Investigating Alternative Approaches for Analyzing Item/Task Model Data

James B. Olsen, Alpine Testing Solutions Joseph A. Olsen, Brigham Young University and Russell W. Smith, Alpine Testing Solutions

Paper Presented at the Annual Meeting of the National Council on Measurement in Education, Denver CO, May, 2010

Abstract

This paper investigates alternative test theory models for use in analyzing item and task model data exemplifying item families, clusters, parcels, bundles, or testlets. The paper summarizes theory and analysis models for generalizing the item difficulty, discrimination, and model misfit parameters (or subsets) and test statistics from score computations based on individual items to groups or sets of items. The study uses an empirical dataset that exemplifies the concepts of item families, item bundles, item parcels, or testlets that may include conditional item/task dependence. The empirical data set is analyzed with multiple test models for computing item and test score statistics. The data set is analyzed first with individual test items and second with a meaningful item family structure. Results from the analyses are presented with item analysis statistics, item parameter estimates, standard errors, model fit indices, test characteristic curves, and test information curves.

Introduction

Scientific and technical advances occur when we pose fundamental investigative problems, decide relevant theories that might be helpful in solving the key problems, implement appropriate design environments and measurement processes and then critically evaluate the results to validate or revise our theories and problems. One fundamental problem in both computerized adaptive testing and statewide educational assessment is the need for creating large banks of well validated test items/tasks that can be produced in a very cost effective manner. A second fundamental problem in educational measurement is effective and efficient test assembly. Relevant educational measurement theories that might useful in addressing these two fundamental problems include: automated test assembly, optimal test design, item generation, item cloning, assessment engineering and item and task modeling and analysis.

The paper presents a theoretical and practical approach for using item and task modeling and analysis. We propose that item and task modeling and analysis will move the educational measurement profession forward in a very significant and meaningful way. This paper provides background theory, testing applications and analysis approaches for generalizing the classical or IRT item difficulty, discrimination, and model misfit parameters (or subsets) by using concepts of item families, clusters, parcels, bundles or testlets. The estimated item difficulties, discriminations, and model misfit parameters and associated parameter standard errors could apply to any child/sibling item or task selected from the item/task model. Test scores are accumulated scores or IRT proficiency estimates over a series of test items or performance tasks. Item and test statistics and parameters can be computed and reported at multiple levels of aggregation. The paper presents alternative theoretical and practical perspectives on the problem and potential solutions

Theoretical Background for Item and Task Modeling and Analysis Approaches

Item Parcels and Factored Homogeneous Item Dimensions

One of the first references to item families in the statistical literature is in the work of Raymond B. Cattell, a pioneering factor analyst of personality data. Raymond Cattell (1965, 1973, 1978; Burdsal & Vaughn, 1974) argued against factor analyzing individual personality items and argued for the use of homogeneous groups of personality items that he called *item parcels*. The item parcels were factor analyzed as groups of items rather than analyzing each of the individual items contributing to each parcel. As input to the factor analysis Comrey (1988) also argued for the use of sets of items, which he defined as *Factored Homogeneous Item Dimension*s (FHIDs; Comrey, 1967, 1984). These analysts believed that the item group score would provide a more stable aggregate score and more theoretically meaningful scoring unit than the individual personality item.

Item Forms, Item Shells and Domain Referenced Testing

Wells Hively's (1974) seminal book on domain referenced testing proposed the need for a better understanding of the behavioral foundations of educational accomplishment and a clear theory and technology to make it operational. Domain referenced testing requires a careful analysis of the universe or domain to be tested and an analysis of the expert's and learner's capabilities within the domain. In constructing a pool of items for the domain the item writers develop an extensive item bank that represents the fundamental characteristics of the universe or domain of knowledge to be tested.

When a learner answers a representative set of test items from the domain then the resulting sample score should allow for generalization to the universe or domain field. The goal of domain referenced testing was to make each concrete tested domain more representative of the total universe of skills within the domain. Domain referenced testing introduced the formalized concepts of item forms and item shells (Hively, 1974, p. 11). The item form or shell is the list of rules for generating or selecting a set of related items from the domain. When the content domain is clearly specified with domains and sub-domains, the testing procedure consisted of drawing representative samples of items from the domain and sub-domains and scoring examinee performance on those samples. With domain-referenced testing reliability is the accuracy with which estimates of probabilities of correct performance are made within the concrete domain and its sub-domains. Validity was the generalization from the probabilities of correct performance on the concrete domains to the larger universe of knowledge from which the concrete domain was specified.

With domain-referenced testing the theoretical and empirical focus is not on the specific test item but on the probabilities of successful performance of the learner within sub-domains for diagnostic formative assessment purposes and the learner probability of successful performance in the concrete domain for summative assessment purposes. The probability of successful performance in the specified and sampled concrete domain was generalized as an expected estimate of the performance of the learner on the universe of items or tasks that could have been administered rather than on the specific samples that were administered. Domain-referenced testing provides another measurement theory link for exploring item and task modeling and analysis for item groups and families.

Item Bundles

Rosenbaum's (1988) entitled his article in *Psychometrika* “item bundles.” He notes, “An item bundle is a small group of multiple choice items that share a common reading passage or graph, or a small group of matching items that share distractors. Item bundles are easily identified by paging through a copy of a test. Bundled items may violate the latent conditional independence assumption of unidimensional item response theory (IRT), but such a violation would not typically suggest the existence of a new fundamental human ability to read one specific reading passage or interpret one specific graph. It is important, therefore, to have theoretical concepts and empirical checks that distinguish between, one the one hand, anticipated violations of latent conditional independence within item bundles, and on the other hand, violations that cannot be attributed to idiosyncratic features of test format and instead suggest departures from unidimensionality (Rosenbaum, 1988, p. 349).”

Rosenbaum used the Mantel-Haenszel statistic to identify conditional independence among 780 possible pairs of multiple choice items in the 40-item population biology subscore of the College Board’s 1982 Advanced Placement Examination in Biology. He identified 17 [of the 40] items that displayed at least one significant negative partial association with another item at the $p < .001$ level. The balance of 23 items showed no negative partial associations among items. His analysis identified two item bundles (items 82 through 85 and items 86 through 88) as separate groups of items that shared common distractors which asked students to link biological terms and their definitions. There were also four other items (13, 14, 49, and 51) which showed significant negative partial associations ($p < .001$) but there were no obvious links among these four items except for relative exam position effect. Rosenbaum posed three questions for consideration.

> “(i). Is there any reasonable sense in which exam responses might be described as unidimensional despite some excessive dependence between small groups of items that share material?
> (ii). If such a notion of unidimensionality exits, what does it imply about observable item response distributions? In other words can we test this broader class of unidimensional models?
> (iii). In particular, how would we interpret the negative partial association between Items 13 and 14? These items do not share materials. Is it possible that this negative partial association is an indirect consequence of the link between Item 13 and the item bundle including Item 88 [the item bundle that includes items 86 to 88]? Or does the negative partial association between two items not in the same bundle indicate a violation of unidimensionality in the wider sense?”

Rosenbaum distinguishes between violations of test item unidimensionality and violations of latent conditional independence. When item bundles share common materials such as a reading passage or science diagram, there is a plausible rationale for conditional item dependence. However, some items may show statistical dependence without sharing any common materials. He notes, “There are many types of items which seem difficult when first attempted, but which seem to become somewhat easier with practice on similar items. Certainly one can construct mathematical word problems or verbal analogies that are so parallel in nature that the sharing of cognitive tasks is almost undeniable (Rosenbaum 1988, p. 358-359)”.

For the perspective of this paper the key theoretical notion introduced by Rosenbaum is that items can form “item bundles” and that psychometric approaches can be used to evaluate characteristics of the “item bundle” and relationships to other “item bundles” or to individual test items.

‘

Item Families, Item Clones and Computerized Adaptive Testing

Glas & van der Linden (2001, 2003) indicate that one major impediment to implementation of computerized adaptive testing (CAT) is the resources needed for item pool development to provide both content item structures and item parameter estimates that are needed for effective and efficient computerized adaptive testing. One of the solutions to this problem is item cloning to generate the required adaptive testing pools. Glas & van der Linden (2001, 2003) suggest two procedures that have been used for generating item clones. One procedure employs a syntactic description of test items with one or more open slots for which replacement option sets may be selected by computer algorithm (Millman and Westman, 1989). The second procedure is to modify parent items and generate cloned sibling items from the parent item by transformation rules. Glas & van der Linden, 2003 note, “examples of such rules are linguistic rules that transform one verbal item into others, geometric rules that present objects from a different angle in spatial ability tests, chemical rules that derive molecular structure from a given structure in tests of organic chemistry, or rules from propositional logic that transform items in analytic reasoning tests into a set of new items.” (Glas & van der Linden, 2003, p. 247).

Glas & van der Linden note that pioneers in the concepts of item cloning included Bormuth (1970), Hively, Patterson and Page (1968), and Osburn (1968). “Common to their approaches is a formal description of a set of ‘parent items’ along with algorithms to derive families of clones from them. These parents are known as ‘item forms,’ ‘item templates,’ or ‘item shells’.
Glas & van der Linden, 2003, p. 247).”

Glas & van der Linden introduce the notions of creating item pools with families of items generated from parents $p = 1,\ldots\ldots.P$. Items within family p will be labeled $i_p = 1,\ldots..I_{p.}$ They use a two stage procedure for adaptive item selection where a family of items is selected that is optimal at the current person proficiency estimate and then an item is randomly sampled from the item family and administered. Items within families are modeled by a three-parameter logistic (3PL) model and the parameters of items within families are modeled by a (joint) distribution that addresses variability within families (Glas & van der Linden, 2003, p. 248).

Their simulation results indicated the value of modeling the family structures of cloned items with the multi-level IRT model with family specified parameter distributions. “It is a statistical fact that ignoring the family structure of the items in the pool is a case of model misspecification, which generally leads to bias in parameter estimation and hence to an increase in the mean absolute estimation error. In the simulation studies, the multilevel IRT model did suffer from this type of bias, but the effects were very small….If all variability in the pool is within the families, the procedure is domain-referenced testing, whereas if all variability is between families, it is CAT from a pool of individually calibrated items (Glas & van der Linden, p. 260).”

Item Families and Family Response Functions

Sinharay, Johnson & Williamson (2003) and Johnson and Sinharay (2005) recommend the investigation of item families/family response functions. Sinharay, Johnson & Williamson (2003) introduced the Family Expected Response Function (FERF) as a way to summarize probabilities of a correct response to an item randomly drawn from an item family. The calibration of item families allows for generation of items on the fly from the family structure. Examinees can also be scored on their performances with new, unscaled items drawn from the defined family structure. Bejar, Lawless, Morley, Wagner, Bennett & Revuelta (2002) also discuss the use of an expected response function for linear-on-the-fly adaptive testing.

Johnson & Sinharay (2005) and Williamson, Johnson, Sinharay, & Bejar (2002) suggest the three approaches for modeling data involving item families using IRT models for either dichotomous or polytomous items: the unrelated siblings model, the identical sibling model, and the related sibling model each of which are briefly summarized below.

Unrelated Siblings Model.

The unrelated siblings model (USM) assumes that the items are mutually independent and each item in the pool or model is calibrated.

Identical Siblings Model.

The identical siblings model (ISM) assumes that the item parameters are the same for all items within the same family. Depending on the degree of variation between sibling items the model provides biased or over confident estimates of examinee scores.

Related Siblings Model.

The related siblings model (RSM) uses a hierarchical model with a separate item response function per item at the lower level and a higher level model that relates the item parameters for each family. Johnson & Sinharary (2005) recommend use of the related siblings model (RSM) to calibrate item families and also address the variability of sibling items within families. The paper graphically compares eleven estimated family response and score functions.

Cluster-and Item Bundling Models

Ellen Boekkooi-Timminga (1990) suggested the use of a cluster-based method for test construction where items within the bank were grouped together based on item information functions and the group clustered items with similar information functions were considered equivalent.

Wilson & Adams (1995) recommend the use of Rasch models for item bundles where the clusters or bundles of test items are identified by "common stimulus materials, common item stems, common item structures, or common item content, such that one might be led to doubt that the usual assumption of conditional independence between items would be an appropriate one to make (Wilson & Adams, 1995, p. 181)."

Testlet Models

In 1987, Wainer and Keily defined a testlet as the aggregation of a packet of test items that are administered together (as a mini test). Testlets provide a way of addressing problems of cross-information from one item to another, unbalanced contexts by controlling presentations of test items that are congruent with test specifications, and providing common item order effects. Testlets can be used for modeling and analysis of item groups that share a common reading passage, a common graphic picture or chart, item groups that do not exhibit conditional independence or other types of departures from standard unidimensional IRT models and assumptions. Testlets provide one method for analyzing and modeling data from item clusters or item families.

Wainer, Bradlow and Wang's book on *Testlet Response Theory and its Applications* (2007) provides multiple measurement models among others for testlet data involving two parameter logistic (2PL), three parameter logistic (3PL) and Bayesian testlet models for analyzing mixtures of dichotomous and polytomous results. The testlet contribution for each of these models is accounted for by using an additional testlet parameter in the standard IRT model parameter estimation. In a 2PL model that is used for analyzing testlet data, an additional third testlet adjustment parameter is estimated in the calibration or scoring process. Likewise, if a 3PL model is used for analyzing testlet data, an additional fourth testlet adjustment parameter is estimated in the calibration or scoring process. Testlet Response Theory provides an alternate approach to assessing item families, clusters, bundles or testlets.

Assessment Engineering

Richard Luecht (2009, 2007, 2006a, 2006b, Luecht, Gierl, Tan & Huff, 2006), has recommended the assessment engineering approach to constructing tests. The assessment engineering approach uses task models and templates to generate structured classes of comparable test items. The items developed with the task models or templates inherit the estimated psychometric characteristics from the task model or templates from which they were selected.

Luecht, Burke & Devore's (2009) initial validation of assessment engineering employs task models that define different regions of the construct-referenced domain measurement scales. The defined task models are positioned along the measurement scale by the joint combination of complexity and difficulty indices. Each task model has multiple templates and each of the associated templates can be used to generate multiple tasks or items. They note, “This provides enormous efficiencies to treat either the task model as a family of templates or a template as a family of items, and lends itself nicely to a hierarchical IRT calibration system (e.g., Glas and van der Linden, 2003) (Luecht, Burke, & Devore, 2009, pp. 4-5).”

Assessment engineering provides an additional linked concept for the analysis of item families, clusters and bundles.

Automated Item Generation

Sandene, et. al., (2005) investigated the use of automated item generation approaches for the NAEP Online Assessment in Mathematics. They note, “Automatic item generation rests on two premises. The first premise is that a class of test items can be described in enough detail for a computer to generate instances of that class. The second is that enough can be known about the determinants of item difficulty so that each of the generated instances does not have to be individually calibrated. The description the computer uses to generate instances of a class is called an item ‘model’ and the instances are called ‘variants.’ ”

The solution to the first premise of automated item generation has been documented in Irvine and Kyllonen’s (2002) book. They demonstrated over a wide range of content that item classes can be sufficiently defined that computers can generate variant or isomorph items. Using automated item generation ETS has used the Mathematics Test Creation Assistant (Singley & Bennett, 2002) for research concerning limited item generation in selected large scale testing programs. The second premise of automated item generation concerns the appropriateness of the calibration of the item model (item family) and verification that the item model parameterization structure can be imputed to the variants (isomorphs or sibling) so that calibration of the variants is not required.

Sandene, et. al., (2005) suggest two approaches for generating comparable item sets. “The calibration can be accomplished by basing the model on an empirically calibrated item and then constraining the model so that it, ideally, produces variants that diverge little in substance and psychometric properties from the original “parent” question. Variants that preserve the underlying problem structure are termed “isomorphs.” Because the variants created by a model are not only isomorphs of one another, but also isomorphs of the parent item, the model’s parameters may, in theory, be imputed from those of the parent. A second calibration method is to pretest a sample of variants from the item model and use that information to establish model parameters (Sandene, et. al., 2005, p 24).”

An empirical investigation of item generation and item family calibration was conducted using the ETS Mathematics Test Creation Assistant for a research project with the NAEP Mathematics Assessment. Item models and isomorphs were created for 15 of 26 items in the Eighth grade paper and pencil forms. Test development specialists reviewed isomorphs that were generated and removed any items that might disadvantage one or more demographic groups. The isomorphs were then randomly selected to be included on Form A, Form B or on Form P&P. The isomorphs were in the same test serial position on all forms. Automatically generated items were also present in each of the test sections and alternate forms.

The three alternate forms were administered to samples of eighth grade students, 954 students received form P&P, 926 students received Form A, and 906 students received Form B. Scaling for the three test forms was conducted using a 20-item anchor test. Within each test form the item parameterization was not constrained. This allowed for comparisons of the item parameters for the 11 items that were common to the three test forms and the 14 sets of isomorphs. Due to scaling problems one isomorph set of items was not analyzed.

Separate plots were analyzed of the pairwise comparisons of the IRT *b* parameters for the 11 identical items on the three grade 8 paper forms and the pairwise comparisons of the IRT b parameters for the 14 isomorphs on the three grade 8 paper forms. The pairwise comparison of parameters for the identical items were fairly well clustered around the diagonal slope but some identical items showed departures from the expected diagonal slope. The plot of the isomorphs showed more variation from the diagonal but many of the isomorphs were still very close to the expected diagonal slope. Mean differences in parameter estimates and mean absolute differences in parameter estimates were computed between the parameter estimates for the identical items and the isomorphs. Pearson correlations between the parameter estimates for the identical items were 0.97, 1.00 and 0.97 for the three pairwise test form results and .80, .85 and .98 for the three isomorph pairwise test form results. As expected the isomorph calibrations were less congruent than the calibrations of the identical items. In their executive summary, Sandene, et. al., (2005) state, "The degree to which the item-parameter estimates from one automatically generated item could be used for related automatically generated items was also investigated. Results suggested that, while the item-parameter estimates varied more than would be expected from chance alone, this added variation would have no statistically significant impact on NAEP scale scores (Sandene, et. al., 2005, p. viii)."

Automated item generation also provides a contributing thread toward the analysis of item families or groups.

In summary, the literature review has shown a variety of related approaches have been used to explore the potential benefit or value for item task modeling with groups of items rather than with individual test items or tasks.

METHODS

The paper presents a theoretical and practical approach for using item and task modeling and analysis. We propose that item and task modeling and analysis will move the educational measurement profession forward in a very significant and meaningful way. This paper provides background theory, testing applications and analysis approaches for generalizing the classical or IRT item difficulty, discrimination, and model misfit parameters (or subsets) by using the concepts of item families, clusters, parcels, bundles or testlets. Test scores are accumulated scores or IRT proficiency estimates over a series of test items or performance tasks. Item and test statistics and parameters can be computed and reported at multiple levels of aggregation. The estimated item difficulties, discriminations, and model misfit parameters and associated parameter standard errors could apply to any child/sibling item or task appropriately and systematically selected from the item/task model. The paper presents alternative theoretical and practical perspectives on the problem and potential solutions.

