ABSTRACT

Title of Document: EXPLORING UNIDIMENSIONAL PROFICIENCY CLASSIFICATION ACCURACY FROM MULTIDIMENSIONAL DATA IN A VERTICAL SCALING CONTEXT

Marc Howard Kroopnick, Doctor of Philosophy, 2010

Directed By: Professor Robert J. Mislevy, Department of Measurement, Statistics, and Evaluation

When Item Response Theory (IRT) is operationally applied for large scale assessments, unidimensionality is typically assumed. This assumption requires that the test measures a single latent trait. Furthermore, when tests are vertically scaled using IRT, the assumption of unidimensionality would require that the battery of tests across grades measures the same trait, just at different levels of difficulty.   Many researchers have shown that this assumption may not hold for certain test batteries and, therefore, the results from applying a unidimensional model to multidimensional data may be called into question. This research investigated the impact on classification accuracy when multidimensional vertical scaling data are estimated with a unidimensional model. The multidimensional compensatory two-parameter logistic model (MC2PL) was the data-generating model for two levels of a test administered to simulees of correspondingly different abilities. Simulated data from the MC2PL model was estimated according to a unidimensional two-parameter logistic (2PL) model and classification decisions were made from a simulated bookmark standard setting procedure based on the unidimensional estimation results. Those unidimensional classification decisions were compared to the “true” unidimensional classification (proficient or not proficient) of simulees in multidimensional space obtained by projecting a simulee’s generating two-dimensional theta vector onto a unidimensional scale via a number correct transformation on the entire test battery (i.e. across both grades). Specifically, conditional classification accuracy measures were considered. That is, the proportion of truly not proficient simulees classified correctly and the proportion of truly proficient simulees classified correctly were the criterion variables. Manipulated factors in this simulation study included the confound of item difficulty with dimensionality, the difference in mean abilities on both dimensions of the simulees taking each test in the battery, the choice of common items used to link the exams, and the correlation of the two abilities. Results suggested that the correlation of the two abilities and the confound of item difficulty with dimensionality both had an effect on the conditional classification accuracy measures. There was little or no evidence that the choice of common items or the differences in mean abilities of the simulees taking each test had an effect.

EXPLORING UNIDIMENSIONAL PROFICIENCY CLASSIFICATION ACCURACY FROM MULTIDIMENSIONAL DATA IN A VERTICAL SCALING CONTEXT

By

Marc Howard Kroopnick

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2010

Advisory Committee:
Professor Robert J. Mislevy, Chair
Associate Professor Robert G. Croninger
Professor Gregory R. Hancock
Assistant Professor Hong Jiao
Professor Robert W. Lissitz

Dedication

To everyone who encourages, motivates, inspires, and loves me.

Acknowledgements

Thank you, Dr. Robert J. Mislevy, for your support, generosity, insight, and advice. I could not have asked for a better academic and dissertation advisor.

Thank you, Dr. Gregory R. Hancock and Dr. Robert W. Lissitz, for always looking out for me.

Thank you to all the members of my dissertation committee for your careful review and suggestions for improving the quality of this dissertation.

Thank you to Cisco Systems Inc. and the Department of Measurement, Statistics, & Evaluation for providing financial support while I was a full-time graduate student.

Table of Contents

List of Tables

List of Figures

Chapter 1: Rationale

The ability to measure student growth over time has become increasingly important especially in the context of the No Child Left Behind Act of 2001 (NCLB). Thus students' test scores need to be placed on a common scale in order for grade to grade growth to be measured and compared even if students take different tests. The process for doing this is called vertical scaling (Harris, 2007; Kolen & Brennan, 2004). One technique used to create a vertical scale, as summarized in Skaggs and Lissitz (1986), is Item Response Theory (IRT).

When IRT is operationally applied for large scale assessments, unidimensionality is typically assumed. This assumption requires that the test measures a single latent trait. Furthermore, when tests are vertically scaled using IRT, the assumption of unidimensionality would require that the battery of tests across grades measures the same trait, just at different levels of difficulty. This assumption of unidimensionality in vertical scaling can be unrealistic and problematic in two very distinct ways.

First, assuming the vertically scaled tests are measuring the same trait may be unrealistic if content areas covered on the tests are somewhat different across grades. For example, one might expect that a 10th grade mathematics test with more emphasis on geometry measures something different than an 11th grade mathematics test with more emphasis on algebra even though both are called "mathematics," or that a 4th grade science test with more emphasis on earth science measures something different than a 5th grade science test with more emphasis on simple machines. While there potentially may be reasoning questions across the pairs of tests mentioned above that assess the same trait, those tests, on the whole, may measure different and/or multiple traits and a unidimensional IRT framework would be inappropriate. From an aptitude testing perspective, however, it is possible to conceptualize that the same trait could be measured across grades, but just at different levels of reasoning. That is, there could be general mathematical reasoning tests designed for, say, third and fourth graders. Also, other academic subjects, even in an achievement testing context, might lend themselves to a more static dimensionality structure. For example, English language usage or reading tests may have the same dimensionality across grades or at least for consecutive grades (Loyd & Hoover, 1980; Skaggs & Lissitz, 1988). The skill sets required for English language usage may be relatively more static compared to, say, mathematics or science tests. Reckase (2004) briefly acknowledged this distinction when he noted that vertically scaled reading tests are more likely unidimensional compared to science tests.

Secondly, within a given test, the items may measure different dimensions (Briggs & Wilson, 2003). For example, on an English language usage test, some items may measure listening skills while others may measure writing skills. These skills are most likely indicators of different, but related, traits. Additionally, items on a given test may assess multiple traits simultaneously (whether intentionally or unintentionally) to varying different degrees (Reckase, 1985; Walker & Beretvas, 2003). Thus, a unidimensional IRT model would be inappropriate. Consequently, many researchers acknowledge that the assumption of unidimensionality is often violated on tests (see, for example, Ackerman 1994; Camilli, Wang, & Fresq, 1995; Reckase, 1997).

Given the potential problems with the unidimensionality assumption in vertical scaling, addressing the multidimensionality of vertical scaling data has become a topic of great interest to many researchers (see, for example, Patz & Yao, 2007; Yon, 2006). Tong and Kolen (2007) suggested and encouraged more research on the topic. When attempting to create a vertical scale (even in a unidimensional framework) there are many factors which must be considered. These factors include, but are not limited to, choosing a data collection design, selecting a measurement model, and choosing a calibration method (Harris, 2007). Moreover, if a multidimensional model is considered, the nature of the dimensionality structure must also be addressed (Wang, 1994; Yon, 2006).

While many researchers and practitioners acknowledge the multidimensional nature of data, a unidimensional model is often used for policy and /or practical reasons. Thus, it is important to further investigate the consequences of using a unidimensional IRT model for vertical scaling calibration when the data are more appropriately modeled according to a multidimensional model. Specifically, this research considered the misclassification consequences when multidimensional vertical scaling data were estimated according to a unidimensional model. There is no clear answer to the appropriate methodology for vertical scaling (Harris, 2007; Kolen & Brennan, 2004) and this research only attempted to address and explore a subset of the factors and issues mentioned above.

Chapter 2: Purpose

This research investigated the impact of unidimensional calibration on the classification accuracy of multidimensional vertical scaling data. The multidimensional compensatory two-parameter logistic model (MC2PL, Reckase, 1985) was the data-generating model for two levels of a test administered to simulees of correspondingly different abilities. Simulated data from the MC2PL model were calibrated according to a unidimensional two-parameter logistic (2PL) model (Birnbaum, 1968) and classification decisions were made from a simulated bookmark standard setting procedure based on the unidimensional calibration results. Those unidimensional classification decisions were compared to the "true" unidimensional classification of simulees in multidimensional space obtained by projecting a simulee's generating two-dimensional theta vector onto a unidimensional scale via a number correct transformation for performance on the entire test battery (i.e. across both grades).

Assessing classification accuracy in the context of model misspecification is the biggest practical application of this research because there can be high stakes decisions made based on vertical scaling results. Because unidimensional models are often applied to multidimensional data in real-world vertical scaling applications, the classification consequences of model misspecification and linking item choices are of extreme importance.

The major manipulated factors in the simulation study were:

- The correlation of the two latent dimensions
- Whether or not dimensionality and difficulty are confounded (e.g., easier items load on dimension one and harder items load on dimension two)
- Choice of linking items (even distribution of items from both lower and upper grades tests, only lower grade test items)
- Difference in mean abilities for the two levels (grades) of simulees

The components of the study that were fixed:

- Concurrent calibration/internal common item linking design

- MML estimation in BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1996)
- Multidimensional compensatory two-parameter logistic model for data generation (MC2PL)
- Unidimensional two-parameter logistic (2PL) estimated model
- Test Length (60 total items per grade, 40 unique, 20 common)
- Grade level test design (relationship of items to the two dimensions)
- Total number of common items
- Sample Size (2000 simulees per grade)
- Variance on dimensions remained equal and constant across grades

Chapter 3: Background

No Child Left Behind

The No Child Left Behind Act of 2001 is federal legislation that requires, in part, schools and school districts to meet certain minimum levels of proficiency for adequate yearly progress (AYP). AYP includes reaching specified minimum levels for annual measurable objectives (AMOs) in language arts and mathematics, participation, and the "other academic indicator." AMOs refers to the percentage of students classified as (at least) proficient on the state's assessments for language arts and mathematics. Note that it is the state, not the federal government, that designs the language arts and mathematics tests and sets the standards for proficient. Thus, the standard-setting for these exams is a critical component in the context of this legislation.

Item response theory

Item response theory includes a class of item response models that describe the relationships of test performance and the unobservable traits or abilities that underlie that performance. Specifically, the models express the probability of a particular response to an item as function of examinee and item parameters which are calibrated onto an unobservable latent trait (ability) continuum. There are item response models for items that are dichotomously and/or polytomously scored as well as for single or multiple traits (dimensions). This research focused on dichotomous item responses modeled according to a multidimensional IRT model, but estimated according to a unidimensional IRT model.

IRT is governed by three major assumptions. The first is that the dimensionality of the response function is properly specified. That is, the appropriate number of dimensions is expressed in the item response model. The second is local independence which means that responses by examinee *n* to item set **I** are independent, conditional on the ability parameter(s) for examinee *n*. The last is examinee response independence which requires that responses by examinees are independent from each other (Hambleton & Swaminathan, 1985). Further discussion of these assumptions (especially in the context multidimensional item response theory) can be found in Embretson and Reise (2000) and Reckase (2009). This research focused on the violation of the first assumption regarding dimensionality.

Two parameter logistic model

The unidimensional item response model for dichotomous responses considered in this study is the 2 parameter logistic (2PL) model (Birnbaum, 1968). The parameterization for this model is as follows:

$$P(x_{ij} = 1|\theta_j, a_i, b_i) = \frac{\exp(a_i(\theta_j - b_i))}{1 + \exp(a_i(\theta_j - b_i))}, \qquad (1)$$

where θ_j is the ability parameter for person *j*, a_i is the discrimination for item *i*, and b_i is the difficulty parameter for item *i*. When this model is estimated in the context of vertical scaling and it is known that different subpopulations are responding to the items, it is necessary to define and control for separate ability distributions in the estimation (Camilli, Yamamoto, & Wang, 1993; Bock & Zimowski, 1997). BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1996) is capable of this type of estimation and was used in this study.

Multidimensional compensatory two parameter logistic model

The multidimensional item response model for dichotomous responses considered in this study was the *m*-dimensional compensatory 2 parameter logistic (MC2PL) proposed by Reckase (1985); it is the multidimensional extension of (1). In this model, each item is allowed to discriminate on all dimensions to varying degrees and a person's ability on one dimension can compensate for a deficiency on the other (specifically, in this study two dimensions were considered). Note that in a noncompensatory model, a person's ability on one dimension cannot compensate for a deficiency on the other. The parameterization for the MC2PL model is as follows (Ackerman, Gierl, & Walker, 2003):

$$P(x_{ij} = 1|\boldsymbol{\theta}_j, \boldsymbol{a}_i, d_i) = \frac{\exp(\sum_{k=1}^{m} a_{ik}\theta_{jk} + d_i)}{1+\exp(\sum_{k=1}^{m} a_{ik}\theta_{jk} + d_i)}, \qquad (2)$$

where $\boldsymbol{\theta}_{\mathrm{j}}$ is the vector of *m* ability parameters for person *j*, $\mathbf{a}_{\mathrm{i}}$ is the vector of discrimination parameters for item *i*, x_{ij} is the response of person *j* to item *i*., and d_i is the parameter related to difficulty. Note, however, the sign on d_i is positive while in the traditional IRT framework, it is negative. It can also be understood as follows (in the equation below) where b_{ik} is interpreted as a unidimensional IRT difficulty parameter, like in the 2PL model. Note, however, that for identification reasons, these *b* parameters are usually not estimated:

$$d_i = -\sum_{k=1}^{m} a_{ik} b_{ik} \;. \qquad (3)$$

Each multidimensional item can be described by three summary characteristics: discrimination, difficulty, and location. Discrimination is a function of the individual discrimination parameters (**a**$_i$), and represents the maximum amount of discrimination. It is referred to as MDSIC and is expressed as follows in the two dimensional case:

$$MDISC_i = \sqrt{a_{i1}^2 + a_{i2}^2}\ . \qquad (4)$$

Difficulty represents the distance from the origin of the two dimensional axes to the line representing the composite of abilities required to have a 50% probability of answering the item correctly. The sign of this value indicates relative difficulty where negative values are relatively easy and positive values relatively hard. It is referred to as *D* and is expressed as follows:

$$D = \frac{-d_i}{MDISC}\ . \qquad (5)$$

Location corresponds to the direction of each item relative to the positive $\boldsymbol{\theta}_1$ axis. Items with a small angle primarily measure $\boldsymbol{\theta}_1$ and those with a larger angle primarily measure $\boldsymbol{\theta}_2$**.** It is referred to as $\boldsymbol{\alpha}$ and expressed as follows:

$$\boldsymbol{\alpha}_i = \arccos \frac{a_{i1}}{MDISC_i}\ . \qquad (6)$$

Vertical scaling

“Vertical scaling refers to the process of linking different levels of an assessment, which measures the same trait, onto a common scale (Harris, 2007, p. 233).” Thus, vertical scaling provides a method for measuring growth across grades which is necessary in the current educational climate where emphasis is placed on student growth through adequate yearly progress measures mandated by the NCLB legislation (Harris, 2007). Note that Kolen and Brennan (2004, p. 372) indicate that tests in a vertical scale are intended to measure *similar* constructs which is slightly in contrast to Harris' notion of tests in a vertical scale measuring the *same* trait. While NCLB does not require vertical scales, they can be utilized not only to measure adequate yearly progress, but also for evaluation and accountability purposes in school systems. Currently only a few states have a vertical scale for their NCLB assessments (Florida and Michigan, for example) and this is perhaps due to the difficulty of developing curricula that in fact measure a single trait over grades; as states revise curricula, this may change. Curricula for English language learners, however, tend to be more consistent with a single trait measured over time (Yen, 2007).

In constructing a vertical scale many decisions need to be made and there is no one standard procedure. These decisions include, but are not limited to, determining the data collection design, measurement model, and calibration method (Harris, 2007). Unidimensional IRT is an increasingly popular class of measurement models used for vertical scaling (other methods include, Thurstone and Hieronymus scaling) and was the focus in this research.

The data collection designs for vertically scaled tests include the common item design, equivalent groups design, and the scaling test design (Kolen & Brennan, 2004). The easiest and most straightforward of these to implement is the common item design (Kolen & Brennan, 2004) and thus was the focus of this research. The common item design simply requires that every pair of adjacent tests include a common block of items; a chaining processing using these common item blocks is used to create the single vertical scale. Determining how many items to include on the common item blocks, however, is not necessarily straightforward. This will especially be the case when common items are indicators of different dimensions. There are, however, some recommendations for creating these common item blocks.

Kolen and Brennan (2004) suggested that at least 20% of a test with at least forty items be common items in a horizontal equating context and Young (2005) suggested that rule can be adapted for vertical scaling by increasing the number to help account for the differences in difficulty and content across grades in a vertical scaling context (McBride & Wise, 2000). For this research 33% of a test was common across the two grades; however, the dimensionality of the items varied. Ideally, the dimensionality of the items should match the dimensionality of the overall test and that condition was considered; that is, where common items function like a “mini-test.” However, some school districts might not have enough items to have a common item “mini-test” or might not put very much thought into their choice of common items, so the situation where common items are not necessarily reflective of the entire test battery was also considered in this investigation.

When using the common item data collection design, there are two item calibration methods available: separate and concurrent. Separate calibration involves calibrating each grade individually and using the common item parameter estimates and a scale transformation method (Mean-Mean or Stocking-Lord, for example) to establish the vertical scale. The concurrent calibration method involves calibrating all items across all grades simultaneously and imposing a multi-group IRT model to account for the multiple grades; items not administered to certain simulees are simply treated as not reached.

In a unidimensional IRT framework, research suggests that separate calibration produces more accurate results relative to concurrent calibration when the measurement model is misspecified. However, concurrent calibration is superior when the model is correctly specified largely because there would be one set of parameter estimates (based on larger sample) for the common items (Patz & Hanson, 2002; Patz & Yao, 2007).

Concurrent calibration rather than separate calibration was considered in this study to reflect applications where the estimated unidimensional model is assumed to be approximately correct and there is reasonable justification for its use in both applied and research contexts.

McCall (2007) presented an overview of vertical scaling entitled “Vertical Scales and the Development of Skills.” Her presentation included a summary of methods used to create and maintain vertical scales. She specifically noted CTB as a company that uses concurrent calibration for vertical scaling.

CTB (2005) developed the Wisconsin Knowledge and Concepts Examination-Criterion Referenced Tests (WKCE-CRT). This test was vertically scaled and CTB evaluated 4 methods to determine the appropriate calibration methodology. These methods spanned a fully concurrent calibration to fully separate calibrations across all grades. They ultimately decided on a compromise solution where both concurrent and separate calibrations were used. Specifically, they conducted a concurrent calibration for grades 5-7, another concurrent calibration for grades 3-4, and yet another concurrent calibration for grades 8-10. Then the results from the grades 3-4 and 8-10 calibrations were placed on the grades 5-7 scale. Thus, this example shows the use of concurrent with adjacent grades in an operational setting.

Lastly, in simulation research addressing the usefulness of multidimensional IRT models relative to unidimensional IRT models to estimate achievement gain in a vertical scaling context, Reckase and Li (2007) used concurrent calibration to develop all of their vertical scales. The motivating research hypothesis for this study was that multidimensional models would be more able to capture achievement gain when math content specifications change over time and the tests may become multidimensional in nature. Note that their research included both of the models considered in this study, the MC2PL and the unidimensional 2PL model.

Vertical scaling also typically requires some assumption on grade-to-grade variability. For this research variance on each dimension was assumed equal and constant across dimensions and grades (i.e., the same on all dimensions for all grades) and there is justification for this assumption. Harris (2007) noted there are many inconsistencies in the literature regarding grade-to-grade variability. On page 234, she specifically cites Bock (1983) as an example where “grade-to-grade variability…..was shown to remain stable across grade levels.” In Bock’s (1983) work, he scaled a cross sectional sample of scores for the Stanford-Binet test. He showed that by not constraining the dispersions of abilities a very plausible mental growth curve emerges. He also noted that the within-age standard deviations of the developmental age scores were homogenous.

Further, research by Linn (1989) and Williams, Pommerich, and Thissen (1998) shows that, in a vertical scaling context, standard deviations do not necessarily systematically increase or decrease across grades. Specifically, Linn observes on the NAEP reading scales a “small” decrease in variability for grades 3-7 and a “slight” increase for grades 7-11. Williams, Pommerich and Thissen found no evidence that the variability of performance on vertically scaled math achievement tests in North Carolina "consistently increased or decreased across grades with IRT scaling techniques." While it wasn't the main focus of their 1998 work, Williams et al. also noted that the standard deviations on North Carolina reading achievement tests also "showed no systematically increasing on decreasing trends."

Standard setting

The criteria for placing examinees into performance categories based on their test scores are the results of a standard-setting procedure. Typically performance standards are established based on either the test items or the examinees taking the test. This research focused on the former which are commonly called "test-centered" approaches to standard setting. Specifically, the Bookmark Procedure (Lewis, Mitzel, & Green, 1996) was considered.

The Bookmark procedure as described by Lewis, Mitzel, and Green (1996) is an IRT based approach where items on a fixed form test are ordered by location on the latent continuum (as determined by their item parameter estimates) into a test booklet. Standard setting panelists are required to place a "bookmark" between the most difficult item a minimally proficient examinee (as defined by established performance level descriptors) would be expected to answer correctly and the easiest item a minimally proficient would be expected to answer incorrectly. Panelists are typically instructed to use a .67 probability of success rate. As with most standard setting procedures, this procedure is iterative and allows for revision and discussion; impact data can also be presented and result in cutscore adjustments (Lewis et al., 1996). Zieky (2001) highlights that the booklet with IRT calibrated items is an effective way of presenting normative data to the panelists and may be useful in their decision making process. Since a cutscore decision is based, essentially, on a group of items after considering the entire item booklet, Lewis et al. (1996) note that the resulting cutscore is based on a "comprehensive understanding of test content." Note also that the Bookmark Procedure can accommodate polytomous and constructed response items by placing the various response categories to a given item at its appropriate location in the item booklet. Since item difficulties are on the same scale as person ability, the location of the bookmark can be easily translated to a cutscore on the score reporting scale using test characteristic curve methods (Lewis et al., 1996). A simulated version of this procedure was implemented for this research study and is described in the Methodology section.

Classification accuracy

Betebenner, Shang, Xiang, Zhao, and Yue (2008) noted that while there is inconsistency on terminology and notation in the classification accuracy measures literature, there are two main approaches. The first approach at determining classification accuracy considers the probability of correct classification across all performance levels and is defined in Betebenner et al. (2008) as follows:

$$\sum_{i=1}^{k} \Pr(A^* = i, A = i)\,, \qquad (7)$$

where *k* represents the number of performance categories, *A* represents the true performance classification, and *A** represents the observed performance classification. The probability values in (5) are based on the joint distribution of observed and true classifications which can be expressed as:

$$\Pr(A^* = i, A = j) \text{ over } _{1 \le i,\, j \le k} \qquad (8)$$

False-positive and false-negative rates can also be calculated; in the two category case, the false positive rate is simply the proportion of individuals in the sample classified as proficient who truly are not proficient and the false-negative rate is simply the proportion of individuals in the sample classified as not proficient who truly are proficient.

The second approach considers not the joint probabilities of observed and true classification, but rather the conditional classification probabilities (Betebenner et al, 2008). A misclassification matrix, **P** (Clauser, Margolis, & Case, 2006), can be created that includes (conditional) false- positive and false negative rates as well as (conditional) correct classification rates. Following from Betebenner et al. (2008) **P** can be expressed as follows:

$$\mathbf{P} = \{p_{ij}\}_{1 \le i, j \le k} \text{ where } p_{ij} = \Pr(A^* = j \mid A = i). \qquad (9)$$

If we consider the circumstance where there are two performance levels, proficient and not proficient, and $i = 1$ for proficient and $i = 2$ for not proficient then the (conditional) false-positive rate can be expressed as $\Pr(A^* = 1 \mid A = 2)$ and the (conditional) false-negative rate can be expressed as $\Pr(A^* = 2 \mid A = 1)$. The conditional correct classification rates would therefore be expressed as $\Pr(A^* = 1 \mid A = 1)$ (or the true positive rate) and $\Pr(A^* = 2 \mid A = 2)$ (or the true negative rate).

The classification accuracy measures based on both approaches can be directly calculated from a standard contingency table. For the two performance level case, Douglas (2007) provided the following contingency table:

Table 1: Classification contigency table

		Observed status		
		Proficient	Not Proficient	Total
TRUE STATUS	Proficient	*a*	*b*	*g*
	Not Proficient	*c*	*d*	*h*
	Total	*e*	*f*	*N*

The classification accuracy measures based on the joint distribution and based on two performance level case can be calculated as follows:

$$\text{Percent Correctly Classified} = \sum_{i=1}^{2} \Pr(A^* = i, A = i) = \frac{(a+d)}{N} \qquad (10)$$

$$\text{False-positive rate} = \Pr(A^* = 1, A = 2) = \frac{c}{N} \qquad (11)$$

$$\text{False-negative rate} = \Pr(A^* = 2, A = 1) = \frac{b}{N} \qquad (12)$$

The misclassification matrix, based on conditional probabilities, can be calculated as follows:

$$\mathbf{P} = \begin{matrix} p_{11} = \dfrac{a}{g} & p_{12} = \dfrac{b}{g} \\ p_{21} = \dfrac{c}{h} & p_{22} = \dfrac{d}{h} \end{matrix} \qquad (13)$$

As Douglas (2007) notes, the measures based on the joint distribution depend on one another. That is, as the classification accuracy measure defined in (7) changes, the measures in (8) and/or (9) will necessarily change. This is not true for measures based on the conditional probabilities; changes in rates associated with truly proficient examinees do not affect rates associated with truly not proficient examinees. Accordingly, results based on the analyses that considered the conditional probabilities were the primary focus of this study; the analyses based on the joint distributions were treated as supplementary.

In a simulation framework these calculations are straightforward because the researcher has the benefit of knowing both the true classifications of the simulees as well as the (simulated) observed classification. However, in operational settings educational statisticians only know the observed classifications and are forced to use models to estimate the true classifications of examinees. As described in Betebenner et al. (2008), Livingston and Lewis (1995), in a classical test theory framework, used a four parameter beta true score distribution with a binomial/compound error distribution. In an IRT framework, Rudner (2001, 2005) used a normal distribution for both the error and true score distributions.

Chapter 4: Literature Review

Literature addressing the question of classification accuracy when a unidimensional model estimated with multidimensional data is sparse. This literature review largely serves to make the case for the presence of multidimensionality in applied vertical scaling settings as well as to describe some of the factors manipulated in this simulation study.

Multidimensionality in vertical scaling

Studies have evaluated the effectiveness of unidimensional IRT models for vertical scaling with the hypothesis that multidimensionality might have an adverse impact. Loyd and Hoover (1980) explored vertical scaling using the 1 parameter logistic model (1PL, also called the Rasch model) using item response data from 6th, 7th, and 8th grade students on three corresponding levels (12, 13, 14) of the mathematics computation portion of the Iowa Tests of Basic Skills (ITBS). Note that the 1PL model is a special case of the 2PL model considered in this study where the discrimination parameter, *a*, is constrained to be equal across all items. In order to facilitate the linking of tests, 30 items were in common for adjacent levels and 15 items were in common for nonadjacent levels and linear transformation constants of the estimated common item parameters were used to establish a common scale. The primary focus of this study was to evaluate differences in the vertical scaling results when different calibration groups are used. That is, they compared the equating functions for levels 12-14 that resulted from calibrations using the 6th, 7th, and 8th grade students. Functions for both adjacent and nonadjacent grades were compared. All three levels shared at least 15 common items, so both direct and indirect linking was compared. Items parameter estimates were calibrated separately for each calibration group on each level.

The resulting equating functions were not invariant. They generally found that when examinees take a lower level of the test and have their scores equated to the higher level raw score scale, the results will be higher if the items were calibrated with the higher ability group. Further, they found that when examinees take a higher level test and have their scores equated to a lower level raw score scale, the results will be higher if the items were calibrated with the lower ability group (Loyd & Hoover, 1980).

The authors suggested that the root of these inconsistencies could be the violation of IRT model assumptions; primarily, they were interested in the degree to which unidimensionality was met. A item-level factor analysis of level 13 item responses from 6th and 7th graders revealed one primary dimension, but nontrivial secondary or minor dimensions. The authors suggested that an item may assess multiple dimensions and an examinee may or may not have been exposed to those dimensions depending on their curriculum. A skills analysis of the items across the test levels suggested that certain topics could have been covered at various points in the 6th through 8th grade mathematics curriculum and that topic emphasis varied across levels. The authors hypothesized that items drawing from different and multiple dimensions could be the reason for the lack of invariance of equating functions (Loyd & Hoover, 1980).

Harris and Hoover (1987) followed-up the Loyd and Hoover (1980) study with an investigation of effectiveness of using the three parameter logistic (3PL) model for vertical scaling using the mathematics computation portion of the ITBS and 3rd through 8th grade students in Iowa. They considered an expanded number of levels, though; in addition to levels 12-14, they considered levels 10 and 11. Their approach differed from Loyd and Hoover (1980) in how the scale was set; in order to establish a single scale for the vertical scaling of the test battery, all items and examinees were estimated simultaneously in LOGIST (Wingersky, Barton, & Lord, 1982) with modifications for omitted and not-reached items. The resulting theta estimates were treated as truth in subsequent item parameter estimation for a given grade and test level. For each grade by test level, test characteristic curves were computed. Results indicated that an examinee would receive a higher theta estimate if the test s/he was administered was calibrated on lower ability students. Across levels, using the equating of level 12 to 13, it was shown that the equating relationship varied based on the groups used to establish it. So, despite a different IRT model, results from this study were consistent with Loyd and Hoover (1980) that vertical scaling calibration results, at least with the ITBS data considered, were not invariant (i.e. person-free). However, it is important to note that while the patterns between the 1PL model and 3PL model were consistent, the actual equating results were not; thus, depending on the estimated model, different conclusions would be made about a given examinee. Again, Harris and Hoover acknowledged that multidimensionality might be a reason for this result, but did not conduct any follow-up dimensionality assessment on the items.

