A COMPARISON OF IRT AND RASCH PROCEDURES IN A MIXED-ITEM FORMAT TEST

Tari L. Kinsey, B.A., M. Ed.

Dissertation Prepared for the Degree of

DOCTOR OF PHILOSOPHY

UNIVERSITY OF NORTH TEXAS

August 2003

APPROVED:

Randall E. Schumacker, Major Professor
Gordon S. Gates, Minor Professor
Robin K. Henson, Committee Member
J. Kyle Roberts, Program Coordinator
Jon I. Young, Department Chair
M. Jean Keller, Dean of the College of Education
C. Neal Tate, Dean of the Robert B. Toulouse School of Graduate Studies

Kinsey, Tari L., A Comparison of IRT and Rasch Procedures in a Mixed Item Format Test. Doctor of Philosophy (Educational Research), August 2003, 110 pp., 29 tables, 19 figures, 6 appendices, 46 titles.

This study investigated the effects of test length (10, 20 and 30 items), scoring schema (proportion of dichotomous ad polytomous scoring) and item analysis model (IRT and Rasch) on the ability estimates, test information levels and optimization criteria of mixed item format tests. Polytomous item responses to 30 items for 1000 examinees were simulated using the generalized partial-credit model and SAS software. Portions of the data were re-coded dichotomously over 11 structured proportions to create 33 sets of test responses including mixed item format tests. MULTILOG software was used to calculate the examinee ability estimates, standard errors, item and test information, reliability and fit indices. A comparison of IRT and Rasch item analysis procedures was made using SPSS software across ability estimates and standard errors of ability estimates using a 3 x 11 x 2 fixed factorial ANOVA. Effect sizes and power were reported for each procedure. Scheffe post hoc procedures were conducted on significant factors. Test information was analyzed and compared across the range of ability levels for all 66-design combinations. The results indicated that both test length and the proportion of items scored polytomously had a significant impact on the amount of test information produced by mixed item format tests. Generally, tests with 100% of the items scored polytomously produced the highest overall information. This seemed to be especially true for examinees with lower ability estimates. Optimality comparisons were made between IRT and Rasch procedures based on standard error rates for the ability estimates, marginal reliabilities and fit indices (-2LL). The only significant differences reported

involved the standard error rates for both the IRT and Rasch procedures. This result must be viewed in light of the fact that the effect size reported was negligible. Optimality was found to be highest when longer tests and higher proportions of polytomous scoring were applied. Some indications were given that IRT procedures may produce slightly improved results in gathering available test information. Overall, significant differences were not found between the IRT and Rasch procedures when analyzing the mixed item format tests. Further research should be conducted in the areas of test difficulty, examinee test scores, and automated partial-credit scoring along with a comparison to other traditional psychometric measures and how they address challenges related to the mixed item format tests.

ACKNOWLEDGMENTS

I'd like to thank Dr. Randall E. Schumacker for his scholarly contributions to this work and for not allowing me to drown before I learned to swim. Special thanks also to Dr. Larry Daniel, Dr. Gordon S. Gates and Bryant DeBord for regularly providing me with candles and matches each time I was lost in the dark.

I will continue to thank my family the rest of my life for their encouragement, patience and many sacrifices that allowed me to complete this journey. Thank you to my mother, Sharon Smith, for reading and editing countless drafts of my early writing. Also, I would especially like to thank Donald and Erin for being a great father-daughter team and more self-sufficient than I ever could have imagined during this process. Your understanding and support helped make achieving this goal possible. I love you both very much.

Above all I am grateful to God for His many blessings. This experience has been humbling beyond words and has made me intimately aware of and comfortable with the fact that the more I continue to learn, the more I realize how little I really know.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Overview

The practice of testing has become increasingly common throughout the twentieth century and a reliance on information gained from test scores has made an indelible mark on our culture. All levels of education, kindergarten through graduate school, most professional license procedures and many employment avenues place a high reliance on test performance to disseminate opportunities, and promote and assure professional standards. Entire industries have evolved to support the proper design, production, administration, analysis, reporting and interpretation of tests and this development has placed an increasing demand on test developers, such as Educational Testing Services (ETS), for establishing quality tests in the field of psychometrics. Additionally, professional organizations such as the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME) have all made contributions to the quality of testing practices. Obviously, test quality is a high priority for those who make tests, those who take tests, and those who rely on test scores for decision making. Now more than ever, it is critical that tests are efficient and effective at measuring ability and that scores are reliable and precise measures of examinee ability. Criteria used to establish test quality generally focus on the areas of test design, test analysis techniques and test score interpretation.

Quality test design is impacted by many elements including format,

length, administration procedures, construction, validity and scoring schema. The most critical element being the items selected to compose the test. The two most common item types are "selected response" and "constructed response." Selected-response items require examinees to recognize and select one response from a group of provided responses (i.e. multiple-choice, matching and true-false items). These items generally have a "dichotomous" scoring schema that allows for two possible outcomes for each item response, "correct" or "incorrect." In contrast, constructed-response items gather information about the incomplete knowledge an examinee may possess, by requiring that the examinee generate a response to an item prompt (e.g. open-response, short- answer and essay items). Constructed-response items award "partial credit" on a hierarchical scale of correctness using a scoring schema known as "polytomous" scoring.

Historically, many large-scale national assessments, such as the Graduate Records Exam (GRE), the Scholastic Achievement Test (SAT) and the General Management Achievement Test (GMAT) have been constructed using only selected-response item formats and have applied a dichotomous scoring schema. Due to budget and time constraints, the sheer volume involved in processing these tests has made automated scoring a preferable method. The popularity of this approach is in spite of the fact that select-response formats have been criticized for the opportunity they provide examinees to "guess" the correct answer. For instance, with four choices provided, an examinee has a 25% chance of guessing the right answer.

Recently however, performance assessments, which usually require

polytomous scoring, have grown in popularity because they are thought to produce a more authentic ability measurement. This new trend in assessment has spurred an increasing practice of combining multiple item types within one testing format. This combination of item types is referred to in the literature as a "mixed-item format test" and an example of this type of test is included in Appendix A.

Once a test has been administered and its responses collected, several techniques exist that can be used for analyzing the item responses. These techniques are generally referred to as "mathematical models" and are used to estimate examinee proficiency and specific item characteristics. These calculations are referred to as "estimates" because of the inherent measurement error associated with all tests of cognitive ability. The reduction of this error is a high priority in quality test design. Error reduction improves the measurement precision of the test resulting in more accurate proficiency and item characteristic estimates.

Various mathematical models exist for analyzing item responses with the goal of producing proficiency level estimates and item characteristics. Item Response Theory (IRT) and the Rasch Model are two techniques developed specifically for these purposes. Both models were originally applied to dichotomously scored responses, but expanded versions of both models also exist for handling polytomous data. The models also have distinct differences. First, the IRT approach applies a probability distribution in determining ability estimates, while the Rasch Model applies a logistic technique. Second, the Rasch Model holds certain item parameters constant while IRT allows them to vary. Finally,

IRT fits an equation to the data while the data must fit the Rasch model in order for psychometric properties to be present.

IRT and Rasch methods have often been compared on their ability to estimate examinee proficiencies and item parameters when analyzing dichotomous or polytomous item responses. However, they have not widely been compared on their ability to analyze mixed-item format tests. Various mathematical models are available in IRT and Rasch for analyzing item responses that possess different scoring rubrics, but not within a mixed-item format test.

Si (2002) applied various item analysis models to simulated dichotomous and polytomous scoring schema and documented that item response models vary greatly in their ability to recover original ability estimates. Test design and analysis are complex procedures and combining multiple item formats in the same test further complicates these tasks. Therefore, it is important that research be done toward optimizing available techniques for designing and analyzing mixed-format tests. The use of optimality theories in the design of mixed-item format tests has been recommended (Li, Lissitz, & Yang, 1999).

Statement of the Problem

Educators have often administered mixed-item format tests. Students benefit when given the opportunity to respond to a variety of item types and various item formats are known to differ in the type of information they gather. Recently, theory was proposed that mixed-item tests provide an opportunity to capitalize on the inherent advantages of various item formats in analysis (Carlson, 1996). A mixed-item test combines multiple item types and various item-scoring

formats. Combining two types of items on one assessment enhances both the reliability and validity of the assessment (Lau & Wang, 1998).

IRT and Rasch techniques are applied in large-scale assessment situations to aid test analysis and test refinement procedures. However, a challenge exists among the available techniques for analyzing response patterns from mixed-item format tests. Most available psychometric techniques are intended for use with examinee responses to similar-scored items. To date, psychometric analysis has been conducted almost exclusively on sets of test items that were of the same score type (all dichotomous or all polytomous.) In addition, Classical Test Theory (CTT) reliability and validity coefficients are based on all test items being of the same item format, e.g. multiple choice and do not readily apply to mixed-item format tests.

Both IRT and Rasch methods have assumptions that must be met by the data before the models may be applied and the results considered valid. While the data assumptions required by IRT and Rasch are similar, IRT methods are considered more robust than Rasch in the face of violated assumptions. The Rasch assumption of a common discrimination parameter across all items is considered strict and difficult to meet, which, in turn, may lead to less valid results when analyzing a mixed-item format test. A benefit of using the Rasch techniques is the allowance for a common logistic scaling of item difficulty and person ability. IRT and Rasch methods have often been compared on their estimates of item parameters and examinee proficiency levels, but not on their ability to analyze mixed-item format tests. Selecting the optimal method for analyzing mixed-item

format responses must also be based on applying the appropriate model for the type of data or "fitting" the model and model assumptions to the data.

Berger's 1998 study of optimal criteria related to mixed-item format tests noted that optimal selection of dichotomous test items for the individual occurs when ability approximates difficulty. The study also stressed that, "little is known about the optimal selection of items for the efficient estimation of [thetas] for a *population* of examinees" (p. 252). With regard to polytomous item selection, Berger noted that discrimination across categories is a complicated issue and that the optimal number of categories has never been determined. In addition to the item selection and format issues, test analysis methods pose a challenge. New psychometric methods must be developed and addressed in light of the mixed-item-format test. Traditional psychometric techniques do not easily apply to the mixed-item format test because none of them were specifically designed to handle the combination of multiple scoring schemas in the one test.

Rationale for the Study

Test design, analysis techniques and optimality issues are all critical factors impacting ability estimates, test information, and ultimately the reliability of the ability estimates produced by any test. It is important to explore and compare (a) test formats of mixed-item format tests, (b) mathematical models used to analyze mixed-item format tests, and (c) optimization techniques for the design and analysis of mixed-item format tests. In the interest of those who take tests, those who make tests, and those who use test scores for decision making, these areas need more attention in the arena of psychometric methods.

Test Formats

Dichotomous scoring has traditionally provided an opportunity to administer longer tests in the interest of increasing score reliability while benefiting from the efficiency of automated scoring options. Recently, polytomous scoring has been recognized as offering a more precise form of measuring the proficiency level possessed by individuals. With multiple response categories and category thresholds, polytomous scoring allows for a greater number of parameters to be estimated. Typically, a mixed-item format test contains a large proportion of dichotomously-scored items and a small proportion of polytomously-scored items. It is important to further investigate the testing format issues that best estimate examinee proficiencies (Berger, 1998.) With the growing popularity of mixed-item format tests it is valuable to conduct research in the area of combined scoring schema.

Mathematical Models

Statistical analysis of a mixed-item format test is complicated. Analysis has been performed through the use of several existing procedures, however, most involve the process of separating the groups of like-scored item types, analyzing them and then constructing a method for combining the results. Conclusions concerning the models depend heavily on the techniques applied in the analysis and the subjective manner of weighting and combining the results. A clear comparison of these techniques has not been conducted on a mixed-item format test. A method for analyzing a complete set of varied item types has not been confirmed as superior to date and item discrimination parameters and item

information are critical issues that must be evaluated.

Comparisons of the Test Information Function (TIF) generated by various item types are common in the existing research. However, comparing the information results after applying various scoring schema and item analysis approaches to a set of specific items is not common. It is important to note that the amount of available information is greatest for items that are scored on a continuous scale, although research regarding the optimal number of response categories for polytomous items has been inconclusive. Overall test-information contributions of polytomous and dichotomous items, relative to each other, is an issue that should be investigated (Carlson, 1996). It is also important to look at information levels across the range of proficiency.

Optimization Techniques

Combining two item types on one test may take advantage of the benefits offered by each, so long as the analysis techniques applied accommodate the combination. Most studies on optimizing test design and analysis procedures that have been conducted in the past focused on tests with all like items. Maximizing both the assumptions of the statistical model and the optimality criteria for that model are the two main requirements of optimal test design (Berger 1998.) If optimization methods were applied toward creating valid guidelines for item selection, tests could be produced that decrease the variance in both the item parameter and examinee proficiency estimates (Berger, 1998.) Optimality theories and mixed-item formats are both fairly new and therefore have not been widely combined..

Research Questions

The present study hypothesized that the proportion of dichotomous and polytomous items, test length, scoring schema and analysis method impact the optimization criteria in mixed-item format tests. Optimization criteria, in turn, have an appreciable effect across a mixed-item format on the following: reliability of ability estimates, item information, test information, proficiency estimates, and item parameters. The research questions investigated in this study were as follows:

1. How does the proportion of dichotomous and polytomous items across test lengths and analysis models impact the proficiency estimates?
2. How does the proportion of dichotomous and polytomous items across test lengths and analysis models impact the Test Information Function (TIF)?
3. How does the proportion of dichotomous and polytomous items across test lengths and analysis models impact the overall test optimality?
4. How do IRT and Rasch techniques compare in their ability to analyze mixed-item format tests?

The present study compares IRT and Rasch analysis techniques across various mixed-item format tests. Investigation and comparison of these approaches for analyzing item responses is conducted on simulated data composed of mixed-item format responses. Test length and scoring proportions

are set at fixed levels based on the literature review. SAS, MULTILOG and SPSS software programs were utilized in the analysis. Proficiency estimates and information levels were calculated and compared. A discussion is presented regarding the generalization of the results of the analysis to actual test data.

Delimitation

This research will analyze data resulting from two scoring formats: items that are scored dichotomously, or binary (0,1), and items that are scored polytomously with a four-category hierarchical rubric (1,2,3,4 with 1=less correct and 4= more correct). Test length was limited to a maximum of 30 items based on the typical time constraints of an assessment with 30 constructed-response items. Also, all mathematical models chosen for this study are limited by the requirements of the software applied. Consequently, guessing and speededness in examinee item responses could not be simulated and therefore beyond the scope of this study. The sample size was fixed at 1,000 examinees.

Definition of Terminology

1. Dichotomous—test items that are scored using a binary scoring schema as either correct (1) or incorrect (0).
2. Item Response Theory—a theory that is based on two ideals: an examinee's performance on a test item can be explained by a set of latent abilities or traits and the level of the traits possessed has a direct relationship to the probability of a correct response to an item designed to measure the trait (Hambleton, Swaminathan and Rogers, 1991).
3. Polytomous/polychotomous—Test items that are scored on a hierarchical scale of correctness, such as with a scoring rubric, where each answer can be judged as relatively more correct or less correct as determined by a set of established criteria and content. Responses are then assigned a scale score, for example a "3" on a scale ranging from 1 to 4.
4. Mixed-item format test— An assessment instrument or test that has a combination of two types of items that require different scoring methods, for instance, multiple-choice items (dichotomous) and items that require partial-credit scoring (polytomous).
5. Model fit—the extent to which the assumptions of the model are met by the item response data.
6. Optimal Test Design—Maximizing both the assumptions of the statistical model and the optimality criteria for that model are the two main requirements of optimal test design (Berger 1998.)

7. Optimality Criteria for the Mixed-Item Format Test—In a composite test, this is achieved when (a) the reliability of scores produced by each subtest is maximized (usually related to test length), (b) the resulting reliability of the scores from the composite test exceeds that of either subtest alone, and (c) the overall length of the composite test is practical to administer.
8. Rasch— The Rasch Model (1960) applies a logistic technique that converts item parameter estimates and person ability estimates into relative logit measurements so they may be compared on a common scale. Georg Rasch had intentions of developing a measurement system that was based on axioms (Hambleton, 1989). If data fit the model criteria, then it could be analyzed.
9. Scoring Rubric—A tool used for scoring items that has a clearly defined scale for partial-credit scoring with a predetermined hierarchical scale. An example would be a scale from 1 to 4 used to rate the quality of an essay, for which specific criteria for each score would be provided to the rater.
10. Weighting approach—When a test is composed of more than one item type, the test is usually divided into subtests by item type. The subtests are then assigned relative weights that are combined to calculate a total composite score. Relative subtest weights are generally based on their contribution to score reliability and affect that of the composite test.
11. Unidimensionality—an assumption that must be met by the data for use in Rash and IRT applications. One dominant construct must exist for any set of test items and item responses that represent a specific ability measured.

CHAPTER TWO

REVIEW OF RELATED LITERATURE

Overview

This chapter provides a comprehensive review of the literature relevant to the present study. Initially, a variety of existing test design procedures, data concerns and scoring formats are presented. Next, a discussion of the history and factors involved in the analysis methods are given. Finally, optimization criteria for mixed-item format tests are discussed.

Test Format and Design Procedures

Cognitive abilities, which are typically referred to as "latent" abilities or "traits," cannot be measured directly, therefore, tests are designed and methods of analysis are applied to help estimate the abilities possessed by examinees. The accuracy of these estimates depends on purposeful and careful test design, construction, administration, analysis and interpretation. Common testing practices involve administering written tests composed of various item types including multiple-choice, essay, matching, open-response, and short-answer. Also increasing in popularity are performance assessments that usually require individuals to execute specific tasks. These assessments are beginning to be considered by some to provide "authentic" measures of ability due to their format. This approach to assessment is fairly new and somewhat labor intensive to administer and therefore not yet common in practice.

The quality of an assessment instrument can be determined in many ways. The careful construction and design of the test instrument itself is critical to establishing score reliability and validity. Designing an assessment instrument is a multi-step procedure with four major

steps as defined by Wright and Stone (1979). They are as follows: *clearly define the variable to be measured by the test; construct items that measure this variable; demonstrate that the items are an accurate measure of the variable; and evaluate the response patterns for expected consistency*. This process is a circular one in that once the response patterns have been evaluated, the results can be used to revise and improve the items and the test design for future administrations. The steps are detailed in the following paragraphs.

The first step in designing an assessment instrument is to *clearly define the variable to be measured by the test*. Defining the variable is necessary in order to construct quality items that will reflect the desired trait when answered correctly. This helps eliminate ambiguity in the test design. The reliability and validity of resulting scores depend on the clear definition of the variable to be measured. It is important that the items are designed to accurately measure the target trait. This is called "unidimensionality," and is especially important when various item formats are combined on a single assessment. Dimensionality differences between polytomous and dichotomous item blocks should be addressed prior to any analyses (Carlson, 1996).

The second step, *item construction*, is then driven by the goal to write items that measure the presence of the defined variable. When test items are well-constructed and valid measures of the targeted trait, the test scores and examinee response patterns provide clear evidence regarding trait level estimates. Many factors impact the quality of a test and people designing tests or basing decisions on test scores need to have a clear understanding of these factors. It is critical that tests produce scores that are valid trait estimates and that these scores are carefully interpreted (Linn, 1990). A variety of item formats exist including selected-response, constructed-response, and performance assessment and different item formats are known to gather different types of information. Item writing involves careful planning and

analysis to ensure unambiguous, efficient, nontrivial questions (Allen & Yen, 1979, p.120). Item development procedures vary depending on the item format and scoring schema. For instance, developing multiple-choice items includes careful attention to wording and selection of each response choice. Composing writing prompts requires considerable time invested in the planning of a scoring rubric, training people to score the essays and the study and confirmation of rater reliability. All tests have a designated item format, and regardless of the item types selected for the format, each item must have a predetermined scoring schema. For instance, each time an examinee responds correctly to an item, a number of points may be awarded toward the total test score. Tests that include more than one item type may also require the design of a scoring schema that combines the item subtest scores into one composite test score. Realistically, cost and time are also important concerns when developing any assessment. However, Wright and Stone (1979) stated that the top priority in determining the accuracy of an item or group of items for measuring ability is dependent on how well the item and test difficulty match the ability of the examinees.

The third step in test design is to *demonstrate that the newly constructed items are an accurate measure of the target variable*. If the targeted variable has been well defined and the items carefully written, then adequate field-testing of the items must occur so that the responses can be analyzed. Once a set of items is approved, they are then combined to make a test. In order to confirm the validity of an assessment, investigations into content and construct validity are conducted. Correlational analyses with the scores from other assessments measuring the same target variable also helps support test score convergent validity.

The fourth step in test design involves *conducting a sound evaluation of the response patterns to identify and resolve any inconsistencies* and support the reliability of the test scores.

For instance, items on a test should be designed so that they are unidimensional (i.e. they measure one single construct or trait). Factor analysis is helpful in evaluating a set of test items for unidimensionality. In addition, individual items can be evaluated for the amount of information they collect. Item information is measurable and each item can be evaluated for its individual contribution to total test information. The maximization of item and total test information is important for constructing a parsimonious and optimally designed test.

Various models exist for conducting item analysis and they are based on specific assumptions about the data and response formats chosen for the items (i.e. selected response, constructed response or a combination of both). The mathematical model used in the analysis, along with the nature of the data collected, both play a major role in determining the extent of the analysis and the potential to optimize the test design. A circular technique of test design and analysis can be used to improve and optimize test performance. Optimization techniques will be discussed in depth later in this chapter.

Data Concerns, Item Types and Scoring Schema

Consideration should be given to gathering data at various information levels. The levels at which the items are both written (nominal, ordinal, interval, and continuous) and scored (dichotomously or polytomously) have an appreciable impact on the resulting amount of test information that can be derived from the responses. As data increases in variability from nominal to continuous, more information is available to contribute to the analyses. The purpose of statistical analysis is to increase the ability to draw relevant conclusions about the information conveyed by the data. Choosing the highest-powered analysis given the level of data allows the researcher to maximize the usefulness of the results.

Dichotomous Scoring Models

Development of multiple-choice items intended for dichotomous scoring is a popular testing approach. Dichotomous scoring of items has advantages and disadvantages. One advantage is that the response time required for test administration is relatively minimal per item so developers are able to plan for administration of longer tests. Longer test length increases the reliability and the precision of the ability estimates by reducing the amount of error in the measurement. In the 1930's, it became possible for objective scoring of multiple-choice responses to be automated and completed in a fraction of the time and cost required for hand scoring. This technological advancement increased the attractiveness of this testing and scoring method. Costs associated with obtaining this technology have continually decreased and now the approach is available to most classroom teachers.

Disadvantages to this testing approach do exist in spite of its growing popularity. The use of multiple-choice items is not considered an "authentic" assessment method by some because it confirms only the ability of the examinee to recall or recognize information as opposed to demonstrating knowledge and ability. Another disadvantage is that guessing by examinees is inevitable and can add a significant nuisance variable to the analysis and to ability estimates.

