2004-12-08

The Strength of Multidimensional Item Response Theory in Exploring Construct Space that is Multidimensional and Correlated

Steven Gerry Spencer
Brigham Young University - Provo

THE STRENGTH OF MULTIDIMENSIONAL ITEM RESPONSE

THEORY IN EXPLORING CONSTRUCT SPACE THAT IS

MULTIDIMENSIONAL AND CORRELATED

by

Steven G. Spencer

A dissertation submitted to the faculty of

Brigham Young University

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Instructional Psychology and Technology

November 19, 2004

BRIGHAM YOUNG UNIVERSITY

GRADUATE COMMITTEE APPROVAL

Of a dissertation submitted by

Steven G. Spencer

This dissertation has been read by each member of the following graduate committee and by majority vote has been found to be satisfactory.

Date	C. Victor Bunderson, Chair
Date	Andy Gibbons
Date	Richard Sudweeks
Date	Stephen Yanchar
Date	Joseph Olsen

BRIGHAM YOUNG UNIVERSITY

As chair of the candidate's graduate committee, I have read the dissertation of Steven G. Spencer in its final form and have found that (1) its format, citations, and bibliographical style are consistent and acceptable and fulfill university and department style requirements; (2) its illustrative materials including figures, tables, and charts are in place; and (3) the final manuscript is satisfactory to the graduate committee and is ready for submission to the university library.

Date

C. Victor Bunderson
Chair, Graduate Committee

Accepted for the Department

Andy Gibbons
Department Chair

Accepted for the College

Richard Young
Dean, College of Education

ABSTRACT

THE STRENGTH OF MULTIDIMENSIONAL ITEM RESPONSE THEORY IN EXPLORING CONSTRUCT SPACE THAT IS MULTIDIMENSIONAL AND CORRELATED

Steven G. Spencer

Department of Instructional Psychology and Technology

Doctor of Philosophy

This dissertation compares the parameter estimates obtained from two item response theory (IRT) models: the 1-PL IRT model and the MC1-PL IRT model. Several scenarios were explored in which both unidimensional and multidimensional item-level and personal-level data were used to generate the item responses. The Monte Carlo simulations mirrored the real-life application of the two correlated dimensions of Necessary Operations and Calculations in the basic mathematics domain. In all scenarios, the MC1-PL IRT model showed greater precision in the recovery of the true underlying item difficulty values and person theta values along each primary dimension as well as along a second general order factor. The fit statistics that are generally applied to the 1-PL IRT model were not sensitive to the multidimensional item-level structure, reinforcing the requisite assumption of unidimensionality when applying the 1-PL IRT model.

ACKNOWLEDGEMENTS

The art of scholarly inquiry is an never-ending journey of discovery and adventure. The path leads to many dead ends as well as to many forks that invite the traveler to further exploration.

As the investigation yields answers, new questions emerge eliciting a new quest to begin. Barely there is time to conclude one search yet another begins. Still, time for reflection is requisite. Both to contemplate the accomplishments made as well as to consider those who made this journey possible.

Dr. C. Vic Bunderson a mentor, colleague, and friend. His guidance and disciplined counsel honed by his own experienced research have given breadth and depth to this project. Dr. Richard Sudweeks whose drive for precision and exactness in every detail gave greater clarity to the dissertation. Dr. Andy Gibbons who years ago demonstrated the process of turning theory into application. Dr. Steve Yanchar whose broadened perspectives made me look beyond this project to the greater application of the theoretical. Dr. Joe Olsen who demonstrated that statistics serves little until applied in an environment to improve humanity.

Ric Luecht, Terry Ackerman, and Mark Wilson for their brief comments as to the theoretical nuances of multidimensionality and the appropriate interpretations of the multidimensional structure of assessments.

To Nona, my dear wife, companion, and friend: this dissertation could not have been complete without your endearing support and enduring patience through all the lonely and frustrating times. Thank you all.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF EQUATIONS

CHAPTER 1

INTRODUCTION

One of the major developments in psychological measurement during the last century is item response theory (IRT). One of item response theory's major advantages over previous measurement theories is the ordered placement of item difficulty values on the same measurement scale as student ability levels, thus facilitating the creation of custom-tailored assessments to meet the unique requirements of individual students. Thus, a new set or subset of items can be added to the item pool without changing the relative ordering of items or persons along the measurement scale.

IRT requires the investigation of several assumptions prior to the application of a particular IRT model to a given data set. Violation of these assumptions results in an improperly applied measurement model and erroneously derived inferences regarding the assessment results.

One of the most important assumptions upon which IRT rests is the assumption of a unidimensional latent trait. Unidimensionality requires that all items within a test measure one specific ability or proficiency (Hambleton, Swaminathan, & Rogers, 1991). This unidimensionality assumption is problematic in that although assessments are intended to measure only one trait or skill, the very nature of statistical testing often introduces multidimensional elements into the measurement process. Although an assumption of IRT, the attainment of unidimensional data is too often the exception rather than the rule (Traub 1983).

Multidimensional item response theory (MIRT) is an extension of unidimensional item response theory. MIRT relaxes the assumption of unidimensionality and allows for the intentional inclusion of items that span multiple abilities or proficiencies.

Statement of Problem

The use of IRT to assess multiple construct-relevant dimensions within the content domain violates not only the statistical assumptions of unidimensionality required by the IRT models, but also the structural aspect of Messick's (1995) construct validity argument.

Item response theory's strength lies in its ability to more accurately estimate the true unidimensional construct structure. The presence of construct-relevant multidimensionality could diminish this strength. Knowing how much IRT's capacity to investigate the true construct structure is diminished and how to recover this construct structure is important to measurement practitioners who use multidimensional data.

Statement of Purpose

This project has two main purposes. The primary purposes are to estimate the accuracy of IRT and MIRT estimation programs when the assumption of unidimensionality is violated and to what degree the misfit would be when a unidimensional model is applied to multidimensional data. The secondary purpose is to determine the degree to which a multidimensional IRT model can recover the underlying construct relevant multidimensional structure within an educational domain.

Audience

The audience for this study are psychometricians who utilize item response theory. They are familiar with the appropriate application of unidimensional IRT, and would like to explore further into multidimensional IRT. A secondary audience are those who are familiar with basic psychometric concepts and who wish to utilize item response theory to improve their assessment instruments. These individuals know enough in general to apply the theory, but may be unfamiliar when the theoretical applications are appropriate or inappropriate.

Research Questions

This project focuses on answering the following four questions:

1. Given unidimensional item-level data and multidimensional person-level data, does the multi-dimensional compensatory one-parameter logistic (MC1-PL) model recover the true generating item and person parameters any more accurately than the one-parameter logistic item response theory (1-PL IRT) or Rasch model?
2. Given simulated data having construct-relevant multidimensionality, how closely can the MC1-PL model recover the true generating values of the items on those multiple dimensions?
3. By applying the Rasch model for calibration of these multidimensional items to obtain a single summary scale, will the resultant model show increasing misfit for those items that lie further from the intersection of the two dimensions than those items that fall closer to the dimensional intersection?
4. By applying the 2-PL IRT model to these multidimensional items, will the value of the discrimination parameter increase for items that lie off the second factor when calibrated one at a time onto the second factor?

Scope

For the first research question, 21 unidimensional items were used. For the purpose of the second and third research questions, a total of 21 items were used on both the primary and the composite dimensions. Seven items were placed on each of the three dimensions. For the fourth research question, each of the seven items on one of the primary dimensions were projected one at a time onto the other primary dimension.

All projections were orthogonal to the target dimension.

Assumptions

This study is based on the following assumptions:

1. For all items, negligible guessing is assumed.
2. The data modeled follows the properties of a mathematics test that is known to have the two primary construct-relevant dimensions of Necessary Operations (NO) and Calculations (C) as well as a composite dimension. Necessary Operations refers to the appropriate selection and ordering of the needed operations to answer the item. Calculations refers to the skills needed to complete each mathematical function. The composite dimension refers to the required utilization of both primary dimensions to solve mathematics problems.
3. Except where noted in the methods and discussion sections, items that are designed to load on one dimension load entirely on that dimension. Items that load on multiple dimensions are assumed for purposes of this study to load approximately equally on both dimensions.

Justification of the Project

The validity argument, as described by Messick (1995) consists of six facets. These facets are: Content, substantive, structural, generalizability, external and consequential. Each of these facets must contain its own evidence and combine with the evidence from other facets to create a foundation that supports the claim of a valid inference from scores for a particular test purpose.

One such evidence, falling under the facet of structural validity, is evidence of dimensionality. Does the assessment cover material from one domain without covering material considered to be ancillary or external to the domain? If all items within an assessment can be shown to measure primarily the same construct, the assessment can be considered unidimensional. However, if some of the items are shown to measure knowledge, skills, or attitudes outside the domain of interest, the assessment must be considered multidimensional. Assessments should not be assumed to be unidimensional, but rather, the dimensional nature of an assessment should always be investigated (Ackerman, 1994). Because many assessments require multiple skills to generate a correct response Traub (1983) argued that unidimensionality is perhaps more the exception rather than the rule. Stout (1990) also notes that several minor abilities may be required to respond to an assessment item. He uses the term *essential unidimensionality* to indicate that an assessment has only one dominant latent trait. Such minor traits may include the ability to read for a mathematics test, or to use a keyboard and mouse during a computer-assisted assessment. Evaluating the degree to which these minor traits remain minor and do not interfere with the dominant latent trait or construct is important in assessing the structural aspect of validity.

A data set that contains multidimensional data can be modeled using a multidimensional model. Researchers who model multidimensional data without accounting for these multidimensional properties will produce inaccurate results and the inferences derived may be invalid.

CHAPTER 2

REVIEW OF RELATED LITERATURE

The purpose of this project is to determine the accuracy of two measurement models when applied to unidimensional and multidimensional data. To provide the necessary background, this literature review will cover item response theory, goodness of fit, dimensionality in item response theory, the use of factor analysis in dimensionality assessment, multidimensional item response theory, MIRT software, and Monte Carlo studies.

Item Response Theory

Item response theory (IRT) is an umbrella of statistical models that attempts to measure the abilities, attitudes, interests, knowledge or proficiencies of respondents as well as specific psychometric characteristics of test items. Hambleton (2000) stated that item response theory places the ability of the respondent and the difficulty of the item on the same measurement scale so direct comparisons between respondents' abilities and items are possible. The ability or proficiency of the respondent is labeled theta (θ). The test item characteristics are described by the difficulty (b), discrimination (a), and pseudo-chance (c) parameters. Not all IRT models utilize all item parameters, and there is a continuing debate about the appropriateness of these parameters. For example, the Rasch model uses only the difficulty parameter and ignores the discrimination and pseudo-chance parameters completely. Because the Rasch model uses only the difficulty parameter as the only item parameter, it is called a 1-PL model (for 1 parameter logistic). Another model uses both the difficulty and discrimination item parameters and is called a

2-PL model. The model that utilizes all three parameters is called the 3-PL model. The formulas for the 1-PL and 2-PL models are shown in Equation 1 and Equation 2

$$P_i(\theta) = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}} \qquad i = 1, 2, 3, \ldots, n \qquad \text{Equation 1}$$

Where:

$P_i(\theta)$ is the probability of an examinee with ability θ answers item *i* correctly.

b_i is the difficulty parameter for the i^{th} item.

n is the number of items within the assessment.

e is a transcendental number (natural log constant) whose value to three decimal places is 2.718.

$$P_i(\theta) = \frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}} \qquad i = 1, 2, 3, \ldots, n \qquad \text{Equation 2}$$

Where:

$P_i(\theta)$, b_i, *n*, and *e* are defined the same as in the 1-PL model.

D is a scaling factor equal to 1.7 and used to approximate the two-parameter normal ogive function.

a_i is the item discrimination parameter the i^{th} item.

The parameters (θ, *b*, *a*, & *c*) are graphed in such a way as to yield important information about the test items themselves. Figure 1 below shows an item characteristic curve for a hypothetical item with the identifying item parameters. The x-axis represents the item's difficulty. Because this is on the same scale as the respondent's ability, we can quickly identify which items are appropriate or "answerable" by a given respondent with a given ability or proficiency.

The item difficulty parameter (*b*) is plotted on the x-axis and is an indicator as to the difficulty of the item. Easier items have a lower value for *b* and the corresponding item traceline is shifted to the left. Harder items have a higher value for *b* and the corresponding item traceline is shifted to the right.

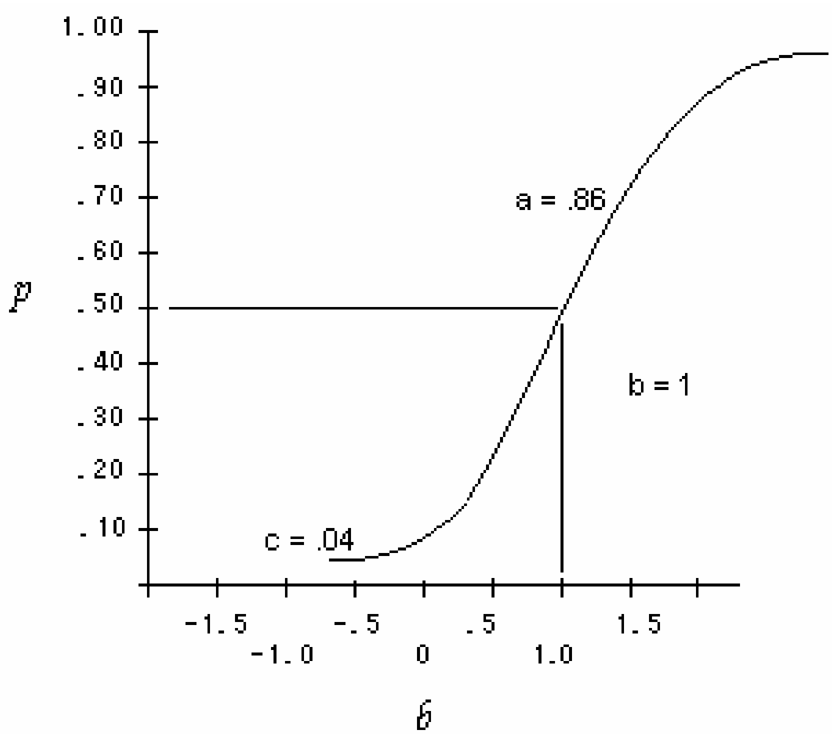

Figure 1. Item characteristic curve

The discrimination parameter (*a*) indicates the slope at the inflection point of the traceline. More discriminating items have a steeper slope. Less discriminating items have a much flatter slope, indicating that respondents of varying abilities have a similar probability in answering the item correctly.

The pseudo-chance parameter (*c*) shifts the lower half of the traceline to a designated point above the x-axis. As the traceline shifts upward, students of lesser ability have a greater probability of a correct response. The pseudo-chance parameter can be construed as the probability of a correct response by an examinee of extremely low ability.

Goodness of Fit

A battery of fit statistics exists that indicate the degree to which a given IRT model adequately fits the empirical data. These are typically called Goodness of Fit Indices (GFI). A poorly fitting model cannot yield theoretically invariant item and ability parameters. Tests for goodness of fit must be performed to ensure that the appropriate model is applied.

All IRT software packages provide goodness of fit statistics. The appropriateness of each fit statistic must be considered when fitting a measurement model to empirical data.

In his presentation at the International Objective Measurement Workshop, Smith (2002) discussed the application of fit statistics. His insight is that there is no single universal fit statistic that is optimal for detecting every type of measurement disturbance. Each statistic has its strengths and weaknesses. By identifying the different types of measurement disturbances, one can select the most appropriate fit statistic. This fit statistic can then be used to determine how adequately the selected IRT model fits the data.

Smith further classifies fit statistics into three categories: total fit, within fit, and between fit. These types differ in their purpose, and in the manner in which they summarize the squared standardized residuals. Another term, misfit, is used to identify when a model fails to adequately fit the data.

The total fit statistic describes misfit due to the interactions of any item/person combination. This statistic works best in identifying random types of measurement disturbances between a target and focal group. The between fit statistic compares logical groups such as gender, ethnicity, or age to detect item bias and is best at identifying systematic measurement disturbances. The within fit statistic is similar to the between fit statistic. Whereas the between fit statistic sums over the entire respondent sample, the within fit statistic is summed over only the group of interest.

Just as no single fit statistic functions optimally to describe the various types of misfit, no single fit statistic functions best for all conditions within these three categories. The fit statistic should be selected based on the specific type of misfit that is of interest.

Each of these types of fit statistics can be calculated as either weighted or unweighted. The weighted calculation attempts to reduce the variation introduced by wide ranges of person abilities or item difficulties.

Goodness of fit indices are largely dependent on the sample size. For some indices, such as the likelihood ratio chi-square, large sample sizes can distort the statistic, artificially inflating its value and leading to erroneous assumptions about the data (Byrne 2001). Small sample sizes are also problematic because of the lack of statistical power

(Hambleton, Swaminathan, & Rogers, 1991). If the sample size is between 100 and 1000, the chi-square can be an appropriate goodness-of-fit indicator. An additional advantage of the chi-square is that of a known distribution.

Monte Carlo studies have shown that any of the chi-square procedures can adequately identify an appropriately-fitted Rasch model with sample sizes of no more than 500 and a test length of approximately 50 items (McKinley & Mills, 1985). McKinley and Mills compared Bock's chi-square, Yen's chi-square, Wright and Mead's chi-square, and the likelihood ratio chi-square to determine whether or not these statistics could identify misfitting items. They tested the three IRT models with three sample sizes of 500, 1000, and 2000 on assessments of 75 items. Their study involved both unidimensional and multidimensional data. They showed that all of the chi-square statistics were distorted with larger sample sizes. This distortion was more apparent with lower-ability respondents than with higher-ability respondents. Multidimensional data caused a greater distortion in the chi-square statistics than did unidimensional data. For sample sizes of 500 responses, all chi-square statistics seemed to adequately show the degree of misfit.

Other indices have been proposed which take into account the fluctuation caused by sample size. Mean-square statistics are chi-square statistics divided by the sample size. Mean-square fit statistics are indicators of the amount of distortion in the measurement system with an expected value of 1.0. Values less than 1.0 indicate either an overfit of the data to the model or redundancy in the data. Values greater than 1.0 indicate random noise. An advantage of the chi-square statistic is that of a known distribution. The mean square statistics do not have a known distribution.

The mean square statistic can be standardized (0,1) by using the Wilson-Hilferty cube root transformation (ZSTD). However, Linacre (2004, page 169) cites Ben Wright's advisement that the ZSTD is useful only in situations in which the MNSQ is greater than 1.5 and either the sample size is small or the test length is short (< 20).

Hulin, Lissak, & Drasgow (1982) use the root mean square error (RMSE) in the recovery of 2-PL and 3-PL item characteristic curves. Drasgow & Parsons (1983) used the root mean squared differences to successfully recover the item parameters for the 2-PL model. In both of these studies, the fit statistic showed little or no distortion for sample sizes of over 2000 candidates on assessments that varied in length from 15 to 65 items.

Zhao, McMorris, Pruzek, and Chen (2002) used both the root mean square error and average standard error estimate (ASE) and determined that the RMSE for the 3-PL model captured the singularity for each dimension of the two-dimensional θs more precisely than the RMSE of the 1-PL ($RMSE_1$) and 2-PL ($RMSE_2$) models. Zhao, McMorris, Pruzek, and Chen reported that $RMSE_2$ was larger than $RMSE_1$, with $RMSE_3$ being the smallest of the three. This trend held true across the maximum likelihood, bayesian sequential, and bayesian EAP (expected a priori) estimation methods.

The RMSEA or Root Mean Square Error of Approximation takes into account the complexity of the model as well as the sample size. RMSEA values of 0 indicate perfect fit. Steiger (1990) defines RMSEA values less than or equal to .05 as being close fit. Brown and Cudeck (1993) further suggest that values between .05 and .08 are fair fit and values between .08 and .10 are mediocre fit.

These are rules of thumb, and no consensus exists. McDonald (1999) states:

> A conventional “rule of thumb” is that the approximation is acceptable when RMSE < .05. The basis of this rule is not clear. It is also not clear if either of these indexes is preferable to the GFI previously defined. At the time of writing the status and utility of the goodness of fit indexes and any “rules of thumb” for them are still unsettled, and it may be questioned whether their use is at all desirable, but the student will certainly encounter them in research reports (p. 171).

Dimensionality of IRT

The topic of dimensionality in assessment precedes the development of item response theory. The focus of dimensionality in this literature review pertains specifically to item response theory.

Item response theory (IRT) entails a statistical assumption of the unidimensionality of an assessment, specifically the measurement of a single latent trait. Although many traits may be necessary to generate a correct response in an assessment, the assumption of unidimensionality is satisfied if only one dominant trait accounts for the largest proportion of variance in the correct responses to a set of test data. Assessments that are not unidimensional risk failing to provide the evidence necessary to support the unified validity concept as developed by Messick. Furthermore, departure from the unidimensionality assumption may result in an incorrect application of the IRT model. Ackerman (1994) wrote that a presumed single trait dimension for any multidimensional test data might jeopardize the invariant feature of the unidimensional IRT models. Furthermore, this could lead to incorrect conclusions about the nature of the test data.

Steinberg, Thissen, and Wainer (2000) identify two distinct categories of multidimensionality: between group and within group.

Between-group multidimensionality. Between-group multidimensionality occurs when the underlying dimensionality of assessment items differs between two target groups of individuals. The assessment measures “different things for different people.” This happens when, all other things being equal, a person who belongs to a particular group has a better or poorer chance of responding correctly to an item than an individual who is not a member of that target group. This is a sensitive issue for racial or ethnic groups. A procedure for detecting between group multidimensionality is called differential item functioning.

Within-group multidimensionality. Within-group multidimensionality occurs when something inherent to the item itself prevents those within the same group from responding the same way.

Within-group multidimensionality can be subdivided into at least three categories. The first category is multidimensionality introduced by the nature of the tasks which make up the assessment instrument. The second category is construct irrelevant multidimensionality. Construct irrelevant multidimensionality within an assessment item is the inclusion of knowledge, skills, or attitudes that lie outside the domain of interest. The third category is construct-relevant multidimensionality that lies within the domain of interest that inherently spans multiple constructs.

The first category, multidimensionality that is introduced by the assessment itself, requires extraneous skills to complete the assessment that do not directly relate to the domain of interest. Items requiring linguistic ability in an oral exam, reading ability in a math test, or mouse and keyboard skills in a computerized exam are all examples of multidimensionality introduced by the assessment instrument. This category violates the structural aspect of Messick's unified validity theory.

The second category is the measurement of knowledge, skills, or attitudes that lie outside the domain of interest. This measurement of extraneous knowledge or skills benefits those respondents who are more capable within this extraneous domain while unfairly penalizing those less capable in this domain although they may be equally competent within the domain of interest. Items in a writing assessment that require a written response to a passage discussing football may unfairly advantage sports enthusiasts who are otherwise lacking in writing skills. This category violates the content aspect of Messick's validity theory.

The third category of within-group multidimensionality is construct-relevant multidimensionality that lies within the domain of interest and inherently spans multiple constructs. This third category is troublesome in that the measurements of knowledge, skills, or attitudes are imprecise indicators of the constructs. The measurements yield ambiguous or erroneous results that can cause incorrect assumptions of a respondent's ability. Items that require skills which span multiple within-domain constructs fail to pinpoint the strengths or weaknesses a respondent may have in relation to a specific construct when measured with a unidimensional measurement model.

The inappropriate application of a unidimensional model to multidimensional data has potentially serious implications. The inferences are likely to be invalid, possibly resulting in respondents who have mastered the subject matter being denied credit for having done so, or respondents who have failed to master the subject matter being given credit when no such credit is due. These false pass/fail decisions have a detrimental effect on the respondents themselves, and can threaten the credibility of the testing instrument.

As a test of unidimensionality, Reckase (1979) suggested the use of an eigenvalue plot of the interitem tetrachoric correlation matrix. Not all agree with this procedure. Steinberg, Thissen, and Wainer (2000) illustrate this lack of consensus. Although there are many different procedures, each has both its advantages and disadvantages. Exploratory and confirmatory factor analytic techniques are the most commonly used methods.

Factor Analysis

Factor analytic techniques are statistical tools used to reduce the number of variables as well as to assess the structure of data. Exploratory factor analysis will attempt to categorize assessment items into dimensions or factors. Confirmatory factor analysis is used to assess how well a set of test items fits a pre-specified model.