Acknowledging the invariance issues related to lack of person-fit with operational vertical scaling data as in Loyd and Hoover (1980) and Harris and Hoover (1987), Skaggs and Lissitz (1988) conducted a simulation study to evaluate this issue where they had control of the data. In their study, vertical data were simulated from a 3PL model, linked via an external anchor item test, estimated with 1PL and 3PL models, and equated with IRT true-score and equipercentile methods. Simulation factors manipulated included the difficulty, discrimination, and guessing item parameters of the 3PL model for each test in the vertical scaling battery as well as the ability distributions of simulees by grade. Largely, they found invariance in vertical scaling with respect to simulee ability. They noted that invariance may not hold when other modeling assumptions are not met such as equal discrimination when the 1PL model is estimated. They concluded by agreeing with previous researchers that multidimensionality could be the reason for lack of invariance of equating functions with respect to simulee ability in vertical equating. More generally, they suggested that vertical equating with unidimensional models should be approached cautiously because of the threat of multidimensionality and that the issue of dimensionality in vertical scaling deserves further investigation. Also note that in addition to lack of invariance of equating functions, multidimensionality has also been hypothesized to be the cause of scale shrinkage in vertical scaling (Camilli, Wang, & Yamamoto, 1993; Yen,1985).

Unidimensional calibration of multidimensional items

Work by Ansley and Forsyth (1985), Way, Ansley, and Forsyth (1988), and Ackerman (1989) provides a reasonable foundation for understanding the consequences of estimating a unidimensional IRT model with multidimensional IRT data. Ansley and Forsyth (1985) considered the implications of estimating a unidimensional 3PL model to data generated from the noncompensatory multidimensional extension of the 3PL model proposed by Sympson (1978). The data generating model had 2 dimensions and data were generated under a variety of levels of correlation between the dimensions. They generally found that the unidimensional discrimination parameter estimate was approximately the average of the true discrimination parameters, the unidimensional difficulty parameter was an over estimate of the true difficulty for dimension 1, and the unidimensional ability estimate was highly related to the average of the true ability parameters. Way, et al. (1988) extended the work of Ansley and Forsyth (1985) by also considering the multidimensional compensatory IRT model proposed by Reckase (1985). They found for data generated from this compensatory model that the unidimensional discrimination parameter estimate appeared to be the sum of the true discrimination parameters, the unidimensional difficulty parameter estimate appeared to be the average of the true parameters, and the unidimensional ability parameter estimate appeared to be highly related to the average of the true ability parameters. They noted that the degree of this relationship remained static regardless of the level of relationship among the latent dimensions. The relationship became stronger in the noncompensatory model as the correlation of latent dimensions increased.

The degree to which the level of difficulty is related to dimensionality and the corresponding consequence on the classification accuracy is also a very important issue to be investigated especially given the risks of multidimensionality in a vertical scaling context. Reckase (1985) showed that dimensionality can be confounded with difficulty and Reckase et al. (1986) illustrated that when difficulty and dimensionality are confounded and a unidimensional model is estimated, the ability estimate has different meanings at different points on the unidimensional latent scale. Furthermore, Reckase (1990) notes that a unidimensional model will fit reasonably well when dimensionality was confounded with difficulty. Many of the results of Reckase (1985, 1986, 1990) are summarized briefly in Ackerman (1989) and Walker and Beretvas (2003).

Ackerman (1989) evaluated the effects of unidimensional IRT calibration of compensatory and noncompensatory multidimensional item response models when difficulty was confounded with dimensionality. He generally found that as the correlation of the latent dimensions increased, the response data became more unidimensional. The results in this study were comparable to the results from Way, Ansley, and Forsyth (1988) with differences attributed to the disparity in the parameter generation used in the two studies. Ackerman (1989) also noted that BILOG (Mislevy & Bock, 1982) appeared to be more sensitive to the confounding of difficulty and dimensionality compared to LOGIST (Wingersky, et al., 1982).

Classification accuracy with multidimensional data

Very little applied or simulation research has been conducted to evaluate the classification accuracy of multidimensional data when a unidimensional model is estimated. Only three relevant studies have been found and they will be briefly described here.

Mignani, Monari, Cagnone, and Ricci (2006) conducted a simulation study to compare the classification results when a unidimensional 2PL model was estimated for data generated from a 2-dimensional MC2PL model versus those estimated from the properly specified model. They considered three distinct types of 2-dimensional models: between-items, within-items, and a mixture of between and within-items. “Between-items” describes the situation where a test measures multiple dimensions, but each item is only an indicator of one. “Within-items” describes the situation where a test measures multiple dimensions and an item can be an indicator of multiple dimensions. The ability parameters were generated from a standard multivariate normal distribution with zero correlation between the dimensions. Classification into two categories was based on whether or not the single ability estimate (unidimensional model) or average of the ability estimates (2-dimensional model) was greater or less than 0. They found the highest correspondence of classification results for the within-items model and the poorest correspondence for the between-items model.

Walker and Beretvas (2003) compared classification results based on ability parameters estimated from unidimensional 3PL model versus those estimated from a 2-dimensional compensatory item response model for mathematics test data. In the 2-dimensional model, all items were indicators of mathematics ability and a subset of those items were also indicators of mathematics communication ability. The authors generally found that examinees with low mathematics communication ability tended to be classified at lower levels under the unidimensional model than on the general mathematics ability dimension of the 2-dimensional model. However their multidimensional classification categorizations relied on response patterns associated with getting the “easiest” items correct based on a unidimensional calibration; those “easiest” items would not necessarily be the same if a multidimensional model were used to rank the items.

Lau (1996) investigated, using Monte Carlo methods, classification accuracy based on the sequential probability ratio testing procedure (SPRT) in the context of computerized mastery testing when data were modeled according to a multidimensional model but item parameters estimated according to unidimensional IRT models.

Specifically he considered a 2-dimensional 3-parameter compensatory IRT model. Essentially, this is an extension of the Reckase (1985) MC2PL where a guessing parameter is added. Item response patterns simulated from this model were calibrated according to both 3PL and 1PL unidimensional models. Additionally, he varied the correlation among the latent dimensions, test length, and cut score. Summarizing from Lau (1996), the findings generally suggested that the SPRT was robust to model-misspecification and resulted in acceptable classification accuracy rates. However, the unidimensional models varied in their test length efficiency where the 3PL model resulted in shorter test lengths required for a mastery decision than the 1PL model. Some bias in the cut-score determined by the unidimensional parameter estimates was detected. Lau noted that this bias could result in differential classification errors.

Chapter 5: Reasonableness of Simulation Conditions

This section serves to provide support for various simulation conditions specified in this research. Specifically, descriptions of research designs and operational testing programs are discussed.

Reckase and Li (2007)

As mentioned briefly earlier, Reckase and Li (2007) investigated, in a vertical scaling context, achievement gain when math content specifications change via a simulation study. Their study argued for the appropriateness of multidimensional IRT models for estimating gain in this context and the generating parameters for their study were 3-dimensional and realistic (i.e. based on an analysis of real 6th and 7th grade data). The MC2PL model was used as the data generating model just like was used in this dissertation. Thus, it was useful to consider their item parameters and the correlation among latent abilities when determining the various elements of the simulation study conducted in this dissertation . The three dimensions of interest in their study were Algebra, Arithmetic, and Problem Solving and the associated correlation matrices for abilities in each grade are as follows:

Table 2: Relationship of mathematics dimensions for 6th and 7th grades

Correlation Matrix: 6th Grade			
	algebra	arithmetic	problem solving
algebra	1.00	0.00	0.00
arithmetic		1.00	0.71
problem solving			1.00
Correlation Matrix: 7th Grade			
	algebra	arithmetic	problem solving
algebra	1.00	0.52	0.60
arithmetic		1.00	0.39
problem solving			1.00

The correlation matrix above used in the Reckase and Li simulation study was based on an actual calibration of vertical scaling item response data. While the algebra dimension was estimated for sixth graders, no sixth grade items loaded strongly on that dimension. Thus, the correlations of algebra with the other dimensions were extremely low; for simplicity, Reckase and Li used a zero correlation of algebra with arithmetic and algebra with problem solving in their simulation data generation. (Note that since the item parameters used in the simulation did have discrimination values (albeit, very low) on the algebra dimension is why variance was modeled on that dimension for 6th graders).

The range of correlations here suggests that .3 and .6 levels for the correlation among dimensions are reasonable. Further, this example suggests that in future research varying the correlation among dimensions across grades might be interesting.

The relationship (correlation) among the difficulty parameters and the three ability dimensions considering just 6th grade items, 7th grade items, or the entire test battery ranged from approximately 0 to .65 (in absolute value). This suggests that the moderate and no confound of difficulty with dimensionality levels are reasonable.

Programme for International Student Assessment

Another resource that is useful for justifying the plausibility of generating item parameters is the PISA (Programme for International Student Assessment) test battery. Following from the 2003 Technical Report, the general goal of PISA is to assess how well 15 year old students are prepared for the real-world. As such, the test items are not necessarily curriculum specific, but rather address the students' ability to apply knowledge to real situations. The tests administered include Reading, Math, Science, and Problem Solving. The multidimensional random coefficients multinomial logit model (Adams, Wilson, & Wang; 1997) was used to scale the items. The test battery was not vertically scaled, however.

A seven-dimensional scaling was conducted. There were four math dimensions, as well as a dimension for reading, problem solving, and science. Each item loaded (discriminated) on only one of the dimensions (simple structure) and because the model used is a Rasch model, the loadings were equal. Only the four math dimensions were considered to justify conditions in this study.

The four math dimensions were Change and Relationships (CR), Uncertainty (C), Space and Shape(SS), and Quantity (Q). The bivariate correlation among any two of these dimensions ranged between .88 and .93. These results suggest a realistic context for a .9 correlation between dimensions. Further, the correlation of difficulty with dimensionality for any of the four dimensions was no greater than .25 in absolute value and the average difficulty on the four dimensions were as follows: 0.1 (CR-22 items), 0.22 (U-20 items), 0.19 (SS-20 items), and -.48 (Q-22 items). So, for the most part, the average item parameter estimates are approximately equal in difficulty (except for the Q dimension) and it follows that there is very little confound of difficulty with dimensionality. To the extent that the four dimensions captured here could be taught in different orders in consecutive grades suggest the reasonableness of a level where difficulty is not confounded with dimensionality.

NAEP

The dimensionality structure of the 1990 NAEP math items (Abedi, 1994) was considered. The five dimensions evaluated were Numbers, Measurement, Geometry, Statistics, and Algebra. Largely, he found that the correlations between dimensions were high, ranging between .83 and 1. However, when controlling for background variables such as the students opinion of their ability at math, the correlations were more varied. For students who were undecided on the phrase "I am good at math", the correlations across dimensions were still high, but ranged between .68 and 1.00. When students agreed with the statement, the correlations between dimensions ranged between .85 and 1. When students disagreed with the statement, the correlations between dimensions ranged between .77 and 1. Thus, it is conceivable for school districts where the majority of students fall into one of these three categories we could expect the range in correlations among certain math dimensions to be between .68 and 1. [Geometry typically had the lowest correlation with the other dimensions.] This finding supports using .6 and .9 levels for the correlation between dimensions.

Paris (2005)

Lastly, reading skills development was considered to help inform certain elements of the simulation design. Paris (2005, p. 184) argues that in reading acquisition, skills can fall into two major categories: constrained and unconstrained. Constrained skills are "learned quickly, mastered entirely, and should not be conceptualized as enduring individual difference variables." Examples of constrained skills include letter knowledge and phonics. Unconstrained skills on the other hand, "continue to develop throughout the life span, are not identical across people, and may benefit from special practice and idiosyncratic experiences at many points in the life course." Examples of unconstrained skills include vocabulary and comprehension. The major point of the article is that different skills develop in different ways (trajectories) and the type of analyses conducted on reading acquisition should adjust to the types of skills. So, unconstrained skills can be analyzed with traditional parametric and normal distribution theory methods, but it might be more appropriate to analyze constrained skills with nonparametric methods such as conditional probability, contingency tables, and log linear models. Paris makes a strong argument that researchers typically (and incorrectly) analyze constrained skills with parametric methods which makes the associated conclusions suspect.

For this dissertation, it is not necessary to accept the substance of Paris's arguments with respect to reading. What is important is that the kinds of patterns he sees in data, and the kinds of patterns that are central to his research, are ones that are consistent with the structures in the proposed simulation design. Implied in Paris' work is that reading tests are multidimensional in nature and their dimensionality may change over time. That is, for consecutive grades, given reading tests may assess phonemic awareness and comprehension and there may be reasonable variability on both dimensions across grades; however, for more advanced grades, similar tests may have little or no variability on the phonemic awareness dimension (i.e. it has been mastered). [Note that Paris argues phonemic awareness is more constrained than comprehension.] So, this article strongly suggests that reading skills are multidimensional and that it is possible for there to be variability on a number of dimensions and no variability on others at different points in reading development; generally, the dimensionality structure could change over time as certain skills are mastered. Thus, this research provides a reasonable context for studying the multidimensionality of tests that are vertically scaled. The extent to which some skills are mastered more quickly (constrained skills) than others is an argument for difficulty confounded with dimensionality—under the presumption that the constrained skills are easier. Alternately, Paris suggests that some skills like comprehension and decoding may develop simultaneously which might indicate the reasonableness of difficulty not confounded with dimensionality.

Chapter 6: Methodology

This section includes the research questions addressed in this study, the manipulated factors, the simulation procedures, and a description of the outcome measures.

Research questions

1. Does the correlation between latent dimensions affect proficiency classification accuracy when vertical scaling data modeled according to the 2-dimensional MC2PL model is calibrated according to the 2PL model?
2. Does the confound of dimensionality with item difficulty affect proficiency classification accuracy when vertical scaling data modeled according to the 2-dimensional MC2PL model is calibrated according to the 2PL model?
3. Does the discrepancy in mean ability of the two groups (grades) affect proficiency classification accuracy when vertical scaling data modeled according to the 2-dimensional MC2PL model is calibrated according to the 2PL model?
4. Does the choice of common items affect proficiency classification accuracy when vertical scaling data modeled according to the 2-dimensional MC2PL model is calibrated according to the 2PL model?

Factors

In order to address the questions posed above, four factors were manipulated:

1. Discrepancy of ability distribution means between grades (2 levels)
2. The confound of difficulty with dimensionality (3 levels)
3. The correlation of the latent ability dimensions (4 levels)
4. The nature of the common item sets (2 levels)

Thus, the simulation study had 48 conditions.

Differences in ability

The ability difference between grades in vertical scaling varies from application to application. This research investigated two levels of ability difference. For the first level, the lower grade ability distribution was MVN with mu = {0.0, -0.2} and the upper grade ability distribution was MVN with mu = {0.4, 0.0}. For the second level, the lower grade ability distribution remained MVN with mu = {0.0, -0.2}, but with the upper grade ability MVN with mu = {0.8, 0.2}. For all levels the variance on any given dimension was 1 and the correlation between dimensions varied as discussed later in this section. The multivariate differences in ability between grades largely followed from Yon (2006).

Confound of item difficulty with dimensionality

Three levels of confound of difficulty with dimensionality were considered; no confound, moderate confound, and high confound. Tests where there was a high confound of difficulty and dimensionality were constructed such that the multidimensional difficulty value, d_i, associated with each item was highly correlated with the dimension for which the item was a primary indicator. Most of the lower grade items were primary indicators of dimension one and most of the upper grade items were primary indicators of dimension two. Thus, easier (positive) *d* values were associated with dimension one and harder (negative) *d* values were associated with dimension two. For tests with no confound the relationship of the *d* parameter and dimension was determined randomly. When there was a moderate confound, the correlation of the difficulty value associated with each item and its primary dimension was less than that of the high confound condition, but much greater than that of the no confound condition. Specific procedures for determining item parameters and how the confound was modeled are included later in this section.

Correlation of the latent ability dimensions

Four levels of correlation between the latent ability dimensions were considered: 0.0, 0.3, 0.6, 0.9. These values represent a range from no association to very strong association. In a given cell of the simulation study, the correlation of latent ability dimensions was kept constant across grades.

Common item sets

Common item sets are usually, but not always, constructed to represent a “mini” version of the test. For this study, two different common item sets were considered. One was a 20 item “mini” lower grade test. The other combined a 10 item version of the lower grade test and a 10 item version of the upper grade test. Common items were treated as “internal” common items in that they were included in the computation of scores for the simulees.

Data generation and model estimation

As described earlier, the two-dimensional MC2PL was the data generating model for the multidimensional vertical scaling data. Data were generated using SAS. The 2PL model was the estimating model for the data. The equations for both the MC2PL and 2PL models were included in the Item Response Theory portion of the Background section. BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1996) was used to estimate the 2PL model and the grouping option was use to define and control for the two separate ability distributions in the estimation. Example SAS data generation code and BILOG-MG model estimation code can be found in appendices A and B, respectively.

Item parameters

Each grade had a 60 item test where 40 items were unique and 20 items were common. So, a test battery for the two grades included 100 items. Note that the use of a 60 item grade level test with 20 common items is consistent with the design used in Lin (2009). The 2-dimensional tests were constructed such that an item primarily discriminates on a single dimension. Items ranged from loading on only one dimension (simple structure) to loading strongly on one dimension while still having a weak relationship with the other. Items that primarily measured dimension one had an angle, α, of 0-25 degrees with the dimension one axis and items that primarily measured dimension two had an, α, of 65-90 degrees with the dimension one axis. Items that primarily measured dimension one represented the majority (approximately 75%) of items on the lower grade test and items that primarily measured dimension two represented the majority (approximately 75%) of items on the upper grade test. The test construction parameters for each item included discrimination parameters, a_1 and a_2 (which will define the item's relationship with the two dimensions) and the parameter related to difficulty, *d*. [Note also that while a test battery was multidimensional on the whole, there were items that were unidimensional; that is, they only loaded on a single dimension]

Parameters for the six 100 item test batteries described in Table 3, below, were determined:

Table 3: Test battery characteristics

Test Battery	Confound Level	Common items
Battery 1	High	lower grade
Battery 2	High	both grades
Battery 3	Moderate	lower grade
Battery 4	Moderate	both grades
Battery 5	No	lower grade
Battery 6	No	both grades

All items had a *MDISC* of 1. By treating the MDISC as fixed at 1 and knowing the location of the items, a_{i1} is simply the cosine of α_i. And a_{i2} is $\sqrt{1-a_{i1}^2}$. For tests where there was no confound of difficulty with dimensionality, each b_{ik} was drawn from a N(0,1) distribution and equation 2 was applied to determine d_i. For tests where difficulty was confounded with dimensionality, b_{i1} was drawn from a N(-1,.25), b_{i2} was drawn from a N(1,.25) distribution, and equation 2 was applied to determine d_i. For tests where difficulty was moderately confounded with dimensionality, b_{i1} was drawn from a N(-.5,.5), b_{i2} was drawn from a N(.5,.5) distribution, and equation 2 was applied to determine d_i. Note that the weighting of each b_{ik} by a_{ik} and the distribution from which the b_{ik} were drawn caused the difficulty and dimensionality confound. The degree of confound of item difficulty and dimensionality was measured by the correlation of the slope parameters and the difficulty parameter. A high correlation close to 1 (in absolute value) indicated a high confound; a moderate correlation close to 0.6 indicated a moderate confound; and a low correlation close to 0 indicated no confound. The correlation between a_{i1} and the d_i represented the confound of item difficulty with dimension one and the correlation between a_{i2} and d_i represented the confound of item difficulty with dimension 2. Item parameters for all exams can be found in Tables C1 through C6 in Appendix C.

The descriptive statistics associated with the difficulty of each test (and based on the tables in Appendix C) broken down by lower grade, upper grade, and entire test battery are included in Tables 4 through 9 below:

Table 4: Test descriptive statistics: Difficulty highly confounded with dimensionality; lower grade common items

	Lower grade test	Upper grade test	Entire test battery
Avg. difficulty	0.22	-0.19	0.00
Stdev. of difficulty	0.79	0.78	0.82
Corr. of difficulty with dim 1	0.96	0.95	0.96
Corr. of difficulty with dim 2	-0.94	-0.93	-0.94

Table 5: Test descriptive statistics: Difficulty highly confounded with dimensionality; both grades common items

	Lower grade test	Upper grade test	Entire test battery
Avg. difficulty	0.17	-0.24	-0.03
Stdev. of difficulty	0.81	0.78	0.83
Corr. of difficulty with dim 1	0.96	0.95	0.96
Corr. of difficulty with dim 2	-0.94	-0.93	-0.94

Table 6: Test descriptive statistics: Difficulty moderately confounded with dimensionality; lower grade common items

	Lower grade test	Upper grade test	Entire test battery
Avg. difficulty	0.17	-0.04	0.05
Stdev. of difficulty	0.60	0.59	0.59
Corr. of difficulty with dim 1	0.65	0.61	0.60
Corr. of difficulty with dim 2	-0.66	-0.58	-0.59

Table 7: Test descriptive statistics: Difficulty moderately confounded with dimensionality; both grades common items

	Lower grade test	Upper grade test	Entire test battery
Avg. difficulty	0.12	-0.09	0.02
Stdev. of difficulty	0.61	0.58	0.59
Corr. of difficulty with dim 1	0.61	0.55	0.58
Corr. of difficulty with dim 2	-0.62	-0.51	-0.56

Table 8: Test descriptive statistics: Difficulty not confounded with dimensionality; lower grade common items

	Lower grade test	Upper grade test	Entire test battery
Avg. difficulty	-0.13	-0.08	-0.02
Stdev. of difficulty	1.02	1.15	1.09
Corr. of difficulty with dim 1	0.06	-0.25	-0.07
Corr. of difficulty with dim 2	-0.04	0.22	0.07

Table 9: Test descriptive statistics: Difficulty not confounded with dimensionality; both grades common items

	Lower grade test	Upper grade test	Entire test battery
Avg. difficulty	0.07	0.11	0.10
Stdev. of difficulty	1.01	1.13	1.07
Corr. of difficulty with dim 1	0.06	-0.16	-0.04
Corr. of difficulty with dim 2	-0.03	0.13	0.04

Simulation steps (for a given cell)

1. Generated 10,000 simulees per grade and simulated item responses (0/1) to ALL items in the test battery (upper grade, lower grade, and common items) based on MIRT generating parameters. The sample size of 10,000 per grade was considered sufficient to represent population level data.
2. Obtained distribution of raw scores for members of each grade.
3. For each grade, found the max raw score for which 40% or fewer simulees fell below. The 60% proficient cut was based on a survey of proportion proficient rates on various state reading and mathematics tests designed to measure student

progress under the NCLB legislation (Tracey Magda, personal communication, August 6, 2008).

4. Targeted 1 plus scores determined above as minimum score for proficient for each grade; this yielded: Target_Cut_Lower_Grade and Target_Cut_Upper_Grade
5. Estimated the resulting response patterns according to a vertically scaled two-group 2PL unidimensional model in BILOG-MG.
6. The resulting unidimensional 2PL item parameter estimates were ranked from smallest to largest according to the theta value required to have a 67% probability of answering the item correctly. Thus, a "simulated" bookmark ordered item booklet (OIB) with response probability (RP) equal to 67% was created; using test characteristic curve methods, expected total scores for a wide range of thetas were computed.
7. Found the two OIB locations and associated expected total scores that captured the target cuts (Target_Cut_Lower_Grade and Target_Cut_Upper_Grade).
8. Rounded the expected total scores and compared them to the target cuts; chose the expected score that has the smallest absolute difference with the target cut. In the event of a tie, used the lower score. The results from this step yielded Cut_Lower_Grade and Cut_Upper_Grade. Took note of the OIB location of each of these cuts.

The steps described above were used to establish the true cut points for proficient (on the entire test battery) for each grade in the given cell; the subsequent steps describe the procedure for running the replications within a given cell

9. Generated 500 replications of 2,000 simulees per grade and simulated item responses to only grade level and common items based on MIRT generating parameters.
10. Estimated item parameters for each replication according to a two-group vertically scaled 2PL unidimensional model in BILOG-MG.
11. For each replication, the resulting unidimensional item parameter estimates were ranked from smallest to largest according to the theta value required to have a 67% probability of answering the item correctly. Thus, a “simulated” OIB was created.
12. For each grade in each replication, the location of the theta cut for proficiency in the OIB was the same as used to determine the true cut for proficiency (step 8). The cut score for proficient in the given replication was determined using test characteristic curve methods with the 2PL parameter estimates for the appropriate grade-level and common items and the associated theta value in the OIB. Standard rounding rules were applied to obtain whole number cut scores.
13. Each simulee in each replication was classified by comparing their “observed” total score to the grade-level appropriate total score cut for proficiency determined in step 12.
14. For each simulee in each replication, their “true” classification was determined by calculating their expected total score (adjusted using standard rounding conventions) on all items in the test battery based on their 2-dimensional theta vector and the MIRT generating parameters and comparing that value to the

"true" grade-level appropriate cut for proficiency determined in step 8. Support for establishing true scores in this manner can be found in Reckase and Li (2007) and Lin (2009).

15. The classifications determined in steps 13 and 14 were used to determine classification success. For each replication, the proportions described in equations 10 through 13 were computed.

Outcome measures

Based on the indices described in Betebenner et al. (2008), for each replication false-positive, false-negative, and correct classification rates based on both the joint and conditional distribution of simulees were calculated. Across replications the mean and standard deviation of these measures were calculated. The percent correctly classified based on the conditional distribution was the primary focus of the subsequent analyses.

Analysis method

A four factor analysis of variance (ANOVA) was the primary method of analysis. Specifically, there were four analyses; for each grade (upper and lower), an ANOVA was conducted on the proportion of not proficient students classified correctly and for the proportion of proficient students classified correctly (i.e. the conditional classification probabilities). Since proportions are typically skewed and do not meet the normal distribution assumptions of ANOVA, all proportions were transformed according to an arcsin transformation as described in Sheskin (2007). Specifically the following transformation was used:

$$Y = 2\arcsin\sqrt{X}\ , \qquad (14)$$

where *X* is a proportion. Note that this transformation was conducted on the radian metric.

For descriptive purposes, the ANOVA results for the conditional raw proportions are also presented. Corresponding ANOVA analyses for the classification accuracy measures based on the joint distributions are included in appendix D. As was described earlier, the four independent levels in the ANOVA were correlation of ability dimensions, confound of difficulty with dimensionality, common item set, and difference in ability of the lower and upper grades.

Because the sample sizes used would likely result in statistically significant results for all main and interaction effects, the partial eta-squared effect size measure was used to assess the degree of relationship of the predictors with the outcome variable. Specifically, the partial eta-squared effect size measure indicates the proportion of variance explained by the main or interaction factor while partialling out all other factors from the nonerror variance (Pierce, Block, & Aguinis; 2004). The formula for the partial eta squared is as follows:

$$\text{partial } \eta^2 = SS_{factor} / (SS_{factor} + SS_{error}) \qquad (15)$$

Chapter 7: Results

Results are presented in three major sections. The first includes the population analysis where performance on the full test battery as well as the "true" cutscores and bookmark locations are presented. The next two sections include the multifactor ANOVA output and associated figures and tables for the conditional probability of correct classification into the proficient or not proficient categories for lower and upper grades. There is a section for each grade, upper and lower.

Population analyses

The population performance across cells was computed and the results, summarized by confound level in Tables 10 through 12 below, were plausible given the population and item generating parameters. Specifically, take note of the larger standard deviations as the correlation between ability dimensions increased.

Table 10: High confound full test battery population performance

			Lower Grade		Upper Grade	
Common Items	Ability Difference	Correlation	AVG	STDEV	AVG	STDEV
Lower Grade	Small	0.0	47.9	16.1	54.3	15.5
		0.3	48.0	17.8	54.4	17.1
		0.6	47.9	19.1	54.0	18.7
		0.9	48.3	20.8	54.4	19.9
	Big	0.0	48.2	15.9	60.2	14.7
		0.3	47.9	17.8	60.0	16.5
		0.6	48.2	19.4	60.1	17.8
		0.9	47.6	20.5	59.6	19.2
Both Grades	Small	0.0	47.5	15.8	53.5	15.6
		0.3	47.4	17.7	53.8	17.2
		0.6	47.2	19.0	53.4	18.7
		0.9	47.2	20.5	53.5	20.1
	Big	0.0	47.3	16.0	59.5	14.8
		0.3	47.5	17.6	59.3	16.5
		0.6	47.2	19.1	58.9	18.1
		0.9	47.3	20.3	59.3	19.4

Table 11: Moderate confound full test battery population performance

			Lower Grade		Upper Grade	
Common Items	Ability Difference	Correlation	AVG	STDEV	AVG	STDEV
Lower Grade	Small	0.0	48.5	16.7	55.7	16.2
		0.3	49.1	18.5	55.4	18.3
		0.6	48.8	20.3	55.3	19.8
		0.9	48.8	21.7	55.4	21.3
	Big	0.0	48.7	16.8	62.0	15.4
		0.3	48.7	18.5	62.1	17.2
		0.6	48.5	20.1	61.6	18.9
		0.9	48.6	21.6	61.5	20.4
Both Grades	Small	0.0	48.3	16.7	55.0	16.4
		0.3	48.3	18.6	55.1	18.2
		0.6	48.4	20.2	54.8	19.8
		0.9	48.1	21.7	54.4	21.3
	Big	0.0	48.4	16.8	61.3	15.7
		0.3	48.0	18.4	61.0	17.6
		0.6	48.4	20.4	60.6	19.2
		0.9	48.5	21.7	60.6	20.4

Table 12: No confound full test battery population performance

			Lower Grade		Upper Grade	
Common Items	Ability Difference	Correlation	AVG	STDEV	AVG	STDEV
Lower Grade	Small	0.0	47.6	15.1	53.7	15.1
		0.3	47.9	16.9	53.9	17.0
		0.6	47.7	18.4	54.1	18.5
		0.9	48.2	19.8	53.7	19.8
	Big	0.0	47.8	15.0	60.0	14.9
		0.3	47.9	16.8	60.0	16.3
		0.6	47.8	18.5	59.5	18.0
		0.9	47.5	19.9	59.7	19.4
Both Grades	Small	0.0	49.7	15.1	55.9	15.1
		0.3	50.0	17.0	55.5	17.0
		0.6	49.5	18.7	56.1	18.4
		0.9	49.7	19.8	55.6	19.7
	Big	0.0	49.7	15.2	61.7	14.6
		0.3	49.7	16.8	62.0	16.3
		0.6	49.8	18.6	61.6	17.7
		0.9	49.3	19.9	61.6	19.0

Using population performance on the test battery for each cell and following the procedure described in simulation steps 1 through 8, true cutscores and bookmark locations were determined. The results are presented in Tables 13 through 15 below by confound level.

