

Identification of Response Latency Threshold

Several methods were used to identify response latency threshold. Schnipke and Scrams (1997a) crossed response distributions for each item by plotting both correct and incorrect responses at each response time level, which were used as the time thresholds to identify rapid-guessing responses. Wise and Kong (2005) established three response thresholds based on each item’s stem and options. Each item was assigned to one of the three thresholds: a) 3 seconds if the item contains less than 200 characters, b) 10 seconds if the item was longer than 1,000 characters or if the item provided with ancillary table or graph, and c) 5 seconds for the remaining items. Wise (2006) and Wise and DeMars (2006) identified time threshold for items in the information literacy test by examining
the response time distribution and then, based on the distribution, they visually chose a value at the end of the first frequency point. Wise (2006, p. 22) provided an example, shown in figure 1, to describe the frequency distribution of examinee response times for one of the test items.

![Graph of response time distribution](image)

FIGURE 1. *Distribution of examinee response times for an information literacy test item.*

The visually scanned suggested threshold for this item would be about 4 seconds.

*Summary*

Overall, the research evidence clearly supports the claim that test-taking motivation affects both test performance and test score validity. The developed approaches to increase student effort varied according to the consequential manipulations and revealed non-consistent results related to using a particular method to alleviate low effort responses. Most of the research conducted in the area of effort filtering involved college students, and effort filtering was based on self-report measures (Sundre, 1999;
Sundre & Moore, 2002; Sundre & Wise, 2003; Wise & DeMars, 2005; Wise & Kong, 2005). The proposed study will be structured to analyze data obtained from high-school examinees using a computer administration of a mathematics test. Response latency was recorded at the item level and was used as an indicator of examinee effort while taking the test.

This study is a response to the dissimilar findings from previous research. On different occasions Wise and his colleagues have called for more research on motivation filtering to better understand the moderated effort model and how to best use this model in assessment settings (Sundre & Wise, 2003; Wise & DeMars, 2005; Wise & Kong, 2005; Wise, Wise, & Bhola, 2006). The goal of this investigation is to evaluate the consequences of using an effort filter and rescoring the rapid-guessed responses accordingly: zero, not presented, omitted, or original response based on CTT and IRT frameworks. Moreover, this study will examine the effects of rescoring methods on students’ Pass/Fail results rather than filtering out rapid-guessing responses.
CHAPTER THREE

Methodology

Rapid-guessing behavior is often observed when examinees do not put forth sufficient effort and time to answer test questions. When examinees are less motivated, they are unlikely to give adequate effort throughout the test which creates problems in presenting what they know and may result in biased estimates for ability and item parameter estimates. The objective of the current study is to evaluate the use of different scoring procedures on CTT and IRT parameter estimates for dichotomously scored items that are obtained from computer-administration after identifying rapid-guessing responses.

This chapter provides a description of the methodology used in the study and will be divided into four parts. Part one presents a description of the sample. Part two will describe the test and the CBT administration. Part three will describe data collection procedures. The last part will describe the analysis plan for the data of this study according to the research questions.

Sample

The current study will employ information from existing data that was collected by teachers from five different schools in Jordan in 2006. Response time latency was recorded during administration of a computer-based mathematics test. The researcher assisted teachers in the collection of these data.

Ninth-grade students from high schools in Amman, Jordan were assessed for mathematics skills. The sample includes 586 students from five high schools of which
three were private schools and two were public schools. Of the sample, 322 were males and 264 were female students. Students ranged in age from 13 to 15 years with an average of 13.99 years and a standard deviation of 0.18.

The students were aware that there were few individual consequences of the testing administration in that only teachers would know the test results and these would be at the class- and not individual-level. There were no consequences for individual students’ test performance in that the test scores will not be counted for their school grades. Each student completed the entire mathematics test in a computer-based setting in the computer lab mandated by their schools. Students were tested in groups of 15 to 20.

**Instruments**

The test consisted of dichotomously-scored, computer-administered, multiple-choice mathematics items. Individual item responses and the amount of time taken by the examinee to read, review, and answer individual items were recorded for each student.  

**The Mathematics Test**

The test used in this study is a portion of mathematics literacy test from the Arabic translation of the TIMMS, 2003 released items (Appendix 1, original items in English, Appendix 2). The test consisted of 35 multiple-choice items with four or five response options per item. Items were classified according to the following content domains: (a) Numerical Operations, (b) Measurement and Geometry, (c) Data Analysis, and (d) Fundamentals of Algebra. Some of the items use graphs and short reading passages. The purpose of the test was to provide teachers with a brief, convenient, computer-administered, class-level measure of student achievement in mathematics that
could be used for a variety of instructional purposes. Since the data would only be used at the class-level, there was concern that some students might not put forward maximum effort. Given that, students perceive the test as a low-stakes test.

The test was administered via a computer-based assessment. The CTT difficulty estimates were reasonable based on the original students’ responses before considering the latency information attached to each item. Additional information included self-reported birth date, school name, students’ mathematics achievement, and GPA in grade eight.

*Computer-based administration program.* A test administration software program was developed. The software recorded both the examinee’s response and response time latency for each item on the test. Response time latency is identified as the number of seconds elapsed between the display of the item on the screen and an examinee’s submission of the response. The intention of using response time latency was to identify items that were answered in less time than the latency threshold, and rescore these responses into one of the suggested scoring procedures. Subsequent to this, analyses were performed to determine whether the rescored data results in different parameter estimates for items and persons based on CTT and IRT models.

Visual Basic 6.0 (VB6.0) software was used to computerize the test administration. The test items were prepared in a way that did not affect the test appearance using different versions of Windows OS and/or different computer configurations within different settings.
The statement of the problem and the response options for each of the 35 questions were converted to a single JPEG image. Radio buttons were inserted on top of each response on the image using Visual Basic 6.0 code. Only one test item was presented on a screen (Appendix 3). The choices were identified based on which radio button was selected. Then the examinee clicks a “NEXT ITEM” button on the lower middle of the screen to obtain the next item. After the student completes the test, a pop-up screen appears to thank him/her for cooperation and to inform the student of the end of the testing session.

For each examinee, the computer recorded the response for each individual item and the time taken to answer that particular item. The time counter starts from the appearance of the item on the screen until the student presses the “NEXT ITEM” button. After the student completed the test, his/her data were compiled and submitted to a dBase file which was created to transfer data for later analysis. The test was administered as a fixed format, giving the same sequence of items to all students.

**Procedures**

The test was delivered as a computer-based test (CBT), administered at the computer lab supervised by the school personnel. Computer labs varied in capacity from 15 to 20 computers. Orientation to the test and the nature of the task that is required from each student was implemented first, followed by the administering of test items. During the test administration, examinees were not allowed to omit items; that is, they were required to provide an answer to each item before they could go on to the next item. Also, examinees were not allowed to go back or revise previous items. Returning to rework
items is problematic because this will generate different decisions with regard to the time latency for the reworked items. However, examinees were allowed to change their responses within each item.

Students were directed to work through the test at their own pace. Students had no prior exposure to computer-based tests as reported by teachers. All students received the following test instructions: *Today you will take a test that measures your mathematics literacy skills. For each item, you are to choose the best answer. Please proceed at your own rate. Note neither you, your parents, nor peers will know the result of this exam.* An example was then presented to allow students to practice selecting the desired response from the options offered for each item. The students were directed to leave the computer lab after finishing the test and return to their classroom. The response time data indicated that the examinees took a reasonable time to complete the test.

*Analysis*

The data for the current study were obtained from a computer-administered, multiple-choice mathematics test. Each student has one record containing his or her responses, time latency, and additional descriptive information.

The time latency thresholds used to distinguish rapid-guessing responses from more thoughtful solution-oriented responses were calculated based on results from an initial and separate group of 20 students. Those students were instructed to read each item and all its possible options, without trying to think about the content or trying to solve the problem, then click the next button to go on for the next item. This method was employed to identify or calibrate individual item latency thresholds. Medians for response times in
seconds for each item were computed. The median is preferred to the mean for this purpose because medians are less susceptible to outliers, thresholds ranged from 4 to 15 seconds, with a median value of 8 seconds (Table 18, Appendix D).

An additional measure of overall effort given during the test was calculated based on RTE index, which shows the proportion of items reflecting solution-oriented behavior during the test (Wise & Kong, 2005). Commonly, RTE scores range between zero and one, with a value of one indicating highest effort (i.e., solution-oriented behavior exhibited on all items) and a value of zero indicating insufficient effort on all items. The intention of using response time latency is to identify responses that were generated in less time than a pre-determined threshold and flag them as rapid-guessing responses.

Four data files were extracted from the original data file; each file contains responses from students who engaged in solution-oriented behavior where items were coded 0 for an incorrect response or 1 for a correct response. The guessed responses was recoded to 9 and scored in the different ways indicated below.

The basic hypothesis is that students who are not motivated to perform well on a test will be likely to show short response latencies compared to motivated students. Student performance on items will likely be different according to the effort put forth on items. Parameter estimates for items and examinees were calibrated using the 3-parameter logistic IRT model using PC-BILOG 3.02 (Mislevy & Bock, 1990).

*Description of IRT.* Item response theory is a mathematical framework developed to model the relationship between abilities (or traits) and item responses. The three parameter logistic IRT model, 3PL-IRT, models the relationship between each examinee
performance and a set of parameters underlying items that can be described by a monotonically increasing function called an item characteristic function or item characteristic curve (ICC). The item parameters of interest are difficulty, discrimination, and pseudo guessing (Hambleton, Swaminathan, & Rogers, 1991).

Each data set was calibrated in a separate run. Four calibrations were done, one for each scoring method, “Not Presented,” “Omitted,” “Zero,” and the last one for the “Original” responses. PC-BILOG 3.02 requires a data set structured such that the input file consists of an examinee identification number with 3 digits and numbers of 0s, 1s, or 9s, representing each examinee’s responses on the 35 items. Where incorrect responses are scored as 0s, correct responses are scored as 1s under solution-oriented behavior, while the guessed responses are replaced by 9s.

Data file 1 contains the original data as they were gathered from the examinees without considering the rapid-guessing behavior. This file consists of 3-digit identification number followed by numbers of 0s or 1s, representing each examinee’s responses to the 35 items.

Data file 2 consists of 3 digit identification number followed by numbers of 0s, 1s, or 9s representing each examinee’s responses to the 35 items where the score 9 is a value given for responses generated in less time than the pre-determined threshold. This score was used later to create a “not presented” key file.

Data file 3 consists of 3 digit identification number followed by numbers of 0s, 1s, or 9s representing each examinee’s responses to the 35 items where the score 9 is a
value given for responses generated in less time than the pre-determined threshold. This score was used later to create an “omitted” key file.

Data file 4 consists of 3-digit identification number followed by numbers of 0s or 1s representing each examinee’s responses to the 35 items where the responses generated in less time than the pre-determined threshold were rescored to 0s instead of 9s.

In addition to these data files, two files were created, a “not presented” key file and an “omitted” key file related to data file 2 and data file 3, respectively. After creating these files, four separate calibrations were executed for data files according to the resoring procedure. The output from phase three of each run were edited to give data for item and person parameter estimates. The edited data were used as input files into SPSS.

The differences in item parameter estimates: difficulty, discrimination, and guessing, and person parameter estimates calibrated with regard to different scoring methods were examined utilizing Pearson-correlation coefficients. Repeated measures analysis were employed to facilitate comparisons and to find the significance of differences, if any, between parameter estimates for persons and items across different scoring procedures. Details of the analyses follow shortly.

A classical item analysis was conducted for the four data files, to find item difficulty estimates, percentage of students selecting the correct answer, and item discrimination indexes. Point-biserial correlation coefficients were computed for each test item. Test score reliability were expressed by the Kuder-Richardson 20 (KR20) coefficient which is equivalent to coefficient alpha for dichotomous data.
For purposes of IRT calibration, the probability of guessing in the omitted procedure was set to \( g \) where \( g \) equals to the reciprocal number of options per item. The assumption is that any examinee who makes rapid guesses would not be expected to have time to eliminate distractors or obviously incorrect options. Therefore, we assume that the student will succeed as if randomly guessing from the given options.

The general plan for the data analysis in SPSS was to compute descriptive statistics for item responses and time latencies. Next, students deemed to show guessed-responses were identified from the sample; according to response latencies lower than a particular threshold for each item. These responses were rescored to either, zero, not presented, omitted, or original selection. The same statistics were computed for the resulting data files.