Test Data

The data employed for the modeling and analysis is from an information technology credentialing and certification test for designing computer databases. The credentialing and certification test included two alternate forms A and B with fifty items each. The test includes a total of nine scenarios case clusters coded as Cases A(11), B(11), C(10), D(12), E(9), F(10), G(14), H(5) and I(7). The parenthesis after the case indicates the total number of items from the item bank included in that scenario case. The nine case scenario groups include between 5 and 14 items from the item pool. The case clusters are based on a database scenario that forms the measurement context for the items. Six of the scenario cases are administered within each test form. The scenario case structure for each of the two test forms and the numbers of items per case are presented below.

Cases and (Items) Per Test Form
Form A: B(6), C(10), E(7), F(10), G(12), and H(5)
Form B: A(11), B(11), D(11), E(4), G(5), and I(7)

The parenthesis after the case indicates the total number of items from the item bank included in that scenario case on each test form. Form A was administered to 630 individuals and Form B was administered to 640 individuals.

There were eleven (11) anchor items within each form. These eleven anchor items were drawn from three of the scenario cases. The anchor items included six (6) items from Case B, two (2) items from Case E and three (3) items from Case G. The anchor items were administered to a total of 1270 individuals.

The anchor items had a *p* value range from 0.74 to 0.94 (mean 0.87, St. Dev. 0.078), point biserial correlations for the anchor items ranged from 0.40 to 0.62 (Mean 0.50, St. Dev. 0.069), and Rasch measures for the anchor items ranged from -1.38 to +1.01 (Mean -0.39, St. Dev. 0.902). Infit and outfit Rasch statistics for the anchor items were less than 1.17 Infit Mean Square Max and 1.35 Outfit Mean Square Max.

ANALYSIS

The test data set was analyzed and modeled using classical test theory, Rasch, Master's partial credit and confirmatory factor analysis. Results are provided from the analysis of the items as individual items and the analysis of the items within the scenario case clusters. The analysis for classical test theory was conducted using standard analysis techniques implemented in Excel spreadsheets. The item response theory modeling was conducted using Winsteps for the Rasch and Master's partial credit analyses. Two parameter item calibrations and confirmatory factor analysis were conducted using MPlus.

RESULTS

Results for the classical item and test analysis are presented first. The Master's partial credit Rasch measurement results are provided and the final analyses presented are for the confirmatory factor analysis.

As shown in Table 1 the mean scores for the two forms A and B were 43.31 and 40.71 respectively showing that form A was relatively easier than form B. The median scores for the two forms were 47.0 and 43.0 points again showing form A was meaningfully easier than Form B, assuming random group equivalence. Figure 1 shows the reverse cumulative distribution (passing rates at each possible score value) for the two forms. Figure 1 shows that Form A is slightly easier than Form B for the upper portion of the score distribution. Figure 2 and 3 show the test characteristic curves and the test information curves for the two test forms. These results show that form A is relatively easier and provides more information than form B particularly in the center of the ability distribution. Form A provides less information at the top end of the distribution. The standard deviations of the scores for Form A and B were 8.40 and 8.68 respectively and the standard errors of the mean for Form A and B were 0.33 and 0.34 points. The Cronbach's alpha reliabilities for Form A and B were 0.939 and 0.928 respectively. The standard error of measurement on the two forms A and B was 2.07 and 2.33 points respectively. With the exception of the mean scores and passing rates the classical test theory statistics for the two test forms and items comprising the forms are comparable. The two test forms show very acceptable internal consistency as measured by Cronbach's alpha reliability.

Table 1. Test Form Statistical Information

Statistic	Form A	Form B
Examinee Count	631	639
Exam Length	50	50
Mean	43.31	40.71
SD	8.40	8.68
Median	47.0	43.0
Mode	49	49
Avg. Time on Test	45.8	51.6
SD of Time on Test	33.4	32.9
Standard Error of the Mean	0.33	0.34
95% confidence interval +/-	0.66	0.67
Minimum	0	0
Maximum	50	50
Alpha Reliability	0.939	0.928
SEM	2.07	2.33
95% confidence interval +/-	4.06	4.57

Figure 1. Pass Rates for the alternative test forms

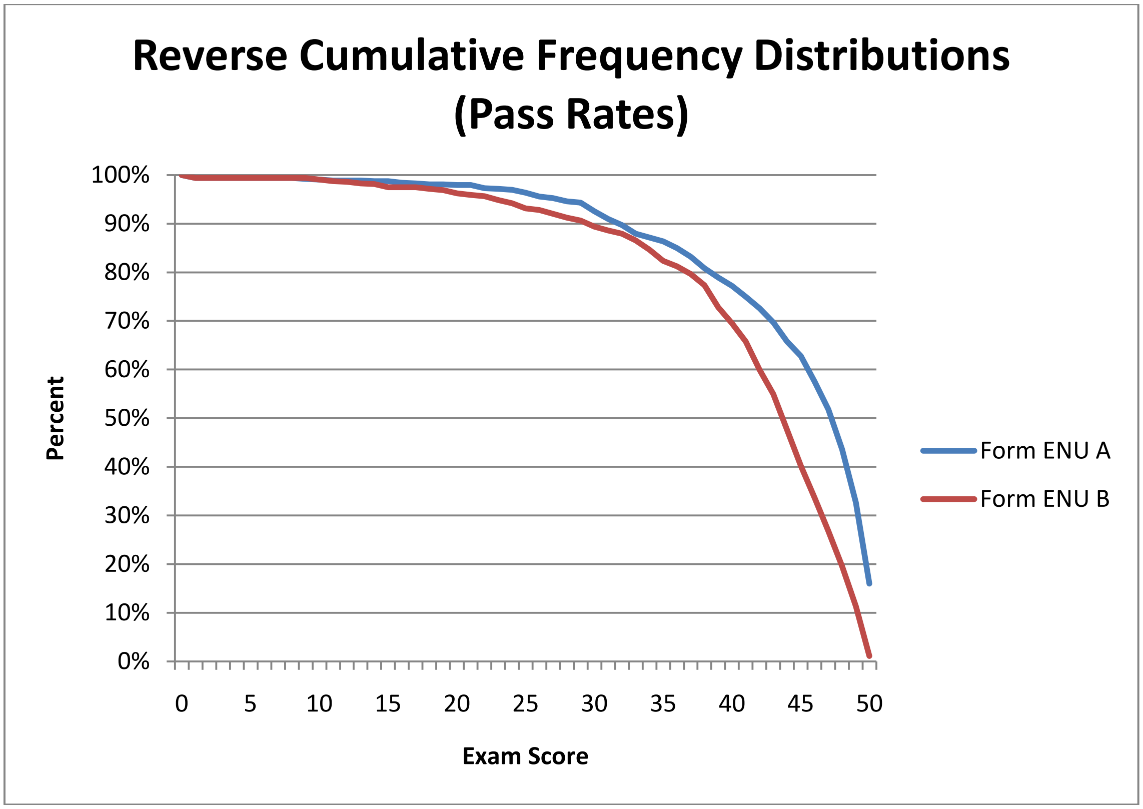


Figure 2. Test Characteristic Curves

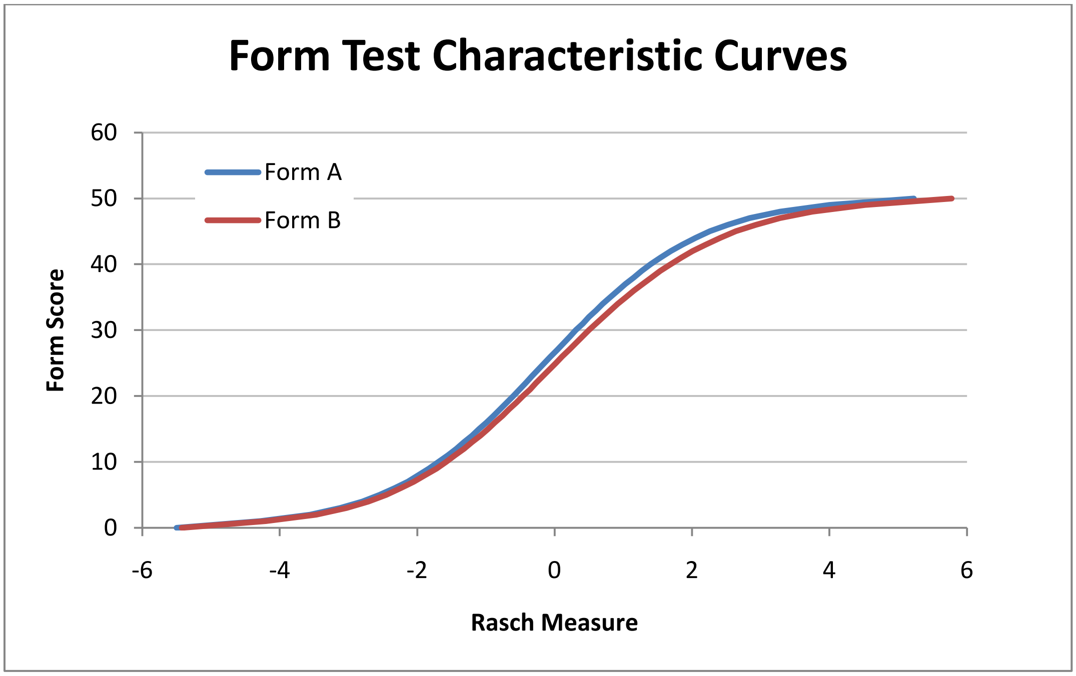
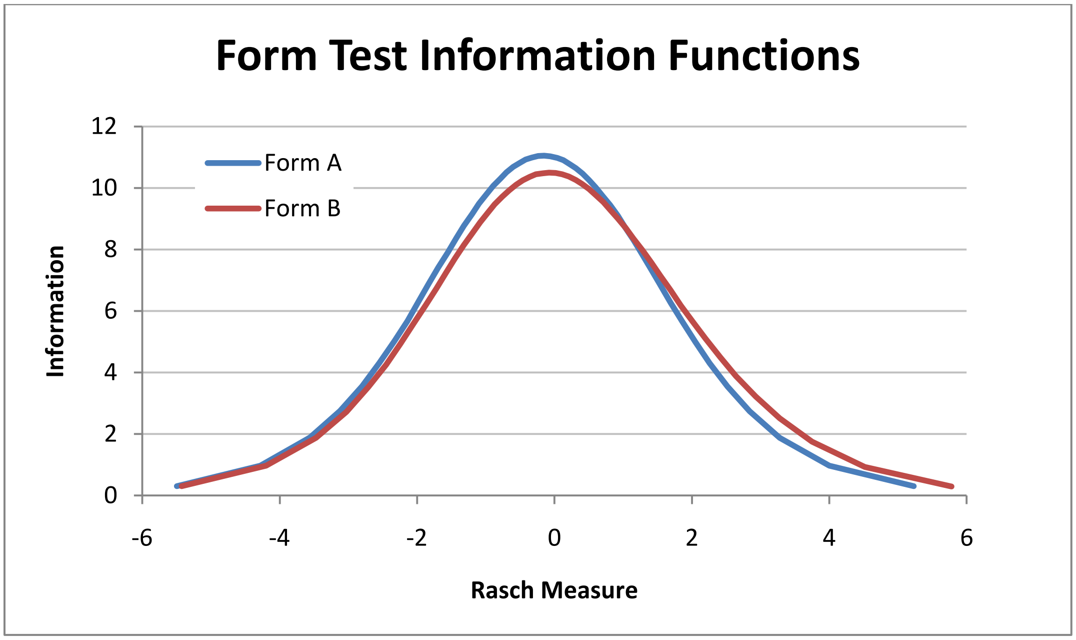


Figure 3. Form Test Information Functions



Classical Item Analysis and Rasch Measures at the Item Level

Table 2 in Appendix A presents classical item analysis results and Rasch Measures at the item level for all eighty-nine items within Form A and Form B. Table 2 includes an item sequence number, the case classification, the item ID, Rasch statistics include the Rasch Measure, Infit Mean Square, Outfit Mean Square, classical item analysis results are given for the p-values, point biserials (item-score correlations), and item reliability which is a function of the p values and the point biserial correlations, the number of examinee responses, and the average number of seconds for the candidate response. Items are flagged in red when the item is too easy or too difficult as indicated by the *p* value statistic and not correlating with the total score as indicated by the low or negative point biserial correlations. The test and item analysis typically includes answer option analysis, the Rasch item file, the Rasch person item map, and the Rasch person file.

Analysis of Test Items by Scenario Case using Master's partial credit analysis

The items within each scenario group were analyzed with Master's partial credit Rasch analysis at the group level. These results are presented in Table 3. Table 3 presents the minimum, maximum and mean and standard deviations for the p values, the point biserial correlations, and the Rasch item level measures. The table also shows the Master's partial credit Rasch analysis for each of the scenario case clusters. Statistics for the Group Rasch analysis include the Item Group measure and the point measure correlations for the case clusters. Cases highlighted in yellow show fairly comparable means for Rasch measurement at the item level and at the group level using the Master’s partial credit analysis. This analysis indicates that for the cases C, E, F and H the Rasch analysis at the individual item level and the Rasch analysis at the case group level resulted in similar mean Rasch Measures. However, each of these cases has wide variance in the minimum and maximum Rasch values at the item level but have comparable mean scores when the item level Rasch mean is compared to the Item Group Rasch mean measure from the partial credit analysis.

Table 3. Item Analysis, Rasch Measurement and Item Group Measurement for Cases

	PValue			Pt. Biserial			Rasch Measure			Item Group Rasch	
CASE	Min	Max	Mean(SD)	Min	Max	Mean(SD)	Min	Max	Mean(SD)	Measure	Pt Meas
A(11)	.78	.92	.86 (.04)	.43	.63	.50 (.06)	-0.94	0.11	-0.20 (0.47)	-0.66	0.85
B(11)	.50	.93	.81(.14)	.32	.57	.42 (.07)	-1.09	2.25	-0.60 (0.09)	0.10	0.17
C(10)	.74	.88	.83 (.05)	.44	.67	.57 (.07)	-0.23	0.92	0.37 (0.47)	0.37	0.76
D(12)	.74	.94	.84 (.07)	.36	.64	.51 (.08)	-1.43	0.84	-0.11 (0.77)	-0.55	0.86
E(9)	.34	.94	.77 (.19)	.05	.62	.45 (.16)	-1.06	3.15	0.50 (1.34)	0.59	0.70
F(10)	.76	.92	.87 (.06)	.27	.62	.48 (.10)	-0.84	1.00	-0.18 (0.68)	0.04	0.79
G(14)	.74	.95	.87 (.12)	.33	.61	.45 (.09)	-1.42	2.16	-0.36 (1.01)	0.40	0.63
H (5)	.79	.87	.83 (.03)	.47	.58	.54 (.05)	-0.01	0.79	0.39 (0.30)	0.24	0.69
I (7)	.71	.92	.88 (.63)	.44	.66	.56 (.07)	-0.99	1.05	-0.10 (0.63)	-0.53	0.82

Table 4 provides the Master’s partial credit Rasch item file for the Group measurement of the cases with statistics for the case identification (number of items per total case), entry into the analysis (the entry numbers have been reordered to correspond to the alphabetical listing of the cases), Rasch partial credit measure, count of examinees, score for all of the examinees per case, standard error, the Infit Mean Square, Infit Mean Standardized Infit Z statistic, the Outfit Mean Square and the Standardized Outfit Z statistic, the point measure correlation, and the estimated item discrimination for the case cluster. Figures 4-12 (one for each case group A-I respectively) provide scenario case item response curves with both estimated and empirical data representations. In Figures 4 to 12 the periods and x’s represent the estimated item response function and the solid line gives the modeled empirical item response functions. The item response functions also show 95% confidence intervals for the expected response functions.

The case response functions that show acceptable fit between the empirical and modeled item response functions for six cases include Case A (Figure 4), Case D (Figure 7), Case E (Figure 8), Case F (Figure 9), Case G (Figure 10), and Case I (Figure 12). The case response functions that show non-acceptable fit between the empirical and estimated item response functions for three cases include Case B (Figure 5), Case C (Figure 6), and Case H (Figure 11).

Figure 3a shows the Test Characteristic Curves for the alternate forms as measured with the Master's Partial Credit Analysis. This analysis shows that Form A is slightly easier than Form B in the middle to upper of the ability scale.

Figure 3a Test Characteristic Curves for Master's Partial Credit Analysis

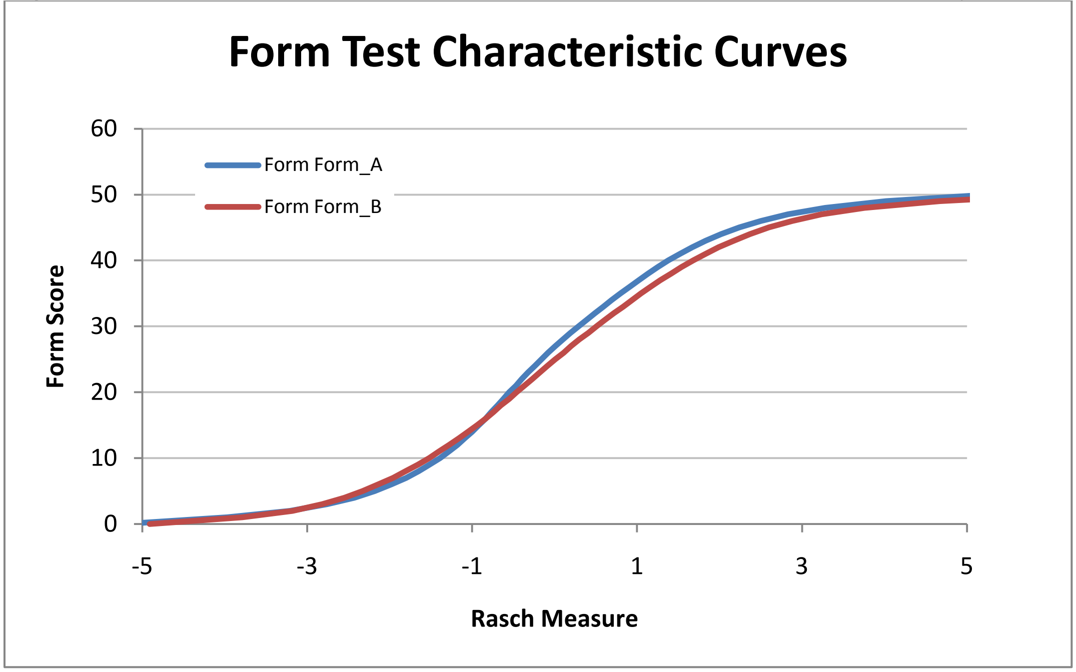


Figure 3b shows the Test Information Function Analysis for the alternate forms as measured with the Master's Partial Credit Analysis. The Test Information Functions shows that Test Form A is more informative than Test Form B at the center of the test score scale.