Table 13: High confound full test battery population cut scores and bookmark locations across factors

			Lower Grade		Upper Grade	
Common Items	Ability Difference	Correlation	Cut Score	Location	Cut Score	Location
Lower Grade	Small	0.0	44	10	51	37
		0.3	43	9	50	35
		0.6	42	10	49	31
		0.9	42	10	49	33
	Big	0.0	44	11	57	52
		0.3	43	10	56	52
		0.6	43	11	57	53
		0.9	41	9	56	52
Both Grades	Small	0.0	42	11	50	36
		0.3	42	10	50	36
		0.6	42	10	49	34
		0.9	41	11	48	31
	Big	0.0	43	12	56	49
		0.3	43	13	56	49
		0.6	41	11	55	49
		0.9	40	11	55	49

Table 14: Moderate confound full test battery population cut scores and bookmark locations across factors

			Lower Grade		Upper Grade	
Common Items	Ability Difference	Correlation	Cut Score	Location	Cut Score	Location
Lower Grade	Small	0.0	45	7	51	19
		0.3	44	7	51	19
		0.6	42	6	51	19
		0.9	41	6	50	17
	Big	0.0	44	7	59	36
		0.3	42	6	59	37
		0.6	42	6	58	33
		0.9	41	6	58	34
Both Grades	Small	0.0	44	6	50	17
		0.3	44	6	51	18
		0.6	43	6	50	17
		0.9	40	5	49	15
	Big	0.0	44	6	58	32
		0.3	41	5	58	33
		0.6	42	6	57	29
		0.9	42	6	57	31

Table 15: No Confound full test battery population cut scores and bookmark locations across factors

			Lower Grade		Upper Grade	
Common Items	Ability Difference	Correlation	Cut Score	Location	Cut Score	Location
Lower Grade	Small	0.0	43	18	50	24
		0.3	43	18	48	23
		0.6	42	17	48	23
		0.9	42	18	48	23
	Big	0.0	45	19	57	35
		0.3	42	18	56	34
		0.6	41	17	55	33
		0.9	41	17	56	35
Both Grades	Small	0.0	45	21	52	27
		0.3	44	21	51	26
		0.6	44	21	52	27
		0.9	44	20	49	25
	Big	0.0	46	22	58	38
		0.3	45	20	59	40
		0.6	44	19	58	39
		0.9	43	19	58	40

Note that as a measure to ensure the reasonableness of the simulated bookmark procedure conducted for each replication and its use of the population based OIB locations, the OIBs for two replications from each of four cells were compared to the corresponding OIBs based on population performance on the entire test battery. The results were very convincing and consistent: Across all replications considered, the items above or below the cutscore (for both grades) differed by no more than three compared to the items above and below the cuts based on the OIB from the population calibration.

Lower grade analysis

Output from the multifactor ANOVAs as well as descriptive statistics and figures are used to describe the results for the lower grade. For ease of presentation the following abbreviations were used for the factor levels and criterion variables:

CONFOUND = factor for item difficulty confounded with dimensionality; C = high confound, M = moderate confound, N = no confound

CORRELAT = factor for the correlation between dimensions; 0 = 0.0 correlation between dimensions, 3 = 0.3 correlation between dimensions, 6 = 0.6 correlation between dimensions, 9 = 0.9 correlation between dimensions

ABILITY = factor for the ability difference between the lower and upper grade; B = big ability difference, S = small ability difference

COMMON = factor for the common items administered to both lower and upper grades; L = lower grade common items, A = both grades common items

CONDPC1 = the raw proportion of those truly not proficient classified as such

CONDPC2 = the raw proportion of those truly proficient classified as such

CONDPC1T = the arcsin transformed proportion of those truly not proficient classified as such

CONDPC2T = the arcsin transformed proportion of those truly proficient classified as such

Note also that a given cell of the study will be referred to as NLB0. This abbreviation would indicate no confound, lower grade common items, big ability difference, and a correlation between dimensions of 0. Other cells will be referred to similarly using the abbreviations above. These abbreviations also apply to the upper grade analysis.

Proportion correctly classified as not proficient

The multifactor ANOVA output, raw and arcsin transformed, for the proportion of lower grade simulees classified correctly as not proficient are found below in Tables 16 and 17, respectively:

Table 16: Lower Grade Multifactor ANOVA on Raw Data for CONDPC1

Tests of Between-Subjects Effects

Dependent Variable: CONDPC1

Source	Type III SS	df	Mean Sq	F	Sig.	Partial η^2
Corrected Model	25.904	47	0.551	1023.336	0.000	0.668
Intercept	18398.318	1	18398.318	34161256.370	0.000	0.999
CONFOUND	6.361	2	3.181	5905.486	0.000	0.330
COMMON	0.857	1	0.857	1591.851	0.000	0.062
ABILITY	0.186	1	0.186	345.620	0.000	0.014
CORRELAT	13.691	3	4.564	8473.811	0.000	0.515
CONFOUND * COMMON	0.206	2	0.103	190.909	0.000	0.016
CONFOUND * ABILITY	0.513	2	0.256	476.239	0.000	0.038
COMMON * ABILITY	0.116	1	0.116	215.375	0.000	0.009
CONFOUND * COMMON * ABILITY	0.099	2	0.049	91.689	0.000	0.008
CONFOUND * CORRELAT	2.019	6	0.336	624.658	0.000	0.135
COMMON * CORRELAT	0.076	3	0.025	46.732	0.000	0.006
CONFOUND * COMMON * CORRELAT	0.374	6	0.062	115.729	0.000	0.028
ABILITY * CORRELAT	0.039	3	0.013	24.420	0.000	0.003
CONFOUND * ABILITY * CORRELAT	0.418	6	0.070	129.266	0.000	0.031
COMMON * ABILITY * CORRELAT	0.086	3	0.029	53.362	0.000	0.007
CONFOUND * COMMON * ABILITY * CORRELAT	0.863	6	0.144	267.069	0.000	0.063
Error	12.900	23952	0.001			
Total	18437.121	24000				
Corrected Total	38.804	23999				

a. R Squared = .668 (Adjusted R Squared = .667)

Table 17: Lower Grade Multifactor ANOVA on Arcsin Transformed Data for CONDPC1

Tests of Between-Subjects Effects

Dependent Variable: CONDPC1T

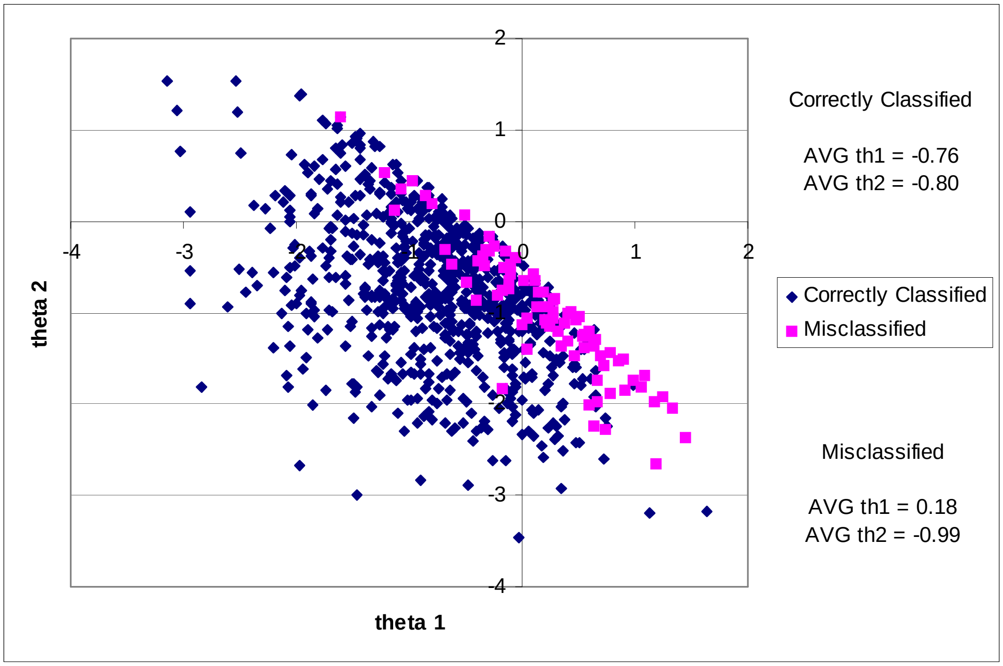
Source	Type III SS	df	Mean Sq	F	Sig.	Partial η^2
Corrected Model	228.855	47	4.869	982.829	0.000	0.659
Intercept	141570.703	1	141570.703	28575214.158	0.000	0.999
CONFOUND	58.258	2	29.129	5879.544	0.000	0.329
COMMON	7.378	1	7.378	1489.237	0.000	0.059
ABILITY	1.393	1	1.393	281.208	0.000	0.012
CORRELAT	122.276	3	40.759	8226.879	0.000	0.507
CONFOUND * COMMON	1.290	2	0.645	130.140	0.000	0.011
CONFOUND * ABILITY	4.635	2	2.318	467.793	0.000	0.038
COMMON * ABILITY	1.072	1	1.072	216.454	0.000	0.009
CONFOUND * COMMON * ABILITY	1.387	2	0.693	139.965	0.000	0.012
CONFOUND * CORRELAT	14.595	6	2.433	491.003	0.000	0.110
COMMON * CORRELAT	0.446	3	0.149	30.021	0.000	0.004
CONFOUND * COMMON * CORRELAT	3.243	6	0.541	109.113	0.000	0.027
ABILITY * CORRELAT	0.230	3	0.077	15.467	0.000	0.002
CONFOUND * ABILITY * CORRELAT	3.246	6	0.541	109.201	0.000	0.027
COMMON * ABILITY * CORRELAT	0.874	3	0.291	58.820	0.000	0.007
CONFOUND * COMMON * ABILITY * CORRELAT	8.530	6	1.422	286.957	0.000	0.067
Error	118.666	23952	0.005			
Total	141918.223	24000				
Corrected Total	347.521	23999				

a. R Squared = .659 (Adjusted R Squared = .658)

Using the ANOVA output for the transformed data (Table 17), it is reasonably clear that three effects have the strongest association with the criterion variable (CONDPC1T) as measured by partial eta squared. They are the confound (partial $\eta^2 = .329$) and correlation (partial $\eta^2 = .507$) main effects and the confound and correlation interaction effect (partial $\eta^2 = .110$). Note that these three effect sizes are all larger than 0.10, which is in between the 0.06 and 0.14 rules of thumb for medium and large effect sizes measured by partial eta squared (Stevens, 1992). Before evaluating the marginal mean differences on proportion correctly classified across the various levels for these factors, it is important to visually appreciate the theta vector distribution of truly not proficient simulees classified as such versus those classified as proficient for each of the two main effects. This evaluation will help us better understand why these factors have a strong association with CONDPC1.

Figures 1 and 2, below, illustrate the distribution of theta vectors for truly not proficient simulees classified correctly and incorrectly from two example replications from the extremes of the confound levels, no confound and high confound. The average values on theta 1 and theta 2 for those classified correctly and incorrectly are also included on the figure. Specifically, a replication from each of the following cells was used: NLB0 and CLB0

Figure 1: Distribution of theta vectors for lower grade truly not proficient simulees; NLB0 cell



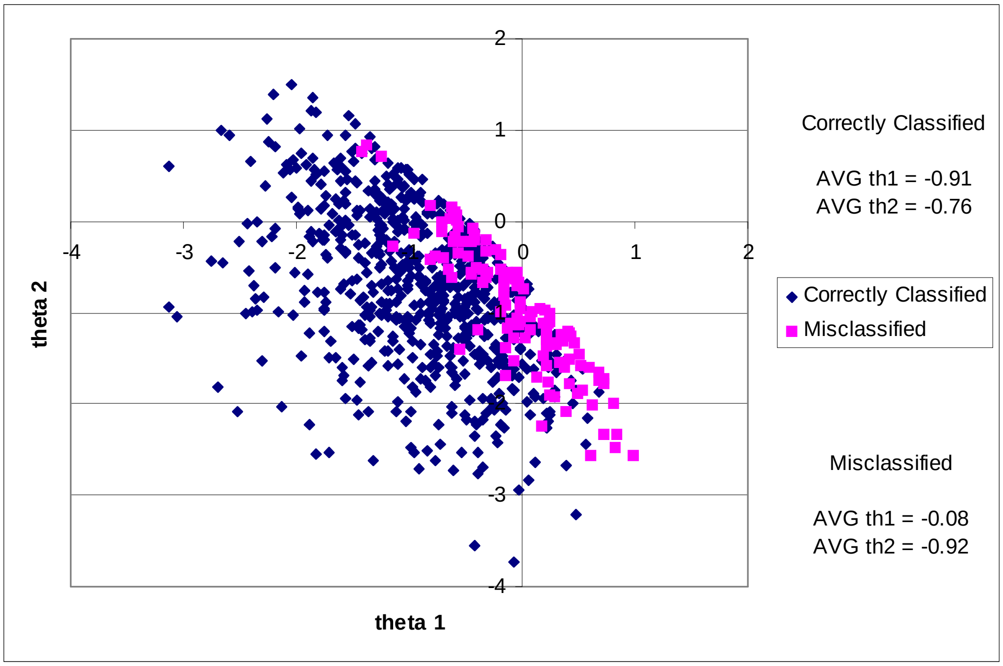
Correctly Classified

AVG th1 = -0.76
AVG th2 = -0.80

Misclassified

AVG th1 = 0.18
AVG th2 = -0.99

Figure 2: Distribution of theta vectors for lower grade truly not proficient simulees; CLB0 cell (I)



In Figure 1, 90% of the truly not proficient simulees were classified correctly and in Figure 2, 84 % of the truly not proficient simulees were classified correctly; so more simulees were classified correctly as not proficient in the no confound case. The theta vectors of those not proficient simulees classified as proficient, however, followed the same pattern in both cells: on average, they were weaker than those classified correctly on dimension two (the upper grade focused items), but much stronger on dimension 1 (the lower grade focused items). In general, it makes sense that those stronger on dimension one were classified as proficient because the majority of the items administered to them were lower grade focused, so those simulees were in a greater position to get those items correct. Remember that their true proficiency classification was based on the entire test battery where there was a greater proportion of upper grade focused items, the dimension on which these incorrectly classified simulees were particularly weak which resulted in their truly not proficient classification. The difference in proportions between these two cases of truly not proficient simulees being classified as such also makes sense. In the high confound case the lower grade items were easier than in the no confound case, so it was more likely that these truly not proficient simulees could get those lower grade items correct; thus a lower CONDPC1 for the confound cell. Note that in the no confound case, difficulty of items was not related to the dimensionality (lower or upper grade focused items), so the lower grade focused items were allowed to be just as hard as the upper grade focused items. Table 18 below includes the marginalized (across all other factors) raw and transformed CONDPC1 average values for all three levels of the confound factor in the lower grade.

Table 18: Average lower grade CONDPC1T and CONDPC1 values for confound levels across all other factors

	No Confound	Moderate Confound	Confound
CONDPC1T	2.498	2.377	2.401
CONDPC1	0.898	0.861	0.867

There is very little difference (third decimal place) in the average values for the high confound or moderate confound levels. This is due to, perhaps, to the coarse nature of the decision being made (proficient or not proficient). The average CONDPC1 values for the high confound and moderate confound levels are less than the no confound level by approximately three percent on the raw metric.

Figures 3 through 6 below illustrate the distribution of theta vectors for truly not proficient simulees classified correctly and incorrectly from four example replications across all four correlation levels, 0.0, 0.3, 0.6, 0.9. Specifically, a replication from each of the following cells was used: CLB0, CLB3, CLB6, and CLB9.

Figure 3: Distribution of theta vectors for lower grade truly not proficient simulees; CLB0 cell (II)

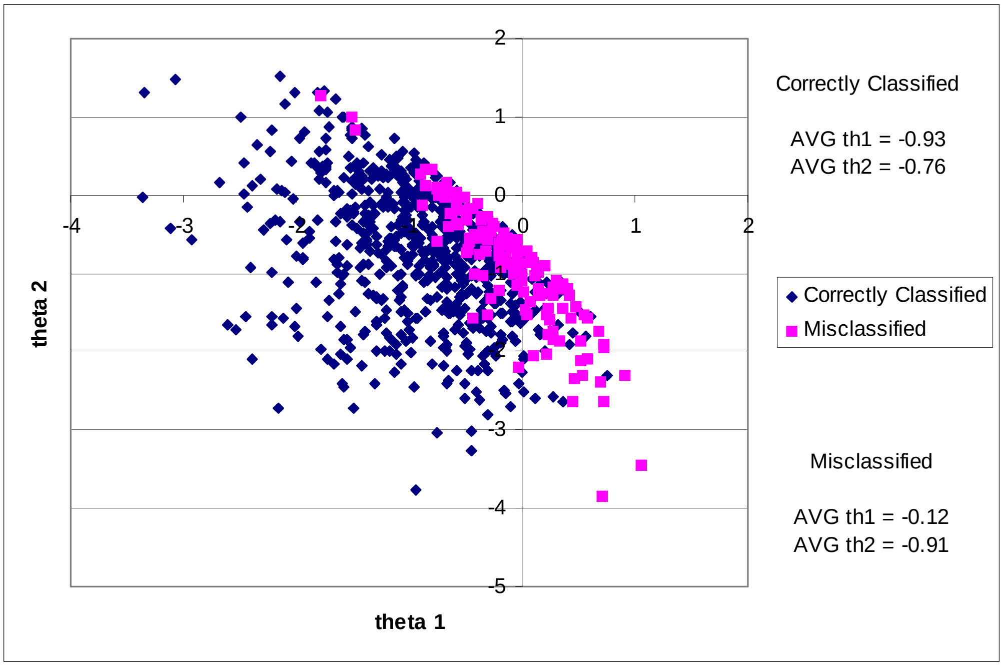


Figure 4: Distribution of theta vectors for lower grade truly not proficient simulees; CLB3 cell

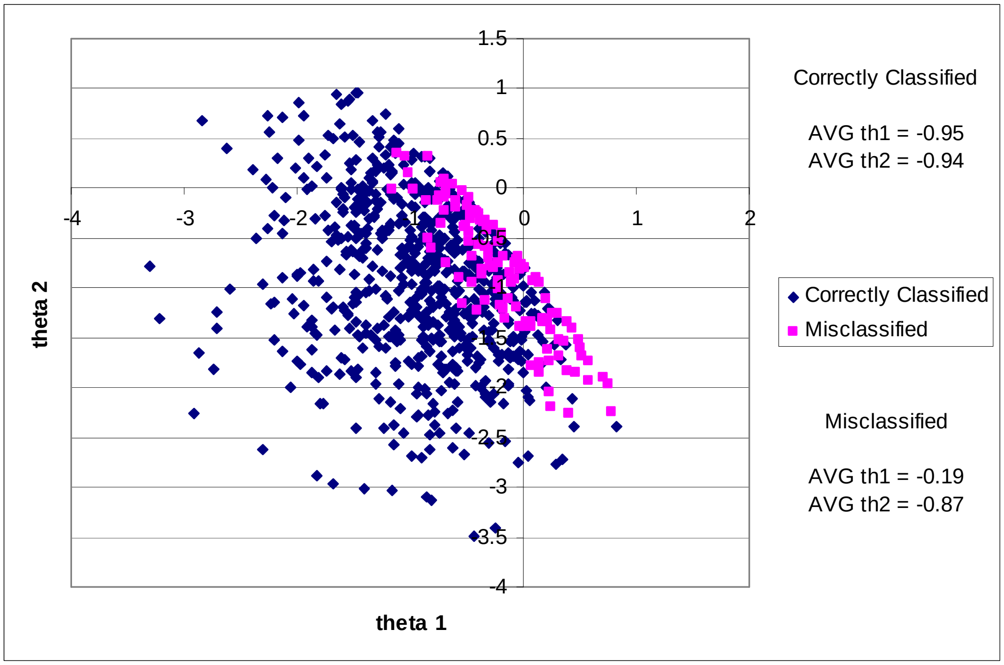


Figure 5: Distribution of theta vectors for lower grade truly not proficient simulees; CLB6 cell

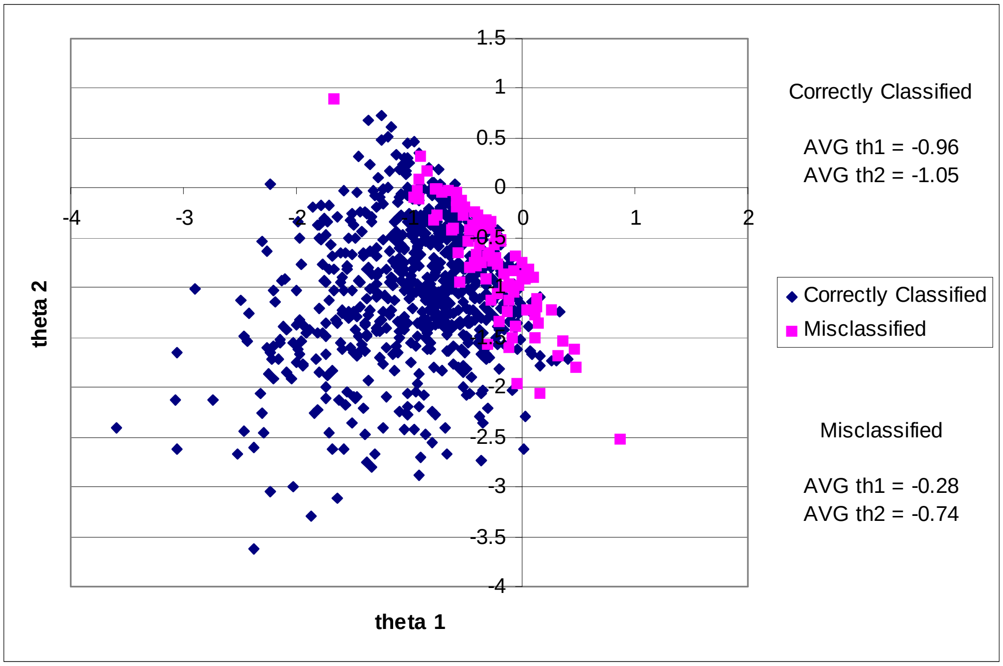
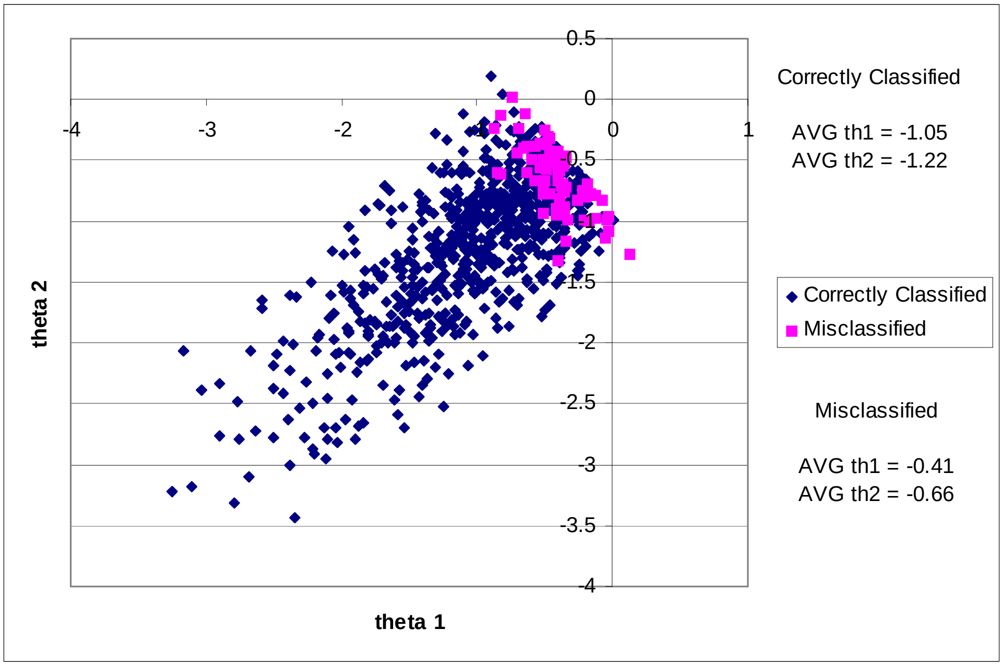


Figure 6: Distribution of theta vectors for lower grade truly not proficient simulees; CLB9 cell



In Figures 3 through 6, 83%, 85%, 88%, and 90%, respectively, of the truly not proficient simulees were classified correctly; so more simulees were classified correctly as not proficient as the correlation between dimensions got stronger. This pattern makes sense. The simulees in each of these four cells were administered the same items to determine their true classification (entire CL test battery) and the same items to determine their "observed" (lower grade and common items from the CL test battery) classification. As the theta 1 and theta 2 values for each of these simulees became more highly related it is reasonable that their relative performance on the entire test battery would match their performance on the lower grade test. Thus, CONDPC1 would be expected to increase as the relationship between the theta values increases. Visually, this increasingly linear relationship can be appreciated across the four figures above. Table 19 below includes the marginalized (across all other factors) raw and transformed CONDPC1 average values for all four levels of the correlation factor in the lower grade.

Table 19: Average lower grade CONDPC1T and CONDPC1 values for correlation levels across all other factors

	0.0	0.3	0.6	0.9
CONDPC1T	2.329	2.404	2.459	2.523
CONDPC1	0.842	0.868	0.887	0.906

There is approximately a six percent CONDPC1 increase on the raw metric between no correlation and 0.9 correlation. Further, the change in raw percentage by correlation level is approximately two percent increasing from no relationship between dimensions to a strong relationship between dimensions.

The plots in Figures 7 and 8 below represent the interaction of the confound and correlation factors. Each point represents the average CONDPC1 or CONDPC1T across the other two factors for a given level of correlation and confound.