In terms of statistical power a minimum of four items are needed to indicate the presence of a factor, resulting in an over-identified model. Three items result in a just-identified model that can neither reject nor fail to reject the null hypothesis. Two items result in an under or non-identified model (Kaplan, 2000). Factor analysis studies the correlations and/or covariances between items. If the percentage of respondents correctly answering each item is between 20 and 80 percent, the covariance matrix should be analyzed. If the percentage correct for any item is more extreme than the 20% to 80% range, the tetrachoric correlation matrix should be analyzed.

Results from a factor analysis can include a scree plot of the eigenvalues, showing the expected number of factors to extract from the data. The factor analysis also shows how much of the variance is accounted for by each extracted factor. Typically, most of the variance will be accounted for by the first or general factor, with the remaining variance explained by a few additional factors.

Although there are many extraction methods used in factor analysis, the two most common are principal components and maximum likelihood. Principal components analysis identifies the linear combinations of the variables that “best” capture the relationships among them with one principal component being extracted for each variable in the data. The single component that accounts for the most variance among the variables is known as the first or principal component. The remaining variance that is not accounted for by the first component is then used to define a subsequent component. The process of extracting subsequent components continues until the data contains only a very little amount of random variability. Because each subsequent component maximizes the variability not captured by preceding components, the components are uncorrelated or mutually orthogonal.

The number of extracted factors to retain is an arbitrary decision. When a correlation matrix is factored, Kaiser (1960) proposed that only those factors with eigenvalues greater than one should be retained. This is equivalent to saying that each factor must extract at least as much variance as one original variable (test item).

Cattell (1966) provides a graphical method to determine the number of factors to retain. Cattell plotted the eigenvalues in order of descending value. As the values of the plotted factors decrease, the decremental variation tapers off to a near-straight line. The factors that taper off are called “factorial scree” which is analogous to the debris that collects at the bottom of a rocky cliff. Only factors that help create the factorial slope are retained. Figure 2 shows a hypothetical scree plot with indicators showing both the retained and the scree factors.

The decision to retain or discard the third factor shown in Figure 2 is unclear. The eigenvalue of factor 3 is less than 1.0. Some practitioners would argue for retention while

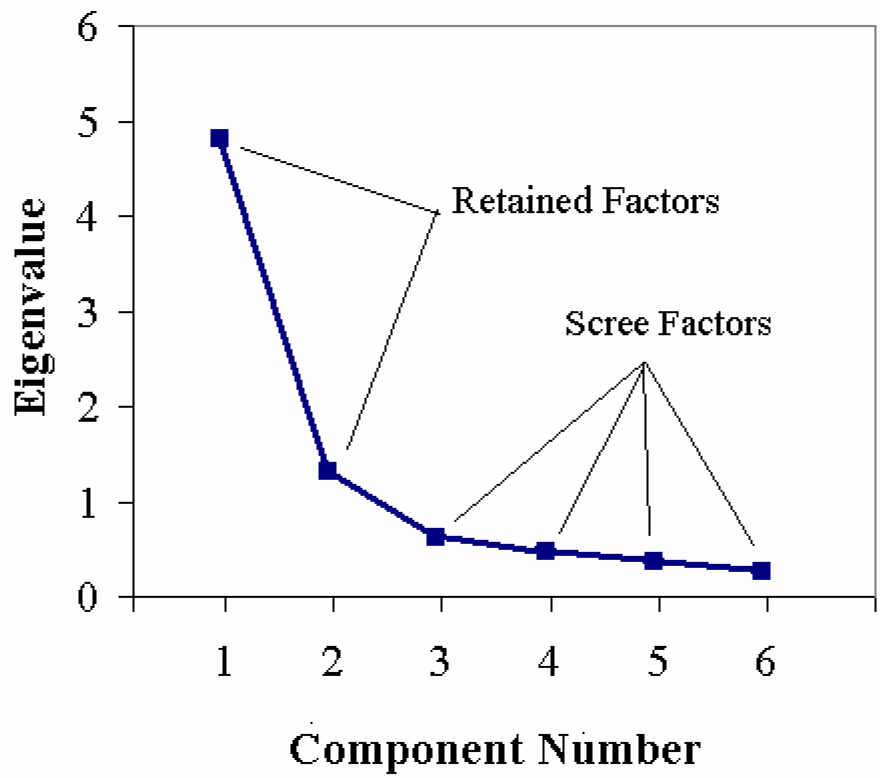

Figure 2. Scree plot of eigenvalues

other practitioners would argue that the third factor should be classified as scree and discarded.

Multidimensional Item Response Theory

Multidimensional item response theory (MIRT) finds its genesis in two different disciplines. Reckase (1997) writes that MIRT can be considered as either an extension of item response theory applied to multidimensional data, or as a special case of confirmatory factor analysis. Unlike unidimensional item response theory, multidimensional item response theory assumes that more than one major trait is necessary and desirable to account for performance on an assessment. An example of a multidimensional assessment may be a problem that asks students how many years were mentioned in the first sentence of Lincoln's Gettysburg Address. A correct response would first require that students be able to recall the first line of the address ("Four score and seven years ago …"), then be able to translate "Four score" to eighty years, and finally correctly add eighty plus seven to obtain an answer of 87. Note however, that knowledge of dates in American history may also aid in generating the response if the respondent is aware of the dates of both the Civil War and the American revolutionary war. Such an item would be considered multidimensional.

MIRT is divided into two branches: Compensatory and noncompensatory. Reckase (1997) explains the differences between these. The formula for compensatory MIRT is additive in nature and therefore a respondent who happens to be weak in one dimension can make up for or compensate for this weakness by a strength in another measured dimension. For example, a child who is familiar with baseball but has poor reading skills may perform well on a test that requires him to read a passage on playing baseball and then write a brief essay about the reading passage.

The noncompensatory MIRT model is multiplicative in nature. Therefore, a respondent who is weak in one area can not make up for this weakness by having a strength in another area. A typist who types 75 words per minute may use a keypad to type 125 characters per minute. Some data entry positions may require that the successful applicant type 50 words per minute and 180 characters per minute on the keypad. In such a scenario, the ability to type 75 words per minute does not compensate for the weakness in the keypad entry of 125 characters per minute.

Currently, all MIRT estimation programs use only the compensatory model. The reason for this, as given by Knol and Berger (1991), is "The disadvantage of noncompensatory models is that no efficient algorithms for estimation of the item parameters are available." Until such algorithms are developed, MIRT calibration software will continue to focus on the compensatory models.

Following our data entry example, if the stakeholders who commissioned the assessment require that strength in one skill not compensate for a weakness in another skill, the assessment results should report separate scores in a composite profile. This profile would require separate items that separately cover typing skills and keypad entry.

Each dimension modeled in MIRT can have the same parameters as the 1-PL, 2-PL, or 3-PL IRT models. Therefore, the graphs that depict MIRT are 3-dimensional. The item characteristic curve in the IRT models is replaced by an item characteristic surface. The discrimination parameter in IRT is replaced by a multiple discrimination parameter

(MDISC) that represents the multidimensional slopes of the surface in different directions.

With the advances in computer software and computing strength, calculations for the more complex MIRT models can be performed, allowing more precise modeling of the test results.

Just as unidimensional item response theory labels the different models by their parameters, multidimensional item response theory labels the models by the parameters and an additional identifier to indicate the compensatory or noncompensatory model. The multidimensional item response theory model that is compensatory in nature and uses two parameters is abbreviated as MC2-PL.

Some scholars assert that rather than simply fitting a model to the data, there must be an underlying reason to apply a multidimensional framework to an assessment. For example, Luecht (1996) states the following:

> That professional certification or licensure tests comprised of complex, integrated content are perhaps multidimensional is not the relevant issue. Rather, the question is whether there is any advantage to attempting to decompose a test into arbitrary and perhaps substantively meaningless statistical multivariate latent structures when the most that could be accomplished would be to estimate a set of (probably unstable) coefficients or loadings for recombining the multivariate scores in some fashion to generate a total test composite score (p. 389).

Still, he argues that if an assessment is fundamentally multidimensional, separate profiles should be developed to report performance on each relevant dimension. He sees a dual purpose in such assessments: The reporting of subscores based on performance in separate categories, and the total pass/fail decision made at a global level as covered by the test. Such an assessment must maintain the content validity at the test level.

Luecht's argument follows Stout's (1990) reasoning that for items measuring multiple traits, the decision must be made as to whether or not one of these traits is primarily dominant and therefore essentially unidimensional. Segal (1996) demonstrated that for correlated traits with items loading primarily on only one trait, unidimensional item parameters that are estimated uniquely for each trait may also be of practical value. The foci then become the unique dimensions, each evaluated independently of each other. Segal's work on the ASVAB (Armed Services Vocational Aptitude Battery) sciences test is composed of chemistry and physics items on one dimension and biology and life science items on a second dimension.

An alternative approach to multidimensionality is best explained by the following example: If two dimensions were apples and oranges, the multidimensional nature would be a fruit salad. Would the stakeholders want to measure simply the number of apples or the oranges within the salad? Or would they want to measure the amounts of ingredients within the entire salad, which includes the interaction between the apples, oranges, and any other additional fruit.

The previous arguments detailed by Stout, Luecht, and Segal to create a set of profiles are analogous to measuring the amount of each individual fruit such as the apples or the oranges.

A philosophical approach that would guide whether to create a single composite score or a set of scales for a profile is to decide whether or not the domain involves a single multidimensional construct or multiple constructs that are intercorrelated.

This philosophical approach will determine not only the assessment strategy, but also the instructional strategy. The single multidimensional construct would best be modeled using a work model approach (Bunderson, Gibbons, Olsen, & Kearsley, 1981). The work-model approach utilizes a concept of increasingly-complex performance microworlds in which a set of elementary constructs are subsumed by a larger, more complex construct, which in turn is later subsumed by an even larger and even more complex construct. An example of such a microworld is the psychomotor construct of the ability to ride a bicycle. The separate subskills of steering, pedaling, balancing, and braking are each subsumed by the greater construct of bicycle riding. With the assistance of another individual to help balance the bike or hold the rear wheel off the ground, each of these component skills can be mastered independently of each other. Additional constructs such as changing gears can be added at later stages. A more scholastic example that lies in the reading domain would be the simple constructs of phonemic awareness and letter recognition being subsumed by word recognition and reading fluency. Once a construct is subsumed, the assessment no longer needs to assess a learner’s ability at that level.

The philosophy of multiple intercorrelated-constructs is best modeled by a more traditional approach that utilizes entry-level objectives, a hierarchical structure of enabling objectives and finished by one or more correlated terminal objectives. A mathematics assessment could be an example of this latter approach. The basic operations of addition and subtraction are not entirely subsumed by the more complex constructs of multiplication, division, and exponentiation but rather continue to more complex levels through advanced math, algebra, and calculus with the introduction of constructs such as derivatives and integrals.

IRT Software

Several software packages have been developed that calculate item parameters using both classical test theory and item response theory.

The most commonly used was BILOG from SSI software. BILOG estimated parameters for dichotomous data using the Rasch, 2-PL and 3-PL models. Additional classical test statistics were provided such as the biserial and point biserial correlations, as well as the classical item difficulty indices.

The functionality of BILOG was incorporated into the release of BILOG MG (Multiple Group) 3.0. With this release, BILOG as a separate program was discontinued.

BIGSTEPS computes item parameters for polytomous or dichotomous data with a Rasch model. WINSTEPS is the Windows version of BIGSTEPS.

Quest and ConQuest (Wilson, 1999) implement the Rasch model as well as many other linear and non-linear models.

MIRT Software

As noted previously, MIRT software uses only the noncompensatory model. Several programs have been developed to model multidimensional data. Among these are TESTFACT, NOHARM II, and MAXLOG. Each of these estimates item parameters from dichotomous data only. Polytomous estimation was evaluated in the program

POLYFACT developed at Educational Testing Service by Eiji Muraki, but is not commercially available.

TESTFACT (Wilson, Wood & Gibbons, 1984) allows the marginal maximum likelihood procedure for item parameter estimation. Furthermore, it implements the EM (expectation – maximization) algorithm developed by Dempster, Laird, and Rubin (1977) to determine which estimation procedure between Maximum Likelihood or Residual Maximum Likelihood is the optimal procedure.

NOHARM II (Fraser, 1988) builds on McDonald's (1985) harmonic non-linear factor analysis. This IRT model uses only information contained in the pairwise proportions. NOHARM II approximates the pairwise probabilities by minimizing the unweighted least squares function.

MAXLOG (McKinley & Reckase, 1983) yields estimates of the MC2-PL model through joint maximum likelihood. This method is susceptible to drift of the discrimination parameters. Also, estimation is cumbersome with large sample sizes.

Each of these programs has limits on the number of dimensions and variables used. As such, they are not as useful for large scale applications.

In a series of Monte Carlo simulations, Knol and Berger (1991) compared these MIRT programs with several factor analytic methods. In all IRT situations, NOHARM and TESTFACT performed better than MAXLOG. When datasets with two dimensions were used, TESTFACT performed better than NOHARM. However, when three or more dimensions were used, NOHARM outperformed TESTFACT. In the factor analytic tests, TESTFACT performed more poorly than the FA methods IPFA (Iterated Principal Factor Analysis) and MINRES (Minimum Residual Analysis) for two or three dimensions. For data sets with six dimensions, TESTFACT performed much worse than IPFA and MINRES. The authors note that extreme data sets were used, and that when difficulty ranges between +2 and –2 are used, that TESTFACT performs almost as well as IPFA and MINRES for factor analysis. (Note that the highly qualitative terms “performed,” “ better,” and “poorly” are those of Knol and Berger).

ConQuest (Wilson, 1999) is a program that implements item response and latent regression models. It implements the Rasch, Partial Credit, Generalized unidimensional, and multidimensional item response models by using marginal maximum likelihood estimates.

Monte Carlo Studies

Because the generating properties of empirical data can’t be sufficiently controlled, this project utilized a Monte Carlo study. By generating and controlling each of the parameters, we can predict what the outcome is expected to be when applying the various IRT and MIRT models. Furthermore, we can compare the predicted and observed outcomes by using descriptive statistics to test the usefulness of the model. Harwell (1997) succinctly described the use of Monte Carlo studies by writing “In the absence of exact mathematical solutions, Monte Carlo studies have been used.”

Harwell highlights the need for results from Monte Carlo studies to be analyzed in ways that clarify the findings. With the volume of data generated from Monte Carlo simulations, simple descriptive statistics tabulated in a chart is often overwhelming. He suggests that both the descriptive and inferential statistics that are utilized in empirical studies be appropriately applied in interpreting and explaining the results of a Monte Carlo study.

Spence (1983) argues that Monte Carlo studies should be treated as statistical sampling experiments and be held to the same principles of experimental design and data analysis as empirical studies.

Perhaps the most common method of generating dichotomous results is based on a normally distributed population. Leucht (1996) compares a uniform random probability π_{ji} to a matrix of examinee X items using the formula for the MC2-PL model. The score, $u_{ji} = 1$ if $P_{ji} \geq \pi_{ji}$, and $u_{ji} = 0$ if $P_{ji} \leq \pi_{ji}$. To generate unidimensional data, one need only apply the formula for the Rasch model. By comparing a normal distribution of ability levels to a uniform random distribution, instances where the uniform distribution is greater than the random distribution results in an incorrect response. Instances where the uniform distribution is less than the random distribution results in a correct response.

This method is detailed in a step-by-step fashion by San-Luis and Sanchez-Bruno (1998). They created 1000 normally-distributed subjects with a mean of zero and a standard deviation of 1. The p_i probability of a correct response was generated using the formula for the 2-PL IRT model. This probability of a correct response was compared to a uniform distribution between 0 and 1. A correct response was generated if the p_i probability was greater than the uniform distribution. If the uniform distribution was greater, then an incorrect response was generated. From this 1 x 1000 vector, a plot was constructed with 250 equidistant points from -3 to +3 on the x and y axis. The log-likelihood values were plotted in a graphical representation.

CHAPTER 3

METHODS

The methods section is divided into three parts: (a) the parameter estimates that are to be recovered, (b) the generation methods to generate the data, and (c) the analysis and comparison of results.

A brief note from Wilson (2004) must be mentioned. In the ConQuest user manual, an example of mathematical ability is provided. Wilson notes that the dimensions do not share a common unit nor point of origin. The multidimensional latent space modeled by ConQuest may or may not share a common origin nor common unit of measurement. The dimensions we observe may simply be a reflection upon a common plane. The data in this project draw upon a simplest-case scenario in which the multidimensional properties can be artificially constrained. Such constraints will hopefully provide a fertile environment in which these questions can be adequately answered.

Parameter Estimates to Be Recovered.

The original parameter estimates that the programs were to attempt recovery came from both person-level and the item-level data. Therefore, both person-level and item-level data sets were required. All four research questions required multidimensional person-level data. Question 1 required unidimensional item-level data. Questions 2 through 4 required multidimensional item-level data.

Table 1 summarizes the data requirements to answer each of the research questions.

Table 1.

Person- and Item- Level Data Required to Answer Research Questions.

Research Question	Person-Level Data	Item-Level Data
1	Multidimensional	Unidimensional
2	Multidimensional	Multidimensional
3	Multidimensional	Multidimensional
4	Multidimensional	Multidimensional

Recovery of Parameter Estimates for Question 1

The estimates to be recovered for the first research question were the original unidimensional item difficulty values for 21 items and the multidimensional person ability values for 1000 simulated respondents (for each iteration).

Recovery of Parameter Estimates for Question 2

The estimates to be recovered for research question 2 were the original multidimensional item difficulty values for the 21 construct-relevant multidimensional items.

Parameter Estimates for Questions 3 and 4

Questions 3 and 4 did not require the recovery of any original parameters. Instead, the data used to answer question 2 were used to answer questions 3 and 4. Question 3 was answered by applying the Rasch model to evaluate the misfit statistics. Question 4 was answered by applying the 2-PL model to the data, then projecting the seven items from the Necessary Operations construct onto the Calculations construct and reapplying the 2-PL model.

Generation Methods

First, an overview explains the logical structure of the data sets. After the overview, the generation processes for both the item-level and the person-level data are explained.

Overview

Although the basis of this dissertation was a Monte Carlo study, the antecessor comes from the real-life domain of a mathematics assessment that covers two correlated constructs. These two constructs were necessary operations and calculations. Necessary operations (NO) was plotted on an oblique (correlated) y-axis. Items within the NO construct assess the respondent's ability to recognize, select, and properly order the mathematical operations of addition, subtraction, multiplication, division, and exponentiation. The second construct is called Calculations (C), and was plotted on the oblique (correlated) x-axis. Items within the C construct require that the respondent properly perform those ordered mathematical operations. Items that load entirely on the NO dimension do not require any calculation skills. Items that load entirely on the C dimension do not require any ordering of the operations, but have these operations already provided in their proper order. Items that do not load univocally on one of the two dimensions, but occupy a location in construct space that requires ability from both dimensions to answer were plotted midway between the correlated oblique x and y axes. For simplicity in this study, these items were assumed to require equal amounts of ability from both the NO and the C dimensions and therefore fell exactly midway between the two. These items form a composite vector. The composite vector was labeled “Z” on the oblique coordinate system. For purposes of this study, the two constructs NO and C were correlated at .50. This .50 correlation is equivalent to an angle of 60°. The 60° angle is obtained by calculating the arc-cosine of .50. The item and person data were plotted in this 60°-degree construct space. The composite vector Z bisected the two dimensions at 30°. The cosine of 30° is .866. Therefore the composite vector Z was correlated at .866 with both the NO and C dimensions. Figure 3 illustrates these three content-related dimensions.

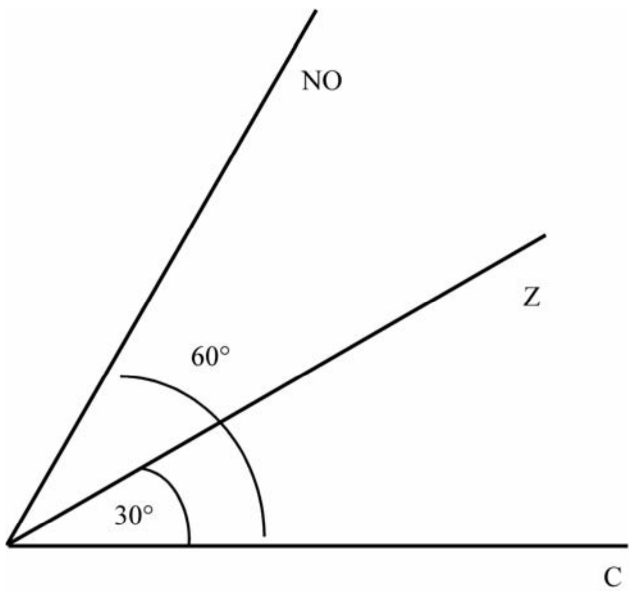

Figure 3. Two dimensions (NO & C) correlated at .50 with a third dimension (Z) bisecting the two.

The item and person-level parameters were drawn from these dimensions. The research questions required the projection of these parameters onto one or another dimension. These projections involved a 3-step process:

1. Conversion from the oblique coordinate system to an orthogonal coordinate system.
2. Projection of the parameters onto the target orthogonal dimension.
3. Conversion of the projected orthogonal parameters onto an oblique coordinate system.

This process will be explained subsequently in greater detail.

Item-Level Data

The research design called for twenty-one items to be placed on these three dimensions. Seven items fell on each of the NO, C and Z dimensions.

Table 2 shows the difficulty values for 21 items that loaded on the dimensions (NO, C) and the orthogonal projection of these difficulty values onto the composite dimension Z.

The trigonometric properties are such that for the projections of the seven NO items onto C and Z, the following conversion algorithms were used: C = (.5 * NO) and Z = (.866 * NO). For the projections of the seven C items onto NO and Z, NO = (.5 * C) and Z = (.866 * C). For the projections of the seven composite Z items onto C, the following conversion algorithms was employed: C = (.866 * Z). Items 7 and 14 did not fall directly on the NO and C dimensions. Therefore, these trigonometric functions did not apply. Because this was an oblique coordinate system, the projections for these items were done after plotting them onto an appropriate orthogonal coordinate system.

Table 2.

Starting Difficulty (b) Values For 21 Items and their Projections from the Dimension of Origin to the Destination Dimension.

		Destination Dimension		
	Dimension			Z
Item	of Origin	NO	C	Composite
1	NO	-2.70	-1.35	-2.34
2	NO	-1.70	-0.85	-1.47
3	NO	-1.00	-0.50	-0.87
4	NO	0.60	0.30	0.52
5	NO	1.70	0.85	1.47
6	NO	2.00	1.00	1.73
7	NO	2.50	0.35	1.65
8	C	-1.25	-2.50	-2.17
9	C	-0.70	-1.40	-1.21
10	C	-0.30	-0.60	-0.52
11	C	0.60	1.20	1.04
12	C	1.10	2.20	1.91
13	C	0.14	2.90	2.51
14	C	1.00	2.00	1.73
15	Z	-2.00	-2.00	-2.30
16	Z	-1.39	-1.39	-1.60
17	Z	-0.80	-0.80	-0.90
18	Z	0.00	0.00	0.00
19	Z	0.87	0.87	1.00
20	Z	1.30	1.30	1.50
21	Z	2.34	2.34	2.70

A graphical representation of these 21 items is shown in Figure 4, and the orthogonal projection of these items onto the composite vector is shown in Figure 5.

Item-level data for question 1. For research question 1, the orthogonal projections for items 1 through 14 onto the composite dimension along with the 7 items already loading on the composite dimension were used as though these projections were unique unidimensional items. The discrimination parameter (a) was 1.0 for all item probability functions used in generating data for calibration. This discrimination parameter constraint conformed to the requirements of the Rasch model. These values are shown in the Z-composite column in Table 2.