In contrast to multiple-choice items, authentic assessment techniques measure the examinee's ability, knowledge and skills without providing the response set. Because knowledge acquisition is a continuous process, tests composed entirely of items that have a provided response set and are scored dichotomously gather limited information about the real ability of the individual. Associating wrong answer choices with "no knowledge" and correct answer choices with "complete knowledge" lead to misrepresentation of a person's ability or

knowledge level. Dichotomous item-response data, with its binary scoring, also limits the analysis options available for response patterns. However, because of its desirable characteristics, such as low relative cost, minimal administration time, automated scoring benefits, and ease of analysis, multiple-choice testing is preferred to other assessment methods in many instances.

Dichotomous item scoring is based on the assumption that the person either knows all or nothing about each item's content. Therefore, when items written at varying degrees of difficulty are combined to build a test, the number of items answered correctly is used to define a person's score. The danger associated with all dichotomous scoring is that correctly responding to all of the items on a very easy test can mean something entirely different from responding correctly to 50% of the items on a very difficult test. In general, item response theory goes beyond the simple percent correct calculation and analyzes the responses and relative difficulty of each item. This information is then used to determine the approximate ability of each examinee.

Polytomous Scoring Models

Analyses of incorrect responses by Thissen (1976) and Bock (1972) found that analyzing the various wrong responses produced useful information and dichotomous scoring makes that information inaccessible by grouping all of the wrong responses into one category. Items with multiple response categories can be scored polytomously. Polytomous item scoring provides a scale for measurement of partial knowledge by each item. Wang and Lee (1998) suggest that incomplete knowledge acquisition is best measured by a hypothesized hierarchical structure provided by polytomous item responses. Published data has been inconsistent and inconclusive regarding polytomous IRT's superiority to dichotomous IRT. A direct relationship

exists between researchers' awareness about the process of knowledge acquisition and the need for assessment to more accurately reflect ability. For item-level information about a person's ability to reflect the continuous scale of knowledge acquisition, each item must measure ability on a continuous scale. These scales generally range between the absolutes of incorrect and correct through the use of a "scoring rubric" for example, ranging in points awarded from "1" to "4". Scoring rubrics define levels of competency or ability based on criteria met at various levels of demonstrated knowledge or skill. In this case, a score of "1" might represent limited evidence of writing ability while a score of "4" represents great ability in writing. Documenting partial knowledge confirms that accurate measures of knowledge or ability can no longer be restricted to the use of traditional dichotomous items. Hambleton (1989, p. 195) noted that there is a "current testing movement away from dichotomously scored multiple-choice items and toward new formats utilizing polychotomous scoring."

Published studies have resulted in contradictory conclusions about polytomous scoring. Muraki and Bock (1996) also found multiple-category scoring to be more informative than binary scoring. A test made entirely of constructed-response items will gather more information about examinee ability than a full multiple-choice test that is scored dichotomously. This increase in information increases the total variance in the response data. Polytomously-scored items have more parameters so a better fit to data is provided, but this improved fit usually accompanies a loss in estimation precision at higher ability levels (Carlson, 1996, Wang and Lee, 1998). In contrast, polytomous scoring of items has also been found to increase the precision of measurement at all levels of ability by Hambleton, Robin and Xing (2000). Muraki and Bock (1996) compared polytomous items to adaptive testing, mentioning that several binary-scored items at different levels of difficulty serve a similar purpose.

Carlson (1996) described polytomously-scored items as multiple ordered categories that hypothesize a hierarchical structure in category difficulties. An examinee may demonstrate partial knowledge on each item administered allowing more information to be collected than with dichotomous scoring. The scale of ability, which requires four dichotomous items, may be delivered through one polytomous item with four hierarchical categories. This process allows for the gathering of an equal amount of data through the use of fewer items. Tests composed entirely of polytomously-scored items have been found to collect between two and three times the amount of information as tests containing strictly dichotomous items (Donoghue, 1994).

Even so, disadvantages exist to administering tests composed entirely of partial-credit items. Included are the increased time and cost involved in each of the following: item construction effort, test administration, design of elaborate scoring rubrics, scoring of the assessments. Due to the more demanding nature of responding to polytomous items, reaching a point of diminishing returns with regard to information is more probable as these tests increase in length. Interest in streamlining the test design process while designing the optimal test (i.e. information-to-cost ratio) has grown, but Wang and Lee (1998) noted that research using empirical data may not be generalizable to actual test data.

Combining Item Types

The mixed-item test format has been mentioned more often in recent research, specifically that dichotomous and polytomous items complement each other when combined in one assessment resulting in improved score reliability, validity and reduced cost (Ericikan, Schwarz, Julian, Burket, Weber and Link, 1998.) The main advantage of utilizing a mixed-item format approach to test design lies in the opportunity to gather more consistent information at all levels of ability while optimizing test length. Many studies have documented the advantages

and disadvantages of both item types and combining them on one test may overcome some of the documented disadvantages. Wainer and Thissen (1993) studied the relative information contribution, test length and administration time of polytomous and dichotomous items. They deemed polytomous items as "inefficient" based on the time invested in item construction, item scoring, administration and cost while having less information per minute of administration than dichotomous items. Two points in their investigation may have contributed to this conclusion: they analyzed field-test response data and they violated the assumption of local independence allowing for multidimensionality problems. Overall, test information contributions for polytomous and dichotomous items relative to each other are an issue that should be further investigated (Carlson, 1996). Gathering data at the nominal level and scoring it dichotomously are both methods that reduce the amount of available information to estimate an examinee's ability. Test items that gather data more toward the natural continuous nature of knowledge and ability, and that are scored on a polytomous scale, offer more precise information about the ability of examinees.

The "all or nothing" scoring of multiple-choice items may misrepresent the true knowledge or ability of examinees. These nominal responses that are scored in a binary format as 100% correct or 100% incorrect ignore the theory of partial knowledge acquisition (Wang & Lee, 1998). Test-item data that can only be scored and analyzed at the dichotomous level limits the information that can be obtained through analysis. Partial knowledge or ability is acquired on the path to full knowledge. Therefore, knowledge acquisition must take place on a sliding continuous and cumulative scale as measured by polytomous scoring models. Polytomous items posses multiple ordered categories that hypothesize a hierarchical structure in category difficulties (Carlson, 1996).

Testing, in general, is designed to measure the amount of a trait possessed by a person. To demonstrate that items are an accurate measure of a trait, a sound evaluation of the response patterns is conducted to verify the desired patterns of consistency. Thus, the accuracy of an item or group of items in measuring ability is dependent on how well the item and test difficulties match the ability of the examinees (Carlson, 1996). In comparing dichotomous and polytomous scoring, it has been stated that dichotomous scoring is clearly a method for "measuring the right thing poorly" and polytomous scoring "measures the wrong thing well," (Wainer and Thissen, 2001, page 42.) The all or nothing approach of binary scoring allows for clear information (the right thing being measured), but is not a precise tool for measurement (hence the poor measurement.) It provides precision at extreme descriptions such as "cold" versus "hot", "on" versus "off", and "wet" versus "dry." On the other hand, polytomous scoring allows more variance across responses (which may stray from the target trait) and a more continuous scale of precision by which to measure (hence measuring well.)

Recently the attempt to measure acquired knowledge and ability through demonstration, performance or direct application has been growing. This form of assessment is typically achieved through performance assessment and is referred to as "authentic measurement". This type of unstructured or open-response assessment is commonly achieved through test items referred to as "constructed-response." For instance, a driving test is administered behind the wheel of a car because the most precise measurement of the ability to drive is best obtained through a demonstration, which is then graded by an instructor using a predetermined and well-defined scoring rubric. Generally, acquiring a driver's license also requires the examinee to obtain an acceptable score on a written test assessing knowledge about driving rules. These two separate tests administered on actual driving skill (performance assessment) and road

knowledge (multiple-choice items scored dichotomously) measure different types of ability and are not combined in a scoring schema or a mixed-item format test. Examinees must obtain a passing score on these tests separately. Psychometrically, a more accurate ability estimate for examinees would be produced if these two forms of assessment and item types were combined into one score.

Some examples of large-scale assessments that have a mixed-item format including both constructed-response and multiple-choice items are the National Assessment of Educational Progress (NAEP), Advanced Placement Exams (AP Exams), and the Test of English as a Foreign Language (TOEFL). The various items are grouped by type, scored with pre-selected schema, and combined by applying a weighting formula to derive one composite score per person. Mixed-item format tests can be challenging to analyze. A technique must be used that addresses multiple scoring schemas contained in one set of test items.

Summary

All assessment methods have documented advantages and disadvantages and each method is more appropriate at some times rather than others. Many factors must be considered when assessments are put in place with careful and purposeful design, analysis and interpretation playing important roles in evaluating the effectiveness of assessment. Budgets, timeframes, test quality and current trends are all critical concerns in the field of assessment.

Analysis Methods

Now that common scoring techniques have been explained in detail, a discussion of test and item analysis methods historically applied across scoring techniques is necessary. The resulting demands on the evolving field of psychometrics will also be discussed.

History

Many traits possessed by individuals have measurable data that is readily available to support conclusions. Brown eyes, marital status, place of residence and employment are all possible to verify through inspection and official documentation. However, cognitive abilities are indirectly measured and inferred, thus difficult to define and evidence is usually documented with an authentic assessment or test. Throughout history standards of testing have changed (Crocker & Algina, 1986). The AERA/APA/NCME *Standards,* published in 1985, addressed several criteria related to publishing tests, the need for the information, and the content that should be assessed. Criteria-referenced testing (based on knowledge of content and minimum passing standards) and norm-referenced testing (based on relative performance and percentile scores) are two very different approaches commonly used in testing. Industry standards exist for the production and use of both. Calculations of test scores and test designs have also become more statistically complex (Embretson & Reise, 2000). Recognition of deficiencies within existing psychometric methods has often been the catalyst inspiring improvement in item construction, test design, test administration, test score interpretation and usage (Embretson & Hershberger, 1999).

It is possible to roughly gauge ability simply based on the number of items each examinee answered correctly. Examinees could be arranged in order of their raw scores, and the expectation would be that those who obtained the highest raw score possessed the highest ability in the area assessed. Responding correctly to 9 out of 10 test items, therefore earning a score of 90%, would qualify an examinee as nearly mastering the content. But what if all examinees obtained a raw score of 10 out of 10 items correct. What would that mean about the test? What would it mean about the examinees ability? What if they all only obtained 2 items

correct out of the possible 10? What if each examinee was correct on a different pair of items? Does that information have implications about the individual items and their quality? Are the scores reliable? Do the scores contain measurement error? If so, how much? Is the same amount of error contained in each score? Many questions are often asked in regard to testing and require solutions to these practical testing problems (Lord, 1980).

Classical Test Theory (CTT), developed in the 1800's, was a move toward recognizing that observed raw scores might not perfectly reflect the true ability or knowledge of an examinee. This theory attempted to partition out the measurement error (E) and the true score (T) from a person's observed score (X). The measurement model that defines CTT is, $X = T + E$, where measurement error represents the sum total of any existing errors in the test design, administration, or examinee responses that impact the examinee's true score.

The deficiencies of CTT are related to the assumption that under this model all examinees have the same amount of error reflected in the observed raw scores. The measurement error is computed for the examinees as a group and not individually. This in turn makes the item characteristics dependent on how the group performed as a whole. The individual scores and ability estimates that result are then dependent on the ability of the group in which they tested and the difficulty of the test relative to that ability.

Latent Trait Theories (e.g. IRT and Rasch) began to develop in the late 1950's in response to researchers recognizing the shortcomings of CTT. Inherent ability or knowledge possessed by an individual does not depend on the ability of the group in which he is tested or on the difficulty of the test taken. These new models focused on removing the group and test dependency of proficiency and item characteristic estimates. This improves the precision of the measurements. Mental traits possessed by a person are difficult to determine and tests provided

supporting estimates and evidence of those traits. The purpose of item response theory, or IRT, is to model the relationship between a latent trait and the probability of persons who possess that trait responding correctly to items designed to measure it (Hambleton et. al., 1991).

Item Response Theory

The following sections focus attention on the various item analysis approaches as they apply to specific scoring methods. A detailed discussion of each IRT approach and the identified mathematical formula are included, explained and compared.

IRT Model for Dichotomous Scoring Format

There are several different analysis models categorized within the scope of item response theory. All of these models include a measure of person ability (theta) and attention to other measurement-related item parameters. The one, two and three-parameter logistic models (1PL, 2PL, 3PL) are commonly used to analyze dichotomous item responses within the IRT framework. These models are named after the number of parameters that have the freedom to vary in the analysis. These parameters include the discrimination index (a), the item difficulty parameter (b), and the guessing parameter (c).

There are several assumptions related to the IRT models that must be evaluated against the data prior to applying the procedures. The level to which these assumptions are met by the data can help determine which model is best fitting and to what degree of accuracy the results may be interpreted. The four general assumptions are related to the areas of dimensionality, speededness, guessing and discrimination. A discussion of each follows.

The first assumption that must be met prior to applying an IRT technique is that a data set must be unidimensional. This means that the items constructed to build the assessment must define one main construct or dimension. If there are many items that do not align with the main

construct it could imply that the assessment is in fact multidimensional and that more than one latent trait appears to be reflected in the assessment responses. For instance, if on a particular mathematics assessment the vocabulary used to construct the items was at too difficult of a reading level for the students to understand the items, this could cause the examinees to respond incorrectly to many items based on the latent trait of language ability rather than their math ability. Failure to pre-assess the responses for unidimensionality could cause inaccurate interpretation of the response data with regard to the math ability of the students.

The second assumption required of a data set for IRT analysis is the absence of "speededness" or a time limit for administration of the assessment. If an assessment is conducted under a limited administration time, one factor that may cause an examinee to incorrectly respond to items is their ability to respond to items quickly. This is a latent ability and could cause interference when interpreting the desired measurement results of the assessment. If two latent abilities, one of them being speed of response, are measured on an assessment, the unidimensionality assumption is violated.

The third assumption is related to the data containing responses that are the result of successful guessing that examinees may have demonstrated. It is desirable for this quality to be minimal in any set of response data, but various IRT models differ on how they address it. Some are set up to assume no successful guessing occurred. Others set the guessing constant across all items based on odds, such as .25 probability for multiple-choice items with four response choices. Another possibility is that some models allow the guessing parameter to vary reflecting the amount of successful guessing that has occurred for that item. This estimated calculation is determined through the use of ability estimates.

The fourth assumption involves the role played by discrimination across response categories. Various IRT models approach the topic of item discrimination from different perspectives. Determining if constant or variant discrimination might be more appropriate to assume for a data set is an important decision and impacts the model selection for each analysis. How various models handle this item parameter should be carefully evaluated when choosing the appropriate model for an analysis.

Fredric M. Lord is credited with introducing the two-parameter normal ogive model in 1952 (Hambleton and Swaminathan, 1985), which later evolved into Birnbaum's (1968) 3-PL logistic model. In the 3-PL model a logistic function is used to relate person ability and item parameters to the probability of correct responses to each item. The equation for the 3-PL model is:

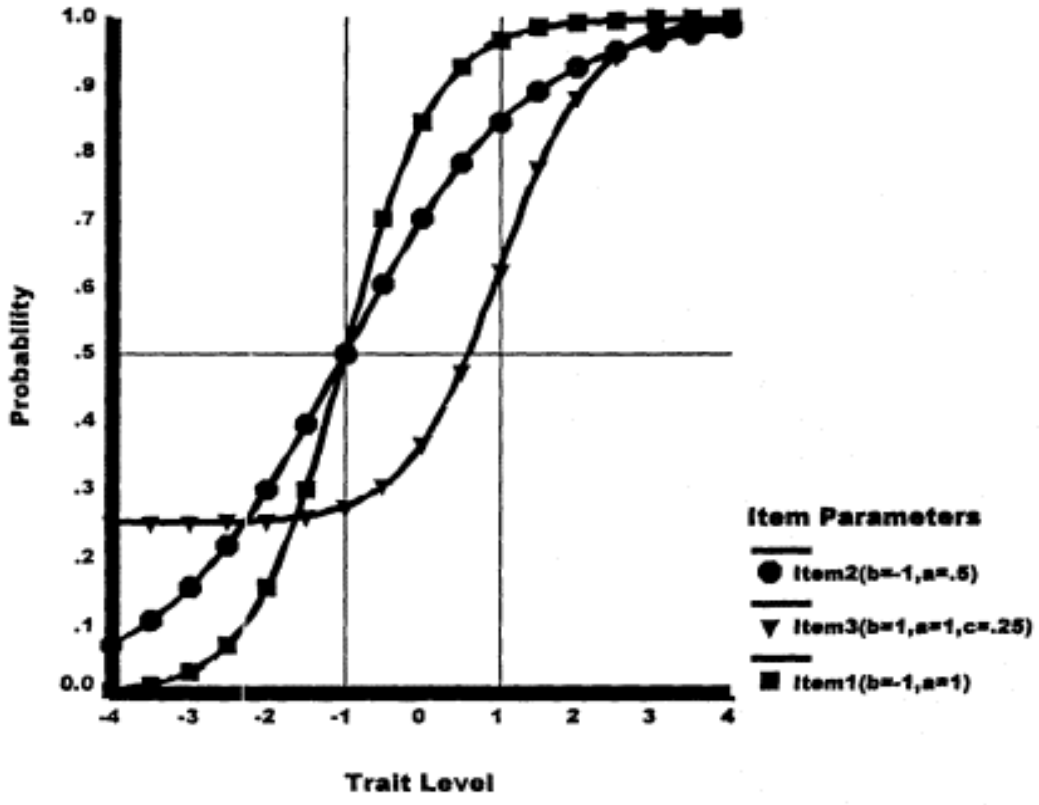
$$P_i(\theta) = c_i + (1 - c_i)\frac{\exp[Da_i(\theta - b_i)]}{1 + \exp[Da_i(\theta - b_i)]},$$

where $P_i(\theta)$ is the probability of an examinee with ability θ answering item *i* correctly. The three parameters characterizing item *i* are a_i, represents item discrimination, b_i representing item difficulty, and c_i representing the guessing parameter. The scaling constant *D* in this formula is equal to 1.702 and is known for its use in item response theory as minimizing the difference between the normal and logistic distribution functions (Camilli, 1994). The three IRT parameters (a, b, and c) are free to vary in the 3-PL model. The equation for the 2-PL model is identical while holding the guessing parameter constant at null ($c_i = 0$) across all items allowing only difficulty and discrimination to vary. In the 1-PL model item difficulty is the only parameter that is allowed to vary, while the guessing parameter is set at null (c = 0) and discrimination (a) is held constant for all items. Hence, the 1-PL model establishes a

relationship between item difficulty and person ability only. The assumptions of no guessing and equal discrimination across all items are somewhat controversial.

These parameter functions can be displayed graphically for each item, and the resulting figure is referred to as an Item Characteristic Curve (ICC). This curve is graphed on a coordinate plane, with person ability, (theta) on the horizontal axis and the probability of a correct response at each level of ability is charted on the vertical axis. Three ICCs are shown in Figure 1. The item parameters determine the shape of the ICC: *a* determines the steepness, *b* determines the placement left or right, and *c* determines the lower asymptote. Figure 1 displays

Figure 1. Item characteristic curves of 3-PL logistic IRT models.[1]



[1]Adopted and modified from Embretson and Reise (2000), 47.

the parameters which help determine that item 3 is the most difficult of this set. The item characteristic curve is created by a set of estimated values plotted as points on a graph. When the actual points (theta, ability) are plotted, they fall near but generally not exactly on the predicted line. A certain amount of error is expected when predicting values. As with the regression line, the error is estimated by the distance between the estimated and the actual location of the point. Minimizing this error is essential when evaluating items for expected consistency of examinee response patterns and in overall performance. The largest error in item performance exists where the actual and predicted values have the largest discrepancy between them.

IRT Model for Polytomous Scoring Format

Polytomous models estimate a greater number of item parameters than dichotomous models based on the multiple, categorical responses to polytomous items. Polytomous response categories can be hierarchically scored and result in parameters that are referred to as "category thresholds." For each *m+1* category item, there are *m* threshold parameters that separate the categories *a(k)*. As with dichotomous models, item responses may be analyzed to produce estimates of person ability and item parameters. It is understood that these "estimates" contain some fraction of measurement error as a result of the indirect estimation of proficiency. Polytomous models have recently received increased attention. "With the growing interest in 'authentic measurement' special attention must be given to IRT models that can handle polytomous scoring, since authentic measurement is linked to performance testing and scoring of examinee performance," (Hambleton, Swaminathan, & Rogers, 1991, p.153.)

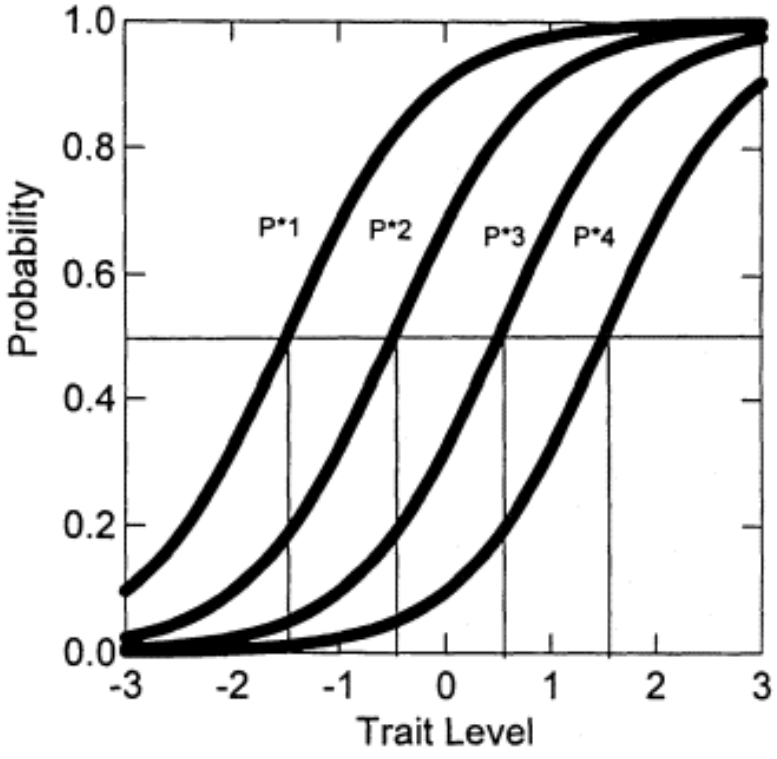
Various mathematical models exist for the purpose of analyzing polytomous item-response data and several are based on IRT theory. Samejima's (1969) Graded Response Model

(GRM) for hierarchically ordered response categories for scoring allows the discrimination parameter to vary across items and between response categories. In this model, the categories (*m*) are represented by k=1,2,…,m, where response *m* reflects the highest θ value:

$$P(k) = \frac{1}{1+\exp[-a(\theta - b_{k-1})]} - \frac{1}{1+\exp[-a(\theta - b_k)]} = P^*(k) - P^*(k+1)$$

in which *a* is the discrimination or slope, b_k is the threshold between categories *(k)*. *P*(k)* describes the probability of a response in category k or higher, for each value of theta. For more specific details on the model refer to Samejima (1969.) This model requires that the category boundaries be a hierarchically-ordered band based on cumulative probabilities.