Figure 7: Lower grade confound and correlation interaction with CONDPC1 as the criterion

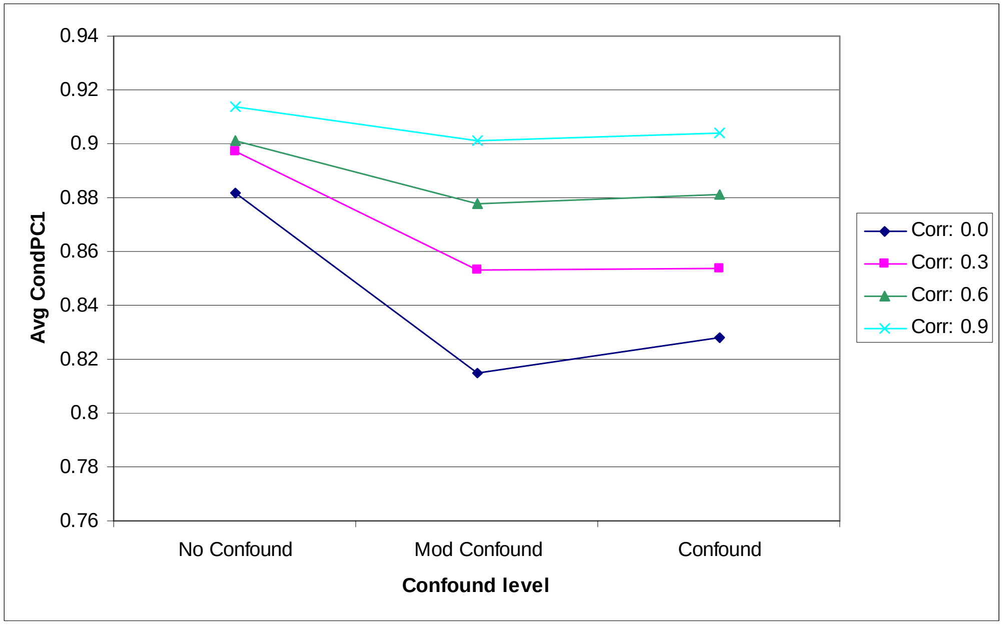
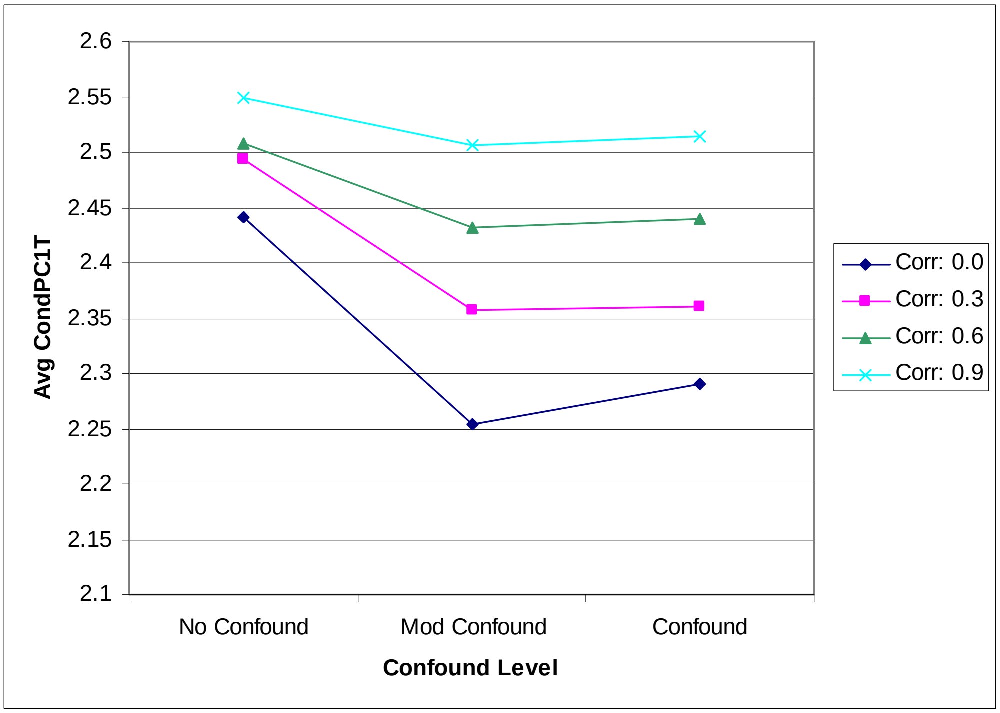


Figure 8: Lower grade confound and correlation interaction with CONDPC1T as the criterion



The plots on these two figures illustrate an ordinal interaction between confound and correlation. Since the raw metric is easier to understand, Figure 7 is discussed; however, the conclusions would also apply to Figure 8. For the no confound level, the differences among the four levels of correlation on CONDPC1 are much smaller than the differences at the moderate confound or high confound levels. Further, the confound of difficulty with dimensionality has a greater effect on CONDPC1 when there is a low relationship among dimensions than when there is a stronger one. As mentioned before, it is also clear that the no confound level will have the greatest success at classifying not proficient students as such. This is simply because in the no confound condition, simulees would be administered relatively harder items and would be more likely to answer them incorrectly resulting in a not proficient "observed" classification. This plot also helps to reinforce the mean differences on CONDPC1 between no confound and the other two levels of confound and across the four levels of correlation. Table 20, below, includes the mean CONDPC1 and CONDPC1T values at each of the levels of the confound and correlation factors.

Table 20: Average lower grade CONDPC1 and CONDPC1T values for confound and correlation levels across all other factors

		Correlation			
	Confound Level	0.0	0.3	0.6	0.9
CONDPC1T	No Confound	2.442	2.493	2.508	2.549
CONDPC1		0.882	0.897	0.901	0.914
CONDPC1T	Mod Confound	2.254	2.357	2.431	2.506
CONDPC1		0.815	0.853	0.878	0.901
CONDPC1T	Confound	2.291	2.361	2.439	2.514
CONDPC1		0.828	0.854	0.881	0.904

As noted earlier, only those effects with partial eta squared values greater than 0.10 were investigated further. In order to feel comfortable with that decision, the marginal means on CONDPC1 and CONDPC1T for the two levels of common items were computed. The partial eta squared for this main effect is 0.059, which, rounded to one decimal place, is the minimum rule of thumb value for a medium effect size. The means are included in Table 21 below:

Table 21: Average lower grade CONDPC1T and CONDPC1 values for common item levels across all other factors

	Lower Grade	Both Grades
CONDPC1T	2.446	2.411
CONDPC1	0.882	0.87

On the raw metric, the difference in these means is approximately 1%. Using this result as a proxy for others with similar partial eta squared effect size values and in terms of efficiency of presentation and to focus on the most important results, it seems reasonable to eliminate effects with partial eta squared values less than 0.10 from further discussion.

Proportion correctly classified as proficient

The multifactor ANOVA output, raw and arcsin transformed, for the proportion of lower grade simulees classified correctly as proficient are found below in Tables 22 and 23, respectively:

Table 22: Lower Grade Multifactor ANOVA on Raw Data for CONDPC2

Tests of Between-Subjects Effects

Dependent Variable: CONDPC2

Source	Type III SS	df	Mean Sq	F	Sig.	Partial η^2
Corrected Model	12.133	47	0.258	1084.343	0.000	0.680
Intercept	20375.749	1	20375.749	85583944.823	0.000	1.000
CONFOUND	6.260	2	3.130	13146.793	0.000	0.523
COMMON	0.396	1	0.396	1662.911	0.000	0.065
ABILITY	0.022	1	0.022	93.861	0.000	0.004
CORRELAT	2.924	3	0.975	4094.256	0.000	0.339
CONFOUND * COMMON	0.021	2	0.011	44.940	0.000	0.004
CONFOUND * ABILITY	0.246	2	0.123	516.986	0.000	0.041
COMMON * ABILITY	0.056	1	0.056	236.462	0.000	0.010
CONFOUND * COMMON * ABILITY	0.105	2	0.052	220.473	0.000	0.018
CONFOUND * CORRELAT	1.269	6	0.211	888.078	0.000	0.182
COMMON * CORRELAT	0.040	3	0.013	55.904	0.000	0.007
CONFOUND * COMMON * CORRELAT	0.178	6	0.030	124.885	0.000	0.030
ABILITY * CORRELAT	0.003	3	0.001	4.280	0.005	0.001
CONFOUND * ABILITY * CORRELAT	0.103	6	0.017	72.159	0.000	0.018
COMMON * ABILITY * CORRELAT	0.065	3	0.022	91.663	0.000	0.011
CONFOUND * COMMON * ABILITY * CORRELAT	0.444	6	0.074	310.580	0.000	0.072
Error	5.702	23952	0.000			
Total	20393.585	24000				
Corrected Total	17.836	23999				

a. R Squared = .680 (Adjusted R Squared = .680)

Table 23: Lower Grade Multifactor ANOVA on Arcsin Transformed Data for CONDPC2T

Tests of Between-Subjects Effects

Dependent Variable: CONDPC2T

Source	Type III SS	df	Mean Sq	F	Sig.	Partial η^2
Corrected Model	149.295	47	3.176	981.915	0.000	0.658
Intercept	159831.980	1	159831.980	49407207.251	0.000	1.000
CONFOUND	78.304	2	39.152	12102.705	0.000	0.503
COMMON	5.862	1	5.862	1812.142	0.000	0.070
ABILITY	0.562	1	0.562	173.621	0.000	0.007
CORRELAT	36.463	3	12.154	3757.178	0.000	0.320
CONFOUND * COMMON	0.420	2	0.210	64.955	0.000	0.005
CONFOUND * ABILITY	3.150	2	1.575	486.849	0.000	0.039
COMMON * ABILITY	0.683	1	0.683	211.031	0.000	0.009
CONFOUND * COMMON * ABILITY	1.195	2	0.598	184.759	0.000	0.015
CONFOUND * CORRELAT	12.187	6	2.031	627.871	0.000	0.136
COMMON * CORRELAT	0.474	3	0.158	48.852	0.000	0.006
CONFOUND * COMMON * CORRELAT	2.408	6	0.401	124.035	0.000	0.030
ABILITY * CORRELAT	0.045	3	0.015	4.652	0.003	0.001
CONFOUND * ABILITY * CORRELAT	1.542	6	0.257	79.468	0.000	0.020
COMMON * ABILITY * CORRELAT	0.619	3	0.206	63.826	0.000	0.008
CONFOUND * COMMON * ABILITY * CORRELAT	5.379	6	0.897	277.150	0.000	0.065
Error	77.485	23952	0.003			
Total	160058.760	24000				
Corrected Total	226.779	23999				

a. R Squared = .658 (Adjusted R Squared = .658)

Using the ANOVA output for the transformed data (Table 23), it is reasonably clear that three effects have the strongest association with the criterion variable (CONDPC1T) as measured by partial eta squared. They are the confound (partial $\eta^2 = .503$) and correlation (partial $\eta^2 = .320$) main effects and the confound and correlation interaction effect (partial $\eta^2 = .136$). None of the other effects reach the 0.10 criterion for further investigation. Before evaluating the marginal differences on proportion correctly classified across the various levels for these factors, it is important to visually appreciate the theta vector distribution of truly proficient simulees classified as such versus those classified as not proficient for each of the two main effects. This evaluation will help us better understand why these factors have a strong association with CONDPC2.

Figures 9 and 10 below illustrate the distribution of theta vectors for truly proficient simulees classified correctly and incorrectly from two example replications from the extremes of the confound levels, no confound and high confound. The average values on theta 1 and theta 2 for those classified correctly and incorrectly are also included on the figure. Specifically, a replication from each of the following cells was used: NLB0 and CLB0.

Figure 9: Distribution of theta vectors for lower grade truly proficient simulees; NLB0 cell (I)

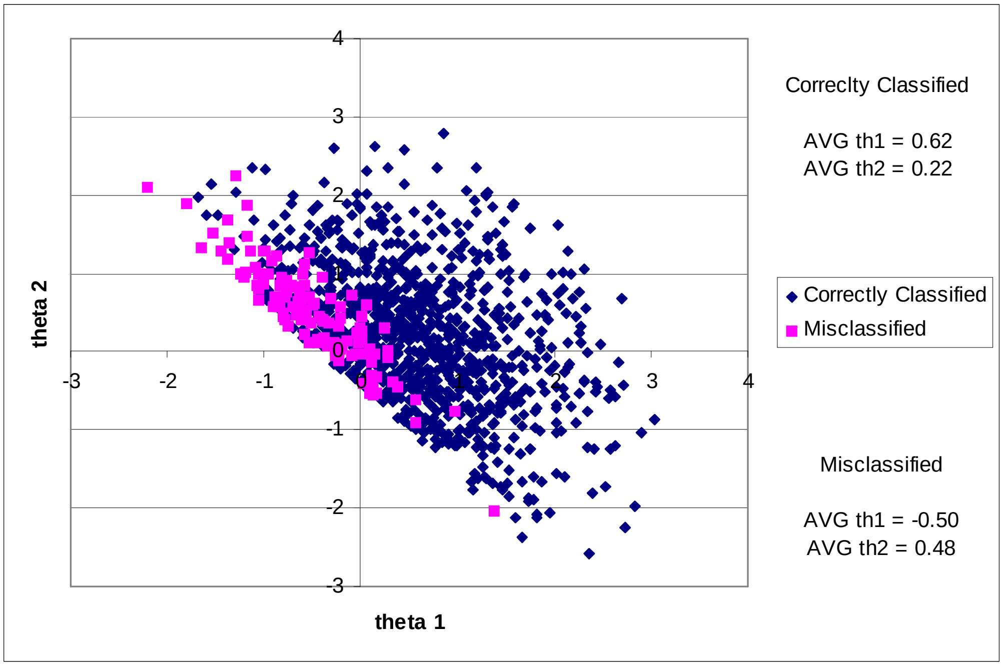
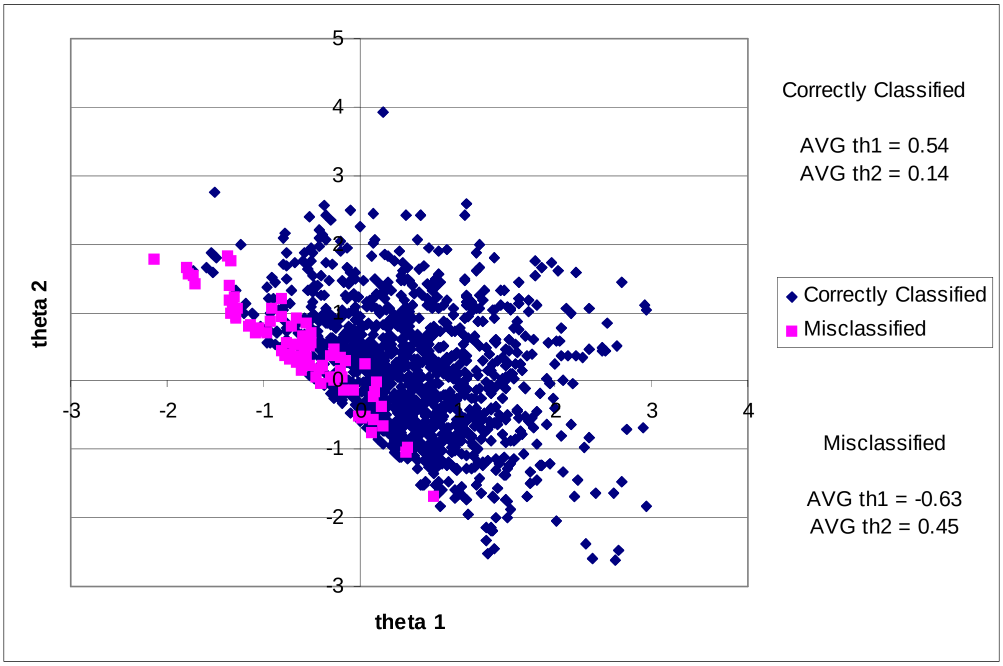


Figure 10: Distribution of theta vectors for lower grade truly proficient simulees; CLB0 cell



In Figure 9, 89% of the truly proficient simulees were classified correctly and in Figure 10, 94 % of the truly proficient simulees were classified correctly; so more simulees were classified correctly as proficient in the high confound case. The theta vectors of those proficient simulees classified as not proficient, however, followed the same pattern in both cells: on average, they were weaker than those classified correctly on dimension one (the lower grade focused items), but much stronger on dimension two (the upper grade focused items). In general, it makes sense that those weaker on dimension one were classified as not proficient because the majority of the items administered to them were lower grade focused, so those simulees were in a greater position to get those items incorrect. Remember that their true proficiency classification was based on the entire test battery where there was a greater proportion of upper grade focused items, the dimension on which these incorrectly classified simulees were particularly strong which caused their truly proficient classification. The difference in proportion between these two cases of truly proficient simulees being classified as such also makes sense. In the high confound case the lower grade items were easier than in the no confound case, so it was more likely that these truly proficient simulees could get those lower grade items correct; thus a higher CONDPC2 for the high confound case. Note that in the no confound case, difficulty of items was not related to the dimensionality (lower or upper grade focused items), so the lower grade focused items were allowed to be just as hard as the upper grade focused items. Table 24 below includes the marginalized (across all other factors) raw and transformed CONDPC2 average values for all three levels of the confound factor in the lower grade.

Table 24: Average lower grade CONDPC2T and CONDPC2 values for confound levels across all other factors

	No Confound	Moderate Confound	Confound
CONDPC2T	2.500	2.624	2.618
CONDPC2	0.899	0.933	0.932

There is very little difference (third decimal place) in the average values for the high confound or moderate confound levels. Again, this could be due to the coarse nature of the decision being made (proficient or not proficient). The average CONDPC2 values for the high confound and moderate confound levels are greater than the no confound level by approximately three percent on the raw metric.

Figures 11 through 14, below, illustrate the distribution of theta vectors for truly proficient simulees classified correctly and incorrectly from four example replications across all four correlation levels, 0, 0.3, 0.6, 0.9. Specifically, a replication from each of the following cells was used: NLB0, NLB3, NLB6, and NLB9

Figure 11: Distribution of theta vectors for lower grade truly proficient simulees; NLB0 cell (II)

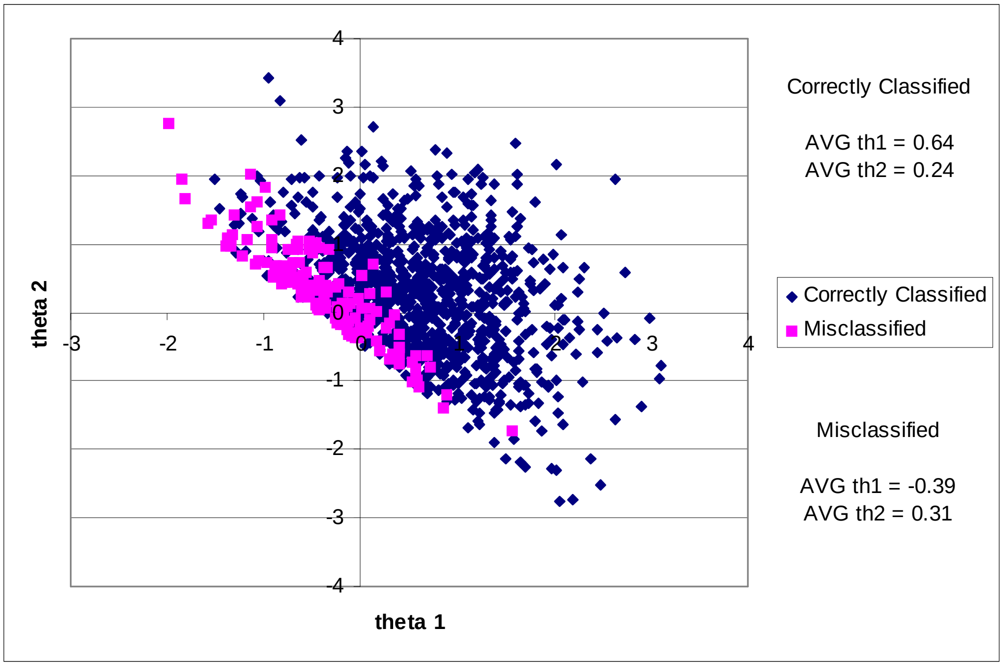


Figure 12: Distribution of theta vectors for lower grade truly proficient simulees; NLB3 cell

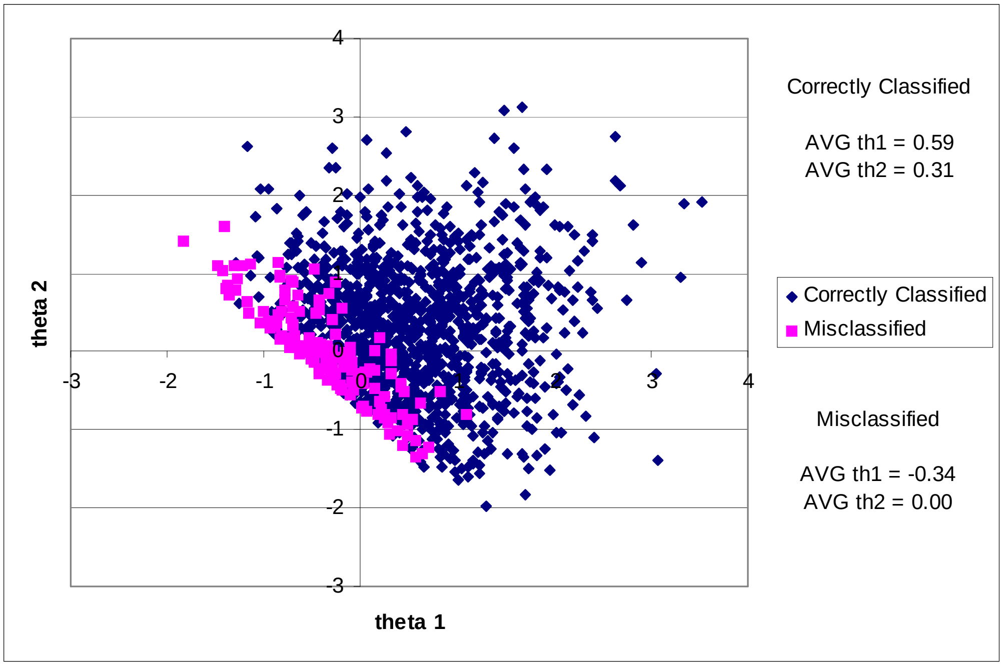


Figure 13: Distribution of theta vectors for lower grade truly proficient simulees; NLB6 cell

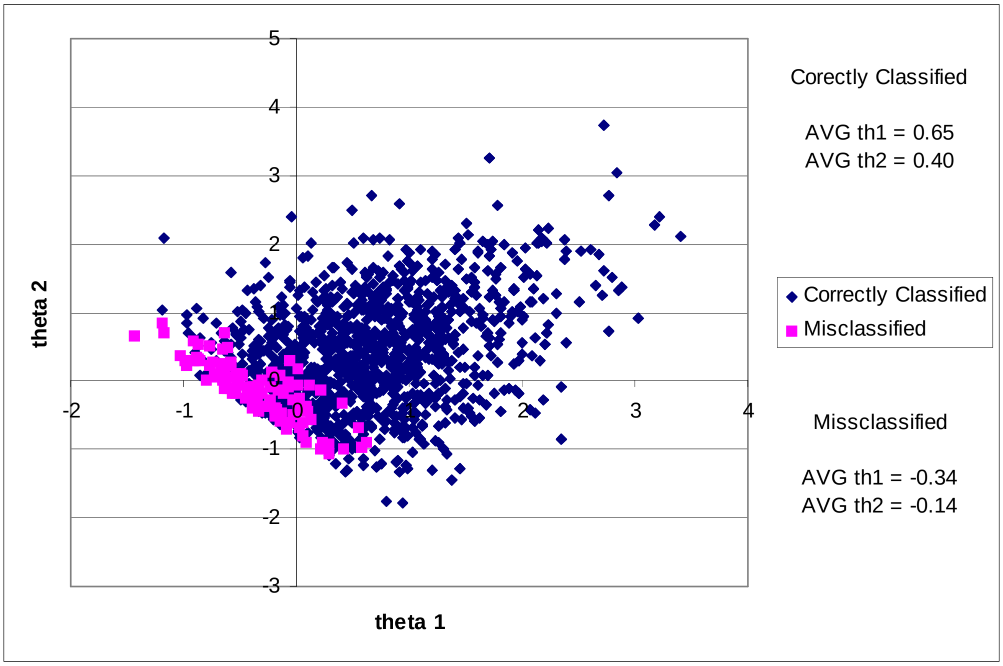
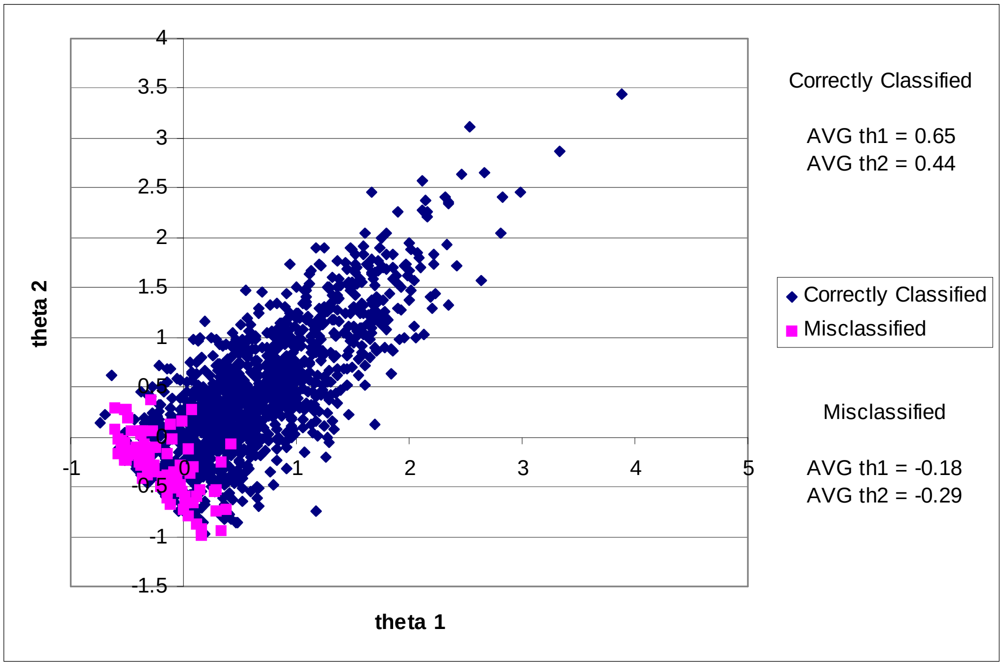


Figure 14: Distribution of theta vectors for lower grade truly proficient simulees; NLB9 cell



In Figures 11 through 14, 87%, 89%, 92%, and 93%, respectively, of the truly proficient simulees were classified correctly; so more simulees were classified correctly as proficient as the correlation between dimensions got stronger. This pattern, and a rationale for its appearance, is similar to that of the CONDPC1 criterion. The simulees in each of these four cells were administered the same items to determine their true classification (entire NL test battery) and the same items to determine their "observed" (lower grade and common items from the NL test battery) classification. As the theta one and theta two for each of these simulees became more highly related it is reasonable that their relative performance on the entire test battery would match their performance on the lower grade test. Thus, CONDPC2 would be expected to increase as the relationship between the theta values increases. Visually, this increasingly linear relationship can be appreciated across the four figures above. Table 25, below, includes the marginalized (across all other factors) raw and transformed CONDPC2 average values for all four levels of the correlation factor in the lower grade.

Table 25: Average lower grade CONDPC2T and CONDPC2 values for correlation levels across all other factors

	0.0	0.3	0.6	0.9
CONDPC2T	2.524	2.566	2.605	2.626
CONDPC2	0.905	0.918	0.929	0.934

There is approximately a three percent average CONDPC2 increase on the raw metric between no correlation and 0.9. Further, the change in raw percentage by correlation level is approximately one to two percent increasing from no relationship between dimensions to a strong relationship between dimensions.

The plots in Figures 15 and 16 below represent the interaction of the confound and correlation factors. Each point represents the average CONDPC2 or CONDPC2T across the other two factors for a given level of correlation and confound.

Figure 15: Lower grade confound and correlation interaction with CONDPC2 as the criterion

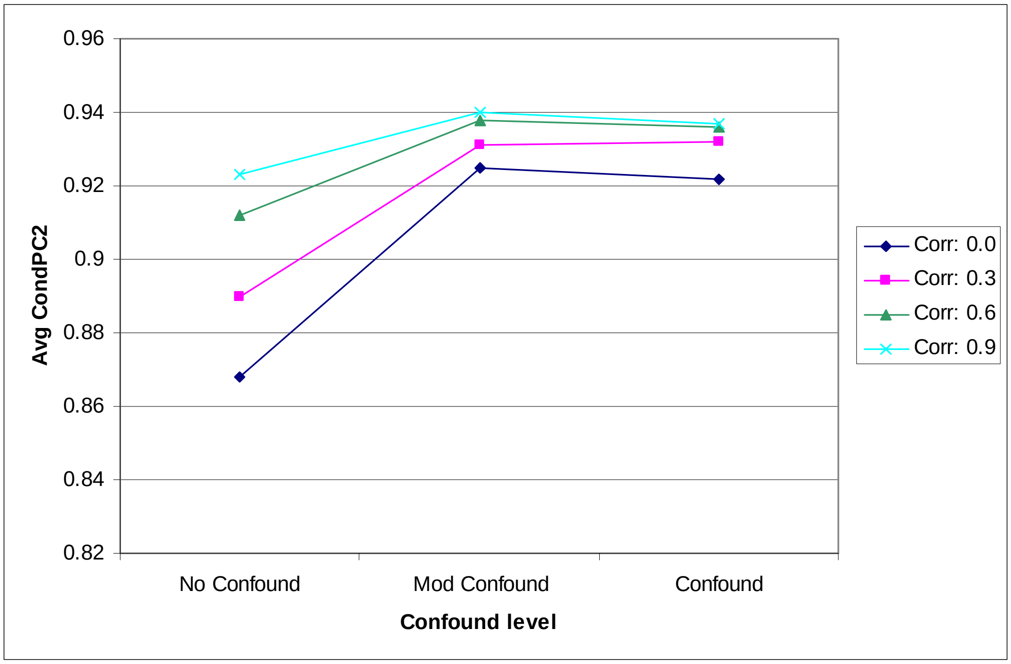
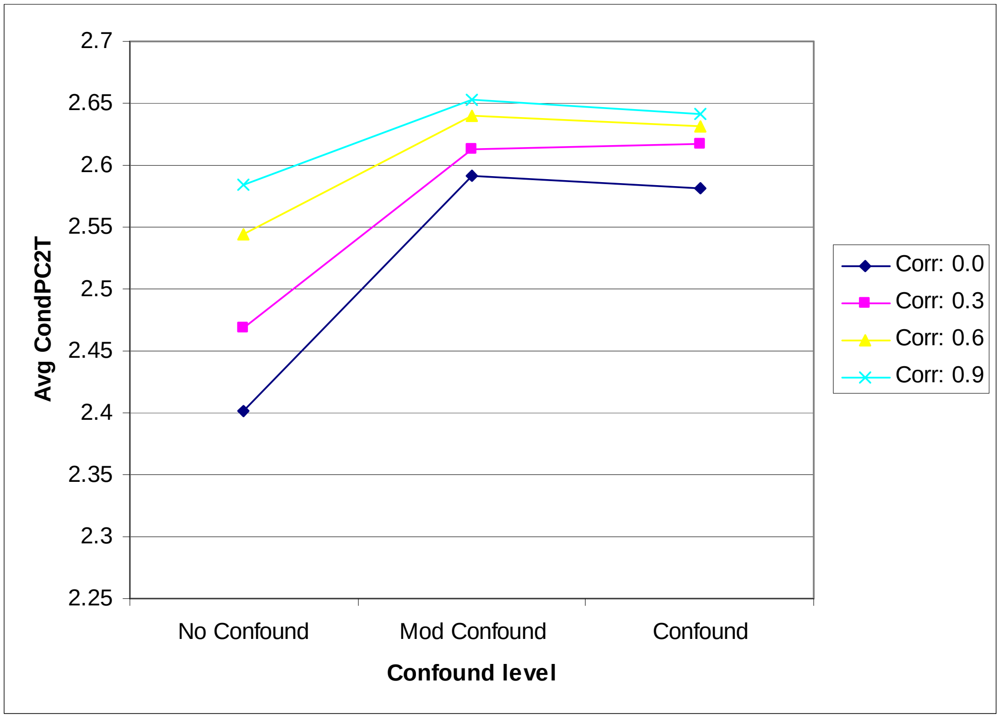


Figure 16: Lower grade confound and correlation interaction with CONDPC2T as the criterion



The plots on these two figures above illustrate an ordinal interaction between confound and correlation similar to that observed for CONDPC1. Since the raw metric is easier to understand, Figure 15 is discussed; however, the conclusions would also apply to Figure 16. For the no confound level, the differences at the four levels of correlation on CONDPC2 are much larger than the differences at the moderate or high confound levels. Again it seems that the confound of difficulty with dimensionality has a greater effect on CONDPC2 when there is a smaller relationship among dimensions than when there is a larger one. It is also clear that having a high confound between difficulty and dimensionality level will result in the greatest success at classifying truly proficient simulees as such. This is simply because in the high confound condition, simulees would be administered relatively easier items and would be more likely to answer them correctly resulting in an “observed” proficient classification. This plot also helps to reinforce the mean differences on CONDPC2 between no confound and the other two levels of confound and across the four levels of correlation. Table 26, below, includes the mean CONDPC2 and CONDPC2T values at each of the levels of confound and correlation.