Figure 3b Test Information Functions for Master's Partial Credit Analysis

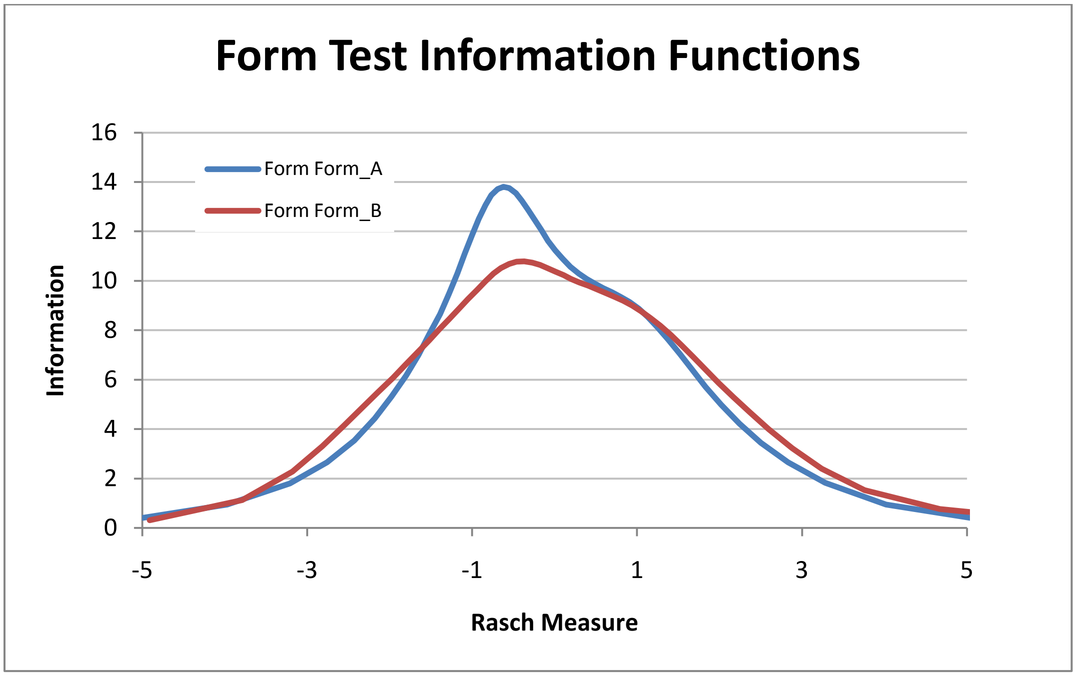


Figure 4. Empirical vs. Model ICC for Case A

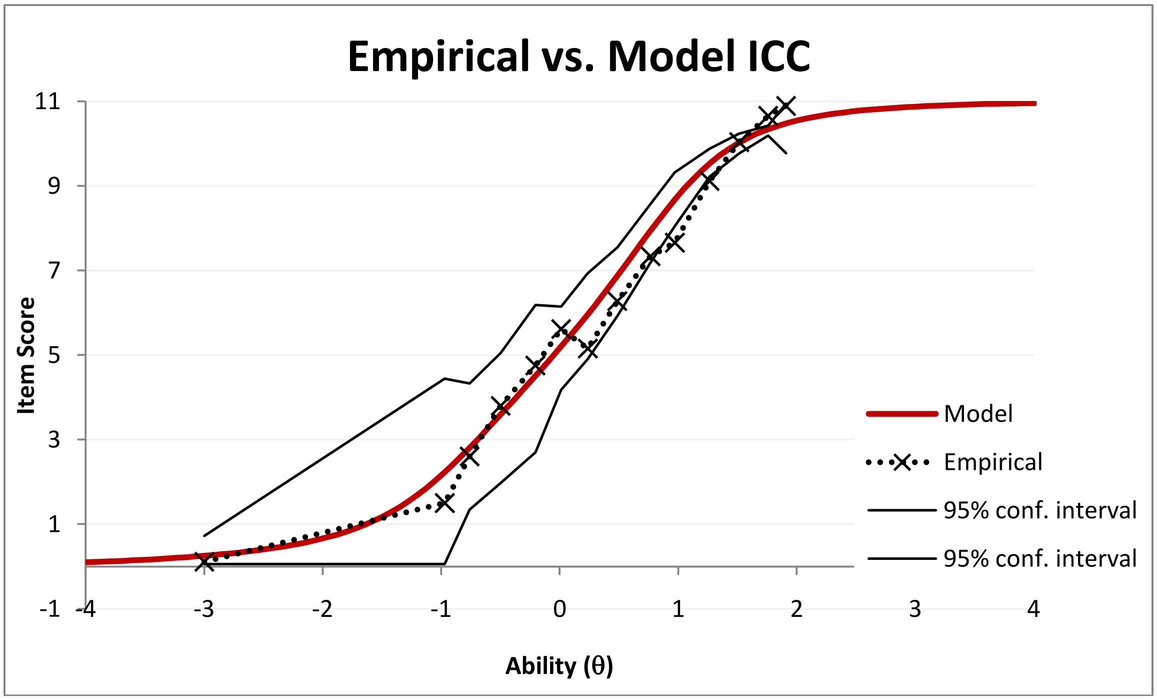


Figure 5. Empirical vs. Model ICC for Case B

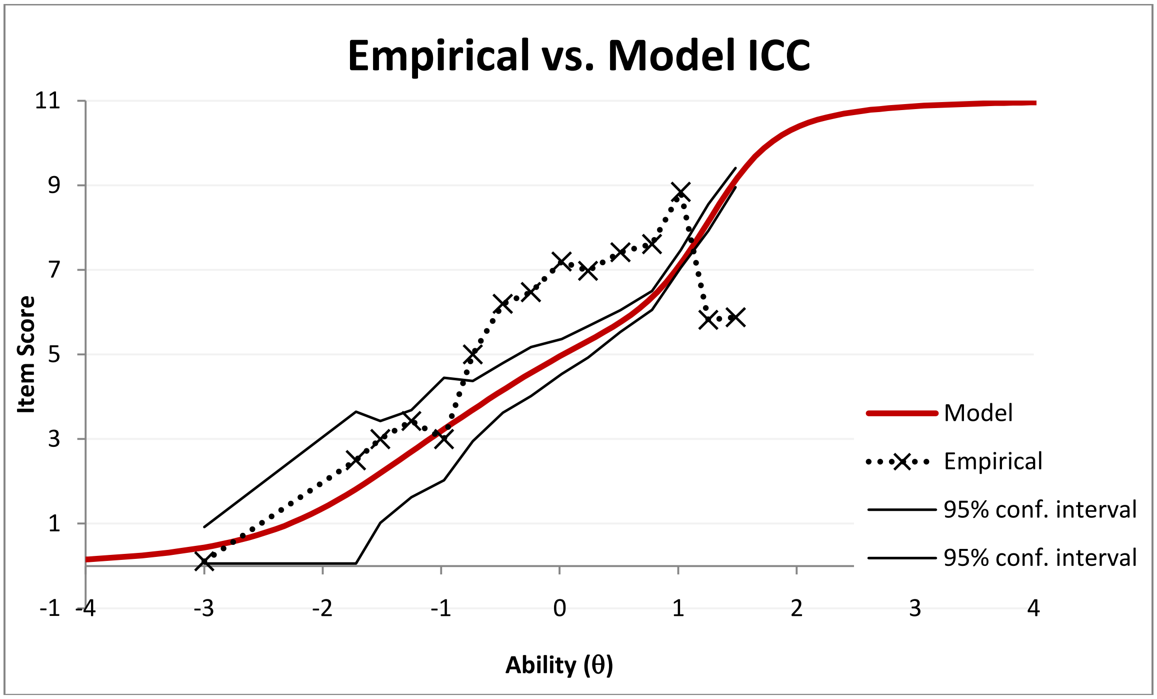


Figure 6. Empirical vs. Model ICC for Case C

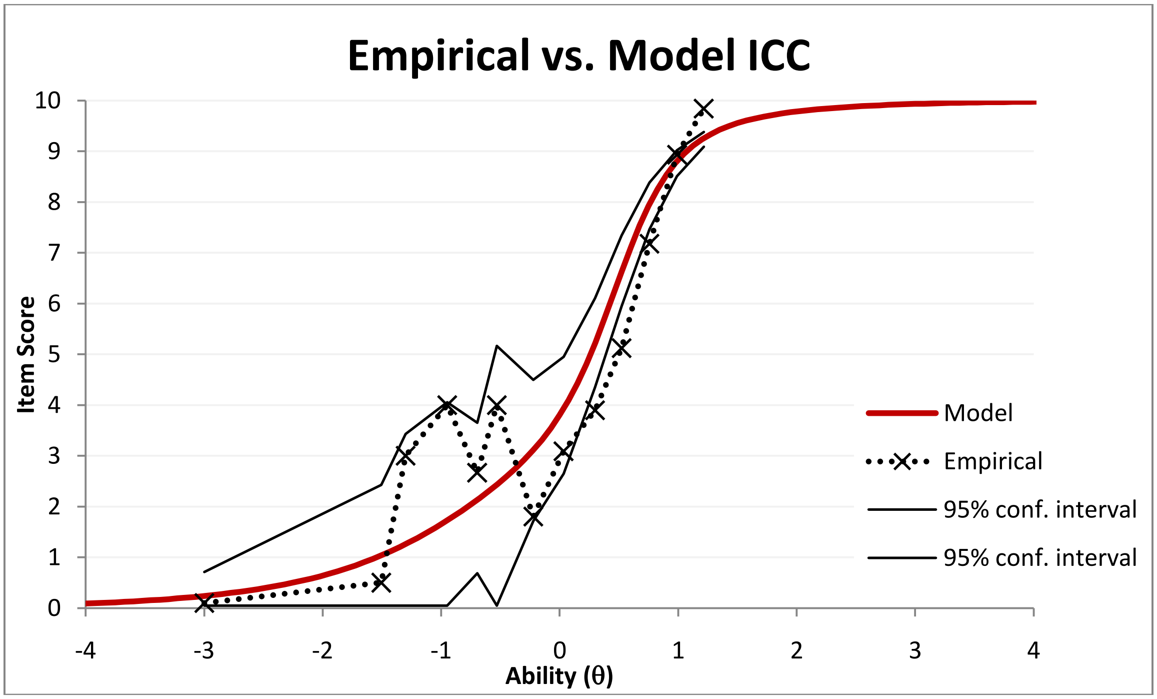


Figure 7. Empirical vs. Model ICC for Case D

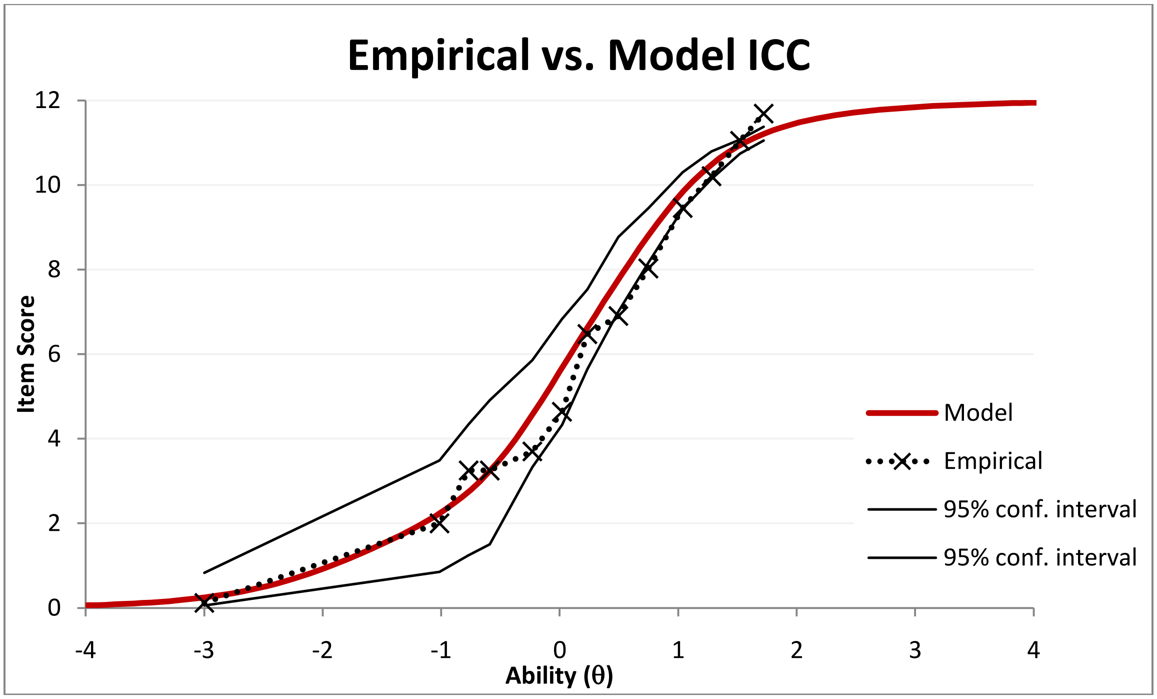


Figure 8. Empirical vs. Model ICC for Case E

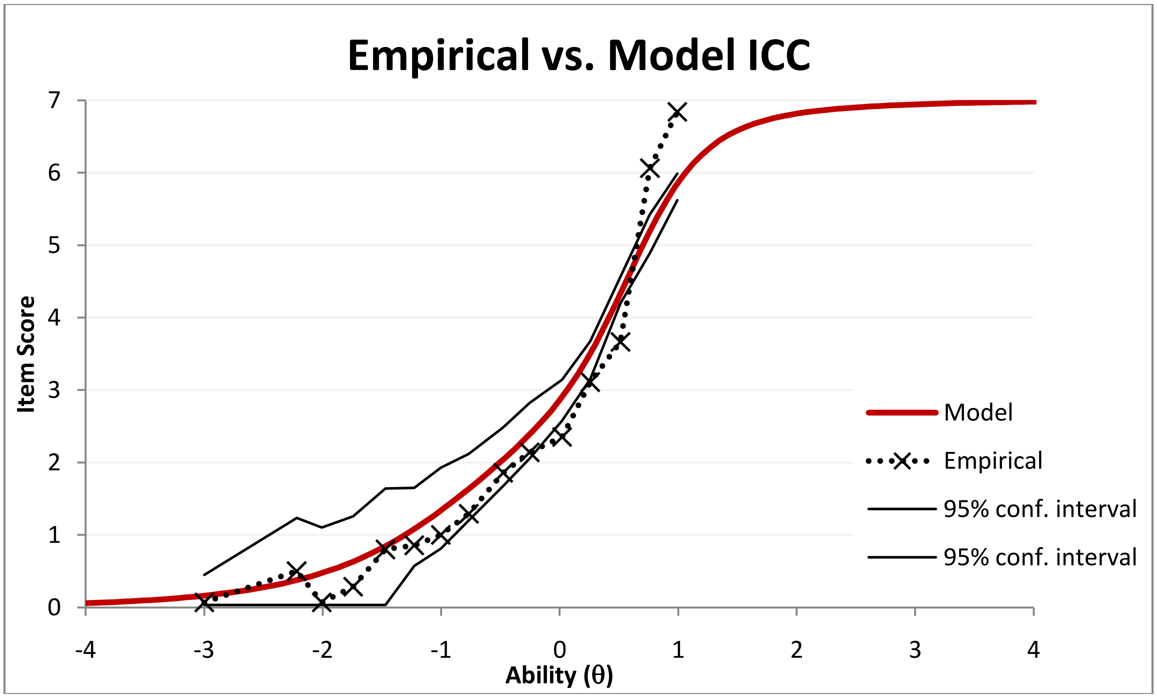


Figure 9. Empirical vs. Model ICC for Case F

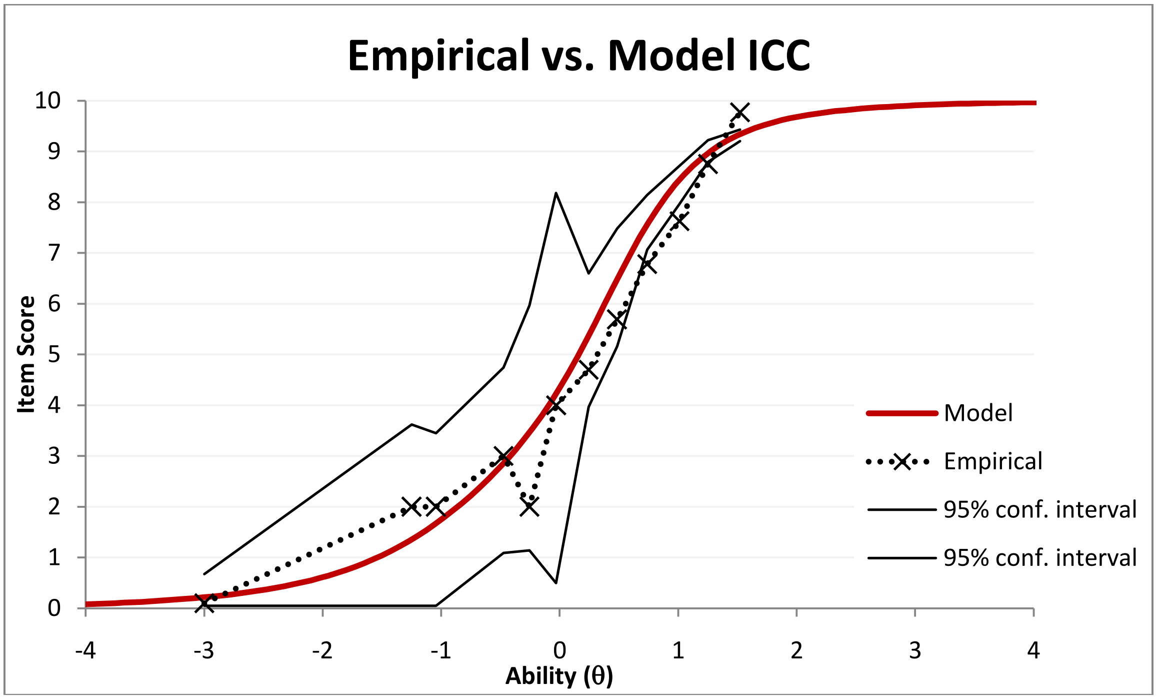


Figure 10. Empirical vs. Model ICC for Case G

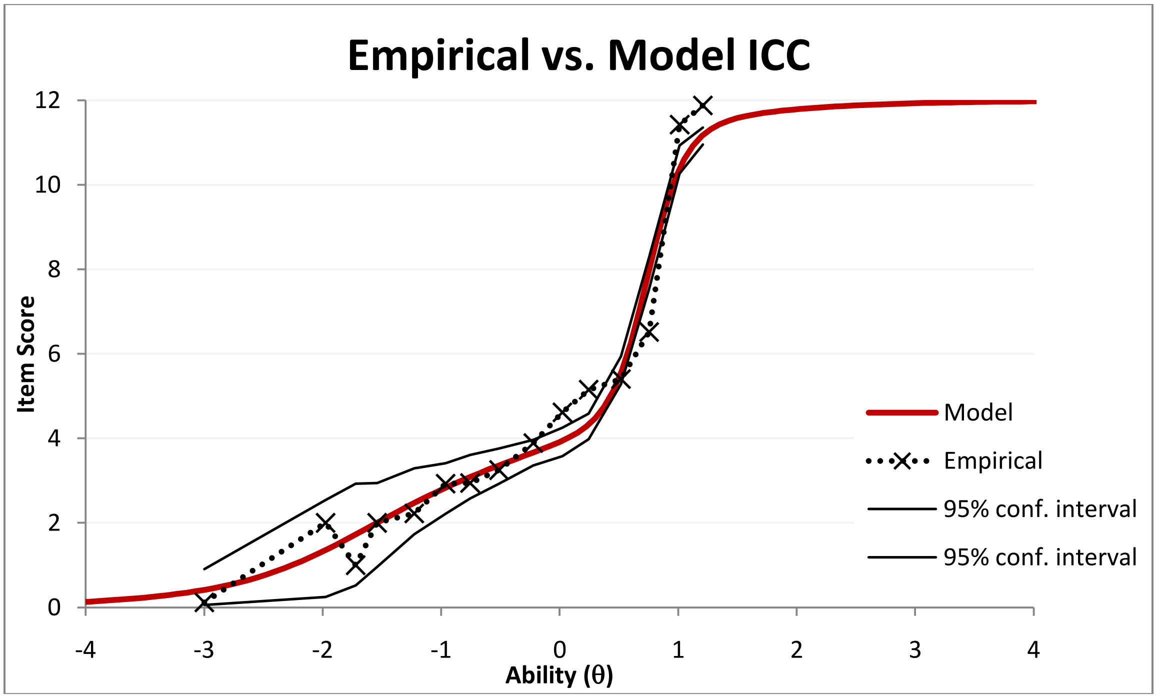


Figure 11. Empirical vs. Model ICC for Case H

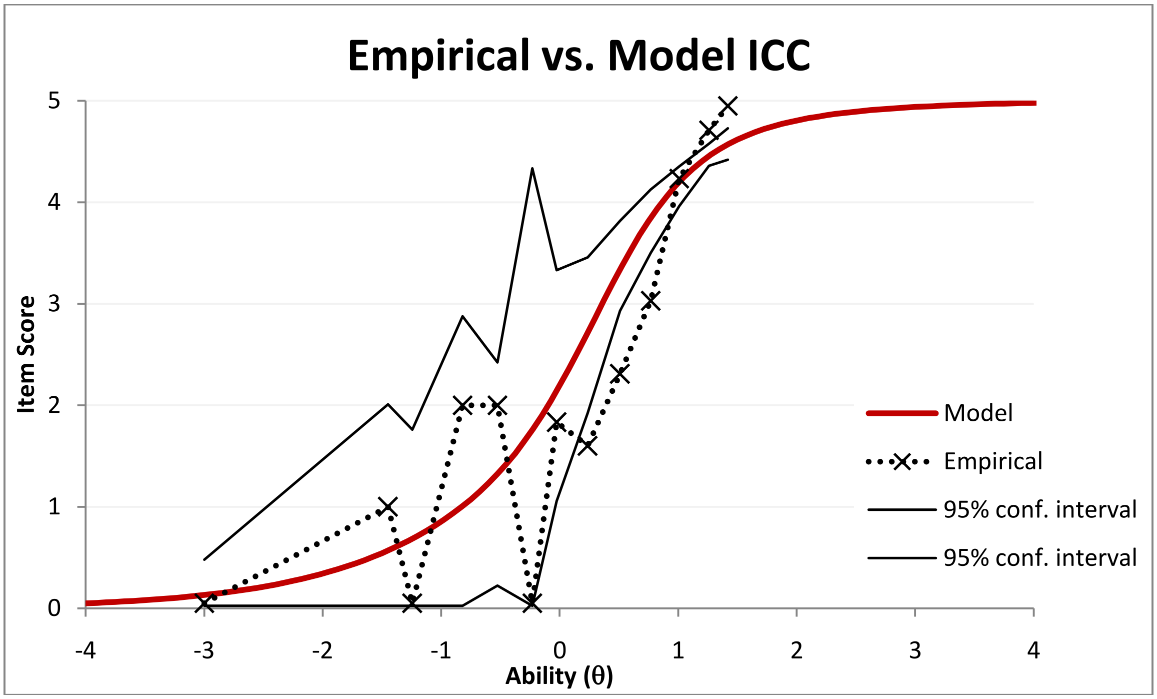


Figure 12. Empirical vs. Model ICC for Case I

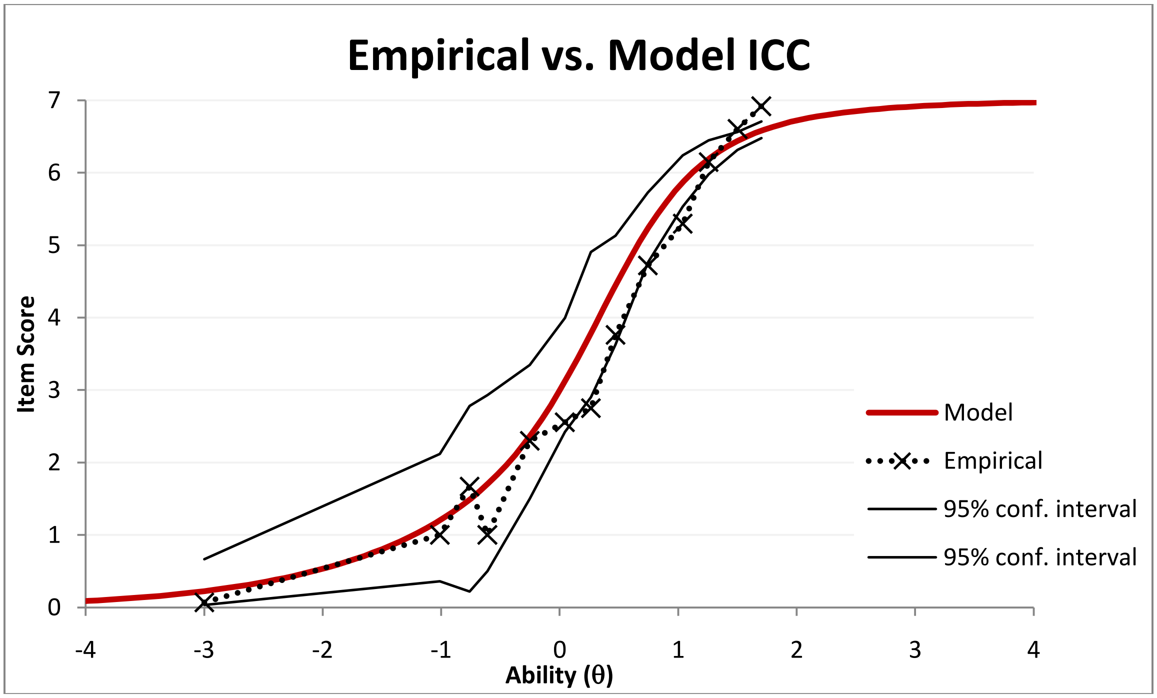


Table 4 shows that the Case Cluster Rasch measures range between a low theta value of -0.66 to a high theta value of 0.59. Table 5 shows the minimum and maximum Rasch measures of the individual test items within each case group. For comparison the Rasch Measures for the individual items range from a low theta value of -1.43 to a high theta value of +3.15.

The Infit and Outfit Mean Squares for Case B show large values indicating that that Case B was not well fit with the partial credit analysis. The significant negative standardized z tests for the infit and outfit statistics may be due to the small number of cluster items which are being measured (nine in total). Table 4 also shows strong point measure correlations for all cases except for Case B which also showed significant Mean Squares and standardized z tests for the Infit and Outfit statistics. Case B has a point measure correlation of only 0.17. The other cases have point measure correlations ranging from a low of 0.63 for Case F to a high of 0.86 for Case D. Table 4 also shows a zero estimated item discrimination value for Case B. All other Cases had estimated item discrimination indices greater than 1.00 with a range from 1.07 for Case G to a high of 1.38 for Case E. The Rasch estimated item discrimination index greater than 1.00 indicates that these Cases have a steeper item response slope (discrimination) than expected.

Table 4. Partial Credit Rasch Item File for Scenario Cases

CASE	ENTRY	MEASURE	COUNT	SCORE	ERROR	IN.MSQ	IN.ZSTD	OUT.MSQ	OUT.ZSTD	PTME	DISCRIM
A(11)	4	-0.66	639	6013	0.03	0.66	-5.56	0.59	-6.19	0.85	1.20
B(11)	1	0.10	1270	9013	0.02	2.53	9.90	2.76	9.90	0.17	-0.07
C(10)	3	0.37	631	5217	0.02	0.47	-9.37	0.40	-8.92	0.76	1.23
D(12)	5	-0.55	639	6434	0.03	0.58	-7.34	0.56	-7.40	0.86	1.29
E(9)	6	0.59	1270	5350	0.02	0.55	-9.90	0.58	-9.90	0.70	1.38
F(10)	8	0.04	631	5505	0.03	0.64	-5.53	0.55	-6.67	0.79	1.20
G(14)	2	0.40	1270	9424	0.01	0.83	-4.57	0.99	-0.19	0.63	1.07
H(5)	9	0.24	631	2613	0.04	0.66	-5.45	0.54	-6.73	0.69	1.19
I(7)	7	-0.53	639	3771	0.04	0.59	-6.49	0.52	-6.78	0.82	1.20

Table 5. Minimum and Maximum Individual item Rasch Measures by Case

Case	A	B	C	D	E	F	G	H	I
Min	-0.94	-1.15	-0.23	-1.43	-1.13	-0.84	-1.42	-0.01	-0.99
Max	+0.52	+2.25	+1.14	+0.84	+3.15	+1.00	+2.16	+0.79	+1.05

The summary Rasch measurement table for the cases indicated that the minimum partial credit exam score was 8 points and the maximum partial credit score was 50 points. The mean case based test score was 42.3 indicating a skewed distribution to the top end of the measurement scale. The standard errors were 0.19 to 0.39 for the test scores. The person to raw score to measure correlation was 0.90 for the analysis. The Rasch measure for the nine cases ranged from -0.66 as a minimum to +0.59 as a maximum with a mean theta score of 0.0 and a standard deviation of 0.44. The standard errors for all of the modeled cases were 0.01 to 0.04. The root mean squared errors were 0.03 for the Cases and 0.29 to 0.31 for the measured Persons indicating fairly accurate measurement for persons and cases.

Confirmatory Factor Analysis and Dimensionality

Principal Components Analysis.

Dimensionality of the test was investigated with principal components analysis using SPSS and confirmatory factor analysis using MPlus. Principal components analysis was conducted for each test form shown in Table 6. The scree plots and a comparison of eigenvalues (λ) indicated that a maximum of three principal components could be extracted for each test form.

Table 6. Summary of Principal Components Eigenvalues and Ratios

	Scree Plot	λ 1 (cumvar)	λ 2 (cumvar)	λ 3 (cumvar)	Ratio λ 1 to λ 2	Ratio λ 2 to λ 3	Ratio Remaining λ
Form A	3	13.07 (26.16%)	2.943 (32.04%)	1.694 (35.43%)	4.4	1.74	1.30
Form B	3	12.38 (24.75%)	2.96 (29.87%)	2.40 (34.66%)	4.2	1.3	1.6

The interpretation of the components for Form A showed that the third principal component was primarily measuring items from Case G since five of the seven of the significant component loadings were from that case. The second component for Form A had twelve of 50 items with significant loadings and three strongly positive loadings, one from Case B (0.443), one from Case F (.587) and one from Case G (0.457). The second component was addressing some elements of several of the cases but not across all cases or consistently strong across all cases. For Form A the first extracted component had significant loadings for each of the 50 items.

For Form B, the third component had seven of 50 component significant loadings and four of these seven were from Case D and two component loadings that were moderately negative (i.e., component adjustment values). The second component appears to be addressing primarily Case B and two loadings from Case G. For Form B the first extracted component had significant loadings for each of the 50 items.

For both Form A and Form B the results from the eigenvalue analysis, percent variance extracted, the scree plot and the substantive interpretation suggest that there is a single component factor that is influencing the full measurement scale.

Confirmatory Factor Analysis