The standard 3PL-IRT parameters were estimated using BILOG 3.02 (Mislevy & Bock, 1990). The full data set of \( N = 589 \) students was used to estimate IRT parameters and CTT indices. The calculations for the total score of each examinee varied according to the scoring procedures. For the Omitted procedure, the total score was composed of the number of correct responses under solution-oriented behavior plus the probability of guessing for the responses under rapid-guessing behavior. For the Not-presented procedure, the total score was composed of the number of correct responses under solution-oriented behavior only. For the Zero procedure, the total score was composed of the number of correct responses under solution-oriented behavior plus zero for the responses under rapid-guessing behavior.
The purpose of this study was to evaluate the consequences of using different scoring procedures on CTT and IRT person and item parameter estimates taking into consideration response latency. In particular, the current study focused on answering the following research questions:

1. Is there a significant proportion of students who exhibit rapid-guessing behavior?

Null hypothesis concerning proportion: proportion of students engaged in rapid-guessing behavior is different from zero.

\[ H_0: \pi_r = 0 \]

where \( \pi_r \) is the proportion of students engaged in rapid-guessing behavior on the test.

A confidence interval for proportions was constructed to examine the difference of the mentioned proportion from zero using \( \alpha = .05 \) and \( DF = 1 \).

The RTE index was used to indicate the percentage of items with solution-oriented behavior. Rapid-guessing responses were detected and rescored differently using different scoring procedures. The resulting parameter estimates for items and persons may differ from those obtained based on the original responses, which include data of rapid-guessed responses. One of the current research goals was to examine the consequences of rapid-guessing behavior and whether this would seriously influence classical item statistics, such as item means (difficulty), item-total correlations (discrimination), and test score reliability; when employing response latency. More specifically:
2. Are there significant differences in classical item indices (item means, item-total correlations, and reliability) across the different scoring procedures?

Null hypothesis concerning CTT item difficulty: classical item difficulty indices are the same across the four scoring procedures.

\[ H_{0a} : \bar{p}_{NP} = \bar{p}_Z = \bar{p}_O = \bar{p}_D \]

Where \( \bar{p} \) is the mean of CTT item difficulty indices,

subscript NP: not presented; Z: Zero; O: Omitted; D: Original Data.

Null hypothesis concerning CTT item discrimination: classical item discrimination indices are the same across the four scoring procedures.

\[ H_{0b} : \bar{D}_{NP} = \bar{D}_Z = \bar{D}_O = \bar{D}_D \]

Where \( \bar{D} \) is the mean of CTT item discrimination indices.

Null hypothesis concerning reliability coefficient: reliability coefficient KR20 is the same across the four scoring procedures.

\[ H_{0c} : KR20_{NP} = KR20_Z = KR20_O = KR20_D \]

Where KR20 is the Kuder-Richardson coefficients for reliability.

Repeated measures analysis of variance were employed to facilitate the comparisons and to find the significance of differences, if any, between CTT estimated indices (difficulty and discrimination) across the four scoring procedures. ALPHATST, a computer program for testing dependent reliability coefficients (Feldt, Woodruff, & Salih, 1987; Woodruff & Feldt, 1986), was used to test for the significance of differences between KR20 reliability coefficients across the four scoring procedures;
3. Do the different scoring procedures yield significantly different IRT item parameter estimates?

Null hypothesis concerning IRT item difficulty estimates: item difficulty parameter estimates based on 3PL-IRT are the same across the four scoring procedures.

\[ H_{0, a} : \bar{b}_{NP} = \bar{b}_Z = \bar{b}_O = \bar{b}_D \]

Where \( \bar{b} \) is the mean of IRT difficulty estimates,

subscript NP: not presented; Z: Zero; O: Omitted; D: Original Data.

Null hypothesis concerning IRT item discrimination estimates: item discrimination parameter estimates based on 3PL-IRT are the same across the four scoring procedures.

\[ H_{0, b} : \bar{a}_{NP} = \bar{a}_Z = \bar{a}_O = \bar{a}_D \]

Where \( \bar{a} \) is the mean of IRT discrimination estimates,

subscript NP: not presented; Z: Zero; O: Omitted; D: Original Data.

Null hypothesis concerning IRT guessing estimates: item guessing parameter estimates based on 3PL-IRT are the same across the four scoring procedures.

\[ H_{0, c} : \bar{c}_{NP} = \bar{c}_Z = \bar{c}_O = \bar{c}_D \]

Where \( \bar{c} \) is the mean of IRT guessing estimates,

subscript NP: not presented; Z: Zero; O: Omitted; D: Original Data.

The standard 3PL-IRT item and person parameter estimates were calibrated using PC-BILOG 3.02 for research questions three and four. The resulting item parameter estimates were edited and examined in SPSS for the differences across the four scoring procedures using repeated measures analysis of variance.
4. Do the different scoring procedures yield significantly different IRT person parameter estimates?

Null hypothesis concerning IRT person estimates: person parameter estimates based on 3PL-IRT are the same across the four scoring procedures.

\[ H_0 : \bar{\theta}_{NP} = \bar{\theta}_Z = \bar{\theta}_O = \bar{\theta}_D \]

Where \( \bar{\theta} \) is the mean of IRT person estimates, subscript NP: not presented; Z: Zero; O: Omitted; D: Original Data.

Person parameter estimates from PC-BILOG 3.02 calibration were edited and examined in SPSS for the differences across the four scoring procedures using repeated measures analysis of variance.

5. Do pass/fail decisions for each examinee differ across the different scoring procedures?

Null hypothesis concerning pass/fail decisions: pass/fail decisions are the same across the four scoring procedures.

Arbitrarily, four cut-scores were derived to build up pass/fail decisions. The cut-score was computed based on the number of items that represents the deciles 50, 60, 70, and 80. The proportions of pass/fail decision consistency were computed to examine if the proportions appear to increase or decrease across the scoring procedures.
CHAPTER FOUR

Results

The current study employed response latency as an indicator of examinees rapid-guessing behavior based on response latencies lower than a particular threshold for each item. The identified responses were rescored differently using the Omitted, the Not-presented, or the Zero scoring procedures. The objective of the study was to evaluate the use of different scoring procedures on CTT and IRT parameter estimates for dichotomously scored items that were obtained from computer-administration after identifying rapid-guessing responses. The full data set of $N = 586$ students was used to calibrate the standard 3PL-IRT parameters estimates using PC-BILOG 3.02.

This chapter provides general descriptive results for examinees rapid-guessing responses followed by description of the study results according to the following research questions:

1. Is there a significant proportion of students who exhibit rapid-guessing behavior?
2. Are there significant differences in classical item indices (item means, item-total correlations, and reliability) across the different scoring procedures?
3. Do the different scoring procedures yield significantly different IRT item parameter estimates?
4. Do the different scoring procedures yield significantly different IRT person parameter estimates?
5. Do pass/fail decisions for each examinee differ across the different scoring procedures?

**General Descriptive Analysis**

On average, 12% of the examinees’ responses were classified as random-guessed responses on the test. The inspection of the percentages across test items revealed that the solution-oriented responses decreased in the last third of the test items in an inconsistent manner (see Figure 2).

![Figure 2. Frequency of solution-oriented responses across items.](image-url)
Figure 3 shows the frequency of rapid-guessed responses during the test. The value 0 on the X-axis represented the non-occurrence of guessing and the value 1 on the X-axis means the occurrence of one rapid-guessing response and so on. Approximately one third of the examinees, 205 of the examinees, engaged in solution-oriented behavior, 110 of the examinees exhibited guessing behavior on one item, while 48 examinees exhibited guessing on two items (see Table 3, Appendix E). A small number of
examinees exhibited more than four guesses during the test and a good number of examinees showed four guessed responses or less, as shown in Figure 3.

*Figure 4.* Frequency of correct and incorrect responses under rapid-guessing behavior.

The total number of responses that produced under rapid-guessing behavior was 2407 of which 533 responses were correct with a percentage of 0.22 and 1874 responses were incorrect with a percentage of 0.78.
Figure 5. Percentage of correct and incorrect responses under rapid-guessing behavior.

The percentages showed in Figure 5 revealed that most of the responses generated under rapid-guessing behavior were incorrect; on average 22% of the guessed responses were correct and 78% were incorrect.

The RTE index was computed to show the proportion of test items for which the examinees exhibited solution-oriented behavior. The total number of responses that were produced under rapid-guessing behavior, based on interaction between the examinees and items, was 2407 across 20510 possible responses. The mean of RTE = 0.88 indicates that about 12% of the total item responses were generated under rapid-guessing behavior.
Table 1 shows the descriptive statistics for the number of guessed responses for both persons and items. Each of the items received rapid-guessing behavior from at least 12 examinees (item 6) and a maximum of 190 examinees (item 34).

Table 1

*Descriptive Statistics for Guessing-behavior*

<table>
<thead>
<tr>
<th>Guessing responses</th>
<th>N</th>
<th>Min</th>
<th>Max</th>
<th>Sum</th>
<th>M</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Based on items</td>
<td>35</td>
<td>12</td>
<td>190</td>
<td>2407</td>
<td>68.77</td>
<td>45.62</td>
</tr>
<tr>
<td>Based on examinees</td>
<td>586</td>
<td>.00</td>
<td>32</td>
<td>2407</td>
<td>4.11</td>
<td>6.09</td>
</tr>
</tbody>
</table>

Two students possibly considered as influential cases because they committed the rapid-guessing behavior on 32 items (see Table 3, Appendix E). It is useful to know if the rapid-guessing responses were executed in the same manner for all items and to test whether the proportion of rapid-guessing responses is different from zero. The confidence interval (CI) for the proportion of guessing committed per item was constructed based on the following formula; the results are shown in Table 2. The CI for a proportion, \( \pi \), is

\[
P - [Z_{1-\alpha/2} \times SE_p] \to P + [Z_{1-\alpha/2} \times SE_p]
\]

Where \( P \) is the proportion of examinees committed rapid-guessing behavior, \( Z_{1-\alpha/2} \) is the percentile from the standard normal distribution. Thus, for a 95% CI \( Z_{1-\alpha/2} = 1.96 \). \( SE_p \), the standard error of the proportion, is equal to

\[
SE_p = \sqrt{\frac{p(1-p)}{N}}
\]

Where \( N \) is the number of observations.
Table 2

Confidence Intervals for Proportion of Guessing Committed Based on Item Level

<table>
<thead>
<tr>
<th>Item</th>
<th>Guessing (p) %</th>
<th>$SE_p$</th>
<th>L95%*</th>
<th>U95%**</th>
</tr>
</thead>
<tbody>
<tr>
<td>q1</td>
<td>0.038</td>
<td>0.008</td>
<td>0.022</td>
<td>0.053</td>
</tr>
<tr>
<td>q2</td>
<td>0.085</td>
<td>0.012</td>
<td>0.063</td>
<td>0.108</td>
</tr>
<tr>
<td>q3</td>
<td>0.038</td>
<td>0.008</td>
<td>0.022</td>
<td>0.053</td>
</tr>
<tr>
<td>q4</td>
<td>0.036</td>
<td>0.008</td>
<td>0.021</td>
<td>0.051</td>
</tr>
<tr>
<td>q5</td>
<td>0.049</td>
<td>0.009</td>
<td>0.032</td>
<td>0.067</td>
</tr>
<tr>
<td>q6</td>
<td>0.020</td>
<td>0.006</td>
<td>0.009</td>
<td>0.032</td>
</tr>
<tr>
<td>q7</td>
<td>0.031</td>
<td>0.007</td>
<td>0.017</td>
<td>0.045</td>
</tr>
<tr>
<td>q8</td>
<td>0.029</td>
<td>0.007</td>
<td>0.015</td>
<td>0.043</td>
</tr>
<tr>
<td>q9</td>
<td>0.039</td>
<td>0.008</td>
<td>0.024</td>
<td>0.055</td>
</tr>
<tr>
<td>q10</td>
<td>0.070</td>
<td>0.011</td>
<td>0.049</td>
<td>0.091</td>
</tr>
<tr>
<td>q11</td>
<td>0.055</td>
<td>0.009</td>
<td>0.036</td>
<td>0.073</td>
</tr>
<tr>
<td>q12</td>
<td>0.048</td>
<td>0.009</td>
<td>0.031</td>
<td>0.065</td>
</tr>
<tr>
<td>q13</td>
<td>0.097</td>
<td>0.012</td>
<td>0.073</td>
<td>0.121</td>
</tr>
<tr>
<td>q14</td>
<td>0.073</td>
<td>0.011</td>
<td>0.052</td>
<td>0.094</td>
</tr>
<tr>
<td>q15</td>
<td>0.048</td>
<td>0.009</td>
<td>0.031</td>
<td>0.065</td>
</tr>
<tr>
<td>q16</td>
<td>0.114</td>
<td>0.013</td>
<td>0.089</td>
<td>0.140</td>
</tr>
<tr>
<td>q17</td>
<td>0.080</td>
<td>0.011</td>
<td>0.058</td>
<td>0.102</td>
</tr>
<tr>
<td>q18</td>
<td>0.099</td>
<td>0.012</td>
<td>0.075</td>
<td>0.123</td>
</tr>
</tbody>
</table>