Table 26: Average lower grade CONDPC2T and CONDPC2 values for confound and correlation levels across all other factors

		Correlation			
	Confound Level	0.0	0.3	0.6	0.9
CONDPC2T	No Confound	2.402	2.468	2.544	2.585
CONDPC2		0.868	0.89	0.912	0.923
CONDPC2T	Mod Confound	2.591	2.613	2.64	2.653
CONDPC2		0.925	0.931	0.938	0.94
CONDPC2T	Confound	2.581	2.617	2.632	2.641
CONDPC2		0.922	0.932	,936	0.937

Upper grade analysis

<u>Proportion correctly classified as not proficient</u>

The multifactor ANOVA output, raw and arcsin transformed, for the proportion of upper grade simulees classified correctly as not proficient are found below in Tables 27 and 28, respectively:

Table 27: Upper Grade Multifactor ANOVA on Raw Data for CONDPC1

Tests of Between-Subjects Effects

Dependent Variable: CONDPC1

Source	Type III SS	df	Mean Sq	F	Sig.	Partial η^2
Corrected Model	23.120	47	0.492	1045.856	0.000	0.672
Intercept	18455.240	1	18455.240	39238256.402	0.000	0.999
CONFOUND	0.627	2	0.314	666.913	0.000	0.053
COMMON	0.092	1	0.092	196.398	0.000	0.008
ABILITY	0.364	1	0.364	774.061	0.000	0.031
CORRELAT	19.459	3	6.486	13791.140	0.000	0.633
CONFOUND * COMMON	0.268	2	0.134	284.722	0.000	0.023
CONFOUND * ABILITY	0.395	2	0.198	420.305	0.000	0.034
COMMON * ABILITY	0.025	1	0.025	53.684	0.000	0.002
CONFOUND * COMMON * ABILITY	0.072	2	0.036	76.994	0.000	0.006
CONFOUND * CORRELAT	0.387	6	0.064	136.985	0.000	0.033
COMMON * CORRELAT	0.084	3	0.028	59.341	0.000	0.007
CONFOUND * COMMON * CORRELAT	0.215	6	0.036	76.046	0.000	0.019
ABILITY * CORRELAT	0.249	3	0.083	176.341	0.000	0.022
CONFOUND * ABILITY * CORRELAT	0.238	6	0.040	84.403	0.000	0.021
COMMON * ABILITY * CORRELAT	0.159	3	0.053	112.408	0.000	0.014
CONFOUND * COMMON * ABILITY * CORRELAT	0.485	6	0.081	171.818	0.000	0.041
Error	11.266	23952	0.000			
Total	18489.625	24000				
Corrected Total	34.385	23999				

a. R Squared = .672 (Adjusted R Squared = .672)

Table 28: Upper Grade Multifactor ANOVA on Arcsin Transformed Data for CONDPC1T

Tests of Between-Subjects Effects

Dependent Variable: CONDPC1T

Source	Type III SS	df	Mean Sq	F	Sig.	Partial η^2
Corrected Model	208.866	47	4.444	1015.642	0.000	0.666
Intercept	141966.405	1	141966.405	32445704.742	0.000	0.999
CONFOUND	5.262	2	2.631	601.328	0.000	0.048
COMMON	1.141	1	1.141	260.696	0.000	0.011
ABILITY	4.083	1	4.083	933.070	0.000	0.037
CORRELAT	175.540	3	58.513	13372.924	0.000	0.626
CONFOUND * COMMON	2.367	2	1.184	270.493	0.000	0.022
CONFOUND * ABILITY	3.613	2	1.806	412.821	0.000	0.033
COMMON * ABILITY	0.335	1	0.335	76.556	0.000	0.003
CONFOUND * COMMON * ABILITY	0.623	2	0.311	71.166	0.000	0.006
CONFOUND * CORRELAT	3.267	6	0.545	124.446	0.000	0.030
COMMON * CORRELAT	1.020	3	0.340	77.678	0.000	0.010
CONFOUND * COMMON * CORRELAT	1.680	6	0.280	63.977	0.000	0.016
ABILITY * CORRELAT	2.744	3	0.915	209.017	0.000	0.026
CONFOUND * ABILITY * CORRELAT	1.964	6	0.327	74.824	0.000	0.018
COMMON * ABILITY * CORRELAT	1.458	3	0.486	111.110	0.000	0.014
CONFOUND * COMMON * ABILITY * CORRELAT	3.770	6	0.628	143.592	0.000	0.035
Error	104.802	23952	0.004			
Total	142280.073	24000				
Corrected Total	313.668	23999				

a. R Squared = .666 (Adjusted R Squared = .665)

Using the ANOVA output for the transformed data (Table 28), it is clear that just the correlation factor has a strong association with the criterion variable (CONDPC1T) as measured by its partial eta squared value of .626. None of the other effects reach the 0.10 criterion for further investigation. Before evaluating the marginal differences on proportion correctly classified across the various correlation levels, it is important to visually appreciate the theta vector distribution of truly not proficient simulees classified as such versus those classified as proficient across the four levels of correlation. This evaluation will help us better understand why correlation has a strong association with CONDPC1.

Figures 17 through 20, below, illustrate the distribution of theta vectors for truly not proficient simulees classified correctly and incorrectly from four example replications across all four correlation levels, 0, 0.3, 0.6, 0.9. Just as was done for the lower grade analysis, the average values on theta 1 and theta 2 for classified and misclassified simulees are included on the figures. Specifically, a replication from following cells were used: CLB0, CLB3, CLB6, and CLB9

Figure 17: Distribution of theta vectors for upper grade truly not proficient simulees; CLB0 cell

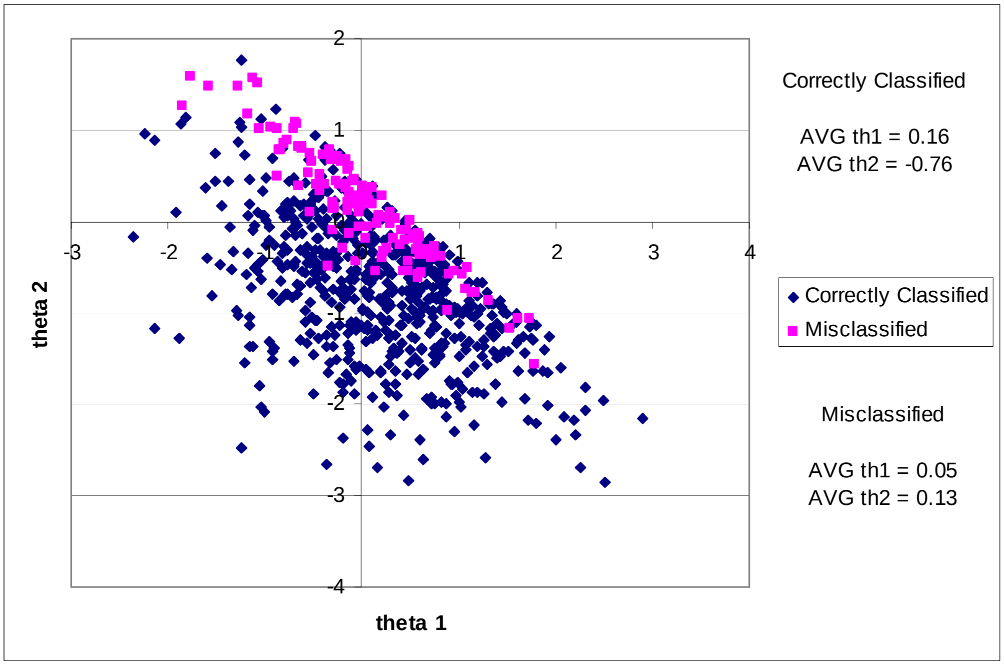


Figure 18: Distribution of theta vectors for upper grade truly not proficient simulees; CLB3 cell

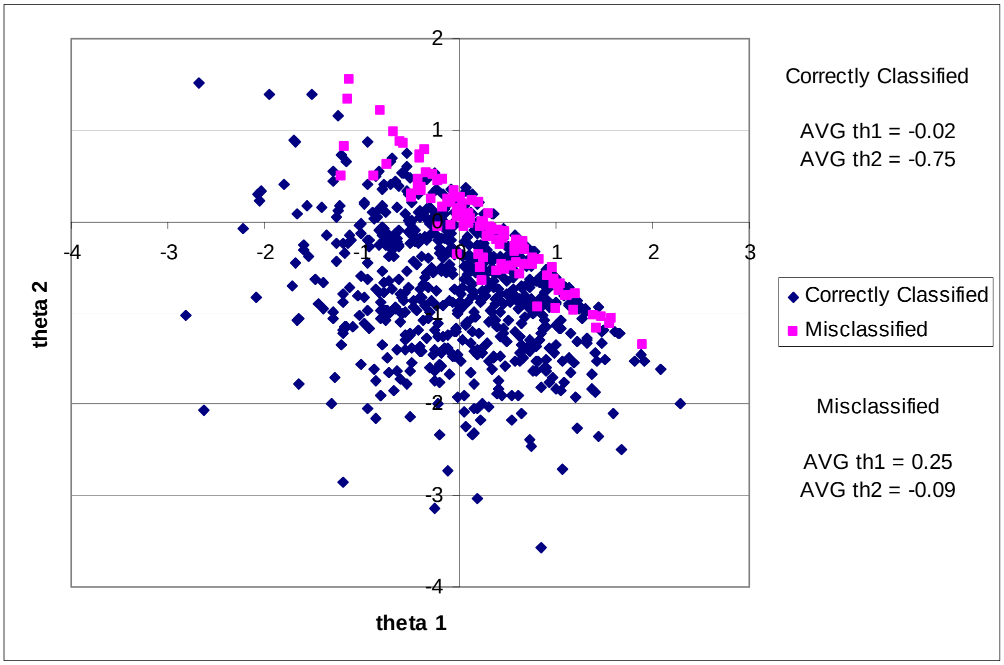


Figure 19: Distribution of theta vectors for upper grade truly not proficient simulees; CLB6 cell

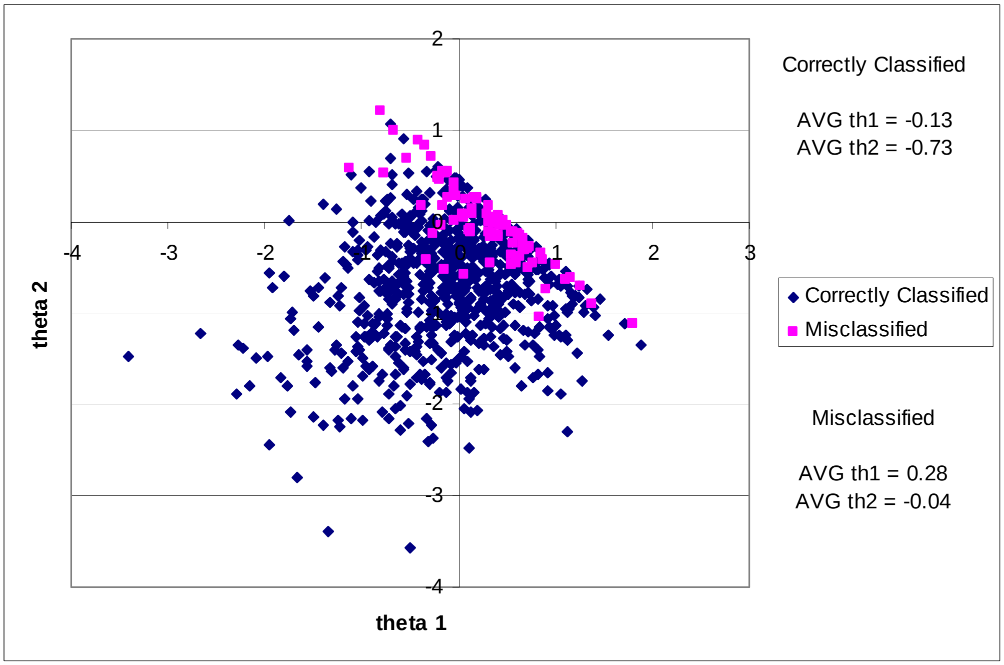
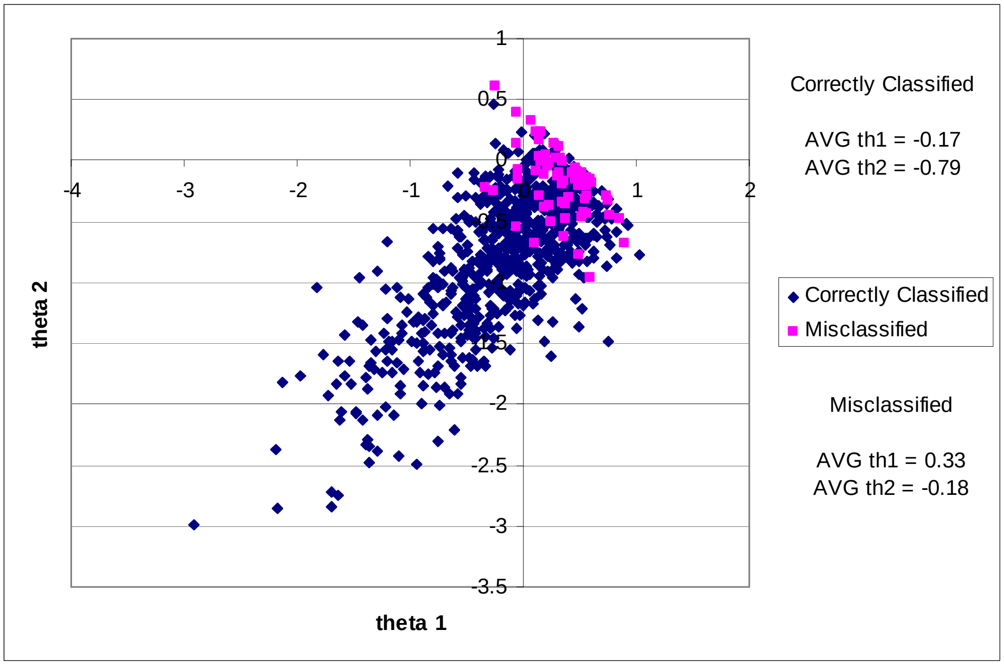


Figure 20: Distribution of theta vectors for upper grade truly not proficient simulees; CLB9 cell



In Figures 17 through 20, 83%, 87%, 90%, and 92%, respectively, of the truly not proficient simulees were classified correctly; so more simulees were classified correctly as not proficient as the correlation between dimensions got stronger. The rationale for this pattern is the same that was used for the similar result in the lower grade analysis. The simulees in each of these four cells were administered the same items to determine their true classification (entire CL test battery) and the same items to determine their "observed" classification (upper grade and common items from the CL test battery). As the theta 1 and theta 2 for each of these simulees became more highly related it is reasonable that their relative performance on the entire test battery would match their performance on the upper grade test. Thus, CONDPC1 would be expected to increase as the relationship between the theta values increases. Visually, this increasingly linear relationship can be appreciated across the four figures above. Note that the average theta

2 value for those that were misclassified is higher than those that were correctly classified. It is this strength and the administration of largely upper grade (dimension 2) items used to determine their "observed" classification that causes their incorrect "observed" classification as proficient. Generally, however, these misclassified simulees are the strongest on both dimensions among those upper grade simulees that are truly not proficient. Table 29, below, includes the marginalized (across all other factors) raw and transformed CONDPC1 average values for all four levels of the correlation factor in the upper grade.

Table 29: Average upper grade CONDPC1T and CONDPC1 values for correlation levels across all other factors

	0.0	0.3	0.6	0.9
CONDPC1T	2.308	2.424	2.471	2.534
CONDPC1	0.832	0.876	0.89	0.909

There is approximately a seven percent increase in CONDPC1 on the raw metric between no correlation and 0.9 correlation. Further, the change in raw percentage by correlation level ranges between approximately 1 to 4 percent. The largest increase is from no correlation to 0.3 correlation.

Proportion correctly classified as proficient

The multifactor ANOVA output, raw and arcsin transformed, for the proportion of upper grade simulees classified correctly as proficient are found below in Tables 30 and 31, respectively:

Table 30: Upper Grade Multifactor ANOVA on Raw Data for CONDPC2

Tests of Between-Subjects Effects

Dependent Variable: CONDPC2

Source	Type III SS	df	Mean Sq	F	Sig.	Partial η^2
Corrected Model	2.604	47	0.055	259.263	0.000	0.337
Intercept	20434.802	1	20434.802	95615262.352	0.000	1.000
CONFOUND	0.006	2	0.003	14.462	0.000	0.001
COMMON	0.148	1	0.148	693.377	0.000	0.028
ABILITY	0.101	1	0.101	473.395	0.000	0.019
CORRELAT	1.125	3	0.375	1754.159	0.000	0.180
CONFOUND * COMMON	0.096	2	0.048	225.132	0.000	0.018
CONFOUND * ABILITY	0.216	2	0.108	504.439	0.000	0.040
COMMON * ABILITY	0.006	1	0.006	28.765	0.000	0.001
CONFOUND * COMMON * ABILITY	0.018	2	0.009	41.259	0.000	0.003
CONFOUND * CORRELAT	0.149	6	0.025	115.924	0.000	0.028
COMMON * CORRELAT	0.027	3	0.009	42.502	0.000	0.005
CONFOUND * COMMON * CORRELAT	0.092	6	0.015	71.480	0.000	0.018
ABILITY * CORRELAT	0.242	3	0.081	377.298	0.000	0.045
CONFOUND * ABILITY * CORRELAT	0.128	6	0.021	99.848	0.000	0.024
COMMON * ABILITY * CORRELAT	0.091	3	0.030	141.624	0.000	0.017
CONFOUND * COMMON * ABILITY * CORRELAT	0.160	6	0.027	124.832	0.000	0.030
Error	5.119	23952	0.000			
Total	20442.525	24000				
Corrected Total	7.723	23999				

a. R Squared = .337 (Adjusted R Squared = .336)

Table 31: Upper Grade Multifactor ANOVA on Arcsin Transformed Data for CONDPC2T

Tests of Between-Subjects Effects

Dependent Variable: CONDPC2T

Source	Type III SS	df	Mean Sq	F	Sig.	Partial η^2
Corrected Model	36.770	47	0.782	260.860	0.000	0.339
Intercept	159983.367	1	159983.367	53343953.148	0.000	1.000
CONFOUND	0.083	2	0.042	13.856	0.000	0.001
COMMON	2.121	1	2.121	707.359	0.000	0.029
ABILITY	1.487	1	1.487	495.828	0.000	0.020
CORRELAT	16.092	3	5.364	1788.504	0.000	0.183
CONFOUND * COMMON	1.308	2	0.654	217.983	0.000	0.018
CONFOUND * ABILITY	2.727	2	1.363	454.628	0.000	0.037
COMMON * ABILITY	0.113	1	0.113	37.616	0.000	0.002
CONFOUND * COMMON * ABILITY	0.215	2	0.107	35.794	0.000	0.003
CONFOUND * CORRELAT	2.201	6	0.367	122.333	0.000	0.030
COMMON * CORRELAT	0.413	3	0.138	45.931	0.000	0.006
CONFOUND * COMMON * CORRELAT	1.168	6	0.195	64.900	0.000	0.016
ABILITY * CORRELAT	3.738	3	1.246	415.404	0.000	0.049
CONFOUND * ABILITY * CORRELAT	1.700	6	0.283	94.487	0.000	0.023
COMMON * ABILITY * CORRELAT	1.175	3	0.392	130.622	0.000	0.016
CONFOUND * COMMON * ABILITY * CORRELAT	2.230	6	0.372	123.901	0.000	0.030
Error	71.834	23952	0.003			
Total	160091.971	24000				
Corrected Total	108.604	23999				

a. R Squared = .339 (Adjusted R Squared = .337)

Using the ANOVA output for the transformed data (Table 31), it appears that only the correlation factor has even a moderate association with the criterion variable (CONDPC2T) as measured by its partial eta squared value of 0.183. None of the other effects reached the 0.10 criterion for further investigation. Before evaluating the marginal differences on proportion correctly classified across the various correlation levels, it is again important to visually appreciate the theta vector distribution of truly proficient simulees classified as such versus those classified as not proficient across for the four levels of correlation. This evaluation will help us better understand how correlation is related to CONDPC2.

Figures 21 through 24 below illustrate the distribution of theta vectors for truly proficient simulees classified correctly and incorrectly from four example replications across all four correlation levels, 0, 0.3, 0.6, 0.9 with average theta values indicated. Specifically, a replication from the following cells was used: NLB0, NLB3, NLB6, and NLB9.

Figure 21: Distribution of theta vectors for upper grade truly proficient simulees; NLB0 cell

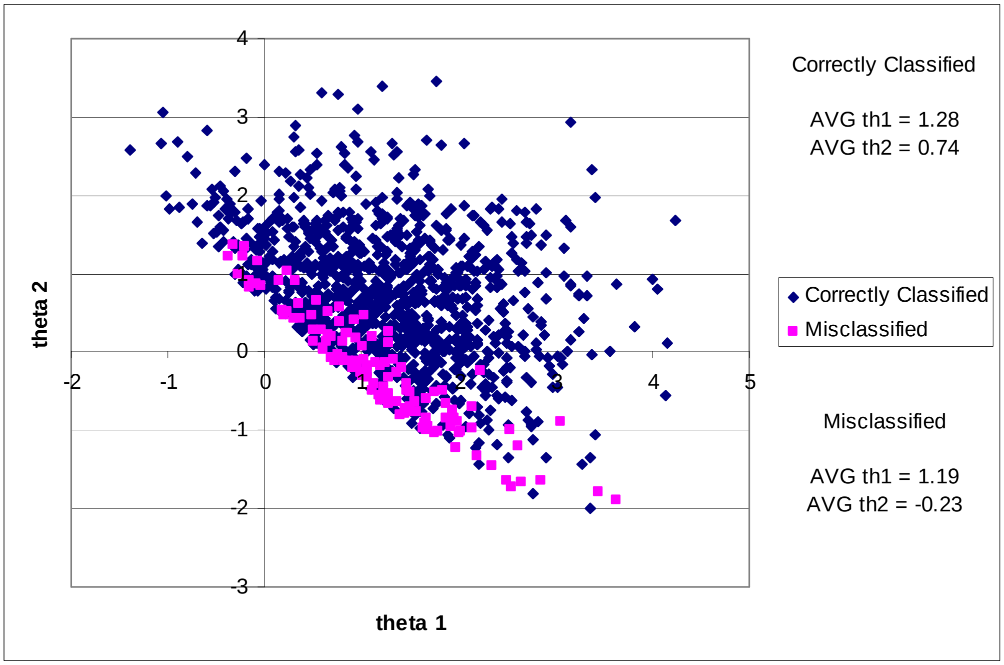


Figure 22: Distribution of theta vectors for upper grade truly proficient simulees; NLB3 cell

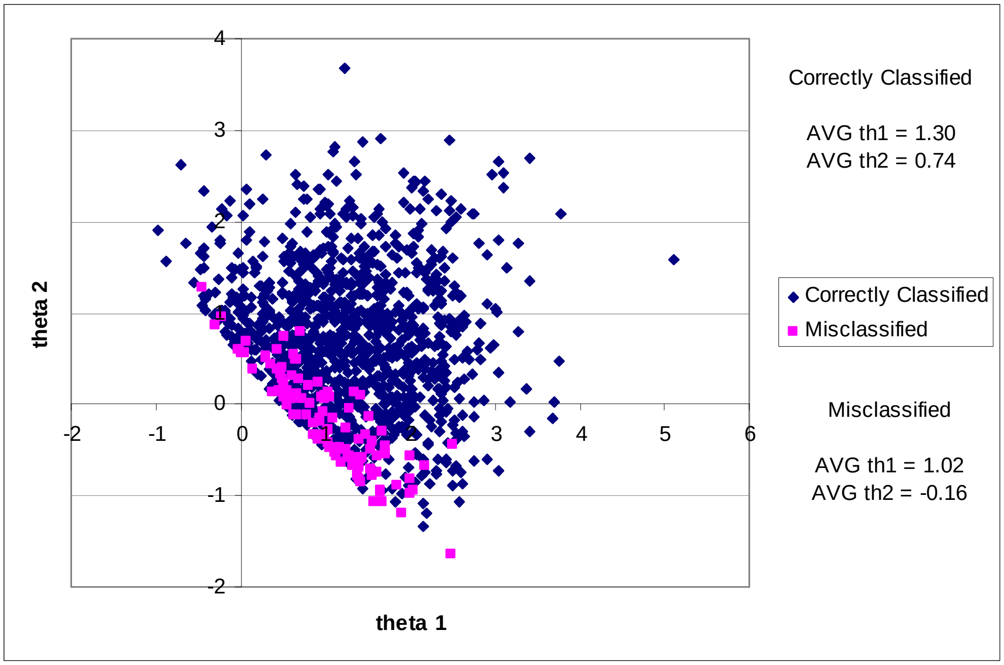


Figure 23: Distribution of theta vectors for upper grade truly proficient simulees; NLB6 cell

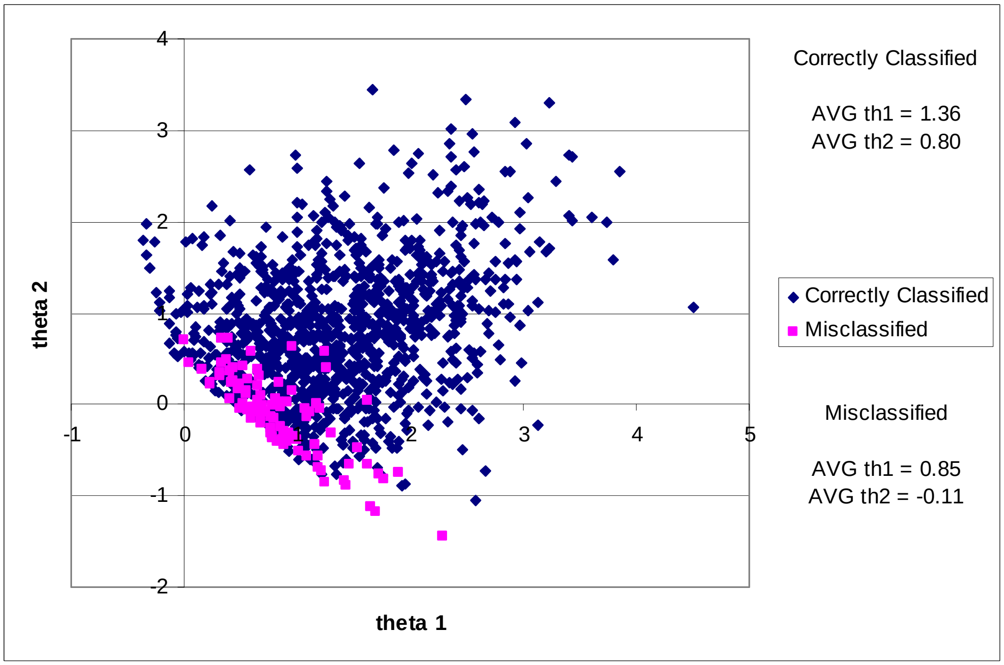
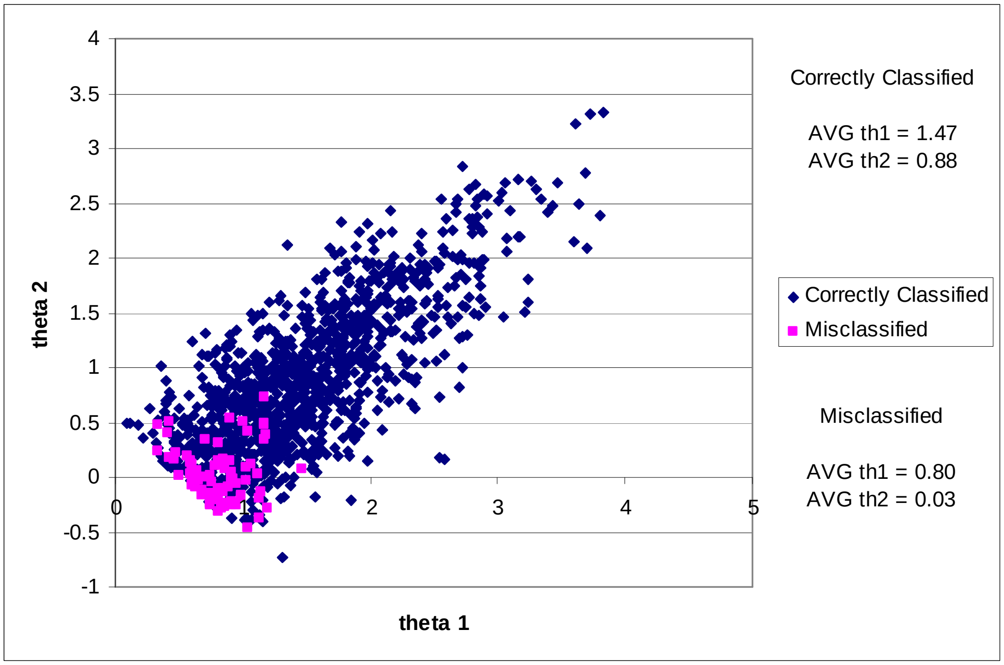


Figure 24: Distribution of theta vectors for upper grade truly proficient simulees; NLB9 cell



In Figures 21 through 24, 90%, 91%, 92%, and 93%, respectively, of the truly proficient simulees were classified correctly; so more simulees were classified correctly as proficient as the correlation between dimensions got stronger. The rationale for this pattern is that same that was used for the similar result in the lower grade analysis and for the CONDPC1 criterion in the upper grade analysis. The simulees in each of these four cells were administered the same items to determine their true classification (entire NL test battery) and the same items to determine their "observed" classification (upper grade and common items from the NL test battery). Consistent with previous results, as the theta 1 and theta 2 for each of these simulees became more highly related, it is reasonable that their relative performance on the entire test battery would match their performance on the upper grade test. Thus, CONDPC2 would be expected to increase as the relationship between the theta values increases. Visually, this increasingly linear relationship can be appreciated across the four figures above. Note that these misclassified simulees are the weakest on both dimensions among those upper grade simulees that are truly proficient; however, it was their relative strength on dimension one (at least 0.8, on average) that resulted in their true classification as proficient (based on the entire test battery) and their weakness on dimension two (at most 0.03, on average) that resulted in their "observed" classification as not proficient (based on largely upper grade focused items). Table 32 below includes the marginalized (across all other factors) raw and transformed CONDPC2 average values for all four levels of the correlation factor in the upper grade.