A confirmatory factor analysis approach was conducted with the case clustered items using MPlus. The confirmatory factor analysis approach was conducted separately by form since each test form had a different mix of cases and number of items per case. The test form data matrices were analyzed with and without the case structure being specified. The hypothesis tested is that a single latent measurement dimension is underlying the scores within each test form. Without a case structure being identified a two-parameter IRT normal ogive model with a single latent dimension model was estimated with weighted least squares with mean and variance corrections. The IRT normal ogive model provides model estimates (loadings) for the one latent factor model, thresholds, item discrimination indices, item difficulties, and RSquared values. MPlus also provides category proportion values for dichotomous variables which are equivalent to the classical test theory p values. Standard errors were computed for each of the statistical estimates. The parameterization for the normal ogive model employs a two parameter probit metric where the probit value is equal to discrimination * (theta-difficulty).

With a structured model being specified the confirmatory factor analysis modeling available in MPlus tests the bi-factor hypothesis that there is a single latent dimension underlying all of the variables within the modeled dataset and separate orthogonal and independent latent dimensions accounting for additional variance beyond the base model specified in the two parameter normal ogive model. Essentially, with the case based items from Form A and Form B the analysis models the single latent dimension or factor and provides measurement model estimates equivalent to factor dimension loadings and standard errors on the primary latent dimension. The analysis is a confirmatory factor analysis to determine if there exist orthogonal, independent latent dimensions that account for supplemental variance for the first case and all remaining variables, for the second case and all remaining variables, etc. to the sixth case group within each form. In standard multi-trait-multi-method terminology, the cases in the analysis can be considered as alternative measurement methods. The bi-factor confirmatory factor analysis sequentially tests the existence of subsequent orthogonal dimensions that account for measurable variance after the primary latent dimension has been modeled and the second latent dimension accounting for the first case and its interactions with the remaining variables. Additional latent orthogonal dimensions are confirmed if they are present for each of the six case groups within the test form. The analysis confirms if there is a primary latent dimension in the data, and if there is a latent dimension that accounts for variance with the first case and all remaining unanalyzed variables, then the second case and remaining unanalyzed variables (cases 2 to 6), then with case

3 and the remaining unanalyzed variables (cases 3 to 6). This procedure continues until the last latent dimension is confirmed with the variables related only to the last case group.

A third confirmatory factor analysis evaluated the presence of a one factor logistic regression model with a two parameter logistic metric using the parameterization where the logit is 1.7*Discrimination*(Theta-Difficulty). The one factor logistic regression model provides model estimates, thresholds, item discriminations, item difficulties, and RSquared values. Standard errors are provided for each statistic estimated.

Multiple tests of model fit were completed for the confirmatory factor analyses, a chi square test, a comparative fit indicator, the Tucker-Lewis indicator, the root mean squared error of approximation (RMSEA) and the weighted root mean square residual (WRMR). As shown in Table 7, each indicator showed very acceptable model fit for both the one factor IRT normal ogive model and the bi-factor model. Each of the analyses was estimated with weighted least squares estimation with mean and variance corrections.

Table 7. Model Fit Tests for Confirmatory Factor Analysis

Tests of Model Fit	**One Factor Normal Ogive WLSMV FormA**	**Bi-Factor WLSMV Form A**	**One Factor Normal Ogive WLSMV Form B**	**Bi-Factor WLSMV Form B**
Chi-Square Test	1409.469	1409.47	2007.319	610.685
Df	30	30	46	30
Probability Value	0.00	0.00	0.00	0.00
Comparative Fit Indicator	0.840	0.862	0.751	0.788
Tucker-Lewis Indicator	0.968	0.972	0.941	0.95
Root Mean Square Error of Approximation (RSMEA)	0.049	0.045	0.063	0.058
Weighted Root Mean Square Residual (WRMR)	1.18	1.087	1.421	1.303

Table 7 shows that for each of the test forms A and B the bi-factor model with the case clusterings was a better fit than the single factor IRT normal ogive model. Essentially this means that there is meaningful measurable variance in the case structure methods dimensions that is not accounted for by the single latent dimension underlying the test items within each form.

Comparing One Factor Weighted Least Squares Model with the Bi-Factor Models

This section compares the loadings for the one factor normal ogive model and the bi-factor loadings for the general loading and the specific cluster or case loading estimates for Forms A (Table 8 to 10) and B (Table 11-13) respectively. The one factor estimated loadings are the factor loadings for each item if there is confirmed only a single latent dimension underlying the examinee performance on the set of items in Form A. The bi-factor general loading is the estimated factor loading on the underlying latent dimension that is measured by all of the items with the structured model. The bi-factor cluster loading is the estimated factor loading on the bi-factor that confirms if there is any supplemental variance contributed by the items in each case and the interactions with the remaining variables in the model. The orthogonal structural factors are sequentially addressed with the remaining items in that case or subsequent cases up to the final latent dimension for the items in the last case. Colors have been added to facilitate comparisons of the cluster loadings for each of the six cases per test form.

Loadings greater or less than 0.200 were classified as significant in Table 8 and 9 for Form A. The significant negative loading are interpreted as the examinee having a greater probability of a lower score with the presence of the item embedded in the case structure than if the case structure was not present. The same interpretation should be given for all of the negative loadings in the following analyses.

