An Examination of the Magnitude of Residual Covariance for Complex Performance Assessments Under Various Scoring and Scaling Methods

Technical Report
August 4, 2009
Number W0901

Joshua T. Goodman

Richard M. Luecht

University of North Carolina at Greensboro

Yanwei O. Zhang

American Institute of Certified Public Accountants

AICPA®
American Institute of Certified Public Accountants

Abstract

Large-scale assessment programs are increasingly including complex performance assessments along with traditional multiple-choice items in a given test. These performance assessments are developed in part to measure sets of skills that are part of the trait to be measured, but are not easily assessed with multiple choice items. One approach to creating and score these items is to create a set of tasks within a scenario that can be objectively scored using set of scoring rules to yield dichotomous responses. Including complex performance items introduces two potential problems. First, the performance items are developed to measure something distinctly different from the multiple-choice items, and may introduce some degree of multidimensionality into the test. Second, as the set of scoring rules or “measurement opportunities” stem from common stimuli and are scored with a set of elaborate rule, contextual and scoring dependencies are likely to arise. Both multidimensionality and statistical dependencies may create a situation where non-zero residual covariances are present. In turn, non-zero residual covariances—which violate tenets of both classical test and item response theories—can have potentially serious psychometric impact.

This study uses real test data to explore the nature and magnitudes of the conditional and residual covariance structures between and among the measurement opportunities of complex performance assessments. Residual covariances are assessed under three different methods for scoring and scaling the simulations. The results indicate that polytomously scoring the measurement opportunities in may be effective in controlling some of the extreme dependencies due to scoring or contextual factors, but changing only the scoring methods does little to reduce the overall amount of conditional covariance among the measurement opportunities. Treating the performance items and multiple choice items as two separate and distinct scales, after polytomous scoring was applied proved most effective in control the magnitude of residual covariance.

Introduction

This research explores options for scoring and scaling the complex performance assessments items in place as part of a large-scale certification examination. This exam consists of four sections, each delivered in a computerized adaptive multistage test (ca-MST; Luecht & Nungester, 1998; Luecht, 2000; Luecht, Brumfield, and Breithaupt, 2006) design[1]. This design uses an adaptive algorithm to match, sequentially, the difficulty of testlets to the apparent proficiency of the examinees. The collection of ca-MST testlets available for adaptive administration to any given examinee is called a "panel". Panels for all four sections include multiple-choice questions (MCQs), and three of the four sections include two “simulations” items—performance assessments that are designed to assess the on the skills (e.g. analysis, research, communication skills, etc.) required to be successful in the workplace. The intent of

[1] The ca-MST structure is only used for the MCQ section. Regardless of ca-MST pathway, candidates receiving the same set of testlets receive the same two simulation items

adding such task is to improve test validity by measuring skills that are required for success in the field.

Boolean rules are applied to examinees' responses as a means of scoring simulations. For example, a Boolean rule may relate a free-response entry in a spreadsheet cell to a stored answer, requiring that the entry be within a prescribed tolerance of the anticipated answer. Each simulation has multiple possible responses opportunities termed "measurement opportunities" (MO). The current approach to scoring the MOs is to treat each as an independent "item". The MCQ items are calibrated using the three parameter logistic model, and the resulting item parameters then serve as a set of anchoring parameters for calibrating the MCQ and simulation MOs using the same unidimensional IRT model.

Using a unidimensional IRT model for the purposes of calibrating the MCQ responses and the simulation MOs poses two potential challenges. First, the simulations, explicitly seeking to measure procedural and problem-solving skills that go beyond the declarative knowledge assessed by the MCQs, raises the possibility that the scored MOs will in fact constitutes a different scale than measured by the longer and more stable MCQ section. Second, because multiple MOs are associated with a common scenario and because scoring rules may share "scorable objects" and further allow explicit relations among the rules, it is likely that many complicated statistical dependencies are present. While the simulations are initially added to improve the relationship of the test to the underlying construct, the local item dependency and dimensionality that may be introduced call into the question the appropriateness of the scores produced. Significant item dependency and dimensionality have practical ramifications as well; directly affecting the estimation of test reliability, estimation of item and ability parameters, estimation of test information, equating processes, and DIF detection.

Defining Local Item Dependency

Local independence of item responses is a notion that appears in numerous forms in test theory. In classical test theory, it is assumed that the errors of measurement are uncorrelated given the true score of an examinee. In item response theory (IRT), a pair of items is considered locally independent in item response theory if, after conditioning on the examinee's proficiency on the measured trait, the joint probability distribution of any two items is equal to the product of the marginal probability distributions (see equation 1). A set of item responses is locally independent when the joint distribution of all item responses is the product of each individual item's response distribution.

$$P\left(X_1 = x_1, X_2 = x_2 \middle| \theta\right) = P\left(X_1 = x_1 \middle| \theta\right) P\left(X_2 = x_2 \middle| \theta\right) \qquad (1)$$

If Equation 1 holds, the trait proficiency (θ) accounts for all of the information relevant for each examinee, thus allowing the items to be evaluated independently (Yen 1984). This is a fundamental assumption in using any unidimensional IRT model. This idea is easily expressed in terms of conditional covariance as well. If items X_i and X_j are locally independent they will have a covariance of zero, after conditioning on some ability θ.

A definition of local item dependence (LID) follows logically from the definition above—the presence of conditional covariance between a set of items. That is,

$$\mathrm{Cov}(X_i,X_j|\theta) = \mathrm{E}(x_i,x_j|\theta) \neq 0, \text{ for } i \neq j. \qquad (2)$$

Conditional covariance in either a positive or negative direction indicates that performance on one item is related to the expected performance of the examinees on the other item. This can also be put in factor analytic terms: LID is present if, after extracting the first factor (roughly equivalent to conditioning on the trait of primary interest θ), there is a non-zero residual covariance between some items. These non-zero covariance indicate that there may be one or more additional factors that explain the remaining variance (Yen 1993). The additional factors are the potential sources of LID and may or may not be vital to the trait or behavior that is being measured.

Sources of Local Item Dependencies

For the purposes of this study the sources of LID are classified into three categories: *contextual*, *scorin*g, and *dimensional*. Items that share contextual circumstances may be related to one another in ways that the primary ability or proficiency of interest cannot be explained. Many dependency studies focus on passages or performance assessments that have an associated cluster of items, resulting in responses showing varying levels of dependencies (Sireci et al, 1991; Yen, 1993; Ferrara et al, 1997; Yan, 1997; Ferrara, Huynh, & Michaels 1999; Zenisky, Hambleton, & Sireci, 2002). This same thinking can be extended to items that share a common setting, stimuli, set of directions, or set of resources.

That choice item-level scoring procedure, particularly on performance items, can also lead to LID between items. Scoring rules that share objects, (i.e. awarding credit in more than one place for a correct response on a particular item), can also lead to dependencies due to scoring. Similarly, the presence of item chains—a series of items where answers depend directly on the responses given previously (i.e. an item requires an explanation of the previous response or the various steps in solving problem are each graded separately). Both scoring and contextual dependencies can be considered to be "nuisance factors"—that is, they are factors that do not have direct implications on the trait or skill being measured, but can have adverse effects on the properties of the scores or equating if ignored.

If more than one skill or trait is required to successfully explain an examinee's response, a test can be considered to have some degree of multidimensionality. LID is an indicator that multiple proficiency traits may be underlying the collective response patterns for a set of items on a test. If the relative magnitude of the multidimensionality is large, the residual covariance cannot be considered as ignorably due to nuisance factors. Ultimately, additional dimensionally relevant score scales may be needed, if the practitioner decides that the skills they represent are essential to the measurement purposes of the test.

Consequences of Ignoring Local Item Dependencies

Ignoring LID, regardless of its cause, is known to affect the psychometric properties of tests. For example, if we assume that a unidimensional IRT model fits the data, but then encounter response patterns that violate the assumption that items are conditionally independent, we will tend to overestimate properties such as the test information and reliability while underestimating the standard errors of the ability estimates (Sireci et al, 1991; Wainer & Thissen 1996, Yen 1993). Reese (1995) describes the underestimation of low score and the overestimation of high scores in sets of items that exhibit LID. Further, the presence of LID is known to affect the estimation and accuracy of item parameters. Wainer and Wang (2002) found that lower asymptotes were overestimated when dependencies were ignored between testlets. Ackerman (1987) found that item discriminations were overestimated when set of items were locally dependent. When items are to be banked for use in creating parallel forms or to be delivered adaptively, inaccurate item parameter estimates can call the fairness of the test into question (Thompson and Pommerich,1996). If residual covariances differ for various population subgroups, differential item functioning (DIF) results. Finally, test scaling and equating practices—which rely on accurate parameter estimates—can be adversely affected by LID (Reese and Pashley, 1999; DeChamplain, 1996).