*(table continues)*
Table 2 (continued)

<table>
<thead>
<tr>
<th>Item</th>
<th>Guessing (p) %</th>
<th>SE_p</th>
<th>L95%*</th>
<th>U95%**</th>
</tr>
</thead>
<tbody>
<tr>
<td>q19</td>
<td>0.111</td>
<td>0.013</td>
<td>0.086</td>
<td>0.136</td>
</tr>
<tr>
<td>q20</td>
<td>0.131</td>
<td>0.014</td>
<td>0.104</td>
<td>0.159</td>
</tr>
<tr>
<td>q21</td>
<td>0.130</td>
<td>0.014</td>
<td>0.102</td>
<td>0.157</td>
</tr>
<tr>
<td>q22</td>
<td>0.130</td>
<td>0.014</td>
<td>0.102</td>
<td>0.157</td>
</tr>
<tr>
<td>q23</td>
<td>0.121</td>
<td>0.013</td>
<td>0.095</td>
<td>0.148</td>
</tr>
<tr>
<td>q24</td>
<td>0.177</td>
<td>0.016</td>
<td>0.147</td>
<td>0.208</td>
</tr>
<tr>
<td>q25</td>
<td>0.147</td>
<td>0.015</td>
<td>0.118</td>
<td>0.175</td>
</tr>
<tr>
<td>q26</td>
<td>0.253</td>
<td>0.018</td>
<td>0.217</td>
<td>0.288</td>
</tr>
<tr>
<td>q27</td>
<td>0.302</td>
<td>0.019</td>
<td>0.265</td>
<td>0.339</td>
</tr>
<tr>
<td>q28</td>
<td>0.128</td>
<td>0.014</td>
<td>0.101</td>
<td>0.155</td>
</tr>
<tr>
<td>q29</td>
<td>0.162</td>
<td>0.015</td>
<td>0.132</td>
<td>0.192</td>
</tr>
<tr>
<td>q30</td>
<td>0.164</td>
<td>0.015</td>
<td>0.134</td>
<td>0.194</td>
</tr>
<tr>
<td>q31</td>
<td>0.206</td>
<td>0.017</td>
<td>0.174</td>
<td>0.239</td>
</tr>
<tr>
<td>q32</td>
<td>0.177</td>
<td>0.016</td>
<td>0.147</td>
<td>0.208</td>
</tr>
<tr>
<td>q33</td>
<td>0.169</td>
<td>0.015</td>
<td>0.139</td>
<td>0.199</td>
</tr>
<tr>
<td>q34</td>
<td>0.324</td>
<td>0.019</td>
<td>0.286</td>
<td>0.362</td>
</tr>
<tr>
<td>q35</td>
<td>0.225</td>
<td>0.017</td>
<td>0.191</td>
<td>0.259</td>
</tr>
</tbody>
</table>

*: L95% is the lower limit for the CI.

**: U95% is the upper limit of the CI.
The proportion of guessing shown in Table 2 showed that rapid-guessing behavior was present in all items since none of the CIs contains the value 0. However, rapid-guessing responses were committed more in the last third of the test and the lower limits of the confidence intervals turn out to be far from capturing the null value of zero.

**Analysis for Research Question One**

Is there a significant proportion of students who exhibit rapid-guessing behavior? A confidence interval (CI) for proportions was constructed to examine if the proportion of examinees committing rapid-guessing behavior, 0.65, is different from zero. Based on the prior formula for calculating confidence intervals, the 95% CI for the population value of the proportion of examinees who exhibited rapid-guessing behavior is from 0.611 to 0.689. Therefore, the 95% CI indicated that the proportion of students who committed rapid-guessing behavior is significantly different from zero and there is a 95% chance the indicated range of CI may include the proportion of students’ committed rapid-guessing behavior in the population.

**Analyses for Research Questions 2, 3, and 4**

A single group within-subjects design (repeated measure) ANOVA was used to assess changes in item parameters across scoring procedures. Separate repeated measures ANOVA analyses were conducted comparing the differences in both CTT and IRT frameworks.

**Assumptions.** Prior to the execution of the repeated measures ANOVA analyses, the following statistical assumptions related to the use of repeated measures analysis of variance were evaluated.
1. All measures are normally distributed. The data were verified to determine if the normality assumption was met using Kolmogorov-Smirnov Z statistic and values of skewness and kurtosis. Table 20 and 21 (Appendix F) show the Kolmogorov-Smirnov results of classical item difficulty and discrimination indices respectively across four scoring procedures.

Kolmogorov-Smirnov Z values were not significant ($\alpha = 0.05$) for both classical difficulty and discrimination indices across scoring procedures; therefore, the normality assumption was met. Table 22, 23, 24, and 25 (Appendix F) show the Kolmogorov-Smirnov results for IRT parameters A, B, C, and ability respectively across four scoring procedures. Table 26 (Appendix F) show the skewness and kurtosis statistics for IRT ability parameter.

Kolmogorov-Smirnov Z results for IRT item parameters across different scoring procedures indicated a non-significant ($\alpha = 0.05$) statistic. In Table 9 the results of Kolmogorov-Smirnov Z for IRT ability parameter show significant ($\alpha = 0.05$) values under three scoring procedures; consequently, skewness and kurtosis statistics were examined. Non-significant deviation from normality is shown in Table 26 (Appendix F). Therefore, IRT item and person parameters across four scoring procedures can be assumed to have a normal distribution.

2. Independence of observations. In the current study, each examinee worked independently from each other examinee.
3. Homogeneous variance-covariance matrices across groups. This assumption is necessary only if there is more than one group of participants (Barcikowski & Robey, 1994); therefore, this assumption does not apply for the current study design.

4. Repeated-measures analysis of variance is also based on assumption of sphericity concerning the variances of the scores and the correlations among the scores. The sphericity assumption is satisfied if (a) the variances of the repeated measurements in the population are equal and (b) the correlations among all pairs of measures in the population are equal. Violation of the assumption of sphericity results in an increase in the type I error rate and one could say that the $F$ value is positively biased.

If the sphericity assumption is violated then an alternative approach can be used to adjust the univariate test degrees of freedom based on an epsilon value as a correction factor. An estimate of epsilon ($\varepsilon$) is provided by Geisser-Greenhouse. The correction is used when $\varepsilon < 0.75$ because this method returns to the limits of the robustness interval given sufficient sample size ($N = 30$ or $N = 60$). Conversely, when $\varepsilon \geq 0.75$, an estimate provided by Huynh and Feldt, also denoted by $\varepsilon$, is used because the Geisser-Greenhouse correction factor has been shown to be too conservative under all sample sizes. It sometimes fails to detect a true difference between group means (Barcikowski & Robey, 1994; Berkovits, Hancock, & Nevitt, 2000). Barcikowski and Robey (1994) commented that the sphericity assumption is rarely met in practice.
Analyses for Research Question Two

Are there significant differences in classical item indices (item means, item-total correlations, and reliability) across the different scoring procedures? Table 3 shows means and standard deviations for the CTT indices across four scoring procedures.

Table 3
Descriptive Statistics for CTT Difficulty and Discrimination Indices

<table>
<thead>
<tr>
<th>CTT indices</th>
<th>Scoring procedure</th>
<th>Min</th>
<th>Max</th>
<th>M</th>
<th>SD</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Difficulty</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Original response</td>
<td>.120</td>
<td>.730</td>
<td>.45200</td>
<td>.154802</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td>Omitted</td>
<td>.150</td>
<td>.810</td>
<td>.52771</td>
<td>.192966</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td>Not-presented</td>
<td>.180</td>
<td>.730</td>
<td>.45571</td>
<td>.149061</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td>Zero</td>
<td>.100</td>
<td>.720</td>
<td>.42629</td>
<td>.157370</td>
<td>35</td>
</tr>
<tr>
<td>Discrimination</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Original response</td>
<td>.043</td>
<td>.448</td>
<td>.26580</td>
<td>.100518</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td>Omitted</td>
<td>-.108</td>
<td>.405</td>
<td>.23446</td>
<td>.108647</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td>Not-presented</td>
<td>-.056</td>
<td>.450</td>
<td>.26491</td>
<td>.103282</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td>Zero</td>
<td>.097</td>
<td>.492</td>
<td>.30960</td>
<td>.100797</td>
<td>35</td>
</tr>
</tbody>
</table>

A classical item analysis indicated that classical item difficulty, $p^\hat{}$, ranged from 0.12 to 0.73 with a mean of 0.45 and a standard deviation of 0.15 for the original responses. Means of $p^\hat{}$ values, according to the scoring procedure, vary from that of the original scoring procedure except for the Not-presented procedure where the $p^\hat{}$ is similar to the original scoring procedure. The point-biserial correlations, $r_{pb}$, ranged from 0.04 to 0.45 with a mean of 0.27 and a standard deviation of 0.10. Similar tendency to $p^\hat{}$, the
mean of $r_{pb}$ varied according to scoring procedure except for the Not-presented procedure where the mean of $r_{pb}$ is almost the same as with the original responses.

The null hypothesis concerning CTT item difficulty is: classical item difficulty indices are the same across the four scoring procedures. Repeated measures analyses of variance were used to determine the significance of the differences of CTT indices across scoring procedures. Mauchly’s W test indicated that the assumption of sphericity has been violated ($\chi^2 (5) = 74.60, p < 0.001$), therefore the relevant degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity ($\varepsilon = 0.438$) (see Table 27 in Appendix F). The univariate analysis results in Table 4 shows that classical difficulty indices across the four scoring procedures differed significantly, $F (1.315, 44.71) = 72.08, p < 0.001$, with an effect size partial-eta squared ($\eta^2$) = 0.68. Simple contrasts were employed to compare difficulty estimates obtained from scoring procedures with a reference measure. The reference measure is the original scoring procedure because much of the interest was to compare the original or default response scoring procedure with the other three scoring procedures.

The simple contrasts shown in Table 5 indicate that there were significant differences in difficulty estimates between the original scoring procedure ($M = 0.452, SD = 0.155$) and the Omitted ($M = 0.528, SD = 0.193$), $F (1, 34) = 57.49, p < 0.001$, $\eta^2 = 0.63$, also with the Zero procedure ($M = 0.426, SD = 0.157$), $F (1, 34) = 69.90, p < 0.001$, $\eta^2 = 0.67$. There was no difference in difficulty estimates between the original scoring procedure and the Not-presented procedure ($M = 0.456, SD = 0.149$), $F (1, 34) = 1.24, p = 0.273$, $\eta^2 = 0.03$. 
The null hypothesis concerning CTT item discrimination is: classical item discrimination indices are the same across the four scoring procedures. Repeated measures analysis of variance was used. Mauchly’s W test for discrimination estimates indicated that the assumption of sphericity has been violated ($\chi^2(5) = 44.97, p < 0.001$), therefore degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity ($\varepsilon = 0.553$) (see Table 28 in Appendix F). The univariate analysis results in Table 6 showed that classical discrimination indices across the four scoring procedures differed significantly, $F(1.66, 56.41) = 25.24, p < 0.001$, with an effect size ($\eta^2$) = 0.43. Simple contrasts were employed to compare discrimination estimates obtained from scoring procedures with the original responses. The simple contrasts in Table 7 indicated that there were significant differences in discrimination estimates between the original scoring procedure and the Omitted, $F(1, 34) = 8.22, p < 0.01$, $\eta^2 = 0.195$, also with the Zero procedure, $F(1, 34) = 46.42, p < .001$, $\eta^2 = 0.58$. There was no difference in discrimination estimates between the Original scoring procedure and the Not-presented procedure, $F(1, 34) = 0.38, p = 0.85$, $\eta^2 = 0.001$. 
Table 4

Univariate ANOVA of Within-Subjects for CTT Difficulty Indices

<table>
<thead>
<tr>
<th>Source</th>
<th>Contrast</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
<th>η²</th>
</tr>
</thead>
<tbody>
<tr>
<td>Difficulty</td>
<td>Greenhouse-Geisser</td>
<td>.199</td>
<td>1.315</td>
<td>.151</td>
<td>72.081</td>
<td>.000</td>
<td>.679</td>
</tr>
<tr>
<td>Error</td>
<td>Greenhouse-Geisser</td>
<td>.094</td>
<td>44.709</td>
<td>.002</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5