Table 8. Form A Comparing One Factor and Bi-Factor with Item Clustering

		One Factor Model		BiFactor Model		BiFactor Model	
Form A		One Factor Estimated Loadings	Standard Errors	General Loading		Cluster Loading	
Item Variable	CASE	Estimate	SE	Estimate	SE	Estimate	SE
V11ABA	**B**	0.712	0.050	0.703	0.051	0.455	0.107
V12ABB	**B**	0.821	0.030	0.836	0.030	-0.361	0.112
V51ABK	**B**	0.705	0.066	0.699	0.067	0.345	0.098
V51ABL	**B**	0.776	0.045	0.776	0.046	0.196	0.101
V52ABM	**B**	0.550	0.045	0.551	0.045	0.128	0.092
V53ABN	**B**	0.594	0.059	0.582	0.060	0.540	0.107
V12AEB	**C**	0.751	0.035	0.747	0.034	0.142	0.076
V13AEC	**C**	0.698	0.037	0.719	0.036	-0.277	0.110
V21AEA	**C**	0.800	0.030	0.798	0.029	0.099	0.073
V31AEE	**C**	0.601	0.053	0.603	0.053	0.032	0.098
V32AEG	**C**	0.788	0.029	0.770	0.031	0.396	0.078
V33AEH	**C**	0.727	0.041	0.724	0.042	0.121	0.086
V33AEI	**C**	0.769	0.038	0.754	0.039	0.330	0.080
V34AEK	**C**	0.679	0.048	0.682	0.048	0.027	0.098

V51AEL	**C**	0.797	0.031	0.782	0.032	0.316	0.071
V52AEO	**C**	0.843	0.024	0.836	0.025	0.191	0.068
V21AIA	**E**	0.682	0.049	0.681	0.049	0.234	0.107
V23AID	**E**	0.715	0.040	0.715	0.041	0.232	0.095
V24AIE	**E**	0.725	0.058	0.729	0.058	0.019	0.112
V26AIH	**E**	0.497	0.051	0.513	0.050	-0.470	0.149
V41AII	**E**	0.699	0.036	0.698	0.037	0.263	0.093
V43AIK	**E**	0.678	0.060	0.686	0.060	-0.159	0.122
V44AIL	**E**	0.796	0.028	0.796	0.029	0.268	0.087
V22AJA	**F**	0.571	0.059	0.533	0.064	0.630	0.097
V23AJB	**F**	0.616	0.044	0.586	0.047	0.585	0.085
V24AJD	**F**	0.719	0.033	0.703	0.035	0.400	0.075
V25AJE	**F**	0.735	0.043	0.747	0.043	-0.199	0.091
V26AJF	**F**	0.698	0.048	0.700	0.049	0.034	0.081
V26AJG	**F**	0.789	0.029	0.800	0.029	-0.138	0.072
V41AJH	**F**	0.776	0.038	0.770	0.040	0.224	0.075
V42AJI	**F**	0.424	0.094	0.427	0.094	-0.023	0.078
V43AJJ	**F**	0.720	0.048	0.713	0.050	0.235	0.079
V44AJL	**F**	0.678	0.057	0.677	0.058	0.117	0.080
V11ADA	**G**	0.650	0.068	0.633	0.070	0.300	0.110
V12ADB	**G**	0.678	0.068	0.639	0.072	0.530	0.095
V31ADD	**G**	0.578	0.062	0.556	0.063	0.371	0.090
V31ADE	**G**	0.611	0.058	0.587	0.059	0.387	0.090
V32ADF	**G**	0.709	0.051	0.680	0.055	0.456	0.086
V33ADG	**G**	0.510	0.050	0.512	0.051	0.039	0.100
V34ADI	**G**	0.794	0.039	0.796	0.040	0.049	0.100
V34ADJ	**G**	0.512	0.066	0.501	0.066	0.222	0.097
V51ADL	**G**	0.831	0.032	0.839	0.032	-0.044	0.102
V52ADM	**G**	0.767	0.035	0.771	0.035	0.013	0.098
V53ADN	**G**	0.803	0.051	0.791	0.052	0.255	0.093
V53ADO	**G**	0.703	0.073	0.666	0.077	0.499	0.085
V23AFD	**H**	0.649	0.049	0.655	0.049	-0.051	0.097
V25AFF	**H**	0.657	0.042	0.671	0.042	-0.301	0.117
V41AFI	**H**	0.783	0.031	0.773	0.033	0.468	0.100
V44AFL	**H**	0.729	0.034	0.724	0.035	0.243	0.077
V44AFM	**H**	0.775	0.033	0.765	0.035	0.366	0.087
AVERAGE		0.697	0.046	0.692	0.047	0.175	0.093
STD DEV		0.095	0.014	0.097	0.015	0.252	0.015
MIN		0.424	0.024	0.427	0.025	-0.470	0.068
MAX		0.843	0.094	0.839	0.094	0.630	0.149

MEDIAN		0.711	0.045	0.703	0.047	0.223	0.093
ITEMS		50					
OBSERVATIONS		630					

Summary information is provided at the bottom of Table 8 for the average, standard deviation, minimum, maximum, median and the number of items and total observations for Form A. This summary information shows that the average loadings for the one factor model are 0.697 with a standard deviation of 0.095 while the average of the loadings for the bi-factor model are slightly less at 0.692 with a standard deviation of 0.097. The average of the loadings for the clustered items within cases is 0.175 with a standard deviation of 0.252. The high and low clustered loadings are -.470 and .630 with a median cluster loading of 0.223. Average standard errors for the one factor model are 0.046 and for the bi-factor underlying factor 0.047 and for the clustered loadings 0.093. There are 50 items and 630 observations for Form A.
Table 9 shows that there are significant positive cluster loadings and a few negative cluster loadings across the cases. This indicates that the confirmatory factor analysis was able to identify independent and orthogonal variables and variance that are contributed by only knowledge of the case clustering as a type of methods variable.

Table 9. Form A cluster loadings for each case.

Case	**Positive cluster loadings**	**Negative cluster loadings**
B	V11ABA, V51ABK, V53ABN (3 items)	V12ABB
C	V32AEG, V33AEI, and V51AEL (3 Items)	V13AEC
E	V21AIA, V23AID, V41AII and V44AIL (4 items)	V26AIH
F	V22AJA, V23AJB, V24AJD, V41AJH. and V43AJJ (5 items)	V25AJE (almost significant)
G	V11ADA, V12ADB, V31ADD, V31ADE , V32ADF, V34ADJ, V53ADN and V53ADO (8 Items)	
H	V41AFI, V44AFL, and V44AFM (3 Items)	V25AFF

Figure 13 provides a scatterplot of the one factor normal ogive estimated factor loadings and the bi-factor estimate loadings for Form A. The linear trendline is also plotted indicating that the one factor normal ogive loading estimates correspond linearly with the bi-factor estimated loadings from the multi-factor confirmatory factor analysis with one independent dimension estimated for each of the six case clusters present in Form A.

Figure 13. One Factor Normal Ogive Loading Estimates and Bi-Factor Loading Estimates

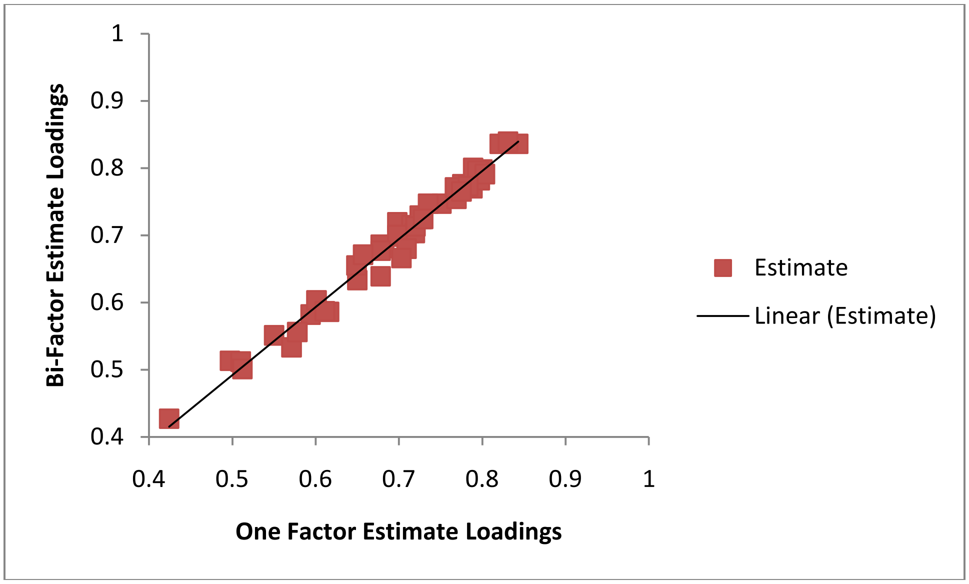
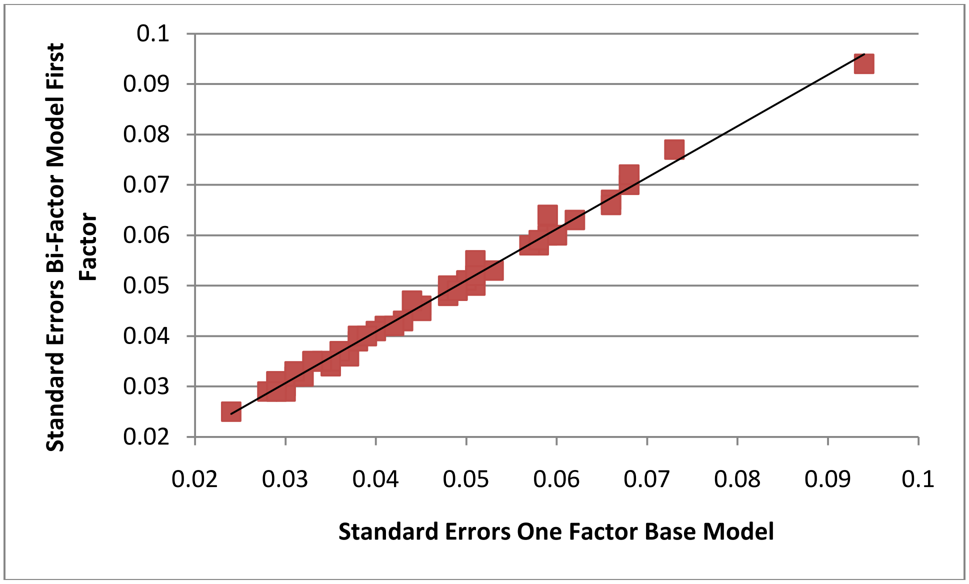
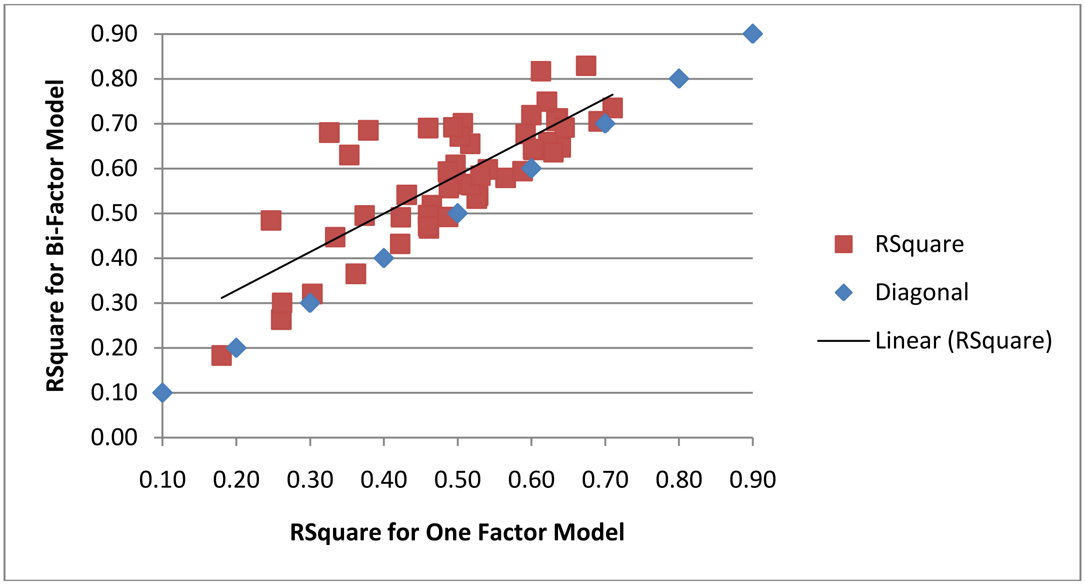


Figure 14. Standard Errors for First Factor One Factor Model and Bi-Factor Model



For Form A Figure 15 presents an analysis of the RSquared model fit for the normal ogive one factor model and the bi-factor model RSquared with the base primary factor and then separate independent dimensions for each of the case scenario clusters. A linear best fitting linear trend line is also displayed by the solid black line. A diagonal line is also represented by the blue diamonds. Since the majority of the RSquared values are above the diagonal line, this indicates that the bi-factor model provides measurable and significant variance beyond the measurement available with the one factor normal ogive model.

Figure 15. RSquared for One Factor Normal Ogive Model and Bi-Factor Model



For Form A, Table 10 provides statistics for the Loadings, Standard Errors and RSquare values for the one factor model and the bi-factor model. The loadings and standard errors are very comparable between models but also the RSquares are slightly larger for the bi-factor model than for the one factor normal ogive model.

Table 10. Form A Loadings, Standard Errors and RSquare for One Factor Model and Bi-Factor Model

	Loadings		Standard Errors		RSquare	
	One Factor Model	Bi-Factor Model	One Factor Model	Bi-Factor Model	One Factor Model	Bi-Factor Model
AVERAGE	0.697	0.692	0.046	0.047	0.495	0.581
STD DEV	0.095	0.097	0.014	0.015	0.126	0.137
MIN	0.424	0.427	0.024	0.025	0.180	0.183
MAX	0.843	0.839	0.094	0.094	0.710	0.829
MEDIAN	0.711	0.703	0.045	0.047	0.505	0.596

Table 11 presents the comparable analysis for the one factor model and the bi-factor model for Form B. Loadings greater or less than 0.200 were classified as significant in Table 11 and 12 for Form B.

Table 11. Form B Comparing One Factor and Bi-Factor with Item Clustering

		One Factor Model		BiFactor Model		BiFactor Model	
Form B		One Factor Estimated Loadings	Standard Errors	General Loading		Cluster Loading	
Item Variable	CASE	Estimate	SE	Loading	SE	Loading	SE
V13AAC	A	0.688	0.041	0.689	0.041	0.072	0.085
V14AAD	A	0.832	0.025	0.832	0.026	0.444	0.078
V31AAE	A	0.673	0.049	0.676	0.049	0.228	0.087
V31AAF	A	0.733	0.034	0.730	0.034	0.208	0.068
V32AAG	A	0.672	0.045	0.669	0.045	0.274	0.096
V33AAH	A	0.604	0.048	0.604	0.048	0.118	0.090
V34AAJ	A	0.621	0.050	0.626	0.049	0.387	0.086
V34AAK	A	0.583	0.056	0.587	0.056	0.311	0.082
V51AAL	A	0.740	0.043	0.741	0.043	0.036	0.082
V52AAM	A	0.706	0.053	0.709	0.053	0.205	0.084
V53AAO	A	0.685	0.053	0.688	0.053	0.245	0.083
V11ABA	B	0.744	0.048	0.748	0.047	0.087	0.066
V12ABB	B	0.784	0.037	0.787	0.037	0.076	0.056
V14ABD	B	0.476	0.075	0.458	0.077	0.357	0.076
V31ABE	B	0.459	0.045	0.418	0.048	0.664	0.064
V32ABG	B	0.510	0.060	0.474	0.064	0.584	0.064
V33ABH	B	0.457	0.047	0.444	0.048	0.240	0.068
V34ABJ	B	0.420	0.050	0.370	0.053	0.727	0.065
V51ABK	B	0.668	0.059	0.674	0.059	0.100	0.070
V51ABL	B	0.748	0.055	0.749	0.054	0.012	0.067
V52ABM	B	0.496	0.046	0.485	0.048	0.238	0.068
V53ABN	B	0.601	0.054	0.611	0.053	0.183	0.074
V21AHA	D	0.691	0.063	0.693	0.063	0.371	0.088
V22AHB	D	0.799	0.048	0.797	0.048	0.146	0.092
V23AHC	D	0.521	0.068	0.524	0.068	0.463	0.097
V23AHD	D	0.541	0.045	0.545	0.045	0.313	0.088
V24AHF	D	0.833	0.026	0.831	0.027	0.302	0.068
V25AHG	D	0.675	0.042	0.679	0.042	0.268	0.081
V26AHH	D	0.718	0.037	0.717	0.038	0.219	0.077
V41AHI	D	0.622	0.041	0.623	0.041	0.079	0.079
V42AHJ	D	0.697	0.049	0.700	0.049	0.204	0.083
V43AHK	D	0.657	0.042	0.657	0.042	0.058	0.079

V44AHL	D	0.836	0.032	0.834	0.032	0.209	0.070
V44AHM	D	0.727	0.038	0.727	0.038	0.136	0.076
V23AIC	E	0.001	0.053	0.026	0.053	0.794	0.241
V41AII	E	0.743	0.034	0.739	0.034	0.204	0.076
V44AIL	E	0.815	0.026	0.811	0.027	0.260	0.081
V44AIM	E	0.571	0.039	0.565	0.040	0.226	0.088
V14ADC	G	0.417	0.052	0.406	0.053	0.576	0.202
V33ADH	G	0.369	0.050	0.363	0.051	0.416	0.147
V52ADM	G	0.740	0.039	0.746	0.039	0.235	0.102
V53ADN	G	0.781	0.052	0.782	0.052	0.063	0.092
V53ADO	G	0.726	0.063	0.725	0.063	0.143	0.100
V22AGB	I	0.800	0.038	0.802	0.038	0.179	0.089
V23AGD	I	0.704	0.043	0.705	0.043	0.055	0.091
V24AGE	I	0.745	0.038	0.746	0.038	0.099	0.091
V25AGF	I	0.776	0.031	0.781	0.031	0.392	0.135
V26AGG	I	0.660	0.061	0.664	0.062	0.393	0.134
V41AGH	I	0.843	0.025	0.844	0.025	0.093	0.076
V43AGJ	I	0.772	0.039	0.774	0.039	0.215	0.095
AVERAGE		0.654	0.046	0.652	0.046	0.258	0.090
STD DEV		0.154	0.011	0.157	0.011	0.180	0.033
MIN		0.001	0.025	0.026	0.025	0.012	0.056
MAX		0.843	0.075	0.844	0.077	0.794	0.241
MEDIAN		0.690	0.046	0.691	0.048	0.223	0.083
ITEMS		50					
OBSERVATIONS		640					

Summary information is provided at the bottom of the Table 11 for Form B for the average, standard deviation, minimum, maximum, median, the number of items and total observations. This summary information shows that the average loadings for the one factor model are 0.654 with a standard deviation of 0.154 while the average of the loadings for the bi-factor model are just slightly less at 0.692 with a standard deviation of 0.157. The average of the loadings for the clustered items within cases is 0.258 with a standard deviation of 0.180. The high and low clustered loadings are 0.012 and .794 with a median cluster loading of 0.223. Average standard errors for the one factor model are 0.046 and for the bi-factor underlying factor 0.046 and for the clustered loadings 0.090. There are 50 items and 640 observations for Form B.

For Form B Table 12 shows that there are significant positive cluster loadings across the cases. No negative cluster loadings were found. This indicates that the confirmatory factor analysis was able to identify independent and orthogonal variables and factor variance that are contributed by only knowledge of the case clustering as a type of methods variable.

Table 12. Form B Summary of cluster loadings for each case.