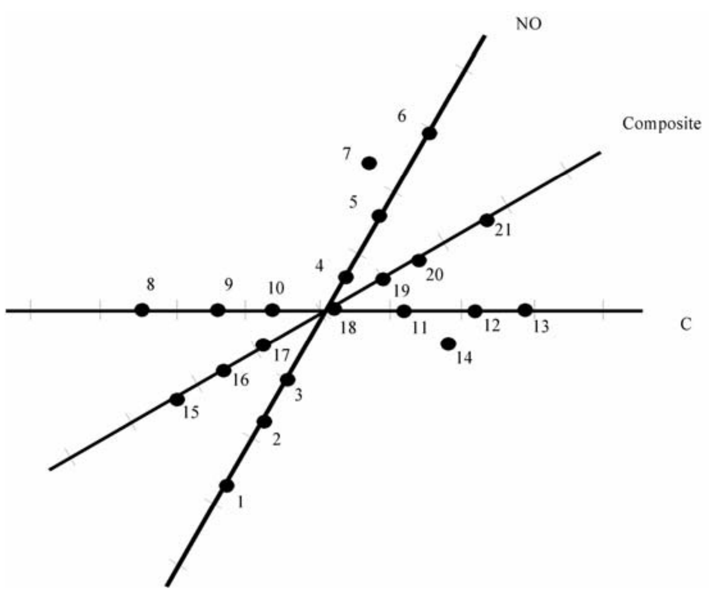

Figure 4. Items that load on two dimensions plotted on an oblique coordinate system

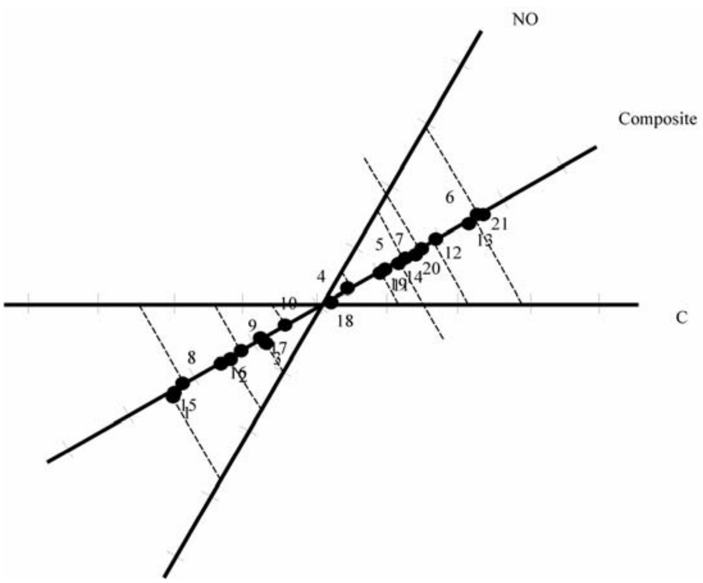

Figure 5. Item difficulties projected onto the composite vector.

Item-level data for question 2. Research question 2 utilized the original multidimensional loadings for all 21 items on their respective dimensions.

Item-level data for question 3. Research question 3 did not require any additional item-level data. Data used in research questions 1 and 2 was used to answer research question 3.

Item-level data for question 4. To establish a baseline for answering research question 4 and to avoid possible distortion of the original generating parameters for items on the C dimension, a total of 21 items were drawn from the hypothetical calculations item pool rather than the original seven items used for research questions 2 and 3. These 21 new items are shown in Table 3 and are identified by indices ordered C1 through C21. These items were unique to the calculations dimension and were not projected onto any other dimension. Their sole purpose was to provide a stable base on the Calculations dimension upon which the seven items from the Necessary Operations dimension could be projected without the potential distortion that would result from calibrating so few items on one dimension. The difficulty and discrimination parameters for the seven Necessary Operations items were calculated for these items prior to projecting the items from NO onto C. The item difficulty and discrimination parameters for the 7 NO items were calibrated one at a time with the same set of respondents. This approach was used to reduce any effects of recalibration with a larger group of items.

Person-Level Data

The person data was generated from probability samples from two normal distributions. These distributions represent the respondents' ability levels for the constructs NO and C. The person sample size was 1000 cases for each iteration. Each of the 1000 ordered pairs (NO, C) inter-correlate at .50. Table 4 shows the ability estimates for 20 randomly selected simulated respondents from one iteration of the person/response generator. The ability estimates shown are for the Necessary Operations and the Calculations constructs.

The 1000 person ability values on the ordered pairs (NO, C) as well as their projection onto the composite (Z) dimension were used to answer the research questions.

Table 3.

Difficulty values for 21 items on the Calculations dimension.

Item	Calculations
C1	-2.39
C2	-2.08
C3	-1.70
C4	-1.36
C5	-1.15
C6	-.79
C7	-.56
C8	-.32
C9	-.06
C10	.05
C11	.10
C12	.12
C13	.22
C14	.54
C15	.70
C16	.98
C17	1.22
C18	1.45
C19	1.95
C20	2.20
C21	2.51

Table 4.

Multidimensional Ability Estimates for 20 Randomly Selected People on the Two Dimensions of Necessary Operations and Calculations.

Person	Ability Estimate on NO	Ability Estimate on C
1	-.91	1.01
2	1.63	1.72
3	1.01	-.59
4	.74	.59
5	.58	1.79
6	.01	-.26
7	-.88	-1.94
8	1.52	1.22
9	-.27	-1.16
10	-.01	.10
11	1.36	.61
12	-.21	.04
13	.70	.47
14	-.98	-1.20
15	.00	.61
16	1.06	1.37
17	-1.38	-.56
18	-.78	-1.82
19	-1.14	-.94
20	1.15	1.03

Person-level data for question 1. An orthogonal projection of these (NO, C) ability values onto the composite dimension was needed to answer the first research question.

The 1-PL IRT model as implemented by Winsteps cannot recover multidimensional theta values, but rather attempts recovery of a unidimensional data structure. The hypothesis was that the person ability values as estimated would more closely align near or on the composite Z dimension. Table 5 shows this hypothesized recovery for the 20 respondents reported in Table 4.

The MC1-PL model as implemented by ConQuest can recover not only the multidimensional theta values, but also can model a unidimensional structure. The hypothesis was two-fold. First, the multidimensional theta estimates as recovered by ConQuest will align with the originating theta values on both the NO and the C dimensions. Second, the unidimensional theta estimates recovered by ConQuest will be similar to Winsteps' unidimensional theta estimates on or near the composite Z dimension. This hypothesized recovery is shown in Table 5. The derivation of the theta estimates on the composite Z dimension is explained in a later section under Generation Procedures.

Table 5.

Unidimensional and Multidimensional Theta Values and their Hypothesized Recovered Estimates for 20 People.

				Hypothesized Winsteps Recovery	Hypothesized ConQuest Recovery		
Person	Ability on NO	Ability on C	Projected Ability onto Z	Ability on Z (1-PL)	Ability on NO (MC1-PL)	Ability on C (MC1-PL)	Ability on Z (1-PL)
1	-.91	1.01	.09	.09	-.91	1.01	.09
2	1.63	1.72	2.91	2.91	1.63	1.72	2.91
3	1.01	-.59	.37	.37	1.01	-.59	.37
4	.74	.59	1.15	1.15	.74	.59	1.15
5	.58	1.79	2.05	2.05	.58	1.79	2.05
6	.01	-.26	-.22	-.22	.01	-.26	-.22
7	-.88	-1.94	-2.44	-2.44	-.88	-1.94	-2.44
8	1.52	1.22	2.37	2.37	1.52	1.22	2.37
9	-.27	-1.16	-1.24	-1.24	-.27	-1.16	-1.24
10	-.01	.10	.07	.07	-.01	.10	.07
11	1.36	.61	1.71	1.71	1.36	.61	1.71
12	-.21	.04	-.15	-.15	-.21	.04	-.15
13	.70	.47	1.01	1.01	.70	.47	1.01
14	-.98	-1.20	-1.89	-1.89	-.98	-1.20	-1.89
15	.00	.61	.52	.52	.00	.61	.52
16	1.06	1.37	2.10	2.10	1.06	1.37	2.10
17	-1.38	-.56	-1.68	-1.68	-1.38	-.56	-1.68
18	-.78	-1.82	-2.25	-2.25	-.78	-1.82	-2.25
19	-1.14	-.94	-1.80	-1.80	-1.14	-.94	-1.80
20	1.15	1.03	1.89	1.89	1.15	1.03	1.89

The conversions to and from the orthogonal coordinate system and the projections onto the composite vector for these 20-person ability values are shown in Table 6. For small data sets, a simple perpendicular projection to the z vector can be done by plotting the coordinates and measuring the distances. For larger data sets, this is not feasible. The projection of person values on an oblique coordinate system was accomplished by first transferring these person values to an orthogonal coordinate system, performing the necessary trigonometric calculations and then transferring the resulting values back to the oblique (correlated) coordinate system. The length of these projections from the origin on vector z is determined on the orthogonal coordinate system.

A plot of these 20 person ability values is shown in Figure 6. Figure 7 shows the orthogonal projection of these 20 person ability levels onto the composite vector.

Person-level data for questions 2 and 3. For research questions 2 and 3, the item difficulty values in both the Necessary Operations and the Calculations columns for Table 2 are to be recovered using MCPL calibration. The person-level data used to generate the response patterns were the projected (NO, C) values onto the Z vector. The person-level data did not need to be recovered for either questions 2 or 3.

Table 6.

Multidimensional ability estimates for 20 randomly selected people on oblique and orthogonal coordinate systems.

	Oblique Coordinate System		Orthogonal Coordinate System		Projection onto Z		Length of Z From Origin
Person	NO	C	NO	C	NO	C	
1	-.91	1.01	-.40	.88	.05	.05	.09
2	1.63	1.72	2.50	1.49	1.68	1.68	2.91
3	1.01	-.59	.72	-.51	.21	.21	.37
4	.74	.59	1.03	.51	.66	.66	1.15
5	.58	1.79	1.48	1.55	1.19	1.19	2.05
6	.01	-.26	-.12	-.23	-.13	-.13	-.22
7	-.88	-1.94	-1.85	-1.68	-1.41	-1.41	-2.44
8	1.52	1.22	2.13	1.06	1.37	1.37	2.37
9	-.27	-1.16	-.85	-1.01	-.71	-.71	-1.24
10	-.01	.10	.03	.09	.04	.04	.07
11	1.36	.61	1.67	.52	.98	.98	1.71
12	-.21	.04	-.20	.03	-.09	-.09	-.15
13	.70	.47	.93	.41	.58	.58	1.01
14	-.98	-1.20	-1.58	-1.04	-1.09	-1.09	-1.89
15	.00	.61	.30	.53	.30	.30	.52
16	1.06	1.37	1.74	1.19	1.21	1.21	2.10
17	-1.38	-.56	-1.67	-.49	-.97	-.97	-1.68
18	-.78	-1.82	-1.69	-1.57	-1.30	-1.30	-2.25
19	-1.14	-.94	-1.61	-.82	-1.04	-1.04	-1.80
20	1.15	1.03	1.66	.90	1.09	1.09	1.89

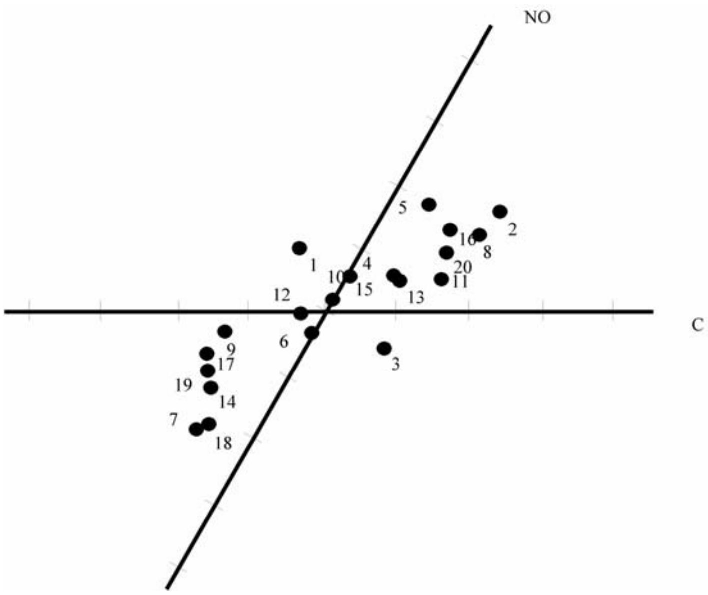

Figure 6. Person ability levels for 20 people on two dimensions.

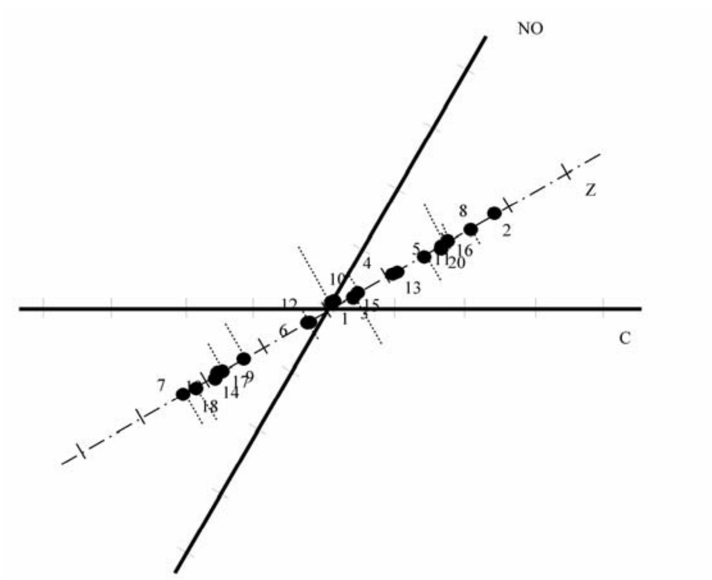

Figure 7. Person ability levels for 20 people after projection onto the composite vector.

Person-level data for question 4. Question 4 requires the 2-PL item calibration of 28 items (this number comes from the needed 21 anchor items, in addition to the seven experimental items) on the NO dimension using person ability values on the NO dimension. After projection onto the C dimension, the seven experimental items were again calibrated along with 21 anchor items on the C dimension. This calibration on the C dimension utilized the person ability values exclusively on the C dimension.

Person-Response Generation for all Items

For items on NO, C and Z, the actual normal distribution deviate for respondent *j* were used in a 1-PL IRT function to calculate a probability p of responding correctly to the *i*th item. This probability was compared to a uniform random number π between 0 and 1. For the score u for person *j* on item *i*, $u_{ji} = 1$ if $P_{ji} \geq \pi_{ji}$, and $u_{ji} = 0$ if $P_{ji} \leq \pi_{ji}$. Item response functions were relative to the NO, C, and Z vectors, respectively for the NO items 1 – 7 the C items 8 – 14, and the Z items 15 – 21.

Generation Procedures

The generation methods are subdivided into two categories. The first category is that which yields the item-level and person-level data. The second category is the procedures that yield the IRT item difficulty parameter estimates.

Item and Person Level Data

The person and item data must have specific properties for the research questions to be answerable. The properties for the item data were unidimensional loadings onto one of two primary dimensions. Each dimension represents a latent trait or construct. The primary dimensions NO and C were correlated at .50. A composite vector called Z represented items that require skills from both primary dimensions to generate a correct response. The value of each item's loading on each dimension was its difficulty parameter for that dimension. For purposes of this project, items that originated on the composite vector were assumed to require equal ability levels from both primary dimensions NO and C.

The properties for the person data were known ability levels for two correlated constructs, and the values for these ability levels when projected orthogonally onto the composite vector. These generating distribution characteristics were:

2 Vectors (NO & C): $\rho = .50$, $\mu = 0$, $\sigma = 1$.

Graybill (1961) provided a solution for generating n-dimensional correlated distributions. A simplification of his formula is shown in Equation 3.

$$X \cap N\left(\mu_1, \sigma_1^2\right) \qquad \left\{ Y \mid X \cap N\left(\left(\mu_2 + \left(\frac{\rho}{\sigma_2^2}\right)(X - \mu_1)\right), \left(\sigma_2^2 - \frac{\rho^2}{\sigma_1^2}\right)\right)\right\} \qquad \text{Equation 3}$$

The application of this formula yields two alternative distributions with a correlation of .5, both with means of .50, and standard deviations of 1. The SPSS syntax used to implement Equation 3 is found in Appendix A. The descriptive statistics that show the correlations of the two distributions for the pilot iteration are found in Appendix B.

The resulting (x,y) values for each case represent the respondent's known theta values for two alternative dimensions. The values can be plotted graphically on an oblique (correlated) coordinate system. The arc cosine of .50 is 60°; therefore, the X and

Y axis were fixed at 60°. The dimension of Necessary Operations fell on the Y axis, and Calculations fell on the X axis. The unit of measurement was in logits, thus facilitating the plotting of item-level and person-level data on the same scale. In this simplest case, the two dimensions bisect at the points of origin. In many instances, such a bisection is not plausible. In fact, multiple dimensions may never share a common point of origin, may never share common units of measure, nor may ever intersect.

The projection of the person-level data to the composite dimension was accomplished with a three-step process. First, coordinates for NO and C were plotted on an orthogonal (Cartesian) coordinate system. Second, these (NO, C) coordinates were projected onto the Z vector. Third, the coordinates on the Z vector were reflected onto the oblique 60° coordinate system. The signed distance from the origin to the plotted point on the Z vector represented the ability or proficiency value on the composite Z dimension.

The implementation of these steps is explained below:

Step 1. Equation 4 shows the plotting of the oblique (NO, C) coordinates on the orthogonal coordinate system. These coordinates on the orthogonal coordinate system are called (P, Q).

$$(x, y) \rightarrow (p, q) = \left(x + \frac{y}{2}, \frac{\sqrt{3}}{2} y \right) \qquad \text{Equation 4}$$

Step 2. The projection of the orthogonal coordinates (P, Q) onto the Z (30°) vector is shown in Equation 5.

$$(P,Q)=\left(\frac{3}{4}p+\frac{\sqrt{3}}{4}q,\frac{\sqrt{3}}{4}p+\frac{q}{4}\right)$$ Equation 5

The length of (P, Q) from the origin is calculated with the formula in Equation 6.

$$length\,Z=\sqrt{\left(\frac{3}{4}P+\frac{\sqrt{3}}{4}Q\right)^2+\left(\frac{\sqrt{3}}{4}P+\frac{Q}{4}\right)^2}$$ Equation 6

Step 3. The plotting of the orthogonal coordinates (P,Q) on the oblique coordinate systems is made using Equation 7.

$$(P,Q)\rightarrow(X,Y)=\left(P-\frac{Q}{\sqrt{3}},\frac{2}{\sqrt{3}}Q\right)$$ Equation 7

The application of Equation 3 through Equation 7 yields the following data:

1. Person-level data with known ability levels on two dimensions.
2. The projection of these ability levels onto a composite vector.
3. The ability or proficiency of each person on this composite Z vector.

The SPSS syntax for generating this data is also found in Appendix A. The descriptive statistics showing the results of this script are shown in Appendix B.

Parameter Estimation

The item difficulty values on the NO, C, and Z dimensions were used with the person ability values on NO, C, and Z to generate the item responses including a normally distributed random error component. Each of the four research questions required item difficulties or person abilities from one or multiple dimensions.

The parameter estimation required the following steps:

1. Generate a correct/incorrect response for each respondent to the items on each dimension as well as the projections for all 21 items onto the composite vector.
2. Apply the MC1-PL and 2-PL IRT models to the values that lie on the composite vector.
3. Apply the MC1-PL IRT model to the multidimensional item responses.
4. Apply the Rasch model to the multidimensional data.
5. Anchor all Calculations items and iteratively project the Necessary Operations items onto the Calculations dimension. Calculate the 2-PL IRT parameters for each replicated iteration.

Each of these steps is detailed below.

Step 1: Generate a correct/incorrect response pattern. The probability of an examinee's correct response to each item was calculated separately for each research question. This probability of a correct response, $P_i(\theta)$ to each item was generated with the Rasch formula. This formula is shown in Equation 8. The probability of a correct response was compared to a uniform distribution between 0 and 1. As previously mentioned, the score for person *j* on item *i*, $u_{ji} = 1$ if $P_{ji} \geq \pi_{ji}$, and $u_{ji} = 0$ if $P_{ji} \leq \pi_{ji}$.

$$P_i(\theta) = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}} \qquad \text{Equation 8}$$

The values for theta were the person (x,y) values for each dimension, and the perpendicular projection onto the composite vector (Z) The values for b were the item (x,y) values for each dimension and the perpendicular projection onto the composite vector (Z).

Steps 2 through 5: Apply the IRT models to the data. With the generated item answers, the MC1-PL IRT and the 2-PL IRT models were separately applied to each dimension using item and person information for only that dimension. For the composite vector, the MC1-PL and 2-PL IRT models were applied to all item and person loadings.

Winsteps was used to calculate the Rasch 1-PL estimates. ConQuest was used to calculate MC1-PL parameter estimates. BILOG MG was used to calculate the 2-PL IRT parameter estimates. To properly answer research questions 1 and 2, multiple item-person data sets were needed.

Research question 1 focused on the ability of each IRT model to recover the unidimensional parameters of interest. Because the two models were compared to each other, the Root Mean Square goodness of fit statistic was the most accurate indicator of this recovery.

Research question 2 used all 21 items calibrated with ConQuest. The logit estimates of these items were compared to the original values. A 95% confidence band was constructed to determine whether the estimate falls within an acceptable range of the original value. The 95% confidence interval was selected because this is a commonly accepted interval for most statistical analyses. To provide a tighter acceptance range, a 65% confidence band was also be used and the differences noted.

Research question 3 was answered by ordering the infit and outfit statistics that were generated in WinSteps. A simple linear rank-order comparison indicated whether or not the items showed an increasing misfit for items that lie further from the point of origin.

Research question 4 was answered by comparing the original discrimination parameter to the discrimination parameter calculated after the item projection. Again, the Root Mean Square fit statistic was the most appropriate indicator of the comparability of the original and projected discrimination parameters.

Previous research shows that approximately fifteen to twenty-five iterations yield sufficiently stable results for similar studies. In light of this heuristic approach, research questions 1 through 3 employed 25 iterations. Research question 4 compared the change that occurred to the discrimination parameter to a group of items. This comparison allowed for the use of a power analysis. A power analysis of $1-\beta$ indicated the number of iterations needed to yield stable results. The number of iterations needed to answer research question four was 19.

CHAPTER 4

RESULTS

Question 1 Results

The first research question attempted to show the robustness of each item response model in recovering the underlying person and item-level structure. The item-level data was unidimensional and the person-level data was multidimensional.

An expectation of Winsteps was that as a unidimensional measurement instrument, both the person-level ability estimates and the item-level difficulty estimates would reflect a unidimensional model. If the unidimensional model were not met, those items and persons not fitting the model would automatically be excluded from the calibration process.

Winsteps has no means of providing multiple theta or difficulty estimates for the same examinee. As such, the Winsteps estimates would not be expected to align univocally with either of the primary correlated dimensions, but instead would be expected to somewhat align in a composite dimension. Using a compensatory model as previously discussed, we first calculated what the expected difficulty and ability or proficiency estimates would be, and then compared these values with the estimates actually obtained from Winsteps.

An attempt could also be made to compare the unidimensional difficulty and ability estimates provided by Winsteps with the originating ability and difficulty values of each of the generating dimensions, but the “gravitational pull” of either dimension would throw off the parameter estimates. The analogy is that the item parameter estimates of one dimension would have an effect on the estimation procedure for the items of the second dimension.

As a multidimensional measurement calibration program, ConQuest can provide theta and difficulty estimates for either a unidimensional model or a multidimensional model. Conquest allows either a comparison of the recovered ability and difficulty estimates to the original values on each dimension and a comparison of the unidimensional estimates to the expected values in a unidimensional setting. A comparison of the unidimensional estimates provided by Winsteps and the unidimensional estimates provided by ConQuest was also possible.

This dissertation not only compared the originating multidimensional difficulty and ability values to the multidimensional difficulty and ability estimates provided by ConQuest, but also the unidimensional difficulty and ability estimates provided by both ConQuest and Winsteps to what they would be expected in a strictly unidimensional model.

Appendix A contains the SPSS script and both the Winsteps and ConQuest command files used to generate the person- and item-level results data sets. These scripts were run once for each of the 25 iterations.

The descriptive statistics generated by SPSS for the first iteration are shown in Appendix B. These statistics show the NO and C distributions to be correlated at .502. For each iteration, the correlation between the NO and C distributions fell between .485 and .530.

Classical Item Analysis

Prior to initiating the research study, a psychometric review of the 1000 x 21 item-person response matrix for the first iteration was conducted to ensure the robustness of the items. Appendix C contains a brief item analysis report for these 21 items. The item difficulty varies from .12 for item RESP21 to .87 for item RESP15. The item discrimination (upper 27% - lower 27%) ranges from .21 for item RESP4 to .08 for item RESP6. Note that the low discrimination for RESP6 was most likely an artifact of the item's difficulty. The item to total score correlations (point biserial correlations) range from a high of .61 for item RESP4 to .38 for item RESP6. The internal consistency of these 21 items as measured by Cronbach's alpha was .86. The classical item analysis indicated that these 21 test items were of sufficient quality for use in this research study.