Tests of Within-Subjects Contrasts for CTT Difficulty Indices

<table>
<thead>
<tr>
<th>Source</th>
<th>Contrast</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
<th>η²</th>
</tr>
</thead>
<tbody>
<tr>
<td>Difficulty</td>
<td>O vs. D</td>
<td>.201</td>
<td>1</td>
<td>.201</td>
<td>57.492</td>
<td>.000</td>
<td>.628</td>
</tr>
<tr>
<td></td>
<td>NP vs. D</td>
<td>.000</td>
<td>1</td>
<td>.000</td>
<td>1.242</td>
<td>.273</td>
<td>.035</td>
</tr>
<tr>
<td></td>
<td>Z vs. D</td>
<td>.023</td>
<td>1</td>
<td>.023</td>
<td>69.898</td>
<td>.000</td>
<td>.673</td>
</tr>
<tr>
<td>Error</td>
<td>O vs. D</td>
<td>.119</td>
<td>34</td>
<td>.003</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>NP vs. D</td>
<td>.013</td>
<td>34</td>
<td>.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Z vs. D</td>
<td>.011</td>
<td>34</td>
<td>.000</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. O: Omitted, D: Original, NP: Not-presented, Z: Zero
Table 6

*Univariate ANOVA of Within-Subjects for CTT Discrimination (a) Indices*

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
<th>η²</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>.100</td>
<td>1.659</td>
<td>.061</td>
<td>25.238</td>
<td>.000</td>
<td>.426</td>
</tr>
<tr>
<td>Error</td>
<td>.135</td>
<td>56.412</td>
<td>.002</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 7

*Tests of Within-subjects Contrasts for CTT Discrimination (a) Indices*

<table>
<thead>
<tr>
<th>Source</th>
<th>Contrast</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
<th>η²</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>O vs. D</td>
<td>.034</td>
<td>1</td>
<td>.034</td>
<td>8.222</td>
<td>.007</td>
<td>.195</td>
</tr>
<tr>
<td></td>
<td>NP vs. D</td>
<td>.00002</td>
<td>1</td>
<td>.00002</td>
<td>.038</td>
<td>.846</td>
<td>.001</td>
</tr>
<tr>
<td></td>
<td>Z vs. D</td>
<td>.067</td>
<td>1</td>
<td>.067</td>
<td>46.422</td>
<td>.000</td>
<td>.577</td>
</tr>
<tr>
<td>Error</td>
<td>O vs. D</td>
<td>.142</td>
<td>34</td>
<td>.004</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>NP vs. D</td>
<td>.024</td>
<td>34</td>
<td>.001</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Z vs. D</td>
<td>.049</td>
<td>34</td>
<td>.001</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note.* O: Omitted, D: Original, NP: Not-presented, Z: Zero

To test for equivalence of internal consistency coefficients, the reliability, based on internal consistency, coefficients were computed across scoring procedures. For the Original responses $KR20 = 0.78$, Omitted $KR20 = 0.74$, Zero $KR20 = 0.83$, and Not-
presented $KR20 = 0.78$. The ALPHATST program (Lautenschlager, 1989), which uses a modified chi-square statistic, was used to test for equivalence of internal consistency coefficients. Statistically significant differences among the four reliability coefficients were found, $\chi^2(3) = 216.83$, $p < 0.001$.

*Analyses for Research Question Three*

Do the different scoring procedures yield significantly different IRT item parameter estimates? Separate repeated measures ANOVA analyses were done comparing the differences for each of the four IRT parameter estimates $a$, $b$, $c$, and ability respectively across different scoring procedures.

Table 8

*Descriptive Statistics for IRT Item Parameter Estimates*

<table>
<thead>
<tr>
<th>IRT parameter</th>
<th>Scoring procedure</th>
<th>Min</th>
<th>Max</th>
<th>$M$</th>
<th>$SD$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Difficulty</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Original response</td>
<td>-.97</td>
<td>3.84</td>
<td>.97</td>
<td>1.18</td>
</tr>
<tr>
<td></td>
<td>Omitted</td>
<td>-.91</td>
<td>8.21</td>
<td>1.12</td>
<td>1.60</td>
</tr>
<tr>
<td></td>
<td>Not-presented</td>
<td>-.93</td>
<td>3.38</td>
<td>.88</td>
<td>.99</td>
</tr>
<tr>
<td></td>
<td>Zero</td>
<td>-.92</td>
<td>3.42</td>
<td>.91</td>
<td>1.00</td>
</tr>
<tr>
<td>Discrimination</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Original response</td>
<td>.58</td>
<td>2.08</td>
<td>1.17</td>
<td>.36</td>
</tr>
<tr>
<td></td>
<td>Omitted</td>
<td>.64</td>
<td>1.95</td>
<td>1.15</td>
<td>.33</td>
</tr>
<tr>
<td></td>
<td>Not-presented</td>
<td>.63</td>
<td>1.82</td>
<td>1.20</td>
<td>.33</td>
</tr>
<tr>
<td></td>
<td>Zero</td>
<td>.63</td>
<td>1.81</td>
<td>1.20</td>
<td>.33</td>
</tr>
</tbody>
</table>

*(table continues)*
Table 18 (continued)

<table>
<thead>
<tr>
<th>Guessing</th>
<th>Original response</th>
<th>.08</th>
<th>.29</th>
<th>.18</th>
<th>.04</th>
</tr>
</thead>
<tbody>
<tr>
<td>(c) Omitted</td>
<td></td>
<td>.13</td>
<td>.27</td>
<td>.18</td>
<td>.03</td>
</tr>
<tr>
<td>Not-presented</td>
<td></td>
<td>.04</td>
<td>.31</td>
<td>.14</td>
<td>.06</td>
</tr>
<tr>
<td>Zero</td>
<td></td>
<td>.04</td>
<td>.31</td>
<td>.14</td>
<td>.06</td>
</tr>
</tbody>
</table>

The means of the IRT $b$ parameter estimates varied according to the scoring procedure from 0.88 for the Not-presented to 1.12 for the Omitted procedure. With regard to $a$ parameter estimates, the means were close to each other and ranged from 1.15 for the Omitted to 1.20 for both the Not-presented and the Zero procedures. With regard to pseudo-guessing parameter, the means for $c$ estimates were similar to each other across all of the different scoring procedures.

The null hypothesis concerning IRT item difficulty estimates: item difficulty parameter estimates based on 3PL-IRT are the same across the four scoring procedures. Repeated Measures Analyses of variance was used to determine the significance of the differences for IRT parameter estimates across scoring procedures. Mauchly’s W test indicated that the assumption of sphericity had been violated, therefore degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity ($\varepsilon = 0.38$) (see Table 29, Appendix F). The univariate analysis of variance results in Table 9 show that the IRT $b$ parameter estimates across the four scoring procedures were not significantly different from each other, $F(1.14, 38.76) = 2.36, p = 0.13$, with an effect size ($\eta^2$) = 0.06.
Table 9

Tests of Within-subjects Effects for IRT Difficulty (b) Parameter Estimates

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
<th>η²</th>
</tr>
</thead>
<tbody>
<tr>
<td>b</td>
<td>Greenhouse-Geisser</td>
<td>1.182</td>
<td>1.140</td>
<td>1.037</td>
<td>2.356</td>
<td>.130</td>
</tr>
<tr>
<td>Error</td>
<td>Greenhouse-Geisser</td>
<td>17.061</td>
<td>38.761</td>
<td>.440</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The null hypothesis concerning IRT item discrimination estimates: item discrimination parameter estimates based on 3PL-IRT are the same across the four scoring procedures. Repeated measures analysis of variance was used to determine the significance of the differences. Mauchly’s W test indicated that the assumption of sphericity has been violated (see Table 30, Appendix F) therefore degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity (ε = 0.44) (see Table 30, Appendix F). The univariate analysis results in Table 10 show that IRT a parameter estimates across the four scoring procedures were not significantly different from each other, $F(1.32, 44.94) = 2.66, p = 0.10$, with an effect size ($\eta^2$) = 0.07.
Table 10

Tests of Within-subjects Effects for IRT Discrimination (a) Parameter Estimates

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
<th>η²</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>Greenhouse-Geisser</td>
<td>.064</td>
<td>1.322</td>
<td>.048</td>
<td>2.662</td>
<td>.100</td>
</tr>
<tr>
<td>Error</td>
<td>Greenhouse-Geisser</td>
<td>.817</td>
<td>44.942</td>
<td>.018</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The null hypothesis concerning IRT pseudo-guessing parameter estimates: item guessing parameter estimates based on 3PL-IRT are the same across the four scoring procedures. Repeated Measures Analyses of variance was used to determine the significance of the differences. Mauchly’s W test indicated that the assumption of sphericity has been violated, therefore degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity (ε = 0.45) (see Table 31, Appendix F). The univariate analysis results in Table 11 show that c parameter estimates across the four scoring procedures were significantly different, $F(1.35, 46.02) = 23.74$, $p < 0.001$, with an effect size ($η^2$) = 0.41.

The simple contrasts in Table 12 indicate that there were non-significant differences in c parameter estimates between the original scoring procedure and the Omitted procedure, $F(1, 34) = 1.67$, $p = 0.205$, $η^2 = 0.047$. By contrast, significant differences were noticed for both the Not-presented, $F(1, 34) = 26.38$, $p < 0.001$, $η^2 = 0.437$ and the Zero scoring procedure, $F(1, 34) = 26.26$, $p < 0.001$, $η^2 = 0.436$. 
Table 11

Tests of Within-subjects Effects for IRT Pseudo-guessing (c) Parameter Estimates

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
<th>η²</th>
</tr>
</thead>
<tbody>
<tr>
<td>c</td>
<td>Greenhouse-Geisser</td>
<td>.045</td>
<td>1.354</td>
<td>.033</td>
<td>23.740</td>
<td>.000</td>
</tr>
<tr>
<td>Error</td>
<td>Greenhouse-Geisser</td>
<td>.065</td>
<td>46.022</td>
<td>.001</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 12

Tests of Within-subjects Contrasts for IRT Pseudo-guessing (c) Parameter Estimates

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
<th>η²</th>
</tr>
</thead>
<tbody>
<tr>
<td>c</td>
<td>Omitted vs. Original</td>
<td>.001</td>
<td>1</td>
<td>.001</td>
<td>1.671</td>
<td>.205</td>
</tr>
<tr>
<td></td>
<td>Not-presented vs. Original</td>
<td>.038</td>
<td>1</td>
<td>.038</td>
<td>26.376</td>
<td>.000</td>
</tr>
<tr>
<td></td>
<td>Zero vs. Original</td>
<td>.038</td>
<td>1</td>
<td>.038</td>
<td>26.256</td>
<td>.000</td>
</tr>
<tr>
<td>Error</td>
<td>Omitted vs. Original</td>
<td>.022</td>
<td>34</td>
<td>.001</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Not-presented vs. Original</td>
<td>.049</td>
<td>34</td>
<td>.001</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Zero vs. Original</td>
<td>.049</td>
<td>34</td>
<td>.001</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Analyses for Research Question Four

Do the different scoring procedures yield significantly different IRT person parameter estimates? Descriptive statistics were computed for the IRT person parameter estimates (θ) as shown in Table 13. Separate repeated measures ANOVA analyses were
done to test for the significance in person parameter estimates across different scoring procedures.

Table 13

*Descriptive Statistics for IRT Person Parameter Estimates*

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Min</th>
<th>Max</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>586</td>
<td>-1.948</td>
<td>2.557</td>
<td>.004</td>
<td>.905</td>
</tr>
<tr>
<td>omitted</td>
<td>586</td>
<td>-1.972</td>
<td>2.488</td>
<td>-.003</td>
<td>.897</td>
</tr>
<tr>
<td>Not-presented</td>
<td>586</td>
<td>-1.720</td>
<td>2.455</td>
<td>.191</td>
<td>.804</td>
</tr>
<tr>
<td>Zero</td>
<td>586</td>
<td>-2.102</td>
<td>2.354</td>
<td>-.011</td>
<td>.922</td>
</tr>
</tbody>
</table>

The means of the $\theta$ estimates shown in Table 13 were similar across the Original, the Zero, and the Omitted scoring procedures. On the other hand, the mean of $\theta$ estimates calibrated using the Not-presented procedure is the largest. Repeated measures analysis of variance was used to determine the significance of the differences for IRT $\theta$ estimates across the scoring procedures. Mauchly’s W test indicated that the assumption of sphericity had been violated, therefore, degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity ($\varepsilon = 0.574$) (see Table 32, Appendix F).

The univariate analysis results in Table 14 show that the $\theta$ parameter estimates across the four scoring procedures were significantly different, $F (1.72, 1007.783) = 183.94, p < 0.001$, with an effect size ($\eta^2$) = 0.24. The simple contrasts in Table 15 indicated that there were non-significant differences in $\theta$ parameter estimates between the
original scoring procedure and both the Omitted procedure, $F(1, 585) = 1.21, p = 0.27, \eta^2 = 0.002$ and the Zero procedure, $F(1, 585) = 2.55, p = 0.11, \eta^2 = 0.004$. On the other hand, a significant difference were shown in $\theta$ parameter estimates for the Not-presented, $F(1, 34) = 294.20, p < 0.001, \eta^2 = 0.335$.