Case	Positive cluster loadings
A	V14AAD, V31AAE, V31AAF, V32AAG, V34AAJ, V34AAK, V52AAM and V53AAO (8 Items)
B	V14ABD,V31ABE, V32ABG, V33ABH, V34ABJ, and V52ABM (6 Items)
D	V21AHA, V23AHC, V24AHF, V25AHG, V26AHH, V42AHJ, V44AHL (7 Items)
E	V23AIC, V41AII, V44AIL and V44AIM (4 Items)
G	V14ADC, V33ADH, V52ADM (3 Items)
I	V25AGF, V26AGG and V43AGJ

Figure 16 provides a scatterplot of the one factor normal ogive estimated factor loadings and the bi-factor estimate loadings for Form B. The Linear trendline is also plotted indicating that the one factor normal ogive loading estimates correspond linearly with the bi-factor estimated loadings from the multi-factor confirmatory factor analysis with one independent dimension estimated for each of the 6 case clusters present in Form B. One item, **V23AIC from Case E,** had numerical estimation problems and was not well estimated in either the one factor normal ogive model or the bi-factor solution and was eliminated from the graphic in Figure 14. This item had very low loading estimates of .001 for the normal ogive model and .026 for the bi-factor model. That item also was the most difficult item on the form and had a near zero point-biserial correlation with the total score.

Figure 16. One Factor Normal Ogive Loading Estimates and Bi-Factor Loading Estimates

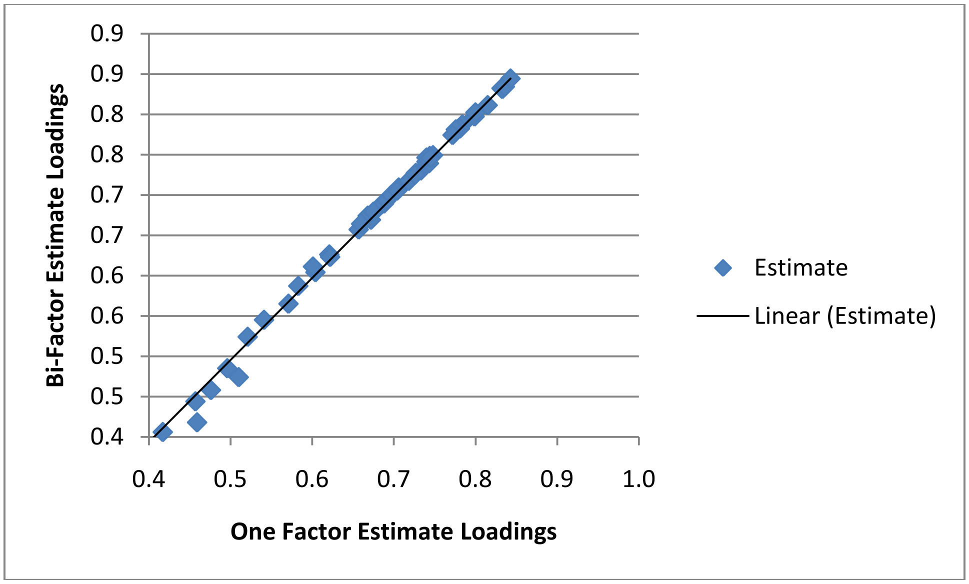
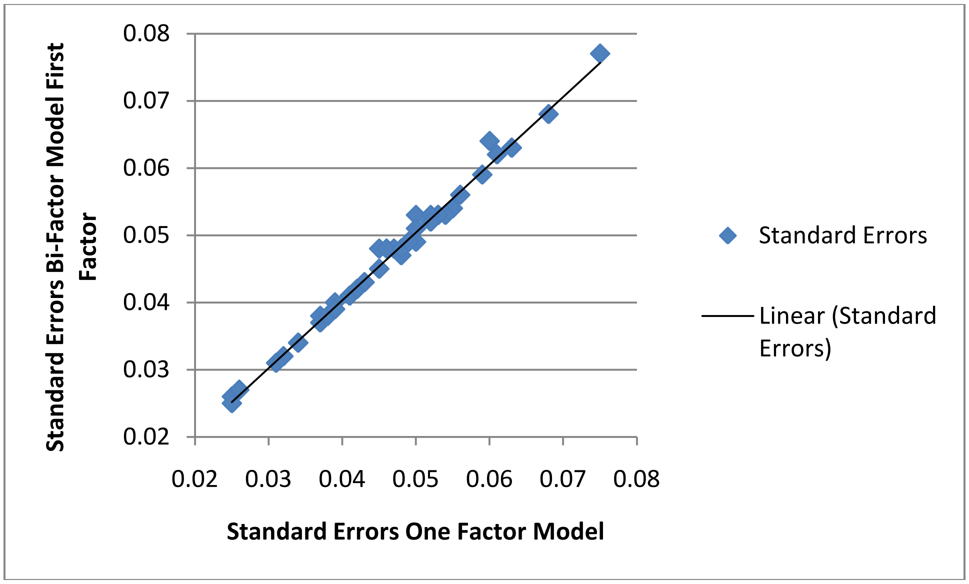


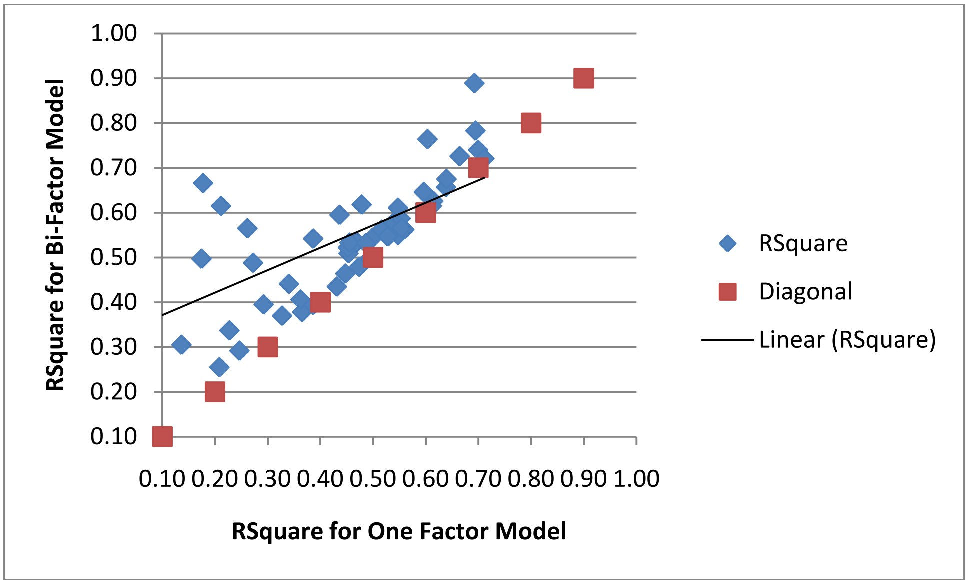
Figure 17 provides a scatterplot of the one factor normal ogive estimated factor loadings and the bi-factor estimate loadings for Form B. The Linear trendline is also plotted indicating that the one factor normal ogive loading estimates correspond linearly with the bi-factor estimated loadings from the multi-factor confirmatory factor analysis with one independent dimension estimated for each of the 6 case clusters present in Form B.

Figure 17. Standard Errors for First Factor of One Factor Model and Bi-Factor Model



For Form B Figure 18 presents a comparison of the RSquared model fit for the normal ogive one factor model and the bi-factor model RSquared with the base primary factor and then separate independent dimensions for each of the case scenario clusters. A linear best fitting linear trend line is also displayed by the solid black line. A diagonal line is also represented by the red squares. Since the majority of the RSquared values are above the equal diagonal line, this indicates that the bi-factor model provides measurable and significant variance beyond the measurement attributable to the one factor normal ogive model.

Figure 18. RSquared for One Factor Normal Ogive Model and Bi-Factor Model



For Form B Table 13 provides statistics for the Loadings, Standard Errors and RSquare for the one factor model and the bi-factor model. The loadings are very comparable between models, the standard errors and the RSquare is slightly larger for the bi-factor model.

Table 13. Form B Loadings, Standard Errors and RSquare for One Factor Model and Bi-Factor Model

	Loadings		Standard Errors		RSquare	
	One Factor Model	Bi-Factor Model	One Factor Model	Bi-Factor Model	One Factor Model	Bi-Factor Model
AVERAGE	0.654	0.652	0.046	0.046	0.451	0.547
STD DEV	0.154	0.157	0.011	0.011	0.167	0.131
MIN	0.001	0.026	0.025	0.025	0.000	0.255
MAX	0.843	0.844	0.075	0.077	0.711	0.889
MEDIAN	0.690	0.691	0.046	0.048	0.476	0.549

Although the bi-factor model accounts for variance for each of the cases, there was a significant residual variance that was not accounted for each variable in the bi-factor model. For Form A the average residual variance was 0.419 with a standard deviation of 0.137. The minimum residual variance for an item variable was 0.171, the maximum residual variance for an item variable was 0.817 and the median residual variance was 0.404. For Form B the average residual variance was 0.453 with a standard deviation of 0.131. The minimum residual variance for an item variable was 0.111, the maximum residual variance for an item variable was 0.745 and the median residual variance was 0.451. These results show that there substantial variance in the item variables for each test form that was not accounted for by the general loading on the first latent dimension and the cluster loading.

One Factor Logistic Regression Model

A one factor logistic regression model was also computed for the confirmatory factor analysis using the item response theory parameterization with the two-parameter logistic metric where the logit is 1.7* Discrimination*(Theta-Difficulty). The analysis for this model is provided in Appendix B. With the logistic regression parameterized model for Form A the average estimated loading on the one latent dimension is 1.861 with a standard deviation of 0.549. The minimum loading is 0.579 and the maximum loading is 3.088 with a median loading of 1.835. Standard errors for Form A are 0.254 with a standard deviation of 0.068.

With the logistic regression parameterized model for Form B the average estimated loading on the one latent dimension is 1.651 with a standard deviation of 0.618. The minimum loading is 0.144 and the maximum loading is 3.020 with a median loading of 1.665. Standard errors for Form A are 0.219 with a standard deviation of 0.068.

Figures 19 and 20 show the estimated primary factor loadings from the one factor normal ogive model and the one factor logistic regression model. Exponential trend lines are also plotted on these graphs showing that there is a logarithmic relationship between the two loadings. This result was expected due to the calibration of the logistic regression model for Forms A and B.

Figure 19. Form A Estimated Primary Factor Loadings from the One Factor Normal Ogive Model and the One Factor Logistic Model

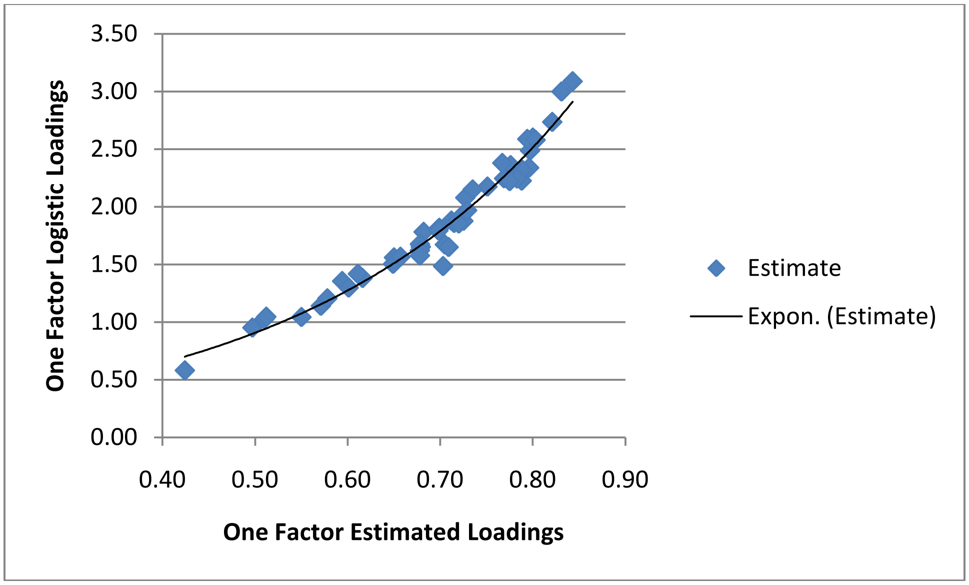


Figure 20. Form B Estimated One Factor Loadings from the One Factor Normal Ogive Model and the One Factor Logistic Model

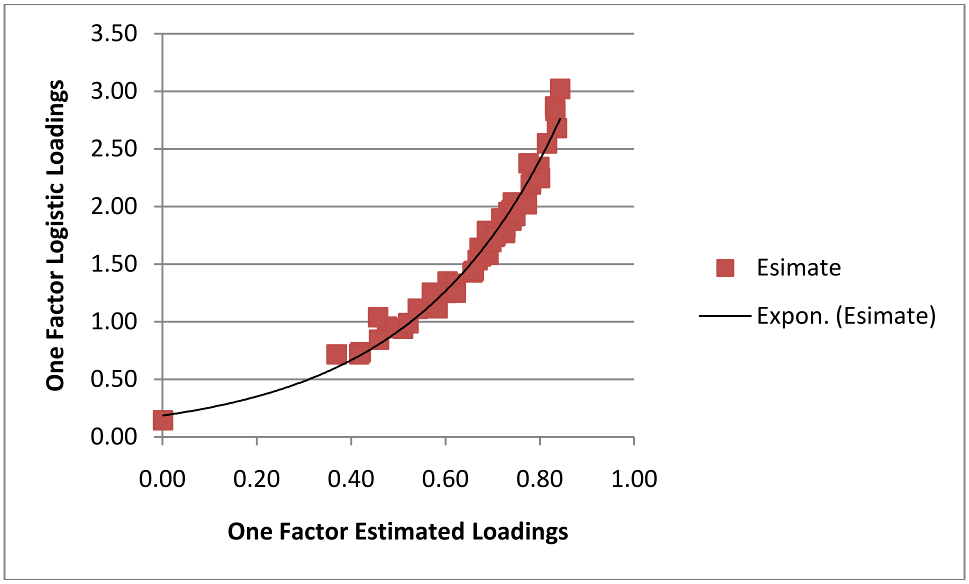


Figure 21 and 22 present results of standard error analysis for the first factor estimation for the weighted least squares model and the one factor logistic model for Forms A and B. These figures both show that the standard errors of the estimated first factor loadings for the logistic model are three to four times larger than the standard errors computed for the first factor loadings of the normal ogive weighted least squares solution with mean and variance correction. There is also a negative slope for the linear (black) and exponential (blue) trendlines for the standard errors.

Figure 21. Form A Standard Errors for one factor solution and the one factor logistic model

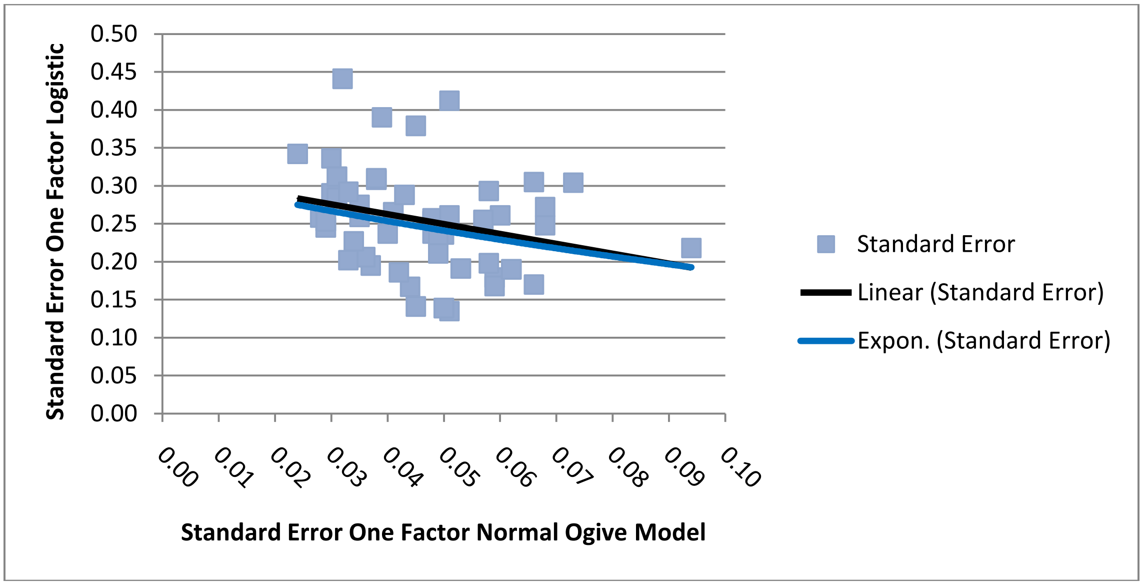
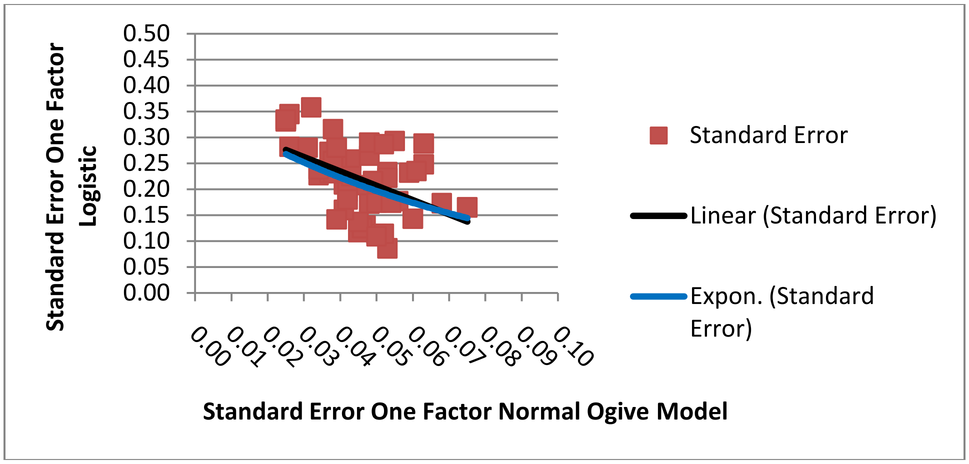


Figure 22. Form B Standard Errors for one factor solution and the one factor logistic model



Figures 23 and 24 present the RSquared comparisons for the first factor normal ogive model and the first factor for the logistic model. The figure shows that the RSquared accounting for variance is very similar for the one factor normal ogive model and the one factor logistic model. The RSquared is well fit by either a linear or exponential trendline.