Question 1: Unidimensional Item-Level Recovery

The root mean square (RMSQ) was calculated for each item across all 25 iterations for both ConQuest and Winsteps. With the criterion of the RMSQ closest to zero (0) indicating the better recovery, ConQuest recovered item difficulty parameters more closely than Winsteps for 15 of the 21 items. A more detailed exploration of this item recovery follows in a subsequent section. Table 7 shows the root mean square for each item's difficulty parameter as estimated by ConQuest and Winsteps.

Table 7.

Item Recovery as Indicated by the Root Mean Square Fit Statistic.

	Original	Root Mean Square		Model attaining
Item	Difficulty	ConQuest	Winsteps	the tightest fit
1	-2.34	.223	.366	MC1-PL
2	-1.47	.178	.341	MC1-PL
3	-0.87	.089	.303	MC1-PL
4	0.52	.085	.192	MC1-PL
5	1.47	.143	.126	1-PL
6	1.73	.192	.150	1-PL
7	1.65	.284	.159	1-PL
8	-2.17	.123	.398	MC1-PL
9	-1.21	.070	.313	MC1-PL
10	-0.52	.079	.262	MC1-PL
11	1.04	.115	.188	MC1-PL
12	1.91	.093	.138	MC1-PL
13	2.51	.145	.122	1-PL
14	1.73	.120	.141	MC1-PL
15	-2.30	.313	.106	1-PL
16	-1.60	.229	.149	1-PL
17	-0.90	.139	.178	MC1-PL
18	0.00	.090	.213	MC1-PL
19	1.00	.164	.308	MC1-PL
20	1.50	.210	.336	MC1-PL
21	2.70	.321	.405	MC1-PL

There was no apparent relationship between the six items that were more accurately recovered by Winsteps. An ordering of the items by difficulty showed that these six items were interspersed throughout the range of –2.30 to 2.51. Items more extreme than these six as well as items more centralized were also more accurately recovered by ConQuest. In other words, the six items more accurately recovered by Winsteps appear to have been recovered at random.

A subsequent test of the recovery of the item difficulty involved the construction of a 95% confidence interval and whether or not the original difficulty value fell within this interval. Using the confidence intervals as a measure of success, the recovery of the item-level information varied greatly between the Winsteps and ConQuest programs. ConQuest was able to recover 74% of the item difficulty values compared to Winstep's 37.7% success rate.

Table 8 shows the confidence intervals for each item difficulty as estimated by ConQuest for one iteration. The confidence intervals for each item difficulty as estimated by Winsteps is shown in Table 9. If the original difficulty parameter fell within the 95% confidence interval that was calculated for each item, the program was considered to have successfully recovered that item's parameter.

Table 8.

ConQuest Item Recovery as Indicated by the 95% Confidence Interval.

Item	Original	ConQuest Estimate	Standard Error	95% Confidence Interval Upper	95% Confidence Interval Lower	Status
1	-2.34	-2.317	.102	-2.117	-2.517	Recovered
2	-1.47	-1.537	.083	-1.374	-1.700	Recovered
3	-0.87	-0.819	.073	-0.676	-0.962	Recovered
4	0.52	0.472	.070	0.609	0.335	Recovered
5	1.47	1.670	.085	1.837	1.503	Failed
6	1.73	1.655	.085	1.822	1.488	Recovered
7	1.65	1.759	.087	1.930	1.588	Recovered
8	-2.17	-2.065	.095	-1.879	-2.251	Recovered
9	-1.21	-1.188	.077	-1.037	-1.339	Recovered
10	-0.52	-0.556	.071	-0.417	-0.695	Recovered
11	1.04	0.971	.074	1.116	0.826	Recovered
12	1.91	2.002	.093	2.184	1.820	Recovered
13	2.51	2.541	.110	2.757	2.325	Recovered
14	1.73	1.744	.087	1.915	1.573	Recovered
15	-2.30	-1.935	.091	-1.757	-2.113	Failed
16	-1.60	-1.443	.081	-1.284	-1.602	Recovered
17	-0.90	-0.856	.073	-0.713	-0.999	Recovered
18	0.00	-0.025	.069	0.110	-0.160	Recovered
19	1.00	0.878	.073	1.021	0.735	Recovered
20	1.50	1.211	.077	1.362	1.060	Failed
21	2.70	2.566	.111	2.784	2.348	Recovered

Table 9.

Winsteps Item Recovery as Indicated by the 95% Confidence Interval.

				95% Confidence Interval		
		Winsteps	Standard			
Item	Original	Estimate	Error	Upper	Lower	Status
1	-2.34	-2.686	.103	-2.484	-2.888	Failed
2	-1.47	-1.793	.084	-1.628	-1.958	Failed
3	-0.87	-1.109	.076	-0.960	-1.258	Failed
4	0.52	0.274	.074	0.419	0.129	Failed
5	1.47	1.487	.087	1.658	1.316	Recovered
6	1.73	1.692	.091	1.870	1.514	Recovered
7	1.65	1.595	.089	1.769	1.421	Recovered
8	-2.17	-2.718	.104	-2.514	-2.922	Failed
9	-1.21	-1.549	.081	-1.390	-1.708	Failed
10	-0.52	-0.865	.075	-0.718	-1.012	Failed
11	1.04	0.86	.078	1.013	0.707	Failed
12	1.91	1.846	.094	2.030	1.662	Recovered
13	2.51	2.56	.115	2.785	2.335	Recovered
14	1.73	1.635	.090	1.811	1.459	Recovered
15	-2.30	-2.3	.094	-2.116	-2.484	Recovered
16	-1.60	-1.916	.086	-1.747	-2.085	Failed
17	-0.90	-1.051	.076	-0.902	-1.200	Failed
18	0.00	-0.145	.073	-0.002	-0.288	Failed
19	1.00	0.698	.077	0.849	0.547	Failed
20	1.50	1.291	.084	1.456	1.126	Failed
21	2.70	2.194	.103	2.396	1.992	Failed

For the 25 iterations used to estimate the difficulty values for the 21 items, ConQuest appropriately placed the mean value within the confidence interval 388 times out of the 525 total items estimated. This was a recovery rate of 73.9%. Winsteps, however, recovered only 198 of the 525 total items estimated over the 25 iterations. This recovery rate was 37.3%.

A regression analysis was attempted to identify the cause of this poor recovery rate. The resultant regression equation is shown in Equation 9. The r^2 for this equation was 99.3%, indicating that almost 100% of the variance was accounted for in this regression equation.

$$\hat{Z} = .209 + .956W \qquad \text{Equation 9}$$

Where:

$\hat{Z}$ = the original generating item difficulty value estimate.

W = the item difficulty parameter as estimated by Winsteps.

The regression equation was used to rescale the item difficulty parameters as estimated by Winsteps. With this rescaling, Winsteps was able to successfully recover 332 of the 525 items for a recovery rate of 63.2%.

The mean of the original difficulty parameters for the 21 items was .21. Winsteps recentered the mean to zero. If the true mean were known, the Winsteps command file could be coded to retain the original mean. In a true-life scenario, the true mean is unknown, rendering such adjustments infeasible. Even with the restoration of the true mean, Winsteps' successful recovery rate was 63% compared to ConQuest's 74%.

A regression analysis was also done on the estimates provided by ConQuest. The regression is shown in Equation 10. The r^2 for this equation is 99.0%,

$$\hat{Z} = -.0079 + 1.04Q \qquad \text{Equation 10}$$

Where:

$\hat{Z}$ = the original generating item difficulty value estimate.

Q = the item difficulty parameter as estimated by ConQuest.

The intercept was extremely close to zero, and the slope was almost one, indicating that ConQuest had already accounted for most of the variance in the model. Furthermore, the mean of the 525 difficulty estimates provided by ConQuest was .204, indicating that ConQuest placed the mean closer to the true mean difficulty as calculated by the original generating items, and estimated by the linear regression model.

Comparisons of item recovery by order of presentation and order of difficulty failed to identify any particular pattern in item recovery between the two calibration programs. A bar chart showing the number of successful recoveries for each item in order of difficulty is listed in Figure 8.

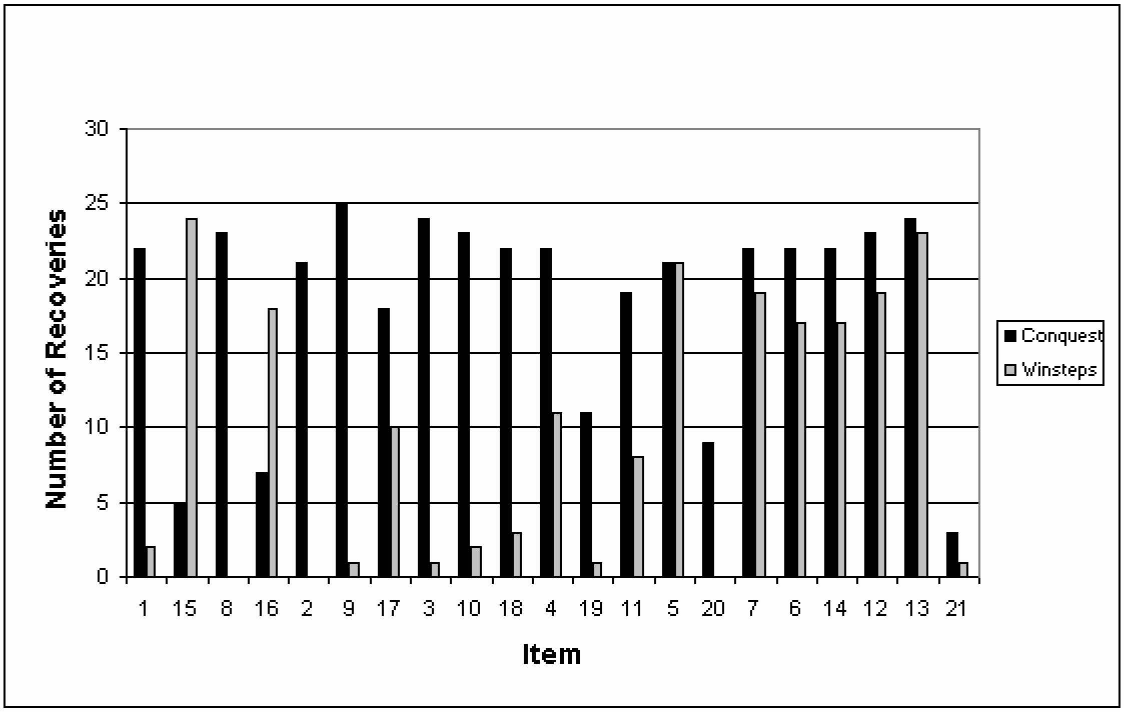

Figure 8. Number of Successful Item Recoveries Sorted by Item Difficulty.

Figure 8 shows that there is no identifiable pattern evident in the order of successful recoveries for either ConQuest or Winsteps.

This procedure of comparing recovery rates by using confidence intervals was repeated with a 68.13% confidence interval instead of the traditional 95% confidence interval. This lower level of confidence was used to create narrower bands and therefore eliminate more of the recovered estimates whose values lie further from the mean. The 68.13% confidence interval accepts only those estimates whose values are within +/- 1 standard error of the mean.

With the tighter confidence bands, ConQuest successfully recovered 218 of the 525 items for a recovery rate of 41.5%. Winsteps successfully recovered only 83 of the

525 items for a recovery rate of only 15.8%. Figure 9 shows a bar chart with the number of successful recoveries for each item in order of item difficulty.

Again, as in Figure 8, no discernible pattern is shown. A higher number of successful recoveries is interspersed with incidents of lower successful recoveries.

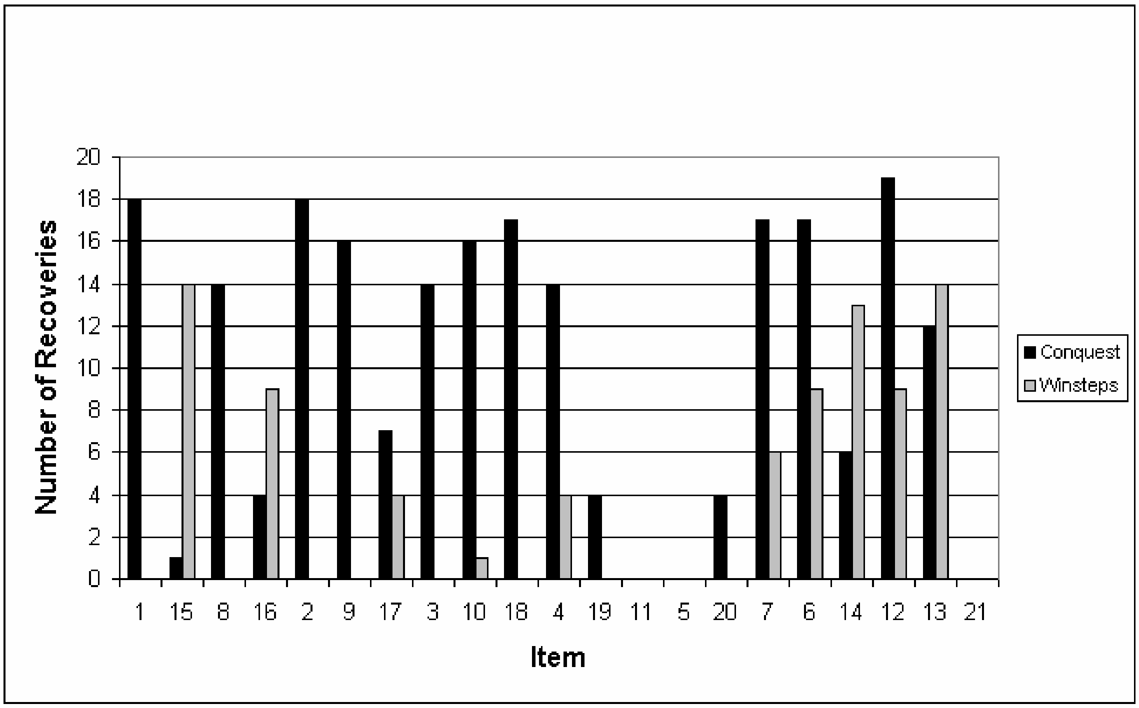

Figure 9. Number of Successful Item Recoveries Sorted by Item Difficulty (68% Confidence Interval).

ConQuest and Winsteps Unidimensional Person-Level Recovery

Because Winsteps does not estimate multidimensional theta values, only unidimensional estimates can be compared between the two calibration programs.

Whereas the item-level recovery was done with fixed item difficulties and random person theta values, the RMSQ fit statistic needed to gauge the person-level recovery requires that the person theta values be fixed across iterations. To this end, the person-item response generator was modified to retain the person theta values for all three dimensions. The correct or incorrect response to any particular item was determined by comparing the appropriate person theta value to a random uniform distribution.

The root mean square (RMSQ) was calculated for each person across all 25 iterations for both ConQuest and Winsteps. With the criterion of the RMSQ closest to zero (0), ConQuest recovered more person ability parameters than Winsteps. ConQuest had a closer RMSQ for 601 of the 1000 (or 60%) simulated people. Winsteps had a closer RMSQ for 399 (or 40%) of the 1000 people.

A 95% confidence interval for each recovered ability estimate showed that 88.4% of the Winsteps estimates fell within the interval. Only 44.2% of ConQuest's ability estimates fell within the 95% confidence interval. The apparent paradox resulting from one index indicating that one calibration program was more accurate in person-level recovery while a different index indicates that the other calibration program was the more accurate stems from the smaller standard errors reported by ConQuest. The standard error estimates for all 1000 examinees across all 25 iterations were larger for Winsteps than for ConQuest. For most cases, Winsteps' standard error estimates were nearly twice the size of ConQuest's standard error estimates. Because ConQuest calculated a smaller standard error, the confidence bands were tighter. The tighter confidence bands force more failures for persons that lie just outside the upper or lower boundaries. The RMSQ fit statistic is based on the sum of the squared differences between actual and observed ability or proficiency values and therefore not susceptible to deviations in the standard error.

ConQuest Multidimensional Person-Level Recovery

The multidimensional ability estimates provided by ConQuest were compared to the original generating values from each dimension. Because the RMSQ multidimensional estimates are not compared to another set of estimates, the RMSQ statistic in this context was meaningless. A 95% confidence interval also was calculated for each of the 1000 respondents across the 25 iterations. The average RMSQ for recovery along the Necessary Operations dimensions all 1000 respondents was .5432. The average RMSQ for recovery along the Calculations dimensions is .5613. The RMSQ for recovery along both dimensions for each of the 1000 respondents is shown in Appendix D.

Table 10 shows the number of successful recoveries on both the Necessary Operations and the Calculations dimensions for each of the 25 iterations. These successful recoveries were determined by whether or not each respondents' ability/proficiency estimates fell within a 95% confidence interval as calculated across all iterations. Of the 25,000 respondent/iteration pairs, ConQuest successfully recovered 17,075 ability/proficiency values along the Necessary Operations dimension, and 17,139 ability/proficiency values along the Calculations dimension. This amounts to a 68.3% and a 68.6% recovery rate respectively. This recovery rate was much higher than the unidimensional recovery rate of 44.2% reported earlier. Again, this low recovery rate of 44.2% was likely an artifact of the smaller standard errors reported by ConQuest.

Table 10.

ConQuest's Successful Recovery of Multidimensional Ability Values for the Necessary Operations and the Calculations Dimensions.

	Number of Successful Ability Recoveries	
Iteration	NO	C
1	733	682
2	699	727
3	721	652
4	646	657
5	588	709
6	680	698
7	631	668
8	696	611
9	642	725
10	696	695
11	715	725
12	696	673
13	698	724
14	733	713
15	710	692
16	640	687
17	657	684
18	725	710
19	696	736
20	707	709
21	662	694
22	701	696
23	665	634
24	671	658
25	667	580
Sum	17,075	17,139
Percent	68.3 %	68.6 %

n = 25,000. 25 iterations, 1000 respondents per iteration.

Issues With the Comparisons of Confidence Intervals Across Estimation Programs

A note of caution is necessary in the comparison of recovery rates across different programs based on a confidence interval. The confidence interval is dependent on the reported standard error. Winsteps reports a much larger standard error than ConQuest and therefore has a larger confidence interval within which to recover the parameter estimates. One could make an argument to apply the standard errors generated by Winsteps to the ConQuest data and vice versa. Another argument can be made to pool the standard errors and apply the pooled values to both confidence intervals. Table 11 shows this cross application of the standard errors to determine item parameter recovery rate. The person ability parameter recovery is shown in Table 12.

Table 11.

Item Parameter Recovery Rates with Standard Error Estimates Applied Across Programs.

	Source of Standard Error		
Program	Winsteps	ConQuest	Pooled
Winsteps	37.7%	35.4%	53.9%
ConQuest	75.8%	73.9%	86.9%

Table 12.

Person Parameter Recovery Rates with Standard Error Estimates Applied Across Programs.

	Source of Standard Error		
Program	Winsteps	ConQuest	Pooled
Winsteps	88.4%	47.9%	100%
ConQuest	86.0%	44.2%	100%

This exercise in applying the standard error estimates obtained from one estimation program to the person ability estimates obtained from a second estimation program shows that the pooled estimate will increase the number of successful estimate recoveries. With a swap of the standard error, both Winsteps and ConQuest increased the number of successful recoveries. The number of recoveries of item parameters by Winsteps was still marginal compared to the number recovered by ConQuest, but in recovering person parameters, there was a small difference in favor of Winsteps. This exercise is provided to demonstrate the effect of changing standard errors. Any interpretation of these results is left to the practitioner.

Answer to Question 1

Given unidimensional item-level data and multidimensional person-level data, and the RMSQ fit statistic as the standard for comparison, the MC1-PL model as utilized by ConQuest successfully recovers both the item difficulty values and the person ability values more frequently than the 1-PL model as utilized by Winsteps.

ConQuest successfully recovered 15 of the 21 original item difficulty values. Winsteps' successfully recovered only six of the 21 original item difficulty values. ConQuest successfully recovered 601 (60%) of the 1000 examinee ability values as compared to Winsteps' recovery of 399 (40%) examinee ability values in an attempt at recovery along a composite dimension.

ConQuest successfully recovered 68.3% of the person ability values along the Necessary Operations dimension and 68.6% of the person ability values along the Calculations dimension. The standard for comparison was a 95% confidence interval using the smaller standard errors reported by ConQuest.

A note by Ben Wright is in needed to place these comparisons in perspective. Linacre (2004) cites Wright and Douglas (1976) in that random discrepancies in calibration as large as .5 logits have negligible effects on measurement. Wright and Douglas qualified this statement with the requirement that the test length be greater than 20 items. If credence is given this statement, all of the parameters estimates for both item difficulty and person ability for both Winsteps and ConQuest fall within +/- .5 logits of the original generating values.

Although such a statement may have merit, the closer to the original parameter the estimate arrives, the more precise the measurement instrument will be.

Practical Considerations Stemming from Question 1

Many, if not most measurement practitioners ignore the possibility of multidimensionality in the person-level ability values. Unidimensionality is generally considered to be an artifact of the assessment item, not the respondent. Respondents of varying abilities along different traits is a more frequent phenomenon than many items of varying difficulties along different scales.

Question 1 has shown that the MC1-PL model can recapture not only the original item difficulty values that lie on a unidimensional scale, but also can recapture the underlying multidimensional person ability values.

The implication of this recovery of person abilities on multiple dimensions is that a properly designed assessment can accurately measure multiple traits on multiple dimensions and report the person theta values on multiple scales far more efficiently and far more precisely than can a unidimensional measurement model.

An analogy in statistics is the use of an independent samples *t*-test to measure differences between two groups and a factorial analysis of variance (ANOVA) to measure for many more differences between multiple groups as well as possible interaction effects. Just as the factorial ANOVA leads to increased precision in more complex statistical tests, the MC1-PL model leads to improved precision over the 1-PL model in the measurement of multiple traits across multiple correlated dimensions.

Question 2

Question 2 asked how closely the MC1-PL model can recover the true generating values of simulated items with construct-relevant multidimensionality. The intent of this research question was to recover the many multidimensional difficulty values for items containing within-item multidimensionality. On the surface, this appears to be feasible. However, after further in-depth probing, current implementations of MIRT are capable of reporting only an average item difficulty value that represents an aggregation of the separate difficulty estimates on each dimension. The aggregation is such that individual difficulty estimates cannot be extracted. An additional observation was that ConQuest can recover multiple theta estimates for person abilities across multiple dimensions, but cannot do the same for item difficulties.

Further communications with both Dr. Mark Wilson and Dr. Terry Ackerman provided additional insights on this intriguing problem. The response surface between the two target dimensions is simply a representation on a flat plane created by the shadows of multiple vectors that are projected through latent space. Each vector not only lacks a common point of origin, but may not even intersect in latent space. Furthermore, the units of measurement are different from one vector to another, creating problems in estimating the anchor points on each respective scale.

The aggregate difficulty value reported by ConQuest represents the within-item multidimensional difficulty estimate which the current model cannot decompose into separate values for each correlated dimension. As such, the answer to question 2 was "No, given current available MIRT programs, the MC1-PL model cannot recover the true generating values of simulated items with construct-relevant multidimensionality." Such an accomplishment will belong to future theoretical and empirical software implementations with greater measurement precision and more efficient estimation algorithms.

Question 3 Results

Question 3 is "By applying the Rasch model to these multidimensional items to get a single summary scale, will the resultant model show increasing misfit for those items that lie further from the intersection of the two dimensions than those items that fall closer to the intersection?"

One commonly accepted method of determining misfit is an analysis of the standard error residuals: the difference between expected and observed SE values for each item. To properly place this analysis in an appropriate context, an analysis of the Winsteps unidimensional fit statistics was necessary.

Winsteps provides two fit statistics as gauges of the appropriateness of a measurement model. Both fit statistics utilize the mean square error with an expected value of 1.0. The first is the MNSQ Infit statistic which is more sensitive to unexpected variations in items near each respondent's ability level. The second is the MNSQ Outfit statistic which is more sensitive to unexpected responses to items further from the respondent's ability level. Values for either fit statistic that are greater than 1.0 indicate random noise. Values for either statistic that are less than 1.0 indicate identifiable dependencies in the data. Table 13 shows Linacre's (2004) guide to interpreting both the infit and outfit MNSQ statistics:

Increasing Misfit as Determined by the Infit MNSQ Statistic

Both the infit and outfit MNSQ fit statistics were examined for increasing misfit across items. The infit MNSQ is shown in Table 14, along with the item difficulty and originating dimension. The table is sorted by the infit MNSQ statistic. Sorting by the MNSQ statistic shows that the items that fell on the Z vector all had an infit MNSQ statistic of less than 1.0. These seven items had the MNSQ statistic clustered between .87 and .94. All remaining items (those that originated on either the Necessary Operations

Table 13.