Strategies for Addressing Local Item Dependencies

In situations where LID is present, or likely to be present, due to contextual and/or scoring concerns, certain courses of action are advisable to reduce the effects and magnitude of LID. The most common solution is to form “testlets” from the sets of related items—effectively creating a single “super” polytomous item from the cluster by summing the individual scored objects. The resulting testlet-based super item can then be scaled using an IRT model for polytomous data. Polytomous scoring of testlets has been demonstrated as effective in reducing LID (Yen, 1993; Stark, Chernyshenko, & Drasgow, 2004; Sireci et al. 1991; Zenisky, Hambleton, & Sireci, 2002). However creating polytomous items from unrelated subsets of items has been shown to decrease reliability and test information (Yen, 1993). If a test has several related sets of items, (i.e. several reading passages with related clusters), then this method is most effective if the created polytomous items can be created so that local independence is maintained across all the newly created set of polytomous items.

A course of action is less clear when a test exhibits apparent multidimensionality. The simplest and perhaps most common practice is to continue to assume that the mixture of multiple dimensions forms an essentially unidimensional measure. In that case, the resulting total-test ability estimate can be shown to represent a weighted composite of some unknown number of traits. If the extent of multidimensionality is small and unrelated to specific features of the items or content of the test, this solution may be reasonable. The composite ability estimate is effectively weighted according to the relative number of items linked to each trait and the average information exhibited by those items. However, as the magnitude of multidimensionality increases, the projection of any ancillary dimensions onto a single reference composite can alter the nature of the total-test composite in unpredictable ways.

Another approach to dealing with multidimensionality can also be employed: scaling the sets of items assumed to represent different traits or skills sets separately. This approach allows separate score for each scale/dimension to be reported and, ideally, ability estimates that adequately explain the responses to related set of items. Again, this course of action is not without practical consequences. Breaking the complete set of test items into separate tests will results in smaller tests. Smaller tests, in turn, yield less reliable scores, and less reliable scores may adversely affect the quality of the ability estimates, themselves. Statistical augmentation can be used to improve the reliability of the multidimensional estimates, but not without regression bias to (e.g., Finally, creating an appropriate and stable total-test, composite score (if one is needed) can become a tedious procedure from an equating perspective.

More complex scaling methods are also available when a test is likely multidimensional. A great number of multidimensional IRT (MIRT) models allow for multiple abilities to be jointly estimated, describe the relationship between set of traits, and allow for factorial complex structures within the test. These scaling methods are more computationally complex, require much larger sample sizes, and software packages to fit the models to data tend to be limited. Technical statistical issues such as rotational indeterminacy also remain largely unresolved for MIRT models.

Lastly, dealing with LID is complicated by an intractable confounding of the interactions between subsets of examinees from the population and subtle or not-so-subtle characteristics of the assessment tasks. Simple statistical methods can help test for the presence of LID. However, conclusively determining the source or sources of LID is not straightforward. For example, it is becoming more common to include complex performance assessment on many types of tests, ranging from certification examinations to educational assessments. These items are typically developed to measure analytical, problem-solving, evaluative and communication skills that are difficult to measure in tradition MCQ questions. Whether these skills align with content to form a ordered continuum is open to empirical investigation. Members of the populations may also differ in their opportunities to learn these new skills. Once items are written for a particular performance assessment, there are additional interactions that may involve prior knowledge for some groups with common contexts or stimuli used to link the items or MOs together. In addition, the items may be scored with complex scoring rules having various implicit or explicit associations among the components used in scoring and those features of how those components are learned may differ for various population subgroups. In short, our capability to detect LID probably surpasses our ability to understand it.

Research Questions

The current study examines the effects that scoring methods and scaling procedure have on the residual covariance structures (indicators of the statistical dependencies due to context/scoring and/or dimensionality) of certification test that includes simulations-based performance assessments. Three different sets of conditional covariance will be compared; the relationships of MOs *within* a single simulations, relationships between MOs on the two different simulation within a single panel, and the relationships between the simulation MOs and the MCQ items. The residual covariances of item pairs are assessed in terms of their magnitude and their direction. In any of the three comparisons above, significant negative (divergent) covariance provides evidence that the items may be measuring different traits. Positive (convergent) covariance between two items may indicate that contextual or scoring associations are leading to dependencies (this is only possible on within simulation comparisons as only those items share scoring rules and contextual settings). Positive covariance is also evidence that the item pair may be associated with some other trait not represented in the assessment.

As mentioned above, clustering related items into sets and forming polytomous items from these sets is one potential way to control the amount of LID due to a common context or scoring rules. In the context of this study, the MOs within a single simulation could be scored in their dichotomous state or grouped to form “super” polytomous items. The distribution of conditional covariance among the MOs within a simulation could then be compared across the different scoring methods. Specifically, the study will compare three methods of scoring the simulations: 1) leaving simulations as dichotomous simulations, 2) grouping the MOs within a single simulation into a smaller number of polytomous items using statistically optimal method, and 3) creating polytomous simulations based the cognitive properties of the MOs. This leads to two separate research questions.

1. Does breaking the simulations MOs into one or more polytomous items reduce or control the amounts of residual covariance between item pairs?

2. Of the two method for creating polytomous simulation items, which controls the amount of residual covariance better?

If conditional covariance is indicative of the presence of more than one scale, the scoring procedure mentioned above will not adequately mitigate the observed LID. Scaling the simulations in using various methods will produce different ability estimates. These estimates can then be used as the conditioning ability in a conditional covariance structures. In this study, three IRT scaling methods (two unidimensional methods and one multidimensional approach) were fully crossed with the scoring MO-level methods. IRT-based ability scores were estimated under each method and, in turn, were used as conditioning variables in the computation of the residual covariances. The distributions and magnitudes of the resulting residual covariances were then compared across conditions. This leads to a final research question:

3. Which of the combinations of scaling methods is most successful in producing conditional covariance structures that meet the assumptions of local independence and unidimensionality used?

Methods

<u>Sample</u>

This study uses real candidate responses from the three content-specific sections of a large-scale, professional certification assessment that contains both multiple-choice and complex performance simulation items. From each these sections, four computer-adaptive multistage test (CA-MST) panels were selected at random. This provided a means of empirically replicating the results across multiple test forms. Table 1 shows the number of candidates who took each of the four sampled panels, the total multiple-choice questions (MCQs) on each panel, and the number dichotomous measurement opportunities (MOs) associated with each of the simulations.

Table 1: Description of the sampled panels

							Sim 1	*Sim 2*	*Total*			
				Sim 1	*Sim 2*	*Total*	*Poly-*	*Poly-*	*Poly-*	*Sim1*	*Sim2*	*Total*
	Panel	*N*	*MCQ*	*MOs*	*MOs*	*MOs*	*Stat*	*Stat*	*Stat*	*TBS*	*TBS*	*TBS*
Section 1	1	316	125	28	24	52	7	7	14	4	4	8
	2	286	125	22	24	46	5	7	12	4	4	8
	3	278	125	24	21	45	6	4	10	4	4	8
	4	322	125	22	24	46	7	5	12	4	4	8
Section 2	1	310	125	19	15	34	3	4	7	4	3	7
	2	323	125	23	17	40	4	4	8	4	3	7
	3	314	124	13	17	30	2	2	4	2	3	5
	4	307	125	23	17	40	3	5	8	4	3	7
Section 3	1	321	100	36	21	57	6	4	10	5	3	8
	2	307	100	22	26	48	6	7	13	3	5	8
	3	274	100	26	23	49	5	4	9	3	3	6
	4	302	100	40	24	64	15	5	20	5	3	8

The MOs for each simulation were scored in three ways: (1) treating each MO as unique “items”; (2) collapsing the MOs into “super” items and scoring each super item polytomously, based on statistical criteria for clustering the MOs together; and (3) combining the MOs into a testlet based on association of the MOs under a common user-interface control known as a “tab”. Additional details are provided below.

Creation of the Polytomous Items

One purpose of the current research is to examine how scoring polytomously might affect the magnitude of conditional residual covariances that are indicative of dependencies and/or dimensionality. For all of the selected panels from, the dichotomous MOs are grouped into polytomous scores using two different methods: (1) a statistically optimal grouping and (2) “tab-based” simulations (TBS). The term “tab-based” simulation refers to the way that the simulation items are presented to examinees. Simulations are presented on single screen with a set number of clickable tabs across the top. For instance, selecting the first tab might present the simulation scenario, the next may contain general directions, and other tabs would contain sets of tasks. Currently, the MOs contained within a tab may or may not be linked to the same requisite set of skills. In that regard, the tabs are viewed as randomly sampling over the mode of presentation. Under TBS scoring, however, the MOs associated with a particular tab ARE logically hypothesized to measure a single skill with associated content and hence should form a natural way of grouping the MOs.