Figure 23 Form A RSquared analysis for one factor normal ogive model and one factor logistic model

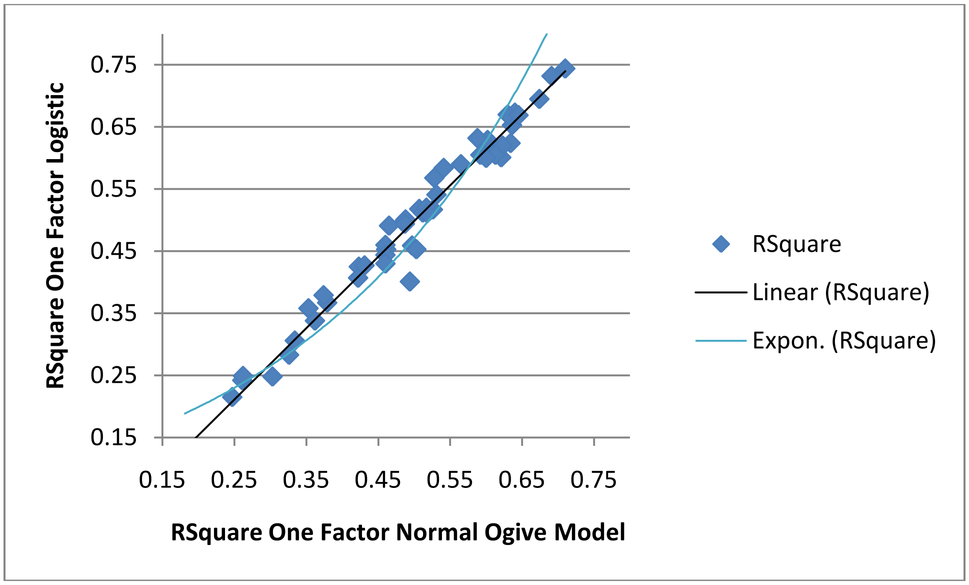
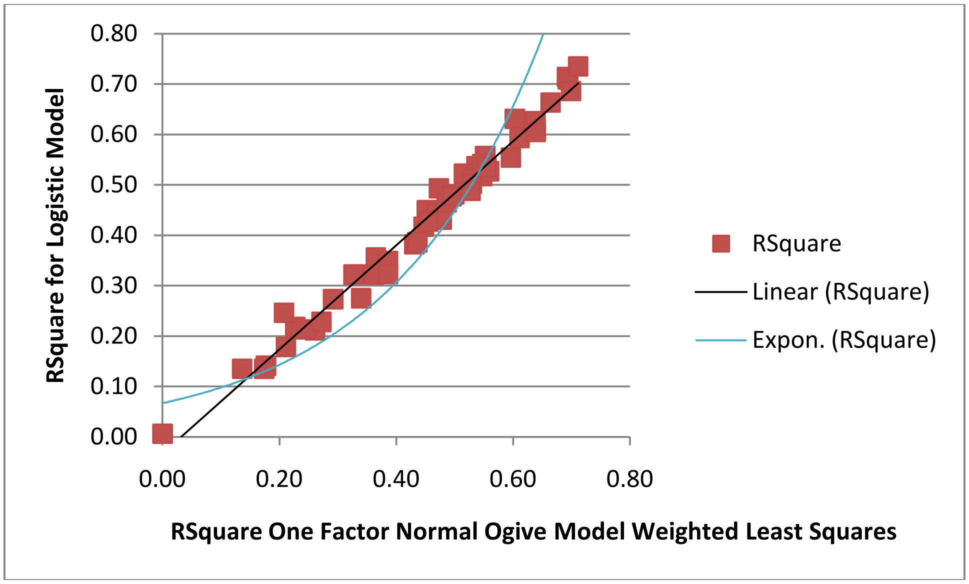


Figure 24 **Form B RSquared analysis for one factor normal ogive model and one factor logistic model.** The linear trendline is a better fit than the exponential trendline for Form B.



The normal ogive model weighted least squares and the logistic model also computed estimates of IRT item difficulty and item discrimination for Forms A and B. Item Difficulty comparisons are provided in Figures 25 and 26 and the Item Discrimination comparisons are provided in Figures 27 and 28. In Figure 25, the one point (-3.376, -4.545) was the item from Form A that had difficulty being numerically estimation.

Figure 25 Form A IRT Difficulty Indices for One Factor Normal Ogive Model and One Factor Logistic Model.

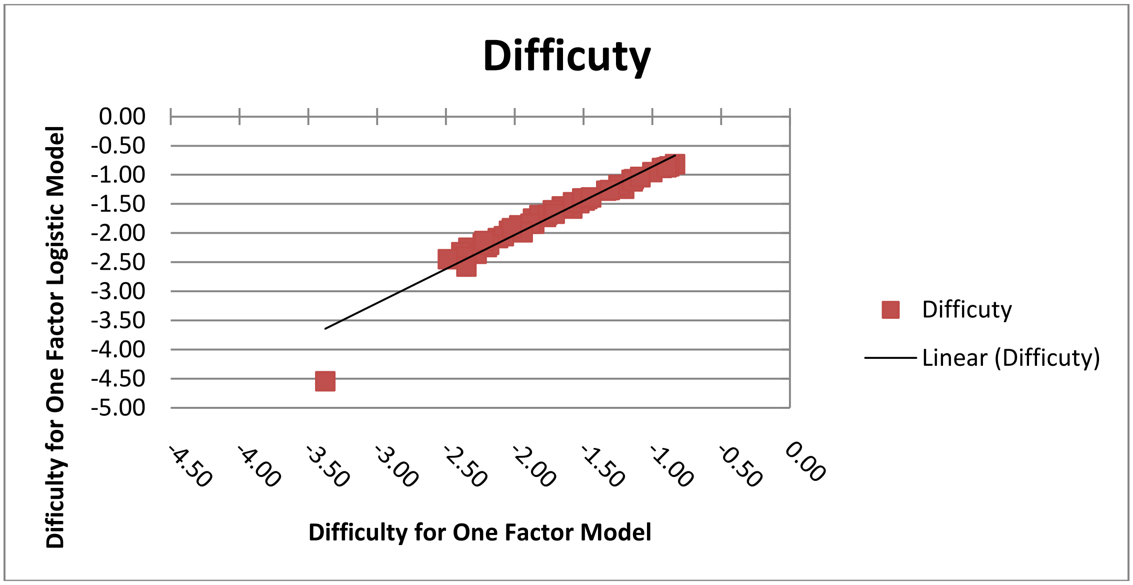


Figure 26 Form B IRT Difficulty for One Factor Normal Ogive Model and One Factor Logistic Regression Model.

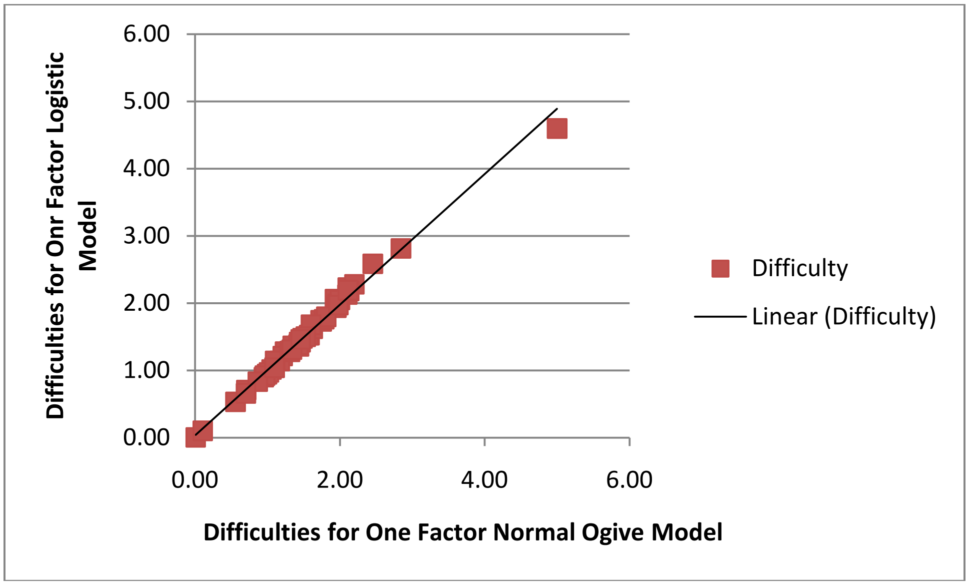
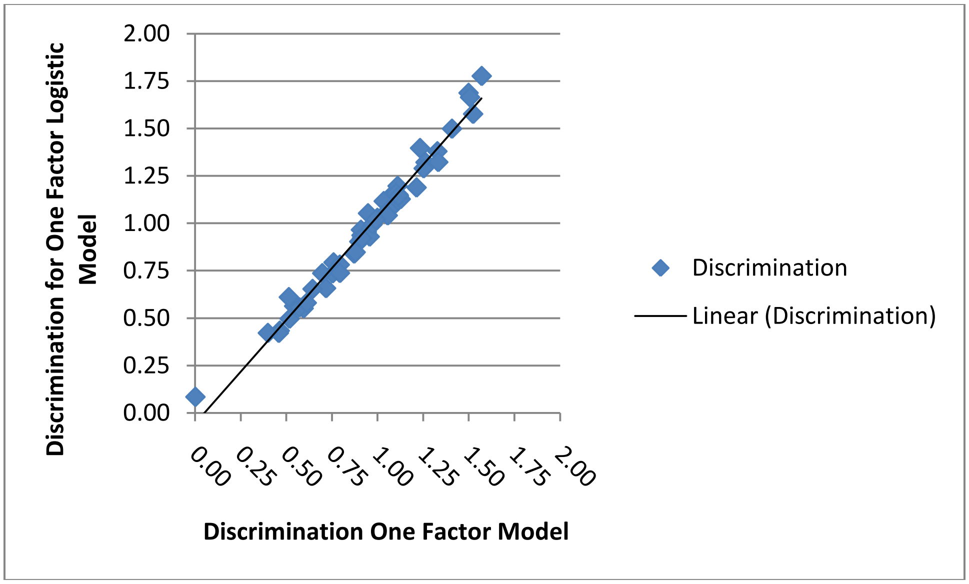
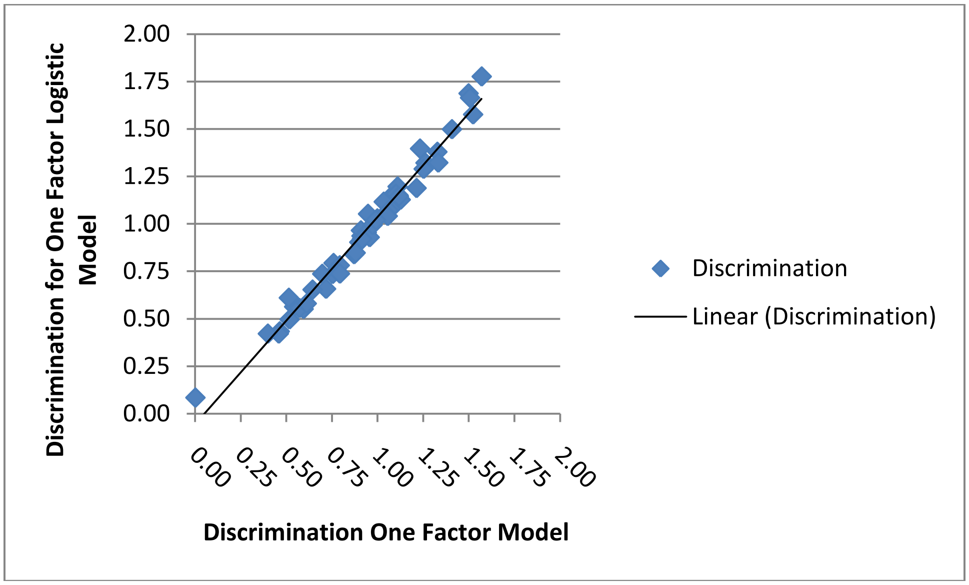


Figure 27 Form A IRT Discrimination Indices for One Factor Normal Ogive Model and One Factor Logistic Model



For Form A, 26 of 50 total items have IRT discrimination indices greater than 1.0 for the one factor model and 29 of 50 total items have IRT discrimination indices greater than 1.0 for the logistic model.

Figure 28 Form B IRT Discrimination Indices for One Factor Normal Ogive Model and One Factor Logistic Model.



For Form B, 21 of 50 total items have IRT discrimination indices greater than 1.0 for the one factor model and 40 of 50 total items have IRT discrimination indices greater than 1.0 for the logistic model.

In summary the confirmatory factor analysis for both Forms A and B provides classical item analysis p value statistics, confirmatory factor loadings on the single dimension underlying the data. If the data are multi-dimensional, the confirmatory factor analysis will indicate the number of underlying primary dimensions present in the data. The confirmatory factor analysis results also provide item thresholds, item discriminations, item difficulties in the theta metric, standardized loading estimates, and RSquared statistics. Each statistic computed has standard errors reported for each variable analyzed.

Correlation Analysis Within and Across Case Clusters.

An analysis was conducted to determine if inter-item correlations were greater within the case clusters rather than across the case clusters. This analysis is presented in Table 14. The table includes the mean correlation, the standard deviation of the correlations, the minimum correlation and the maximum correlation. For each of the test forms, the mean correlations are higher within the cases by about .04 on average than across the cases. The standard deviations show a mixed pattern with a smaller standard deviation across than within cases for Form A. For Form B the standard deviation is greater across cases than within cases. For Form A the minimum correlation is the same within as across cases. For Form B the correlation is smaller across cases than within cases. For both Form A and Form B, the maximum correlation is greater across cases than within cases. The general result from this analysis was that there was a slightly larger correlation (but only 0.04) within the case groups then across the cases.

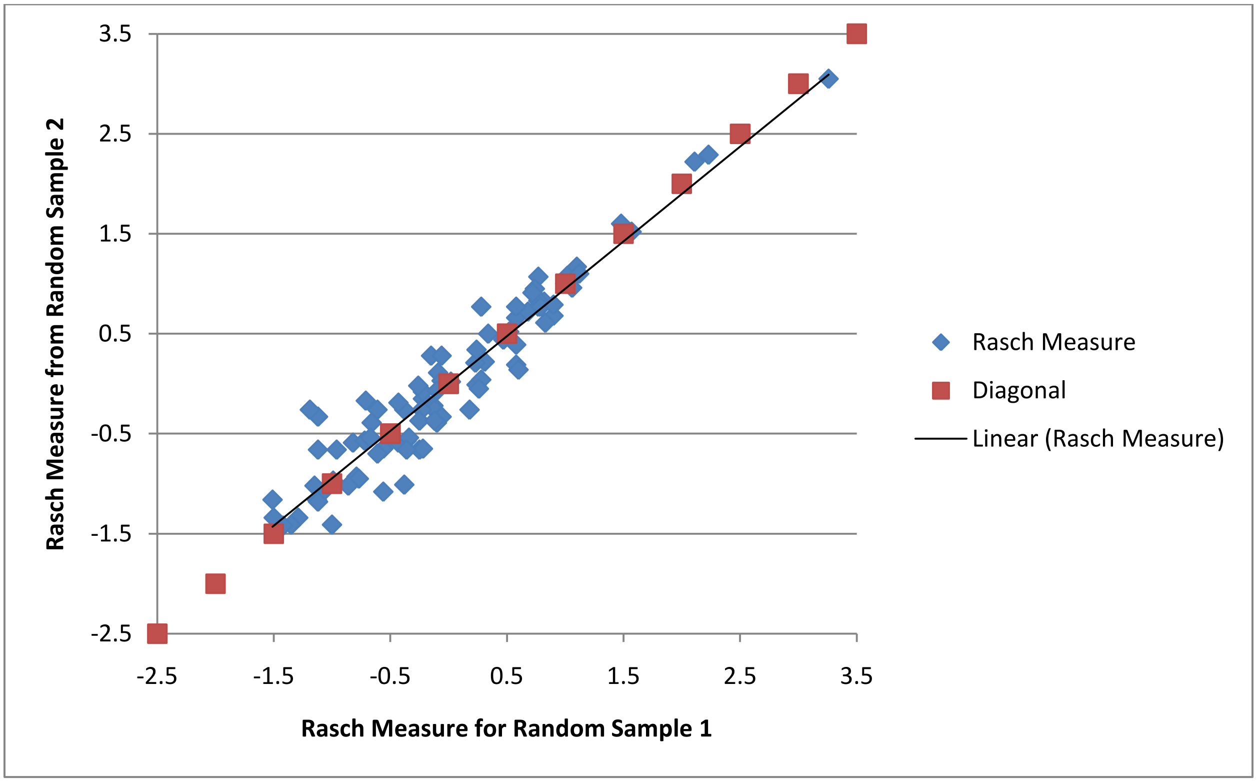
Table 14. Analyses of Across and Within Correlations of Cases by Form

	Form A		Form B	
	Across	**Within**	**Across**	**Within**
Mean	0.230	0.264	0.214	0.257
S.D.	0.080	0.089	0.102	0.096
Min	0.010	0.037	-0.174	-0.116
Max	0.550	0.505	0.524	0.469

Stability of Item Parameter Estimation by Differential Item Functioning (DIF)

A Differential Item Functioning analysis using Rasch item meassures and Mansel/Haenszel statistics was conducted to verify the stability of item parameter estimation from two random samples of examinees. The analysis was conducted by fixing person ability estimates for the test forms, estimating the Master's partial credit parameters for each items on each form, estimating item measures for two random samples of examinees for each item on each form and then evaluating the t statistic and Mansel/Haenszel statistics for each item for the two random samples. Figure 29 shows that Rasch Measures from two random samples of examinees receiving each of the 89 test items. This plot shows a strong linear trend between the Rasch measures from the two independent samples. The plot also shows the diagonal indicated by the red squares which would indicate an identical measure for each sample.

Figure 29 Rasch Measures for Differential Item Functioning (DIF) from Two Random Samples



There were four items of the 89 test items that exhibited significant or near significant Differential Item Functioning as measured by the Welch's (Student's) t statistic and the Mantel Haenszel statistic which would indicate a lack of local item independence for these items given two random examinee samples. Two items were from Case D, one from case G and one from Case H. Eighty-five of the eighty-nine items (96%) showed no differential item functioning and would indicate stability of item parameter estimation.

Table 15 Items Showing Significant DIF from Two Random Samples

PERSON DIF	DIF	PERSON DIF	DIF		Welch		Mantel	Haenszel	ITEM	CASE
SAMPLE 1 MEASURE	S.E.	SAMPLE 2 MEASURE	S.E.	t	df	Prob.	Prob.	Size	Name	CASE
-1.12	0.24	-0.33	0.19	-2.56	622	0.0107	0.0132	-0.27	2.3.ahc	D
0.58	0.15	0.19	0.17	1.7	622	0.0893	0.0412	0.34	2.4.ahf	D
-1.19	0.26	-0.26	0.21	-2.77	517	0.0057	0.0091	-0.80	3.4.ADI	G
0.28	0.18	0.77	0.17	-2.01	523	0.0449	0.0841	-0.43	4.1.afi	H

Conclusions

The measurement dataset employed in the analysis was from an information technology certification test administered in two test forms of 50 items in length and administered to approximately 630-640 individuals. The fifty items per form were also grouped into six case scenarios that give the context setting for the individual items.

Classical item and test analysis was conducted on the two test forms. The analysis showed that test Form A was measurably easier than Form B as indicated by the mean score which was higher by two points than Form A and the median score which was higher by four points than Form A. The test score ranges were the same and the alpha reliabilities of the forms were 0.939 for Form A and 0.928 for Form B. The standard error of measurement for the two test forms was 2.07 for Form A and 2.33 for Form B out of the 50 possible items. The reverse cumulative score distribution for the two forms also showed that Form A was easier than Form B. Test characteristic curves for the two forms showed that Form A was slightly easier in the middle to upper portion of the ability continuum. Test information curves also indicated that Form A has higher information at the center portion of the ability continuum.