Linacre's (2004) Guide to Interpreting the Winsteps Infit and Outfit MNSQ Fit Statistics

Value	Meaning
> 2.0	Off-variable noise is greater than useful information. Degrades measurement.
> 1.5	Noticeable off-variable noise. Neither constructs nor degrades measurement.
0.5 – 1.5	Productive of measurement.
< 0.5	Overly predictive. Misleads us into thinking we are measuring better than we really are.

or the Calculations dimensions) had an infit MNSQ statistic greater than 1.0. The MNSQ statistics for these 14 items were clustered between 1.03 and 1.06. All of these values were well within the boundaries specified by Linacre for productive measurement. Although these values fell within the specified boundaries, an important note is that all seven of the composite items fell on the side that is considered to contain some dependencies in the data and the remaining 14 multidimensional items fell on the side that is considered to contain off-variable noise. This phenomenon was sustained across

Table 14.

Item Difficulty and the Infit MNSQ for 21 Items, Sorted by MNSQ.

Item	Originating Dimension	Item Difficulty	Infit MNSQ
19	Z	1.00	0.87
18	Z	0.00	0.88
16	Z	-1.60	0.90
17	Z	-0.90	0.90
20	Z	1.50	0.90
15	Z	-2.30	0.92
21	Z	2.70	0.93
6	NO	1.73	1.03
7	NO	1.65	1.03
5	NO	1.47	1.04
12	C	1.91	1.04
14	C	1.73	1.04
1	NO	-2.34	1.05
2	NO	-1.47	1.05
4	NO	0.52	1.05
8	C	-2.17	1.05
11	C	1.04	1.05
13	C	2.51	1.05
9	C	-1.21	1.06
3	NO	-0.87	1.07
10	C	-0.52	1.07

all 25 iterations. Although a unidimensional measurement tool, Winsteps appears to be segregating the items by their inherent multidimensionality: items that fell on the composite dimension were separate from items that fell on either of the two remaining dimensions.

Figure 10 shows the average item infit MNSQ for all 21 items across all 25 iterations. As previously noted, all items fell within the 0.5 and 1.5 range specified by Linacre. These items were considered to be productive to measurement. The seven items that fell noticeably below the other 14 items were the seven items that lie on the composite (Z) dimension.

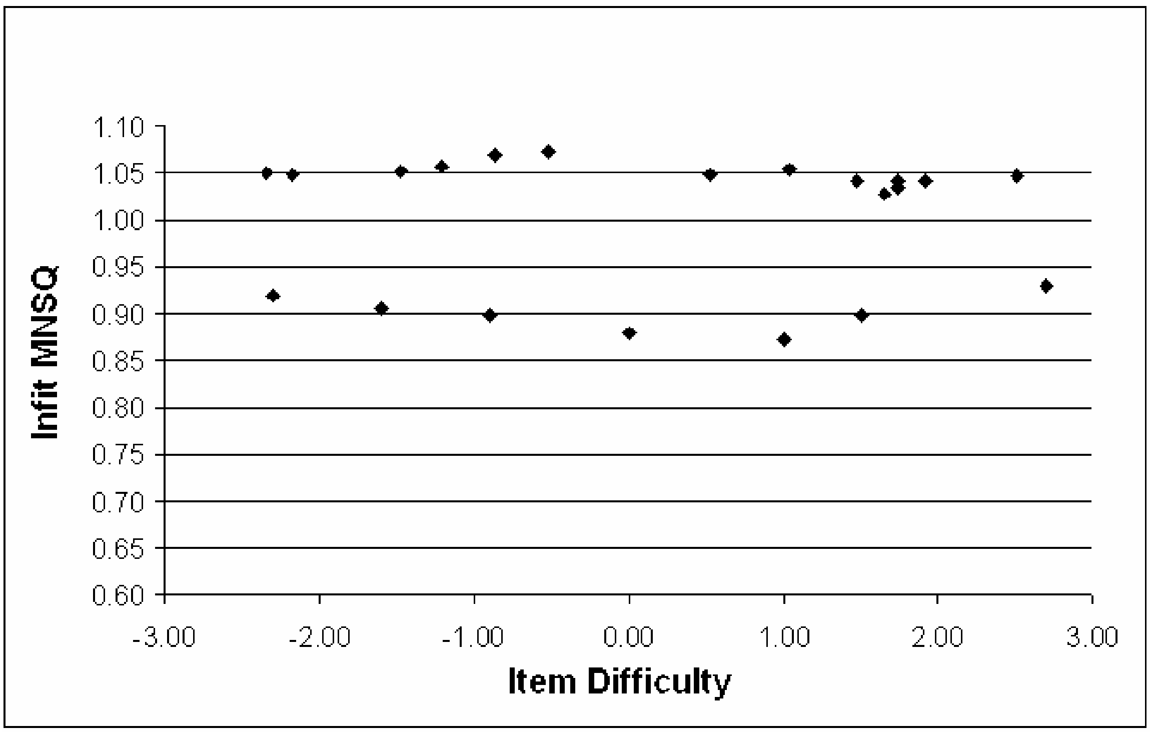

Figure 10. Determining Item Misfit: Inflation to the Infit MNSQ.

The scatter plot of the 21 MNSQ infit statistics does not show any noticeable change for items with the more extreme difficulty values as compared to items that are of a more moderate difficulty. If there were an increase in the misfit, items with extreme difficulty values would be expected to show more variation away from the centered placement than is observed in Figure 10.

Increasing Misfit as Determined by the Outfit MNSQ Statistic

The second fit statistic, the outfit MNSQ shows more segregation between items than the infit MNSQ statistic. The outfit MNSQ is sensitive to respondents' answers to items far from the person's ability level. Table 15 shows the average outfit MNSQ for all 21 items across 25 iterations. As with the infit MNSQ, the outfit MNSQ separates the seven items on the composite vector from the fourteen items that fell on either of the two primary dimensions. Again, as with the infit MNSQ, all values for the outfit MNSQ fell within the boundaries specified by Linacre.

The relationship between the item difficulty and the change in the outfit MNSQ statistic becomes apparent with a scatter plot. Figure 11 shows this relationship.

The seven items that fell below 1.0 were the seven items originating on the composite vector. The 14 items that fell above 1.0 were the 14 items that originated on one of the two primary dimensions. The item difficulty range for the seven composite items was from .79 to .85. The item difficulty range for the 14 NO and C items was from 1.08 to 1.27. Figure 11 shows that the distortion in the MNSQ increases for items that fell further from the origin.

Table 15.

Item Difficulty and the Outfit MNSQ for 21 Items, Sorted by MNSQ.

Item	Originating Dimension	Item Difficulty	Outfit MNSQ
15	Z	-2.30	0.79
21	Z	2.70	0.79
19	Z	1.00	0.81
20	Z	1.50	0.81
16	Z	-1.60	0.83
18	Z	0.00	0.83
17	Z	-0.90	0.85
4	NO	0.52	1.08
11	C	1.04	1.09
10	C	-0.52	1.11
5	NO	1.47	1.12
7	NO	1.65	1.12
9	C	-1.21	1.12
2	NO	-1.47	1.13
3	NO	-0.87	1.13
6	NO	1.73	1.13
14	C	1.73	1.15
12	C	1.91	1.16
8	C	-2.17	1.18
1	NO	-2.34	1.20
13	C	2.51	1.27

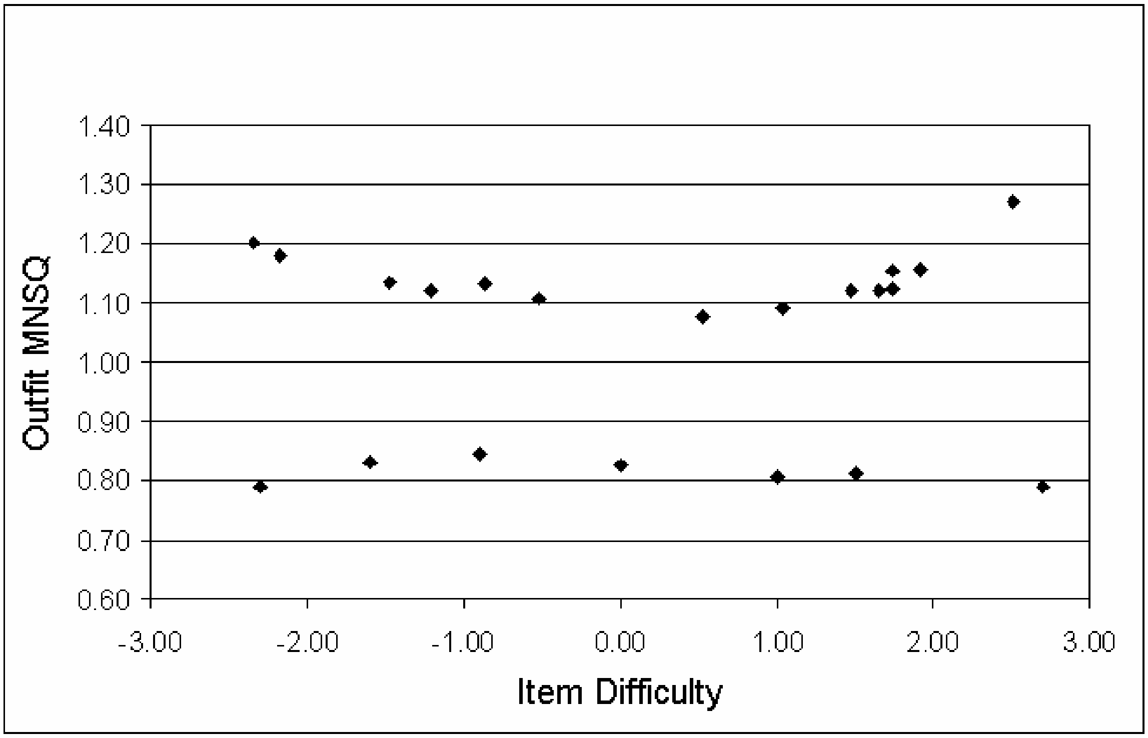

Figure 11. Determining Item Misfit: Inflation to the Outfit MNSQ.

Although both the infit and outfit mean square fit statistics fell within the range specified by Linacre, attention must be brought to this dissertation's primary focus: that of construct-related multidimensionality. With a correlation of .50 between primary dimensions, the MNSQ did not exceed the boundaries specified by Linacre. In the context of construct-irrelevant multidimensionality in which the correlation between dimensions is less than .50, the outfit statistic may show greater distortion for items that lie further from the point of origin. Such hypothesis and confirmation is beyond the scope of this dissertation and remains a question to be answered by future research.

Another possibility is perhaps the variation at the tails of the distribution was due to a lack of sufficient items at the upper and lower tails of the item difficulty scale. Test practitioners will generally author more items within one logit above or below zero (0). This region takes in two-thirds of the test respondents. Much fewer items are targeted at the region between one and two logits beyond zero and even fewer items are targeted between two and three logits beyond zero. Such an item authoring strategy creates an item bank that targets the ability distribution of the target respondent population.

Distortion to the Standard Error

Allusions to distortions to the standard error were noted in the answer to research question 1. Specifically, the standard errors that were calculated by Winsteps were shown to be considerably larger than the standard errors that were calculated by ConQuest. A consequence of this larger standard error was the false recovery of several item difficulty values and person ability values. An examination of the standard error estimates for both items and people shows that for entities that fell further from the point of origin, the size of the standard error increased. This is a common psychometric phenomenon and is to be expected. A plot of the standard errors against the item difficulty estimates invariably yields a parabolic pattern somewhat in the shape of the letter “u”. The measure of misfit is whether or not the increase to the standard error falls within or without an expected range.

Figure 12 shows the distortion to the standard error for items that lie further from the point of origin. The values used in Figure 12 come from one of the twenty-five iterations.

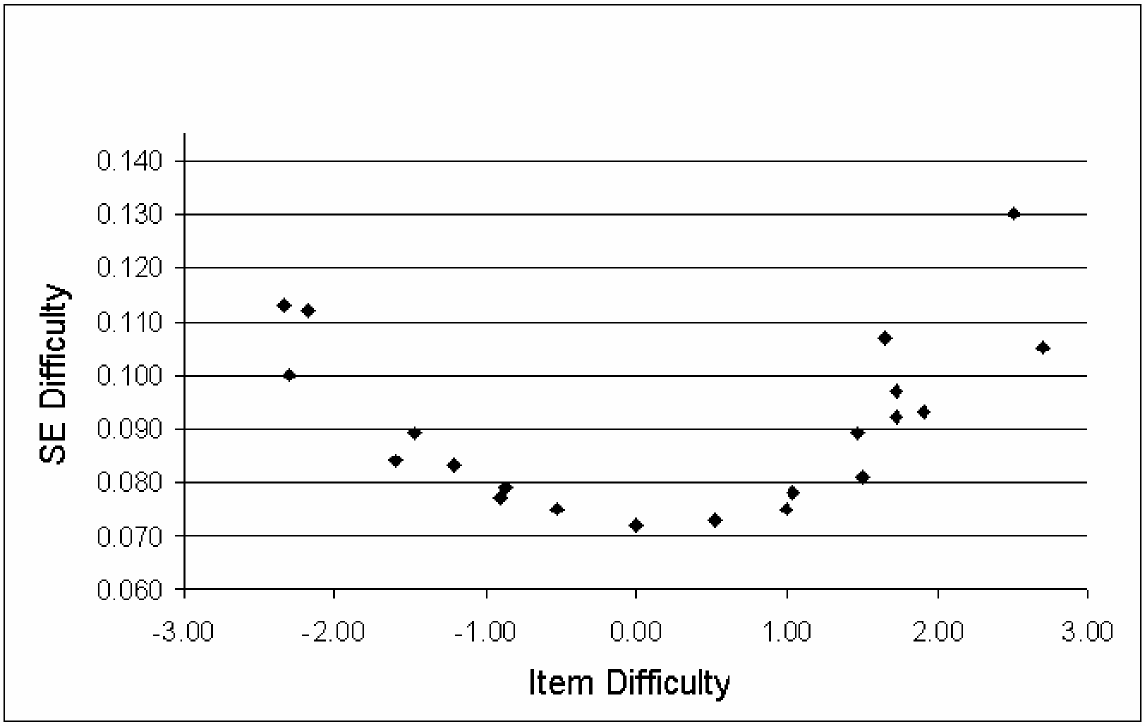

Figure 12. Inflation to the Standard Error for Items With Difficulty Values Further from the Origin.

Although more complex, Figure 12 provides additional information as to the distortion of the standard error. Figure 12 shows the range of distortion for each item across 25 iterations. The high, low, and mid-points are shown for each item. Each dot in Figure 13 indicates the mean point for the standard error for each item. The tick marks above and below each dot indicates respectively the high and low points for the standard error for each item. These are the observed values across 25 iterations.

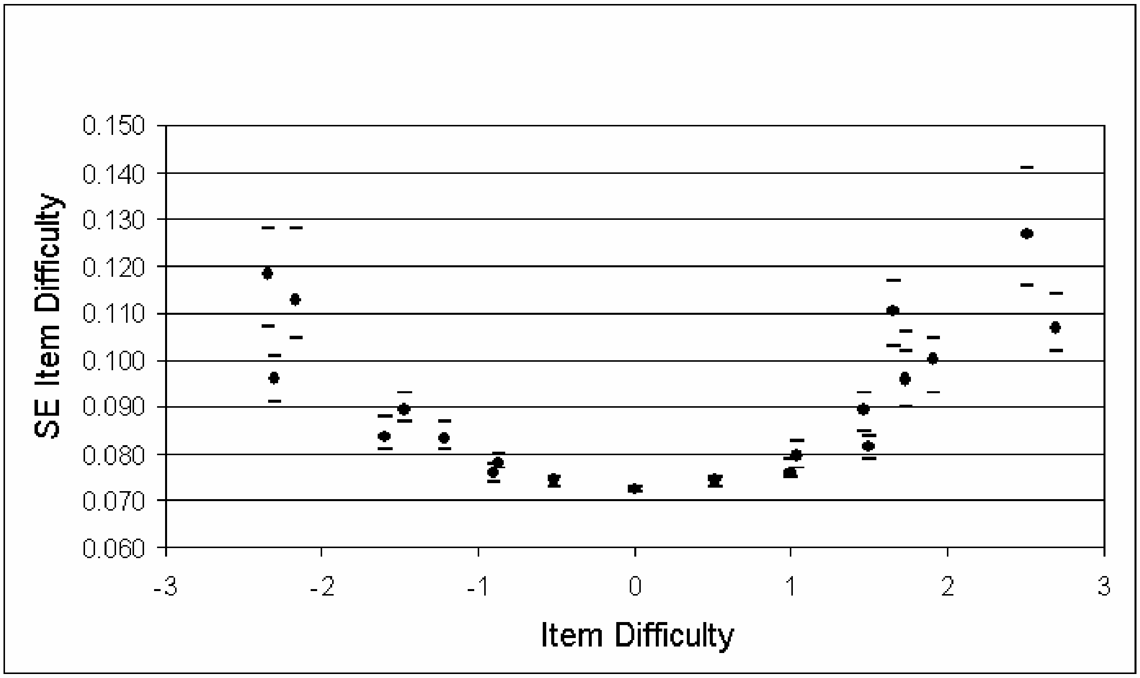

Figure 13. Inflation to the Standard Error Across 25 Iterations.

Not only does the standard error increase for items with difficulty values more extreme than one logit from the mean, but the amount of variation within this distortion also increases the further from the mean the difficulty values happen to lie.

A regression analysis on the standard errors yielded a regression equation that is shown in Equation 11. The equation has an adjusted r^2 of 78.3%.

$$Y = .0734 + .000434X + .00673X^2 \quad \text{Equation 11}$$

Where:

Y = The Standard Error Estimate.

X = The item difficulty.

A plot of the fitted regression equation is shown in Figure 14.

The most precise indicator of increasing misfit is an analysis of the SE residuals: a comparison between the expected and observed standard error estimates. If the item difficulty values did not show increasing misfit, a plot of the expected versus fitted values

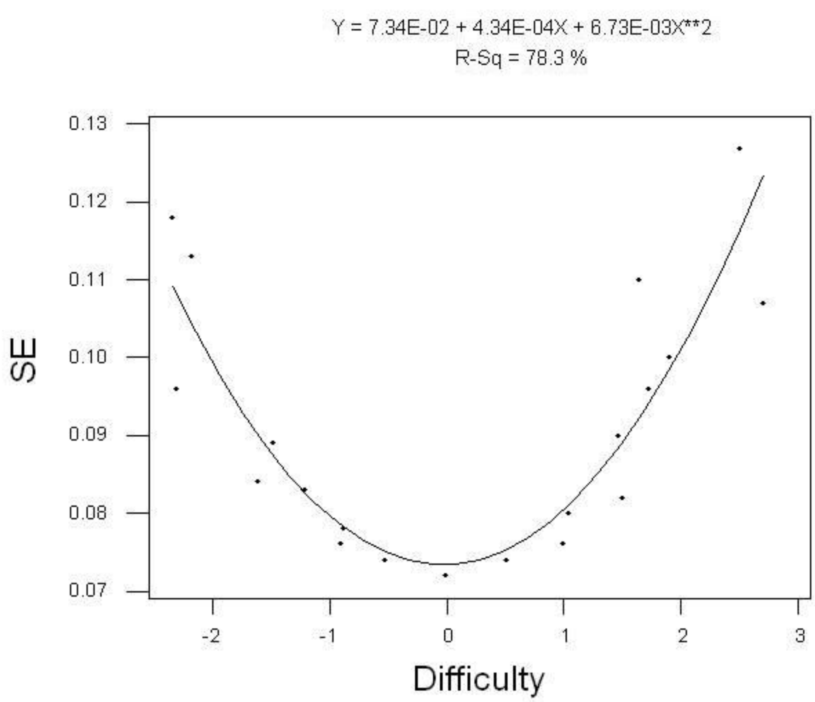

Figure 14. Regression Plot of the Item Difficulty Standard Errors.

should align in approximately a 45° angle. A plot of these values is shown in Figure 15. standard errors for approximately 15 to 18 of these 21 items fell on or near a 45° line. The remaining three to six items have standard errors that fell beyond what would be expected. These items merit further exploration.

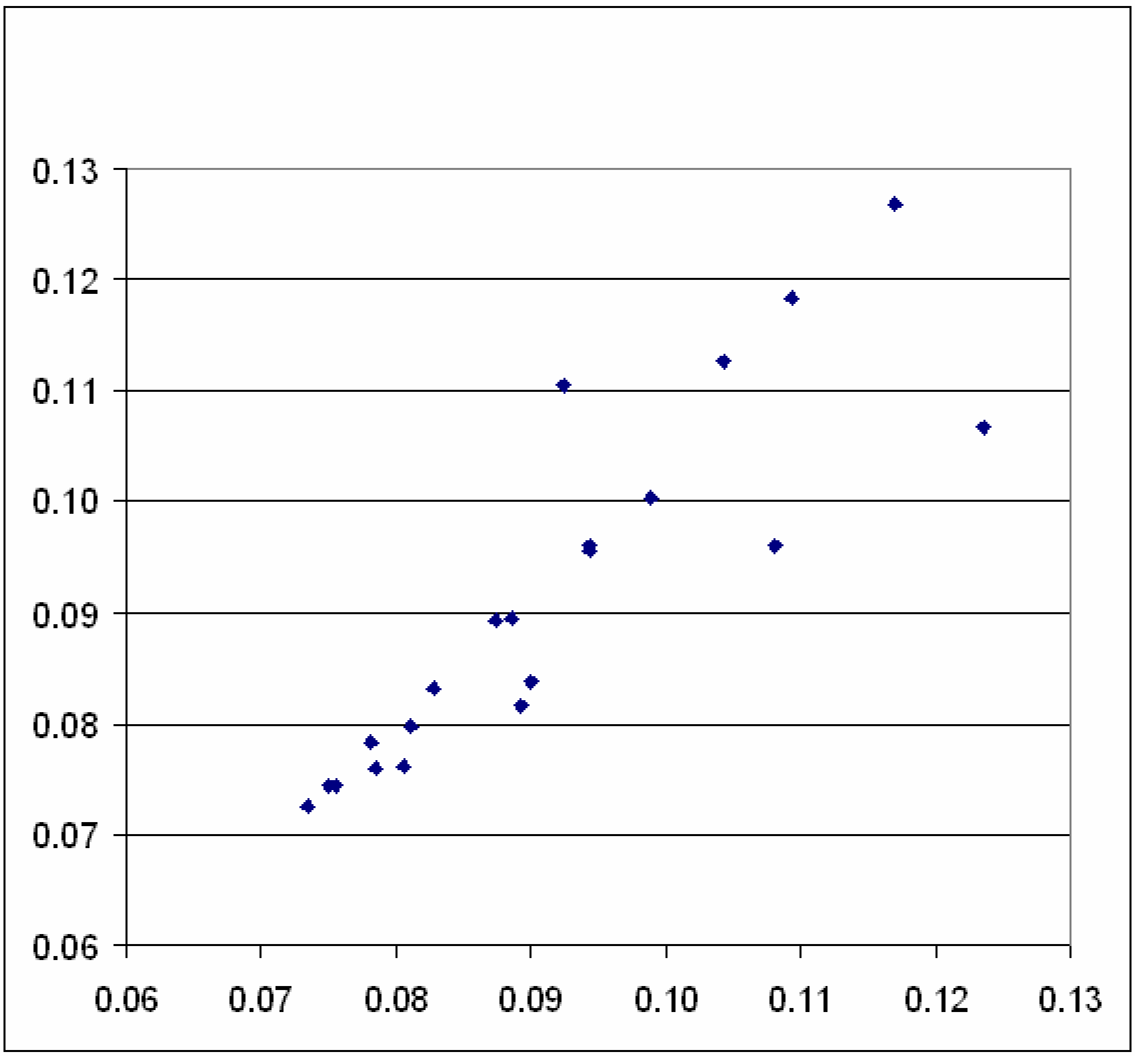

Figure 15. Observed vs. Expected SE Values for 21 Items.