Table 14

*Tests of Within-subjects Effects for IRT Person ($\theta$) Parameter Estimates*

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
<th>$\eta^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta$ Greenhouse-Geisser</td>
<td>16.667</td>
<td>1.723</td>
<td>9.675</td>
<td>183.937</td>
<td>.000</td>
<td>.239</td>
</tr>
<tr>
<td>Error Greenhouse-Geisser</td>
<td>53.010</td>
<td>1007.783</td>
<td>.053</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 15

*Tests of Within-subjects Contrasts for IRT Person ($\theta$) Parameter Estimates*

<table>
<thead>
<tr>
<th>Source</th>
<th>Contrast</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
<th>$\eta^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta$ Omitted vs. Original</td>
<td>.027</td>
<td>1</td>
<td>.027</td>
<td>1.214</td>
<td>.271</td>
<td>.002</td>
<td></td>
</tr>
<tr>
<td>Not-presented vs. Original</td>
<td>20.565</td>
<td>1</td>
<td>20.565</td>
<td>294.199</td>
<td>.000</td>
<td>.335</td>
<td></td>
</tr>
<tr>
<td>Zero vs. Original</td>
<td>.121</td>
<td>1</td>
<td>.121</td>
<td>2.553</td>
<td>.111</td>
<td>.004</td>
<td></td>
</tr>
<tr>
<td>Error Omitted vs. Original</td>
<td>13.113</td>
<td>585</td>
<td>.022</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not-presented vs. Original</td>
<td>40.893</td>
<td>585</td>
<td>.070</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zero vs. Original</td>
<td>27.625</td>
<td>585</td>
<td>.047</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure 6. The relationship between scores based on Original responses and the Not-presented scoring procedure.

Distribution of true scores calibrated from the Original responses and the Not-presented procedure revealed that lower ability examinees (i.e., those with scores lower than 17.5) tend to get higher scores when considering their responses that generated under solution-oriented behavior.
Analyses for Research Question Five

Do pass/fail decisions for each examinee differ across the different scoring procedures? Percentages of students who passed the test were computed based on four arbitrary cut-scores as shown in Table 16.

Table 16

<table>
<thead>
<tr>
<th>Scoring procedure/ Cut-score</th>
<th>50%</th>
<th>60%</th>
<th>70%</th>
<th>80%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original response</td>
<td>37.7</td>
<td>23.4</td>
<td>8.2</td>
<td>2.0</td>
</tr>
<tr>
<td>Omitted</td>
<td>36.5</td>
<td>22.4</td>
<td>7.7</td>
<td>1.9</td>
</tr>
<tr>
<td>Not-presented</td>
<td>44.4</td>
<td>26.6</td>
<td>9.9</td>
<td>2.7</td>
</tr>
<tr>
<td>Zero</td>
<td>34.1</td>
<td>21.8</td>
<td>7.7</td>
<td>1.9</td>
</tr>
</tbody>
</table>

The proportion of students who pass the test, according to four cut-scores, were similar across the Original, the Omitted, and the Zero scoring procedures. However, the proportions of students who passed the test were dissimilar for the Not-presented procedure when 50% correct was used as a cut-score. At least 7% of the students were classified differently based on the Not-presented procedure compared with the original responses. To test if the proportion difference of 7% is significantly different from zero, a 95% CI for the difference between paired proportions was constructed using Confidence Interval Analysis software (Bryant, 2000). The result showed that the CI limits does not contain the null value of zero with 95% CI ranged from 0.046 to 0.087. The resulting
interval suggests that in the population at least 5% to 9% of students may be classified differently if we utilize the original responses rather than using the Not-presented procedure with cut-score 50% of the items.

Table 17

<table>
<thead>
<tr>
<th>Cut-score</th>
<th>50% L95% *</th>
<th>50% U95% **</th>
<th>60% L95%</th>
<th>60% U95%</th>
<th>70% L95%</th>
<th>70% U95%</th>
<th>80% L95%</th>
<th>80% U95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original response</td>
<td>0.338</td>
<td>0.416</td>
<td>0.200</td>
<td>0.268</td>
<td>0.060</td>
<td>0.104</td>
<td>0.009</td>
<td>0.031</td>
</tr>
<tr>
<td>Omitted</td>
<td>0.326</td>
<td>0.404</td>
<td>0.190</td>
<td>0.258</td>
<td>0.055</td>
<td>0.099</td>
<td>0.008</td>
<td>0.030</td>
</tr>
<tr>
<td>Not-presented</td>
<td>0.404</td>
<td>0.484</td>
<td>0.230</td>
<td>0.302</td>
<td>0.075</td>
<td>0.123</td>
<td>0.014</td>
<td>0.040</td>
</tr>
<tr>
<td>Zero</td>
<td>0.303</td>
<td>0.379</td>
<td>0.185</td>
<td>0.251</td>
<td>0.055</td>
<td>0.099</td>
<td>0.008</td>
<td>0.030</td>
</tr>
</tbody>
</table>

*: L95% is the lower limit for the CI.

**: U95% is the upper limit of the CI.

The calculated CIs for the percentages of students who succeeded in the test appeared to be the same and they were not significantly different from each other at the confidence level 95%. However, when constructing the CIs based on a lesser confidence level value, for example, 68% the Not-presented procedure appeared to be statistically different from other scoring procedures (see table 33, Appendix J).
CHAPTER FIVE

Discussion

Students differ in the amount of effort they put forth during testing. Some students may not give tests sufficient effort if they recognize that the test has no impact on their grades in school or if scores will not be reported back to the students themselves. Consequently, performance will be affected and will not represent their true abilities (Wise & DeMars, 2003, 2005). This research investigated the effects of utilizing response latency to detect rapid-guessing behavior on CTT and IRT parameter estimates. Any response that was generated under short response latency less than a particular threshold for that item was rescoring as either Omitted, Not-presented, Zero, or maintained the Original response. The study employed data obtained from a sample of 586 ninth-grade students from five high schools in Jordan who took a computer-administration of a mathematics test. It was assumed that some students would not put forth reasonable effort toward the test items when they perceived that the test scores had a minor personal consequence associated with their scores.

The current study employed a relatively different method of establishing response thresholds. The median time needed to read items was recorded based on results from a group of 20 students, assuming that one threshold was not appropriate for all of the items. There were some indications that the calibrated thresholds were suitable since they were established based on empirical data and the accuracy of responses generated under rapid-guessing behavior was near the chance level.
Summary of Results

The analysis of the current study started with describing rapid-guessing responses in the sample and constructing confidence intervals for the proportion of examinees who exhibited rapid-guessing behavior for each item. The results revealed that rapid guessing may have occurred fairly early in the test. Further, it was found that examinees showed rapid-guessing behavior on every test item with different proportions of examinees; on the other hand, the proportion of solution-oriented responses declined in the final third of the test. This result was evident from the line graph (Figure 2) that represented solution-oriented behavior across items. This finding suggests a low level of effort put forth on the test items by some students. The proportion of students who exhibited at least one rapid-guessed response was significantly different from zero. This finding implies that more attention should be given to consider these responses if decisions will be made based on student performance on such tests or if the purpose of the test administration is to establish norms.

The analysis of students’ responses showed that 22% of the rapid-guessed responses were correct. This finding was similar to earlier findings that under rapid-guessing behavior the probability of a correct response did not significantly differ from what was expected and remained near the level of chance (Wise, 2006; Wise & Kong, 2005).

The first research question investigated the proportion of students who exhibited rapid-guessing behavior. Confidence intervals for proportions were constructed to determine whether the proportion of examinees who committed rapid-guessing behavior
was different from zero. The result showed that 12% of the total responses were generated under rapid-guessing behavior. This proportion was more than the 6% reported by Wise and DeMars (2006) and the 95% CI indicated that the proportion of students who committed rapid-guessing behavior was significantly different from zero.

With regard to research question 2 (are there significant differences in classical item indices and reliability across the different scoring procedures?), the three scoring procedures were contrasted with the Original scoring procedure. A classical item analysis indicated that both classical item difficulty and discrimination indices were significantly different for the Omitted and the Zero procedures contrasted with the default scoring procedure. This difference indicated the amount of rapid guessing on an item influenced the item’s mean and correlation with scores of another item or the entire test. Conversely, there was no significant difference in difficulty and discrimination estimates between the Original scoring procedure and the Not-presented procedure.

The results for CTT difficulty estimates were consistent with Wise (2006) in that rescoring rapid-guessing responses had little effect on item difficulty. In contrast, the findings of discrimination estimates were inconsistent with Wise’s results where he had found an impact on item-total correlations, and estimates decreased after rescoring. DeMars (2000) asserted that researchers and test developers wanted item difficulties estimated from pilot tests or low-stakes tests to reflect the relative difficulties that the items would have under final administration.

Different values of reliability coefficients were observed across scoring procedures except for the Not-presented procedure. The internal consistency coefficient
for the Not-presented procedure was similar to the Original scoring procedure. This may be attributed to reduction of item variance and the number of valid cases. That is, removing noise from data that may add a systematic influence on the scores may be responsible for the differences in reliability. In conclusion, the same level of reliability in the original responses was retained after removing the noise from the data using the Not-presented procedure. Increased reliability was indicated by the Omitted or Zero procedures. Wolf and Smith (1995) noted that changing consequences for the test did not affect the reliability of the scores; although it does affect the mean level of the scores. Wise and Sundre (2003) found similar reliability coefficients for different degrees of motivation filtering. Conversely, findings from Wise, Wise, and Bhola (2006) showed a simultaneous decrease in the internal consistency coefficients and the corresponding standard error of measurement. The variation in internal consistency coefficients was explained by the decrease in the observed score standard deviations. Paradoxically, the standard error of measurement decreased in the same direction as the reliability.

To respond to research question 3 (do the different scoring procedures yield significantly different IRT item parameter estimates?), repeated measure analyses of variance showed that the IRT difficulty and discrimination parameter estimates were not significantly different across scoring procedures. For the pseudo guessing parameter, there was no significant difference contrasting the Original scoring with the Omitted procedure. When calibrating parameters, Bilog 3.02 (Mislevy & Bock, 1990) replaced the guessed responses, as they were keyed, with the inverse number of options per item. This procedure matches with the probability of accuracy under rapid-guessing behavior. On
the other hand, lower means of guessing parameter were shown for both the Zero and the Not-presented procedures contrasted with the default procedure.

The results for the IRT parameter estimates were consistent with the findings of Schnipke (1996) that items appear more difficult than they really are before removing rapid-guessing responses from low-effort examinees. Gaviria (2005) reported lower difficulty estimates after considering response latency. Results for discrimination and pseudo guessing parameter estimates were consistent with Gaviria (2005). The estimates for a model that considers response latency are always higher for the IRT discrimination item parameter with the exception of the pseudo guessing parameter.

To respond to research question 4 (do the different scoring procedures yield significantly different IRT person parameter estimates?), the distribution of person parameter estimates calibrated from the Original responses and the Not-presented procedure revealed that lower ability examinees tend to get higher estimates on the Not-presented procedure compared with the Original responses. This result is consistent with findings from Wise, 2006 and Sundre and Wise (2003) where the person estimates are influenced by both true proficiency and effort. It is expected that we will have higher ability estimates after considering the rapid-guessing responses.

Responding to research question 5 (do pass/fail decisions for each examinee differ across the different scoring procedures?), the findings showed different proportions of students who passed the cut-scores across scoring procedures. Some of the examinees who previously failed using the default scoring procedure passed after rescoring the less thoughtful responses following the Not-presented procedure. The analysis based on
confidence intervals showed at least 5% to 9% of students in the population may have been classified differently. As a result, misclassification may occur if we utilize the original responses rather than rescoring them as Not-presented. It appears that identifying individual examinee rapid-guessing responses and rescoring them may ultimately influence the scores and, therefore, the decision taken upon performance might be changed accordingly. A possible implication of this result may be beneficial for norming and equating studies especially when conducted under low-stakes settings. It is not surprising that norms sometimes appear lower than expected because under low-stakes settings the scores underestimate the students’ ability level as they fail to capture the full effort of the examinees (Wolf, Smith, & Birnbaum, 1995).

Previous research has shown that response latency can be useful in measuring both examinee effort in the test (Wise & Kong, 2005) and the effort received by items (Wise, 2006). Additionally, Wise and DeMars (2006) used an effort moderated IRT model based on response latency to consider examinee behavior and were able to better estimate both person and item parameter estimates.