Figure 2. Item characteristic curves for GRM example item.[2]



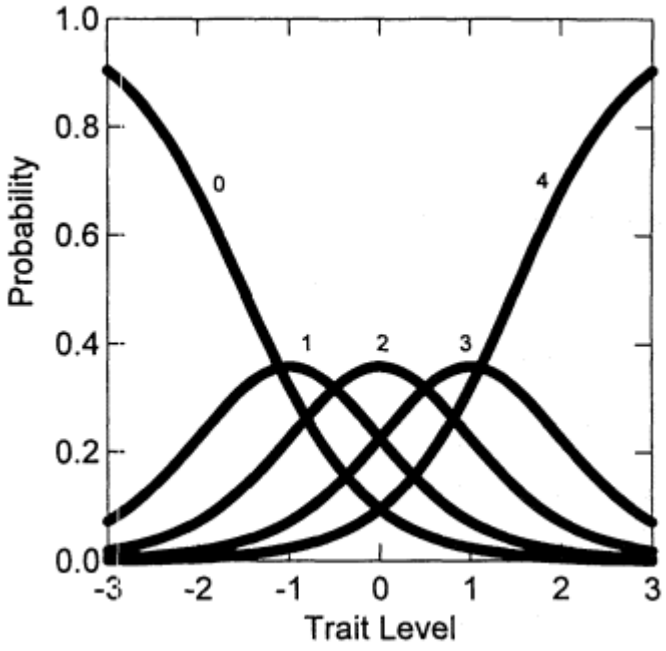
[2]Adopted from Embretson and Reise (2000), 100.

Rasch Techniques

Rasch Model for Dichotomous Scoring Format

Rasch Models were designed by Georg Rasch (1960) to analyze dichotomous item responses are by separately estimating person ability and item difficulty. Georg Rasch derived this approach to item analysis for the purpose of modeling test behavior specifically at the item level and for analyzing dichotomously scored responses. The use of "sufficient statistics" in calculating item and person estimates eliminates the interdependency between them. This is a significant difference between the IRT and Rasch approaches. Rasch models apply a logistic

Figure 3. Five-category threshold parameters of a GRM item[3]



technique to estimate item parameters and person abilities into relative logit measurements. This allows person ability and item difficulty to be compared on a common scale. The Rasch model includes measures of person ability and item difficulty, while holding other item parameters of

[3]Adopted from Embretson and Reise (2000), 101.

discrimination and guessing constant across all items. The assumptions required by Rasch are considered strict and difficult to meet so few data sets are thought to adequately meet the assumptions for analysis.

The Rasch Model has been described as a special case of Birnbaum's 1-PL logistic model (Hambleton, 1989.) However, it is important to remember that IRT models have four assumptions placed on the data, different estimation algorithms, and the 1-PL model has two additional strict assumptions (equal discrimination and no guessing). The Rasch model does have three qualities that make it more attractive to use than some other models: *the ease of use due to fewer parameters, fewer estimation problems because of the fewer parameters, and the specific objectivity regarding the estimation of the item and ability parameters,* which was the reason for its creation (Rasch, 1960.) In IRT applications, an equation is fit to the data, hence more parameters allow for better fit. In Rasch, the data must fit the model to possess the properties of specific objectivity and sufficiency.

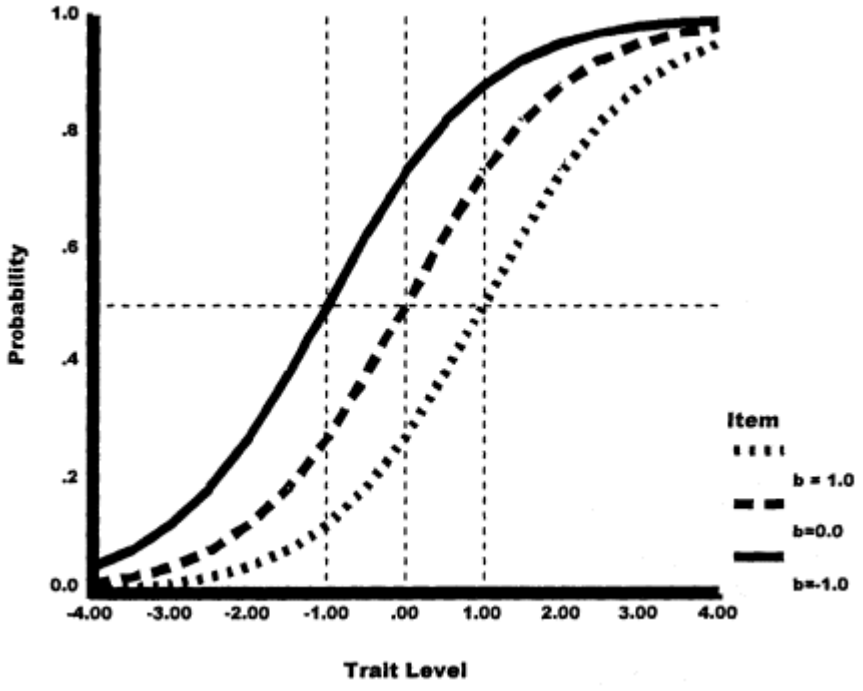
The equation for the Rasch model that represents the probability of a correct response to an item is:

$$P(\theta) = \frac{e^{(\beta_v - \delta_i)}}{1 + e^{(\beta_v - \delta_i)}},$$

where β represents the person ability and δ represents the item difficulty. This logistic function provides an equal interval, linear scale by which measurement of persons and items can be accomplished separately. Rasch referred to this quality as "specific objectivity" (Rasch, 1960.) Sometimes the 1-PL model is used to refer to the "Rasch Model". This general use of the 1-PL term is not advocated by the Rasch supporters (Wright & Stone, 1979, Wright & Masters, 1982). In fact, the Rasch approach has adopted several specific mathematical models for the

analysis of dichotomous, binary, poisson, rating scale, partial credit and facets item-scoring models. The scoring rubrics essentially define the different mathematical models. The Rasch approach uses the unconditional estimation procedure (UCON) which provides great accuracy by taking into account the separate estimation of ability and difficulty and requires sufficient statistics. Other available estimation methods exist for use with item analysis such as joint maximum likelihood (JML), conditional maximum likelihood (CMLE) and marginal maximum likelihood (MML). A detailed discussion of these methods is beyond the scope of this study.

Figure 4. Rasch Model[4]



The theoretical basis of Rasch and IRT methods are very different. The Rasch model is intended for use with data that fits the model. This implies that if the fit is poor, then the data

[4] Adopted from Embretson and Reise (2000), 68.

should be altered to improve the fit by deleting items with poor fit. The IRT approach is meant to provide a model that can be modified to fit data by adding or removing parameters.

Rasch Model for Polytomous Scoring Format

The Partial Credit Model (PCM) (Masters, 1982) is one model that is based on the Rasch approach. The PCM model exhibits many of the characteristics of the dichotomous Rasch model, and provides an opportunity to separate the person and item parameters so that sufficient statistics are achieved in the estimation procedures as described earlier. It requires that the discrimination across all response categories for each item and across all items is held constant and that the category responses for items are ordered in step correspondence to the level of knowledge acquisition. The equation for the PCM is:

$$P_{kni}(\theta) = \frac{e^{(\beta_n - \delta_{ik})}}{1 + e^{(\beta_n - \delta_{ik})}}$$

where $P_{kni}(\theta)$ is the probability of person n responding in category k to item i, β_n is the ability of person n, and δ_{ik} is the difficulty of the k'th "step" in item i. From this expression, adding the requirement that person must respond in one of the available categories, it follows that

$$P_{kni}(\theta) = \frac{e^{\sum_{j=0}^{k}(\beta_n - \delta_{ik})}}{\sum_{h=0}^{m} e^{\sum_{j=0}^{h}(\beta_n - \delta_{ik})}} \qquad k = 0,1,...,m_i.$$

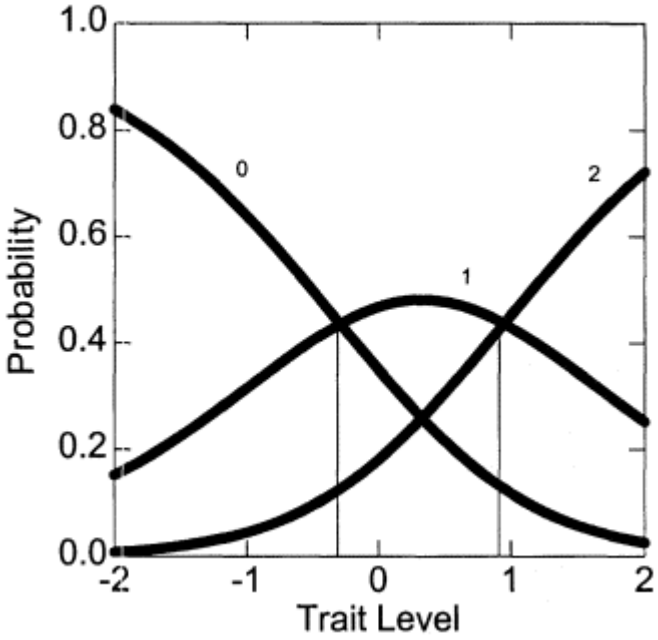
The Partial Credit Model assumes that all item and category discriminations are constant. This is reflected in the category response curves in Figure 5. This model requires that not only is the category boundaries hierarchically ordered but also based on cumulative probabilities, and cannot be separated from the person parameters in the model (Masters, 1982.) The PCM is

expressed at the "step" difficulty level and this allows for the model to separate person and item parameters.

Comparison of IRT and Rasch Theories

A long running controversy exists concerning model selection for item analysis. Rasch and IRT have similarities, but they are conceptually different (Hambleton 1989). IRT models begin with huge item response banks and seek a model to fit the data so that the data patterns may be understood and explained. Georg Rasch, on the other hand, had intentions of developing a measurement system that was based on axioms (Hambleton, 1989). If data fit the model criteria, then it possessed "objectivity, sufficiency and concatenation properties" and could be analyzed. While the theories operate under different assumptions, they both have valuable contributions to make in the field of measurement practice (Hambleton, 1989.)

Figure 5. Category response curves under the partial credit model[5]



[5] Adopted from Embretson and Reise (2000), 107.

The two mathematical models are based on different assumptions about item-response data. How well the data meets those assumptions is an important consideration when choosing an analysis model that has appropriate assumptions for the data, provides an acceptable "fit" to the data and improves the quality of the analysis results. Also, as with the dichotomous models, the polytomous IRT and Rasch Models differ in their consideration of the discrimination parameter. In general, IRT allows the item parameter of discrimination to vary across items and categories while Rasch Models hold discrimination constant.

IRT typically requires very large data sets for estimation of person ability in contrast to Rasch models that require fewer items and fewer examinees for estimation. Additionally, Rasch and IRT use different estimation algorithms. Rasch models create a common linear, equal interval scale and calculate separate item and person measures, while IRT models don't possess these qualities. The Rasch Model is more theory-driven than data-driven (van der Linden & Hambleton, 1997), and therefore has strict prerequisite conditions placed on data for the model-data fit to produce acceptable analysis results.

Rasch is a latent trait model and has been criticized for limited robustness based on the strict assumptions data must meet for the data to have a good fit to the model. Sykes and Yen (2000) found that even when the data-model fit is good and all assumptions are met, an inflated amount of information per item (up to 300%) is predicted. The exclusion of guessing in the Rasch Model combined with a broad range of discrimination across the items brought Sykes & Yen (2000) to the conclusion that the Rasch model was a poor fit for their mixed item model. The constant discrimination required by Rasch forced items with very high or very low discrimination to be labeled as misfitting items. Glas and Verhelst (1989) recommended a two-stage analysis approach applying the Rasch model first, with its higher standard placed on the

data, before resorting to IRT models for item analysis. However, their research also concluded that the Rasch model posed more challenges during parameter estimation, data-fit assessment and interpretation of results. In defense of these conclusions, they cited the fit of the model to their particular data as poor and strongly emphasized the value of attempting to use Rasch in initial analyses to inspect the results under more stringent assumptions. Glas and Verhelst (1989) recommended applying the Rasch model and, "if the model fails, two directions can be followed depending on the practitioners' motives. If the motive is to construct a Rasch-homogeneous measurement instrument, items must be removed. If the motive is fitting a model to the data, alternative models must be chosen by relaxing the assumptions of the RM" (pg. 94).

Another area of concern that should be addressed when comparing item analysis procedures involves the estimation method chosen. The Conditional Maximum Likelihood (CML), Joint Maximum Likelihood (JML), Unconditional estimation procedure (UCON) and other estimation procedures all described by Masters (1982) and Wright and Masters (1982.) All estimation procedures have differences and similarities that make them more or less desirable depending on the situation. CML procedures do not require an assumption about population distribution to be met, but do require the use of sufficient statistics and therefore cannot be applied with all item analysis procedures. JML depends on prior knowledge of ability (theta) and therefore may not provide consistent estimates as the number of persons in the data set increases. The Marginal Maximum Likelihood procedure seeks to maximize log likelihood equations and is an alternative to the CML and JML estimation methods for models that do not require sufficient statistics (Masters & Wright, 1997).

Methods of Analyzing Mixed-Item Format Tests

The challenges involved in applying a mixed-item format test approach to test design become evident when executing the stages of item construction, scoring and analysis. Constructing both dichotomous and polytomous items, determining scoring rubrics, schema and weighting scales can all prove to be challenging, only to be followed by the difficulties apparent during analysis of the results. Traditional psychometric techniques do not easily apply to the mixed-item format test because none of them were specifically designed to handle the combination of multiple scoring schemas in one test. It is possible to combine models in order to analyze responses from both dichotomous and polytomous items. Several software packages are available for analyzing combined response types such as MULTILOG (Thissen, 1991), PARSCALE (Muraki and Bock, 1996) and WINSTEPS (Linacre and Wright, 1991). Each package contains various options for combining item formats in an analysis. Analysis options are limited and technical manuals provided with each software package review available options in detail. Mixed-item format tests, in practice however, are not currently analyzed in a sophisticated manner using this software.

Weighting Methods

Birnbaum (1968) studied weighting theories and proposed the idea that a mixed-item format test could be described as a composite test composed of subtests, each one defined by a single item type. The items in these subtests may then be subjectively assigned a weight or proportion of the entire test score, such as the decision to assign multiple-choice items a 4-percentage-point weight and essay items a 10-percentage-point weight. There is some disagreement in the measurement community about how this weighting proportion can best be

determined and if subjectivity can reasonably be applied when the goal is to achieve optimal test scoring.

Concerns were raised by Li, Lissitz, and Yang (1999) that IRT techniques required considerable model unidimensionality and appropriate model fit to the data. Additionally, they noted that the proportion of dichotomous-to- polytomous items on a mixed-item format test impacted the equating coefficient precision; therefore, special attention should be given to IRT weighting issues, which have been evaluated but not resolved.

If the model-data fit is good, complete response-pattern scoring is thought to produce optimal ability (theta) estimates compared to scores based on sections of summed scores. Scores based on sections of summed scores are thought to be more precise than approximations of these summed scores. In analysis, an increasing amount of information is lost as analysis techniques move further away from analysis of item response patterns and toward summed scores. Ability estimates, as compared to test scores, do not lend themselves to ease of interpretation by the layman. A recent trend in mixed-item format testing is to represent scores as a linear combination of weighted scores, one from each section separated by item type (Thissen, Nelson and Swygert, 2001.)

In a study by Ercikan, Schwarz, Julian, Burket, Weber, and Link, (1998), weighting procedures were criticized and labeled as problematic for not using information about student performance optimally. This criticism was based on the observation that most constructed-response test sections generally have lower accuracy in ability estimation than selected-response sections. Therefore, the reliability of the ability estimates resulting from a composite of multiple sections cannot exceed those resulting from the individual subtests. In spite of the fact that constructed-response items generally result in a lower reliability coefficient, these items gather

more information and are capable of more precise measurement than selected-response items which provide examinees with an opportunity to guess correct responses regardless of ability.

Logit Methods

An alternative to the weighting method is an approach that scales the item parameters together using a combination of logistic models and a common estimation method. Published studies typically combined item types by applying dichotomous logistic models to dichotomous items and polytomous models to polytomous items. Then the items were scaled together so that a common logistic scale could be determined for item, ability and test statistics. However, reliability and validity statistics commonly applied to single item-format tests have yet to be fully applied to mixed-item format tests.

Summary

Various analysis techniques for conducting test item analysis exist. Test format is an important data consideration when selecting appropriate mathematical models to apply. The mixed-item format test brings new considerations to test design and selection of analysis approaches since CTT was not designed to handle the unique challenges of mixed-item format tests. The ease and familiarity of dividing test items up into subtests and developing a composite score is attractive, however, consideration must be given to the accuracy of the results produced by this approach. Current interest in developing tests containing multiple scoring formats is growing and the subjectivity with which composite scores are derived encourages the consideration of methods that accommodate multiple item formats and scoring schema in one analysis, using IRT and Rasch models rather than CTT.

Optimization of Mixed-Item Format Tests

Test Design

Considering the paradigm shift in analysis from estimating trait levels based on total test scores to item responses, it has been stated that the idea of test score "reliability" should be abandon in favor of the more appropriate measurements of item and test information (Thissen, 1991.) Item and test information have been described using Fisher's formulas by Birnbaum who began work in the field of item analysis in the 1950's. Applying information functions in the stages of test construction and item selection is based on item contribution to overall test information. Upon examination of the relative amounts of information gathered by various item types, it has been proposed by Birnbaum to weight item contributions to the total test score based on their contributions to total test information. This weighting approach can be done through the use of a scale score. Birnbaum has also applied his weighting approach to the optimality theories of test scoring.

Test design efforts involve the manipulation of the following factors: item format, item selection, test length, scoring methods, test refinement procedures, timelines and budgetary concerns. Efforts to optimize test design include writing items that clearly measure the target ability and that efficiently collect maximum information across all levels of ability. The mixed-item format may be an optimal test design because it may take advantage of the pros and cons of both types of items (Carlson, 1996). Achieving a good fit of the selected model to the collected data is also critical if response pattern scores are to achieve optimal theta estimates (Thissen, Nelson and Swygert, 2001).

Lau and Wang (1998) determined that when test-item selection is based on information, mastery classification and administration efficiency are maximized while error rates are

minimized. They also found that combining two types of items enhances reliability and validity of the scores produced, while allowing test length to vary and holding test difficulty constant may reduce error rates especially in dichotomous items. These findings lead to a more accurate mastery decision with SPRT (Lau and Wang, 1998), but the length of response time for test administration is still an issue.

Item and Test Information

The Item Information Function (IIF) is calculated by finding the amount of information produced by each individual item on a test. The formula, derived by Birnbaum (1968), is:

$$I_i(\theta) = \frac{[P'_i(\theta)]^2}{P_i(\theta)Q_i(\theta)} \qquad (i = 1,2,...,n).$$

where $P_i(\theta)$ is the probability that a randomly selected examinee with ability θ will answer item *i* correctly, $Q_i(\theta)$ is the probability that a randomly selected examinee with ability θ will answer item i incorrectly, and $P'_i(\theta)$ is the first derivative of $P_i(\theta)$ with respect to θ.

The Test Information Function (TIF) is calculated by finding the sum of all of the Item Information Functions on the test. The equation is:

$$I(\theta) = \sum_{i=1}^{n} \frac{[P'_i(\theta)]^2}{P_i(\theta)Q_i(\theta)} \quad \text{or} \quad I(\theta) = \sum_{i=1}^{n} I_i(\theta)$$

Items scored on the polytomous level increase the available information provided by any item, and therefore any assessment containing such items, and enhance the ability to evaluate item response patterns. Test information is directly related to the items that are selected to compose the test (Carlson, 1996). Carlson (1996) also concluded that polytomous items are more accurate (collect more information) and efficient (require fewer items to gather set amounts of information than dichotomous items). Also, the amount of information provided by

a test at specific ability (theta) level is inversely related to the precision with which ability is estimated at that point:

$$SE(\theta_0) = \frac{1}{\sqrt{I(\theta_o)}},$$

where $SE(\theta_o)$ is the *standard error of estimation* and is a good measure of the precision of measurement. It is dependent on the level of ability across the range of θ's.

Test Information Function (TIF) is calculated by summing the IIF's for a test. Carlson (1996) found that polytomously scored items tended to provide a relatively larger amount of information than dichotomous items and provided information over a broader range of proficiency estimates than did dichotomous items. He also found that the dichotomous items seem to produce an information curve that more closely matches the examinee proficiency curve than that of polytomous items. Given the time involved in writing, responding to and scoring polytomous items, it seems their increased information comes at a high price, therefore Carlson generalized from his findings that tests combining multiple types of items would capitalize on the advantages of each while reducing the disadvantages.

Donoghue (1994) found that polytomous items might collect between 2.1 to 3.1 times as much information as dichotomous items. This is support for recommending weighting of polytomous items heavier on an assessment. When polytomous items were scored dichotomously, the amount of information gathered decreased considerably but was still greater than that produced by using strictly dichotomous items. Polytomous items take more time and money to administer and score.

Evaluation Criteria for Optimization

As a result of documented benefits attributed to both dichotomous and polytomous item scoring, it is important to give considerations to the design of tests containing a combination of these scoring types. The mixed-item format test approaches optimization when the test design and analysis methods are chosen simultaneously. Test design and analysis methods also both have an impact on the available amount of item and test information (Ercikan et al 1998; Si, 2002). Optimality research by Plake (1993) stated that the most optimal results are produced by test items are those that "best" match the ability of the examinee. Polytomously scored items are expected to estimate ability with the highest level of precision.

Berger (1998) noted that optimal test design research depends on the assumed statistical model applied and the optimality criterion for that model, such as meeting assumptions and data fit to the model. The desired result would be more efficient trait and parameter estimation (i.e. less variance in the estimates). There are two challenges noted by Berger in achieving optimality. The first challenge is related to optimal test design and its dependence on Fischer's information function. This function is dependent on theta values, which in any test design process are initially unknown. This dependence can be resolved by using a complicated, computerized, sequential procedure for estimating thetas in which the item parameter estimates are updated at each step. The second challenge to achieving optimality is the impact of the number of categories and number of category parameters on the test design procedure. The optimal number of categories has not been determined in the literature. Berger also noted that more research is needed concerning optimal item selection for effective ability estimation for a population.

Summary

Wright and Stone (1979) defined the best test design as the test that carefully matches the following qualities: the average item difficulty with the average ability of examinees; the range of item difficulty with the range of examinee ability; and the test length with the target distribution of content. This matching is obviously a challenge when designing a test for a target population if the ability of the presumed examinees is not known. Designing a test with multiple item types is possibly the most appropriate way to attempt matching the item difficulty to examinee ability. These mixed-item formats however can pose a challenge in the item analysis stage.