To statistically group the MOs into polytomous “super” units, a tetrachoric correlation matrix (as the data is ordinal in nature) was created for the simulations using the dichotomous MOs within each sampled panel. A principal component analysis (PCA) was conducted on the tetrachoric correlations to determine the number of components (factors) to be extracted. The number of factors to be extracted was determined by counting the number of eigenvalues greater than one[2]. A common-factor model, with the number of factors predetermined from the PCA results, was then fit to the tetrachoric matrix. Finally, a Varimax rotation was applied to the matrix of loadings. The rotated loading matrix was then used to perform a hierarchical cluster analysis (see Luecht & Miller, 1992). The cluster analysis effectively groups together items that share a similar direction of maximum information in the latent factor space. The results of the cluster analysis were used to group MO's into polytomous score units (PSUs). An arbitrary guideline was established to allow no more than nine MOs per PSU, yielding a score range from 0-9 (10 possible score categories). To create the polytomous scores from the dichotomous MOs, the binary responses were summed within each cluster for each candidate to produce a polytomous score.

As noted above, the tab-based simulation (TBS) MOs were grouped into polytomous items by making use of the content-based tab assignments created by subject matter experts (SMEs) and test developers. Small clusters of tasks were identified by SMEs as measuring a predetermined skill. Within these clusters, MO scoring was performed using Boolean rules. The TBS "super" items were formed by grouping together the MOs associated with the common skill and computing a polytomous score as the sum of the MO scores for that skill. Although, in practice, the TBS MOs for each "super" item would actually be administered under the same tab, here, the MOs were merely linked to a tab identifier, and these identifiers used to group the MOs. Polytomous items that had more than ten score categories were randomly split in half to meet the ten-category maximum mentioned above. For all TBS polytomous items created, categories with no or few candidates present were collapsed into adjacent categories. Table1 (shown earlier) also shows the number of simulation items created for each panel for the dichotomous MOs, the statistically grouped, and TBS methods.

Creating residual covariance structures

If the assumption of local independence holds, it can be presumed that

$$Cov\left(X_i, X_j \middle| \theta^*\right) = E\left(x_i, x_j \middle| \theta^*\right) = 0 \tag{3}$$

Where $x = X - \mu$ forall *i,j* item pairs, with i≠j, and θ^* is any conditioning variable—usually a proficiency score (or *vector* of conditioning variables in the multidimensional case) . The presence of LID due to contextual and scoring or dimensional effects would cause some conditional covariance between item responses. One way to assess the magnitude and direction of LID present between items is to estimate the unconditional covariance, by integrating over the conditioning variable, θ*. Yen's Q_3 statistic (1984), a standardized residual covariance statistic, is typically applied to the residual errors for each pair for items. A number of studies have

[2] Though there are better methods for determining the number factors, this method was used by the researcher so that the process could be automated and consistent across sampled panels

shown Q_3 to be an effective way to describe the presence and magnitude of LID (Yen,1984; Yen,1993; Habing, Finch, & Roberts , 2005; Chen & Thissen, 1997).

The residual errors were calculated with:

$$d_{ik} = x_{ik} - E\left(X_{ik}\right) \tag{2}$$

Where x_{ikc} is the observed score on the i^{th} for the k^{th} examinee and $E\left(X_{ki}\right)$ is the expected score of the k^{th} examinee on the i^{th} item given some estimate of ability (The 4 IRT score and 3 raw-core used in conditioning are described below). For the dichotomous IRT conditions, $E\left(x_{ik}\right) = P\left(\hat{\theta}_k, \hat{\boldsymbol{\xi}}_{\mathbf{i}}\right)$ where $\hat{\theta}_k$ is an estimated latent ability of the candidate and $\hat{\boldsymbol{\xi}}_{\mathbf{i}}$ is a vector of item parameters.

In the polytomous case, $E\left(X_{ik}\left|\hat{\theta}_k, \hat{\xi}_k\right.\right) = \sum_{c=1}^{M}\left(C-1\right)P\left(\hat{\theta}_k, \hat{\boldsymbol{\xi}}_{\mathbf{i}}\right)$ where $\hat{P}_{ic}\left(\hat{\theta}_k\right)$ is the probability of falling into the c^{th} category, and $\hat{\theta}_k$ is a candidate's estimated proficiency. Note that here any appropriate estimate of proficiency, $\hat{\theta}_k$ (e.g. based on MCQ only or all items) could be used. Q_3 is calculated as the Pearson product-moment correlation of all pairs of item residuals— $Q_3 = r_{d_i d_j}$. Q_3 correlations matrices were created for all panels under all three scoring methods and the conditioning variables below.

A limited set of IRT-based ability estimates[3] were used as conditioning variables in the computing the Q_3 statistic. The scaling methods used to produce these score were selected to demonstrate some very different, but operationally feasible, approaches that could be taken to account for residual covariance. In theory, any score estimate or vector of score estimates (in multidimensional model) could be used as conditioning variable. Three unidimensional scaling methods and one multidimensional scaling method were selected to produce EAP score estimates. All of the scaling methods were fully crossed with the three scoring conditions described above, producing a conditional covariance structure for each scaling method by scoring method combination.

The three unidimensional IRT scaling approaches used for obtaining the conditioning variables are: (1) estimating a proficiency score, θ_{MCQ}, using only the responses to the MCQ questions; (2) scaling using the combined, complete responses string (MCQ and simulations) with the MCQ item parameters from Approach #1 used as an anchor set of items (i.e., fixing the IRT parameter estimates, based only on MCQ calibration), and (3) scaling all of the responses together (MCQ and simulation MOs in a large, concurrent calibration). The MCQ-only is the most limited

[3] Each conditioning All IRT item calibrations were performed using Hanson's IRT Control Language (ICL) program (Hanson, 2002). ICL was also used to generate EAP theta estimates for each candidate. For the IRT calibrations, dichotomous MCQ and simulations were scaled using the 3pl model. Muraki's (1992) Generalized Partial Credit Model (GPCM) was used to fit the polytomous simulations.

approach in that it ignores the simulation MOs. Nonetheless, this approach does provide a baseline, of sorts, for the Q_3 analyses. The anchored of the simulation to the MCQ scale represents the current operational scaling practices used for this testing program. In this scaling method, the underlying trait is still assumed that of the MCQ trait, but the responses to the simulation are allowed to contribute to the overall score estimation. The last IRT unidimensional scaling method estimates abilities and item parameters based on a joint calibration the simulation and MCQ items. This type of calibration results in a composite unidimensional trait, where the contribution of the individual items to the score estimate are weighted by their relative influence or sensitivity to the underlying trait being estimated (and affected both by the number of MOs compared to the MCQ and the statistical discrimination of the MCQ items and simulation MOs).

Only one multidimensional approach to scaling/conditioning was considered. It assumes a rather simple multidimensional structure for the test—a two trait model with simple structure, and ignoring the covariance among the traits. To model this, each subset of items (MCQ and simulations) are treated as loading to two separate scales. A unidimensional IRT calibration is completed for each section, and two separate Bayes *expected a posteriori* (EAP) abilities produced, corresponding to the two scales. Expected scores for MCQ and simulation items are found by conditioning on the appropriate IRT score (i.e., MCQ responses were conditioned on the estimated θ_{MCQ} and the simulation MOs were conditioned on the estimated θ_{SIM}). For this condition, a mean correlation of zero for the residuals indicates that the ability estimate produced by each separate scale explains the response patterns reasonably well for that particular section of the test.

The Q_3 indices for each scaling-scoring combination were then decomposed into four subsets for comparison—the residual correlations for the items composing simulation one, the residual correlations for the items composing simulation two[4], the correlations between the items of simulation one and two, and the correlations between simulation and MCQ items. The minimum, maximum, and mean were recorded Q_3 from these subsets as indictors of the range and the magnitude of dependences. Each Q_3 a matrix was also transformed to Z scores the Fisher r-to-z transformation (see equation 4).

$$z_{ij} = \frac{\ln\left(1 + r_{ij}\right)}{2} - \frac{\ln\left(1 - r_{ij}\right)}{2} \qquad (3)$$

The distribution of $z_{ij} \sim N\left(0, \frac{1}{N-3}\right)$, where N is the number of cases used to compute the correlation. The transformed values were used to count the number of significant correlation coefficients at an alpha level of 0.05. In addition, the number of divergent (negative) and convergent (positive) elements in the correlation matrix were collected and summarized.