**Conclusions**

The analysis revealed that not all items were affected equally by rapid guessing behavior; some items were less affected than other items. In addition, findings support the notion that rapid-guessing responses influence test performance and that this influence may be substantially different for different scoring procedures.

The results of this research suggest that response latency may profitably be considered and integrated into the scoring process, because response latency
differentiates between the more thoughtful responses and the raid-guessed responses. The implications for educational measurement researchers are to examine response latency and to take into account student scores and item parameter changes after considering rapid-guessing responses. Rapid guessing behavior may also be an issue in high-stakes assessment since raising the stakes does not always contribute to a corresponding increase in effort and achievement (Baumert & Demmrich, 2001; O’Neil, Abedi, Miyoshi, & Mastergeorge, 2005).

Further research to explore the properties of rapid-guessing items in terms of cognitive load, content, number of options, item response format, and other characteristics that may influence the appearance of this behavior is recommended. In addition, further research is needed to explore the rapid-guessing behavior with regard to gender, age, and minority groups.

Limitations

The major limitations of the current research are as follows:

1. This research examined data from high school ninth-grade students from Jordan. The results, therefore, should not be generalized without considering the sample characteristics.

2. The examinees’ responses were collected based upon a portion of multiple-choice released items from TIMMS 2003. Therefore, the results may not be generalized to different types of items in achievement tests and in other content areas.
3. Time thresholds for items were calibrated from the administration of the test on a
group of 20 students; therefore, the accuracy of the obtained thresholds may affect
the identification of rapid-guessed items.

4. The classification of students to pass/fail was based on an arbitrary percentage of
items; the validity of decisions depends on the reasonableness of the cut-scores.
Different classifications such as basic, proficient, and advanced levels may reveal
different decisions. The study is not aimed at generalizing to basic, proficient, and
advanced levels.

Recommendations and Suggestions for Future Research

In light of the findings and limitations of this research, the following
recommendations may be helpful in conducting similar studies in the future:

1. Since the proportion of rapid-guessing responses was noticed in a significant
amount, future studies might consider filtering responses to get an accurate
estimation of students’ knowledge. Furthermore, other content areas (e.g.,
language, science...), other item formats, manipulation of the item order in the
test, and examinees from different grades could be utilized.

2. Future studies might investigate parameter accuracy and bias using different
scoring procedures and examine their consequences on test validity. More
research using simulated or real data is needed to investigate the psychometric
properties of tests that consider response latency. For example, identifying rapid-
guessing responses may impact item fit and test information.
3. This research used the 3PL-IRT model; future research might focus on other models of IRT, such as the one- and two-parameter logistic models of IRT. Doing so would extend the existing body of research and provide a more general view on the reasonableness of rescoring.

4. Wise and DeMars (2006) recommended professionals require a minimum percentage of solution-oriented responses for an examinee to obtain a valid score. Test administrators and test givers might set the minimum at 75%, which implies that any examinee with an RTE score of less than that percentage would not receive a valid score. Lower percentages of RTE may yield insufficient information for such responses to be trustworthy. Therefore, professionals might consider the utility of being able to measure response latency using computer-based testing. Response latency could be used to monitor the examinees’ effort during the test and measurement professionals can examine different levels of RTE scores to set the minimum acceptable level of effort to obtain legitimate scores. In addition, test administrators may choose short tests to reduce the chances of committing rapid-guessing behavior.


Bhola, D. S. (1994). *An investigation to determine whether an algorithm based on response latencies and number of words can be used in a prescribed manner to reduce measurement error.* (Doctoral dissertation, University of Nebraska - Lincoln). Abstract retrieved March 19, 2007 from ProQuest database.


APPENDICES
APPENDIX A Mathematics Test (Arabic Translation)

A portion of released items from TIMMS 2003

1. The items in the balance are balanced completely. On the left side of the balance, there is a jar of water. What is the weight of the water?

- 0.5 kg
- 1 kg
- 2 kg
- 3 kg

2. The box is 9 cm long. What is the approximate length in centimeters? Any answer can be an approximation.

- 10 cm
- 9.9 cm
- 9.6 cm
- 8.6 cm
3

يدور علي حول الملعب 4 مرات في الوقت نفسه الذي يدور فيه كمال 3 دورات.
كم دورة يمكنها علي حول الملعب عندما ينفي كمال 12 دورة؟

9 1
11 4
12 3
16 5

4

حصلت خديجة على الدرجات 78، 76، 74 في ثلاثة امتحانات. حصلت مريم على الدرجات 74، 72، 82 في الامتحانات نفسها. قارن متوسط الدرجات مع متوسط الدرجات مريم.

1 متوسط الدرجات أعلى بعلامة واحدة.
2 متوسط الدرجات أدنى بعلامة واحدة.
3 المتوسطان متساويان.
4 متوسط الدرجات أعلى بعلامات.
5 متوسط الدرجات أدنى بعلامات.
ببين الجدول علاماتطلبة أحد الصفوف في اختبار نهاية الفصل 10 علامات.

<table>
<thead>
<tr>
<th>الانتظام</th>
<th>التكرار</th>
</tr>
</thead>
<tbody>
<tr>
<td>٠</td>
<td>//</td>
</tr>
<tr>
<td>١</td>
<td>///</td>
</tr>
<tr>
<td>٢</td>
<td>/</td>
</tr>
<tr>
<td>٣</td>
<td>//</td>
</tr>
<tr>
<td>٤</td>
<td>///</td>
</tr>
<tr>
<td>٥</td>
<td>/</td>
</tr>
<tr>
<td>٦</td>
<td>//</td>
</tr>
<tr>
<td>٧</td>
<td>///</td>
</tr>
<tr>
<td>٨</td>
<td>/</td>
</tr>
<tr>
<td>٩</td>
<td>//</td>
</tr>
<tr>
<td>١٠</td>
<td>///</td>
</tr>
</tbody>
</table>

كم عدد الطلبة الذين حصلوا على علامة أعلى من ٩٧؟

٢ ٢٠٨،١٧٦٢
في الشكل أ ب  ج ه مستقيمان متتقاطعان.

ما قيمة س + ص؟

16 4
12 5
8 6
18 8
32 9

إذا كان 

فإن قيمة ن تساوي:

2 4
7 6
26 8
23 9
إذا كانت ن عدداً صحيحًا سالباً، فأي ما يلي هو العدد الأكبر؟

1. \(3 + n\)
2. \(2 \times n\)
3. \(2 - n\)
4. \(2 ÷ n\)

يمثل الشكل الآتي توزيع المحاصيل في إحدى الدول.

طبقاً للمعلومات الواردة في الشكل، أي العبارات الآتية صحيحة؟

1. محصول الشعير أكثر من محصول القمح.
2. محصول القمح أكثر من نصف المحاصيل بالدولة.
3. محصول الشعير أكثر من ثلث المحاصيل بالدولة.
4. مجموع محصول الشعير ومحصول القمح أكثر من محصول القمح.
س ص ع ل (غير مبين) شبه منحرف آخر يطابقه (له الشكل والمساحة نفسها).

إذا علم أن الزاوية س = الزاوية ل = 70°، فما مما يأتي عبارة صحيحة؟

1. س ص = أب.
2. الزاوية ص زاوية قائمة.
3. أطول أضلاع الشكل س ص ع ل متساوية.
4. محيط الشكل س ص ع ل = امثال محيط الشكل أب ج د.
5. مساحة المنطقة س ص ع ل أصغر من مساحة المنطقة أب ج د.

في أي من أزواج الأعداد الأتية يكون 2 أكبر من العدد الأول وأصغر من العدد الثاني؟

1. 7، 1
2. 8، 2
3. 11، 8
4. 11، 7
نظمت أعداد التماثل لتكوين الأشكال الآتية:

الشكل ۱
الشكل ۲
الشكل ۳

إذا استمر تكوين الأشكال على المنوال نفسه، فكم عددًا من التماثل يلزم لتكوين الشكل العاشر؟
۱ ۴۷ ۸ ۳۶ ۸ ۷۲ ۴۱ ۳۷ ۴۶

ملك أحمد ملكي ما يملكه سعيد من الكتب ويمتلك خليل ۶ كتب زيادة عما يملكه سعيد، إذا كانت س تمثل عدد الكتب التي يملكها سعيد، أي مما يظل يمثل مجموع أعداد الكتب التي يملكها الأرائد الثلاثة؟
۱ ۳۲ + ۶ ۸ ۳ ۸ ۶ ۲ ۶ ۸ ۲
أسماء المواد لم تظهر على الرسم البياني. أقالام البحر عادة أكثر المواد مبيعًا، والملاحبات أقلها مبيعًا وأقلام الرصاص البيضاء أكثر من المساطر. فكم قلم رصاص بيع؟

1. 40
2. 80
3. 120
4. 140

أي من الأعداد التالية مرتبة من الأكبر إلى الأصغر؟

1. 332, 322, 232
2. 323, 332, 222
3. 332, 322, 232
4. 323, 332, 222
يوضح الشكل الآتي العلاقة بين المسافة والأولم من سير كل من جمال وخالد سيراً على الأقدام.

إذا كان كل منهما قد بدأ الحركة من المكان نفسه، وسار في الاتجاه نفسه، عند أي وقت يلتقيان معاً؟

8:00 ①
8:20 ②
9:00 ③
10:00 ④
11:00 ⑤
في الشكل التالي، المثلثات أن - ج ، د هما متساويتان ، ب - ج = د - ه.

ما قياس الزاوية د - ج؟

1. 30°
2. 40°
3. 50°
4. 60°
5. 70°
6. 80°
7. 90°
8. 100°
9. 110°

المطلوب ترتيب الأرقام الأربعة الموضحة من الأكبر إلى الأصغر لتكون عدداً يتالف من أربع منازل. ثم إعادة ترتيب الأرقام نفسها من الأصغر إلى الأكبر لتكون عدداً آخر من أربع منازل أيضاً.

ما الفرق بين العددين المتكونين؟

2736 ①
4726 ②
8732 ③
8182 ④
8192 ⑤
أي مما ينطوي على مساحة المحصول على العدد الثاني من العدد الأول في كل من الأزواج المرتبة المذكورة؟

1. إضافة 2
2. طرح 3
3. الضرب في 2
4. الضرب في 2 ثم إضافة 2
5. الضرب في 2 ثم طرح 2

سوف ينقص الوقت الذي تحتاجه حافلة لقطع المسافة بين مدينتين من 25 دقيقة إلى 20 دقيقة بعد أن يكتمل إنشاء الطريق السريع. ما النسبة المئوية للنقصان في الوقت لقطع المسافة بين المدينتين؟

1/4
1/5
1/20
1/25
عدد طلاب الصف الثامن في مدرسة ما 30 طالباً، واحتمال اختيار طالب بصورة عشوائية عمره أقل من 12 سنة يساوي $\frac{1}{3}$. ما عدد الطلبة في الصف الثامن الذين تقل أعمارهم عن 12 سنة؟

اثنان

ثلاثة

أربعة

خمسة

ستة

يمكن تدوير المستطيل على عل لينطبق على المستطيل المربك.

ما نقطة مركز الدوران؟

ا

ب

ج

د

ه
عدد الزجاجات التي سعة كل منها 250 ملليترا والتي يمكن ملؤها بـ 400 لتراً من الاماء يساوي؟

16 1
16.7 2
160 3
1600 4
16... 5
بركة مستطيلة الشكل محاطة بممر للمشاه كما في الشكل.

ما مساحة ممر المشاه؟

1. 100 م
2. 166 م
3. 162 م
4. 170 م
5. 161 م

يمر خط مستقيم بالنقطتين (2,2) و (7,7). أي من النقاط التالية تقع على الخط؟

1. (2,0)
2. (2,1)
3. (4,2)
4. (5,0)
5. (0,4)
بين الشكل أدناه قرصاً دائرياً مقسم إلى 24 قطعةً. وعندما يحرك شخص ما السهم، فإن فرصة توقفه على أي قطاع متساوية.

\[ \frac{1}{24} \text{ القطعات زرقاء}, \frac{1}{24} \text{ أرجوانية}, \frac{1}{24} \text{ برتقالي}, \frac{1}{24} \text{ حمراء} \]

ما لون القطاع الذي يمكن أن يكون السهم عند نقل أقل ما يمكن؟

1. أزرق
2. أرجواني
3. برتقالي
4. أحمر

سعة خزان الوقود لسيارة ٤٥ لترًا من الوقود تستهلك السيارة ٨ لترًا لكل مسافة ١٠٠ كم تقريبًا. بدأ السيارة رحلة مساحتها ٣٥٠ كم. وكان الخزان مليئًا بالوقود.