In addition to test format, analysis methods also require consideration of the model selection, model assumptions, model-data fit, and the model impact on ability estimates. Test design and analysis procedures are mutually dependent and this supports selecting them simultaneously. Optimizing test design and analysis procedures both depend on the fit between the data and the model. The analysis of mixed-item format tests is a recent challenge on the horizon of psychometric theory. Few published studies have attempted to analyze the proportion of scoring formats and optimal analysis method for mixed-item format tests. The design of this study attempts to address this issue.

CHAPTER THREE

MATERIALS AND PROCEDURES

Data Simulation and Design

This study employed an 11 x 3 x 2 fixed effects factorial ANOVA design with eleven test scoring formats (structured proportions of dichotomous and polytomous items), three test lengths (short=10 items, medium=20 items & long=30 items) and two mathematical models (Rasch and IRT). The first of the eleven structured scoring formats consisted of 100% polytomous item scoring with 0% dichotomous scoring, followed by 90% polytomous scoring with 10% dichotomous, and so on until reaching the eleventh format consisting of 0% polytomous and 100% dichotomous scoring. The Rasch model applied in the study consisted of a combination of the IRT 1-PL model and Master's (1982) Partial Credit Model (PCM). The IRT model applied consisted of the 2-PL IRT model and Samejima's (1969) Graded Response Model. The three-factor design resulted in 66 data combinations and they are listed in Table 1.

Table 1

Factor Combinations for Study Design

		Percent of Polytomous/Dichotomous Items										
		100/0	90/10	80/20	70/30	60/40	50/50	40/60	30/70	20/80	10/90	0/100
10	Rasch	R1001	R901	R801	R701	R601	R501	R401	R301	R201	R101	R001
	IRT	Q1001	Q901	Q801	Q701	Q601	Q501	Q401	Q301	Q201	Q101	Q001
20	Rasch	R1002	R902	R802	R702	R602	R502	R402	R302	R202	R102	R002
	IRT	Q1002	Q902	Q802	Q702	Q602	Q502	Q402	Q302	Q202	Q102	Q002
30	Rasch	R1003	R903	R803	R703	R603	R503	R403	R303	R203	R103	R003
	IRT	Q1003	Q903	Q803	Q703	Q603	Q503	Q403	Q303	Q203	Q103	Q003

Item Construction and Simulation

A data simulation was performed to generate polytomous response data for 1000 examinees to 30 items. The program used to generate response patterns (see Appendix B) was based on Muraki's generalized partial credit model and was used to define a "true" ability for each of the examinees and was adopted and modified from the work of Susan Chen (1996). The item parameter values for the 30 items were defined and read into the SAS software program (see Appendix C.)

The item parameters used in this study to simulate the examinee item responses reflect typical item discrimination and difficulty of real test data. These parameters were defined by Si (2002) and generated to model partial credit items in his analysis of various scoring models and their impact on ability estimation. Discrimination parameters ranging from .75 to 2 were randomly assigned to the 30 items representing adequate and varied discrimination. Assigning varied item difficulty (10 easy, 10 medium and 10 difficult items) and varying threshold values from –2.25 to 2.25 both helped ensure enough range to match examinee ability. Examinee abilities were created in a normal distribution using the item parameters and the random seed (seed 3 in the SAS program in Appendix B.) These ability estimates were then used to generate item responses based on the item parameters and the probability of each examinee responding between each pair of category thresholds. Through the programmed transformations examinees with higher ability levels were more likely to record responses from higher scoring response categories. The item response data generated polytomous responses that ranged from 1 to 4 (example response data is displayed in Appendix D). In this study, guessing and speededness in examinee item responses could not be simulated and therefore beyond the scope of this study.

Constructing the Data Sets

Ten-item and twenty-item response sets were randomly selected to build the medium and shorter length tests. Based on test lengths observed throughout the review of literature and considering the length of administration time for typical items scored with a polytomous rubric, the 10-item test was considered short, the 20-item test was medium length and the 30 item set was considered long. The 10, 20 and 30 item response sets were then duplicated 11 times and a predetermined proportion of items were randomly selected from each test and re-scored dichotomously. Eleven proportioned formats across three test lengths were then defined by the percent of items on the test scored polytomously and dichotomously, respectively (100%/0%, 90%/10%, 80%/20%, 70%/30%, 60%/40%, 50%, 50%, 40%/60%, 30%/70%, 20%/80%, 10%/90% and 0%/100%.) For re-scoring polytomous items that ranged from 1-4, dichotomization was based on the criteria that only responses in the highest polytomous response category (4) were binary coded as "correct" while responses in all lower categories (1,2 and 3) were coded as "incorrect" (see Appendix E for scoring proportion.) Therefore, each dichotomous score was obtained by re-coding 1, 2 and 3 as "0" and 4 as "1" (0,1). These manipulations of the data resulted in eleven different mixed-item format tests, each composed of a predetermined proportion of both dichotomous and polytomous data. These combinations of mixed format responses across three test lengths resulted in 33 data sets of response formats which were analyzed using two mathematical approaches: IRT and Rasch.

Analysis of Mixed Item Formats

All item responses were analyzed for unidimensionality in SPSS using the factor analysis procedure with the maximum likelihood extraction method. A single factor structure

was revealed for each set of items. The entire population of 1000 examinees was evaluated in each test format and was performed through the use of the software program MULTILOG (Thissen, 1991). For analysis of the dichotomous item responses the Rasch model was achieved by applying the 1-parameter logistic model with the MULTILOG software program. For analyzing polytomous item responses in a Rasch-based approach the Master's Partial Credit Model (1982) was applied. At this point an acknowledgement to the reader should be made regarding this approach. A different Rasch approach uses the unconditional estimation procedure (UCON) which provides great accuracy by requiring sufficient statistics and taking into account the distributions of ability and difficulty (Wright & Stone, 1979). The UCON estimation procedure was not available in the software used in this study. Marginal maximum likelihood estimation (MML) was available in lieu of UCON in the MULTILOG package and was the method chosen for this study across all models. It was decided that all models applied in the study should be executed within the same software package and with the same estimation procedure for the benefit of comparison purposes. Varying methods, models and packages can result in different results for ability estimates. (Refer to Appendix F for examples of MULTILOG program code.)

Another requirement that must be met by MULTILOG users is the re-coding of dichotomous or binary response categories such that the higher or correct response is coded "2" and the lower or incorrect response is coded "1." A compatibility issue exists with the graded response model that precludes the use of zero for response coding. With partial credit or hierarchical scoring in MULTILOG requires "non-zero" values for use in calculations and will recode any zero categories to meet this requirement.

The Models

The IRT model selected for analysis of dichotomized item responses was the 2-parameter logistic approach allowing discrimination to vary across items. Samejima's Graded Response Model (1969) was the IRT model selected for this analysis for use in analyzing polytomous item responses. The IRT approaches were used in combination for analysis of mixed-item format tests. The marginal-maximum likelihood estimation procedure was applied in both approaches and was verified by Thissen and Steinberg (1986) as a viable approach when comparing across models.

A two-step procedure was applied in the MULTILOG program on the item responses. First, the item parameters, individual item and total test information levels, log-likelihood fit statistics and marginal reliability statistics were calculated and saved for each of the 66 design combinations. Then, the calibrated item parameters were applied during calculation of ability estimates and standard errors of those estimates across each design combination. This two-step procedure was conducted twice across each of the eleven proportioned scoring formats and across all three test lengths; once using the IRT model (2-PL/GRM) and once using the Rasch Model (1-PL/PCM). Further analysis was conducted on the 66 sets of ability estimates and standard errors calculated in MULTILOG by importing them into the SPSS software program. SPSS was used to calculate the fixed effects, interaction effects, power, effect sizes and significance levels first for the ability estimates and then for the standard errors. These analyses were conducted across the 11 proportioned formats by conducting 22 (2 x 3) factorial ANOVAs (11 on the ability estimates and 11 on the standard errors.)

Criteria for Evaluation

Total test information was calculated with MULTILOG for each of the sixty-six factor combinations. Optimality criteria such as relative standard error rates, marginal reliability, fit statistics, ability estimates, and test information are compared and discussed across three test lengths and both analysis models. Total test information was also compared across a range of ability levels. The negative-twice-the-log-likelihood (-2LL) statistic was used to assess model-fit across each of the design combinations. Comparisons were made to determine if significant differences existed between the information levels generated by IRT and Rasch. ANOVA statistics included F value comparison to critical F for significance, effect size estimates (eta squared) and power. The test formats were compared based on their relative total test information and the average amount contributed over the width of the range of proficiency. The test information was summed across all nine quadrature points used during estimation and was analyzed and reported for comparison across models.

The study investigated optimal test design features based on overall test information and error rates for the predetermined proportions of scoring in the mixed-item format models. Ability estimates were compared across all eleven mixed-format test designs and across the three test lengths. Marginal reliability calculations were compared for the purpose of investigating optimal test design and analysis procedures. Ideal design and analysis procedures were proposed with regard to the mixed-item format test.

CHAPTER FOUR

RESULTS

Overview

As described in Chapter 3, the SAS software program was used to simulate true ability estimates (thetas) for a population of 1000 examinees. These theta values were then used to generate response sets to 30 test items that were scored on a polytomous scale of 1 through 4. These response sets were randomly separated into tests of 3 different lengths that were then systematically scored over 11 scoring formats. Each of these formats combined predetermined proportions of dichotomous and polytomous scoring. The 33 response sets were each analyzed with both IRT and Rasch techniques using MULTILOG software. With this software a two-stage approach was applied and 66 sets of results were generated. Within multiple files, these results contained theta estimates, test information, marginal reliabilities and fit indices for each design combination. Using SPSS software descriptive statistics and related analyses were performed on the MULTILOG output and are explained for each research question.

The simulated theta distribution is displayed in Figure 6 with frequency and percent distributions listed in Table 2. The mean of the simulated "true" theta values was 0.07 with a standard deviation of 1.03. The expected simulated true mean should be zero with a standard deviation of 1. Simulated theta values spanned a ranged from –3.514 to 3.886. Skewness (symmetry of the distribution) and kurtosis (peakedness of the distribution) were both in the normal range from –0.039 to 0.225, respectively.

Figure 6

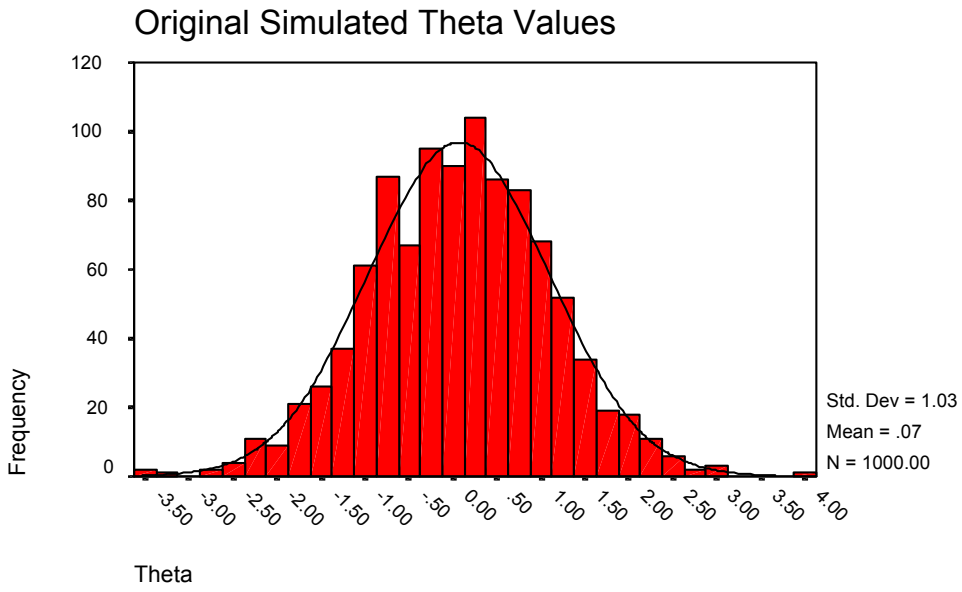


Table 2

Simulated Examinee Ability Distribution

Theta Quartile Midpoints	Frequency	Cumulative Frequency	Percent	Cumulative Percent
-3.5	3	3	0.3	0.3
-3	0	3	0	0.3
-2.5	12	15	1.2	1.5
-2	27	42	2.7	4.2
-1.5	46	88	4.6	8.8
-1	133	221	13.3	22.1
-0.5	145	366	14.5	36.6
0	201	567	20.1	56.7
0.5	179	746	17.9	74.6
1	139	885	13.9	88.5
1.5	61	946	6.1	94.6
2	37	983	3.7	98.3
2.5	12	995	1.2	99.5
3	4	999	0.4	99.9
3.5	0	999	0	99.9
4	1	1000	0.1	100

Comparable theta statistics for each IRT factor combination are listed in Table 3 and for Rasch design combinations in Table 4. Each design combination is represented by the specific analysis

model (Rasch or IRT), test length (10, 20 or 30 items) and percentage of items on the test scored polytomously (10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90% or 100%). The entire population (N=1000) of examinee responses was used in the calculations of each design combination to generate theta estimates, item and test information across 9 quadrature points, Table 3

Descriptive Statistics for IRT Combinations across Test Lengths and Proportions

Model	Test Length	% Poly	Range	Min	Max	Sum	Mean	Std. Error	Std. Deviation	Variance
Original Thetas	30	100	7.400	-3.514	3.886	66.755	0.067	0.032	1.028	1.056
IRT	10	0	3.669	-1.360	2.309	33.519	0.034	0.027	0.852	0.727
IRT	10	10	3.847	-1.541	2.306	23.849	0.024	0.028	0.880	0.774
IRT	10	20	3.911	-1.595	2.316	35.806	0.036	0.028	0.890	0.792
IRT	10	30	3.974	-1.654	2.320	34.462	0.034	0.029	0.906	0.820
IRT	10	40	4.064	-1.667	2.397	28.564	0.029	0.029	0.915	0.837
IRT	10	50	4.315	-1.923	2.392	28.863	0.029	0.029	0.919	0.845
IRT	10	60	4.717	-2.329	2.388	17.953	0.018	0.029	0.933	0.870
IRT	10	70	4.971	-2.584	2.387	18.947	0.019	0.030	0.939	0.881
IRT	10	80	4.932	-2.581	2.351	24.170	0.024	0.030	0.935	0.875
IRT	10	90	4.973	-2.604	2.369	10.497	0.010	0.030	0.942	0.887
IRT	10	100	5.145	-2.769	2.376	5.128	0.005	0.030	0.952	0.906
IRT	20	0	4.585	-1.896	2.689	-44.231	-0.044	0.030	0.935	0.873
IRT	20	10	5.006	-2.206	2.800	-9.298	-0.009	0.031	0.968	0.937
IRT	20	20	5.125	-2.320	2.805	-26.845	-0.027	0.031	0.980	0.961
IRT	20	30	5.503	-2.644	2.859	-30.599	-0.031	0.032	1.003	1.006
IRT	20	40	5.799	-2.901	2.898	-12.975	-0.013	0.032	1.027	1.055
IRT	20	50	5.688	-2.772	2.916	-19.321	-0.019	0.033	1.042	1.086
IRT	20	60	5.757	-2.763	2.994	45.587	0.046	0.034	1.062	1.128
IRT	20	70	5.859	-2.764	3.095	84.278	0.084	0.034	1.087	1.181
IRT	20	80	6.133	-3.041	3.092	78.750	0.079	0.035	1.094	1.197
IRT	20	90	6.249	-3.139	3.110	74.287	0.074	0.035	1.096	1.202
IRT	20	100	6.371	-3.208	3.163	92.838	0.093	0.035	1.108	1.227
IRT	30	0	4.517	-2.016	2.501	0.842	0.001	0.029	0.928	0.861
IRT	30	10	4.699	-2.092	2.607	2.321	0.002	0.030	0.940	0.883
IRT	30	20	5.183	-2.468	2.715	2.313	0.002	0.030	0.958	0.918
IRT	30	30	5.530	-2.724	2.806	4.814	0.005	0.031	0.971	0.942
IRT	30	40	5.648	-2.807	2.841	-27.978	-0.028	0.030	0.962	0.926
IRT	30	50	5.698	-2.824	2.874	-28.933	-0.029	0.030	0.964	0.929
IRT	30	60	6.758	-3.946	2.812	218.358	0.218	0.030	0.942	0.887
IRT	30	70	5.787	-2.928	2.859	-54.531	-0.055	0.031	0.973	0.946
IRT	30	80	5.528	-2.732	2.796	-46.690	-0.047	0.030	0.961	0.923
IRT	30	90	5.886	-3.079	2.807	-74.149	-0.074	0.030	0.954	0.910
IRT	30	100	6.043	-3.103	2.940	-66.665	-0.067	0.030	0.959	0.920

reliability and fit statistics. The mean calculations for each set of theta estimates ranged from –0.074 to 0.218.The range of the estimated theta values for each of the 66 design combinations spanned from 2.847 to 6.758, all smaller than the original simulated theta range of 7.4.

Table 4

Descriptive Statistics for Rasch Combinations across Test Lengths and Proportions

Model	Test Length	% Poly	Range	Min	Max	Sum	Mean	Std. Error	Std. Deviation	Variance
Original Thetas	30	100	7.400	-3.514	3.886	66.755	0.067	0.032	1.028	1.056
Rasch	10	0	3.769	-1.363	2.406	16.776	0.017	0.027	0.845	0.715
Rasch	10	10	3.920	-1.502	2.418	9.128	0.009	0.027	0.862	0.743
Rasch	10	20	4.015	-1.575	2.440	22.923	0.023	0.028	0.872	0.760
Rasch	10	30	4.098	-1.654	2.444	21.145	0.021	0.028	0.884	0.782
Rasch	10	40	4.040	-1.682	2.358	22.016	0.022	0.028	0.901	0.812
Rasch	10	50	4.384	-2.000	2.384	22.305	0.022	0.029	0.910	0.829
Rasch	10	60	4.658	-2.262	2.396	23.583	0.024	0.029	0.923	0.852
Rasch	10	70	4.813	-2.408	2.405	23.449	0.023	0.029	0.933	0.870
Rasch	10	80	4.747	-2.413	2.334	23.380	0.023	0.029	0.928	0.861
Rasch	10	90	3.410	-1.028	2.382	103.450	0.103	0.016	0.510	0.260
Rasch	10	100	4.914	-2.594	2.320	-2.335	-0.002	0.030	0.942	0.887
Rasch	20	0	4.766	-1.842	2.924	-13.585	-0.014	0.029	0.931	0.866
Rasch	20	10	5.039	-2.107	2.932	1.250	0.001	0.031	0.967	0.935
Rasch	20	20	5.136	-2.213	2.923	-34.497	-0.034	0.031	0.965	0.932
Rasch	20	30	5.481	-2.512	2.969	-6.424	-0.006	0.031	0.986	0.972
Rasch	20	40	5.63	-2.698	2.932	-14.420	-0.014	0.032	1.017	1.034
Rasch	20	50	5.542	-2.564	2.978	1.650	0.002	0.033	1.031	1.062
Rasch	20	60	5.632	-2.580	3.052	49.251	0.049	0.033	1.046	1.094
Rasch	20	70	5.651	-2.604	3.047	64.623	0.065	0.034	1.065	1.135
Rasch	20	80	5.874	-2.792	3.082	91.982	0.092	0.034	1.072	1.150
Rasch	20	90	5.918	-2.886	3.032	88.407	0.088	0.034	1.078	1.162
Rasch	20	100	6.054	-2.980	3.074	113.251	0.113	0.035	1.092	1.193
Rasch	30	0	4.861	-1.950	2.911	6.110	0.006	0.029	0.932	0.869
Rasch	30	10	4.699	-2.092	2.607	2.321	0.002	0.030	0.940	0.883
Rasch	30	20	5.183	-2.468	2.715	2.313	0.002	0.030	0.958	0.918
Rasch	30	30	5.530	-2.724	2.806	4.814	0.005	0.031	0.971	0.942
Rasch	30	40	5.648	-2.807	2.841	-27.978	-0.028	0.030	0.962	0.926
Rasch	30	50	5.472	-2.677	2.795	-11.150	-0.011	0.030	0.962	0.926
Rasch	30	60	2.847	-1.521	1.326	154.132	0.154	0.017	0.549	0.302
Rasch	30	70	5.560	-2.740	2.820	-28.975	-0.029	0.031	0.973	0.947
Rasch	30	80	5.528	-2.732	2.796	-46.690	-0.047	0.030	0.961	0.923
Rasch	30	90	5.643	-2.878	2.765	-35.286	-0.035	0.031	0.967	0.936
Rasch	30	100	5.785	-2.948	2.837	-22.053	-0.022	0.031	0.970	0.941

Skewness and kurtosis statistics for theta estimates generated with IRT and Rasch procedures were generally reflective of normal distributions. Skewness figures ranged from -0.651 to 0.313 and kurtosis ranged from -0.689 to 0.479 with the single exception of IRT for 30 items at 60% polytomous scoring which was slightly leptokurtic at 1.156.

Research Questions

The present study hypothesized that the scoring schema (proportion of dichotomous and polytomous items), test length and analysis method impact the optimization criteria in mixed-item format tests. Optimization criteria, in turn, have an appreciable effect across a mixed-item format on the following: reliability, item information, test information, ability estimates, and item parameters. The research questions investigated in this study were as follows:

Research Question 1

How does the proportion of dichotomous and polytomous items across test lengths and analysis models impact the examinee ability estimates?

Eleven three-by-two ANOVAs (one for each proportion level) were conducted on the 66 sets of theta estimates with test length and model as independent variables. Main and interaction effects were evaluated for significance by comparing the observed F values to the critical F value. This critical F value is determined by the degrees of freedom for each effect and the alpha level which was set at $\alpha=.05$. Each $F_{observed}$ was compared to $F_{critical\ (.05,\ 1,\ \infty)}=3.84$ for testing model significance and $F_{critical\ (.05,\ 2,\ \infty)}=3.00$ for testing significance of test length and interaction between model and test length. Effect size was also evaluated. Effect size is the ratio of the between-groups sum of squares and the total sum of squares, and in these analyses

represents the proportion of variance in the theta estimates explained by the difference among groups. Effect size estimates were measured using partial eta squared values and were evaluated based on what Cohen (1988) reported as the general reference for effect size for F tests in ANOVA (small=.1, medium=.25 and large=.4).Observed power calculations were also calculated using SPSS and are included in the ANOVA tabled results.

ANOVAs calculated for proportion levels 0% through 50% polytomous scoring resulted in non-significant F values for all main and interaction effects. These F values ranged from .003 to 2.054 for all effects. The ANOVA for one proportion level (90% polytomous) produced a significant F value for the model main effect $F_{observed}$=3.977 with an effect size of .001 and power equal to .524. The ANOVAs conducted on proportion levels 60% through 100% polytomous scoring all resulted in significant F values for the test length main effect. These F values ranged from $F_{observed}$=6.813 to $F_{observed}$=18.455 with effect size remaining equal to or below .006 and power remaining equal to or above .921. None of the 11 ANOVAs produced a significant F value for the model-by-test length interaction effect. Results are displayed in Tables 5, 6, 7, 8 and 9.