[4] The results for within simulation 1 and simulation 2 were later collapsed to make overall within simulation comparisons

Results

As the distribution of Q_3 for all the panels within any given sub-examination are very similar, the results in the tables and figures below reflect an aggregation of all the panels within a given sub-examination. Tables 1-3 show the descriptive statistic retained form the analysis of each test section. Each table is organized by the conditioning variable used to generate the Q_3 correlation matrix (column labeled "Condition"), the comparison reported (dependency within any given simulations, between the simulation on a panel, or between the MCQ and simulations in a panel), and the scoring method used —dichotomous, statistically grouped polytomous (Poly Stat), or tab-based polytomous (Poly TBS). The results in terms of the three research questions are discussed separately below.

Effects of dichotomous and polytomous scoring Q_3

In general, the results are consistent across all sections. For the single multidimensional condition (the two separate scales approach), all mean Q_3 correlations are negative and the magnitude of difference between dichotomous and polytomous scoring is relatively small. This finding was an exception to the general findings (i.e., compared to the other unidimensional scaling conditions). These differences are discussed in greater detail in the last part of the results section. The discussion comparing scoring types that follows in this section considers all of the scaling conditions except two separate scales approach (the top row in figure 1-3 below).

Tables 1-3 show clearly that the results of the Q_3 analysis for each of the three sections, aggregated across the sampled panels within each, of the examination. The mean Q_3 value for polytomous scoring tends to be farther from zero than the corresponding mean Q_3 value under dichotomous scoring. This result is at first counterintuitive, as the purpose of adding the polytomous scoring condition was to *reduce* residual covariance ,and will be explored in greater detail below. For all conditions, the range of Q_3 is much larger than under dichotomous scoring. This is particularly true in the item-pair correlations within a simulation (where contextual and scoring dependencies are most likely to occur). These tables also show that the residual covariance is primarily convergent in the within and between simulation comparisons under polytomous scoring. In the dichotomous condition scoring condition, a majority of the residual correlations are convergent, but with a larger proportion of divergent elements when compared to the polytomous scoring methods. The between simulation and MCQ comparisons (denoted MCQ and Sims in Tables 1-3) are similar results regardless of scoring method, with mean Q_3 values near zero and nearly equally amount of positive and negative correlations.

Figures 1-3 show the distributions of Q3 for the aggregated results of each section of the examination. The magnitude of Q_3 is plotted on the y-axis. The three columns of each of these figures display, respectively, the residual correlations of each pair of MOs within a single simulation, the residual correlations of the MOs in the two different simulations in single panel, and the residual correlations between all of the simulation MOs and the MCQ items in a single panel. Each row represents one of the scaling methods employed. In the within-simulation comparisons (column one of Figures 1-3), across all scaling methods, it is clear that polytomous scoring (whether statistical or tab-based) results in far fewer extreme dependencies and far fewer divergent correlations than dichotomous scoring. Comparing this pattern with the plots in the second column (between simulation comparisons), the range of Q_3 correlations is still noticeably smaller for polytomous scoring methods than for dichotomous scoring, with fewer extreme values for all three scoring methods.

Recall that contextual and scoring dependencies arise only within a single simulation. In the dichotomous scoring condition, the extreme Q_3 values apparent when moving from the within- to between– simulation comparisons (columns one and two of the figures) are likely due to the contextual/scoring dependencies within a simulation. The fact that no such extreme values appear in the corresponding polytomous scoring conditions in the within simulation comparisons is relatively strong evidence that extreme dependencies due to contextual/scoring are indeed mitigated by moving toward polytomous scoring. Note that forming sets of polytomous items based on logical, content, or other extraneous criteria, will likely not account for *all* contextual dependencies.

Because (theoretically speaking) No contextual or scoring dependencies exist in the between-simulation correlations (items from different simulations share no common context or scoring rules), it is assumed any residual covariance is largely evident of inherent multidimensionality in the test. The magnitude of Q_3 for the between-simulation Q_3 statistics is remains somewhat comparable to the Q_3 values shown for the within-simulation conditions. Comparatively, this finding suggests that there is some multidimensionality present in the data—dimensionality not effectively explained by the unidimensional composite trait.

Comparing TBS and Statistically Optimum Sims

Figures 1-3 can also be used to compare the two methods of polytomous scoring. Overall, the TBS and the statistically grouped simulations have very similar distributions of Q_3. Distributions of statistically "optimal" polytomous simulations tend to be slightly less variable than the TBS simulations. Comparing the mean Q_3 values for these two methods (see Tables 2-4) or the median values of Q_3 (depicted in Figure 1-3) leads to inconclusive results; neither method of grouping for purposes of polytomous scoring clearly outperforms the other in terms of reducing the magnitude of the residual covariance.

Table 2: Dependency Analysis Results Across Section 1

				Mean	*Min*	*Max*	*Num*		*Num*	*Num*
Conditioning	*Comparison*	*Scoring*	*N*	*Q3*	*Q3*	*Q3*	*Sig*	*% sig*	*+*	*-*
Theta (MCQ)	**Within Simulations**	*Dichotomous*	2154	0.06	-0.19	0.89	493	22.89%	1587	567
		Poly (Stat)	125	0.12	-0.07	0.35	49	39.20%	111	14
		Poly (TBS)	48	0.14	-0.07	0.44	27	56.25%	46	2
	Between Simulations	*Dichotomous*	2232	0.04	-0.17	0.57	278	12.46%	1564	668
		Poly (Stat)	143	0.11	-0.12	0.42	61	42.66%	129	14
		Poly (TBS)	64	0.15	-0.04	0.50	39	60.94%	60	4
	MCQ and Sims	*Dichotomous*	23625	0.00	-0.52	0.57	5367	22.72%	12002	11623
		Poly (Stat)	6000	0.00	-0.54	0.44	1386	23.10%	2968	3032
		Poly (TBS)	4000	0.00	-0.45	0.48	925	23.13%	2000	2000
Theta (Anchored)	**Within Simulations**	*Dichotomous*	2154	0.04	-0.21	0.88	373	17.32%	1385	769
		Poly (Stat)	125	0.10	-0.08	0.36	39	31.20%	103	22
		Poly (TBS)	48	0.12	-0.08	0.43	20	41.67%	43	5
	Between Simulations	*Dichotomous*	2232	0.01	-0.19	0.53	184	8.24%	1271	961
		Poly (Stat)	143	0.09	-0.12	0.42	49	34.27%	122	21
		Poly (TBS)	64	0.12	-0.08	0.49	28	43.75%	58	6
	MCQ and Sims	*Dichotomous*	23625	-0.01	-0.53	0.54	5587	23.65%	10338	13287
		Poly (Stat)	6000	-0.01	-0.54	0.44	1433	23.88%	2769	3231
		Poly (TBS)	4000	0.00	-0.47	0.48	932	23.30%	1939	2061
Theta (Unanchored)	**Within Simulations**	*Dichotomous*	2154	0.03	-0.21	0.88	332	15.41%	1312	842
		Poly (Stat)	125	0.09	-0.08	0.36	34	27.20%	100	25
		Poly (TBS)	48	0.11	-0.10	0.42	20	41.67%	41	7
	Between Simulations	*Dichotomous*	2232	0.01	-0.19	0.51	167	7.48%	1171	1061
		Poly (Stat)	143	0.08	-0.13	0.41	42	29.37%	121	22
		Poly (TBS)	64	0.12	-0.08	0.49	26	40.63%	56	8
	MCQ and Sims	*Dichotomous*	23625	-0.02	-0.53	0.53	5642	23.88%	9923	13702
		Poly (Stat)	6000	-0.02	-0.54	0.43	1474	24.57%	2454	3546
		Poly (TBS)	4000	-0.02	-0.46	0.48	997	24.93%	1654	2346
Thetas (Two-scales)	**Within Simulations**	*Dichotomous*	2154	0.00	-0.26	0.81	363	16.85%	968	1186
		Poly (Stat)	125	-0.07	-0.29	0.21	40	32.00%	29	96
		Poly (TBS)	48	-0.10	-0.37	0.11	17	35.42%	6	42
	Between Simulations	*Dichotomous*	2232	-0.02	-0.23	0.37	230	10.30%	878	1354
		Poly (Stat)	143	-0.08	-0.29	0.10	44	30.77%	27	116
		Poly (TBS)	64	-0.09	-0.30	0.35	33	51.56%	11	53
	MCQ and Sims	*Dichotomous*	23625	0.00	-0.52	0.58	5178	21.92%	11977	11648
		Poly (Stat)	6000	0.00	-0.53	0.44	1321	22.02%	2996	3004
		Poly (TBS)	4000	0.00	-0.49	0.47	856	21.40%	1925	2075