كم بقي من الوقود في نهاية الرحلة؟

1. ١٥.٢٥ لترًا
2. ١٦.٢٥ لترًا
3. ٢٤.٧٥ لترًا
4. ٢٧.٧٥ لترًا
أي الأشكال التالية يمكن تثبيتها لتكون شكلًا ذات أبعاد كالشكل المرسوم في الأعلى؟
في المستوى الإحداثي المرسوم أعلاه، أي نقطة يمكن أن يكون إحداثياتها (2، 4)؟

1. 
2. 
3. 
4. 

إذا كانت ص = 5 و كانت $rac{x}{2} = 3$، ما قيمة ص؟

1. 
2. 
3. 
4. 
5. 
6.
إذا كان \( \frac{1}{72} = 0.70 \), فإن \( \frac{1}{24} = \) 

أي من الوحدات التالية تستخدم عادة لقياس مساحة ملعب كرة القدم؟

1. سنتيمترات مربعة
2. سنتيمترات مكعبة
3. أمतار مربعة
4. أمتار مكعبة
أي مما يلي هو الأقرب لـ $\sqrt{11 + 9}$؟

- ① $2 + 2$
- ② $8 + 3$
- ③ $2 + 12$
- ④ $8 + 12$

ثلاثة أخوة بدر، زيد وعمر. تلقوا هدية من والدهم مقدارها 4500 ديناراً. وضع المبلغ بين الأخوة بنسبة عدد الأطفال لكل منهم. بدر لديه طفلان، زيد لديه 3 أطفال، عمر لديه 4 أطفال.

كم ديناراً كان نصيب عمر؟

- ① 1500$
- ② 1000$
- ③ 1500$
- ④ 2000$
في الشكل أعلاه، المثلثات الصغيرة لها نفس المساحة. ما نسبة مساحة المنطقة المظللة إلى المساحة غير المظللة؟

1. 6:5
2. 5:6
3. 8:5
4. 5:3
APPENDIX B Mathematics Test (in English)

A portion of released items from TIMMS, 2003

1. The objects on the scale make it balance exactly. On the left pan there is a 1 kg weight (mass) and half a brick. On the right pan there is one brick.

What is the weight (mass) of one brick?

A. 0.5 kg
B. 1 kg
C. 2 kg
D. 3 kg

2. The length of a box is 9 cm to the nearest centimeter. Which of these could be the actual length of the box?

A. 10 cm
B. 9.9 cm
C. 9.6 cm
D. 8.6 cm
3. Alice can run 4 laps around a track in the same time that Carol can run 3 laps. When Carol has run 12 laps, how many laps has Alice run?

A. 9
B. 11
C. 13
D. 16

4. Joe had three test scores of 78, 76, and 74, while Mary had scores of 72, 82, and 74. How did Joe’s average (mean) score compare with Mary’s average (mean) score?

A. Joe’s was 1 point higher.
B. Joe’s was 1 point lower.
C. Both averages were the same.
D. Joe’s was 2 points higher.
E. Joe’s was 2 points lower.
The table shows scores for a class on a 10-point test.

<table>
<thead>
<tr>
<th>Test Score</th>
<th>Tally</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>/</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>///</td>
<td>3</td>
</tr>
<tr>
<td>6</td>
<td>///// /</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>//</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>////</td>
<td>4</td>
</tr>
<tr>
<td>9</td>
<td>///</td>
<td>3</td>
</tr>
<tr>
<td>10</td>
<td>/</td>
<td>1</td>
</tr>
</tbody>
</table>

How many in the class made a score greater than 7?

A) 2
B) 8
C) 10
D) 12
E) 20
In the figure, $PQ$ and $RS$ are intersecting straight lines.

What is the value of $x + y$?

- A 15
- B 30
- C 60
- D 180
- E 300

If $\frac{12}{n} = \frac{36}{21}$, then $n$ equals

- A 3
- B 7
- C 36
- D 63
If \( n \) is a negative integer, which of these is the largest number?

A. \( 3 + n \)
B. \( 3 \times n \)
C. \( 3 - n \)
D. \( 3 \div n \)

9

The graph shows the distribution of crops grown in a certain country.

According to the information in the graph, which of these statements is true?

A. More oats are grown than wheat.
B. Corn is more than one-half of the country’s crop.
C. Oats are more than one-third of the country’s crop.
D. The total crop of oats and wheat is greater than the corn crop.
$ABCD$ is a trapezoid.

Another trapezoid, $GHIJ$ (not shown), is congruent (the same size and shape) to $ABCD$. Angles $G$ and $J$ each measure $70^\circ$. Which of these could be true?

(A) $GH = AB$
(B) Angle $H$ is a right angle.
(C) All sides of $GHIJ$ are the same length.
(D) The perimeter of $GHIJ$ is 3 times the perimeter of $ABCD$.
(E) The area of $GHIJ$ is less than the area of $ABCD$. 
In which of these pairs of numbers is 2.25 larger than the first number but smaller than the second number?

A  1 and 2
B  2 and \( \frac{5}{2} \)
C  \( \frac{5}{2} \) and \( \frac{11}{4} \)
D  \( \frac{11}{4} \) and 3

Matchsticks are arranged as shown in the figures.

If the pattern is continued, how many matchsticks would be used to make Figure 10?

A  30
B  33
C  36
D  39
E  42
Graham has twice as many books as Bob. Chan has six more books than Bob. If Bob has $x$ books, which of the following represents the total number of books the three boys have?

A. $3x + 6$
B. $3x + 8$
C. $4x + 6$
D. $5x + 6$
E. $8x + 2$

The graph shows the number of pens, pencils, rulers, and erasers sold by a store in one week.

The names of the items are missing from the graph. Pens were the item most often sold, and fewer erasers than any other item were sold. More pencils than rulers were sold. How many pencils were sold?

A. 40
B. 80
C. 120
D. 140
In which list are the numbers ordered from greatest to least?

A. 0.233, 0.3, 0.32, 0.332
B. 0.3, 0.32, 0.332, 0.233
C. 0.32, 0.233, 0.332, 0.3
D. 0.332, 0.32, 0.3, 0.233

The graph represents the distance and time of a hike taken by Joshua and Liam.

If they both started from the same place and walked in the same direction, at what time did they meet?

A. 8:00
B. 8:30
C. 9:00
D. 10:00
E. 11:00
In this figure, triangles $ABC$ and $DEF$ are congruent with $BC = EF$.

What is the measure of angle $EGC$?

A 20°  
B 40°  
C 60°  
D 80°  
E 100°
18

The four digits above are to be arranged from largest to smallest to form a four-digit number. The same four digits are then to be arranged from smallest to largest to form another four-digit number. What is the difference between the two resulting four-digit numbers?

- **A** 3726
- **B** 4726
- **C** 8082
- **D** 8182
- **E** 8192

19

(3, 6), (6, 15), (8, 21)

Which of these describes how to get the second number from the first number in every ordered pair above?

- **A** Add 3
- **B** Subtract 3
- **C** Multiply by 2
- **D** Multiply by 2 and then add 3
- **E** Multiply by 3 and then subtract 3
20

When a new highway is built, the average time it takes a bus to travel from one town to another is reduced from 25 minutes to 20 minutes. What is the percent decrease in time taken to travel between the two towns?

A  4%
B  5%
C  20%
D  25%

21

In an eighth-grade class of 30 students, the probability that a student chosen at random will be less than 13 years old is \(\frac{1}{5}\). How many students in the class are less than 13 years old?

A  Two
B  Three
C  Four
D  Five
E  Six
Rectangle $PQRS$ can be rotated (turned) onto rectangle $UVST$.

What point is the center of rotation?

A) $P$
B) $R$
C) $S$
D) $T$
E) $V$

The number of 250 milliliter bottles that can be filled from 400 liters of water is

A) 16
B) 160
C) 1600
D) 16000
A rectangular shaped swimming pool has a paved walkway around it as shown.

What is the area of the paved walkway?

A) 100 m²  
B) 161 m²  
C) 710 m²  
D) 1610 m²

A straight line passes through the points (2,3) and (4,7). Which of these points is also on the line?

A) (0,2)  
B) (1,2)  
C) (2,4)  
D) (3,5)  
E) (4,5)
The figure below shows a spinner with 24 sectors. When someone spins the arrow, it is equally likely to stop on any sector.

\[ \frac{1}{8} \text{ of the sectors are blue, } \frac{1}{24} \text{ are purple, } \frac{1}{2} \text{ are orange, and } \frac{1}{3} \text{ are red.} \]

If a person spins the arrow, on which color sector is the spinner LEAST likely to stop?

A. blue  
B. purple  
C. orange  
D. red

---

A car has a fuel tank that holds 45 L of fuel. The car consumes 8.5 L of fuel for each 100 km driven. A trip of 350 km was started with a full tank of fuel. How much remained in the tank at the end of the trip?

A. 15.25 L  
B. 16.25 L  
C. 24.75 L  
D. 29.75 L
Which of these could be folded to make a shape like the 3-D figure above?

A

B

C

D

In the coordinate plane above, which point could have coordinates (2, -4)?

A  P
B  Q
C  R
D  S
30. If $x - y = 5$ and $\frac{x}{2} = 3$, what is the value of $y$?

- A 6
- B 1
- C -1
- D -7

31. If $\frac{a}{b} = 70$, then $\frac{a}{2b} =$

- A 35
- B 68
- C 72
- D 140
Which of these units would usually be used for an area the size of a soccer field?

A) square centimeters
B) cubic centimeters
C) square meters
D) cubic meters

Which of these is closest to $11^2 + 9^2$?

A) $20 + 20$
B) $20 + 80$
C) $120 + 20$
D) $120 + 80$

Three brothers, Bob, Dan, and Mark, receive a gift of 45 000 zeds from their father. The money is shared between the brothers in proportion to the number of children each one has. Bob has 2 children, Dan has 3 children, and Mark has 4 children.

How many zeds does Mark get?

A) 5000
B) 10 000
C) 15 000
D) 20 000
In the figure above, each of the smaller triangles has the same area. What is the ratio of the shaded area to the unshaded area?

A) 5:3  
B) 8:5  
C) 5:8  
D) 3:5
APPENDIX C Sample of Items from the Computer Administration

Item 1

1. What is the weight of the rod?
   a. 5.0 kg
   b. 15 kg
   c. 20 kg
   d. 30 kg
   e. 35 kg

Item 4

The average of the scores of 78 items in a test is 82.72. If the average of the scores of 3 items is 80, what is the average of the scores of the other two items?
   a. 80
   b. 82.72
   c. 84
   d. 86
   e. 88
Item 20

سوف ينقص الوقت الذي تحتاجه حافلة لقطع المسافة بين مدينتين من 20 دقيقة إلى 15 دقيقة بعد أن يكمل إنشاء الطريق السريع. ما النسبة المئوية للنقص في الوقت لقطع المسافة بين المدينتين؟

أ) 1/2
ب) 1/3
ج) 1/4
د) 1/5

Item 24

بيرة مستطيلة الشكل محاطة بعمق للماء كما في الشكل.