Table 5

Two-Way Fixed Effects ANOVA on Theta Estimates: Proportion 60% (Model, Test Length)

Source	Type III Sum of Squares	df	Mean Square	F	P value	Partial Eta Squared	Non-Central Parameter	Observed Power(a)
Corrected Model	33.662(b)	5	6.732	7.869	.000	.007	39.347	1.000
Intercept	43.157	1	43.157	50.446	.000	.008	50.446	1.000
MODEL	.503	1	.503	.588	.443	.000	.588	.120
LENGTH	31.577	2	15.789	18.455	.000	.006	36.910	1.000
MODEL * LENGTH	1.582	2	.791	.925	.397	.000	1.849	.211
Error	5127.947	5994	.856					
Total	5204.766	6000						
Corrected Total	5161.609	5999						

a Computed using alpha = .05; b R Squared = .007 (Adjusted R Squared = .006)

Table 6

Two-Way Fixed Effects ANOVA on Theta Estimates: Proportion 70% (Model, Test Length)

Source	Type III Sum of Squares	df	Mean Square	F	P Value	Partial Eta Squared	Non Central Parameter	Observed Power(a)
Corrected Model	14.064(b)	5	2.813	2.832	.015	.002	14.159	.843
Intercept	1.936	1	1.936	1.950	.163	.000	1.950	.287
MODEL	.018	1	.018	.018	.893	.000	.018	.052
LENGTH	13.535	2	6.767	6.813	.001	.002	13.626	.921
MODEL * LENGTH	.512	2	.256	.258	.773	.000	.515	.091
Error	5953.929	5994	.993					
Total	5969.930	6000						
Corrected Total	5967.993	5999						

a Computed using alpha = .05; b R Squared = .002 (Adjusted R Squared = .002)

Table 7

Two-Way Fixed Effects ANOVA on Theta Estimates: Proportion 80% (Model, Test Length)

Source	Type III Sum of Squares	df	Mean Square	F	P Value	Partial Eta Squared	Non Central Parameter	Observed Power(a)
Corrected Model	17.553(b)	5	3.511	3.553	.003	.003	17.763	.923
Intercept	2.600	1	2.600	2.631	.105	.000	2.631	.368
MODEL	.026	1	.026	.026	.872	.000	.026	.053
LENGTH	17.465	2	8.733	8.837	.000	.003	17.675	.972
MODEL * LENGTH	.062	2	.031	.031	.969	.000	.063	.055
Error	5922.938	5994	.988					
Total	5943.091	6000						
Corrected Total	5940.491	5999						

a Computed using alpha = .05; b R Squared = .003 (Adjusted R Squared = .002)

Table 8

Two-Way Fixed Effects ANOVA on Theta Estimates: Proportion 90% (Model, Test Length)

Source	Type III Sum of Squares	df	Mean Square	F	P Value	Partial Eta Squared	Non Central Parameter	Observed Power(a)
Corrected Model	26.230(b)	5	5.246	5.877	.000	.005	29.386	.995
Intercept	4.660	1	4.660	5.220	.022	.001	5.220	.627
MODEL	3.550	1	3.550	3.977	.046	.001	3.977	.514
LENGTH	21.055	2	10.527	11.794	.000	.004	23.588	.995
MODEL * LENGTH	1.625	2	.813	.910	.402	.000	1.821	.208
Error	5350.280	5994	.893					
Total	5381.169	6000						
Corrected Total	5376.510	5999						

a Computed using alpha = .05; b R Squared = .005 (Adjusted R Squared = .004)

Table 9

Two-Way Fixed Effects ANOVA on Theta Estimates: Proportion 100% (Model, Test Length)

Source	Type III Sum of Squares	df	Mean Square	F	P Value	Partial Eta Squared	Non Central Parameter	Observed Power(a)
Corrected Model	24.000(b)	5	4.800	4.741	.000	.004	23.707	.979
Intercept	2.407	1	2.407	2.377	.123	.000	2.377	.338
MODEL	.552	1	.552	.545	.460	.000	.545	.114
LENGTH	22.769	2	11.385	11.245	.000	.004	22.491	.993
MODEL * LENGTH	.679	2	.340	.335	.715	.000	.671	.104
Error	6068.224	5994	1.012					
Total	6094.631	6000						
Corrected Total	6092.225	5999						

a Computed using alpha = .05; b R Squared = .004 (Adjusted R Squared = .003)

To answer the first research question, it is important to remind the reader that effect sizes are relative to the variables studied. These analyses indicated that as the proportion of polytomous items increased beyond 50% of the test, test length did impact the ability estimates although effect size for each comparison was negligible. Further analyses were conducted to investigate the effect of the combinations within the test lengths and analysis models. Figures 7

through 12 display graphically the effects that the combinations of the two factors had on the theta estimates.

Post hoc tests were conducted following significant F tests to determine where the differences could be found using the Scheffe homogeneous test of subsets at an alpha level equal to .05. At the 60% proportion of polytomous items the 10- and 20-item length tests were similar in marginal means. However the 30-item test length performance was significantly different from the shorter tests (see Figure 7). At the 70% and 80% proportion level the 20- and 30-item tests were significantly different (Figure 8 and Figure 9). At the 90% proportion level the 30-item test differed significantly in marginal mean from the two shorter tests (Figure 10) and the Rasch and IRT models differed significantly across the test lengths (Figure 11). At the 100% proportion level the marginal means of the 20-item test differed significantly from the two other tests (Figure 12).

Figure 7. Marginal means of theta estimation by model across test length (60% proportion)

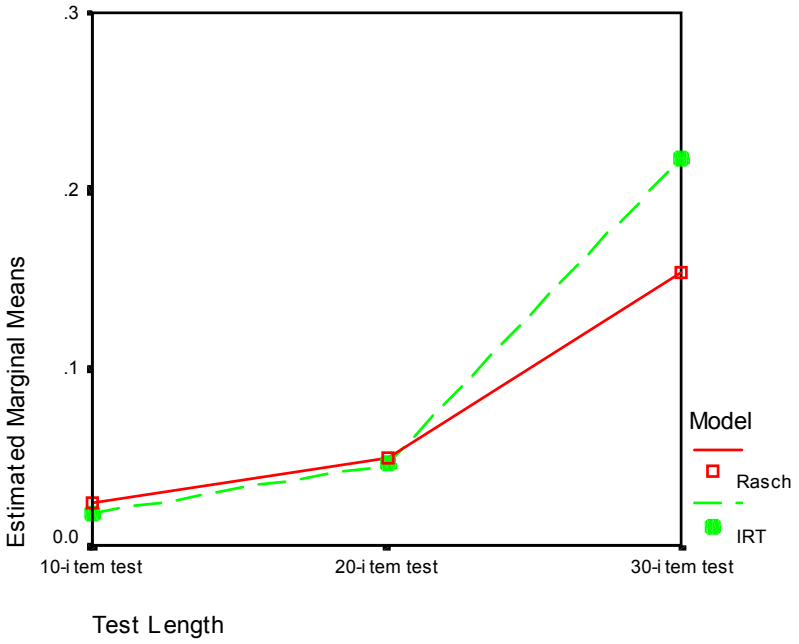


Figure 8. Marginal means of theta estimation by model across test length (70% proportion)

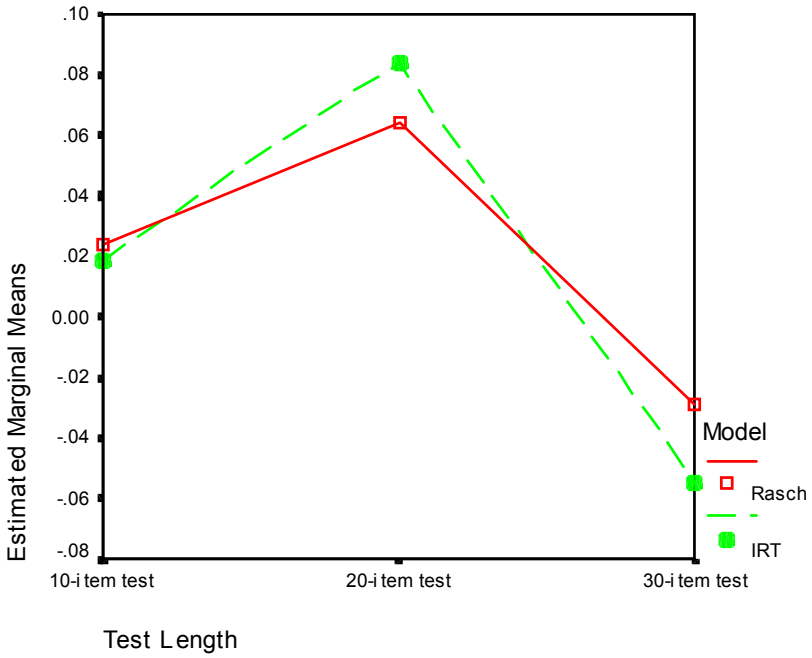


Figure 9. Marginal means of theta estimation by model across test length (80% proportion)

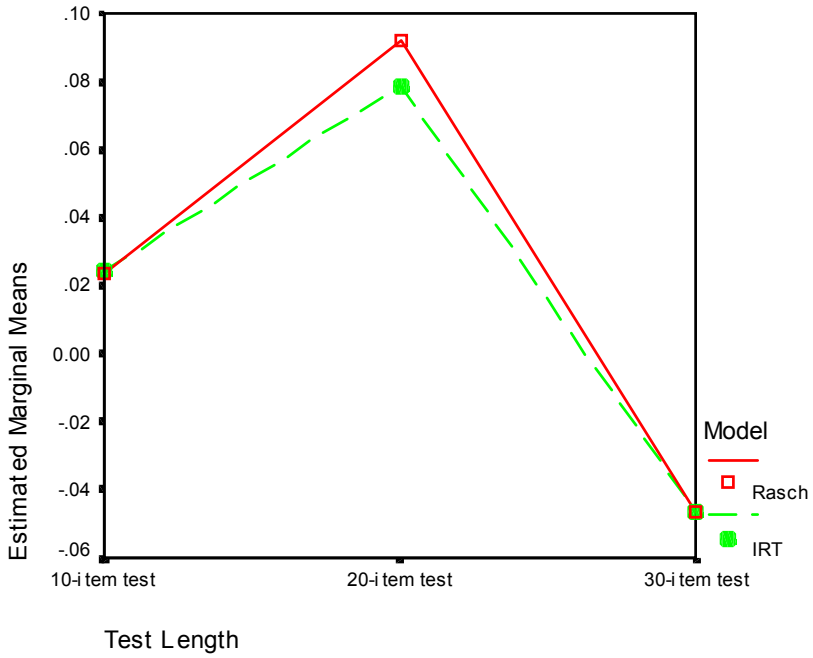


Figure 10. Marginal means of theta estimation by model across test length (90% proportion)

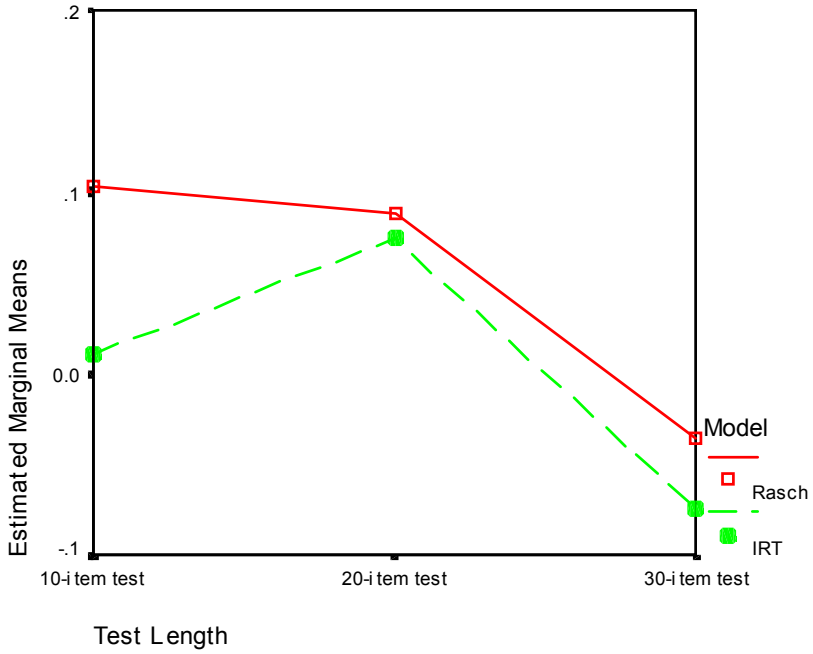


Figure 11. Marginal means of theta estimation by test length across models (90% proportion)

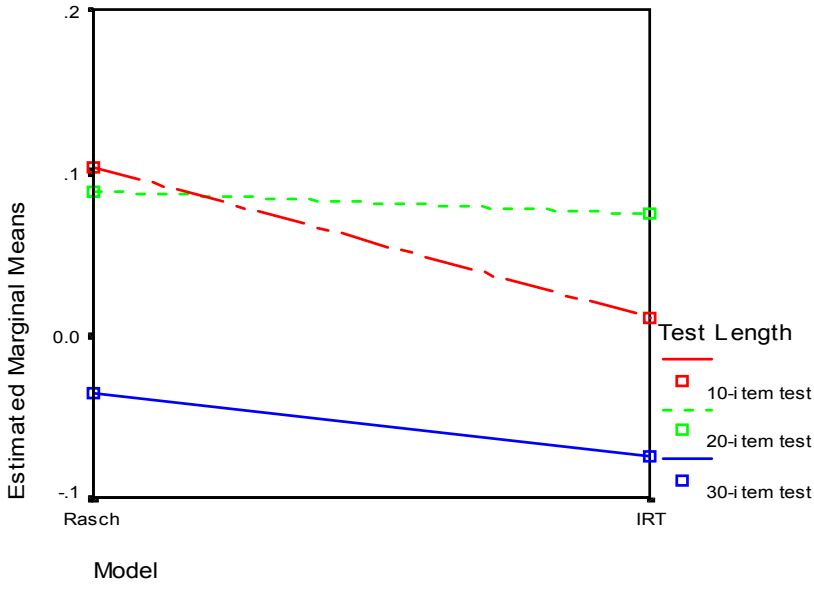
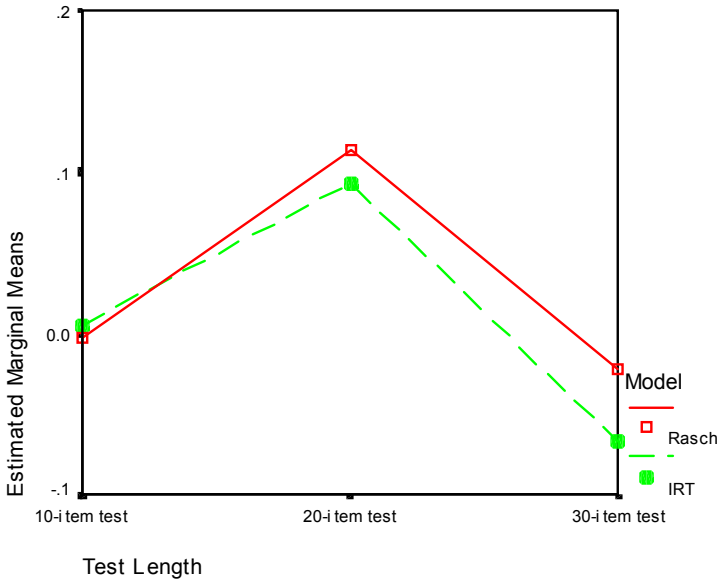


Figure 12. Marginal means of theta estimation by model across test length (100% proportion)



Research Question 2

How does the proportion of dichotomous and polytomous items across test lengths and analysis models impact the Total Test Information?

The test information across theta levels, information sum and proportion improvement (between analysis models) for each design combination are displayed in Tables 10, 11 and 12. These tables show that for all test lengths, in general, information is highest toward the middle of the theta distribution, information across theta and overall has a direct relationship to the proportion of polytomous items, and IRT analysis improves total information over Rasch.

From the information in Tables 10, 11 and 12, it is evident that tests with a larger proportion of items scored polytomously gather more information than those scored dichotomously. This is especially noticeable in the lower ability levels where the highest reported information was almost always at the 100% proportion polytomous level. At higher ability levels, the highest information reported tended to be at less than 100% polytomous

scoring. Figures 13, 14 and 15 display information levels across the theta distribution for tests of 10, 20 and 30 items in length.

Table 10

Test Information Distribution, Sum and Improvement (10-Item Tests)

Model	% Poly	Information Over Range of Theta Levels									Info Sum	Proportion Improvement
		-2.0	-1.5	-1.0	-0.5	0.0	.5	1.0	1.5	2.0		
IRT	0	1.8	3.5	3.5	4.5	5.6	6.7	6.5	6.5	5.6	44.2	11.5%
IRT	10	2.2	3.2	4.4	5.3	6.1	6.9	6.7	6.7	5.7	47.2	11.0%
IRT	20	2.3	3.7	5.7	6.9	7.3	7.4	6.9	6.8	5.8	52.8	11.9%
IRT	30	2.4	4.2	6.9	8.4	8.4	7.9	7.1	**6.9**	**5.9**	58.1	11.2%
IRT	40	2.5	4.3	7.1	8.9	9.3	9.1	7.7	6	5	59.9	4.0%
IRT	50	2.8	4.6	7.3	9.1	9.5	9.2	7.8	6	5	61.3	4.9%
IRT	60	3.5	5.2	7.7	9.2	9.5	9.2	7.8	6	5	63.1	3.2%
IRT	70	4.1	5.6	8	9.4	9.6	9.3	7.8	6	5	64.8	0.5%
IRT	80	4.1	5.7	8.2	9.8	10.1	9.8	**8.2**	6.1	4.7	66.7	
IRT	90	4.3	6	8.5	10.1	**10.4**	**9.9**	8.1	5.8	4.4	67.5	1.8%
IRT	100	**5.4**	**6.8**	**9**	**10.3**	**10.4**	9.8	8	3.8	4.4	**67.9**	
Rasch	0	1.8	2.7	4.1	5.3	5.7	5.6	5.2	4.6	4.1	39.1	
Rasch	10	2.1	3.2	4.8	6.2	6.5	5.9	5	4.4	3.9	42	
Rasch	20	2.3	3.5	5.5	7.6	8.1	6.6	5	4.2	3.7	46.5	
Rasch	30	2.5	3.9	6.2	9	9.9	7.6	5.2	3.9	3.4	51.6	
Rasch	40	2.5	4	6.3	9.2	**10.4**	**8.9**	7	5.3	3.9	57.5	
Rasch	50	3.5	5.1	7.1	9.2	9.7	8.1	6.5	5.1	4	58.3	
Rasch	60	4.5	6.2	7.8	9.4	9.6	8	6.4	5.2	4	61.1	
Rasch	70	5.4	7.2	8.7	9.6	9.3	7.8	6.7	5.4	4.4	64.5	
Rasch	80	5.4	7.2	8.6	**9.5**	9.5	8.3	7.2	**6.3**	**4.9**	66.9	0.3%
Rasch	90	5.3	7	8.4	**9.5**	9.6	8.6	7.4	6	4.5	66.3	
Rasch	100	**6.3**	**8.1**	**9**	**9.5**	9.5	8.6	**7.5**	6.1	4.5	**69.1**	1.7%

Note: bolded values are the highest information value within model, proportion and ability.

Table 11

Test Information Distribution, Sum and Improvement (20-Item Tests)

Model	% Poly	Range of Theta Levels									Info Sum	Proportion Improvement
		-2.0	**-1.5**	**-1.0**	**-0.5**	**0.0**	**.5**	**1.0**	**1.5**	**2.0**		
IRT	0	3.4	5.3	7	8	9.4	11.5	11.6	11.1	9.7	77.0	14.2%
IRT	10	4.3	5.7	7.2	8.2	9.7	11.9	12.1	11.1	9.4	79.6	6.7%
IRT	20	4.9	6.7	8.4	9.4	10.6	12.2	11.9	10.8	9.2	84.1	8.5%
IRT	30	5.9	7.3	8.8	9.6	10.9	12.7	12.8	**11.7**	9.7	89.4	3.2%
IRT	40	6.6	8.1	9.9	10.7	11.5	12.8	12.7	**11.7**	9.8	93.8	5.7%
IRT	50	7.3	9	10.7	11.3	11.8	12.6	12.4	11.5	9.8	96.4	6.1%
IRT	60	7.7	9.9	12.2	12.8	12.6	12.6	12.3	11.4	**10.2**	101.7	4.6%
IRT	70	7.7	10.1	12.8	13.9	13.8	13.7	13	11.1	9.1	105.2	0.7%
IRT	80	8.3	10.5	13.1	14	13.9	13.7	13	11.1	9.1	106.7	0.6%
IRT	90	8.8	11	13.5	14.4	14.3	**14.1**	**13.3**	11.1	8.8	109.3	0.6%
IRT	100	**9.6**	**11.7**	**13.9**	**14.7**	**14.5**	**14.1**	13.2	11.1	8.7	**111.5**	2.7%
Rasch	0	3.3	5.3	7.4	8.9	9.4	9.6	9.1	7.8	6.6	67.4	
Rasch	10	4.9	7	8	8.7	9.2	9.9	10.1	9.3	7.5	74.6	
Rasch	20	5	7.1	9	10.4	10.9	10.5	9.6	8.2	6.8	77.5	
Rasch	30	7.3	9.1	9.4	9.8	10.2	10.5	10.7	**10.5**	**9.1**	86.6	
Rasch	40	7.8	9.2	9.9	11	11.7	11.3	10.4	9.4	8	88.7	
Rasch	50	8.3	9.9	10.9	11.9	12.2	11.2	9.9	8.9	7.7	90.9	
Rasch	60	8.6	10.5	12	13.6	13.9	12.3	10.1	8.7	7.5	97.2	
Rasch	70	8.9	10.9	12.4	**14.4**	**15.2**	**13.6**	11.4	9.7	8	104.5	
Rasch	80	10.1	12.1	13.3	**14.4**	14.7	13.1	11	9.4	8	106.1	
Rasch	90	10.7	12.6	**13.5**	14.3	14.4	13	11.3	10.1	8.7	**108.6**	
Rasch	100	**11.1**	**12.8**	13.4	13.9	14.1	13.1	**11.5**	10.1	8.6	**108.6**	

Table 12

Test Information Distribution, Sum and Improvement (30-Item Tests)