A scatterplot of the SE residuals against the original item difficulty values shows which items exhibit the most misfit. If there were no discernible pattern to the degree of misfit, then the scatter-plot would also show no discernible pattern. If there were a pattern to the degree of misfit, the pattern would be exhibited in the scatter-plot. Table 16 contains the item difficulty values and the accompanying SE residual. This table is sorted by item difficulty, not by order of item presentation.

A cursory glance indicates that the more extreme residual values were associated with item difficulty values that were further from the point of origin. Furthermore, the largest residual values were associated with items that fell along the composite dimension. Generally, residuals are generally considered small if they are less than 0.01. Those items with residuals than 0.01 were those items whose item difficulties were at the extreme ranges of the scale. Figure 16 shows a plot of the standard error residuals against the generating item difficulty value. To facilitate understanding, items have been marked according to their originating dimension.

Table 16.

Item Difficulty Standard Error Residuals, Sorted by Residual.

Item ID	Originating Dimension	Difficulty	Residual
7	NO	1.65	-.0180
13	C	2.51	-.0099
1	NO	-2.34	-.0090
8	C	-2.17	-.0085
2	NO	-1.47	-.0020
14	C	1.73	-.0017
12	C	1.91	-.0016
6	NO	1.73	-.0014
5	NO	1.47	-.0009
9	C	-1.21	-.0004
3	NO	-0.87	.0000
10	C	-0.52	.0006
18	Z	0.00	.0009
4	NO	0.52	.0011
11	C	1.04	.0015
17	Z	-0.90	.0025
19	Z	1.00	.0044
16	Z	-1.60	.0063
20	Z	1.50	.0076
15	Z	-2.30	.0120
21	Z	2.70	.0170

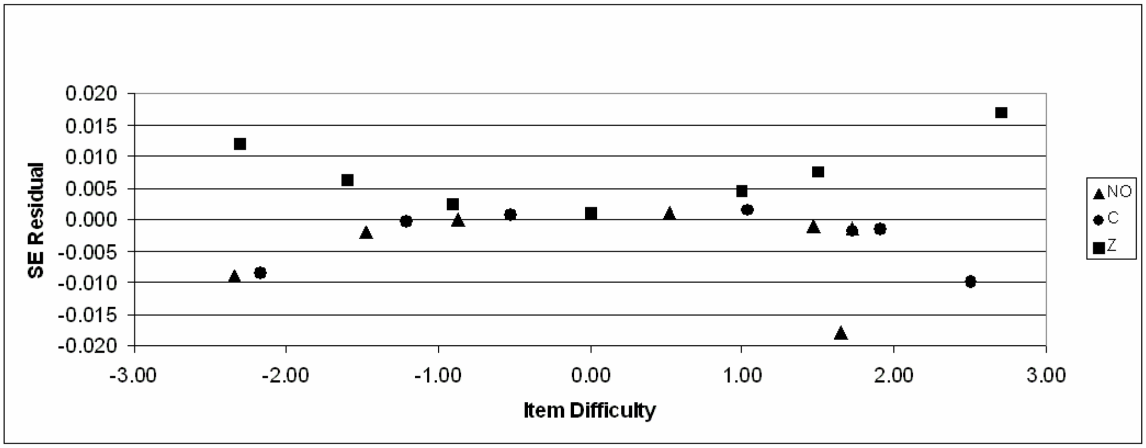

Figure 16. Standard Error Residuals vs. Item Difficulty.

Figure 16 shows that items within one logit of zero exhibit very little misfit. Items more extreme than one logit exhibit increasing misfit. Items that originated along the composite dimension had positive residuals, indicating that the expected value was greater than the observed value. Items that originated along either the Necessary Operations or the Calculations dimensions tended to have negative residuals, indicating that the observed value was greater than the expected value. This trend remained fairly consistent across all iterations.

Distortion to the Standard Error With Unidimensional Data

The focus of question 3 is whether or not the multidimensional data applied to the 1-PL model results in increasing misfit. The answer to this question is an apparent “yes.” One final consideration to question 3 is a comparison of the misfit due to multidimensional data applied to a unidimensional model and the misfit that normally occurs with unidimensional data applied to a unidimensional model. If the distortion to the standard error of multidimensional data is greater than the distortion to the standard error of unidimensional data, the unequivocal answer must be yes, that a unidimensional IRT model shows greater misfit when known multidimensional data is applied to that model.

To finish question 3, the SPSS item/person response generator was modified to provide answers to 21 hypothetical unidimensional items that required a unitary ability that aligns with these 21 items. The same process used for the multidimensional data was used for the unidimensional data. The standard errors reported by both processes were compared with an independent samples t-test. A 95% confidence interval for the difference in means was (.0877, .0923). Because zero is not within the interval, the *t*-test was significant. The p-value for this test was zero up to four decimal places, indicating that if the means between standard errors were the same, only one time in ten thousand iterations would result in more disparate standard errors than was encountered in this study.

This final test can be taken as evidence that the distortion to the standard errors of multidimensional data applied to a unidimensional model was different than the distortion to the standard errors of unidimensional data applied to the same unidimensional model.

Answer to Question 3

The application of the Rasch model to multidimensional item-level data to obtain a single summary scale results in a model with increasing misfit for items that lie further from the intersection of the two dimensions than for those items that fell closer to the intersection.

The answer to question 3 was determined by an analysis of the standard error residuals. A prior analysis of the infit and outfit statistics as reported by Winsteps failed to identify any appreciable misfit. All fit statistics provided by Winsteps showed that the model appropriately fit the data.

Practical Considerations Stemming from Question 3

Perhaps the most important implication to rise from question 3 is the realization that the fit statistics reported by an IRT calibration program designed to model unidimensional data indicate a properly-fitting model although the underlying data were not unidimensional. An analogy would be trusting that the gauges on your automobile were reporting safe fluid levels when in fact the car is running with no oil and is almost out of gasoline. Such a degree of misplaced trust could be catastrophic not only to the vehicle but also possibly to the passengers riding inside. However, the IRT calibration can be robust with regard to departures from the target assumptions.

Prior to the utilization of a unidimensional measurement model, appropriate measures must be followed to ensure that the data are truly unidimensional. Failure to observe this precaution will not be flagged by the unidimensional fit statistics.

Question 4 Results

Question 4 asked if the size of the discrimination parameter would increase for items that lie off the second factor when calibrated one at a time onto the second factor.

The rationale behind this question is that as each value used to plot the item characteristic curve is subjected to an orthogonal projection from the originating dimension to a second dimension, the resultant geometric shape appears to draw the inflection points inward towards the value of the difficulty parameter. The logical extension of this observation is with tighter inflection points the slope should be steeper and therefore the discrimination parameter should be larger.

Pilot Study for Question 4

To estimate the number of iterations needed for stable results, a pilot run was conducted. The results of this pilot run are reported first, followed by the power analysis and results of the subsequent iterations.

Discrimination parameter estimates for seven items on both the necessary operations and the calculations dimensions were needed. The original design used for research questions 1 through 3 contained seven items on each of the NO and C dimensions. To provide stable item parameter estimates, the number of items on both these dimensions was increased to 21. An initial calibration of these 21 items on the Necessary Operations dimension provided a stable framework upon which the experimental seven items could be calibrated.

Each iteration consists of a series of calibrations. Table 17 summarizes the number of calibration runs for each iteration.

Table 18 shows the discrimination parameter estimates for these seven experimental items on both the Necessary Operations and the Calculations dimensions. These estimates come from the initial pilot iteration.

The change in the discrimination (a) parameter estimates were largest for item 7. Item 7 had a positive change in discrimination of .07658. The interpretation is that the item becomes a more discriminating measure of ability on dimensions other than the original dimension. Item 5 also experienced a positive change in the discrimination parameter estimates. Items 1, 2, 3, 4, and 6 all experienced a reduction in the discrimination parameter after projection from the original NO dimension to the C dimension.

A scatter plot showing the change in the discrimination estimates for each of these seven items is shown in Figure 17. If the discrimination parameter did not shift during the projection from the NO to the C dimension, all values would fall on the 45° diagonal line.

Table 17.

Number of Calibration Runs per Iteration.

Items Within Each Calibration Run	Total Number of Runs
21 Necessary Operations	1
21 Calculations	1
21 Necessary Operations + NO_i	7
21 Calculations + Cp_i	7
Total Number of Calibration Runs per Iteration:	16

Where i = 1 – 7

NO_i = Each of the seven experimental NO items prior to projection onto the C dimension.

Cp_i = Each of the seven experimental NO items after projection onto the C dimension.

Table 18.

Change in the a *parameter for Multidimensional Items Projected from the NO to the C Dimension.*

	NO		C			
Item	*a* Parameter	SE	*a* Parameter	SE	Change in *a*	Change in SE
1	.67027	.09567	.60364	.06305	-.06663	-.16230
2	.67159	.06936	.62246	.05972	-.04913	-.11849
3	.54398	.05508	.50348	.05230	-.04050	-.09558
4	.55988	.05716	.52712	.05397	-.03276	-.08992
5	.51905	.06495	.57714	.06064	.05809	-.00686
6	.56754	.06860	.56284	.05973	-.00470	-.07330
7	.55378	.07841	.63036	.06903	.07658	-.00183

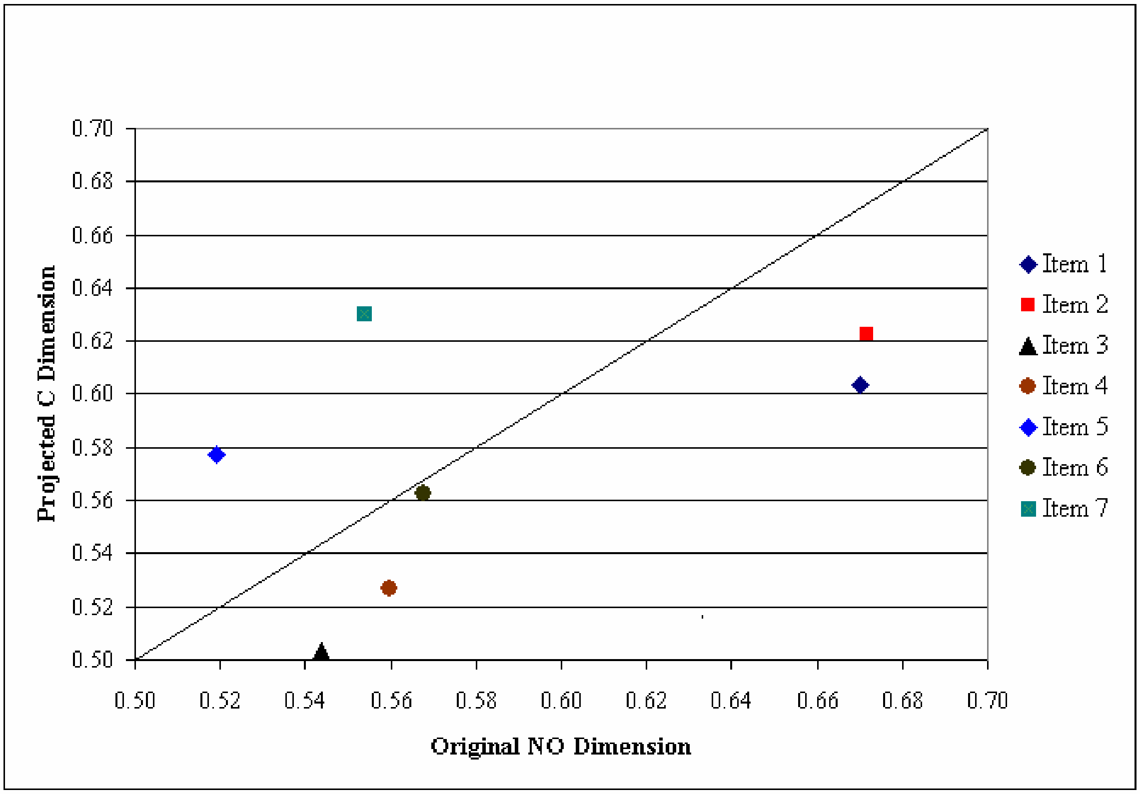

Figure 17. *a*-Parameter estimates for seven items projected from the NO to the C dimension.

The standard error for each item's discrimination parameter decreased slightly. One possible interpretation is that the estimate of the discrimination parameter becomes more accurate when the item measures a dimension other than the original intended dimension.

An important note is that these measurements reflect one iteration of seven items that were projected from one dimension to a second dimension. Subsequent iterations may yield different results.

Power Analysis to Determine the Appropriate Number of Iterations

The results of this pilot iteration and a subsequent second iteration were used to conduct a power analysis. The power analysis determined the appropriate number of iterations needed to achieve stable results. Table 19 shows that 17 iterations are required to achieve stable estimates of the discrimination parameter. These estimates were calculated using α = .05 and power = .80.

Table 19.

Number of Iterations Needed to Achieve Stable Parameter Estimates.

a-Parameter Estimate	Value
Iteration 1 Mean	0.5761
Iteration 2 Mean	0.6139
Difference in Means	0.0378
St dev.	0.0522
Iterations*:	17

*α = .05, power = .80

For simplicity in minor calculations, 20 iterations were performed. The results of these iterations are summarized in Table 20. These results are the averaged values for both the discrimination parameter and the standard error for the discrimination parameter across all 20 iterations.

Change in the Discrimination Parameter after Projection from the NO to the C Dimension

As was found in the pilot data, the discrimination parameter increased for some items, and decreased for other items. One possible interpretation is that an item's discriminating power will shift as the item measures constructs on correlated dimensions. The magnitude and direction of the shift is not clear and varies from item to item. A regression analysis on the original and projected values on both dimensions showed an adjusted R^2 of only 0.3%, indicating little or no linear relationship between the two variables.

One observation in Table 20 is the magnitude and direction of the change in the standard errors for the discrimination parameter for each item. The standard error for each item was reduced as a result of the projection from one dimension to the other. Item four experienced the smallest change in the standard error of only –3.21%. The change in the standard error for the first item was -27.45%. One interpretation is that the precision of the estimated discrimination parameter is increased by as a result of using the item to measure ability on a different dimension.

Table 20.

Average Change in the a *parameter for Multidimensional Items Projected from the NO to the C Dimension across 20 Iterations.*

	NO		C			
Item	*a* Parameter	SE	*a* Parameter	SE	Change in *a*	Change in SE
1	.5947	.0853	.5943	.0619	-.0004	-.0234
2	.5818	.0657	.6091	.0600	.0273	-.0056
3	.5729	.0588	.5711	.0567	-.0018	-.0021
4	.5858	.0574	.5583	.0556	-.0274	-.0018
5	.5758	.0654	.5835	.0599	.0077	-.0064
6	.6050	.0730	.6052	.0602	.0002	-.0128
7	.5831	.0779	.5808	.0628	-.0023	-.0152

Change in Each Item's Difficulty after Projection from the NO to the C Dimension

As a verification of the measurement process, a similar comparison was made with the difficulty parameter for each of the seven projected items. Table 21 summarizes these estimates.

Table 21.

Average Change in the Difficulty Parameter for Multidimensional Items Projected from the NO to the C Dimension across 20 Iterations.

	NO		C			
Item	*b* Parameter	SE	*b* Parameter	SE	Change in *b*	Change in SE
1	-2.7376	.3195	-1.3168	.1296	1.4208	-.1900
2	-1.7573	.1738	-0.8192	.0925	0.9382	-.0812
3	-1.0497	.1114	-0.5079	.0833	0.5419	-.0281
4	0.6126	.0863	0.3037	.0790	-0.3090	-.0073
5	1.7751	.1754	0.8609	.0989	-0.9142	-.0765
6	2.0622	.2032	1.0064	.1039	-1.0558	-.0994
7	2.5569	.2841	1.3506	.1368	-1.2063	-.1472

The most important observation stemming from Table 21 is that the percent change in difficulty is centered around 50%. This is the value predicted by the trigonometric projections. Such a finding is indicative that the fundamental hypothesis of the research question is sound. Another interesting observation is that the standard error for each item also decreased. Item four experienced the smallest change to the standard error. Item one experienced the greatest change to the standard error.

Practical Considerations Stemming from Question 4

Question 4 has shown that the discrimination parameter estimates shifted for items that were calibrated on a dimension other than the original target dimension. The discrimination parameter indicates an item's usefulness in separating different respondents into different ability levels. The shift in the discrimination parameter estimate showed that the item is more or less useful in separating each respondent into appropriate ability levels.

Although the standard error was smaller for items on the calculations dimension, the standard error is simply a more-precise indicator of a less-precise discrimination parameter. The practitioner must decide whether the potential loss in item discrimination is a worthwhile sacrifice to obtain the greater precision in measurement.

CHAPTER 5

CONCLUSIONS AND RECOMMENDATIONS FOR FUTURE RESEARCH

Each of these four research questions presented a unique set of challenges. Some of these challenges were seen early on in the research phase. Other challenges did not appear until late in the analysis phase. Each of these challenges presented opportunities for additional research or for more in-depth thought, analysis and discovery.

Research question 1 shows that a multidimensional measurement model can more accurately measure the multidimensional latent traits of respondents with varying abilities than can a unidimensional measurement model. Furthermore, a multidimensional IRT model and calibration program is more precise in the recovery of unidimensional item difficulties than is a unidimensional model and calibration program.

The principle focus of this dissertation was on the reconstruction of underlying multiple latent trait structure of an assessment using the two measurement models. The multidimensional compensatory one-parameter logistic model as implemented in ACER ConQuest can more accurately recover the underlying structure of not only the assessment itself, but also the underlying latent traits, abilities, or proficiencies of the respondents.

The intent of research question 2 was to simultaneously recover the item difficulty parameters on two separate dimensions. ConQuest provides only one item difficulty parameter. This difficulty parameter represents the within-item multidimensional difficulty estimate. The MC1-PL model as currently interpreted by ConQuest cannot provide more than one difficulty parameter per item.

Research question 3 targeted the all-too-frequent assumption that a unidimensional measurement model can adequately estimate item difficulty parameters even if the items themselves were multidimensional. The results indicate that such an assumption can lead not only to incorrect estimates, but also to incorrect fit statistics. The reliance on these fit statistics can result in diminished measurement precision and an increase in false passes or false fails for the respondent population.

Question 4 was intended to determine whether or not the discrimination parameter for items known to measure ability on one scale shifted as the ability measurement shifted to another correlated scale. A quick look at the shift in the difficulty parameter during an orthogonal projection from one dimension to another shows that the difficulty parameter is related to the correlation of the two dimensions. For the 21 items observed in this Monte Carlo study, the *a*-parameter showed little or no relationship to the correlation of the two dimensions. The original discrimination parameter estimates and the discrimination estimates for the projected items were correlated at –.055. In most instances, the discrimination parameter became smaller, indicating that the item becomes less discriminating as a measure on any dimension other than the intended dimension.

This finding that an item becomes less discriminating in a multidimensional environment strengthens the statements made by Luecht, Ackerman, and Stout regarding essential unidimensionality and the creation and reporting of separate construct-linked scales to measure each dimension separately. An example of such an implementation is the Armed Services Vocational Aptitude Battery (ASVAB) in which Physics and Chemistry were measured along one dimension and General Science and Biology were measured along a second dimension. In light of the multidimensionality of these construct scales, the researcher decided to use two concurrent unidimensional scales and then combine the scores across the dimensions at the end of the assessment.

Completed Statement of Purpose

The first purpose of this dissertation was to evaluate the accuracy of both IRT and MIRT estimation programs when the assumption of unidimensionality is violated. The research has shown that the MC1-PL model is superior to the 1-PL IRT model in accurately estimating the both the multidimensional and unidimensional person parameters as well as the unidimensional item parameters.

This dissertation has also shown that the fit statistics used in the 1-PL model are not sensitive to distortions caused by a multidimensional data structure. If the data can be shown to be unidimensional, the 1-PL IRT model is sufficiently robust to recover the person ability and item difficulty values. If the data are multidimensional, the 1-PL IRT model provides not only unstable parameter estimates, but also inflated standard error values and less-accurate fit statistics.

The final purpose of this dissertation was to determine whether or not the MC1-PL IRT model as implemented by ConQuest can correctly recover the underlying construct relevant multidimensional structure within the educational domain. ConQuest can effectively recover the underlying multidimensional person-level ability structure and report accurate theta estimates on each of the latent dimensions. ConQuest is not capable of reporting multiple item difficulty values that span multiple correlated dimensions for a single item. Instead, a single difficulty value is provided that represents a single point in latent space that represents an aggregate difficulty value across dimensions. Wilson (personal communication, August 8, 2004) points that test practitioners can use this point value as the difficulty parameter for each of the latent constructs measured by the item.

Considerations for Future Research

This project intentionally maintained a focus on two correlated dimensions with construct-relevant multidimensionality. The correlation was constrained to be 0.50, a realistic assumption given the nature of scholastic assessments within a content domain. To maintain a simplest-case scenario, the dimensions were constrained with a common point of origin and a common measurement scale. The calibration programs may yield disparate results if the correlation between dimensions is decreased to a value less than 0.50.

A Monte Carlo study provides a stable foundation in which the variables of interest can be controlled. The next step is to bridge the theoretical and real-life scenario with a practical application of theory. The next logical step is the creation of a multidimensional assessment that covers the mathematics domain. The blueprint for such an assessment has been designed with the intent to gather empirical evidence to bolster this project's findings.

Research question 2 could not be answered because of software and theoretical limitations. The proofs for a next generation of a multidimensional model will be an important step to answering question 2. This multidimensional model will allow the estimation of item difficulty parameters on each of *n* dimensions in latent space.

REFERENCES

Ackerman, T.A. (1994). Using multidimensional item response theory to understand what items and tests are measuring, *Applied Measurement in Education*, *7*, 255-278.

Byrne, B.M. 2001. *Structural Equation Modeling with Amos*. Mahwah, New Jersey: Lawrence Erlbaum.

Brown, M.W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen, & J. S. Long, (Eds), *Testing Structural Equation Models.* Newbury Park, CA: Sage. 136-162.

Bunderson, C.V., Gibbons, A.S., Olsen, J.B., & Kearsley, G.P. (1981). Work Models: Beyond Instructional Objectives, *Instructional Science*, *10*, 205-215.

Cattell, R.B. (Ed.). (1966). *Handbook of multivariate experimental psychology*. Chicago: Rand McNally.

Drasgow, F., & Parsons, C. K. (1983). Application of Unidimensional Item Response Theory Models to Multidimensional Data, *Applied Psychological Measurement*. *7*, 189-199.

Graybill, F. A. (1964). *An Introduction to Linear Statistical Models, Vol 1*. New York: McGraw Hill.

Hambleton, R.K. (2000). *Introduction to Item Response Theory*. Breakout session presented at the Sylvan Prometric Results 2000 Conference, Tucson AZ.

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of Item Response Theory*. Thousand Oaks: Sage Publications.

Harwell, M.R. (1997). Analyzing the Results of Monte Carlo Studies in Item Response Theory, *Educational and Psychological Measurement*. *58*, 266-279.

Hulin, C. L., Lissak, R. I., Drasgow, F. (1982). Recovery of Two- and Three-Parameter Logistic Item Characteristic Curves: A Monte Carlo Study, *Applied Psychological Measurement*. *6*, 249-260.

Kaplan, D. (2000). *Structural Equation Modeling: Foundations and Extensions*. Thousand Oaks: Sage Publications.

Knol, D.L. & Berger, M.P. (1991). Empirical Comparison Between Factor Analysis and Multidimensional Item Response Models, *Multivariate Behavioral Research*. *26*, 457-477.

Linacre, J. M. (2004). *A User's Guide to Winsteps Ministep Rasch-Model Computer Programs*. Chicago, IL.

Luecht, R.M. (1996). Multidimensional Computerized Adaptive Testing in a Certification or Licensure Context, *Applied Psychological Measurement*. *20*, 389-404.

McDonald, R. P.. (1999). *Test Theory*. Mawhah, New Jersey: Lawrence Erlbaum Associates.

McKinley, R.L. & Mills, C.N. (1985). A Comparison of Several Goodness-of-Fit Statistics, *Applied Psychological Measurement*. *9*, 49-57.

Messick, S. (1995). Validity of Psychological Assessment, *American Psychologist*. *50* 741-749.

Reckase, M.D. (1997). The Past and Future of Multidimensional Item Response Theory, *Applied Psychological Measurement*. *21*, 25-36.