ما مساحة بيرك الماء؟

أ) 200 م²
ب) 120 م²
ج) 210 م²
د) 1610 م²
إذا كانت $x - 5 = 3$ وكانت $\frac{x}{2} = 6$، ما قيمة $x$؟

$\begin{array}{l}
1 \quad 0 \\
0 \quad 3 \\
2 \quad 0 \\
\end{array}$
APPENDIX D Medians of response latency thresholds

Table 18

*Meditches of Response Latency Thresholds*

<table>
<thead>
<tr>
<th>Item</th>
<th>Median of threshold</th>
<th>Item</th>
<th>Median of threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10</td>
<td>19</td>
<td>7</td>
</tr>
<tr>
<td>2</td>
<td>13</td>
<td>20</td>
<td>7</td>
</tr>
<tr>
<td>3</td>
<td>14</td>
<td>21</td>
<td>8</td>
</tr>
<tr>
<td>4</td>
<td>18</td>
<td>22</td>
<td>6</td>
</tr>
<tr>
<td>5</td>
<td>15</td>
<td>23</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>8</td>
<td>24</td>
<td>8</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>25</td>
<td>6</td>
</tr>
<tr>
<td>8</td>
<td>7</td>
<td>26</td>
<td>15</td>
</tr>
<tr>
<td>9</td>
<td>12</td>
<td>27</td>
<td>10</td>
</tr>
<tr>
<td>10</td>
<td>11</td>
<td>28</td>
<td>5</td>
</tr>
<tr>
<td>11</td>
<td>8</td>
<td>29</td>
<td>6</td>
</tr>
<tr>
<td>12</td>
<td>8</td>
<td>30</td>
<td>5</td>
</tr>
<tr>
<td>13</td>
<td>11</td>
<td>31</td>
<td>5</td>
</tr>
<tr>
<td>14</td>
<td>11</td>
<td>32</td>
<td>4</td>
</tr>
<tr>
<td>15</td>
<td>4</td>
<td>33</td>
<td>4</td>
</tr>
<tr>
<td>16</td>
<td>10</td>
<td>34</td>
<td>11</td>
</tr>
<tr>
<td>17</td>
<td>7</td>
<td>35</td>
<td>6</td>
</tr>
<tr>
<td>18</td>
<td>9</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
APPENDIX E Number of Rapid-guessing Responses across Examinees

Table 19

*Percentages of Examinees According to Number of Guesses*

<table>
<thead>
<tr>
<th>Number of guesses</th>
<th>Frequency</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>.00</td>
<td>205</td>
<td>35.0</td>
</tr>
<tr>
<td>1.00</td>
<td>110</td>
<td>18.8</td>
</tr>
<tr>
<td>2.00</td>
<td>48</td>
<td>8.2</td>
</tr>
<tr>
<td>3.00</td>
<td>34</td>
<td>5.8</td>
</tr>
<tr>
<td>4.00</td>
<td>24</td>
<td>4.1</td>
</tr>
<tr>
<td>5.00</td>
<td>17</td>
<td>2.9</td>
</tr>
<tr>
<td>6.00</td>
<td>19</td>
<td>3.2</td>
</tr>
<tr>
<td>7.00</td>
<td>14</td>
<td>2.4</td>
</tr>
<tr>
<td>8.00</td>
<td>9</td>
<td>1.5</td>
</tr>
<tr>
<td>9.00</td>
<td>11</td>
<td>1.9</td>
</tr>
<tr>
<td>10.00</td>
<td>15</td>
<td>2.6</td>
</tr>
<tr>
<td>11.00</td>
<td>8</td>
<td>1.4</td>
</tr>
<tr>
<td>12.00</td>
<td>5</td>
<td>.9</td>
</tr>
<tr>
<td>13.00</td>
<td>7</td>
<td>1.2</td>
</tr>
<tr>
<td>14.00</td>
<td>7</td>
<td>1.2</td>
</tr>
<tr>
<td>15.00</td>
<td>6</td>
<td>1.0</td>
</tr>
<tr>
<td>16.00</td>
<td>5</td>
<td>.9</td>
</tr>
</tbody>
</table>

*(table continues)*
Table 19 (continued)

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>17.00</td>
<td>7</td>
<td>1.2</td>
</tr>
<tr>
<td>18.00</td>
<td>8</td>
<td>1.4</td>
</tr>
<tr>
<td>19.00</td>
<td>2</td>
<td>0.3</td>
</tr>
<tr>
<td>20.00</td>
<td>7</td>
<td>1.2</td>
</tr>
<tr>
<td>21.00</td>
<td>5</td>
<td>0.9</td>
</tr>
<tr>
<td>22.00</td>
<td>4</td>
<td>0.7</td>
</tr>
<tr>
<td>23.00</td>
<td>1</td>
<td>0.2</td>
</tr>
<tr>
<td>24.00</td>
<td>1</td>
<td>0.2</td>
</tr>
<tr>
<td>25.00</td>
<td>1</td>
<td>0.2</td>
</tr>
<tr>
<td>27.00</td>
<td>4</td>
<td>0.7</td>
</tr>
<tr>
<td>32.00</td>
<td>2</td>
<td>0.3</td>
</tr>
<tr>
<td>Total</td>
<td>586</td>
<td>100.0</td>
</tr>
</tbody>
</table>
APPENDIX F Statistics for Repeated Measures ANOVA Assumptions

Table 20

*One-sample Kolmogorov-Smirnov Test for Classical Difficulty Index*

<table>
<thead>
<tr>
<th></th>
<th>Original responses</th>
<th>Omitted</th>
<th>Not-presented</th>
<th>Zero</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kolmogorov-Smirnov Z</td>
<td>.658</td>
<td>.742</td>
<td>.645</td>
<td>.744</td>
</tr>
<tr>
<td>Asymp. Sig. (2-tailed)</td>
<td>.779</td>
<td>.641</td>
<td>.800</td>
<td>.637</td>
</tr>
</tbody>
</table>

Table 21

*One-sample Kolmogorov-Smirnov Test for Classical Discrimination Index*

<table>
<thead>
<tr>
<th></th>
<th>Original responses</th>
<th>Omitted</th>
<th>Not-presented</th>
<th>Zero</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kolmogorov-Smirnov Z</td>
<td>1.063</td>
<td>.747</td>
<td>.827</td>
<td>.565</td>
</tr>
<tr>
<td>Asymp. Sig. (2-tailed)</td>
<td>.209</td>
<td>.632</td>
<td>.502</td>
<td>.907</td>
</tr>
</tbody>
</table>

Table 22

*One-sample Kolmogorov-Smirnov Test for IRT Discrimination Parameter*

<table>
<thead>
<tr>
<th></th>
<th>Original responses</th>
<th>Omitted</th>
<th>Not-presented</th>
<th>Zero</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kolmogorov-Smirnov Z</td>
<td>.933</td>
<td>.972</td>
<td>.496</td>
<td>.496</td>
</tr>
<tr>
<td>Asymp. Sig. (2-tailed)</td>
<td>.349</td>
<td>.301</td>
<td>.967</td>
<td>.966</td>
</tr>
</tbody>
</table>
Table 23

*One-sample Kolmogorov-Smirnov Test for IRT Difficulty Parameter*

<table>
<thead>
<tr>
<th></th>
<th>Original responses</th>
<th>Omitted</th>
<th>Not-presented</th>
<th>Zero</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kolmogorov-Smirnov Z</td>
<td>.787</td>
<td>.942</td>
<td>.746</td>
<td>.745</td>
</tr>
<tr>
<td>Asymp. Sig. (2-tailed)</td>
<td>.566</td>
<td>.338</td>
<td>.634</td>
<td>.636</td>
</tr>
</tbody>
</table>

Table 24

*One-sample Kolmogorov-Smirnov Test for IRT Pseudo-Guessing Parameter*

<table>
<thead>
<tr>
<th></th>
<th>Original responses</th>
<th>Omitted</th>
<th>Not-presented</th>
<th>Zero</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kolmogorov-Smirnov Z</td>
<td>.557</td>
<td>.803</td>
<td>.977</td>
<td>.974</td>
</tr>
<tr>
<td>Asymp. Sig. (2-tailed)</td>
<td>.916</td>
<td>.539</td>
<td>.295</td>
<td>.299</td>
</tr>
</tbody>
</table>

Table 25

*One-sample Kolmogorov-Smirnov Test for IRT Ability Parameter*

<table>
<thead>
<tr>
<th></th>
<th>Original responses</th>
<th>Omitted</th>
<th>Not-presented</th>
<th>Zero</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kolmogorov-Smirnov Z</td>
<td>1.559</td>
<td>1.707</td>
<td>1.042</td>
<td>1.578</td>
</tr>
<tr>
<td>Asymp. Sig. (2-tailed)</td>
<td>.015</td>
<td>.006</td>
<td>.228</td>
<td>.014</td>
</tr>
</tbody>
</table>
Table 26

*Skewness and Kurtosis for IRT Ability Parameter*

<table>
<thead>
<tr>
<th>Scoring procedure</th>
<th>Skewness</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original data</td>
<td>.148</td>
<td>-.608</td>
</tr>
<tr>
<td>Omitted</td>
<td>.182</td>
<td>-.667</td>
</tr>
<tr>
<td>Not resented</td>
<td>.077</td>
<td>-.457</td>
</tr>
<tr>
<td>Zero</td>
<td>-.008</td>
<td>-.703</td>
</tr>
</tbody>
</table>

Table 27

*Mauchly's Test of Sphericity for CTT Difficulty Indices*

<table>
<thead>
<tr>
<th>Within Subjects Effect</th>
<th>Mauchly's W</th>
<th>Approx. Chi-Square</th>
<th>df</th>
<th>Sig.</th>
<th>Epsilon Greenhouse-Geisser</th>
<th>Huynh-Feldt</th>
<th>Lower-bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>Difficulty</td>
<td>.102</td>
<td>74.596</td>
<td>5</td>
<td>.000</td>
<td>.438</td>
<td>.449</td>
<td>.333</td>
</tr>
</tbody>
</table>
Table 28

*Mauchly's Test of Sphericity for CTT Discrimination Indices*

<table>
<thead>
<tr>
<th>Within Subjects Effect</th>
<th>Mauchly's W</th>
<th>Approx. Chi-Square</th>
<th>df</th>
<th>Sig.</th>
<th>Epsilon</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discrimination</td>
<td>.253</td>
<td>44.972</td>
<td>5</td>
<td>.000</td>
<td>.553</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.578</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.333</td>
</tr>
</tbody>
</table>

Table 29

*Mauchly's Test of Sphericity for IRT Difficulty Parameter Estimates*

<table>
<thead>
<tr>
<th>Within Subjects Effect</th>
<th>Mauchly's W</th>
<th>Approx. Chi-Square</th>
<th>df</th>
<th>Sig.</th>
<th>Epsilon</th>
</tr>
</thead>
<tbody>
<tr>
<td>b</td>
<td>.000</td>
<td>345.674</td>
<td>5</td>
<td>.000</td>
<td>.380</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.384</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.333</td>
</tr>
</tbody>
</table>
### Table 30

*Mauchly's Test of Sphericity for IRT Discrimination Parameter Estimates*

<table>
<thead>
<tr>
<th>Effect</th>
<th>Mauchly's W</th>
<th>Approx. Chi-Square</th>
<th>df</th>
<th>Sig.</th>
<th>Epsilon</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>.002</td>
<td>209.710</td>
<td>5</td>
<td>.000</td>
<td>.441</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Epsilon</th>
<th>Greenhouse-Geisser</th>
<th>Huynh-Feldt</th>
<th>Lower-bound</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>.452</td>
<td>.333</td>
<td></td>
</tr>
</tbody>
</table>

### Table 31

*Mauchly's Test of Sphericity for IRT Pseudo-guessing Parameter Estimates*

<table>
<thead>
<tr>
<th>Effect</th>
<th>Mauchly's W</th>
<th>Approx. Chi-Square</th>
<th>df</th>
<th>Sig.</th>
<th>Epsilon</th>
</tr>
</thead>
<tbody>
<tr>
<td>c</td>
<td>.000</td>
<td>296.627</td>
<td>5</td>
<td>.000</td>
<td>.451</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Epsilon</th>
<th>Greenhouse-Geisser</th>
<th>Huynh-Feldt</th>
<th>Lower-bound</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>.463</td>
<td>.333</td>
<td></td>
</tr>
</tbody>
</table>
Table 32

*Mauchly's Test of Sphericity for IRT Person Parameter Estimates*

<table>
<thead>
<tr>
<th>Within subjects effect</th>
<th>Mauchly's W</th>
<th>Approx. Chi-Square</th>
<th>df</th>
<th>Sig.</th>
<th>Epsilon</th>
<th>Greenhouse-Geisser</th>
<th>Huynh-Feldt</th>
<th>Lower-Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>θ</td>
<td>.297</td>
<td>708.770</td>
<td>5</td>
<td>.000</td>
<td>.574</td>
<td>.576</td>
<td>.333</td>
<td></td>
</tr>
</tbody>
</table>
APPENDIX G 68% CI across Scoring Procedures

Table 33

68% CI for the Proportion of Succeeded Students across Scoring Procedures

<table>
<thead>
<tr>
<th>Cut of score</th>
<th>50%</th>
<th>60%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>L68%</td>
<td>U68%</td>
</tr>
<tr>
<td>Original response</td>
<td>0.344</td>
<td>0.409</td>
</tr>
<tr>
<td>Omitted</td>
<td>0.332</td>
<td>0.398</td>
</tr>
<tr>
<td>Not-presented</td>
<td>0.410</td>
<td>0.478</td>
</tr>
<tr>
<td>Zero</td>
<td>0.309</td>
<td>0.373</td>
</tr>
</tbody>
</table>

(Table continues)

Table 33 (continued)

<table>
<thead>
<tr>
<th>Cut of score</th>
<th>70%</th>
<th>80%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>L68%</td>
<td>U68%</td>
</tr>
<tr>
<td>Original response</td>
<td>0.063</td>
<td>0.101</td>
</tr>
<tr>
<td>Omitted</td>
<td>0.059</td>
<td>0.095</td>
</tr>
<tr>
<td>Not-presented</td>
<td>0.079</td>
<td>0.119</td>
</tr>
<tr>
<td>Zero</td>
<td>0.059</td>
<td>0.095</td>
</tr>
</tbody>
</table>