Model	% Poly	Range of Theta Levels									Info Sum	Proportion Improvement
		-2.0	-1.5	-1.0	-0.5	0.0	.5	1.0	1.5	2.0		
IRT	0	4.3	7.1	9.3	10.3	13.2	18.7	21.6	**32.4**	**24.3**	141.2	34.7%
IRT	10	5	8.6	11.8	12.9	15	18.8	20.3	27.7	22.7	142.8	25.0%
IRT	20	6.5	10	14.1	15.7	17.6	20.8	20.8	21.2	20.6	147.3	13.0%
IRT	30	7.5	10.7	14.7	16.6	18.9	22.1	21.6	20.4	18.9	151.4	7.1%
IRT	40	9.1	12.2	16.3	18.6	21.8	25.4	23.6	20.2	16.7	163.9	
IRT	50	9.8	13.7	18	20.3	23.4	26.9	25	21.2	16.9	175.2	4.9%
IRT	60	11.7	14.9	19.3	21.4	29.7	28.2	**35.1**	23.1	16.5	199.9	11.6%
IRT	70	13.5	17.6	22.7	24.5	26.2	27.7	24.9	20.9	16.7	194.7	4.1%
IRT	80	16.3	21.3	28.4	30.8	31.7	31.6	26.7	20.9	16.4	224.1	6.8%
IRT	90	16.1	20.2	23.2	28.4	30.7	28	25.5	22.1	15.6	209.8	
IRT	100	**17.9**	**22.8**	**29.6**	**31.8**	**32.5**	**32.2**	27	20.7	15.9	**230.4**	5.7%
Rasch	0	4.2	7.3	10.7	12.7	13.8	15	14.8	13.5	12.8	104.8	
Rasch	10	4.7	8.3	12.2	14.9	16	16.2	15.4	13.8	12.7	114.2	
Rasch	20	6.9	9.8	13.4	17.1	18.9	18.9	17.3	15.2	12.8	130.3	
Rasch	30	8.6	10.9	13.6	17.1	19.4	20	19.9	17.8	14.1	141.4	
Rasch	40	16.3	13.2	15	17.8	20.5	22	23.1	21.2	15.6	164.7	0.5%
Rasch	50	10.3	13.5	16.3	20	22.6	23.1	23.2	21.6	**16.4**	167	
Rasch	60	13.8	16.6	17.9	21.2	23.9	24	23.6	21.8	**16.4**	179.2	
Rasch	70	14.9	18.3	20.5	24.1	25.9	24.3	22.7	20.6	15.8	187.1	
Rasch	80	16.1	20.2	23.2	28.4	**30.7**	**28**	**25.5**	22.1	15.6	209.8	
Rasch	90	18.9	23.6	**25.7**	**28.7**	29.5	26.5	24.3	21.6	16	214.8	2.4%
Rasch	100	**19.3**	**23.9**	25.6	28.2	29.5	27.5	25.4	**22.3**	16.2	**217.9**	

Figure 13. Information across theta levels for 10-item tests (0%, 50% and 100% proportions)

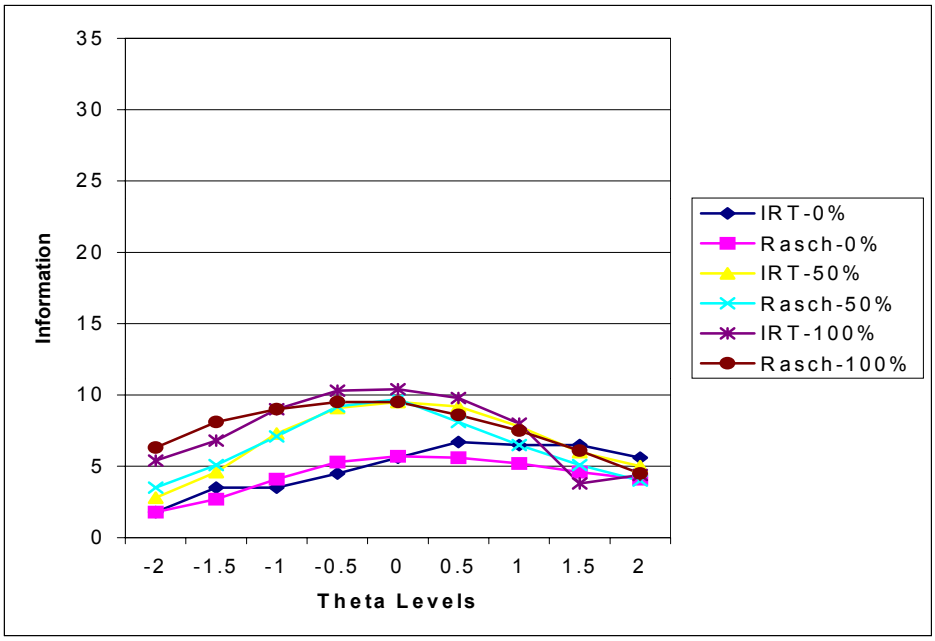


Figure 14. Information across theta levels for 20-item tests (0%, 50% and 100% proportions)

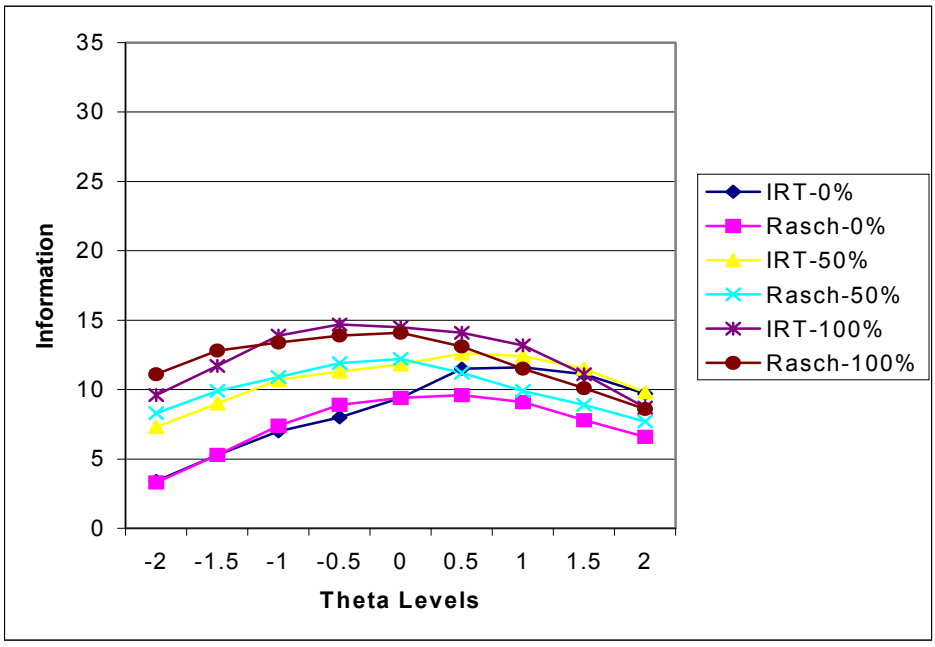


Figure 15. Information across theta levels for 30-item tests (0%, 50% and 100% proportions)

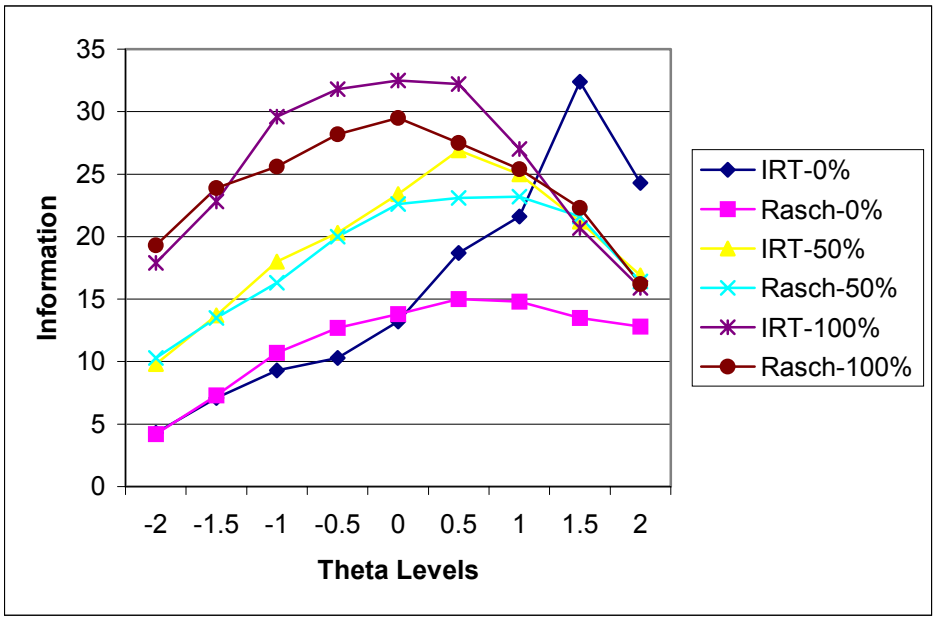


Table 13 displays the percentage increase in information levels when test length is increased from 10 to 20 items for both models at three extreme proportions (0%, 50%, 100% polytomous scoring.) It can be seen that increasing the test length from 10 to 20 items, or doubling the length, increases the average information across models and proportion of polytomous items by 71%. This rate seems to be fairly stable across both models and extreme scoring proportions. Table 14 displays the calculated percentages when test length is increased from 20 to 30 items. The percent increase on average is 82% with the lowest value for Rasch 0% polytomous at 54% and IRT 100% polytomous the highest at 104%. Table 15 displays the calculated percentages when test length is increased from 10 to 30 items (tripled). This increase in test length increases the possible information available by 207% on average across all models and proportions, ranging from 167% for Rasch 0% polytomous to 254% for IRT 100% polytomous.

Table 13

Percentage Increase in Information from 10-Item to 20-Item Test

	-2	-1.5	-1	-0.5	0	0.5	1	1.5	2	Average increase
IRT-0%	89%	51%	100%	78%	68%	72%	78%	71%	73%	76%
Rasch-0%	83%	96%	80%	68%	65%	71%	75%	70%	61%	74%
IRT-50%	161%	96%	47%	24%	24%	37%	59%	92%	96%	71%
Rasch-50%	137%	94%	54%	29%	26%	38%	52%	75%	93%	66%
IRT-100%	78%	72%	54%	43%	39%	44%	65%	192%	98%	76%
Rasch-100%	76%	58%	49%	46%	48%	52%	53%	66%	91%	60%
Overall										71%

Table 14

Percentage Increase in Information from 20-Item to 30-Item Test

	-2	-1.5	-1	-0.5	0	0.5	1	1.5	2	Average increase
IRT-0%	26%	34%	33%	29%	40%	63%	86%	192%	151%	73%
Rasch-0%	27%	38%	45%	43%	47%	56%	63%	73%	94%	54%
IRT-50%	34%	52%	68%	80%	98%	113%	102%	84%	72%	78%
Rasch-50%	24%	36%	50%	68%	85%	106%	134%	143%	113%	84%
IRT-100%	86%	95%	113%	116%	124%	128%	105%	86%	83%	104%
Rasch-100%	74%	87%	91%	103%	109%	110%	121%	121%	88%	100%
Overall										82%

Table 15

Percentage Increase in Information from 10-Item to 30-Item Test

	-2	-1.5	-1	-0.5	0	0.5	1	1.5	2	Average increase
IRT-0%	139%	103%	166%	129%	136%	179%	232%	398%	334%	202%
Rasch-0%	133%	170%	161%	140%	142%	168%	185%	193%	212%	167%
IRT-50%	250%	198%	147%	123%	146%	192%	221%	253%	238%	196%
Rasch-50%	194%	165%	130%	117%	133%	185%	257%	324%	310%	202%
IRT-100%	231%	235%	229%	209%	213%	229%	238%	445%	261%	254%
Rasch-100%	206%	195%	184%	197%	211%	220%	239%	266%	260%	220%
Overall										207%

To answer the second research question, the calculations indicated that the relationship between higher proportions of polytomous items and the amount of available information produced by a test are directly related. All analyses indicated that the test length plays a major role in determining the amount of information that is gathered by the items. This seems obvious

since the total test information is the sum of all of the item information values. With more items present, more information is available.

Research Question 3

How does the proportion of dichotomous and polytomous items across test lengths and analysis models impact the overall optimality of the test?

Three criteria were used to evaluate test optimality across the design combinations in this study: marginal reliability, ANOVA analyses of the standard error rates for each design combination, and information levels. Table 16 displays the marginal reliability values across the design combinations. Within each test length reliability increases across proportions of polytomous items. It should also be noted that across all three test lengths reliability increased as the test lengthened. Looking at marginal reliability coefficients across the models, differences between the IRT and Rasch are negligible.

The second criteria analyzed for determining test optimality consisted of eleven 3 x 2 ANOVAs (one for each proportion level) conducted on the 66 sets of standard error rates with test length and model as independent variables. Main and interaction effects were evaluated for significance by comparing the observed F values to the critical F value. This critical F value is determined by the degrees of freedom for each effect and the alpha level which was set at α=.05. Each $F_{observed}$ was compared to $F_{critical\ (.05,\ 1,\ \infty)}$=3.84 for testing model significance and $F_{critical\ (.05,\ 2,\ \infty)}$=3.00 for testing significance of test length and that of interaction between model and test length. Effect size was also evaluated. Effect size is the ratio of the between-groups sum of squares and the total sum of squares, and in these analyses represents the proportion of variance in the standard error estimates explained by the difference among groups. And as reported earlier, effect size estimates were measured by partial eta squared values and were

evaluated based against Cohen's standards (small=.1, medium=.25 and large=.4).Observed power calculations were also calculated using SPSS and are included in the ANOVA tabled results.

Table 16

Marginal reliability across test length, scoring proportion and model

Items	% Poly	Rasch	IRT	Difference
	0%	0.78	0.78	0
	10%	0.80	0.81	-0.01
	20%	0.82	0.84	-0.02
	30%	0.84	0.85	-0.01
	40%	0.86	0.86	0
10	50%	0.86	0.87	-0.01
	60%	0.87	0.87	0
	70%	0.87	0.87	0
	80%	0.88	0.88	0
	90%	0.88	0.88	0
	100%	0.88	0.88	0
	0%	0.88	0.88	0
	10%	0.89	0.89	0
	20%	0.89	0.90	-0.01
	30%	0.90	0.90	0
	40%	0.90	0.91	-0.01
20	50%	0.91	0.91	0
	60%	0.91	0.92	-0.01
	70%	0.92	0.92	0
	80%	0.92	0.92	0
	90%	0.92	0.92	0
	100%	0.92	0.92	0
	0%	0.92	0.92	0
	10%	0.93	0.93	0
	20%	0.94	0.94	0
	30%	0.94	0.94	0
	40%	0.95	0.95	0
30	50%	0.95	0.95	0
	60%	0.95	0.96	-0.01
	70%	0.96	0.96	0
	80%	0.96	0.96	0
	90%	0.96	0.96	0
	100%	0.96	0.96	0

ANOVAs calculated for all eleven proportion levels for three effects resulted in 30 significant F values for model (11), test length (11) and model-by-test length interaction (8) effects with the only three exceptions: interaction effects for the 0%, 50% and 100% proportion models were not significant. The 0% and 100% proportions were the only tests analyzed that were not mixed-item format tests as the 0% proportion level was 100% dichotomously scored and the 100% proportion model was 100% polytomously scored. The significant F values from the other ANOVAs ranged from 11.019 to 950.319 for the model main effect, from 2265.213 to 32216.698 for the test length main effect, and from 4.139 to 7348.11 for the interaction effect between model and test length. Results are displayed in Tables 17 through 27.

Effect size across all proportions for model main effect ranged from .002 to .016 with two exceptions: the 60% proportion had effect size of .137 and the 90% proportion had effect size of .591. Effect size across all proportions for test length main effect ranged from .608 to .781 with the same two exceptions: the 60% proportion had effect size of .430 and the 90% proportion had effect size of .915. Effect size across all proportions for model-by-length interaction effect ranged from .000 to .012 with the same two exceptions: the 60% proportion had effect size of .160 and the 90% proportion had effect size of .710.

Power across all proportions consistently remained between .913 and 1 for model main effect and equal to 1 for all test-length main effects. For the interaction effect power ranged from .733 to 1 for eight of the proportion levels with the 0%, 50% and 100% proportion levels falling far below the rest at .375, .343 and .149, respectively.

Table 17

Two-Way Fixed Effects ANOVA on Standard Errors: Proportion 0% (Model, Test Length)

Source	Type III Sum of Squares	df	Mean Square	F	P Value	Partial Eta Squared	Non Central Parameter	Observed Power(a)
Corrected Model	28.664(b)	5	5.733	1990.358	.000	.624	9951.788	1.000
Intercept	771.970	1	771.970	268017.996	.000	.978	268017.996	1.000
MODEL	.059	1	.059	20.515	.000	.003	20.515	.995
LENGTH	28.595	2	14.297	4963.852	.000	.624	9927.703	1.000
MODEL * LENGTH	.010	2	.005	1.785	.168	.001	3.570	.375
Error	17.264	5994	.003					
Total	817.899	6000						
Corrected Total	45.929	5999						

a Computed using alpha = .05; b R Squared = .624 (Adjusted R Squared = .624)

Table 18

Two-Way Fixed Effects ANOVA on Standard Errors: Proportion 10% (Model, Test Length)

Source	Type III Sum of Squares	df	Mean Square	F	P Value	Partial Eta Squared	Non Central Parameter	Observed Power(a)
Corrected Model	27.574(b)	5	5.515	2426.215	.000	.669	12131.074	1.000
Intercept	703.077	1	703.077	309315.367	.000	.981	309315.367	1.000
MODEL	.031	1	.031	13.484	.000	.002	13.484	.957
LENGTH	27.525	2	13.762	6054.656	.000	.669	12109.312	1.000
MODEL * LENGTH	.019	2	.009	4.139	.016	.001	8.278	.733
Error	13.624	5994	.002					
Total	744.275	6000						
Corrected Total	41.198	5999						

a Computed using alpha = .05; b R Squared = .669 (Adjusted R Squared = .669)

Table 19

Two-Way Fixed Effects ANOVA on Standard Errors: Proportion 20% (Model, Test Length)

Source	Type III Sum of Squares	df	Mean Square	F	P Value	Partial Eta Squared	Non Central Parameter	Observed Power(a)
Corrected Model	23.515(b)	5	4.703	2296.209	.000	.657	11481.046	1.000
Intercept	629.415	1	629.415	307302.905	.000	.981	307302.905	1.000
MODEL	.086	1	.086	41.979	.000	.007	41.979	1.000
LENGTH	23.365	2	11.683	5703.850	.000	.656	11407.700	1.000
MODEL * LENGTH	.064	2	.032	15.684	.000	.005	31.368	.999
Error	12.277	5994	.002					
Total	665.207	6000						
Corrected Total	35.792	5999						

a Computed using alpha = .05; b R Squared = .657 (Adjusted R Squared = .657)

Table 20

Two-Way Fixed Effects ANOVA on Standard Errors: Proportion 30% (Model, Test Length)

Source	Type III Sum of Squares	df	Mean Square	F	P Value	Partial Eta Squared	Non Central Parameter	Observed Power(a)
Corrected Model	19.445(b)	5	3.889	1876.586	.000	.610	9382.930	1.000
Intercept	578.143	1	578.143	278976.181	.000	.979	278976.181	1.000
MODEL	.023	1	.023	11.019	.001	.002	11.019	.913
LENGTH	19.272	2	9.636	4649.712	.000	.608	9299.424	1.000
MODEL * LENGTH	.150	2	.075	36.244	.000	.012	72.487	1.000
Error	12.422	5994	.002					
Total	610.010	6000						
Corrected Total	31.867	5999						

a Computed using alpha = .05; b R Squared = .610 (Adjusted R Squared = .610)

Table 21

Two-Way Fixed Effects ANOVA on Standard Errors: Proportion 40% (Model, Test Length)

Source	Type III Sum of Squares	df	Mean Square	F	P Value	Partial Eta Squared	Non Central Parameter	Observed Power(a)
Corrected Model	18.955(b)	5	3.791	2000.403	.000	.625	10002.015	1.000
Intercept	543.057	1	543.057	286552.138	.000	.980	286552.138	1.000
MODEL	.042	1	.042	22.094	.000	.004	22.094	.997
LENGTH	18.892	2	9.446	4984.412	.000	.625	9968.824	1.000
MODEL * LENGTH	.021	2	.011	5.549	.004	.002	11.097	.856
Error	11.359	5994	.002					
Total	573.372	6000						
Corrected Total	30.315	5999						

a Computed using alpha = .05; b R Squared = .625 (Adjusted R Squared = .625)

Table 22

Two-Way Fixed Effects ANOVA on Standard Errors: Proportion 50% (Model, Test Length)

Source	Type III Sum of Squares	df	Mean Square	F	P Value	Partial Eta Squared	Non Central Parameter	Observed Power(a)
Corrected Model	19.671(b)	5	3.934	2481.864	.000	.674	12409.319	1.000
Intercept	529.345	1	529.345	333932.236	.000	.982	333932.236	1.000
MODEL	.101	1	.101	63.740	.000	.011	63.740	1.000
LENGTH	19.565	2	9.782	6171.178	.000	.673	12342.356	1.000
MODEL * LENGTH	.005	2	.003	1.611	.200	.001	3.223	.343
Error	9.502	5994	.002					
Total	558.518	6000						
Corrected Total	29.173	5999						

a Computed using alpha = .05; b R Squared = .674 (Adjusted R Squared = .674)

Table 23

Two-Way Fixed Effects ANOVA on Standard Errors: Proportion 60% (Model, Test Length)

Source	Type III Sum of Squares	df	Mean Square	F	P Value	Partial Eta Squared	Non Central Parameter	Observed Power(a)
Corrected Model	14.117(b)	5	2.823	1325.302	.000	.525	6626.512	1.000
Intercept	561.275	1	561.275	263455.446	.000	.978	263455.446	1.000
MODEL	2.025	1	2.025	950.319	.000	.137	950.319	1.000
LENGTH	9.652	2	4.826	2265.213	.000	.430	4530.427	1.000
MODEL * LENGTH	2.441	2	1.220	572.883	.000	.160	1145.766	1.000
Error	12.770	5994	.002					
Total	588.162	6000						
Corrected Total	26.887	5999						

a Computed using alpha = .05; b R Squared = .525 (Adjusted R Squared = .525)

Table 24

Two-Way Fixed Effects ANOVA on Standard Errors: Proportion 70% (Model, Test Length)

Source	Type III Sum of Squares	df	Mean Square	F	P Value	Partial Eta Squared	Non Central Parameter	Observed Power(a)
Corrected Model	21.211(b)	5	4.242	3402.595	.000	.739	17012.977	1.000
Intercept	484.489	1	484.489	388599.480	.000	.985	388599.480	1.000
MODEL	.058	1	.058	46.298	.000	.008	46.298	1.000
LENGTH	21.138	2	10.569	8477.007	.000	.739	16954.014	1.000
MODEL * LENGTH	.016	2	.008	6.333	.002	.002	12.666	.900
Error	7.473	5994	.001					
Total	513.174	6000						
Corrected Total	28.684	5999						

a Computed using alpha = .05; b R Squared = .739 (Adjusted R Squared = .739)

Table 25

Two-Way Fixed Effects ANOVA on Standard Errors: Proportion 80% (Model, Test Length)

Source	Type III Sum of Squares	df	Mean Square	F	P Value	Partial Eta Squared	Non Central Parameter	Observed Power(a)
Corrected Model	22.193(b)	5	4.439	3779.672	.000	.759	18898.362	1.000
Intercept	462.381	1	462.381	393732.701	.000	.985	393732.701	1.000
MODEL	.019	1	.019	16.270	.000	.003	16.270	.981
LENGTH	22.163	2	11.082	9436.311	.000	.759	18872.622	1.000
MODEL * LENGTH	.011	2	.006	4.735	.009	.002	9.471	.793
Error	7.039	5994	.001					
Total	491.614	6000						
Corrected Total	29.232	5999						

a Computed using alpha = .05; b R Squared = .759 (Adjusted R Squared = .759)

Table 26

Two-Way Fixed Effects ANOVA on Standard Errors: Proportion 90% (Model, Test Length)

Source	Type III Sum of Squares	df	Mean Square	F	P Value	Partial Eta Squared	Non Central Parameter	Observed Power(a)
Corrected Model	96.361(b)	5	19.272	17555.889	.000	.936	87779.444	1.000
Intercept	581.684	1	581.684	529881.811	.000	.989	529881.811	1.000
MODEL	9.495	1	9.495	8649.715	.000	.591	8649.715	1.000
LENGTH	70.733	2	35.366	32216.698	.000	.915	64433.396	1.000
MODEL * LENGTH	16.133	2	8.067	7348.166	.000	.710	14696.332	1.000
Error	6.580	5994	.001					
Total	684.625	6000						
Corrected Total	102.941	5999						

a Computed using alpha = .05
b R Squared = .936 (Adjusted R Squared = .936)

Table 27

Two-Way Fixed Effects ANOVA on Standard Errors: Proportion 100% (Model, Test Length)

Source	Type III Sum of Squares	df	Mean Square	F	P Value	Partial Eta Squared	Non Central Parameter	Observed Power(a)
Corrected Model	23.003(b)	5	4.601	4287.854	.000	.782	21439.270	1.000
Intercept	446.861	1	446.861	416485.310	.000	.986	416485.310	1.000
MODEL	.103	1	.103	96.287	.000	.016	96.287	1.000
LENGTH	22.898	2	11.449	10670.903	.000	.781	21341.806	1.000
MODEL * LENGTH	.001	2	.001	.588	.555	.000	1.177	.149
Error	6.431	5994	.001					
Total	476.295	6000						
Corrected Total	29.434	5999						

a Computed using alpha = .05; b R Squared = .782 (Adjusted R Squared = .781)

To answer the third research question, analysis of the first criteria indicated that the 30-item tests with higher proportions of polytomous items regardless of test length resulted in higher marginal reliability rates. No significant differences in reliability calculations existed across the IRT and Rasch models.