Table 3: Dependency Analysis Results Across Section 2

				Mean	*Min*	*Max*	*Num*		*Num*	*Num*
Conditioning	***Comparison***	***Scoring***	***N***	***Q3***	***Q3***	***Q3***	***Sig***	***% sig***	***+***	***-***
Theta (MCQ)	**Within Simulations**	*Dichotomous*	1268	0.23	-0.14	0.95	946	74.61%	1232	36
		Poly (Stat)	36	0.24	-0.03	0.64	33	91.67%	35	1
		Poly (TBS)	48	0.14	-0.07	0.44	27	56.25%	46	2
	Between Simulations	*Dichotomous*	1288	0.06	-0.14	0.35	351	27.25%	1043	245
		Poly (Stat)	47	0.15	-0.12	0.31	34	72.34%	42	5
		Poly (TBS)	64	0.15	-0.04	0.50	39	60.94%	60	4
	MCQ and Sims	*Dichotomous*	17970	0.00	-0.29	0.34	3122	17.37%	8878	9092
		Poly (Stat)	3371	0.00	-0.26	0.29	575	17.06%	1620	1751
		Poly (TBS)	4000	0.00	-0.45	0.48	925	23.13%	2000	2000
Theta (Anchored)	**Within Simulations**	*Dichotomous*	1268	0.16	-0.23	0.94	681	53.71%	1126	142
		Poly (Stat)	36	0.21	-0.06	0.60	29	80.56%	34	2
		Poly (TBS)	48	0.12	-0.08	0.43	20	41.67%	43	5
	Between Simulations	*Dichotomous*	1288	0.00	-0.19	0.34	140	10.87%	606	682
		Poly (Stat)	47	0.12	-0.16	0.29	26	55.32%	40	7
		Poly (TBS)	64	0.12	-0.08	0.49	28	43.75%	58	6
	MCQ and Sims	*Dichotomous*	17970	-0.03	-0.38	0.27	3990	22.20%	6175	11795
		Poly (Stat)	3371	-0.01	-0.27	0.30	612	18.15%	1487	1884
		Poly (TBS)	4000	0.00	-0.47	0.48	932	23.30%	1939	2061
Theta (Unanchored)	**Within Simulations**	*Dichotomous*	1268	0.12	-0.27	0.90	511	40.30%	998	270
		Poly (Stat)	36	0.26	-0.04	0.59	29	80.56%	34	2
		Poly (TBS)	31	0.27	0.01	0.80	25	80.65%	31	0
	Between Simulations	*Dichotomous*	1288	-0.03	-0.21	0.31	177	13.74%	400	888
		Poly (Stat)	47	0.10	-0.15	0.29	25	53.19%	40	7
		Poly (TBS)	42	0.15	-0.02	0.37	28	66.67%	41	1
	MCQ and Sims	*Dichotomous*	17970	-0.03	-0.38	0.28	4028	22.42%	6385	11585
		Poly (Stat)	3371	-0.02	-0.31	0.29	698	20.71%	1306	2065
		Poly (TBS)	3245	-0.02	-0.30	0.30	632	19.48%	1299	1946
Thetas (Two-scales)	**Within Simulations**	*Dichotomous*	1268	0.07	-0.40	0.86	478	37.70%	760	508
		Poly (Stat)	36	-0.10	-0.61	0.46	23	63.89%	19	17
		Poly (TBS)	31	-0.04	-0.40	0.29	19	61.29%	10	21
	Between Simulations	*Dichotomous*	1288	-0.04	-0.30	0.32	195	15.14%	337	951
		Poly (Stat)	47	-0.15	-0.36	0.05	27	57.45%	3	44
		Poly (TBS)	42	-0.14	-0.54	0.28	22	52.38%	10	32
	MCQ and Sims	*Dichotomous*	17970	0.00	-0.30	0.30	2881	16.03%	9156	8814
		Poly (Stat)	3371	0.00	-0.27	0.31	511	15.16%	1674	1697
		Poly (TBS)	3245	0.00	-0.29	0.29	492	15.16%	1604	1641

Table 4: Dependency Analysis Results Across Section 3

Conditioning	*Comparison*	*Scoring*	*N*	*Mean Q3*	*Min Q3*	*Max Q3*	*Num Sig*	*% sig*	*Num +*	*Num -*
Theta (MCQ)	**Within Simulations**	*Dichotomous*	3030	0.20	-0.12	0.91	2058	67.92%	2837	193
		Poly (Stat)	188	0.19	0.01	0.83	173	92.02%	188	0
		Poly (TBS)	48	0.14	-0.07	0.44	27	56.25%	46	2
	Between Simulations	*Dichotomous*	2886	0.07	-0.15	0.31	796	27.58%	2378	508
		Poly (Stat)	161	0.16	-0.06	0.44	113	70.19%	154	7
		Poly (TBS)	64	0.15	-0.04	0.50	39	60.94%	60	4
	MCQ and Sims	*Dichotomous*	21800	0.00	-0.34	0.40	3570	16.38%	10552	11248
		Poly (Stat)	5200	-0.01	-0.30	0.36	898	17.27%	2444	2756
		Poly (TBS)	4000	0.00	-0.45	0.48	925	23.13%	2000	2000
Theta (Anchored)	**Within Simulations**	*Dichotomous*	3030	0.10	-0.30	0.88	1213	40.03%	2178	852
		Poly (Stat)	188	0.16	-0.05	0.77	144	76.60%	184	4
		Poly (TBS)	48	0.12	-0.08	0.43	20	41.67%	43	5
	Between Simulations	*Dichotomous*	2886	-0.01	-0.27	0.22	250	8.66%	1182	1704
		Poly (Stat)	161	0.09	-0.11	0.26	53	32.92%	133	28
		Poly (TBS)	64	0.12	-0.08	0.49	28	43.75%	58	6
	MCQ and Sims	*Dichotomous*	21800	-0.04	-0.43	0.38	5150	23.62%	7480	14320
		Poly (Stat)	5200	-0.03	-0.35	0.35	1269	24.40%	1839	3361
		Poly (TBS)	4000	0.00	-0.47	0.48	932	23.30%	1939	2061
Theta (Unanchored)	**Within Simulations**	*Dichotomous*	3030	0.06	-0.33	0.87	1082	35.71%	1798	1232
		Poly (Stat)	188	0.22	-0.12	0.76	125	66.49%	175	13
		Poly (TBS)	45	0.23	-0.07	0.75	32	71.11%	42	3
	Between Simulations	*Dichotomous*	2886	-0.03	-0.24	0.21	332	11.50%	931	1955
		Poly (Stat)	161	0.06	-0.15	0.20	40	24.84%	120	41
		Poly (TBS)	54	0.12	-0.11	0.26	27	50.00%	50	4
	MCQ and Sims	*Dichotomous*	21800	-0.02	-0.43	0.38	4576	20.99%	8701	13099
		Poly (Stat)	5200	-0.04	-0.35	0.30	1389	26.71%	1685	3515
		Poly (TBS)	3000	-0.03	-0.36	0.30	679	22.63%	1065	1935
Thetas (Two-scales)	**Within Simulations**	*Dichotomous*	3030	0.05	-0.38	0.88	1103	36.40%	1715	1315
		Poly (Stat)	188	0.02	-0.30	0.71	77	40.96%	83	105
		Poly (TBS)	45	-0.02	-0.48	0.64	23	51.11%	16	29
	Between Simulations	*Dichotomous*	2886	-0.01	-0.22	0.26	295	10.22%	1238	1648
		Poly (Stat)	161	-0.09	-0.29	0.20	53	32.92%	33	128
		Poly (TBS)	54	-0.13	-0.41	0.22	25	46.30%	11	43
	MCQ and Sims	*Dichotomous*	21800	0.00	-0.32	0.38	3405	15.62%	11215	10585
		Poly (Stat)	5200	0.00	-0.29	0.36	834	16.04%	2598	2602
		Poly (TBS)	3000	0.00	-0.32	0.38	501	16.70%	1523	1477