Table 32: Average upper grade CONDPC2T and CONDPC2 values for correlation levels across all other factors

	0.0	0.3	0.6	0.9
CONDPC2T	2.558	2.557	2.500	2.616
CONDPC2	0.916	0.916	0.927	0.931

There is approximately a one percent increase in CONDPC2 on the raw metric between no correlation and 0.9 correlation. This data suggest that there is very little gain in the upper grade CONDPC2 as the correlation between dimensions increases. Note that the overall R- square value for this multifactor ANOVA (0.339) was considerably less than for the other three (at least 0.650).

Summary of study

This study helped to gain insight into the factors that affect classification accuracy in vertical scaling when a multidimensional model is misspecified with a unidimensional model. Classification accuracy was measured by the probability of successfully classifying not proficient simulees (or the true negative rate) and the probability of successfully classifying proficient simulees (or the true positive rate). The relationship of item difficulty and dimensionality, the relationship between ability dimensions, choice of common items, and difference in mean abilities between grades were the factors considered. Generally, it was only the relationship between item difficulty and dimensionality and the relationship between ability dimensions that had an effect on the conditional classification accuracy.

Across both grades and both criterion variables, the correlation between ability dimensions had an effect on classification accuracy in the direction one would expect. As the relationship became stronger, the values increased on both true positive and true negative rates. The magnitude of the increase across correlation levels and classification rates ranged from one to seven percent. These results make intuitive sense because it is to be expected that as the relationship among dimensions increases performance on the complete test battery (used to determine truth) would match performance on the grade level test (used to determine the observed classification). The larger percentages (6 and 7) were observed for the true negative rates in both grades. The magnitudes of the change in conditional probabilities across correlation levels are addressed in the limitations and implications section.

In only the lower grade was an effect observed for the relationship between item difficulty and dimensionality. The magnitudes (3 percent) of the change across levels were approximately the same for the true negative and positive rates; however, the direction of the change differed. The true negative rate was the highest when there was no relationship and the true positive rate was the highest when there was at least a moderately strong relationship. (Note that there was no meaningful difference in either of these rates between the moderate confound and confound levels.) The direction of these results is also intuitive. The true negative rate was the highest when the difficulty of items was the hardest on the grade level test because it is more likely that one will be able to classify truly not proficient simulees as such when they are administered more difficult items. Conversely, the true positive rate was the highest when the difficulty of the items was relatively easy because it is more likely that one will be able to classify truly proficient simulees as such when they are administered relatively easier items. An ordinal interaction for the relationship between item difficulty and dimensionality and the relationship between ability dimensions was also observed. This interaction simply indicated that the relationship between item difficulty and dimensionality has a greater effect on the conditional classification rates when there is no relationship between ability dimensions than when there is a strong relationship.

In addition to those factors for which we observed an effect, it is equally important to acknowledge those for which we did not. First, there was no meaningful effect for choice of common items. Certainly, this null result is tied to this particular research design (as are the previously discussed results) where one third of each grade level test was common items; however it does suggest that common item choice may have no more than a minimal effect on conditional classification accuracy given this testing particular situation (60 total items, 20 common). This finding is not without precedent given that work on mixed-format test equating by Cao (2008) indicated that content representativeness had a minimal effect on classification consistency. This result could be informative for test developers who may be stressed for finding common items that function as a “mini-test”; although, best practices in testing do suggest that common items be reflective of the test as a whole (Kolen & Brennan, 2004). Note that while there was no meaningful effect for common items in this study using conditional classification rates as the criterion variables other criteria such as equating functions could indicate a meaningful effect for choice of common items as shown in Loyd and Hoover (1980) and Harris and Hoover (1987).

The results also suggested that the small and large differences in mean abilities for the lower and upper grades had no more than a minimal effect on conditional classification accuracy. That is, from the classification perspective, the differences in abilities between the lower and upper grades did not affect the concurrent calibration of the item response data and the subsequent standard setting.

Chapter 8: Discussion

Limitations of the research design

Generalization of these results is closely tied to the limitations of this study. There are three main limitations to the research design:

1. The 60% proficient standard
2. The choice of average abilities on both dimensions for the lower and upper grades
3. The method for establishing true status (proficient or not proficient) for the simulees

In order to appreciate the magnitudes and meaningfulness of the effects discovered (and not discovered) in this study a sensitivity analysis would be warranted. This analysis would consider variations to both the percentage proficient used in determining the population based cutscores (perhaps, 40% and 80%) as well as the average ability values for each grade on each dimension. The average values for each grade on each dimension could be raised or lowered. Doing this analysis would provide further support for the results presented and help researchers gain an appreciation for the degree to which the magnitudes of the effect change as ability levels and standards change. Making such changes would, perhaps, result in a confound effect and/or a confound and correlation interaction effect for the upper grade similar to what was shown in this study for the lower grade. However, the choices for this research design were reasonable and justification was provided; the results, therefore, can be interpreted as potentially real. Lastly, note that as states strive to meet the 100% proficient goal of

NCLB by 2014, the results of this study will become less relevant in the NCLB context because the proportion of students near cut points may depart increasingly from the ones used here.

There are likely alternatives to establishing the true classification of simulees. One such alternative would be to establish truth based simply on the two-dimensional generating parameters of each simulee. The choice here was to project a simulee's generating two-dimensional theta vector onto a unidimensional scale via a number correct transformation on the entire test battery (i.e. across both grades). A similar procedure was used in Reckase and Li (2007) and this procedure has meaning from an operational perspective. That is, scores are typically reported on a unidimensional scale and one could conceptualize administering an entire test battery to students. In fact, administering a full test battery to students is done to a certain degree in a scaling test linking design (Kolen & Brennan, 2004). It is fully acknowledged, though, that establishing truth in a different way could cause different results and conclusions.

Implications

The most substantial implication of this research is for the community using vertically scaled tests to be aware of a test's dimensionality characteristics and its impact on the use of the test scores (in this case, for classification purposes). It was shown that the relationship of test item difficulty with the ability dimensions as well as the relationship between the ability dimensions themselves has an effect on classification accuracy. Certainly it is always expected that there will be some level of misclassification, but understanding the degree to which the test itself and the construct(s) the test measures contribute to the ability of using test scores to classify students is important. For test developers, knowing that that two abilities are highly correlated might minimize the need for them to be concerned about the degree to which item difficulty is confounded with the two abilities tested on their exam (even if the test was originally designed to just test one of the two abilities). Additionally, knowing whether or not difficulty is confounded with dimensionality on their exam might help test developers inform test users how to appropriately use test results. That is, from the lower grade perspective in this study, a high confound of difficulty with dimensionality will minimize the chance of not passing a truly proficient student; however, it could also result in increasing the chances of passing a truly not proficient student. Further, results suggest a confound of item difficulty with dimensionality has a greater impact on conditional classification accuracy rates when there is a smaller correlation among dimensions. The magnitudes of the differences in conditional classification accuracy rates across the levels of confound and correlation can certainly help inform test design and use as different types of classification errors might have varied consequences across test users. Thus, test developers can advise their users of the pros and cons of their tests based on the importance and consequences of different decision made from the associated test scores. Above all, though, these results argue for the test developer to first be aware of the item and ability relationships in the domain they are testing.

While these results are important to the developer in administration to the masses, these results can also help inform teachers and administrators who often have to deal with students on an individual level. When students are held back in school (because of, say, being classified as not proficient), it is necessary to try to figure out why. Knowledge of the dimensionality of the test used to classify students can help explain the reasons. Using the various theta 1 by theta 2 plots presented in this study, teachers can find where their students fall. In some cases, they could be legitimately held back. In other cases, they could be a borderline proficient student held back because the test was more difficult than it should have been. Heubert and Hauser (1999) describe in *High Stakes: Testing for Tracking, Promotion, and Graduation* that an assessment must lead to decisions that are educationally beneficial. Thus, it is important for schools to know how to make the appropriate decisions from test scores and when further investigation for a given student is necessary. Heubert and Hauser argue that effective remedial support services should be available for low-performing students. Knowing the dimensionality of the assessment used for classification decisions and why mistakes could be made can certainly inform the appropriate course of action for remediation. Of course, this would require that the teacher have additional academic information about the student (i.e. performance on other tests that focus on both ability dimensions 1 and 2) and knowledge of the relationship of the test items to the various ability dimensions. Due to the effort and financial resources that would likely have to be devoted to the efforts described above, it is less likely that schools on their own would be able to use and apply the results of this study relative to the ability of test developers to do so and share the appropriate information with the schools.

Teachers, school administrators, and test developers, however, can work collaboratively to better understand and potentially revise the tests they use to classify students. When the dimensionality and the associated classification implications of a given test are explained to school officials, they could simply decide that the test meets their needs.   Alternately, school systems might have different philosophies on the values of passing truly not proficient students and/or holding back truly proficient students. Based on the results of this study, test developers could potentially manage the confound of item difficulty with dimensionality to meet a school system’s needs. As mentioned earlier (from the perspective of the lower grade), increasing the confound of item difficulty with dimensionality would increase the true positive rate and decreasing the confound of item difficulty with dimensionality would increase the true negative rate. The impact of changing the item difficulty’s relationship with the dimensions would be based on how strongly the ability dimensions are related. There is less “bang for the buck” of change when the ability dimensions are strongly related.

Schools systems would benefit from understanding that a given test that is perceived as testing a single ability might, in fact, be measuring multiple abilities. Knowing this, may contribute to a revision of the curriculum. If school systems learned that a considerable amount of reading comprehension is measured on their math proficiency test (in the context of word problems, for example) changing the emphasis or ordering of reading comprehension topics in the grade level reading or English might be warranted to ensure that their students are better prepared for the test.

As the complexities of measuring a single construct across grades comes to light, school systems could decide that using multidimensional models is worthwhile. Work on this has already begun. Reckase and Martineau (2004) concluded that multidimensional models should be used in the vertical scaling of science tests. They supported their conclusion by observing that students grow on different dimensions in science at different rates over time and that the knowledge and skills assessed on tests can vary significantly across grades. Patz and Yao (2007) also illustrated the usefulness of multidimensional models for vertical scaling for a writing assessment across 5 grades. They showed that four dimensions emerged. They also presented evidence that suggested that item type (multiple-choice versus constructed response) was related to dimensionality.

Lastly, exploring the dimensionality of assessments may provide new opportunities. If test developers do spend time understanding the dimensionality of their large scale assessments, they can, perhaps, leverage what they learn into the development of formative assessments. Formative assessments can be especially useful when there is a strong match to what is taught in the classroom and what is being assessed in a summative context (Stout, 2007). This relationship is all the more reason that the large-scale (summative) test developers could be interested in such an opportunity. Specifically, it has been argued that multidimensional IRT can be used in IRT-based Cognitive Diagnosis Models (ICDM) and these models can be used in classroom-based formative assessment (Stout, 2007). One such example of an IRT based formative assessment is the SEPUP (Science Education for Public Understanding)-Embedded Assessment Project which uses a multidimensional Rasch model (Sloane, Wilson, Samson, 1996).

In consideration of the multidimensionality arguments presented by Paris (2005) regarding reading skills, one could envision that a dimensionality analysis of a summative reading assessment (especially one that spans many grades) could provide insight into the construction of appropriate formative assessments in reading. Specifically, different assessments could be created to assess skills (dimensions) that are attained and mastered quickly (constrained) and those that are continually developing (unconstrained). Knowing the levels and/or the attainment of any of these skills for the students in their classes could be of great value to classroom teachers for lesson planning and summative test preparation.

Extensions

In addition to the sensitivity analysis proposed in the *Limitations* section, extensions to this research would largely revolve around the various decision points involved in conducting vertical scaling as well as a manipulation of some the simulee population assumptions.

The first extension would be to conduct separate grade calibration instead of concurrent calibration. While there is no consensus in the literature which method is correct, both methods are typically considered when developing a vertical scaling design and it would be informative to be able to compare results. Further separate calibration is suggested as superior to concurrent calibration by researchers (e.g. Kolen & Brennan, 2004) when multidimensionality is suspected. Of course, by conducting separate calibration across many grades the accumulation of linking errors would be of concern. As noted earlier, concurrent calibration is preferred when there is a strong assumption of unidimensionality in the item response data.

In this study, there was a reasonable proportion (33%) of common items in each of the grade level tests and one of the common item conditions was a true “mini” test. However, these ideal conditions and best practices are not always achieved by test developers. Therefore, it would be useful to extend this research to conditions where there are fewer common items (either in absolute number or in proportion to the total test) that may or may not function as a “mini” test. Those results could inform the work of those responsible for tests where the ideal conditions for common items are not met.

The correlation of ability dimensions and the variance of ability on each dimension was always the same for both grade levels in each cell of this study. To address concerns of scale shrinkage, it would be useful to reduce the variance of ability on the upper grade test. Additionally, work by Reckase and Li (2007), for example, showed that the relationships among dimensions can change from grade to grade. Thus, varying the correlation between dimensions across grades should be addressed in future work.

Lastly, the multidimensional IRT item generating models and unidimensional IRT estimating models could be varied. Given that many tests are multiple-choice, introducing a guessing parameter to both the generating model and estimating model would likely be the first step. Thus, the three parameter extension to the MC2PL model (Reckase, 1997) could be used for the item response generating model and the 3PL model could be used for the unidimensional estimating model.

Appendix A: Example data generation code

```
/*get log to a text file*/
proc printto log="c:\dissertation2\CL.log";
Run;

DATA lowerC; /*read in lower grade confound generating parameters*/
INFILE "C:\dissertation\parameters\lowerC.txt";
input a b c;
run;
DATA upperC; /*read in upper grade confound generating parameters*/
INFILE "C:\dissertation\parameters\upperC.txt";
input a b c;
run;
DATA lowerCommonC;/*read in lower grade common with confound generating
parameters*/
INFILE "C:\dissertation\parameters\lowerCommonC.txt";
input a b c;
run;

%macro GENERATE (lower_ab1, lower_ab2, upper_ab1, upper_ab2, corr,
lowercut, uppercut, test, cell);
%do it=1 %to 500;

PROC IML; /*get data into matrix form*/

/*get lowerCommonC into a matrix*/
USE lowerCommonC;
READ ALL INTO lowerCommonCmat;
CLOSE lowerCommonC;

/*get upperC into a matrix*/
USE upperC;
READ ALL INTO upperCmat;
CLOSE upperC;

/*get lowerC into a matrix*/
USE lowerC;
READ ALL INTO lowerCmat;
CLOSE lowerC;

/*generate lower grade 2D thetas*/
mu_lower = {&lower_ab1, &lower_ab2};
sigma_lower = {1.0 &corr, &corr 1.0};
call vnormal (lower_thetas, mu_lower, sigma_lower, 2000);

/*generate upper grade 2D thetas*/
mu_upper = {&upper_ab1, &upper_ab2};
sigma_upper = {1.0 &corr, &corr 1.0};
call vnormal (upper_thetas, mu_upper, sigma_upper, 2000);
```

```
/*generate MC response vectors for lower grade on unique items*/

LOWER_UNIQUE_RESPONSES = J(2000,40,.);
DO S = 1 TO 2000;
   DO J = 1 TO 40;
IF
((exp(lower_thetas[s,1]*lowerCmat[J,1]+lower_thetas[s,2]*lowerCmat[J,2]
+ lowerCmat[J,3]))/(1 +
(exp(lower_thetas[s,1]*lowerCmat[J,1]+lower_thetas[s,2]*lowerCmat[J,2]+
lowerCmat[J,3]))) >= RANUNI(0))THEN DO; LOWER_UNIQUE_RESPONSES[S,J]= 1;
END;
ELSE LOWER_UNIQUE_RESPONSES[S,J] = 0;
END;
END;


/*generate MC response vectors for upper grade on unique items*/
UPPER_UNIQUE_RESPONSES = J(2000,40,.);
DO S = 1 TO 2000;
   DO J = 1 TO 40;
IF
((exp(upper_thetas[s,1]*upperCmat[J,1]+upper_thetas[s,2]*upperCmat[J,2]
+ upperCmat[J,3]))/(1 +
(exp(upper_thetas[s,1]*upperCmat[J,1]+upper_thetas[s,2]*upperCmat[J,2]+
upperCmat[J,3]))) >= RANUNI(0))THEN DO;
UPPER_UNIQUE_RESPONSES[S,J]= 1;
END;
ELSE UPPER_UNIQUE_RESPONSES[S,J] = 0;
END;
END;

/*generate MC response vectors for lower grade on all 20 lower grade
common items*/
LOWER_common_20_RESPONSES = J(2000,20,.);
DO S = 1 TO 2000;
   DO J = 1 TO 20;
IF ((exp(lower_thetas[s,1]*lowerCommonCmat[J,1]+lower_thetas[s,2]*
lowerCommonCmat[J,2]+ lowerCommonCmat[J,3]))/(1 +
(exp(lower_thetas[s,1]*lowerCommonCmat[J,1]+lower_thetas[s,2]*lowerComm
onCmat[J,2]+ lowerCommonCmat[J,3]))) >= RANUNI(0))THEN DO;
LOWER_common_20_RESPONSES[S,J]= 1;
END;
ELSE LOWER_common_20_RESPONSES[S,J] = 0;
END;
END;
```

```
/*generate MC response vectors for upper grade on all 20 lower grade
common items*/
UPPER_common_20_RESPONSES = J(2000,20,.);
DO S = 1 TO 2000;
   DO J = 1 TO 20;
IF ((exp(upper_thetas[s,1]*lowerCommonCmat[J,1]+upper_thetas[s,2]*
lowerCommonCmat[J,2]+ lowerCommonCmat[J,3]))/(1 +
(exp(upper_thetas[s,1]*lowerCommonCmat[J,1]+upper_thetas[s,2]*lowerComm
onCmat[J,2]+ lowerCommonCmat[J,3]))) >= RANUNI(0))THEN DO;
UPPER_common_20_RESPONSES[S,J]= 1;
END;
ELSE UPPER_common_20_RESPONSES[S,J] = 0;
END;
END;

/*generate person ids; can be applied to both grades, lower and upper
*/
IDEN=J(2000,1,.);
DO I = 1 TO 2000;
   IDEN[I,1]=I + 1000;
END;

/*create group ID for lower grade*/
lowergradeID=J(2000,1,1);
/*create group ID for upper grade*/
uppergradeID=J(2000,1,2);

/*create not administered matrix*/
notadmin=J(2000,40,9);

/*create complete set of lower grade responses*/
lower_responses = IDEN || lowergradeID || LOWER_UNIQUE_RESPONSES ||
LOWER_common_20_RESPONSES || notadmin;
upper_responses = IDEN || uppergradeID || notadmin ||
UPPER_common_20_RESPONSES || UPPER_UNIQUE_RESPONSES;
all_responses = lower_responses // upper_responses;


/*create the SAS dataset of the responses*/
CREATE responses FROM all_responses;
APPEND FROM all_responses;


/*create a matrix of all item parameters*/
items = lowerCmat // lowerCommonCmat // upperCmat;

/*get probabilities of correct response to each item for lower grade
students across entire test*/
lowergradetrue = J(2000,100,.);
DO S = 1 TO 2000;
   DO J = 1 TO 100;
lowergradetrue[S,J] =
(exp(lower_thetas[s,1]*items[J,1]+lower_thetas[s,2]*items[J,2]+
items[J,3]))/(1 +
(exp(lower_thetas[s,1]*items[J,1]+lower_thetas[s,2]*items[J,2]+
items[J,3])));
END;
END;

/*get probabilities of correct response to each item for upper grade
students across entire test*/
uppergradetrue = J(2000,100,.);
DO S = 1 TO 2000;
   DO J = 1 TO 100;
uppergradetrue[S,J] =
(exp(upper_thetas[s,1]*items[J,1]+upper_thetas[s,2]*items[J,2]+
items[J,3]))/(1 +
(exp(upper_thetas[s,1]*items[J,1]+upper_thetas[s,2]*items[J,2]+
items[J,3])));
END;
END;

/*get expected total score on entire test for lower grade students*/
lowergradetruetotal = lowergradetrue[,+];
/*PRINT lowergradetruetotal;*/
/*get expected total score on entire test for upper grade students*/
uppergradetruetotal = uppergradetrue[,+];
/*PRINT uppergradetruetotal;*/

/*round the lower grade expected scores on entire test*/
lowergradetruetotalround = J(2000,3,1);
DO S = 1 to 2000;
   lowergradetruetotalround[S,1] = lowergradetruetotal[S,1];
   lowergradetruetotalround[S,2] = ROUND (lowergradetruetotal[S,1],1);
END;
/*PRINT lowergradetruetotalround;*/

/*round the upper grade expected scores on entire test*/
uppergradetruetotalround = J(2000,3,1);
DO S = 1 to 2000;
   uppergradetruetotalround[S,1] = uppergradetruetotal[S,1];
   uppergradetruetotalround[S,2] = ROUND (uppergradetruetotal[S,1],1);
END;
/*PRINT uppergradetruetotalround;*/

/*get true classifications for lower grade*/
DO S = 1 to 2000;
      IF (lowergradetruetotalround[S,2] >= &lowercut) THEN DO;
            lowergradetruetotalround[S,3] = 2;
                  END;
END;


/*get true classifications for upper grade*/
DO S = 1 to 2000;
      IF (uppergradetruetotalround[S,2] >= &uppercut) THEN DO;
            uppergradetruetotalround[S,3] = 2;
                  END;
END;
/*PRINT uppergradetruetotalround;*/
```

```
/*get lower grade observed scores*/
lowergraderesponses = LOWER_UNIQUE_RESPONSES ||
LOWER_common_20_RESPONSES;
uppergraderesponses = UPPER_common_20_RESPONSES ||
UPPER_UNIQUE_RESPONSES;

lowergradeobserved = lowergraderesponses[,+];
uppergradeobserved = uppergraderesponses[,+];


lowerall = lowergradetruetotalround || lowergradeobserved;
upperall = uppergradetruetotalround || uppergradeobserved;

/*in the lowerall and upperall matrices there are 4 columns: expected
score on all items unrounded, rounded, classification, observed score
on grade level test*/

/*create a sas dataset for lowerall*/
CREATE lowerdata FROM lowerall;
APPEND FROM lowerall;
/*create a sas dataset for upperall*/
CREATE upperdata FROM upperall;
APPEND FROM upperall;

/*get thetas with Identification number*/
lower_ID_thetas = IDEN || lower_thetas;
upper_ID_thetas = IDEN || upper_thetas;

/*get thetas into a SAS dataset*/
CREATE lowerthetas FROM lower_ID_thetas;
APPEND from lower_ID_thetas;

CREATE upperthetas FROM upper_ID_thetas;
APPEND from upper_ID_thetas;

Quit;
Run;

/*creating the data file of lower grade simulee scores and
classifications*/
/*FILENAME mydata1
"G:\dissertationcode\dissertation\lowerscores&it..txt";*/
FILENAME mydata1
"C:\dissertation2\&test\&cell\lowerscores&cell&it..txt";
DATA DUMMY1;
SET lowerdata;
FILE mydata1 NOPRINT NOTITLES;
PUT @1 COL1 @17 COL2 @22 COL3 @27 COL4;
RUN;

/*creating the data file of upper grade simulee scores and
classifications*/
/*FILENAME mydata2 FILENAME mydata2
"C:\dissertation2\&test\&cell\upperscores&cell&it..txt";
DATA DUMMY2;
SET upperdata;
```

```
FILE mydata2 NOPRINT NOTITLES;
PUT @1 COL1 @17 COL2 @22 COL3 @27 COL4;
RUN;



/*creating the datafile for the respones*/
/*FILENAME mydata "C:\Documents and Settings\Marc\My
Documents\dissertation code\data&it..txt";*/
FILENAME mydata "C:\dissertation2\&test\&cell\data&cell&it..txt";
DATA DUMMY;
SET responses;
FILE mydata NOPRINT NOTITLES;
PUT @1 COL1 @6 COL2 @8 COL3 @9 COL4 @10 COL5 @11 COL6 @12 COL7 @13 COL8
@14 COL9 @15 COL10 @16 COL11 @17 COL12 @18 COL13 @19 COL14 @20 COL15
@21 COL16 @22 COL17 @23 COL18 @24 COL19 @25 COL20 @26 COL21 @27 COL22
@28 COL23 @29 COL24 @30 COL25 @31 COL26 @32 COL27 @33 COL28 @34 COL29
@35 COL30 @36 COL31 @37 COL32 @38 COL33 @39 COL34 @40 COL35 @41 COL36
@42 COL37 @43 COL38 @44 COL39 @45 COL40 @46 COL41 @47 COL42 @48 COL43
@49 COL44 @50 COL45 @51 COL46 @52 COL47 @53 COL48 @54 COL49 @55 COL50
@56 COL51 @57 COL52 @58 COL53 @59 COL54 @60 COL55 @61 COL56 @62 COL57
@63 COL58 @64 COL59 @65 COL60 @66 COL61 @67 COL62 @68 COL63 @69 COL64
@70 COL65 @71 COL66 @72 COL67 @73 COL68 @74 COL69 @75 COL70 @76 COL71
@77 COL72 @78 COL73 @79 COL74 @80 COL75 @81 COL76 @82 COL77 @83 COL78
@84 COL79 @85 COL80 @86 COL81 @87 COL82 @88 COL83 @89 COL84 @90 COL85
@91 COL86 @92 COL87
@93 COL88 @94 COL89 @95 COL90 @96 COL91 @97 COL92 @98 COL93 @99 COL94
@100 COL95 @101 COL96 @102 COL97 @103 COL98 @104 COL99 @105 COL100 @106
COL101 @107 COL102;
RUN;

/*create files for thetas*/
FILENAME mydata3
"C:\dissertation2\&test\&cell\lowerthetas&cell&it..txt";
DATA DUMMY3;
SET lowerthetas;
FILE mydata3 NOPRINT NOTITLES;
PUT @1 COL1 @6 COL2 @30 COL3;
RUN;

FILENAME mydata4
"C:\dissertation2\&test\&cell\upperthetas&cell&it..txt";
DATA DUMMY4;
SET upperthetas;
FILE mydata4 NOPRINT NOTITLES;
PUT @1 COL1 @6 COL2 @30 COL3;
RUN;

%end;
%mend Generate;


/*data DUMMY;*/
/* cell naming Confound(C)/NoConfound(N), Lower grade common items
(L)/Both grade common items(A), Big ability difference (B)/Small
Ability difference (S), Correlation of Dimensions 0, .3, .6, .9
(0,3,6,9)*/
```

```
%GENERATE(0,-.2,.8,.2,0,44,57,CL,CLB0);
%GENERATE(0,-.2,.8,.2,.3,43,56,CL,CLB3);
%GENERATE(0,-.2,.8,.2,.6,43,57,CL,CLB6);
%GENERATE(0,-.2,.8,.2,.9,41,56,CL,CLB9);
%GENERATE(0,-.2,.4,0,0,44,51,CL,CLS0);
%GENERATE(0,-.2,.4,0,.3,43,50,CL,CLS3);
%GENERATE(0,-.2,.4,0,.6,42,49,CL,CLS6);
%GENERATE(0,-.2,.4,0,.9,42,49,CL,CLS9);
/*run;*/

/*reset log back to normal location*/
proc printto;
run;
```