The second criteria, ANOVA analyses of the standard error rates, indicated that main effects for test length and model were significant for all design combinations and 8 out of 11 tests for interaction. Figures 17 and 18 display the typical relationship between the factors for most of the models. The ANOVAs indicated that mixed-item format tests tended to have standard error rates for theta that decreased as test length increased from 10 to 20 and then to 30 items. Results also show that error rates tended to be significantly different for the IRT and Rasch models at all test lengths included in this study. However, this result must be viewed in light of the fact that the effect size was negligible. Figures 19 and 20 display graphically the relationship between 60% and 90% proportion levels that differed in significance from the

others across models. Both figures indicate that the error rate for Rasch and IRT differed greatly for one test length. In both cases the IRT model performed better than Rasch in reducing standard error rates of theta estimates. Table 28 shows the effect size estimates for all proportion levels. The pattern displayed across the effect size estimates shows an increase in standard errors. This increase is due to test length and an increase in the proportion of polytomous scoring.

The third criteria analyzed for evaluating optimization was the information level across test lengths, models and proportion levels. Referring back to Tables 10, 11 and 12 it was noted that, in general, IRT produced higher overall information across all proportion levels and improved the amount of information collected as test length was increased. Also, greater amounts of information were produced as the proportion of polytomous items increased.

Figure 16. Model standard error rates across test lengths (40% proportion)

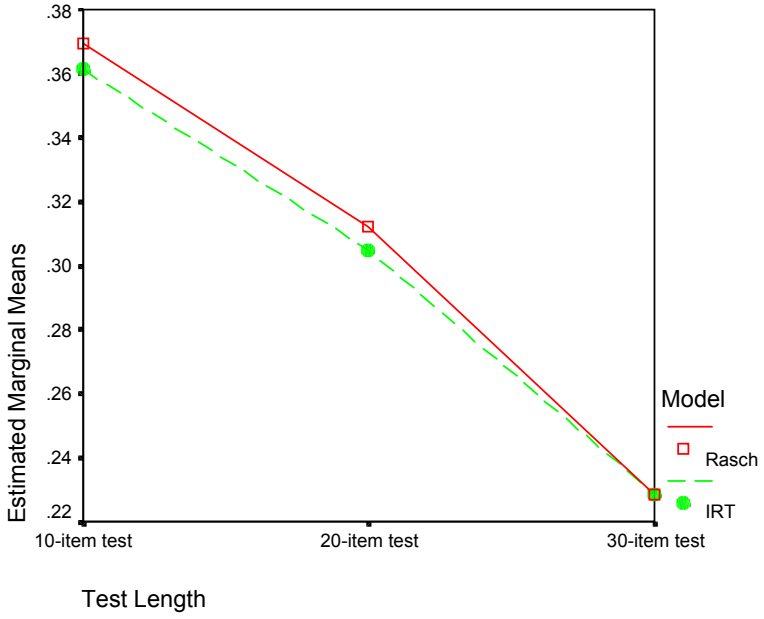


Figure 17. Test length standard error rates across models (40% proportion)

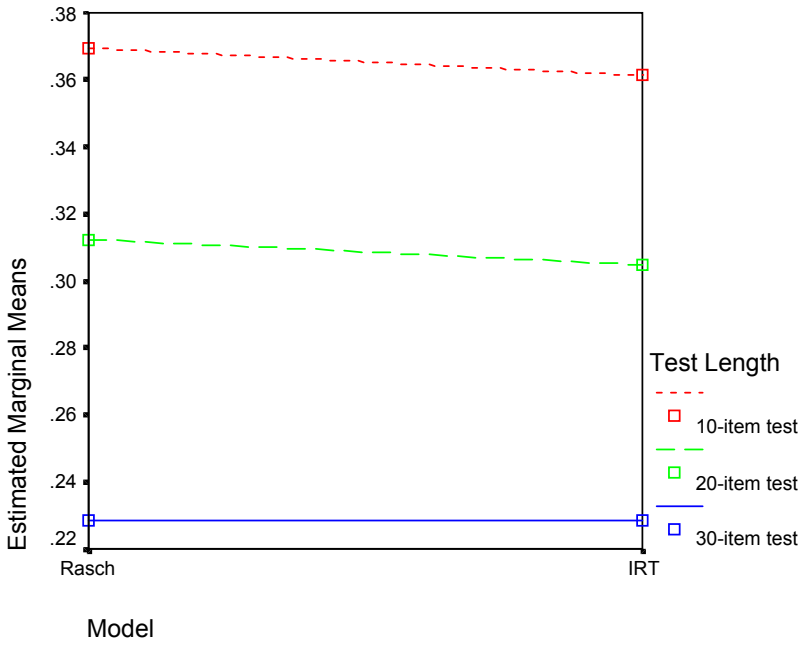


Figure 18. Model standard error rates across test lengths (60% proportion)

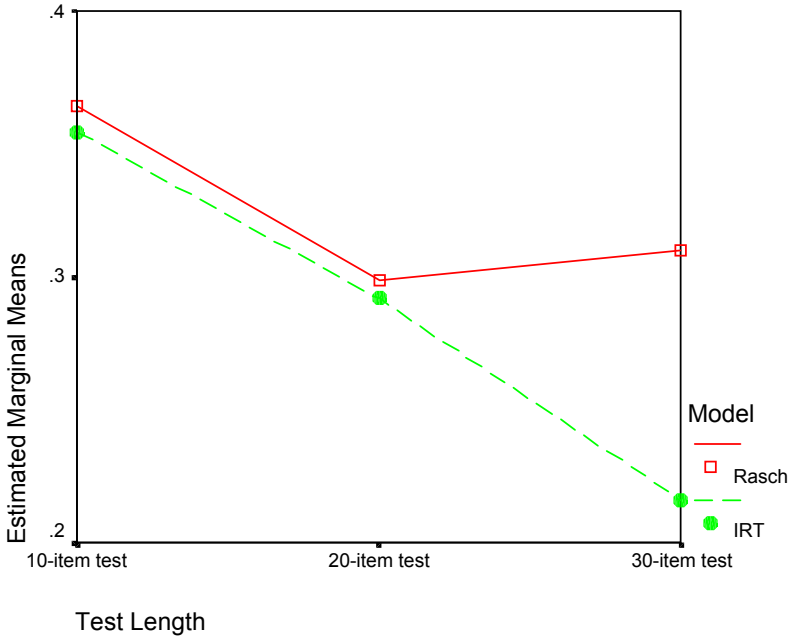


Figure 19. Model standard error rates across test lengths (90% proportion)

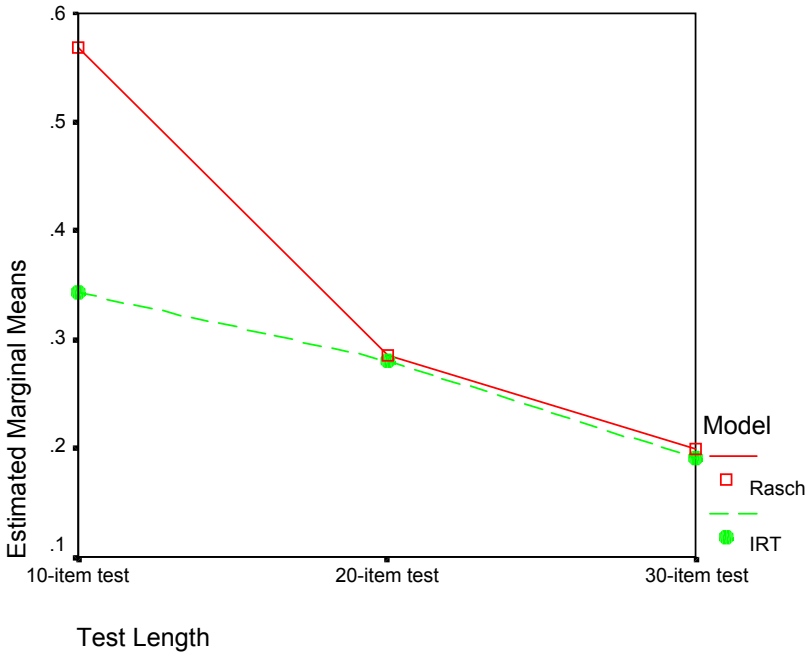


Table 28

Comparison of Effect Size Across Standard Error Results

Design Combination	Model	Length	Model by Length Interaction
0%	.003	.624	.001
10%	.002	.669	.001
20%	.007	.656	.005
30%	.002	.608	.012
40%	.004	.625	.002
50%	.011	.673	.001
60%	.137*	.430*	.160*
70%	.008	.739	.002
80%	.003	.759	.002
90%	.591*	.915*	.710*
100%	.016	.781	.000

* Effect sizes for the 60% and 90% model vary from the pattern of other output because they reflect non-normal test data distributions that fall into the category of Johnson Curves not easily analyzed in the MULTILOG software program (Thissen, 1991).

Research Question 4

How do the IRT and Rasch techniques compare in their ability to analyze mixed-item format tests?

Five criteria were evaluated to compare the IRT and Rasch techniques: assumptions for each model (fit statistics), quality of true score estimation (theta estimates and size of standard errors), reliability and information. Several of these topics were addressed in the analyses for previous research questions and are referred to here. The four general assumptions are related to the areas of dimensionality, speededness, guessing and discrimination. The theta distribution was controlled and confirmed to be normal, and since the original theta estimates were simulated, speededness and guessing are concerns beyond the scope of this study.

Discrimination parameters for the data simulation were representative of typical test data; therefore, the data was more appropriate for the IRT model which allowed discrimination to vary. The –2LL statistic, which is an estimate of the chi-squared statistic, was evaluated for model fit to the data by comparing the observed χ^2 difference to the critical χ^2. The –2LL results were compared to the null model to determine which test had the best overall fit to the data and results are displayed in Table 29. Significant differences existed between the fit of the data to the Rasch and the IRT models. Table 29 shows that across design combinations IRT procedures maintained a better fit to the data set across most of the combinations. It should be emphasized to the reader that the best fit of these data for both IRT and Rasch across all test lengths and proportions cannot be determined by using the –2LL statistic. This statistic is strictly useful when comparing analysis models when the same data set is used. Since categories are collapsed when the proportion of dichotomous and polytomous items change and test length differs across similar proportion levels, none of these –2LL values can be compared between the rows.

For quality of true ability estimation, Rasch and IRT had similar performance across 10 of the 11 proportion levels. Figures 7 to 12 graphically displays the differences in model performance across test length and proportion levels. The only significant difference in theta estimates across models existed at the 90% proportion level and it was slight. The ANOVAs performed on standard error rates indicated that mixed-item format tests tended to have standard error rates for theta that decreased as test length increased. Results also showed that error rates tended to be higher for the Rasch model at all test lengths. Figures 19 and 20 indicated that the error rate for Rasch and IRT differed greatly for one of the 3 test lengths. In both cases the IRT model performed better than Rasch in reducing standard error rates of theta estimates. No significant differences existed in marginal reliability across design combinations (see Table 16). However, information is a greater indication of optimality and therefore should be weighed more heavily than reliability. To answer the fourth research question, IRT analysis improved total information over Rasch. IRT in general seemed to collect greater overall information across theta values and IRT seemed to benefit more from increased test length with regard to information gathered.

Table 29

Negative Twice the Loglikelihood (-2LL)

Items	% Poly	Rasch -2LL	IRT -2LL	$\chi 2$ Difference	df	p	Better Fit
10	0%	-4943.5	-5017.5	74	9	0.001	Rasch
10	10%	-3573.1	-3650	76.9	9	0.001	Rasch
10	20%	-2355.4	-2423.6	68.2	9	0.001	Rasch
10	30%	-1203.4	-1247.5	44.1	9	0.001	Rasch
10	40%	11.3	-6.5	17.8	9	0.050	IRT
10	50%	714.5	677.2	37.3	9	0.001	IRT
10	60%	1213.3	1194.7	18.6	9	0.050	IRT
10	70%	1843.5	1848.1	4.6	9	0.900	Similar
10	80%	3221.1	3230.9	9.8	9	0.500	Similar
10	90%	4891.3	4875.6	15.7	9	0.100	Similar
10	100%	5438.6	5419	19.6	9	0.050	IRT
20	0%	2323.8	2156.1	167.7	19	0.001	IRT
20	10%	3853	3713	140	19	0.001	IRT
20	20%	6739	6559.3	179.7	19	0.001	IRT
20	30%	7819.1	7738	81.1	19	0.001	IRT
20	40%	9503	9339.7	163.3	19	0.001	IRT
20	50%	11640.8	11496.9	143.9	19	0.001	IRT
20	60%	14062.4	13949.6	112.8	19	0.001	IRT
20	70%	16344.9	16300.0	44.9	19	0.001	IRT
20	80%	17533.2	17482.7	50.5	19	0.001	IRT
20	90%	19560.1	19525.4	34.7	19	0.020	IRT
20	100%	21765.8	21676.3	89.5	19	0.001	IRT
30	0%	7138	6819.5	318.5	29	0.001	IRT
30	10%	11189.2	10891.7	297.5	29	0.001	IRT
30	20%	13625.3	13371	254.3	29	0.001	IRT
30	30%	16390.8	16174.5	216.3	29	0.001	IRT
30	40%	18949.5	18811.4	138.1	29	0.001	IRT
30	50%	22657.3	22459.7	197.6	29	0.001	IRT
30	60%	24504.2	24347.8	156.4	29	0.001	IRT
30	70%	27877.5	27760.9	116.6	29	0.001	IRT
30	80%	31182.9	32970.7	1787.8	29	0.001	Rasch
30	90%	32986.2	31182.9	1803.3	29	0.001	IRT
30	100%	36561.5	36465	96.5	29	0.001	IRT

CHAPTER FIVE

CONCLUSIONS AND RECOMMENDATIONS

The findings concern the impact of test length, scoring format (proportion of polytomous and dichotomous items) and analysis model (IRT and Rasch) on the ability estimates, total test information and overall optimality of tests. A comparison across models is discussed in the first part of this chapter. Some conclusions are drawn from the discussion and the practical educational importance of these findings are addressed. Recommendations for future study are made in the second part of this chapter.

Findings of the Present Study

Ability Estimates

In the present study, theta estimates were compared across IRT and Rasch analysis methods over 33 different test formats (three test lengths by eleven fixed scoring proportions of polytomous and dichotomous items). One set of original ability estimates of 1000 examinees was used to generate simulated item response data. The response set was then reconfigured into three test lengths and 11 scoring formats (including mixed formats.) IRT and Rasch techniques were then applied to each of the 33 sets of response data to generate ability estimates. These ability estimates were then compared using a fixed effects factorial ANOVA technique. Main and interaction effects were compared for analysis model and test length. Significant results were noted for tests with high proportions of polytomous items. Model and interaction effects, in general, were not significant.

The differences in ability estimates across test lengths existed when 60% to 100% of the items were scored polytomously. These differences must be viewed in light of the fact that the effect size on ability estimates was negligible and over the five proportion levels one particular

test length did not consistently result in different ability estimates. Therefore, for mixed-item format tests, no solid conclusions can be made from these analyses that ability estimates are significantly different regardless of the test length, the analysis model or the proportion of items on the test scored polytomously. This finding is interesting when considered with Si's (2002) findings that item response models vary greatly in their ability to recover original ability estimates. Si's analyses were not on mixed-item format tests but on entirely dichotomous or polytomous response sets. Interestingly, the Rasch and IRT models compared here resulted in similar ability estimates because the test scoring formats were mixed.

Test Information

This study has demonstrated that on mixed-item format tests, the test length and the proportion of items scored polytomously both have direct relationships to the amount of test information available. In general, longer tests with larger proportions of polytomous items make available more information across the distribution of ability levels. This is especially noticeable in the lower ability levels where the highest reported information was almost always at the 100% proportion level of polytomous items. For examinees with higher ability levels, highest information tends to be at proportion levels with less than 100% polytomous scoring. Highest overall test information resulted from 100% polytomous scoring across all test lengths and across both analysis models applied within this study.

It has been discussed in the literature review that polytomous items allow examinees to receive partial credit for demonstrating partial knowledge. Dichotomous items tend to ignore the existence of partial knowledge in lieu of defining all responses in an "all or nothing" context. Polytomous responses allow a continuum of measurement for ability levels and therefore provide an opportunity for gathering more information from the mid-range of ability

on each item. It seems obvious that this approach, in turn, would allow for greater information to be gathered. Awareness of this finding should serve test designers who strive to measure the true ability of examinees more accurately through the design of more effective test items. Especially for polytomous items, the costs incurred by test administration time, item writing, field testing of items and item scoring, along with the practicality of test lengths are important to consider in balancing the desire for high information levels. However, recent technological developments, such as electronic essay scoring (Shermis & Berstein, 2003; Shermis, Koch, Page, Keith & Harrington, 2002) has been investigated and shown to be equally effective if not more effective than manual scoring. This area of research may prove to lend needed efficiency to the use of polytomous scoring across a wide array of testing scenarios. Improved reliability, validity and cost per unit of information gathered are all issues that might be made to perform more optimally if stakeholders of tests and test results begin to accept automated scoring of polytomous items.

Optimality

Test optimality is dependent on maximizing both the assumptions of the statistical model and the optimality criteria for that model, which includes the fit between the model and the data. Given that the original responses were simulated, most model assumptions were designed into the simulation and meeting them was predetermined. Discrimination across the configuration of category thresholds for the polytomous items was designed to reflect actual test data and therefore, the constant discrimination required by the Rasch model was not met through the simulation. The fit statistic analyzed (-2LL) determined that the data fit both the IRT and Rasch model best on the 10-item test with 40% polytomous scoring. IRT displayed a better fit to most of the data sets when compared to Rasch.

Standard error rates of the ability estimates were analyzed under optimality criteria and significant results were obtained across test length with large effect sizes. Longer test lengths were found to produce smaller standard errors across all proportion levels. Impacts on standard errors were also noted from model effects and model-by-test length interaction but effect sizes for both were small. With few exceptions, standard error rates differed across all test lengths and both models. Standard errors decreased as test length increased displaying an indirect or inverse relationship. For these comparisons effect size of test length on standard error increased as the proportion of polytomous items increased. This was true for all three test lengths and high power values were also present.

Reliability results were found to increase across test lengths and proportion levels. The 30-item test at 100% polytomous scoring had the highest reliability rates across both models. This finding aligns with the research by Lau and Wang (1998) who found that the combination of two types of items on one assessment enhances the reliability of the assessment (Lau & Wang, 1998). It should be noted that information levels provided more significant measures of the optimality of a mixed-item format test as mentioned in the literature review and was discussed in this chapter under the previous section.

Comparison of IRT and Rasch

Overall, significant differences were not noted between the performance of IRT and Rasch on the analyses in this study. Both models effectively generated ability estimates similar to the original simulated values. Generally across the analyses of information in this study, the use of IRT analysis on a mixed-item format tests did slightly improved the available test information over Rasch techniques across the range of abilities for most test lengths and proportion levels of polytomous scoring. IRT also seemed to benefit more from increases in test

length with regard to information gathered. Differences in reliability and fit statistics between the IRT and Rasch models are negligible. Standard error rates tended to be higher for the Rasch model at all test lengths analyzed in this study. Berger (1998) noted that discrimination across categories is a complicated issue and the differences noted in this study support that conclusion. The Rasch model requires a constant discrimination parameter which was not met by the simulated data. Also, the guessing parameter was set to zero across both models and in actual response data it is likely that high ability examinees have a higher probability of guessing correctly on multiple choice items.

Recommendations

No known superior analysis method exists for analyzing the mixed-item format test. This study investigated IRT and Rasch methods and how they compared across three test lengths and an array of mixed-item formats. The results and conclusions of this study were based on analyses conducted on simulated data. There are conveniences and limitations involved in the use of simulated data and the possibility of generalizing these results to actual mixed-item format test data is encouraged by this researcher and by Li, Lissitz & Yang (1999).

The analyses of this study were sufficient in producing results that addressed the research questions posed here, however, further analysis is needed in the following areas:

- looking at test difficulty across a variety of mixed scoring options,
- looking at examinee test scores across a variety of mixed scoring options,
- developing more appropriate ways of analyzing reliability and fit of statistical models to mixed-item format tests,
- testing the robustness of the models when assumptions are violated (speededness and guessing) in a mixed-item format test,
- pursuing the use of automated scoring of polytomous items.
- investigate the ideal mix of dichotomous and polytomous items that yield reliable and valid test scores.

APPENDIX

APPENDIX A

Example of a Mixed-Item Format Test
(70% Dichotomous and 30% Polytomous)

Directions: Choose the option that best answers the question or completes the statement.