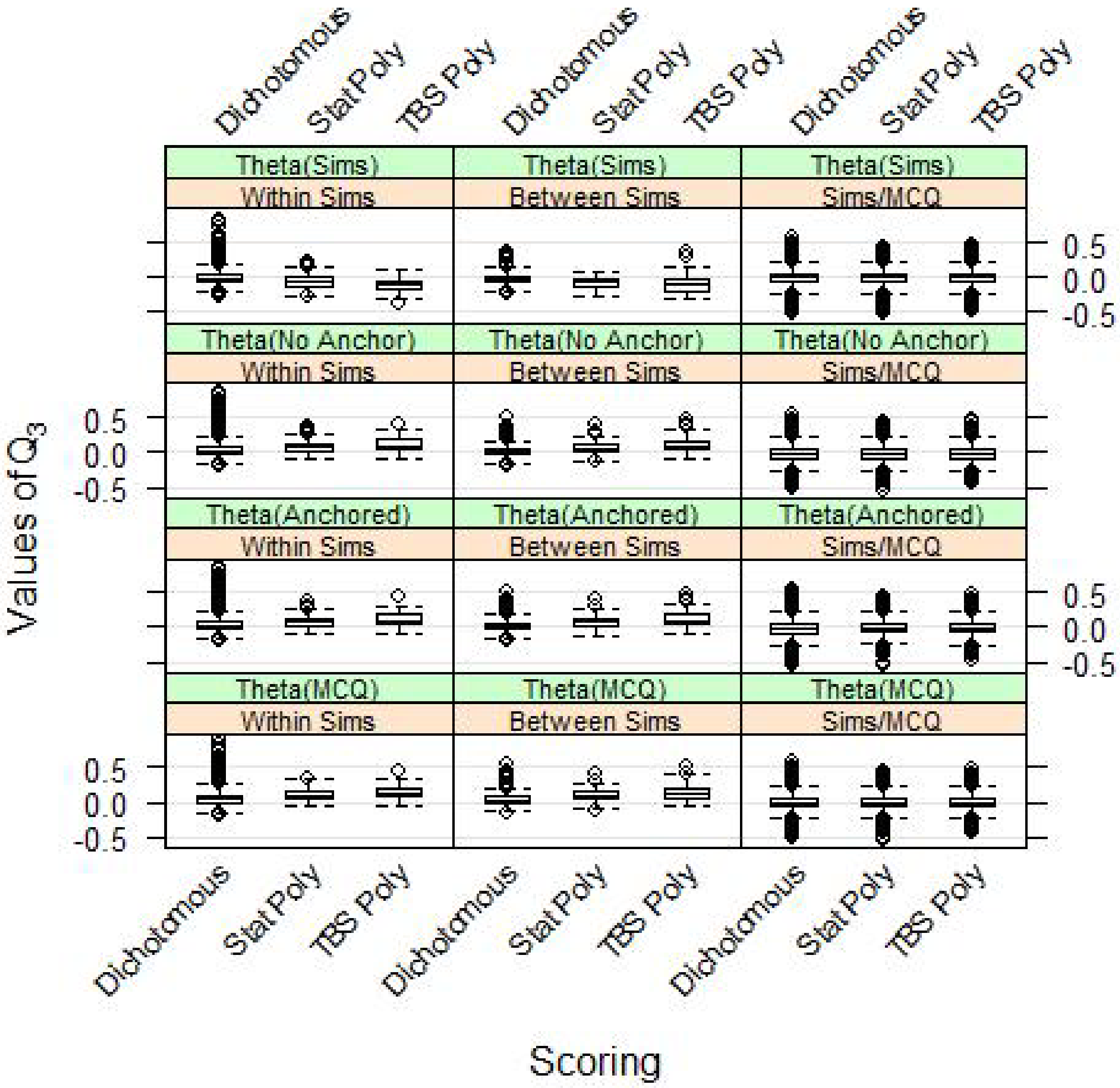

Figure 1: Distribution of Q_3 Statistics for exam section 1. Note: The values of Q_3 for each item pair are plotted on the y axis, on the x-axis are the three scoring procedure employed—dichotomous, polytomous items grouped by statistical methods (stat poly) and expert constructed polytomous items (TBS poly). Each row of panels shares a common conditioning variable—Theta(MCQ) for the ability estimate based only on MCQ items, Theta(Anchored)for the ability estimate based all items with simulation anchored to the MCA scale, Theta(Unanchored) for the ability estimate for all items, and Theta(Sims) for the separate scaling of the simulations and MCQ items. Each column of panels represents one of the comparisons of interest—residual correlations for MO within a single simulation (Within Sims) , residual correlations between the MOs of simulations of a single panel (Between Sims), and residual correlations between MCQ items and simulations within a single panel (MCQ/Sims).

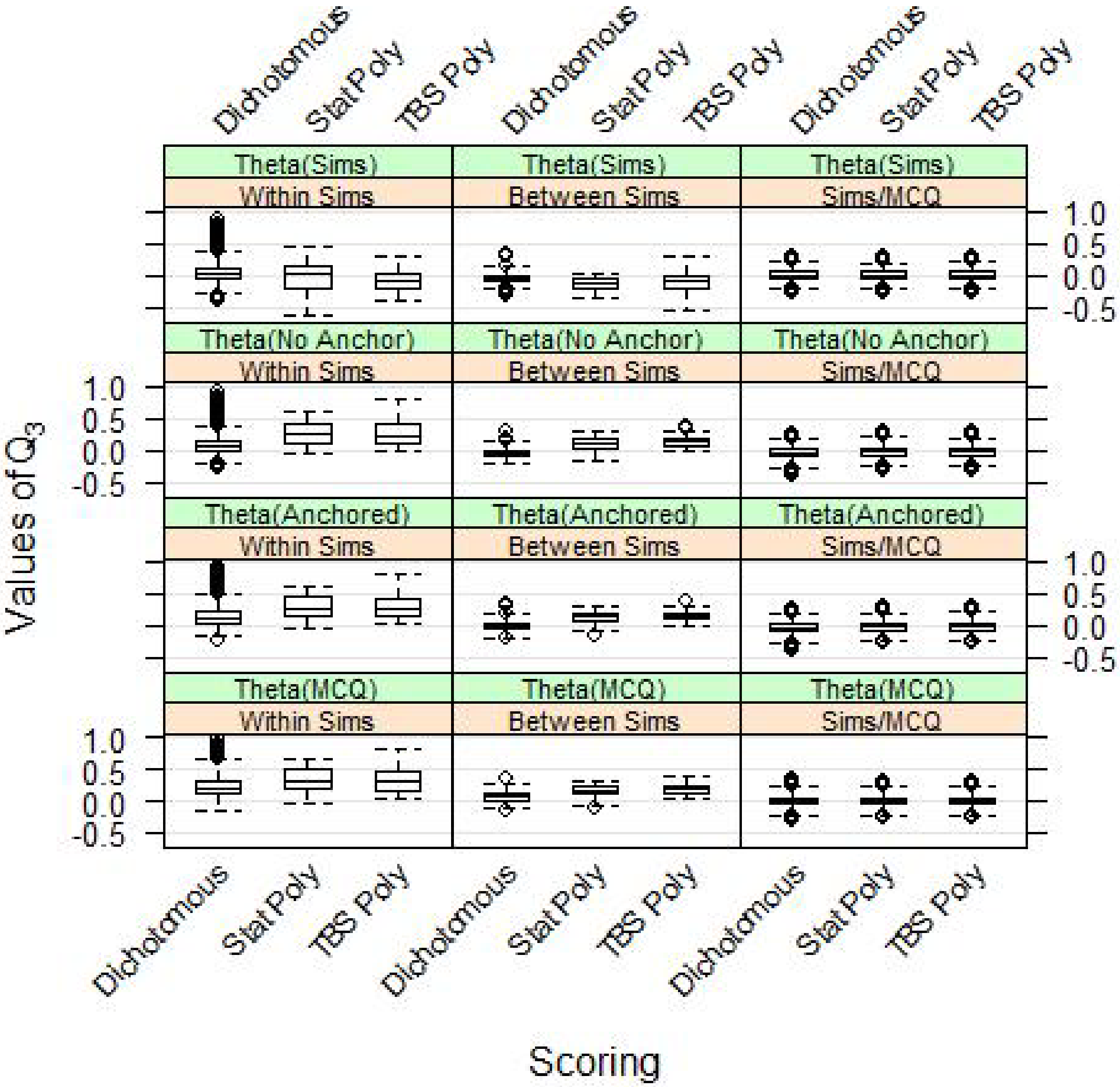

Figure 2: Distribution of Q_3 Statistics for exam section 2. Note: The values of Q_3 for each item pair are plotted on the y axis, on the x-axis are the three scoring procedure employed—dichotomous, polytomous items grouped by statistical methods (stat poly) and expert constructed polytomous items (TBS poly). Each row of panels shares a common conditioning variable—Theta(MCQ) for the ability estimate based only on MCQ items, Theta(Anchored)for the ability estimate based all items with simulation anchored to the MCA scale, Theta(Unanchored) for the ability estimate for all items, and Theta(Sims) for the separate scaling of the simulations and MCQ items. Each column of panels represents one of the comparisons of interest—residual correlations for MO within a single simulation (Within Sims) , residual correlations between the MOs of simulations of a single panel (Between Sims), and residual correlations between MCQ items and simulations within a single panel (MCQ/Sims).