Appendix B: Example Bilog-MG IRT estimation code

```
>GLOBAL DFNAME = 'C:\dissertation2\CL\CLB0\dataCLB01.txt',
       NPArm=2,
       LOGistic,
       SAVe;
>SAVE SCOre='C:\dissertation2\CL\CLB0\bilogrunCLB01.SCO',
PARM='C:\dissertation2\CL\CLB0\bilogrunCLB01.PAR';
>LENGTH NITems = (100),
       NVAriant = (0);
>INPUT NTOtal = 100,
     NALt = 5,
     NGROUPS = 2,
     NIDchar = 4,
     NFNAME= 'C:\dissertation\EXAMPL05testNOT1.nfn';
>ITEMS INUM = (1(1)100), INAMES=(M01(1)M100);
>TEST TNAme = CLB01, INUM = (1(1)100);
>GROUP1 GNAME='LOWER', LENGTH=60, INUM=(1(1)60);
>GROUP2 GNAME='UPPER', LENGTH=60, INUM=(41(1)100);
(4A1, 1X, I1, 1X, 100A1)
>CALIB NQPt = 51,
      NORMAL,
      CYClE = 30,
      TPRIOR,
      REFERENCE=1;
>SCORE METHOD=2,
       IDIST=3,
       NOPRINT,
       RSCTYPE=0;
```

Appendix C: Item generating parameters

Table C1: Item parameters for confound and lower grade common items test battery

lower grade item number	angle with dim 1	a1	a2	d
1	0	1.000	0.000	0.698
2	0	1.000	0.000	0.769
3	0	1.000	0.000	1.111
4	0	1.000	0.000	0.672
5	0	1.000	0.000	0.824
6	5	0.996	0.087	1.206
7	5	0.996	0.087	0.804
8	5	0.996	0.087	1.111
9	5	0.996	0.087	0.972
10	5	0.996	0.087	1.145
11	10	0.985	0.174	0.566
12	10	0.985	0.174	1.015
13	10	0.985	0.174	0.652
14	10	0.985	0.174	0.444
15	10	0.985	0.174	0.694
16	15	0.966	0.259	0.340
17	15	0.966	0.259	0.308
18	15	0.966	0.259	0.674
19	15	0.966	0.259	0.907
20	15	0.966	0.259	0.234
21	20	0.940	0.342	0.549
22	20	0.940	0.342	0.698
23	20	0.940	0.342	0.652
24	20	0.940	0.342	0.752
25	20	0.940	0.342	0.750
26	25	0.906	0.423	0.424
27	25	0.906	0.423	0.370
28	25	0.906	0.423	0.900
29	25	0.906	0.423	0.394
30	25	0.906	0.423	0.594
31	85	0.087	0.996	-0.795
32	85	0.087	0.996	-1.270
33	85	0.087	0.996	-1.148
34	85	0.087	0.996	-1.033
35	85	0.087	0.996	-0.872
36	90	0.000	1.000	-0.715
37	90	0.000	1.000	-1.190
38	90	0.000	1.000	-1.210
39	90	0.000	1.000	-0.850

lower grade item number	angle with dim 1	a1	a2	d
40	90	0.000	1.000	-1.396
common item number				
1	0	1.000	0.000	0.929
2	0	1.000	0.000	1.081
3	5	0.996	0.087	0.561
4	5	0.996	0.087	0.833
5	10	0.985	0.174	0.603
6	10	0.985	0.174	0.816
7	15	0.966	0.259	0.184
8	15	0.966	0.259	0.374
9	20	0.940	0.342	0.678
10	20	0.940	0.342	0.397
11	25	0.906	0.423	0.688
12	25	0.906	0.423	0.362
13	30	0.866	0.500	0.561
14	30	0.866	0.500	0.179
15	90	0.000	1.000	-1.132
16	90	0.000	1.000	-1.149
17	85	0.087	0.996	-1.141
18	85	0.087	0.996	-0.820
19	80	0.174	0.985	-0.933
20	80	0.174	0.985	-0.529
upper grade item number				
1	90	0.000	1.000	-1.247
2	90	0.000	1.000	-1.025
3	90	0.000	1.000	-0.869
4	90	0.000	1.000	-1.058
5	90	0.000	1.000	-0.987
6	85	0.087	0.996	-0.813
7	85	0.087	0.996	-0.950
8	85	0.087	0.996	-1.218
9	85	0.087	0.996	-0.894
10	85	0.087	0.996	-0.870
11	80	0.174	0.985	-0.786
12	80	0.174	0.985	-0.591
13	80	0.174	0.985	-0.858
14	80	0.174	0.985	-0.160
15	80	0.174	0.985	-1.315
16	75	0.259	0.966	-0.872
17	75	0.259	0.966	-0.641
18	75	0.259	0.966	-1.178
19	75	0.259	0.966	-0.638
20	75	0.259	0.966	-0.380
21	70	0.342	0.940	-0.212
22	70	0.342	0.940	-0.771
23	70	0.342	0.940	-0.390
24	70	0.342	0.940	-0.806

upper grade item number	angle with dim 1	a1	a2	d
25	70	0.342	0.940	-0.244
26	65	0.423	0.906	-0.610
27	65	0.423	0.906	-0.583
28	65	0.423	0.906	0.138
29	65	0.423	0.906	-0.648
30	65	0.423	0.906	-0.533
31	15	0.966	0.259	0.717
32	15	0.966	0.259	0.906
33	15	0.966	0.259	0.778
34	15	0.966	0.259	0.918
35	15	0.966	0.259	0.674
36	0	1.000	0.000	0.849
37	0	1.000	0.000	1.291
38	0	1.000	0.000	0.406
39	0	1.000	0.000	0.927
40	0	1.000	0.000	0.784

Table C2: Item parameters for confound and both grades common items test battery

lower grade item number	angle with dim 1	a1	a2	d
1	0	1.000	0.000	0.698
2	0	1.000	0.000	0.769
3	0	1.000	0.000	1.111
4	0	1.000	0.000	0.672
5	0	1.000	0.000	0.824
6	5	0.996	0.087	1.206
7	5	0.996	0.087	0.804
8	5	0.996	0.087	1.111
9	5	0.996	0.087	0.972
10	5	0.996	0.087	1.145
11	10	0.985	0.174	0.566
12	10	0.985	0.174	1.015
13	10	0.985	0.174	0.652
14	10	0.985	0.174	0.444
15	10	0.985	0.174	0.694
16	15	0.966	0.259	0.340
17	15	0.966	0.259	0.308
18	15	0.966	0.259	0.674
19	15	0.966	0.259	0.907
20	15	0.966	0.259	0.234
21	20	0.940	0.342	0.549
22	20	0.940	0.342	0.698
23	20	0.940	0.342	0.652
24	20	0.940	0.342	0.752
25	20	0.940	0.342	0.750
26	25	0.906	0.423	0.424
27	25	0.906	0.423	0.370
28	25	0.906	0.423	0.900
29	25	0.906	0.423	0.394
30	25	0.906	0.423	0.594
31	85	0.087	0.996	-0.795
32	85	0.087	0.996	-1.270
33	85	0.087	0.996	-1.148
34	85	0.087	0.996	-1.033
35	85	0.087	0.996	-0.872
36	90	0.000	1.000	-0.715
37	90	0.000	1.000	-1.190
38	90	0.000	1.000	-1.210
39	90	0.000	1.000	-0.850
40	90	0.000	1.000	-1.396
common item number				
1	0	1.000	0.000	0.929
2	5	0.996	0.087	0.561
3	10	0.985	0.174	0.603
4	15	0.966	0.259	0.184

common item number	angle with dim 1	a1	a2	d
5	20	0.940	0.342	0.678
6	25	0.906	0.423	0.688
7	30	0.866	0.500	0.561
8	90	0.000	1.000	-1.132
9	85	0.087	0.996	-1.141
10	80	0.174	0.985	-0.933
11	90	0.000	1.000	-0.801
12	85	0.087	0.996	-0.737
13	80	0.174	0.985	-0.666
14	75	0.259	0.966	-0.689
15	70	0.342	0.940	-0.589
16	65	0.423	0.906	-0.798
17	60	0.500	0.866	-0.080
18	0	1.000	0.000	1.086
19	5	0.996	0.087	0.835
20	10	0.985	0.174	1.049
upper grade item number				
1	90	0.000	1.000	-1.247
2	90	0.000	1.000	-1.025
3	90	0.000	1.000	-0.869
4	90	0.000	1.000	-1.058
5	90	0.000	1.000	-0.987
6	85	0.087	0.996	-0.813
7	85	0.087	0.996	-0.950
8	85	0.087	0.996	-1.218
9	85	0.087	0.996	-0.894
10	85	0.087	0.996	-0.870
11	80	0.174	0.985	-0.786
12	80	0.174	0.985	-0.591
13	80	0.174	0.985	-0.858
14	80	0.174	0.985	-0.160
15	80	0.174	0.985	-1.315
16	75	0.259	0.966	-0.872
17	75	0.259	0.966	-0.641
18	75	0.259	0.966	-1.178
19	75	0.259	0.966	-0.638
20	75	0.259	0.966	-0.380
21	70	0.342	0.940	-0.212
22	70	0.342	0.940	-0.771
23	70	0.342	0.940	-0.390
24	70	0.342	0.940	-0.806
25	70	0.342	0.940	-0.244
26	65	0.423	0.906	-0.610
27	65	0.423	0.906	-0.583
28	65	0.423	0.906	0.138
29	65	0.423	0.906	-0.648
30	65	0.423	0.906	-0.533

upper grade item number	angle with dim 1	a1	a2	d
31	15	0.966	0.259	0.717
32	15	0.966	0.259	0.906
33	15	0.966	0.259	0.778
34	15	0.966	0.259	0.918
35	15	0.966	0.259	0.674
36	0	1.000	0.000	0.849
37	0	1.000	0.000	1.291
38	0	1.000	0.000	0.406
39	0	1.000	0.000	0.927
40	0	1.000	0.000	0.784

Table C3: Item parameters for moderate confound and lower grade common items test battery

lower grade item number	angle with dim 1	a1	a2	d
1	0	1.000	0.000	0.237
2	0	1.000	0.000	0.798
3	0	1.000	0.000	0.417
4	0	1.000	0.000	-0.742
5	0	1.000	0.000	0.668
6	5	0.996	0.087	0.765
7	5	0.996	0.087	1.262
8	5	0.996	0.087	0.736
9	5	0.996	0.087	0.145
10	5	0.996	0.087	0.545
11	10	0.985	0.174	0.104
12	10	0.985	0.174	0.640
13	10	0.985	0.174	0.344
14	10	0.985	0.174	0.248
15	10	0.985	0.174	1.288
16	15	0.966	0.259	-0.040
17	15	0.966	0.259	0.649
18	15	0.966	0.259	0.267
19	15	0.966	0.259	0.281
20	15	0.966	0.259	0.047
21	20	0.940	0.342	-0.117
22	20	0.940	0.342	0.086
23	20	0.940	0.342	-0.159
24	20	0.940	0.342	0.402
25	20	0.940	0.342	1.122
26	25	0.906	0.423	1.001
27	25	0.906	0.423	-1.072
28	25	0.906	0.423	0.232
29	25	0.906	0.423	0.693
30	25	0.906	0.423	0.090
31	85	0.087	0.996	-0.178
32	85	0.087	0.996	0.072
33	85	0.087	0.996	-0.832
34	85	0.087	0.996	0.429
35	85	0.087	0.996	-0.360
36	90	0.000	1.000	-0.426
37	90	0.000	1.000	-0.588
38	90	0.000	1.000	-0.642
39	90	0.000	1.000	-0.475
40	90	0.000	1.000	-0.120
common item number				
1	0	1.000	0.000	0.826
2	0	1.000	0.000	0.632

common item number	angle with dim 1	a1	a2	d
3	5	0.996	0.087	0.735
4	5	0.996	0.087	1.009
5	10	0.985	0.174	0.160
6	10	0.985	0.174	0.459
7	15	0.966	0.259	0.218
8	15	0.966	0.259	1.231
9	20	0.940	0.342	0.809
10	20	0.940	0.342	0.253
11	25	0.906	0.423	0.339
12	25	0.906	0.423	-0.158
13	30	0.866	0.500	0.070
14	30	0.866	0.500	-0.050
15	90	0.000	1.000	-0.956
16	90	0.000	1.000	-0.323
17	85	0.087	0.996	-1.294
18	85	0.087	0.996	-0.580
19	80	0.174	0.985	-0.320
20	80	0.174	0.985	-0.462
upper grade item number				
1	90	0.000	1.000	-0.099
2	90	0.000	1.000	-0.624
3	90	0.000	1.000	-0.721
4	90	0.000	1.000	-0.045
5	90	0.000	1.000	-0.643
6	85	0.087	0.996	-1.149
7	85	0.087	0.996	-0.732
8	85	0.087	0.996	0.277
9	85	0.087	0.996	-0.482
10	85	0.087	0.996	0.012
11	80	0.174	0.985	-0.278
12	80	0.174	0.985	-0.507
13	80	0.174	0.985	-0.360
14	80	0.174	0.985	-0.489
15	80	0.174	0.985	-0.566
16	75	0.259	0.966	0.445
17	75	0.259	0.966	0.542
18	75	0.259	0.966	0.040
19	75	0.259	0.966	-0.459
20	75	0.259	0.966	0.376
21	70	0.342	0.940	0.173
22	70	0.342	0.940	0.369
23	70	0.342	0.940	0.360
24	70	0.342	0.940	-0.871
25	70	0.342	0.940	-0.340
26	65	0.423	0.906	-0.159
27	65	0.423	0.906	0.636
28	65	0.423	0.906	-1.153

upper grade item number	angel with dim 1	a1	a2	d
29	65	0.423	0.906	-0.165
30	65	0.423	0.906	-0.692
31	15	0.966	0.259	-0.116
32	15	0.966	0.259	1.264
33	15	0.966	0.259	0.400
34	15	0.966	0.259	0.407
35	15	0.966	0.259	-0.521
36	0	1.000	0.000	-0.134
37	0	1.000	0.000	-0.315
38	0	1.000	0.000	0.837
39	0	1.000	0.000	0.043
40	0	1.000	0.000	0.441

Table C4: Item parameters for moderate confound and both grades common items test battery

lower grade item number	angle with dim 1	a1	a2	d
1	0	1.000	0.000	0.237
2	0	1.000	0.000	0.798
3	0	1.000	0.000	0.417
4	0	1.000	0.000	-0.742
5	0	1.000	0.000	0.668
6	5	0.996	0.087	0.765
7	5	0.996	0.087	1.262
8	5	0.996	0.087	0.736
9	5	0.996	0.087	0.145
10	5	0.996	0.087	0.545
11	10	0.985	0.174	0.104
12	10	0.985	0.174	0.640
13	10	0.985	0.174	0.344
14	10	0.985	0.174	0.248
15	10	0.985	0.174	1.288
16	15	0.966	0.259	-0.040
17	15	0.966	0.259	0.649
18	15	0.966	0.259	0.267
19	15	0.966	0.259	0.281
20	15	0.966	0.259	0.047
21	20	0.940	0.342	-0.117
22	20	0.940	0.342	0.086
23	20	0.940	0.342	-0.159
24	20	0.940	0.342	0.402
25	20	0.940	0.342	1.122
26	25	0.906	0.423	1.001
27	25	0.906	0.423	-1.072
28	25	0.906	0.423	0.232
29	25	0.906	0.423	0.693
30	25	0.906	0.423	0.090
31	85	0.087	0.996	-0.178
32	85	0.087	0.996	0.072
33	85	0.087	0.996	-0.832
34	85	0.087	0.996	0.429
35	85	0.087	0.996	-0.360
36	90	0.000	1.000	-0.426
37	90	0.000	1.000	-0.588
38	90	0.000	1.000	-0.642
39	90	0.000	1.000	-0.475
40	90	0.000	1.000	-0.120
common item number				
1	0	1.000	0.000	0.826
2	5	0.996	0.087	0.735

common item number	angel with dim 1	a1	a2	d
3	10	0.985	0.174	0.160
4	15	0.966	0.259	0.218
5	20	0.940	0.342	0.809
6	25	0.906	0.423	0.339
7	30	0.866	0.500	0.070
8	90	0.000	1.000	-0.956
9	85	0.087	0.996	-1.294
10	80	0.174	0.985	-0.320
11	90	0.000	1.000	-0.329
12	85	0.087	0.996	-0.809
13	80	0.174	0.985	-0.054
14	75	0.259	0.966	0.075
15	70	0.342	0.940	0.087
16	65	0.423	0.906	0.119
17	60	0.500	0.866	-1.307
18	0	1.000	0.000	-0.010
19	5	0.996	0.087	0.223
20	10	0.985	0.174	1.080
upper grade item number				
1	90	0.000	1.000	-0.099
2	90	0.000	1.000	-0.624
3	90	0.000	1.000	-0.721
4	90	0.000	1.000	-0.045
5	90	0.000	1.000	-0.643
6	85	0.087	0.996	-1.149
7	85	0.087	0.996	-0.732
8	85	0.087	0.996	0.277
9	85	0.087	0.996	-0.482
10	85	0.087	0.996	0.012
11	80	0.174	0.985	-0.278
12	80	0.174	0.985	-0.507
13	80	0.174	0.985	-0.360
14	80	0.174	0.985	-0.489
15	80	0.174	0.985	-0.566
16	75	0.259	0.966	0.445
17	75	0.259	0.966	0.542
18	75	0.259	0.966	0.040
19	75	0.259	0.966	-0.459
20	75	0.259	0.966	0.376
21	70	0.342	0.940	0.173
22	70	0.342	0.940	0.369
23	70	0.342	0.940	0.360
24	70	0.342	0.940	-0.871
25	70	0.342	0.940	-0.340
26	65	0.423	0.906	-0.159
27	65	0.423	0.906	0.636
28	65	0.423	0.906	-1.153

upper grade item number	angel with dim 1	a1	a2	d
29	65	0.423	0.906	-0.165
30	65	0.423	0.906	-0.692
31	15	0.966	0.259	-0.116
32	15	0.966	0.259	1.264
33	15	0.966	0.259	0.400
34	15	0.966	0.259	0.407
35	15	0.966	0.259	-0.521
36	0	1.000	0.000	-0.134
37	0	1.000	0.000	-0.315
38	0	1.000	0.000	0.837
39	0	1.000	0.000	0.043
40	0	1.000	0.000	0.441

Table C5: Item parameters for no confound and lower grade common items test battery

lower grade item number	angle with dim 1	a1	a2	d
1	0	1.000	0.000	1.887
2	0	1.000	0.000	-1.115
3	0	1.000	0.000	0.677
4	0	1.000	0.000	-0.152
5	0	1.000	0.000	-0.908
6	5	0.996	0.087	-0.958
7	5	0.996	0.087	-0.918
8	5	0.996	0.087	-0.908
9	5	0.996	0.087	0.832
10	5	0.996	0.087	1.552
11	10	0.985	0.174	-1.233
12	10	0.985	0.174	-0.012
13	10	0.985	0.174	0.988
14	10	0.985	0.174	1.170
15	10	0.985	0.174	0.535
16	15	0.966	0.259	1.152
17	15	0.966	0.259	-0.464
18	15	0.966	0.259	0.282
19	15	0.966	0.259	1.888
20	15	0.966	0.259	0.503
21	20	0.940	0.342	1.321
22	20	0.940	0.342	0.921
23	20	0.940	0.342	-0.942
24	20	0.940	0.342	-0.342
25	20	0.940	0.342	0.555
26	25	0.906	0.423	-0.752
27	25	0.906	0.423	-0.762
28	25	0.906	0.423	-0.116
29	25	0.906	0.423	-0.189
30	25	0.906	0.423	0.510
31	85	0.087	0.996	-0.281
32	85	0.087	0.996	1.301
33	85	0.087	0.996	0.381
34	85	0.087	0.996	1.544
35	85	0.087	0.996	-0.444
36	90	0.000	1.000	-1.602
37	90	0.000	1.000	0.296
38	90	0.000	1.000	-0.899
39	90	0.000	1.000	0.111
40	90	0.000	1.000	-1.954
common item number				
1	0	1.000	0.000	0.083
2	0	1.000	0.000	-0.819
common item number	angle with dim 1	a1	a2	d
3	5	0.996	0.087	0.152

4	5	0.996	0.087	-1.454
5	10	0.985	0.174	-1.774
6	10	0.985	0.174	-2.971
7	15	0.966	0.259	0.611
8	15	0.966	0.259	-1.154
9	20	0.940	0.342	0.684
10	20	0.940	0.342	-0.550
11	25	0.906	0.423	-0.284
12	25	0.906	0.423	-1.271
13	30	0.866	0.500	-1.330
14	30	0.866	0.500	0.485
15	90	0.000	1.000	-0.723
16	90	0.000	1.000	-0.402
17	85	0.087	0.996	0.360
18	85	0.087	0.996	0.262
19	80	0.174	0.985	-1.486
20	80	0.174	0.985	0.308
upper grade item number				
1	90	0.000	1.000	-0.088
2	90	0.000	1.000	-1.695
3	90	0.000	1.000	1.804
4	90	0.000	1.000	0.360
5	90	0.000	1.000	1.139
6	85	0.087	0.996	0.315
7	85	0.087	0.996	0.838
8	85	0.087	0.996	1.682
9	85	0.087	0.996	-0.520
10	85	0.087	0.996	0.595
11	80	0.174	0.985	-0.397
12	80	0.174	0.985	-0.305
13	80	0.174	0.985	0.603
14	80	0.174	0.985	-0.534
15	80	0.174	0.985	1.403
16	75	0.259	0.966	0.173
17	75	0.259	0.966	-0.011
18	75	0.259	0.966	3.034
19	75	0.259	0.966	2.182
20	75	0.259	0.966	0.759
21	70	0.342	0.940	-1.902
22	70	0.342	0.940	0.084
23	70	0.342	0.940	0.309
24	70	0.342	0.940	0.364
25	70	0.342	0.940	1.208
26	65	0.423	0.906	-0.504
27	65	0.423	0.906	-2.039
28	65	0.423	0.906	-0.903
upper grade item number	angle with dim 1	a1	a2	d
29	65	0.423	0.906	-1.143

30	65	0.423	0.906	-0.951
31	15	0.966	0.259	0.598
32	15	0.966	0.259	-0.818
33	15	0.966	0.259	1.295
34	15	0.966	0.259	1.773
35	15	0.966	0.259	-1.447
36	0	1.000	0.000	-0.760
37	0	1.000	0.000	-0.954
38	0	1.000	0.000	-0.779
39	0	1.000	0.000	-0.042
40	0	1.000	0.000	1.579

Table C6: Item parameters for no confound and both grades common items test battery

lower grade item number	angle with dim 1	a1	a2	d
1	0	1.000	0.000	1.887
2	0	1.000	0.000	-1.115
3	0	1.000	0.000	0.677
4	0	1.000	0.000	-0.152
5	0	1.000	0.000	-0.908
6	5	0.996	0.087	-0.958
7	5	0.996	0.087	-0.918
8	5	0.996	0.087	-0.908
9	5	0.996	0.087	0.832
10	5	0.996	0.087	1.552
11	10	0.985	0.174	-1.233
12	10	0.985	0.174	-0.012
13	10	0.985	0.174	0.988
14	10	0.985	0.174	1.170
15	10	0.985	0.174	0.535
16	15	0.966	0.259	1.152
17	15	0.966	0.259	-0.464
18	15	0.966	0.259	0.282
19	15	0.966	0.259	1.888
20	15	0.966	0.259	0.503
21	20	0.940	0.342	1.321
22	20	0.940	0.342	0.921
23	20	0.940	0.342	-0.942
24	20	0.940	0.342	-0.342
25	20	0.940	0.342	0.555
26	25	0.906	0.423	-0.752
27	25	0.906	0.423	-0.762
28	25	0.906	0.423	-0.116
29	25	0.906	0.423	-0.189
30	25	0.906	0.423	0.510
31	85	0.087	0.996	-0.281
32	85	0.087	0.996	1.301
33	85	0.087	0.996	0.381
34	85	0.087	0.996	1.544
35	85	0.087	0.996	-0.444
36	90	0.000	1.000	-1.602
37	90	0.000	1.000	0.296
38	90	0.000	1.000	-0.899
39	90	0.000	1.000	0.111
40	90	0.000	1.000	-1.954
common item number				
1	0	1.000	0.000	0.083
2	5	0.996	0.087	0.152
3	10	0.985	0.174	-1.774
4	15	0.966	0.259	0.611

common item number	angle with dim 1	a1	a2	d
5	20	0.940	0.342	0.684
6	25	0.906	0.423	-0.284
7	30	0.866	0.500	-1.330
8	90	0.000	1.000	-0.723
9	85	0.087	0.996	0.360
10	80	0.174	0.985	-1.486
11	90	0.000	1.000	0.833
12	85	0.087	0.996	1.245
13	80	0.174	0.985	0.752
14	75	0.259	0.966	-1.355
15	70	0.342	0.940	0.586
16	65	0.423	0.906	2.303
17	60	0.500	0.866	-0.218
18	0	1.000	0.000	-0.690
19	5	0.996	0.087	-0.623
20	10	0.985	0.174	1.386
upper grade item number				
1	90	0.000	1.000	-0.088
2	90	0.000	1.000	-1.695
3	90	0.000	1.000	1.804
4	90	0.000	1.000	0.360
5	90	0.000	1.000	1.139
6	85	0.087	0.996	0.315
7	85	0.087	0.996	0.838
8	85	0.087	0.996	1.682
9	85	0.087	0.996	-0.520
10	85	0.087	0.996	0.595
11	80	0.174	0.985	-0.397
12	80	0.174	0.985	-0.305
13	80	0.174	0.985	0.603
14	80	0.174	0.985	-0.534
15	80	0.174	0.985	1.403
16	75	0.259	0.966	0.173
17	75	0.259	0.966	-0.011
18	75	0.259	0.966	3.034
19	75	0.259	0.966	2.182
20	75	0.259	0.966	0.759
21	70	0.342	0.940	-1.902
22	70	0.342	0.940	0.084
23	70	0.342	0.940	0.309
24	70	0.342	0.940	0.364
25	70	0.342	0.940	1.208
26	65	0.423	0.906	-0.504
27	65	0.423	0.906	-2.039
28	65	0.423	0.906	-0.903
29	65	0.423	0.906	-1.143
30	65	0.423	0.906	-0.951

upper grade item number	angle with dim 1	a1	a2	d
31	15	0.966	0.259	0.598
32	15	0.966	0.259	-0.818
33	15	0.966	0.259	1.295
34	15	0.966	0.259	1.773
35	15	0.966	0.259	-1.447
36	0	1.000	0.000	-0.760
37	0	1.000	0.000	-0.954
38	0	1.000	0.000	-0.779
39	0	1.000	0.000	-0.042
40	0	1.000	0.000	1.579

Appendix D: Multifactor ANOVAs for percent correctly classified, false negative rate, and false positive rate by grade

Table D1: Lower grade multifactor ANOVA on raw data for percent correctly classified (PC)

Tests of Between-Subjects Effects

Dependent Variable: PC

Source	Type III SS	df	Mean Sq	F	Sig.	Partial η^2
Corrected Model	6.625	47	0.141	2861.857	0.000	0.849
Intercept	19598.256	1	19598.256	397875195.133	0.000	1.000
CONFOUND	0.335	2	0.167	3400.314	0.000	0.221
COMMON	0.000	1	0.000	7.705	0.006	0.000
ABILITY	0.002	1	0.002	46.296	0.000	0.002
CORRELAT	6.154	3	2.051	41646.316	0.000	0.839
CONFOUND * COMMON	0.019	2	0.010	197.827	0.000	0.016
CONFOUND * ABILITY	0.014	2	0.007	137.644	0.000	0.011
COMMON * ABILITY	0.000	1	0.000	6.203	0.013	0.000
CONFOUND * COMMON * ABILITY	0.008	2	0.004	83.244	0.000	0.007
CONFOUND * CORRELAT	0.040	6	0.007	136.644	0.000	0.033
COMMON * CORRELAT	0.002	3	0.001	11.789	0.000	0.001
CONFOUND * COMMON * CORRELAT	0.005	6	0.001	17.627	0.000	0.004
ABILITY * CORRELAT	0.005	3	0.002	33.220	0.000	0.004
CONFOUND * ABILITY * CORRELAT	0.017	6	0.003	58.021	0.000	0.014
COMMON * ABILITY * CORRELAT	0.016	3	0.005	105.856	0.000	0.013
CONFOUND * COMMON * ABILITY * CORRELAT	0.007	6	0.001	23.951	0.000	0.006
Error	1.180	23952	0.000			
Total	19606.062	24000				
Corrected Total	7.805	23999				

a. R Squared = .849 (Adjusted R Squared = .849)

Table D2: Lower grade multifactor ANOVA on transformed data for percent correctly classified (PC_T)