1. To produce viable research results, what type of procedures must be used?
   a) deductive
   b) inductive
   c) systematic
   d) ethical

2. Which of these characteristics is *not* a requirement of educational research?
   a) purposeful ethical enterprise
   b) systematic processes used to obtain outcomes
   c) seeking solutions to education problems
   d) publication of findings

3. Which of the following can be considered as variable(s)?
   a) gender
   b) favorite color
   c) age
   d) a, b, and c

4. Which characteristic describes *applied* research?
   a) advances fundamental knowledge and theory
   b) provides insights into societal concerns
   c) solves problems in particular settings
   d) benefits the educational community generally

5. Which characteristic describes *action* research?
   a) advances fundamental knowledge and theory
   b) provide insights into societal concerns
   c) solve problems in particular settings
   d) benefits the educational community generally

Directions: mark each statement as T or F for false.

6. T F Although qualitative and quantitative research methods are based on different philosophical assumptions, they can be combined as complementary processes.

7. T F Devising a research plan is explicit within the systematic processes of research.

Directions: Supply an answer to each question.

8. Explain why most researchers devise a research plan.

9. How does the literature review of a report benefit the readers?

10. Why is it important for qualitative researchers to use purposeful sampling procedures?

Adopted and modified from Educational Research: An Integrative Introduction (Instructor's Manual and Test Bank) by Evelyn J. Sowell and Tari L. Kinsey (2001) McGraw-Hill Higher Education, New York: NY.

APPENDIX B

SAS Program for Data Simulation

```
/*********************************************************************/
/*  This program is to generate response patterns                    */
/*  based on Muraki's generalized partial credit model.              */
/*  A normal distribution is created for two scoring schema.         */
/*  The possible responses for item i are 1 to ncat for              */
/*  polytomous responses and binary for dichotomous responses.       */
/*********************************************************************/

options ps=52 ls=72;
%let ne=1000;                    /*no. of examinees*/
%let ni=30;                     /*no. of items*/
%let thlist=sd(i,1) sd(i,2) sd(i,3);   /*threshold input format*/
%let ncb=3;                     /*no. of category boundries*/
%let ncat=4;                    /*no. of categories*/
%let ipmfl='c:\multilog\working\iparm.dat';  /* item param input path*/
%let ipmfmt=a(i) 8.2 (&thlist) (&ncb*8.2);  /*item param input format*/
%let outfl='c:\multilog\working\workingbothout1.dat';     /*gen data */
%let seed1=6734;                /* seed number for normal dist.. */
%let seed3=3422;                        /*seed number for r*/
%let putfmt=@1 id 4.0 @6 z 7.3 @14 (r1-r&ni) (&ni*1.0) @45 (dr1-dr&ni) (&ni*1.0);;
/*********************************************************************/
data know (keep=z);
    do i=1 to &ne;
      z=rannor(&seed1);
      output;
    end;

filename ipm &ipmfl;          /* read in the parameter input file */
data iparm;
    array sd(&ni, &ncb);
    array a(&ni);
    do i=1 to &ni;
        infile ipm;
        input &ipmfmt;
    end;

filename resp &outfl;
data respdat (keep= z r1-r&ni dr1-dr&ni e1 den pbc1-pbc&ncat r tot);
    if _n_=1 then set iparm; set know;
    array sd(&ni, &ncb);
    array a(&ni);
```

```
    array pbc(&ncat);
    array rr(*) r1-r&ni;
    array drr(*) dr1-dr&ni;
       do i=1 to &ni;
         rr(i)=1.0;
         drr(i)=0.0;
       end;
       do i=1 to &ni;
         den=1.0;
         do j=1 to &ncb;
           e1=0.0;
           do k=1 to j;
             e1=e1+a(i)*(z-sd(i,k));

           end;
           pbc(j+1)=exp(e1);
           den=den+exp(e1);
         end;
         pbc(1)=1/den;
         do j=2 to &ncat;
            pbc(j)=pbc(j)/den;
         end;
         tot=0.0;
         r=ranuni(&seed3);
         do k=1 to &ncat;
           tot=tot+pbc(k);
           if r>tot then rr(i)=k+1;
           if rr(i)>3 then drr(i)=1;
        end;

       end;
       id=_n_;
       file resp;
       put &putfmt;
proc univariate normal vardef=n; var z;
proc chart;
  hbar z;

run;
```

Adopted and modified from Chen (1996).

APPENDIX C

Item Parameters used in Data Simulation

Item	a_i	b_{i1}	b_{i2}	b_{i3}
1	1.60	-2.25	-1.75	-1.25
2	.80	-2.25	-1.55	-.85
3	2.00	-2.20	-1.75	-1.30
4	1.50	-2.00	-1.60	-1.20
5	1.70	-1.65	-1.20	-.75
6	1.80	-2.10	-1.60	-1.10
7	1.20	-1.90	-1.30	-.70
8	.90	-1.55	-1.10	-.75
9	1.70	-1.80	-1.40	-1.00
10	1.20	-2.05	-1.50	-.95
11	.90	-.65	.00	.65
12	1.90	-.60	.10	.80
13	1.40	-.55	.00	.55
14	1.80	-.45	-.05	.35
15	1.30	-.50	.00	.50
16	1.60	-.50	.05	.60
17	1.10	-.45	.00	.45
18	1.80	-.40	.00	.40
19	1.20	-.30	.20	.70
20	1.80	-.35	.00	.35
21	1.70	1.30	1.75	2.20
22	1.30	.85	1.55	2.25
23	.90	.50	.90	1.30
24	1.00	1.15	1.65	2.15
25	1.60	.80	1.40	2.00
26	1.60	1.00	1.45	1.90
27	1.90	.75	1.20	1.65
28	1.10	1.05	1.50	1.95
29	2.00	.75	1.30	1.85
30	1.80	.95	1.50	2.05

Adopted from Si (2002).

APPENDIX D

SAS simulated Thetas and Item Responses

Person	Theta	30 item responses
1	-0.984	311143112133424121311114132431
2	0.333	411344134334444143422224143331
3	1.734	444444344444434334432334244443
4	-1.118	111443111214313111411114141211
5	-0.088	411244223433424124411114144133
6	-0.058	211443133434424132411124144431
7	-0.656	422332112211414143311224141121
8	-0.338	421444112311323132211114232111
9	0.724	412444134234434142411124143232
10	-0.357	411342111334413122411113142321
11	-0.981	411443111122414121411113121221
12	0.682	421444234343434142411114244321
13	0.167	311444143414414132412114142243
14	-0.491	111433123424323121411114142111
15	0.759	433434244344444222411114144442
16	-0.679	411433131211313122411113141121
17	-0.233	412344233322434133411213141321
18	-0.718	422244112214414113311213143131
19	-0.839	321342111113314111311114141111
20	0.326	311434334244444244411224144331
21	1.367	413444244443444144433324344443
22	1.767	442444444444444344443414344444
23	0.275	411433232324424142413114144442
24	0.895	432443133344444234421134144421
25	0.547	411344143134444122412214144231
26	-0.281	421424222333424111411124141321
27	-0.590	441143113321214131411114132121
28	0.179	431434134314344142421214143331
29	0.230	431344122434414142412114144331
30	-0.788	411424131133414111412214142111
31	0.141	311444143134434122411114143422
32	-0.288	321433121144424131411114142332
33	-0.745	312343122224313111211112142311
34	-0.476	321333113134424112411214143242
35	-0.996	332433213212414121211113141111

(continued to 1000 persons)

APPENDIX E

Response Frequencies and Proportions for 30 Items

Items	POLYTOMOUS								DICHOTOMOUS			
	FREQUENCY				PROPORTION				FREQUENCY		PROPORTION	
	1	2	3	4	1	2	3	4	0	1	0	1
1	50	90	187	673	.05	.09	.19	.67	327	673	.33	.67
2	457	253	176	114	.46	.25	.18	.11	886	114	.89	.11
3	659	224	84	33	.66	.22	.08	.03	967	33	.97	.03
4	54	102	246	598	.05	.10	.25	.60	402	598	.40	.60
5	35	86	216	663	.03	.09	.22	.66	337	663	.34	.66
6	46	114	258	582	.05	.11	.26	.58	418	582	.42	.58
7	625	198	110	67	.62	.20	.11	.07	933	67	.93	.07
8	279	192	209	320	.28	.19	.21	.32	680	320	.68	.32
9	275	179	229	317	.28	.18	.23	.32	683	317	.68	.32
10	242	238	233	287	.24	.24	.23	.29	713	287	.71	.29
11	251	232	235	282	.25	.23	.24	.28	718	282	.72	.28
12	76	151	234	539	.08	.15	.23	.54	461	539	.46	.54
13	23	50	144	783	.02	.05	.14	.79	217	783	.21	.79
14	269	255	242	234	.27	.25	.24	.23	766	234	.77	.23
15	26	50	158	766	.03	.05	.16	.77	234	766	.23	.77
16	790	183	50	22	.79	.14	.05	.02	978	22	.98	.02
17	237	221	256	286	.24	.22	.26	.29	714	286	.71	.29
18	289	249	241	221	.29	.25	.24	.22	779	221	.78	.22
19	47	56	153	744	.05	.06	.15	.74	256	744	.26	.74
20	698	184	77	41	.70	.18	.08	.04	959	41	.96	.04
21	696	206	67	41	.69	.21	.07	.04	959	41	.96	.04
22	649	213	87	51	.65	.21	.09	.05	949	51	.95	.05
23	643	240	86	31	.64	.24	.09	.03	969	31	.97	.03
24	21	44	139	796	.02	.04	.14	.80	204	796	.20	.80
25	653	207	85	55	.65	.21	.09	.06	945	55	.94	.06
26	20	33	119	828	.02	.03	.12	.83	172	828	.17	.83
27	277	192	199	332	.27	.19	.20	.33	668	332	.67	.33
28	250	243	254	253	.25	.24	.25	.25	747	253	.75	.25
29	250	219	245	286	.25	.22	.24	.29	714	286	.71	.29
30	634	223	98	45	.63	.22	.10	.05	955	45	.95	.05

APPENDIX F

MULTILOG Programs for Data Analysis

FILE R003

```
>PRO RA IN NI=30 NG=1 NE=1000 NCHARS=13;
>TEST ALL L1;
>ESTIMATE NCYCLES=100;
>SAVE;
>END;
        2
01
1111111111111111111111111111111
N
(13A1,3X,30A1)

>PRO SCORE RA IN NI=30 NG=1 NE=1000 NCHARS=13;
>TEST ALL L1;
>ESTIMATE NCYCLES=100;
>START ALL;
Y

>END;
        2
01
1111111111111111111111111111111
N
(13A1,3X,30A1)
```

FILE R1003

```
>PRO RA IN NI=30 NG=1 NE=1000 NCHARS=12;
>TEST ALL NO NC=(4(0)30) HI=(4(0)30);
>TMATRIX ALL AK POLY;
>EQUAL ALL AK=1;
>FIX ALL AK=(2,3) VALUE=0.0;
>TMATRIX ALL CK TRIANGLE;
>ESTIMATE NCYCLES=100;
>SAVE;
>END;
        4
1234
```

```
111111111111111111111111111111
222222222222222222222222222222
333333333333333333333333333333
444444444444444444444444444444
(12A1,3X,30A1)


>PRO SCORE RA IN NI=30 NG=1 NE=1000 NCHARS=12;
>TEST ALL NO NC=(4(0)30) HI=(4(0)30);
>TMATRIX ALL AK POLY;
>EQUAL ALL AK=1;
>FIX ALL AK=(2,3) VALUE=0.0;
>TMATRIX ALL CK TRIANGLE;
>ESTIMATE NCYCLES=100;
>START ALL;
Y

>END;
        4
1234
111111111111111111111111111111
222222222222222222222222222222
333333333333333333333333333333
444444444444444444444444444444
(12A1,3X,30A1)
```

FILE R503

```
 CALIBRATION PARTIAL CREDIT MODEL for mixed 50/50
>PRO RA IN NI=30 NG=1 NE=1000 NCHARS=11;
>TEST IT=(1(1)15) L1;
>TEST IT=(16(1)30) NO NC=(4(0)30) HI=(4(0)30);
>TMATRIX IT=(16(1)30) AK POLY;
>EQUAL IT=(16(1)30) AK=1;
>FIX IT=(16(1)30) AK=(2,3) VALUE=0.0;
>TMATRIX IT=(16(1)30) CK TRIANGLE;
>ESTIMATE NCYCLES=100;
>SAVE;
>END;
        5
01234
111111111111110000000000000000
222222222222221111111111111111
000000000000000222222222222222
000000000000000333333333333333
```

```
00000000000000044444444444444
(11A1,11X,30A1)


>PRO SCORE RA IN NI=30 NG=1 NE=1000 NCHARS=11;
>TEST IT=(1(1)15) L1;
>TEST IT=(16(1)30) NO NC=(4(0)30) HI=(4(0)30);
>TMATRIX IT=(16(1)30) AK POLY;
>EQUAL IT=(16(1)30) AK=1;
>FIX IT=(16(1)30) AK=(2,3) VALUE=0.0;
>TMATRIX IT=(16(1)30) CK TRIANGLE;
>ESTIMATE NCYCLES=100;
>START ALL;
Y

>END;
        5
01234
11111111111111000000000000000
22222222222222111111111111111
00000000000000222222222222222
00000000000000333333333333333
00000000000000444444444444444
(11A1,11X,30A1)
```

FILE Q003

```
CALIBRATION IRT 0% POLY
>PRO RA IN NI=30 NG=1 NE=1000 NCHARS=13;
>TEST ALL L3;
>ESTIMATE NCYCLES=100;
>SAVE;
>END;
        2
01
1111111111111111111111111111
N
(13A1,3X,30A1)
```

FILE Q003

```
SCORING IRT 0% POLY
>PRO SCORE RA IN NI=30 NG=1 NE=1000 NCHARS=13;
>TEST ALL L3;
>ESTIMATE NCYCLES=100;
```

```
>START ALL;
Y

>END;
        2
01
11111111111111111111111111111
N
(13A1,3X,30A1)
```

FILE Q503

```
CALIBRATION RUN IRT 50% POLY
>PRO RA IN NI=30 NG=1 NE=1000 NCHARS=11;
>TEST IT=(1(1)15)L3;
>TEST IT=(16(1)30) GRaded NC=(4(0)15);
>ESTIMATE NCYCLES=100;
>SAVE;
>END;

        5
01234
11111111111111111111111111111
22222222222222211111111111111
00000000000000022222222222222
00000000000000033333333333333
00000000000000044444444444444
(11A1,11X,30A1)
```

FILE Q503

```
>PRO SCORE RA IN NI=30 NG=1 NE=1000 NCHARS=11;
>TEST IT=(1(1)15)L3;
>TEST IT=(16(1)30) GRaded NC=(4(0)15);
>ESTIMATE NCYCLES=100;
>START ALL;
Y

>END;
        5
01234
11111111111111111111111111111
22222222222222211111111111111
00000000000000022222222222222
```

```
000000000000000033333333333333
000000000000000044444444444444
(11A1,11X,30A1)
```

FILE Q1003

```
> RA IN NI=30 NG=1 NE=1000 NCHARS=12;
>TEST ALL GRaded NC=(4(0)30);
>ESTIMATE NCYCLES=100;
>SAVE;
>END;

        4
1234
111111111111111111111111111111
222222222222222222222222222222
333333333333333333333333333333
444444444444444444444444444444
(12A1,3X,30A1)

Q10032A
>PRO SCORE RA IN NI=30 NG=1 NE=1000 NCHARS=12;
>TEST ALL GRaded NC=(4(0)30);
>ESTIMATE NCYCLES=100;
>START ALL;
Y

>END;
        4
1234
111111111111111111111111111111
222222222222222222222222222222
333333333333333333333333333333
444444444444444444444444444444
(12A1,3X,30A1)
```

REFERENCES

Allen, M. J., & Yen, W. M. (1979). Introduction to measurement theory. Monterey, CA: Brooks/Cole Publishing Company.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). Standards for educational and psychological testing. Washington, D.C: Author.

Berger, M. P. (1998). Optimal design of tests with dichotomous and polytomous items. Applied Psychological Measurement, 22(3) 248-258.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In Lord, F. M. & Novick, M. R. (Eds.), Statistical theories of mental test scores, (pp. 395-479). Reading, MA: Addison-Wesley Publishing Company, Inc.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. Psychometrika, 37(1) 29-51.

Camilli, G. (1994). Origin of the scaling constant d=1.7 in item response theory. Journal of Educational and Behavioral Statistics, 19(3) 293-295.

Carlson, J. E. (1996, April). Information provided by polytomous and dichotomous items on certain NAEP instruments. Paper presented at the annual meeting of the American Educational Research Association, New York, NY.

Chen, S. (1996). A comparison of maximum likelihood estimation and expected a posteriori estimation in computerized adaptive testing using the generalized partial credit model. Unpublished doctoral dissertation, University of Texas at Austin.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Crocker, L. & Algina, J. 1986). Introduction to Classical & Modern Test Theory. Belmont, CA: Wadsworth Group.

Donoghue, J. R. (1994). An empirical examination of the IRT information of polytomously scored reading items under the generalized partial credit model. Journal of Educational Measurement, 31(4) 295-311.

Embretson, S.E. & Hershberger, S. (1999). The New Rules of Measurement. Mahwah, NJ: Lawrence Erlbaum

Embretson, S. E. & Reise, S. P. (2000). Psychometric Methods: Item response theory for psychologists. Mahwah, NJ: Lawrence Erlbaum.

Ercikan, K., Schwarz, R. D., Julian, M. W., Burket, G. R., Weber, M. M. & Link, V. (1998). Calibration and scoring of tests with multiple-choice and constructed-response item types. Journal of Educational Measurement, 35(2), 137-154.

Glas, C. A. W. & Verhelst, N. D.(1989). Extensions of the Partial Credit Model. Psychometrika; 54(4), 635-659.

Hambleton, R. K. (1994). The rise and fall of criterion-referenced measurement? Educational Measurement: Issues and Practice, 13(4) 21-26.

Hambleton, R. K. (1989). Principles and selected applications of item response theory. R.L. Linn (Ed.), Educational Measurement: Third Edition, (pp. 147-200). American Council on Education and Macmillan Publishing Company.

Hambleton, R. K. & Murphy E. (1992). A psychometric perspective on authentic measurement. Applied measurement in Education, 5(1), 1-16.

Hambleton, R. K., Robin, F. & Xing, D. (2000). Item response models for the analysis of educational and psychological test data. In H. E. A. Tinsley & S. D. Brown (Eds.), Handbook of applied multivariate statistics and mathematical modeling (pp. 553-581). San Diego, CA: Academic Press.

Hambleton, R. K. & Swaminathan, H. (1985). Item response theory: Principles and applications. Norwell, MN: Kluwer-Nijhoff Publishing.

Hambleton, R. K., Swaminathan, H., Rogers, H. J. (1991). MMSS: Fundementals of item response theory. (Vol. 2). Newbury Park, CA: Sage.

Lau, C. A. & Wang, T. (1998, April). Comparing and combining dichotomous and polytomous items with SPRT procedure in computerized classification testing. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Li, Y. H., Lissitz, R. W., & Yang, Y. N. (1999) Estimating IRT equating coefficients for tests with polytomously and dichotomously scored items. (ERIC Document Reproduction Service No. ED 431 800).

Linden, W. J. van der (1998). Optimal assembly of psychological and educational tests. Applied Psychological measurement, 22(3) 195-211.

Linn, R.L. (1990). Has item response theory increased the validity of achievement test scores? Applied Measurement in Education, 3(2), 115-141.

Lord, F. M. (1980). Applications of Item Response Theory to Practical Testing Problems. Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers, Inc.

Masers, G.N. (1982). A rasch model for partial credit scoring. Psychometrika 47(2), 149-174.

Masters, G. N. & Wright B. D. (1997). The partial credit model. In W. J. van der Linden & R. K. Hambleton (Eds.), Handbook of modern item response theory (pp. 101-121). New York, NY: Springer-Verlag.

Muraki, E. & Bock R. D. (1996). PARSCALE: IRT based test scoring and item analysis for graded open-ended exercises and performance tasks (third ed.). Chicago, IL: Scientific Software International.

Plake, B. (1993). Applications of educational measurement: Is optimum optimal? Educational Measurment: Issues and Practices, 5-10.

Rasch, G (1977). On specific objectivity an attempt at formalizing the request for generality and validity of scientific staatements. The Danish Yearbook of Philosophy. Copenhagan: Munksgaard.

Roskam, E. E. & Jansen, P. G. W. (1989). Conditions for rasch-dichotomizability of the unidimensional polytomous rasch model. Psychometrika, 54(2) 317-332.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. Psychometrika Monograph Supplement No. 17, Vol. 34 (4, Pt. 2).

Shermis, M. D. & Burstein, (2003). Automated essay scoring: A cross disciplinary perspective. Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.

Shermis, M. D., Koch, C. M., Page, E. B., Keith, T. Z. & Harrington, S. (2002). Trait ratings for automated essay grading. Educational and Psychological Measurement, 62(1) 5-18.

Si, C. B. (2002). Ability estimation under different item parameterization and scoring models. Unpublished doctoral dissertation, University of North Texas.

Sowell, E. J. & Kinsey, T. L. (2001). Educational Research: An Integrative Introduction- Instructor's Manual and Test Bank: McGraw-Hill Higher Education, New York: NY.

Sykes, R. C. & Yen, W. M. (2000). The scaling of mixed-item format tests with the one-parameter and two-parameter partial credit models. Journal of Educational Measurement, 37(3), 221-244.

Thissen, D. M. (1976). Information in wrong responses to the raven progressive matrices. Journal of Educational Measurement, 13(3), 201-214.

Thissen, D. (1991). MULTILOG User's Guide: Multiple, Categorical item Analysis and Test Scoring using item Response Theory; version 6.0. Scientific Software International Chicago: Illinois.

Thissen, D., Nelson, L. and Swygert, K. (2001). Item response theory applied to combinations of multiple-choice and constructed-response items- Approximation method for scaled scores. In Thissen, D. & Wainer, H. (Eds.), Test Scoring, (pp. 293-341). Mahwau, New Jersey: Lawrence Erlbaum Associates, Inc.

Wainer, H. & Thissen, D. (2001). True score theory: The traditional method. In Thissen, D. & Wainer, H. (Eds.), Test Scoring, (pp. 253-292). Mahwau, New Jersey: Lawrence Erlbaum Associates, Inc.

Wainer, H. & Thissen, D. (1993). Combining multiple-choice and constructed-response test scores: toward a marxist theory of test construction. Applied Measurement in Education, 6(2), 103-118.

Wang L. L. & Lee, C. S., (1998). Cognitive-psychometric modeling in computer adaptive testing. Paper presented at the Annual Conference of the American Psychological Association, San Francisco, CA.

Wright, B. D. & Masters, G. N. (1982). Rating Scale Analysis. Chicago, IL: Mesa Press.

Wright, B. D., & Stone, M. H. (1979). Best test design: Rasch measurement. Chicago, IL: Mesa Press.