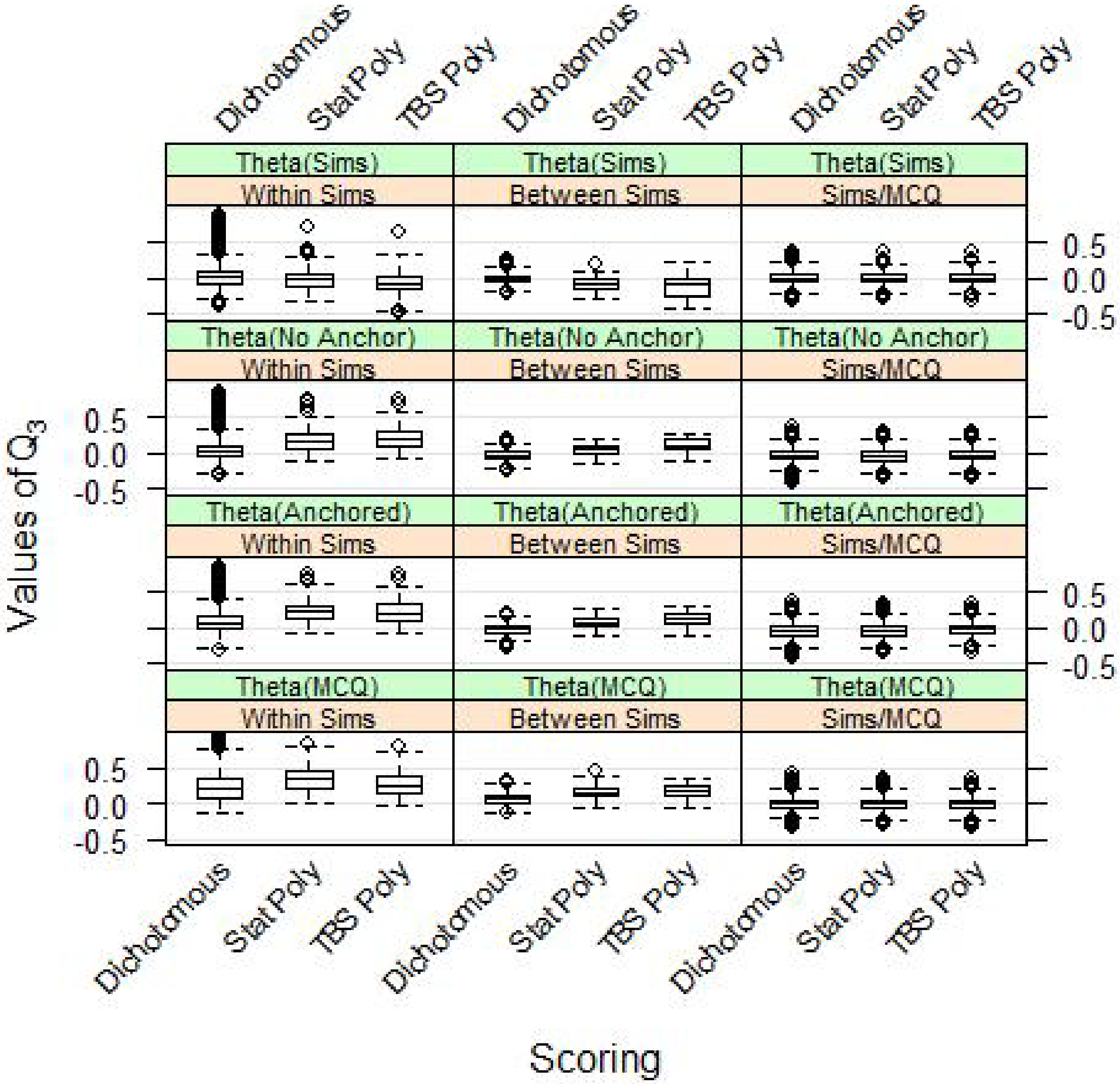

Figure 3: Distribution of Q_3 Statistics for exam section 3. Note: The values of Q_3 for each item pair are plotted on the y axis, on the x-axis are the three scoring procedure employed—dichotomous, polytomous items grouped by statistical methods (stat poly) and expert constructed polytomous items (TBS poly). Each row of panels shares a common conditioning variable—Theta(MCQ) for the ability estimate based only on MCQ items, Theta(Anchored)for the ability estimate based all items with simulation anchored to the MCA scale, Theta(No Anchor) for the ability estimate for all items, and Theta(Sims) for the separate scaling of the simulations and MCQ items. Each column of panels represents one of the comparisons of interest—residual correlations for MO within a single simulation (Within Sims), residual correlations between the MOs of simulations of a single panel (Between Sims), and residual correlations between MCQ items and simulations within a single panel (MCQ/Sims).

Comparison of scaling approaches

Tables 2-3 and Figures 1-3 (shown earlier) were also used to compare competing scaling methods across all three scoring methods. All of the unidimensional approaches to scaling produced positive mean Q_3 correlations (see Tables 2-4). In contrast, the separate scales approach (a pseudo-multidimensional method included in this study) produces mean Q_3 that are closer to zero and slightly negative.

Looking first at the unidimensional conditions (denoted 'Theta(MCQ)', 'Theta(Anchored)', & 'Theta(No Anchor)' in Figures 1-3), it is clear that, as the simulation responses contribute more to the estimation of ability, the dependencies decrease. The θ_{mcq} is the cleanest unidimensional scaling condition. This condition merely assumes that the MCQ items are measuring the same trait and are adequate to use as a conditioning variable for computing the residual statistics. However, this MCQ-only scaling method produces the highest mean Q_3 correlation in the within- and between-simulation comparisons, indicating that the MCQ trait does not adequate explain responses to the simulation items. A slight decrease in mean Q_3 is observed in the $\theta_{anchored}$ condition. The underlying trait is still directionally determined by MCQ items, but the inclusion of simulation responses in the estimation of the final latent scores produces a conditioning variable that incorporates some of the measurement information present in the simulation responses. The last unidimensional IRT conditioning score, $\theta_{No\ Anchor}$, is based on a simultaneous calibration of all of the simulation MOs and MCQ items. Here, the underlying trait is not assumed to defined by solely the MCQ items, but is instead jointly determined by both item types. As one might expect, this scaling condition produces the smallest mean Q_3 of the three unidimensional scaling conditions. This finding again suggests that there is multidimensionality present in the data, at least separable by item type.

Even with the reductions in mean Q_3 over the three unidimensional conditions, the mean Q_3 in both the between-simulations and within-simulation comparisons is still convergent and sizeable. The separate scales condition (denoted as Theta(Sims) in Tables 2-3 and Figures 1-3) produces Q_3 correlations that tend to be slightly negative[5] and closer to zero (on average) that the Q_3 values produced under the corresponding scoring conditions with unidimensional scaling of θ. This is additional, strong evidence that the simulations are likely measuring something distinct from the MCQ items.

[5] This divergent covariance is evidence that within the simulations items scale, the various items may appear to be distinct, each measuring at least in some part, something unique. The two separate scales model is used as an illustration, but in reality a model that accounts for the uniqueness of the different simulation items (e.g. a testlet model or bi-factor model) may be more appropriate.

Discussion

Comparing Polytomous and Dichotomous Scoring

Overall, dichotomous scoring produces average residual correlations that are closer to zero. On the surface, this would suggest that polytomous scoring does little to reduce LID. That finding is counter-intuitive and somewhat contradicts other research. However, it is also important to consider not just the location of the Q_3 distribution, but also the range and variance of these correlations when comparing scoring methods. Dichotomous score produces a much more varied set of correlations and a great number of item pairs with extreme dependencies. The reduced range and reduced number of large dependency indicators produced under polytomous score *within* a particular simulation suggest that, while not eliminating non-zero covariance, polytomous scoring IS effective in controlling some of the extreme dependencies likely due to contextual and scoring issues. The similar amounts of residual covariance in the within-simulation and between-simulation (which has no contextual dependencies) comparisons further suggest that the amount of residual covariance after scoring polytomously may be largely due to issues of dimensionality.