Reckase, M.D. (1979). Unifactor latent trait models applied to multi-factor tests: Results and implications, *Journal of Educational Statistics*. *4*, 207-230.

San-Luiz, C. Sanchez-Bruno, A. (1998). Graphical Analysis of the Two Parameters Logistic Function (A Methodological Proposal), *Quality and Quantity*. *32*, 433-439.

Segal, D.O. (1996). Multidimensional adaptive testing, *Psychometrika*. *61*, 331-354.

Smith, R.M. (2002, April). *The Family Approach to Assessing Fit in Rasch Measurement*. Paper presented at the 11th International Objective Measurement Workshop, New Orleans.

Spence, I. (1983). Monte Carlo Simulation Studies, *Applied Psychological Measurement*. *7*, 405-425.

Steinberg, L., Thissen, D. & Wainer, H. (2000) Validity. In H. Wainer (Ed), *Computerized Adaptive Testing: A Primer*. Mahwah, Jew Jersey: Lawrence Erlbaum.

Steiger, J.H. (1990). Structural Model Evaluation and Modification: An Interval Estimation Approach, *Multivariate Behavioral Research*. *25*, 173-180.

Stout, W. (1990). A new Item Response Theory Modeling Approach with Applications to Unidimensionality Assessment and Ability Estimation, *Psychometrika*. *55*, 293-325.

Traub, R.E. 1983. A priori considerations in choosing an item response model. In R. K. Hambleton, (Ed), *Applications of Item Response Theory*. Educational Research Institute of British Columbia.

Tucker, L.R., Lewis, C. (1973). A Reliability Coefficient for Maximum Likelihood Factor Analysis, *Psychometrika*. *38*.

Wilson, M.R. (1998). *ACER ConQuest User Manual*. Melbourne, Australia: Quintec Group.

Zhao, J. C., McMorris, R. F., Pruzek, R. M., & Chen, R. (2002, April). *The Robustness of the Unidimensional 3-PL IRT Model When Applied to Two-dimensional Data in Computerized Adaptive Tests.* Paper presented at the Annual Meeting of the American Educational Research Association in New Orleans, LA.

APPENDICES

Appendix A

Syntax and Command Files Used in Question 1

SPSS Test Results Syntax Used to Generate Test Responses to Answer Question 1

```
new file.
comment     Search for the following symbols: <*>
comment     Follow the instructions in the comments that
comment     follow the symbol.

comment     Save and run the script within SPSS. A new data sheet
comment     will be created. The simulated test results will be
comment     listed in variables resp1 - resp21.


input program.

NUMERIC i (F4.0)

comment     <*>
comment     Change the following value from 1000 to the
comment     number of cases you want generated.

  loop i=1 to 1000.
        COMPUTE x=rv.normal (0,1).
        end case.
  end loop.
  end file.
end input program.
EXECUTE.

COMPUTE x2=sqrt(1-(.5**2)/1).
COMPUTE x1=(0+.5/1)*(x-0).
COMPUTE y=rv.normal (x1,x2).
EXECUTE.

comment     The following correlations command verifies that
comment     the two distributions are correlated.

CORRELATIONS
  /VARIABLES=x y
  /PRINT=TWOTAIL NOSIG
  /STATISTICS DESCRIPTIVES
  /MISSING=PAIRWISE .
```

```
COMPUTE p=(x+(y/2)).
COMPUTE q=(sqrt(3)/2)*y.
COMPUTE p1=(.75*p+sqrt(3)/4*q).
COMPUTE q1=sqrt(3)/4*p+(q/4).
COMPUTE xz=p1-(q1/sqrt(3)).
COMPUTE yz=(2/sqrt(3))*q1.
COMPUTE lvz=sqrt((3/4*p1+sqrt(3)/4*q1)**2+(sqrt(3)/4*p1+q1/4)**2).
EXECUTE.

comment         lvz: 'Length Vector Z' (from origin)
comment        slvz: 'Signed Length Vector Z' (from origin)

IF (lvz > 3) lvz = 3.
IF ((xz   <=   0)) slvz = 0 - lvz.
IF ((xz   >=   0)) slvz = lvz.

EXECUTE .

comment       These are all projections onto the Composite vector.
comment       For Q1.

comment       <*>
comment       For each of the following 21 items/variables, type in the
comment       logit values for each item difficulty.

COMPUTE diff1 = -2.34.
COMPUTE diff2 = -1.47.
COMPUTE diff3 = -.87.
COMPUTE diff4 = .52.
COMPUTE diff5 = 1.47.
COMPUTE diff6 = 1.73.
COMPUTE diff7 = 1.65.
COMPUTE diff8 = -2.17.
COMPUTE diff9 = -1.21.
COMPUTE diff10 = -.52.
COMPUTE diff11 = 1.04.
COMPUTE diff12 = 1.91.
COMPUTE diff13 = 2.51.
COMPUTE diff14 = 1.73.
COMPUTE diff15 = -2.3.
COMPUTE diff16 = -1.6.
COMPUTE diff17 = -.9.
COMPUTE diff18 = 0.0.
COMPUTE diff19 = 1.0.
COMPUTE diff20 = 1.5.
```

```
COMPUTE diff21 = 2.7.

execute .

COMPUTE ptheta1 = (2.718**(x - diff1))/(1+2.718**(x - diff1)) .
COMPUTE ptheta2 = (2.718**(x - diff2))/(1+2.718**(x - diff2)) .
COMPUTE ptheta3 = (2.718**(x - diff3))/(1+2.718**(x - diff3)) .
COMPUTE ptheta4 = (2.718**(x - diff4))/(1+2.718**(x - diff4)) .
COMPUTE ptheta5 = (2.718**(x - diff5))/(1+2.718**(x - diff5)) .
COMPUTE ptheta6 = (2.718**(x - diff6))/(1+2.718**(x - diff6)) .
COMPUTE ptheta7 = (2.718**(x - diff7))/(1+2.718**(x - diff7)) .
COMPUTE ptheta8 = (2.718**(y - diff8))/(1+2.718**(y - diff8)) .
COMPUTE ptheta9 = (2.718**(y - diff9))/(1+2.718**(y - diff9)) .
COMPUTE ptheta10 = (2.718**(y - diff10))/(1+2.718**(y - diff10)) .
COMPUTE ptheta11 = (2.718**(y - diff11))/(1+2.718**(y - diff11)) .
COMPUTE ptheta12 = (2.718**(y - diff12))/(1+2.718**(y - diff12)) .
COMPUTE ptheta13 = (2.718**(y - diff13))/(1+2.718**(y - diff13)) .
COMPUTE ptheta14 = (2.718**(y - diff14))/(1+2.718**(y - diff14)) .
COMPUTE ptheta15 = (2.718**(slvz - diff15))/(1+2.718**(slvz - diff15)) .
COMPUTE ptheta16 = (2.718**(slvz - diff16))/(1+2.718**(slvz - diff16)) .
COMPUTE ptheta17 = (2.718**(slvz - diff17))/(1+2.718**(slvz - diff17)) .
COMPUTE ptheta18 = (2.718**(slvz - diff18))/(1+2.718**(slvz - diff18)) .
COMPUTE ptheta19 = (2.718**(slvz - diff19))/(1+2.718**(slvz - diff19)) .
COMPUTE ptheta20 = (2.718**(slvz - diff20))/(1+2.718**(slvz - diff20)) .
COMPUTE ptheta21 = (2.718**(slvz - diff21))/(1+2.718**(slvz - diff21)) .

COMPUTE un1=rv.uniform(0,1).
COMPUTE un2=rv.uniform(0,1).
COMPUTE un3=rv.uniform(0,1).
COMPUTE un4=rv.uniform(0,1).
COMPUTE un5=rv.uniform(0,1).
COMPUTE un6=rv.uniform(0,1).
COMPUTE un7=rv.uniform(0,1).
COMPUTE un8=rv.uniform(0,1).
COMPUTE un9=rv.uniform(0,1).
COMPUTE un10=rv.uniform(0,1).
COMPUTE un11=rv.uniform(0,1).
COMPUTE un12=rv.uniform(0,1).
COMPUTE un13=rv.uniform(0,1).
COMPUTE un14=rv.uniform(0,1).
COMPUTE un15=rv.uniform(0,1).
COMPUTE un16=rv.uniform(0,1).
COMPUTE un17=rv.uniform(0,1).
COMPUTE un18=rv.uniform(0,1).
COMPUTE un19=rv.uniform(0,1).
COMPUTE un20=rv.uniform(0,1).
```

```
COMPUTE un21=rv.uniform(0,1).

EXECUTE .

IF (ptheta1 >= un1) resp1 = 1.
IF (ptheta1 < un1) resp1 = 0.
IF (ptheta2 >= un2) resp2 = 1.
IF (ptheta2 < un2) resp2 = 0.
IF (ptheta3 >= un3) resp3 = 1.
IF (ptheta3 < un3) resp3 = 0.
IF (ptheta4 >= un4) resp4 = 1.
IF (ptheta4 < un4) resp4 = 0.
IF (ptheta5 >= un5) resp5 = 1.
IF (ptheta5 < un5) resp5 = 0.
IF (ptheta6 >= un6) resp6 = 1.
IF (ptheta6 < un6) resp6 = 0.
IF (ptheta7 >= un7) resp7 = 1.
IF (ptheta7 < un7) resp7 = 0.
IF (ptheta8 >= un8) resp8 = 1.
IF (ptheta8 < un8) resp8 = 0.
IF (ptheta9 >= un9) resp9 = 1.
IF (ptheta9 < un9) resp9 = 0.
IF (ptheta10 >= un10) resp10 = 1.
IF (ptheta10 < un10) resp10 = 0.
IF (ptheta11 >= un11) resp11 = 1.
IF (ptheta11 < un11) resp11 = 0.
IF (ptheta12 >= un12) resp12 = 1.
IF (ptheta12 < un12) resp12 = 0.
IF (ptheta13 >= un13) resp13 = 1.
IF (ptheta13 < un13) resp13 = 0.
IF (ptheta14 >= un14) resp14 = 1.
IF (ptheta14 < un14) resp14 = 0.
IF (ptheta15 >= un15) resp15 = 1.
IF (ptheta15 < un15) resp15 = 0.
IF (ptheta16 >= un16) resp16 = 1.
IF (ptheta16 < un16) resp16 = 0.
IF (ptheta17 >= un17) resp17 = 1.
IF (ptheta17 < un17) resp17 = 0.
IF (ptheta18 >= un18) resp18 = 1.
IF (ptheta18 < un18) resp18 = 0.
IF (ptheta19 >= un19) resp19 = 1.
IF (ptheta19 < un19) resp19 = 0.
IF (ptheta20 >= un20) resp20 = 1.
IF (ptheta20 < un20) resp20 = 0.
IF (ptheta21 >= un21) resp21 = 1.
IF (ptheta21 < un21) resp21 = 0.
```

```
EXECUTE .

comment       Format each response to a single digit integer (1 or 0).

FORMAT resp1 to resp21 (F1.0).
EXECUTE .

comment       Change the name of the .SAV file to the iteration number.

SAVE OUTFILE='Q1-all.SAV'
/COMPRESSED.

WRITE OUTFILE = 'Q1.DAT'
TABLE
 /i resp1 resp2 resp3
        resp4 resp5 resp6 resp7 resp8 resp9
        resp10 resp11 resp12 resp13 resp14
        resp15 resp16 resp17 resp18 resp19
        resp20 resp21.

SAVE TRANSLATE OUTFILE='Q1-ZYZTheta.xls'
  /TYPE=XLS /MAP /REPLACE
  /keep i x y slvz .

EXECUTE .

FACTOR
  /VARIABLES resp1 resp2 resp3 resp4 resp5 resp6 resp7 resp8 resp9 resp10
  resp11 resp12 resp13 resp14 resp15 resp16 resp17 resp18 resp19 resp20 resp21
   /MISSING LISTWISE /ANALYSIS resp1 resp2 resp3 resp4 resp5 resp6 resp7 resp8
  resp9 resp10 resp11 resp12 resp13 resp14 resp15 resp16 resp17 resp18 resp19
  resp20 resp21
  /PRINT INITIAL EXTRACTION ROTATION
  /PLOT EIGEN ROTATION
  /CRITERIA MINEIGEN(1) ITERATE(25)
  /EXTRACTION PC
  /CRITERIA ITERATE(25)
  /ROTATION PROMAX(4)
  /METHOD=CORRELATION .

EXECUTE.
new file.
```

Comments:

x = the original theta level for the Calculations dimension on the oblique coordinate system.

y = the original theta level for the Necessary Operations dimension on the oblique coordinate system.

p = the value of x plotted on the orthogonal coordinate system.

q = the value of y plotted on the orthogonal coordinate system.

p1 = the perpindicular projection of p onto vector Z on the orthogonal coordinate system. This is the same as P in the equations found in the methods section.

q1 = the perpindicular projection of q onto vector Z on the orthogonal coordinate system. This is the same as Q in the equations found in the methods section.

xz = the value of p1 plotted on the oblique coordinate system.

yz = the value of q1 plotted on the oblique coordinate system.

lvz = the distance of (P,Q) or (p1,q1) from the origin.

slvz = the signed distance of (P,Q) or (p1,q1) from the origin. This is the person's theta level on the composite vector.

ConQuest Command File for Question 1.

```
datafile q1.dat;
format id 1-4 responses 5-25;
set constraint=cases,update=yes,warnings=no;
score (0,1) (0,1)  ! item(1);
score (0,1) (0,1)  ! item(2);
score (0,1) (0,1)  ! item(3);
score (0,1) (0,1)  ! item(4);
score (0,1) (0,1)  ! item(5);
score (0,1) (0,1)  ! item(6);
score (0,1) (0,1)  ! item(7);
score (0,1) (0,1)  ! item(8);
score (0,1) (0,1)  ! item(9);
```

```
score (0,1) (0,1)  ! item(10);
score (0,1) (0,1)  ! item(11);
score (0,1) (0,1)  ! item(12);
score (0,1) (0,1)  ! item(13);
score (0,1) (0,1)  ! item(14);
score (0,1) (0,1)  ! item(15);
score (0,1) (0,1)  ! item(16);
score (0,1) (0,1)  ! item(17);
score (0,1) (0,1)  ! item(18);
score (0,1) (0,1)  ! item(19);
score (0,1) (0,1)  ! item(20);
score (0,1) (0,1)  ! item(21);
model item;
export parameters >> q1.prm;
export reg_coefficients >> q1-regression.reg;
export covariance >> q1.cov;
estimate !method=montecarlo,fit=yes,iterations=500,conv=.001;
show parameters !tables=1:2:3 >> q1.shw;
show cases !estimates=eap >> q1_person.prs;
quit;
```

Winsteps Command File for Question 1.

```
&INST
 TITLE = "Q1"
;Input Data Format
NAME1  = 1
NAMLEN = 4
ITEM1  = 5
NI     = 21
XWIDE  = 1
PERSON = Person
ITEM   = Item
DATA   = Q1.DAT
CODES  = "01"
TFILE=*
10.1
14.1
18.1
25.1
*
CLFILE = *
0 Wrong
1 Right
*
```

```
UMEAN  = 0.00          ; item mean - default is 0.00
USCALE = 1.00          ; measure units - default is 1.00
UDECIM = 3             ; reported decimal places - default is 2
MRANGE = 0             ; half-range on maps - default is 0 (auto-scaled)

&END          ; item IDs
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
END LABELS          ;data
```

Appendix B

Descriptive Statistics for Two Distributions Correlated at .50.

Descriptive Statistics

	Mean	Std. Deviation	N
X	.0166	.96219	1000
Y	.0498	1.00255	1000

Correlations

		X	Y
X	Pearson Correlation	1	.502**
	Sig. (2-tailed)	.	.000
	N	1000	1000
Y	Pearson Correlation	.502**	1
	Sig. (2-tailed)	.000	.
	N	1000	1000

**. Correlation is significant at the 0.01 level

Figure B1. Correlation between two distributions.

Appendix C

Item Analysis for 21 Items

The classical statistics for the 21 items used to answer question 1 are shown in Table C1.

A review of the item p-values indicates that the most difficult item has a p-value of .12. The easiest item is RESP15 with a p-value of .87. These items are within the target difficulty ranges for most assessments.

The range for the item discrimination values (upper 27% - lower 27%) is .21 for RESP4 to .08 for RESP6. Note that 14% of the respondents answered item RESP6 correctly. The lower discrimination value is most likely an artifact of the item's difficulty. The items with p-values more extreme than the range of .15 and .85 exhibit poor discrimination between knowledgeable and less-knowledgeable respondents.

The item to total score correlation for these items ranges from a high of .613 for RESP4 to .376 for item RESP6. With this range of item to total score correlations, a response to each of these items can be a predictor of the total raw score.

Based on this hypothetical item analysis of the difficulty, item discrimination, and item to total score correlation, each of these items performs adequately in this assessment. These 21 items should be retained for use in future assessments.

The internal consistency of these 21 items as measured by Cronbach's alpha is .86. This is a surprisingly high value given that the underlying data structure is multidimensional in nature.

Table C1. Classical Statistics for 21 Items.

Classical Statistics for 21 Items.

Item ID	p-value	Discrimination	Correlation
RESP1	.84	.10	.444
RESP2	.75	.15	.510
RESP3	.82	.11	.424
RESP4	.43	.21	.613
RESP5	.27	.16	.526
RESP6	.14	.08	.376
RESP7	.27	.15	.503
RESP8	.84	.11	.444
RESP9	.70	.18	.565
RESP10	.58	.17	.519
RESP11	.35	.18	.561
RESP12	.22	.13	.480
RESP13	.14	.10	.399
RESP14	.27	.15	.504
RESP15	.87	.09	.408
RESP16	.76	.14	.497
RESP17	.67	.16	.518
RESP18	.49	.21	.601
RESP19	.35	.19	.580
RESP20	.26	.16	.539
RESP21	.12	.09	.431

Appendix D

Question 1: RMSQ for 1000 Respondents Across 25 Iterations

	RMSQ			RMSQ			RMSQ	
ID	NO	C	ID	NO	C	ID	NO	C
1	0.5888	0.9744	26	0.8127	0.4562	51	0.8596	0.6866
2	0.3771	0.5518	27	0.5066	0.3675	52	1.0860	0.5544
3	0.4857	0.3835	28	0.4424	0.3740	53	0.6096	0.4337
4	0.3938	0.4157	29	0.4890	0.4014	54	0.6716	0.4345
5	0.8313	0.3657	30	0.5641	0.5190	55	0.4693	0.4425
6	1.0838	0.7530	31	0.4887	0.5623	56	0.7330	0.6562
7	0.3937	0.4733	32	0.4905	0.5716	57	0.3818	0.3919
8	0.5208	0.6246	33	0.7719	0.5650	58	1.3483	0.4938
9	0.3804	0.4803	34	0.5996	0.5310	59	0.4243	0.4872
10	0.4890	0.4048	35	0.4638	0.8669	60	0.4373	0.4871
11	0.5762	1.0951	36	0.5097	0.8976	61	0.7261	0.5705
12	0.6159	0.3236	37	0.6793	0.3842	62	0.4442	0.6485
13	0.8058	0.3747	38	0.5054	0.4263	63	0.6289	0.4049
14	0.6201	0.5594	39	0.4712	0.8966	64	0.9493	0.6817
15	0.4805	0.4976	40	0.7030	0.3794	65	0.3204	0.6370
16	0.7094	0.3671	41	0.4061	0.3932	66	0.4335	0.4653
17	0.8931	0.4385	42	0.4219	0.6769	67	0.3524	0.4644
18	0.7714	0.6959	43	0.4664	0.5113	68	0.8314	0.8808
19	0.4579	0.5260	44	0.4644	0.5063	69	0.3872	0.6617
20	0.4036	0.8315	45	0.3565	0.7236	70	1.1436	0.6807
21	0.3408	0.3948	46	0.4036	0.5104	71	0.7049	0.5536
22	0.3792	0.9328	47	0.4658	0.3282	72	0.4925	0.4557
23	0.6266	0.5293	48	0.3810	0.3724	73	0.3347	0.4712
24	0.4040	0.7668	49	0.4663	0.6153	74	0.4134	0.5345
25	0.3105	0.4372	50	0.3799	0.2871	75	0.5393	0.4278

(Table Continued)

	RMSQ			RMSQ			RMSQ	
ID	NO	C	ID	NO	C	ID	NO	C
76	0.5285	0.3911	101	0.9493	0.5715	126	0.4388	0.9107
77	0.4641	0.4801	102	0.4918	0.3567	127	0.3650	0.5895
78	0.7065	0.8100	103	0.9119	0.6395	128	0.6484	0.7236
79	1.1283	0.9040	104	0.4247	0.5450	129	1.2388	0.4337
80	0.3352	1.2049	105	0.3842	0.4413	130	0.5126	0.8966
81	0.4794	0.5573	106	0.6182	0.4201	131	0.3733	0.3815
82	0.3054	0.8703	107	0.5567	0.3943	132	0.5310	0.4765
83	0.3775	0.4714	108	0.3860	0.3952	133	0.5887	0.3435
84	0.9398	0.4831	109	0.7419	1.0025	134	0.7401	0.2951
85	0.3322	0.3194	110	0.6553	0.5872	135	0.4333	0.4213
86	0.7241	0.4283	111	0.4324	0.3451	136	0.2722	0.5122
87	0.7031	0.5565	112	0.4414	0.4472	137	0.4403	0.5319
88	0.4284	0.4559	113	0.4143	0.6530	138	0.7467	0.5095
89	0.9035	0.9020	114	0.7291	0.4278	139	0.4743	0.9236
90	0.4850	0.4155	115	0.3943	0.6148	140	0.4907	0.6038
91	0.5103	0.4044	116	0.5620	0.5443	141	0.3921	0.4373
92	1.0286	0.4763	117	0.4261	0.4401	142	0.3941	0.4893
93	0.4494	0.5357	118	0.7881	1.0652	143	0.9086	0.4109
94	0.5599	0.5191	119	0.3026	0.3639	144	0.4322	0.3647
95	0.6670	0.3892	120	0.6776	0.4869	145	0.3989	0.3573
96	0.4387	0.7395	121	0.2704	0.5221	146	0.8669	0.7876
97	0.3625	0.6217	122	0.6496	0.6923	147	0.4456	0.6265
98	0.6258	0.9552	123	0.6340	0.3435	148	1.6577	0.8792
99	0.4276	0.4102	124	0.4656	1.0248	149	0.4096	0.4213
100	0.5193	0.5107	125	0.4731	0.3630	150	0.5536	0.4971

(Table Continued)

	RMSQ			RMSQ			RMSQ	
ID	NO	C	ID	NO	C	ID	NO	C
151	0.4134	0.7147	176	0.3348	0.6630	201	0.5974	0.4259
152	0.4536	0.4713	177	0.4476	0.4537	202	0.4743	0.3897
153	0.4745	0.5991	178	0.4408	0.4225	203	0.4898	0.5267
154	0.6365	0.4467	179	1.1229	0.4246	204	0.4375	0.3731
155	0.7121	0.3926	180	0.7649	0.4283	205	0.4374	0.3181
156	1.2731	0.8304	181	0.3840	0.3693	206	0.4375	0.8874
157	0.5325	0.4224	182	0.5927	0.3644	207	0.3805	0.4024
158	0.5428	0.5276	183	0.5484	0.3612	208	0.6997	0.6911
159	0.4677	0.6506	184	0.4080	0.7666	209	0.3795	0.3903
160	0.4728	0.4251	185	0.3593	0.6366	210	0.3341	0.7463
161	0.6076	0.6033	186	0.4005	0.9737	211	0.5249	0.5148
162	0.4366	0.4717	187	0.5346	0.3531	212	0.5149	0.4977
163	0.5083	0.6407	188	0.5595	0.3771	213	0.4016	0.4835
164	0.5648	0.3437	189	0.5471	0.5607	214	0.4845	0.6636
165	0.2830	0.8393	190	0.2804	0.6281	215	0.3835	0.4865
166	0.7984	0.4455	191	0.9783	0.3958	216	0.7285	0.3205
167	0.5738	0.3590	192	0.4212	0.4833	217	0.5191	0.3928
168	0.4678	0.5343	193	0.5696	0.4346	218	0.3324	0.5291
169	0.3996	0.7055	194	0.4565	0.5567	219	0.3581	0.7371
170	0.4634	0.4818	195	0.6442	0.5949	220	0.4132	0.5075
171	0.6755	0.4087	196	1.1982	1.0496	221	0.7991	0.5911
172	0.9624	0.8024	197	0.4635	0.4730	222	0.3881	0.5084
173	0.5495	0.4112	198	0.7050	0.6674	223	1.0572	0.9417
174	0.5682	0.3924	199	0.5286	0.4867	224	0.4572	0.4012
175	0.8585	0.3452	200	0.5976	0.8074	225	0.5198	0.5841