Tests of Between-Subjects Effects

Dependent Variable: PC_T

Source	Type III SS	df	Mean Sq	F	Sig.	Partial η^2
Corrected Model	75.220	47	1.600	2843.739	0.000	0.848
Intercept	151552.717	1	151552.717	269288594.932	0.000	1.000
CONFOUND	3.712	2	1.856	3297.857	0.000	0.216
COMMON	0.004	1	0.004	7.276	0.007	0.000
ABILITY	0.022	1	0.022	39.194	0.000	0.002
CORRELAT	70.159	3	23.386	41554.513	0.000	0.839
CONFOUND * COMMON	0.221	2	0.111	196.515	0.000	0.016
CONFOUND * ABILITY	0.139	2	0.069	123.214	0.000	0.010
COMMON * ABILITY	0.004	1	0.004	7.223	0.007	0.000
CONFOUND * COMMON * ABILITY	0.089	2	0.045	79.201	0.000	0.007
CONFOUND * CORRELAT	0.318	6	0.053	94.260	0.000	0.023
COMMON * CORRELAT	0.017	3	0.006	10.289	0.000	0.001
CONFOUND * COMMON * CORRELAT	0.057	6	0.009	16.793	0.000	0.004
ABILITY * CORRELAT	0.048	3	0.016	28.486	0.000	0.004
CONFOUND * ABILITY * CORRELAT	0.190	6	0.032	56.171	0.000	0.014
COMMON * ABILITY * CORRELAT	0.165	3	0.055	97.669	0.000	0.012
CONFOUND * COMMON * ABILITY * CORRELAT	0.074	6	0.012	22.038	0.000	0.005
Error	13.480	23952	0.001			
Total	151641.417	24000				
Corrected Total	88.700	23999				

a. R Squared = .848 (Adjusted R Squared = .848)

Table D3: Lower grade multifactor ANOVA on raw data for the false negative rate (FN)

Tests of Between-Subjects Effects

Dependent Variable: FN

Source	Type III SS	df	Mean Sq	F	Sig.	Partial η^2
Corrected Model	4.613	47	0.098	1079.234	0.000	0.679
Intercept	56.169	1	56.169	617615.738	0.000	0.963
CONFOUND	2.359	2	1.179	12966.908	0.000	0.520
COMMON	0.146	1	0.146	1609.324	0.000	0.063
ABILITY	0.008	1	0.008	83.864	0.000	0.003
CORRELAT	1.047	3	0.349	3835.740	0.000	0.325
CONFOUND * COMMON	0.016	2	0.008	88.897	0.000	0.007
CONFOUND * ABILITY	0.081	2	0.040	443.428	0.000	0.036
COMMON * ABILITY	0.028	1	0.028	311.035	0.000	0.013
CONFOUND * COMMON * ABILITY	0.045	2	0.023	248.151	0.000	0.020
CONFOUND * CORRELAT	0.460	6	0.077	843.694	0.000	0.174
COMMON * CORRELAT	0.019	3	0.006	70.531	0.000	0.009
CONFOUND * COMMON * CORRELAT	0.073	6	0.012	133.679	0.000	0.032
ABILITY * CORRELAT	0.010	3	0.003	34.963	0.000	0.004
CONFOUND * ABILITY * CORRELAT	0.056	6	0.009	103.043	0.000	0.025
COMMON * ABILITY * CORRELAT	0.033	3	0.011	121.574	0.000	0.015
CONFOUND * COMMON * ABILITY * CORRELAT	0.232	6	0.039	425.684	0.000	0.096
Error	2.178	23952	0.000			
Total	62.960	24000				
Corrected Total	6.791	23999				

a. R Squared = .679 (Adjusted R Squared = .679)

Table D4: Lower grade multifactor ANOVA on transformed data for the false negative rate (FN_T)

Tests of Between-Subjects Effects

Dependent Variable: FN_T

Source	Type III SS	df	Mean Sq	F	Sig.	Partial η^2
Corrected Model	89.308	47	1.900	979.020	0.000	0.658
Intercept	4595.231	1	4595.231	2367602.446	0.000	0.990
CONFOUND	46.235	2	23.118	11910.838	0.000	0.499
COMMON	3.490	1	3.490	1798.109	0.000	0.070
ABILITY	0.281	1	0.281	145.020	0.000	0.006
CORRELAT	20.507	3	6.836	3521.958	0.000	0.306
CONFOUND * COMMON	0.473	2	0.236	121.808	0.000	0.010
CONFOUND * ABILITY	1.685	2	0.843	434.181	0.000	0.035
COMMON * ABILITY	0.525	1	0.525	270.265	0.000	0.011
CONFOUND * COMMON * ABILITY	0.797	2	0.399	205.399	0.000	0.017
CONFOUND * CORRELAT	6.972	6	1.162	598.664	0.000	0.130
COMMON * CORRELAT	0.342	3	0.114	58.805	0.000	0.007
CONFOUND * COMMON * CORRELAT	1.472	6	0.245	126.394	0.000	0.031
ABILITY * CORRELAT	0.192	3	0.064	32.916	0.000	0.004
CONFOUND * ABILITY * CORRELAT	1.469	6	0.245	126.172	0.000	0.031
COMMON * ABILITY * CORRELAT	0.479	3	0.160	82.296	0.000	0.010
CONFOUND * COMMON * ABILITY * CORRELAT	4.388	6	0.731	376.799	0.000	0.086
Error	46.488	23952	0.002			
Total	4731.026	24000				
Corrected Total	135.795	23999				

a. R Squared = .658 (Adjusted R Squared = .657)

Table D5: Lower grade multifactor ANOVA on raw data for the false positive rate (FP)

Tests of Between-Subjects Effects

Dependent Variable: FP

Source	Type III SS	df	Mean Sq	F	Sig.	Partial η^2
Corrected Model	4.184	47	0.089	1083.446	0.000	0.680
Intercept	55.219	1	55.219	672017.790	0.000	0.966
CONFOUND	0.944	2	0.472	5745.261	0.000	0.324
COMMON	0.132	1	0.132	1604.416	0.000	0.063
ABILITY	0.018	1	0.018	222.082	0.000	0.009
CORRELAT	2.139	3	0.713	8677.491	0.000	0.521
CONFOUND * COMMON	0.057	2	0.028	346.578	0.000	0.028
CONFOUND * ABILITY	0.083	2	0.041	503.839	0.000	0.040
COMMON * ABILITY	0.023	1	0.023	276.414	0.000	0.011
CONFOUND * COMMON * ABILITY	0.015	2	0.008	93.491	0.000	0.008
CONFOUND * CORRELAT	0.317	6	0.053	643.877	0.000	0.139
COMMON * CORRELAT	0.012	3	0.004	46.809	0.000	0.006
CONFOUND * COMMON * CORRELAT	0.060	6	0.010	121.029	0.000	0.029
ABILITY * CORRELAT	0.028	3	0.009	112.588	0.000	0.014
CONFOUND * ABILITY * CORRELAT	0.125	6	0.021	253.393	0.000	0.060
COMMON * ABILITY * CORRELAT	0.020	3	0.007	81.188	0.000	0.010
CONFOUND * COMMON * ABILITY * CORRELAT	0.212	6	0.035	429.446	0.000	0.097
Error	1.968	23952	0.000			
Total	61.372	24000				
Corrected Total	6.152	23999				

a. R Squared = .680 (Adjusted R Squared = .679)

Table D6: Lower grade multifactor ANOVA on transformed data for the false positive rate (FP_T)

Tests of Between-Subjects Effects

Dependent Variable: FP_T

Source	Type III SS	df	Mean Sq	F	Sig.	Partial η^2
Corrected Model	87.370	47	1.859	1039.102	0.000	0.671
Intercept	4559.084	1	4559.084	2548415.738	0.000	0.991
CONFOUND	20.482	2	10.241	5724.425	0.000	0.323
COMMON	2.674	1	2.674	1494.868	0.000	0.059
ABILITY	0.303	1	0.303	169.593	0.000	0.007
CORRELAT	45.187	3	15.062	8419.461	0.000	0.513
CONFOUND * COMMON	0.904	2	0.452	252.561	0.000	0.021
CONFOUND * ABILITY	1.632	2	0.816	456.179	0.000	0.037
COMMON * ABILITY	0.520	1	0.520	290.657	0.000	0.012
CONFOUND * COMMON * ABILITY	0.572	2	0.286	159.996	0.000	0.013
CONFOUND * CORRELAT	5.349	6	0.892	498.368	0.000	0.111
COMMON * CORRELAT	0.255	3	0.085	47.459	0.000	0.006
CONFOUND * COMMON * CORRELAT	1.168	6	0.195	108.783	0.000	0.027
ABILITY * CORRELAT	0.475	3	0.158	88.495	0.000	0.011
CONFOUND * ABILITY * CORRELAT	2.509	6	0.418	233.758	0.000	0.055
COMMON * ABILITY * CORRELAT	0.465	3	0.155	86.604	0.000	0.011
CONFOUND * COMMON * ABILITY * CORRELAT	4.875	6	0.812	454.137	0.000	0.102
Error	42.850	23952	0.002			
Total	4689.304	24000				
Corrected Total	130.220	23999				

a. R Squared = .671 (Adjusted R Squared = .670)

Table D7: Upper grade multifactor ANOVA on raw data for percent correctly classified (PC)

Tests of Between-Subjects Effects

Dependent Variable: PC

Source	Type III SS	df	Mean Sq	F	Sig.	Partial η^2
Corrected Model	5.197	47	0.111	2502.227	0.000	0.831
Intercept	19663.353	1	19663.353	444945480.558	0.000	1.000
CONFOUND	0.105	2	0.052	1184.995	0.000	0.090
COMMON	0.017	1	0.017	381.808	0.000	0.016
ABILITY	0.001	1	0.001	24.355	0.000	0.001
CORRELAT	5.038	3	1.679	37998.797	0.000	0.826
CONFOUND * COMMON	0.001	2	0.000	5.724	0.003	0.000
CONFOUND * ABILITY	0.003	2	0.001	29.806	0.000	0.002
COMMON * ABILITY	0.001	1	0.001	19.575	0.000	0.001
CONFOUND * COMMON * ABILITY	0.002	2	0.001	17.530	0.000	0.001
CONFOUND * CORRELAT	0.004	6	0.001	14.560	0.000	0.004
COMMON * CORRELAT	0.002	3	0.001	13.459	0.000	0.002
CONFOUND * COMMON * CORRELAT	0.002	6	0.000	5.908	0.000	0.001
ABILITY * CORRELAT	0.014	3	0.005	106.030	0.000	0.013
CONFOUND * ABILITY * CORRELAT	0.005	6	0.001	20.427	0.000	0.005
COMMON * ABILITY * CORRELAT	0.003	3	0.001	23.042	0.000	0.003
CONFOUND * COMMON * ABILITY * CORRELAT	0.001	6	0.000	5.573	0.000	0.001
Error	1.059	23952	0.000			
Total	19669.609	24000				
Corrected Total	6.256	23999				

a. R Squared = .831 (Adjusted R Squared = .830)

Table D8: Upper grade multifactor ANOVA on transformed data for percent correctly classified (PC_T)

Tests of Between-Subjects Effects

Dependent Variable: PC_T

Source	Type III SS	df	Mean Sq	F	Sig.	Partial η^2
Corrected Model	60.194	47	1.281	2471.775	0.000	0.829
Intercept	152116.025	1	152116.025	293579403.020	0.000	1.000
CONFOUND	1.234	2	0.617	1190.853	0.000	0.090
COMMON	0.196	1	0.196	377.870	0.000	0.016
ABILITY	0.008	1	0.008	14.580	0.000	0.001
CORRELAT	58.353	3	19.451	37540.081	0.000	0.825
CONFOUND * COMMON	0.006	2	0.003	5.655	0.004	0.000
CONFOUND * ABILITY	0.029	2	0.015	28.107	0.000	0.002
COMMON * ABILITY	0.008	1	0.008	15.688	0.000	0.001
CONFOUND * COMMON * ABILITY	0.017	2	0.009	16.470	0.000	0.001
CONFOUND * CORRELAT	0.033	6	0.005	10.573	0.000	0.003
COMMON * CORRELAT	0.021	3	0.007	13.264	0.000	0.002
CONFOUND * COMMON * CORRELAT	0.019	6	0.003	6.135	0.000	0.002
ABILITY * CORRELAT	0.160	3	0.053	102.814	0.000	0.013
CONFOUND * ABILITY * CORRELAT	0.060	6	0.010	19.398	0.000	0.005
COMMON * ABILITY * CORRELAT	0.036	3	0.012	22.900	0.000	0.003
CONFOUND * COMMON * ABILITY * CORRELAT	0.015	6	0.003	4.881	0.000	0.001
Error	12.411	23952	0.001			
Total	152188.630	24000				
Corrected Total	72.605	23999				

a. R Squared = .829 (Adjusted R Squared = .829)

Table D9: Upper grade multifactor ANOVA on raw data for the false negative rate (FN)

Tests of Between-Subjects Effects

Dependent Variable: FN

Source	Type III SS	df	Mean Sq	F	Sig.	Partial η^2
Corrected Model	0.997	47	0.021	260.097	0.000	0.338
Intercept	53.984	1	53.984	661877.819	0.000	0.965
CONFOUND	0.006	2	0.003	38.171	0.000	0.003
COMMON	0.043	1	0.043	533.273	0.000	0.022
ABILITY	0.027	1	0.027	329.665	0.000	0.014
CORRELAT	0.456	3	0.152	1861.958	0.000	0.189
CONFOUND * COMMON	0.040	2	0.020	242.883	0.000	0.020
CONFOUND * ABILITY	0.075	2	0.038	459.791	0.000	0.037
COMMON * ABILITY	0.004	1	0.004	45.795	0.000	0.002
CONFOUND * COMMON * ABILITY	0.008	2	0.004	46.694	0.000	0.004
CONFOUND * CORRELAT	0.059	6	0.010	121.302	0.000	0.029
COMMON * CORRELAT	0.015	3	0.005	61.878	0.000	0.008
CONFOUND * COMMON * CORRELAT	0.037	6	0.006	75.575	0.000	0.019
ABILITY * CORRELAT	0.085	3	0.028	345.813	0.000	0.042
CONFOUND * ABILITY * CORRELAT	0.053	6	0.009	108.404	0.000	0.026
COMMON * ABILITY * CORRELAT	0.036	3	0.012	145.446	0.000	0.018
CONFOUND * COMMON * ABILITY * CORRELAT	0.054	6	0.009	110.628	0.000	0.027
Error	1.954	23952	0.000			
Total	56.934	24000				
Corrected Total	2.951	23999				

a. R Squared = .338 (Adjusted R Squared = .337)

Table D10: Upper grade multifactor ANOVA on transformed data for the false negative rate (FN_T)

Tests of Between-Subjects Effects

Dependent Variable: FN_T

Source	Type III SS	df	Mean Sq	F	Sig.	Partial η^2
Corrected Model	22.202	47	0.472	262.255	0.000	0.340
Intercept	4565.773	1	4565.773	2534753.249	0.000	0.991
CONFOUND	0.139	2	0.070	38.711	0.000	0.003
COMMON	0.995	1	0.995	552.665	0.000	0.023
ABILITY	0.640	1	0.640	355.310	0.000	0.015
CORRELAT	10.270	3	3.423	1900.453	0.000	0.192
CONFOUND * COMMON	0.852	2	0.426	236.589	0.000	0.019
CONFOUND * ABILITY	1.475	2	0.738	409.571	0.000	0.033
COMMON * ABILITY	0.105	1	0.105	58.532	0.000	0.002
CONFOUND * COMMON * ABILITY	0.148	2	0.074	40.978	0.000	0.003
CONFOUND * CORRELAT	1.388	6	0.231	128.401	0.000	0.031
COMMON * CORRELAT	0.356	3	0.119	65.856	0.000	0.008
CONFOUND * COMMON * CORRELAT	0.763	6	0.127	70.559	0.000	0.017
ABILITY * CORRELAT	2.042	3	0.681	377.863	0.000	0.045
CONFOUND * ABILITY * CORRELAT	1.087	6	0.181	100.606	0.000	0.025
COMMON * ABILITY * CORRELAT	0.734	3	0.245	135.914	0.000	0.017
CONFOUND * COMMON * ABILITY * CORRELAT	1.207	6	0.201	111.686	0.000	0.027
Error	43.144	23952	0.002			
Total	4631.120	24000				
Corrected Total	65.346	23999				

a. R Squared = .340 (Adjusted R Squared = .338)

Table D11: Upper grade multifactor ANOVA on raw data for the false positive rate (FP)

Tests of Between-Subjects Effects

Dependent Variable: FP

Source	Type III SS	df	Mean Sq	F	Sig.	Partial η^2
Corrected Model	3.213	47	0.068	963.787	0.000	0.654
Intercept	53.962	1	53.962	760730.122	0.000	0.969
CONFOUND	0.070	2	0.035	491.866	0.000	0.039
COMMON	0.006	1	0.006	87.220	0.000	0.004
ABILITY	0.039	1	0.039	545.906	0.000	0.022
CORRELAT	2.683	3	0.894	12606.914	0.000	0.612
CONFOUND * COMMON	0.044	2	0.022	312.688	0.000	0.025
CONFOUND * ABILITY	0.061	2	0.030	428.561	0.000	0.035
COMMON * ABILITY	0.008	1	0.008	115.533	0.000	0.005
CONFOUND * COMMON * ABILITY	0.011	2	0.006	78.284	0.000	0.006
CONFOUND * CORRELAT	0.070	6	0.012	163.603	0.000	0.039
COMMON * CORRELAT	0.022	3	0.007	102.220	0.000	0.013
CONFOUND * COMMON * CORRELAT	0.035	6	0.006	81.699	0.000	0.020
ABILITY * CORRELAT	0.031	3	0.010	143.593	0.000	0.018
CONFOUND * ABILITY * CORRELAT	0.050	6	0.008	116.665	0.000	0.028
COMMON * ABILITY * CORRELAT	0.027	3	0.009	125.406	0.000	0.015
CONFOUND * COMMON * ABILITY * CORRELAT	0.058	6	0.010	136.726	0.000	0.033
Error	1.699	23952	0.000			
Total	58.874	24000				
Corrected Total	4.912	23999				

a. R Squared = .654 (Adjusted R Squared = .653)

Table D12: Upper grade multifactor ANOVA on transformed data for the false positive rate (FP_T)

Tests of Between-Subjects Effects

Dependent Variable: FP_T

Source	Type III SS	df	Mean Sq	F	Sig.	Partial η^2
Corrected Model	69.216	47	1.473	931.300	0.000	0.646
Intercept	4525.926	1	4525.926	2862139.472	0.000	0.992
CONFOUND	1.338	2	0.669	422.918	0.000	0.034
COMMON	0.210	1	0.210	132.492	0.000	0.006
ABILITY	1.045	1	1.045	660.674	0.000	0.027
CORRELAT	57.766	3	19.255	12176.733	0.000	0.604
CONFOUND * COMMON	0.959	2	0.479	303.196	0.000	0.025
CONFOUND * ABILITY	1.296	2	0.648	409.720	0.000	0.033
COMMON * ABILITY	0.224	1	0.224	141.643	0.000	0.006
CONFOUND * COMMON * ABILITY	0.245	2	0.123	77.505	0.000	0.006
CONFOUND * CORRELAT	1.356	6	0.226	142.916	0.000	0.035
COMMON * CORRELAT	0.611	3	0.204	128.735	0.000	0.016
CONFOUND * COMMON * CORRELAT	0.739	6	0.123	77.851	0.000	0.019
ABILITY * CORRELAT	0.800	3	0.267	168.653	0.000	0.021
CONFOUND * ABILITY * CORRELAT	0.921	6	0.153	97.057	0.000	0.024
COMMON * ABILITY * CORRELAT	0.601	3	0.200	126.586	0.000	0.016
CONFOUND * COMMON * ABILITY * CORRELAT	1.108	6	0.185	116.759	0.000	0.028
Error	37.876	23952	0.002			
Total	4633.017	24000				
Corrected Total	107.091	23999				

a. R Squared = .646 (Adjusted R Squared = .646)

References

Abedi, J. (1994). *NAEP TRP Task 3e: Achievement Dimensionality, Section A*. Los Angelas, CA: National Center for Research on Evaluation, Student Testing, and Standards.

Ackerman, T. A. (1989). Unidimensional IRT calibration of compensatory and noncompensatory multidimensional items. *Applied Psychological Measurement, 13*(2), 113-127.

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

Ackerman, T. A., Gierl, M. J., & Walker, C. M. (2003). Using multidimensional item response theory to evaluate educational and psychological tests. *Educational Measurement: Issues and Practice, 22*(3), 37-53.

Adams, R. J., Wilson, M. R., & Wang, W. C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*, 1-24.

Ansley, T. N., & Forsyth, R. A. (1985). An examination of the characteristics of unidimensional IRT parameter estimates derived from two-dimensional data. *Applied Psychological Measurement, 9*, 37-48.

Betebenner, D. W., Shang, Y., Xiang, Y., Zhao, Y., & Yue, X. (2008). The impact of performance level misclassification on the accuracy and precision of percent at performance level measures. *Journal of Educational Measurement, 45*, 119-138.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical Theories of Mental Test Scores* (pp. 397-472). Reading, MA: Addison-Wesley.

Briggs, D. C., & Wilson, M. (2003). An introduction to multidimensional measurement using Rasch models. *Journal of Applied Measurement, 4*, 87-100.

Bock, D. R. (1983). The mental growth curve reexamined. In D. J. Weiss (Ed.), *New Horizons In Testing* (pp. 205-218). New York: Academic Press.

Bock, R. D. & Zimowski, M. F. (1997). Multiple group IRT. In R. K. Hambleton & W.J. van der Linden (Eds.), Handbook of Modern Item Response Theory (pp. 433-448). New York: Springer-Verlang.

Camilli, G., & Wang, M.-m., & Fresq, J. (1995). Effects of dimensionality on equating the Law School Admission Test. *Journal of Educational Measurement, 32*, 79.

Cao, Y. (2008). Mixed-format test equating: Effects of dimensionality and common item sets. Unpublished Doctoral Dissertation. University of Maryland.

Douglas, K. M. (2007). A general method for estimating the classification reliability of complex decisions based on configural combinations of multiple assessment scores. Unpublished Doctoral Dissertation. University of Maryland.

Drasgow, F., & Parsons, C. K. (1983). Application of unidimensional item response theory models to multidimensional data. *Applied Psychological Measurement,* 7, 189-199.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, New Jersey: Lawrence Erlbaum.

Fraser, C. (1988). NOHARM [Computer program]. Armidale, New South Wales, Australia.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer Nijhoff Publishing.

Harris, D. J. (2007). Practical issues in vertical scaling. In N. J. Dorans, M. Pommerich & P. W. Holland (Eds.), *Linking and aligning scores and scales.* (pp. 233-251). New York: Springer

Harris, D. J., & Hoover, H. D. (1987). An Application of the three-parameter IRT model to vertical equating. *Applied Psychological Measurement, 11*, 151-159.

Herbert, J. P., & Hauser, R. M. (Eds.). (1999). *High stakes: Testing for tracking, promotion, and graduation*: National Research Council.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: methods and practices* (2nd ed.). New York: Springer.

Lau, A. (1996). Robustness of a unidimensional computerized mastery testing procedure with multidimensional testing data. Unpublished Doctoral Dissertation. University of Iowa.

Li, Y. H., & Lissitz, R. W. (2000). An evaluation of the accuracy of multidimensional IRT linking. *Applied Psychological Measurement, 24*, 115-138.

Lin, P. (2009). IRT versus factor analysis approaches in analyzing multigroup multidimensional binary data: The effect of structural orthogonality and the equivalence in test structure, item difficulty, and examinee groups. Unpublished Doctoral Dissertation. University of Maryland.

Linn, R. (1989). *Has item response theory increased the validity of achievement test scores?* (No. 302): UCLA Center for Research on Evaluation, Standards, and Student Testing.

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement, 17*, 179.

McBride, J., & Wise, L. (2000). *Developing a vertical scale for the Florida comprehensive assessment test (FCAT)*. A HummRRO report under subcontract to Harcourt Assessment, San Antonio, TX.

McCall, M. (2007). *Vertical scaling and the development of skills*. Paper presented at the Washington State Assessment Conference, Seattle, WA.

Mignani, S., Monari, P., Cagnone, S., & Ricci, R. (2006). Multidimensional versus unidimensional models for ability testing. In S. Zani, A. Cerioli, M. Riani & M. Vichi (Eds.), *Data analysis, classification, and the forward search* (pp. 339-348). New York: Springer.

Mislevy, R. J., & Bock, R. D. (1982). Bilog, maximum likelihood item analysis and test scoring: Logistic model [Computer software]. Mooresville, IN: Scientific Software, Inc.

No Child Left Behind Act of 2001. Public Law No. 107-110, 115 Stat. 1425.

Paris, S. G. (2005). Reinterpreting the development of reading skills. *Reading Research Quarterly, 40*, 184-202.

Patz, R. J., & Yao, L. (2007). Methods and models for vertical scaling. In N. J. Dorans, M. Pommerich & P. W. Holland (Eds.), *Linking and aligning scores and scales.* (pp. 253-272). New York: Springer.

Pierce, C. A., Block, R. A., & Aguinis, H. (2004). Cautionary note on reporting eta-squared values from multifactor ANOVA designs. *Educational & Psychological Measurement, 64*, 916-924.

Program for International Student Assessment Technical Report. (2003).). Paris: Organization for Economic Co-operation and Development.

Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement, 9*, 401-412.

Reckase, M. D. (1990). *Unidimensional Data from Multidimensional Tests and Multidimensional Data from Unidimensional Tests*. Paper presented at the annual meeting of the American Educational Research Association, Boston, MA.

Reckase, M. D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement, 21*, 25-36.

Reckase, M. D. (2009). *Multidimensional item response theory*. New York: Springer.

Reckase, M. D., Carlson, J. E., Ackerman, T. A., & Spray, J. A. (1986). *The interpretation of unidimensional IRT parameters when estimated from multidimensional data*. Paper presented at the annual meeting of the Psychometric Society, Toronto.

Reckase, M. D., & Li, Y. H. (2007). Estimating change in achievement when content specifications change: A Multidimensional Item Response Theory Approach. In R. W. Lissitz (Ed.), *Assessing and Modeling Cognitive Development in School* (pp. 189-204). Maple Grove, MN: JAM Press.

Reckase, M. D., & Martineau, J. A. (2004). *The vertical scaling of science achievement tests.* Unpublished Report, Michigan State University, East Lansing, MI.

Sheskin, D. J. (2007). *Handbook of Parametric and Nonparametric Statistical Procedures*. Boca Raton: Chapman & Hall/CRC.

Skaggs, G., & Lissitz, R. W. (1986). IRT test equating: relevant issues and a review of recent research. *Review of Educational Research, 56*(4), 495-529.

Skaggs, G., & Lissitz, R. W. (1988). Effect of examinee ability on test equating invariance. *Applied Psychological Measurement, 12*, 69-82.

Sloane, K., Wilson, M., & Samson, S. (1996). *Designing an embedded assessment system: From principles to practice*: University of California, Berkley.

Stevens, J. C. (1992). *Applied multivariate statistics for the social sciences*. Hillside, NJ: Lawrence Erlbaum.

Stout, W. (2007). Skills diagnosis using IRT-based continuous latent trait models. *Journal of Educational Measurement, 44*, 313-324.

Sympson, J. B. (1978). *A model for testing with multidimensional items.* Paper presented at the Computerized Adaptive Testing Conference, Minnesota.

Tong, Y., & Kolen, M. J. (2007). Comparisons of methodologies and results in vertical scaling for educational achievement tests. *Applied Measurement in Education, 20*, 227.

Walker, C. M., & Beretvas, S. N. (2003). Comparing multidimensional and unidimensional proficiency classifications: multidimensional IRT as a diagnostic aid. *Journal of Educational Measurement, 40*, 255-275.

Wang, W.-C. (1994). *Implementation and application of the multidimensional random coefficient logit model.* Unpublished doctoral dissertation, University of California, Berkley.

Way, W. D., Ansley, T. N., & Forsyth, R. A. (1988). The comparative effects of compensatory and noncompensatory two-dimensional data on unidimensional IRT estimates. *Applied Psychological Measurement, 12*, 239-252.

Williams V. S., Pommerich M., & Thissen D. (1998). A comparison of developmental scales based on Thurstone methods and item response theory. *Journal of Educational Measurement, 35*, 93-107.

Wingersky, M. S., Barton, M. A., & Lord, F. M. (1982). LOGIST [Computer program]. Princeton, NJ: Educational Testing Service.

Wisconsin Knowledge and Concept Examinations-CRT: December 2004 Field Test/Standardization Technical Report. (2005).). Monterey, CA: CTB McGraw Hill.

Yao, L. (2003). BMIRT: Bayesian multivariate item response theory [Computer software]. Monterey, CA: CTB/McGraw-Hill.

Yon, H. (2006). *Multidimensional item response theory (MIRT) approaches to vertical scaling.* Unpublished doctoral dissertation, Michigan State University, East Lansing, MI.

Young, M. J. (2006). Vertical Scales. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development.* (pp. 469-485). Mahwah, NJ, US: Lawrence Erlbaum Associates

Zieky, M. J. (2006). So much has changed: How the setting of cutscores has evolved since the 1980s. In G. Cizek (Ed.). *Setting performance standards: Concepts, methods, and perspectives* (pp. 53-88). Mahwah, NJ: Erlbaum.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). BILOG-MG: Multiple-group IRT analysis and test maintenance for binary items [Computer software]. Chicago: Scientific Software International.