TBS and Statistically Grouped Simulation

The results indicate that statically group polytomous simulations and TBS grouped items perform similarly (with perhaps a slight advantage to statically grouped polytomous score units). The process of creating statistically group is based solely on the empirical response to the simulation items, which could be highly variable in different panels that contain the same simulations. The result is a grouping that works locally, but is not stable across panels or time. Tab-based simulations are more stable over time, given their logical clustering basis as coded by SMEs. Because the TBS-grouped polytomous scoring performed about as well as the statistically grouped polytomous items, TBS may provide a promising method for creating smaller simulations (either scored dichotomously or clustered into polytomous items for scoring). From a test development perspective, the TBS approach to scoring represents a proactive process that identifies the particular skills that need to be measured by the simulations, and designing items that assess the skills. Though considerable research must still be done on TBS scoring , this study provides some preliminary evidence to suggest that a cognitively based process for creating complex performance assessment might be successful from an operational perspective.

Comparing Scaling Methods

The results in the comparison of the scaling comparison strongly suggest that the simulations and MCQ section constitute distinct scales. While the combination of polytomous scoring and separate scaling produced the most optimal residual covariance matrix (located near zero with a very limited range of dependencies), separate scaling was also a great benefit to the dichotomously scored simulations. The benefits of separate scaling in terms of conditional covariance and assumptions of local independence are clear, but such a move would require careful consideration.

Separate scaling would require that validation evidence must be produced for *both* scales. For the long existing MCQ section, this is not problematic, but a newly formed simulation scale would require more attention. A simulation "trait" would have to carefully defined, and represented within each panel of each sub-examination. The current work with tab-based simulations might prove effective in accomplishing this task. In a two-scale test, a decision on how many scores to report would have to be considered. If a single score would be reported and/or used to make pass/fail decisions, the method for creating a composite is non-trivial and must be carefully considered. Operational practices and procedures would also have to be considered (equating, item-banking, etc.).

References

Ackerman, T. A. (1987). *The robustness of LOGIST and BILOG IRT estimation programs to violations of local independence.* Paper presented at the annual meeting of the American Educational Research Association, Washington, D.C. (ERIC Document Reproduction Service No. ED284902). Retrieved September 15, 2006 from the ERIC database.

De Champlain, A., & Gessaroli, M. (1991). Assessing Test Dimensionality Using an Index Based on Nonlinear Factor Analysis. Retrieved Sunday, October 8, 2006 from the ERIC database.

De Champlain, A. (1996). The Effect of Multidimensionality on IRT True-Score Equating for Subgroups of Examinees. *Journal of Educational Measurement*, *33*(2), 181-201.

Ferrara, S., & Others, A. (1997). Contextual Characteristics of Locally Dependent Open-Ended Item Clusters in a Large-Scale Performance Assessment. *Applied Measurement in Education*, *10*(2), 123-144.

Ferrara, S., Huynh, H., & Michaels, H. (1999). Contextual Explanations of Local Dependence in Item Clusters in a Large Scale Hands-On Science Performance Assessment. *Journal of Educational Measurement*, *36*(2),

Finch, H., & Habing, B. (2005). Comparison of NOHARM and DETECT in Item Cluster Recovery: Counting Dimensions and Allocating Items. *Journal of Educational Measurement*, *42*(2), 149-169.

Gessaroli, M., & De Champlain, A. (1996). Using an Approximate Chi-Square Statistic To Test the Number of Dimensions Underlying the Responses to a Set of Items. *Journal of Educational Measurement*, *33*(2), 157-179.

Habing, B., Finch, H., & Roberts, J. (2005). A Q3 Statistic for Unfolding Item Response Theory Models: Assessment of Unidimensionality with Two Factors and Simple Structure. *Applied Psychological Measurement*, *29*(6), 457-471.

Hanson, B (2001). IRT Control Language [Computer Program]

Jöreskog, K. & Sorebom, D. (2005). LISREL 8.72 (Including PRELIS 2.0) Computer program. Chicago: Scientific Software International, Inc.

Keller, L., Swaminathan, H., & Sireci, S. (2003). Evaluating scoring procedures for context-dependent item sets. *Applied Measurement in Education*, *16*(3), 207-222.

Li, Y., Bolt, D., & Fu, J. (2006). A Comparison of Alternative Models for Testlets. *Applied Psychological Measurement*, *30*(1), 3-21.

Luecht., R. & Miller, T. (1992). Unidimensional Calibrations and Interpretation of Composite Traits for Multidimensional Tests. *Applied Psychological Measurement*, *16*(3), 279-294.

Luecht R. (October 2003). *Calibration and Scoring for the Uniform CPA Examination.* AICPA Technical Report Series 2, no. 14.

Luecht, R. M. (2000, April). *Implementing the computer-adaptive sequential testing (CAST) framework to mass produce high quality computer-adaptive and mastery tests.* Symposium paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Luecht, R. M.; Brumfield, T.; and Breithaupt, K. (2006). A Testlet Assembly Design for the Uniform CPA Examination. *Applied Measurement in Education*, 19(3), 189-202.

Luecht, R. M. & Nungester, R. J. (1998). Some practical applications of computerized adaptive sequential testing. *Journal of Educational Measurement*, 35, 229-249.

Mattern, K. (May 2004). *Incorporating Innovative Items into a Licensing Exam: An Analysis of Psychometric Properties of Simulations.* AICPA Technical Report Series 2, no. 5.

Miller, T., & Hirsch, T. (1992). Cluster Analysis of Angular Data in Applications of Multidimensional Item-Response Theory. *Applied Measurement in Education*, *5*(3), 193-211.

Muraki, E. (1992). A generalized partial credit model: application of an EM algorithm. *Applied Psychological Measurement*, *16*(2), 159-176.

Muraki, E. & Bock, D. (2002) PARSCLE 4.1 Computer program. Chicago: Scientific Software International, Inc.

Reese, L. (1999a). A Classical Test Theory Perspective on LSAT Local Item Dependence. LSAC Research Report Series. Statistical Report. Retrieved Sunday, October 15, 2006 from the ERIC database.

Reese, L. (1999b). Impact of Local Item Dependence on Item Response Theory Scoring in CAT. Law School Admission Council Computerized Testing Report. LSAC Research Report Series. Retrieved October 1, 2006 from the ERIC database.

Reese, L., & Pashley, P. (1999). Impact of Local Item Dependence on True-Score Equating. LSAC Research Report Series. Retrieved September 15, 2006 from the ERIC database.

Stark, S., Chernyshenko, O., & Drasgow, F. (April 2002). *Investigating the Effects of Local Dependence on the Accuracy of IRT Ability Estimation.* AICPA Technical Report Series 2, no. 15.

Stout, W., & Others, A. (1996). Conditional Covariance-Based Nonparametric Multidimensionality Assessment. *Applied Psychological Measurement*, *20*(4), 331-354. Retrieved Sunday, October 1, 2006 from the ERIC database.

Tate, R. (2004). Implications of Multidimensionality for Total Score and Subscore Performance. *Applied Measurement in Education*, *17*(2), 89-112.

Thompson, T., & Pommerich, M. (1996). Examining the Sources and Effects of Local Dependence.

Wainer, H. & Wang C. (2000) Using a new statistical model for testlets to score TOEFL. *Journal of Educational Measurement*, *37*(2), 203-220.

Wainer. H., Vevea, J. L., Camacho, F., Reeve, B. B., Rosa, K., Nelson, L., Swygert, K. A. & Thissen, D. (2001). Augmented scores—"borrowing strength" to compute scores based upon small numbers of items, In H. Wainer & D. Thissen (Eds.), *Test Scoring* (pp. 343-387). Mahwah, NJ: Lawrence Erlbaum Associates.

Xie, Y. (2001). Dimensionality, Dependence, or Both? An Application of the Item Bundle Model to Multidimensional Data. Paper presented at the Annual Meeting of the American Educational Research Association, Seattle, WA.

Yan, J. (1997). Examining Local Item Dependence Effects in a Large-Scale Science Assessment by a Rasch Partial Credit Model. Retrieved Sunday, October 15, 2006 from the ERIC database.

Yen, W. (1984). Effects of Local Item Dependence on the Fit and Equating Performance of the Three-Parameter Logistic Model. *Applied Psychological Measurement*, 8(2), 125-145.

Yen, W. (1993). Scaling Performance Assessments: Strategies for Managing Local Item Dependence. *Journal of Educational Measurement*, *30*(3), 187-213.

Zenisky, A., Hambleton, R., & Sireci, S. (2002). Identification and evaluation of local item dependencies in the Medical College Admissions Test. *Journal of Educational Measurement*, *39*(4), 291-309.