(Table Continued)

	RMSQ			RMSQ			RMSQ	
ID	NO	C	ID	NO	C	ID	NO	C
226	0.4100	0.4462	251	0.3579	0.4593	276	0.9132	0.5298
227	0.9650	0.8149	252	0.5624	0.7641	277	1.2346	0.3569
228	0.4082	0.5191	253	0.6118	0.7020	278	0.5066	0.4235
229	0.3881	0.6785	254	0.5280	0.5101	279	0.3829	0.3576
230	0.4004	0.9941	255	0.6672	0.4524	280	0.3960	0.6319
231	0.7512	0.6448	256	0.3421	0.4433	281	0.5957	0.6201
232	0.5055	1.0040	257	0.4608	0.4195	282	0.6652	0.9346
233	0.4500	0.4837	258	0.3393	0.2277	283	0.4187	0.6896
234	0.7363	0.4951	259	0.3927	0.5196	284	0.4106	0.5002
235	0.5047	0.6754	260	1.0866	0.5662	285	0.3870	0.4404
236	0.4631	0.8074	261	0.4564	0.4587	286	0.7731	0.5344
237	0.4032	0.3185	262	0.7447	0.4867	287	0.5435	0.4033
238	0.6986	1.1140	263	0.5712	0.4882	288	0.6243	0.5455
239	0.8499	0.6895	264	0.3875	0.4193	289	1.0785	0.4093
240	0.5473	0.4760	265	0.3927	0.4219	290	0.5276	0.5189
241	0.5866	0.3485	266	0.8559	0.8527	291	0.7897	0.9131
242	0.4642	0.3362	267	0.3898	0.5808	292	0.6543	0.5546
243	0.5959	0.3365	268	0.8349	0.4500	293	0.4065	0.4339
244	0.4324	0.4342	269	0.3809	0.4788	294	0.5075	0.8186
245	0.5354	1.0682	270	0.7510	0.7144	295	0.3303	0.3916
246	0.3809	0.6615	271	0.4664	0.3419	296	1.0549	1.2342
247	0.4269	0.5225	272	0.5654	0.3934	297	0.6296	0.4649
248	0.4473	0.5374	273	0.4787	0.3899	298	0.3670	0.3872
249	0.4365	0.6870	274	0.5239	0.7678	299	0.3773	0.4247
250	0.3719	0.4267	275	0.4047	0.4612	300	0.5177	0.5701

(Table Continued)

	RMSQ			RMSQ			RMSQ	
ID	NO	C	ID	NO	C	ID	NO	C
301	0.2992	0.6014	326	0.3961	0.3433	351	0.4592	0.4536
302	0.6041	0.6275	327	0.3750	0.4491	352	0.3495	0.7295
303	0.4153	0.5673	328	0.4749	0.3861	353	0.6774	0.4048
304	0.8408	0.4120	329	0.3901	0.6144	354	0.5078	0.4747
305	0.6509	0.8137	330	0.5500	0.5474	355	0.8032	0.5349
306	0.4840	0.3782	331	0.4690	0.4889	356	0.4611	0.7939
307	0.3603	0.3269	332	0.4877	0.7565	357	0.3566	0.4557
308	0.5374	0.5290	333	0.4375	0.8057	358	0.4188	0.4875
309	0.3541	0.3724	334	0.3688	0.4768	359	0.4871	0.6190
310	0.5258	0.4668	335	0.3847	0.5987	360	0.2936	0.6477
311	0.8736	0.4404	336	0.5895	0.4506	361	0.3931	0.3532
312	0.4557	0.5913	337	0.3969	0.5414	362	0.3896	0.4707
313	0.5862	0.5762	338	0.4098	0.4548	363	0.5043	0.5112
314	0.3889	0.4153	339	0.4367	0.5376	364	0.5752	0.4593
315	0.4543	0.4706	340	0.3424	0.3754	365	0.3715	0.3923
316	0.5522	0.3793	341	0.3874	0.5940	366	0.9182	0.4177
317	0.6590	0.7744	342	0.7875	0.4566	367	0.4472	0.4365
318	0.4009	0.7723	343	0.3512	0.4949	368	0.5482	0.3452
319	0.4753	0.4354	344	0.8921	1.2234	369	0.3848	0.3016
320	0.3910	0.3858	345	0.3338	0.8572	370	0.8140	0.6269
321	0.4282	0.4837	346	0.3246	0.3121	371	0.4026	0.7182
322	0.5024	0.5668	347	1.2991	0.6403	372	0.4375	0.7344
323	0.4667	0.6283	348	0.5464	0.5035	373	0.7825	0.3384
324	0.4327	0.4494	349	0.4145	0.4159	374	0.7181	0.3982
325	0.3865	0.5338	350	0.5036	1.1085	375	0.3557	0.5958

(Table Continued)

	RMSQ			RMSQ			RMSQ	
ID	NO	C	ID	NO	C	ID	NO	C
376	0.3282	0.5235	401	0.4144	0.4561	426	0.5260	0.3991
377	0.6309	0.5957	402	0.7275	0.5024	427	0.3129	0.3804
378	0.6103	0.3670	403	0.3689	0.8483	428	0.5044	1.3078
379	0.4025	0.3953	404	0.5456	0.3197	429	0.6430	0.5115
380	0.8143	0.4905	405	0.5854	0.3388	430	0.3810	0.5047
381	0.6755	0.5778	406	0.6293	0.5284	431	0.4667	0.9677
382	0.3912	0.4513	407	0.3546	0.4361	432	0.4210	0.3398
383	0.4389	0.4022	408	0.3424	0.5757	433	0.3787	0.6615
384	0.7906	0.4451	409	0.5271	0.4777	434	0.4575	0.6590
385	0.4954	0.4810	410	0.3893	0.9037	435	0.3957	0.4559
386	0.4562	0.8393	411	0.6769	0.4549	436	0.5615	0.7507
387	0.4420	0.5236	412	0.3900	0.4637	437	0.6752	0.6207
388	0.4362	1.0986	413	0.3660	0.5186	438	0.7705	1.1251
389	0.5637	0.5137	414	0.3353	0.4586	439	0.4693	0.4293
390	0.4299	0.3644	415	0.6983	0.5000	440	0.6488	1.0852
391	0.7332	0.6130	416	0.2903	0.7166	441	0.4744	0.5881
392	0.5590	0.4792	417	0.5657	0.8230	442	0.3864	0.4075
393	0.3878	0.7778	418	0.4401	0.3561	443	0.9374	0.8657
394	0.3591	0.3433	419	0.7547	0.5429	444	0.3816	0.8586
395	0.8849	0.5624	420	0.5663	0.6833	445	0.3595	0.5633
396	0.4349	0.4350	421	0.6789	0.5063	446	0.5717	0.4606
397	1.1578	0.8280	422	0.9321	0.4356	447	0.4423	0.3667
398	0.4570	0.6102	423	0.8295	0.4376	448	0.3813	0.4208
399	0.3538	0.3247	424	0.7737	0.6439	449	1.0084	0.4032
400	0.5285	0.4290	425	0.5968	0.9366	450	0.6735	1.2930

(Table Continued)

	RMSQ			RMSQ			RMSQ	
ID	NO	C	ID	NO	C	ID	NO	C
451	0.4446	0.3525	476	0.4230	0.7250	501	0.3851	0.4749
452	0.6234	0.5066	477	0.4374	0.5076	502	0.3395	0.4108
453	0.5604	0.8059	478	0.6572	0.4214	503	0.5826	0.3961
454	0.5079	0.3798	479	0.4349	0.6611	504	0.9799	0.5596
455	0.3385	1.5419	480	0.3649	0.3022	505	0.5255	0.7427
456	0.3476	0.6533	481	0.3315	0.5939	506	0.5679	0.8621
457	0.2896	0.4339	482	0.9574	0.9968	507	0.4124	0.3455
458	0.4155	0.8299	483	0.8655	0.5255	508	0.7747	0.4732
459	1.0908	0.8343	484	0.4429	0.4723	509	0.4580	0.7782
460	0.4446	1.1125	485	0.4705	0.4767	510	0.6523	0.5440
461	0.5107	0.6059	486	1.1499	0.9091	511	0.5122	0.2785
462	0.4161	0.8541	487	0.4328	0.5057	512	0.5842	0.5130
463	0.6382	0.3858	488	0.3617	0.5627	513	1.0138	0.5062
464	0.5349	0.7071	489	0.5017	0.4695	514	1.0029	0.9251
465	0.9448	0.4075	490	0.6490	0.8549	515	0.3334	0.3455
466	0.6742	0.5450	491	1.4141	0.3266	516	0.4110	0.7566
467	0.4014	0.3874	492	0.7866	0.4996	517	0.4330	0.4301
468	0.4377	0.4843	493	1.0404	1.0335	518	0.4184	0.6740
469	0.4660	0.4095	494	0.5958	0.4707	519	0.4319	0.5241
470	0.4099	1.0101	495	0.4375	0.2963	520	0.4016	0.5608
471	0.3659	1.0544	496	0.4150	0.3446	521	0.7237	0.7279
472	0.4088	0.3862	497	0.4276	0.3856	522	0.2363	0.4078
473	0.5902	0.8858	498	0.3902	0.3543	523	0.3800	0.3878
474	0.9026	0.6534	499	0.6518	0.4647	524	0.3845	0.4029
475	0.3587	0.4049	500	0.4240	0.5122	525	0.6732	0.6339

(Table Continued)

	RMSQ			RMSQ			RMSQ	
ID	NO	C	ID	NO	C	ID	NO	C
526	0.4574	0.6517	551	0.3632	0.4670	576	0.8774	1.0042
527	1.1232	0.6820	552	0.5778	0.5371	577	1.0141	0.6986
528	0.7038	0.7313	553	0.4742	0.6038	578	0.7950	0.3295
529	0.2857	0.3033	554	0.3571	0.3909	579	0.3556	0.6043
530	0.5849	0.8057	555	0.4719	0.4423	580	0.6534	0.5724
531	0.5158	0.4589	556	0.3444	0.3850	581	0.4469	0.3984
532	0.4488	0.4917	557	0.5266	0.4319	582	0.5509	0.6736
533	0.4990	0.4621	558	0.3926	0.4436	583	0.3969	0.6570
534	0.7889	0.5086	559	0.3060	0.3606	584	0.2772	0.3767
535	0.4124	0.3343	560	0.4256	0.8017	585	0.3308	0.9394
536	0.5293	0.8682	561	0.4190	0.5996	586	0.4039	0.8853
537	0.5014	0.4636	562	0.3690	0.4730	587	0.4283	0.5150
538	0.4511	0.3922	563	0.4993	0.5808	588	1.0950	0.5652
539	0.5914	0.5727	564	0.3870	0.7183	589	0.3683	0.6238
540	0.7779	0.7028	565	0.4234	0.4325	590	0.2972	0.4123
541	0.3386	0.6240	566	0.3693	0.4383	591	0.4897	1.3868
542	0.5121	0.6445	567	0.6326	0.7247	592	1.0028	0.4721
543	0.3173	0.6724	568	0.5738	0.4642	593	0.4433	0.3791
544	0.4003	0.4694	569	0.3893	0.5322	594	0.5422	1.4432
545	0.6769	0.4458	570	0.5067	1.2544	595	0.3674	0.4449
546	0.3725	0.7556	571	0.3704	0.4125	596	0.6797	0.3444
547	0.5271	0.6609	572	0.3774	0.7378	597	0.5602	0.4345
548	0.5542	0.4687	573	0.3141	0.3323	598	0.3935	0.6358
549	0.5638	0.4080	574	0.5951	0.4546	599	0.4676	0.3232
550	0.8277	0.5822	575	0.6890	0.7386	600	0.4706	0.5247

(Table Continued)

	RMSQ			RMSQ			RMSQ	
ID	NO	C	ID	NO	C	ID	NO	C
601	0.6249	0.5269	626	0.7059	0.5398	651	1.1991	0.4930
602	0.4491	0.6377	627	0.4579	0.4358	652	0.5510	0.4932
603	1.1529	0.6699	628	0.5022	0.4026	653	0.5774	1.4121
604	0.5049	0.3698	629	0.4475	0.5384	654	0.3598	0.4539
605	0.5005	0.5053	630	0.7656	0.5064	655	0.4715	0.3862
606	0.3292	0.3017	631	0.5460	0.5113	656	0.4475	0.4763
607	0.9754	1.0358	632	0.3485	0.4566	657	0.4557	0.7117
608	0.4234	0.8385	633	0.2818	0.4223	658	0.3573	0.5921
609	0.5907	1.5618	634	0.3928	0.4365	659	0.4518	0.4272
610	0.6371	0.3905	635	1.1083	1.3433	660	0.4081	0.6852
611	0.4733	0.6173	636	0.4974	0.4411	661	0.3548	0.4314
612	1.1163	0.6740	637	0.4602	0.5806	662	0.2886	0.4323
613	0.4271	0.4235	638	0.5119	0.6070	663	0.4734	0.4623
614	0.3378	0.3654	639	0.5554	0.4642	664	0.4608	0.6313
615	0.4022	0.5304	640	0.4695	0.3685	665	0.3755	0.6327
616	0.3959	1.1019	641	0.4666	0.5108	666	0.5010	0.3883
617	0.5285	0.3678	642	0.4562	0.3280	667	0.4633	0.7825
618	0.5073	0.6479	643	0.5733	0.5348	668	0.6365	0.4477
619	0.4006	0.8545	644	0.4425	0.4611	669	0.4923	0.4048
620	0.3590	0.5018	645	0.3450	0.3647	670	0.4591	0.3824
621	0.7944	0.4595	646	0.5552	0.3767	671	0.4045	0.8442
622	0.4091	0.4333	647	0.5860	0.8306	672	0.4496	0.4095
623	0.6151	0.5923	648	1.1545	1.0340	673	0.4438	1.2428
624	0.3853	0.3297	649	0.8813	0.5734	674	0.6878	0.3825
625	0.4013	0.3695	650	0.4162	0.7106	675	0.3520	0.4247

(Table Continued)

ID	RMSQ NO	RMSQ C	ID	RMSQ NO	RMSQ C	ID	RMSQ NO	RMSQ C
676	0.5175	0.4360	701	0.3492	0.3766	726	0.4518	0.4591
677	0.3869	0.5271	702	0.3729	0.7366	727	0.4318	0.4596
678	0.8130	0.7478	703	0.7336	0.4154	728	1.0000	0.8259
679	1.1144	0.7321	704	0.3999	0.7488	729	0.4849	0.4412
680	0.4193	0.7533	705	0.3867	0.4493	730	0.8224	0.4688
681	0.4540	0.5462	706	0.7550	0.3321	731	0.6698	0.3329
682	0.4498	0.4457	707	0.3698	0.4777	732	0.3699	0.4892
683	0.7404	0.6320	708	0.3950	0.3131	733	0.6396	0.9016
684	0.3877	0.3957	709	0.5041	0.3538	734	0.6466	0.5727
685	0.5720	0.3315	710	0.3533	0.3701	735	0.3909	0.5951
686	0.6231	0.5114	711	0.6525	0.4807	736	0.3706	0.4592
687	0.4688	0.5679	712	0.4464	0.4661	737	0.3690	0.6153
688	0.4263	0.5242	713	0.4115	0.6227	738	0.2948	1.2515
689	0.4529	0.5456	714	0.5460	0.6682	739	0.4030	0.6052
690	0.3602	0.3761	715	0.5334	0.6178	740	0.6016	0.7011
691	0.6794	0.5675	716	0.5288	0.4098	741	0.3767	0.5149
692	0.3642	0.4764	717	0.4042	0.5252	742	0.3553	0.5773
693	0.3663	0.4190	718	0.3956	0.5175	743	0.5196	0.4092
694	0.7400	0.4890	719	0.6725	0.5424	744	0.6271	0.4214
695	0.4849	0.5245	720	0.3974	0.3931	745	0.6052	0.5933
696	0.5322	0.5375	721	0.9197	0.5540	746	0.8383	0.4806
697	0.7365	0.8902	722	0.3715	0.4707	747	0.5737	0.3375
698	0.5565	0.4070	723	0.5454	0.5239	748	0.4540	0.5480
699	0.6482	0.5719	724	0.4148	0.3599	749	0.3708	0.4612
700	0.7406	0.5889	725	0.9162	0.3455	750	0.6816	0.7706

(Table Continued)

	RMSQ			RMSQ			RMSQ	
ID	NO	C	ID	NO	C	ID	NO	C
751	0.6221	0.5005	776	0.5238	0.5925	801	0.7842	0.6616
752	0.5390	0.5050	777	0.6032	0.3239	802	0.6477	0.7517
753	0.3569	0.4634	778	0.5284	0.4662	803	0.3373	0.5056
754	0.8784	0.6829	779	0.5376	0.4522	804	0.5881	0.5003
755	1.3547	1.2501	780	0.4307	0.8155	805	0.6023	0.4463
756	0.3686	0.3302	781	0.4929	0.4370	806	0.3521	0.4912
757	0.5497	0.5929	782	0.4445	1.0200	807	0.4548	0.4278
758	0.4757	0.6380	783	0.5958	0.3399	808	0.3813	0.6361
759	0.3136	0.5091	784	0.9547	0.7772	809	0.5897	0.4557
760	0.4509	0.4232	785	0.2800	0.4644	810	0.4164	0.4252
761	0.8061	0.4100	786	0.3385	0.9645	811	0.8275	0.4426
762	0.4752	0.9119	787	0.4366	0.5107	812	0.6154	0.4936
763	0.7292	0.5031	788	0.3941	0.4654	813	0.3319	0.6167
764	0.3232	0.3584	789	0.4794	0.5444	814	0.7523	0.5085
765	0.4159	0.5251	790	0.4478	0.6175	815	0.7244	0.5693
766	0.6076	0.4219	791	0.4887	0.3834	816	0.4931	0.4851
767	0.3621	0.5435	792	0.3789	0.6747	817	0.3898	0.4271
768	0.3969	0.4962	793	0.6900	0.2906	818	0.6619	0.4874
769	0.4260	0.6033	794	0.3999	0.5068	819	0.3917	0.6823
770	0.8918	0.6093	795	0.6291	1.0081	820	0.6069	1.2301
771	0.6726	0.8187	796	0.3677	0.4228	821	1.2522	1.0234
772	0.4594	0.5909	797	0.3572	0.4684	822	0.4067	0.3784
773	0.4409	0.4070	798	0.3533	0.3886	823	0.4823	0.5596
774	0.3103	0.4189	799	0.5082	0.5431	824	0.4635	0.4270
775	0.9705	0.8194	800	0.2999	0.4827	825	0.3837	0.4966

(Table Continued)

	RMSQ			RMSQ			RMSQ	
ID	NO	C	ID	NO	C	ID	NO	C
826	0.5298	0.3503	851	0.3929	0.4581	876	0.3642	0.5607
827	0.4863	0.7517	852	0.9524	0.5504	877	0.3734	0.5069
828	0.6748	0.5064	853	0.3969	0.7104	878	1.0135	0.4584
829	0.4349	0.4936	854	0.4200	0.6755	879	0.6694	0.3920
830	1.0445	0.5614	855	1.2949	0.4303	880	0.4521	0.5028
831	0.5086	0.6112	856	0.3975	0.4854	881	0.2746	0.4565
832	0.5174	0.4481	857	0.4453	0.5897	882	1.1361	1.1384
833	0.4911	0.5854	858	0.8359	0.4544	883	0.3522	0.8940
834	0.3985	0.2992	859	0.4021	0.8887	884	0.6468	0.4026
835	0.5191	0.7925	860	0.9390	0.9472	885	1.0181	0.4715
836	0.4322	0.9739	861	0.3375	0.5091	886	0.2983	0.3909
837	0.3779	0.7614	862	0.7245	0.5634	887	0.3326	0.6006
838	0.4476	0.4728	863	0.4166	0.6111	888	0.3867	0.7507
839	0.5493	0.5337	864	0.6740	0.4580	889	0.6281	0.5477
840	0.7100	0.6557	865	0.3320	0.6040	890	0.4843	0.3100
841	0.4250	0.4503	866	0.5490	0.5406	891	0.3119	0.3372
842	0.3942	0.3845	867	0.6974	0.5256	892	0.4075	0.6590
843	0.4365	0.3221	868	0.7183	0.3878	893	0.5587	0.9326
844	0.6598	0.4357	869	0.4899	0.3354	894	0.9261	0.8012
845	0.9200	0.7209	870	0.7681	0.5590	895	0.5130	0.4553
846	0.4414	0.5893	871	0.6039	0.3512	896	0.5837	0.3049
847	0.9161	0.8800	872	0.3709	0.4127	897	0.6801	0.4710
848	0.8186	0.9051	873	0.6360	0.3677	898	0.3362	0.5553
849	0.5009	0.3319	874	0.4426	0.4847	899	0.6574	0.8146
850	0.2913	0.3831	875	0.3836	0.3197	900	0.4051	0.5800

(Table Continued)

	RMSQ			RMSQ			RMSQ	
ID	NO	C	ID	NO	C	ID	NO	C
901	0.3360	0.3900	926	0.8495	0.6931	951	0.5127	0.4511
902	0.4255	0.8205	927	0.4769	0.7735	952	0.4140	0.5139
903	0.5873	0.4663	928	0.5902	0.4006	953	0.4113	0.8127
904	0.4421	0.6898	929	0.9146	0.8510	954	0.8527	0.3448
905	0.4841	0.6168	930	0.3789	0.3331	955	0.3997	0.3791
906	0.4425	0.9773	931	0.4146	0.7439	956	0.6181	0.3323
907	0.3284	0.7909	932	1.1840	1.1645	957	0.5202	0.4448
908	1.0297	0.8107	933	0.4841	0.5015	958	0.4648	0.6604
909	0.5233	0.4282	934	0.4329	0.4139	959	0.4270	0.8167
910	0.3646	0.4590	935	0.4460	0.6004	960	0.6548	0.5716
911	0.5718	1.1404	936	0.6069	0.3523	961	0.4169	0.4954
912	0.4119	0.7588	937	0.4369	0.6674	962	0.4382	0.4592
913	0.3000	0.3181	938	0.6042	0.6207	963	0.5261	0.4653
914	0.6721	0.4646	939	0.4763	0.6461	964	0.3752	0.4103
915	0.4186	0.3936	940	0.5105	0.3619	965	0.4197	0.3442
916	0.7116	0.4704	941	0.4663	0.4165	966	0.5452	0.4572
917	0.3816	0.4044	942	0.6901	0.4084	967	0.6680	0.9679
918	0.3854	0.4938	943	0.3977	0.9915	968	0.3500	0.5552
919	0.5313	0.5010	944	0.5896	0.5299	969	1.0227	0.5203
920	0.4843	0.3286	945	0.5365	0.5055	970	0.5952	0.8635
921	0.4084	0.5856	946	0.4403	0.6431	971	0.4278	0.4883
922	0.4803	0.5812	947	0.5139	0.6539	972	0.4672	0.5677
923	0.5593	0.4101	948	0.7540	0.4305	973	0.4391	0.3404
924	0.8543	0.6233	949	0.4663	0.4752	974	0.4223	0.4903
925	0.3971	0.5205	950	0.6541	0.4002	975	0.4081	0.3028

(Table Continued)

	RMSQ			RMSQ			RMSQ	
ID	NO	C	ID	NO	C	ID	NO	C
976	0.3600	0.3885						
977	1.2621	0.4226						
978	0.4923	0.9849						
979	0.5017	0.4226						
980	0.4629	1.0356						
981	0.4623	0.8314						
982	0.3515	0.5372						
983	0.3950	0.5238						
984	0.7750	0.5366						
985	0.2698	0.9122						
986	0.4821	0.4610						
987	0.3497	0.5732						
988	0.4831	0.4373						
989	0.4653	0.5042						
990	0.6151	0.7958						
991	0.7047	0.6176						
992	0.5494	0.4488						
993	0.5666	0.6346						
994	0.6409	0.4778						
995	0.8120	0.4451						
996	0.7609	0.3720						
997	0.5822	0.4769						
998	0.3609	0.4839						
999	0.5851	0.5459						
1000	0.2684	0.6380						