Classical item analysis and Rasch analysis was completed for the 89 unique items on Forms A and B. Eleven anchor items were administered in both Form A and Form B. The item analysis provided the following summary statistics. These item analysis statistics indicate that the test items are relatively easy with average p values of .84 and median p values of 0.86. The items show good correlations to the total score with average and median point biserials of 0.49. The items also show a Rasch measurement range from -1.43 to a maximum of +3.15 with an average of 0.0 and a standard deviation of 0.871.

Statistic	**Average**	**Std Dev**	**Minimum**	**Maximum**	**Median**
P value	0.837	0.102	0.34	0.95	0.86
Point Biserial	0.490	0.100	0.05	0.67	0.49
Rasch Measure	0.00	0.871	-1.43	3.15	-0.16

In this paper the Master's partial credit Rasch analysis was computed. The partial credit analysis indicated that six of the nine case scenario clusters showed acceptable fit between the empirical and modeled item response functions. Three of the case scenario clusters showed poor fit between the empirical and modeled item response functions. A comparison of the average Rasch measures at the item level and at the case group level indicated that four cases with individual item average Rasch measures were within measure values of 0.20 of the case group average Rasch measures. Each of the nine cases showed wide range variance of the Rasch measures.

Dimensionality of the results was verified with principal components analysis and a unidimensional construct was found to account for the examinee responses within each test form. In each form all items had significant loadings on the first principal component. The unidimensional first principal component accounted for approximately 25% of the variance in the item responses and the ratio of the first to second eigenvalues of the test forms were 4.4 and

4.2 for Forms A and B respectively. Subsequent eigenvalue ratios dropped to 1.74 and 1.3 for the ratio of second to third eigenvalues and 1.3 to 1.6 for all subsequent eigenvalues. These results suggest that a unidimensional principal measurement component is present within each of the two test forms.

Confirmatory factor analysis demonstrated that the item scores within each test form were unidimensional with two different IRT models: 1) a one factor weighted least squares with mean and variance correction using a two parameter IRT normal ogive model and 2) a one factor logistic regression model using the logistic form of the two parameter IRT model. All items loaded significantly on the primary unidimensional factor for both the IRT normal ogive model and the IRT logistic regression model. The most difficult item in Form A with a p value of 0.34 and an item total correlation of 0.05 had numerical estimation problems and was set to a difficulty value of +5.0 on the theta metric in comparison to the rest of the items.

The confirmatory factor analysis included a bi-factor analysis and a unidimensional factor was estimated. From the remaining residual variance six independent and orthogonal factors were estimated and modeled. For each test form the first dimension of the bi-factor model included items within the first case group and all remaining five case groups included in that form. The second dimension of the bi-factor model included the second case group and the remaining four case groups. The same factor dimension and modeling procedure was computed and the last dimension modeled included only the items from the sixth case group per form.

Form A included several significant positive cluster loadings for each case and a few negative cluster loadings for four of the six cases and an additional case with a near significant item case loading. Form B included several significant positive cluster loadings for each case and no negative loadings. The bi-factor analysis results demonstrated from the residual variance after the primary factor extraction, measurable variance that could attributed to each case group and separate independent and orthogonal dimensions could be modeled and item loadings could be estimated for each item on each of the separate independent orthogonal dimensions.

For both test forms there was a linear relationship between the factor loadings of the one factor normal ogive model and the bi-factor model indicating that each model was extracting the primary unidimensional latent factor for the examinees responses. A linear relationship was found between the standard errors for the one factor normal ogive model and the bi-factor model. The RSquared comparisons of the normal ogive and the bi-factor models showed that the IRT bi-factor model accounted for 8% in Form A and 10% in Form B more average variance and a median of 10% in Form A and 7% in Form B more variance than the IRT normal ogive model.

As expected, exponential relationship were found between the primary factor loadings and between the standard errors for the one factor IRT normal ogive model and the IRT logistic regression model.

Discussion

This paper investigates alternative approaches for analyzing item and task models to better understand the statistical relationships among individual test items and items that are grouped in parcels, bundles, clusters, families, or testlets. The literature review has identified a long standing interest in exploring items that are related within a domain or presentation unit and how these items and clusters can be analyzed for measurement purposes. We feel the systematic investigation of item and task models for item families, bundles or clusters has the potential to make a significant and meaningful scientific contribution to educational measurement. If the item family, bundle or cluster can be calibrated rather than individual test items then sibling items drawn from the family cluster could be administered without recalibration if the sibling item falls within the expected item family parameters. To provide meaningful educational or proficiency measurement value, the item family or cluster concept must satisfy the following conditions: 1) include representative samples from the target content or trait domains, 2) the families or clusters of items should measure different portions rather than the same portions of the ability or proficiency domain, and 3) the range and variation of the item family parameters for item difficulty, discrimination and model misfit should be sufficiently small to permit allocation and estimation of the parameters for each specific family or item cluster. When the item family or cluster is sufficiently defined then it is hypothesized that the item difficulty, discrimination and model misfit parameters for the item family is applicable to each sibling item drawn or generated from the item family or cluster.

Research within the same family or item cluster groups often shows substantial variations in p values and IRT proficiency estimates from items within the same family or testlet group. Research has also shown that items that were developed to be isomorphs often exhibited wider variation than expected in statistical difficulty and calibrations.

Test scores or IRT proficiency estimates are accumulated over a series of items or performance tasks administered. These item and score estimates can be reported at multiple levels of aggregation. The item family can be considered one of the aggregation levels for the test scores. The p values, point biserial correlations and Rasch measures were calculated for each of the individual test items. In a similar manner p values, point biserial correlations and Rasch measures can be calculated from aggregated scores from the individual case clusters rather than individual item scores. The p value can be computed as the average item score and referenced against the total number of points possible per item family or cluster. If the average item score is divided by the number of points possible an adjusted p value can be computed for the item family or cluster. The biserial correlation can be computed between the item cluster score and the total test score. With clustered item scores, the biserial correlation has a stronger theoretical base than the point biserial correlation. A partial credit or polytomous Rasch measure can also be computed for the item cluster score.

Since scores on each of the alternate forms showed high examinee performance, one alternative explanation for the presence of an underlying overall latent exam dimension and separate dimensions for the case clusters could be the presence of pre-knowledge or over exposure of items and case clusters. Some examinees may be responding to the exam based on their level of underlying knowledge of the content and other examinees may be responding based on pre-knowledge of the item or case based on content overexposure.

In preparing this paper the researchers noted the difficulty in finding and receiving permission for using common data sets for scientific investigation and cross-validation from different researchers for analysis of the item family concept. The research dataset used was not designed to test the item family concept but provided a good model for investigation of alternative item and task modeling approaches.

Appendix A

Table 2. Item Analysis and Rasch Measures at the Item Level.

(Lines in the body of the table separate the measurement of specific content objectives from the blueprint)

Count	Case	Item ID	Rasch Measure	MNSQ Infit	MNSQ Outfit	P-value	Item-Score Correlation	Item Reliability	Number of Responses	Time
1	B	1.1.aba	-0.85	0.96	0.66	0.92	0.49	0.14	1270	27.0
2	G	1.1.ADA	-1.17	1.03	1.26	0.94	0.41	0.10	631	26.0
3	B	1.2.abb	-0.17	0.89	0.69	0.87	0.57	0.19	1270	22.0
4	B	1.2.ADB	-1.32	1.02	1.14	0.95	0.42	0.09	631	25.0
5	C	1.2.AEB	0.43	0.91	0.85	0.83	0.59	0.22	631	34.0
6	A	1.3.aac	0.02	0.99	0.79	0.84	0.50	0.19	639	33.0
7	C	1.3.aec	0.80	0.99	0.82	0.78	0.54	0.22	631	53.0
8	A	1.4.aad	0.48	0.81	0.62	0.79	0.63	0.26	639	37.0
9	B	1.4.abd	-0.89	1.18	1.17	0.91	0.32	0.09	639	34.0
10	G	1.4.ADC	0.77	1.23	1.07	0.75	0.34	0.15	639	33.0
11	C	2.1.aea	0.92	0.83	0.68	0.77	0.64	0.27	631	44.0
12	D	2.1.aha	-1.32	0.98	0.83	0.94	0.45	0.11	639	37.0
13	E	2.1.AIA	-0.43	1.01	0.91	0.90	0.49	0.15	631	28.0
14	I	2.2.AGB	-0.46	0.84	0.71	0.88	0.60	0.19	639	29.0
15	D	2.2.ahb	-1.43	0.93	0.46	0.94	0.50	0.12	639	32.0
16	F	2.2.aja	-0.50	1.17	1.14	0.90	0.36	0.11	631	25.0
17	H	2.3.afd	-0.01	1.05	1.26	0.87	0.47	0.16	631	35.0
18	I	2.3.AGD	-0.19	0.95	1.08	0.86	0.53	0.18	639	32.0
19	D	2.3.ahc	-0.70	1.12	1.36	0.90	0.36	0.11	639	31.0
20	D	2.3.ahd	0.84	1.10	1.07	0.74	0.43	0.19	639	56.0
21	E	2.3.AIC	3.15	1.71	2.39	0.34	0.05	0.02	639	60.0
22	E	2.3.AID	0.13	0.98	0.96	0.85	0.53	0.19	631	38.0
23	F	2.3.ajb	0.47	1.12	1.08	0.82	0.46	0.18	631	29.0
24	I	2.4.AGE	0.00	0.92	0.82	0.84	0.55	0.20	639	28.0
25	D	2.4.ahf	0.40	0.80	0.65	0.80	0.64	0.26	639	51.0
26	E	2.4.AIE	-1.13	0.96	0.81	0.94	0.48	0.12	631	30.0
27	F	2.4.ajd	0.82	0.98	0.83	0.78	0.55	0.23	631	31.0
28	H	2.5.aff	0.40	1.03	1.00	0.83	0.51	0.19	631	43.0
29	I	2.5.AGF	1.05	0.84	0.72	0.71	0.59	0.27	639	40.0
30	D	2.5.ahg	0.06	0.98	0.97	0.83	0.51	0.19	639	56.0
31	F	2.5.aje	-0.60	0.97	0.75	0.91	0.52	0.15	631	30.0
32	I	2.6.AGG	-0.99	0.99	1.37	0.92	0.44	0.12	639	39.0
33	D	2.6.ahh	0.62	0.91	0.88	0.77	0.55	0.23	639	54.0
34	E	2.6.AIH	0.80	1.27	1.53	0.78	0.39	0.16	631	49.0
35	F	2.6.ajf	-0.32	0.98	0.93	0.89	0.51	0.16	631	30.0
36	F	2.6.ajg	1.00	0.87	0.79	0.76	0.62	0.27	631	40.0
37	A	3.1.aae	-0.44	0.98	0.93	0.88	0.49	0.16	639	36.0
38	A	3.1.aaf	0.52	0.92	0.84	0.78	0.56	0.23	639	29.0
39	B	3.1.abe	1.54	1.14	1.08	0.63	0.38	0.19	639	80.0
40	G	3.1.ADD	-0.40	1.10	1.52	0.90	0.40	0.12	631	26.0
41	G	3.1.ADE	-0.65	1.09	1.08	0.91	0.41	0.12	631	31.0
42	C	3.1.AEE	-0.01	1.10	1.43	0.87	0.44	0.15	631	68.0
43	A	3.2.aag	-0.16	1.01	0.84	0.86	0.49	0.17	639	53.0
44	B	3.2.abg	-0.19	1.15	0.98	0.86	0.38	0.13	639	57.0
45	G	3.2.ADF	-0.70	0.97	1.13	0.92	0.48	0.13	631	28.0
46	C	3.2.AEG	1.14	0.89	0.83	0.74	0.61	0.27	631	35.0

Count	Case	Item ID	Rasch Measure	MNSQ Infit	MNSQ Outfit	P-value	Item-Score Correlation	Item Reliability	Number of Responses	Time
47	A	3.3.aah	0.11	1.06	0.99	0.83	0.45	0.17	639	49.0
48	B	3.3.abh	2.25	1.06	1.24	0.50	0.37	0.18	639	59.0
49	G	3.3.ADG	0.82	1.25	1.42	0.78	0.40	0.16	631	51.0
50	G	3.3.ADH	2.16	1.16	1.42	0.52	0.33	0.16	639	56.0
51	C	3.3.AEH	-0.23	0.97	0.77	0.88	0.53	0.17	631	27.0
52	C	3.3.AEI	-0.07	0.91	0.81	0.87	0.57	0.19	631	29.0
53	A	3.4.aaj	-0.19	1.03	0.92	0.86	0.47	0.16	639	45.0
54	A	3.4.aak	-0.31	1.06	1.20	0.87	0.43	0.14	639	44.0
55	B	3.4.abj	1.10	1.21	1.18	0.70	0.34	0.16	639	35.0
56	G	3.4.ADI	-0.67	0.89	0.59	0.91	0.57	0.16	631	25.0
57	G	3.4.ADJ	-0.43	1.18	1.74	0.90	0.35	0.11	631	26.0
58	C	3.4.AEK	-0.23	1.01	1.21	0.88	0.49	0.16	631	38.0
59	H	4.1.afi	0.53	0.92	0.89	0.81	0.58	0.23	631	29.0
60	I	4.1.AGH	0.17	0.77	0.53	0.82	0.66	0.25	639	32.0
61	D	4.1.ahi	0.72	1.04	1.16	0.75	0.48	0.21	639	42.0
62	E	4.1.AII	0.70	0.93	0.91	0.78	0.56	0.23	1270	55.0
63	F	4.1.ajh	-0.43	0.88	0.61	0.90	0.58	0.17	631	26.0
64	D	4.2.ahj	-0.65	0.97	0.74	0.90	0.49	0.15	639	34.0
65	F	4.2.aji	-0.84	1.20	4.06	0.92	0.27	0.07	631	41.0
66	I	4.3.AGJ	-0.31	0.86	0.93	0.87	0.58	0.20	639	34.0
67	D	4.3.ahk	0.26	0.98	0.99	0.81	0.51	0.20	639	35.0
68	E	4.3.AIK	-1.06	1.01	1.14	0.94	0.44	0.11	631	26.0
69	F	4.3.ajj	-0.65	0.96	0.84	0.91	0.50	0.14	631	29.0
70	H	4.4.afL	0.79	0.96	0.90	0.79	0.56	0.23	631	58.0
71	H	4.4.afm	0.23	0.92	0.93	0.84	0.58	0.21	631	37.0
72	D	4.4.ahl	-0.24	0.81	0.62	0.86	0.63	0.22	639	31.0
73	D	4.4.ahm	0.11	0.93	0.89	0.83	0.55	0.21	639	35.0
74	E	4.4.AIL	0.84	0.83	0.73	0.76	0.62	0.27	1270	54.0
75	E	4.4.AIM	1.54	1.00	0.89	0.63	0.47	0.23	639	58.0
76	F	4.4.ajl	-0.79	1.00	1.21	0.92	0.46	0.12	631	39.0
77	A	5.1.aaL	-0.50	0.94	0.79	0.89	0.53	0.17	639	43.0
78	B	5.1.abk	-1.09	0.98	1.16	0.93	0.45	0.12	1270	28.0
79	B	5.1.abl	-1.15	0.94	0.71	0.93	0.50	0.12	1270	19.0
80	G	5.1.ADL	-0.52	0.83	0.63	0.90	0.61	0.18	631	17.0
81	C	5.1.AEL	0.30	0.87	0.74	0.84	0.61	0.22	631	31.0
82	A	5.2.aam	-0.94	0.97	0.67	0.92	0.48	0.13	639	34.0
83	B	5.2.abm	1.01	1.17	1.29	0.74	0.42	0.19	1270	43.0
84	G	5.2.ADM	-0.14	0.91	0.66	0.87	0.56	0.19	1270	20.0
85	C	5.2.AEO	0.68	0.78	0.58	0.80	0.67	0.27	631	35.0
86	A	5.3.aao	-0.81	0.99	0.81	0.91	0.47	0.14	639	26.0
87	B	5.3.abn	-0.60	1.10	0.95	0.90	0.40	0.12	1270	23.0
88	G	5.3.ADN	-1.38	0.90	0.56	0.94	0.51	0.12	1270	20.0
89	G	5.3.ADO	-1.42	0.96	1.35	0.94	0.44	0.10	1270	26.0

Appendix B
Form A Weighted Least Squares Normal Ogive Model with Mean and Variance Correction

Item Variable	CASE	Estimate	SE	Discrimination	Difficulty	RSquare	SE
V11ABA	**B**	0.712	0.050	1.013	-2.042	0.507	0.071
V12ABB	**B**	0.821	0.030	1.439	-1.335	0.674	0.049
V51ABK	**B**	0.705	0.066	0.994	-2.202	0.497	0.093
V51ABL	**B**	0.776	0.045	1.232	-1.966	0.603	0.070
V52ABM	**B**	0.550	0.045	0.659	-1.204	0.303	0.049
V53ABN	**B**	0.594	0.059	0.739	-2.336	0.353	0.070
V12AEB	**C**	0.751	0.035	1.139	-1.246	0.565	0.052
V13AEC	**C**	0.698	0.037	0.974	-1.135	0.487	0.052
V21AEA	**C**	0.800	0.030	1.332	-0.930	0.640	0.047
V31AEE	**C**	0.601	0.053	0.753	-1.859	0.362	0.063
V32AEG	**C**	0.788	0.029	1.280	-0.834	0.621	0.046
V33AEH	**C**	0.727	0.041	1.058	-1.657	0.528	0.060
V33AEI	**C**	0.769	0.038	1.205	-1.473	0.592	0.058
V34AEK	**C**	0.679	0.048	0.926	-1.773	0.461	0.065
V51AEL	**C**	0.797	0.031	1.321	-1.237	0.636	0.050
V52AEO	**C**	0.843	0.024	1.566	-0.999	0.710	0.041
V21AIA	**E**	0.682	0.049	0.933	-1.866	0.465	0.066
V23AID	**E**	0.715	0.040	1.023	-1.473	0.512	0.057
V24AIE	**E**	0.725	0.058	1.054	-2.122	0.526	0.084
V26AIH	**E**	0.497	0.051	0.573	-1.581	0.247	0.050
V41AII	**E**	0.699	0.036	0.976	-1.149	0.488	0.051
V43AIK	**E**	0.678	0.060	0.923	-2.231	0.460	0.082
V44AIL	**E**	0.796	0.028	1.315	-0.876	0.634	0.045
V22AJA	**F**	0.571	0.059	0.695	-2.277	0.326	0.068
V23AJB	**F**	0.616	0.044	0.782	-1.490	0.379	0.054
V24AJD	**F**	0.719	0.033	1.034	-1.086	0.517	0.048
V25AJE	**F**	0.735	0.043	1.085	-1.819	0.541	0.063
V26AJF	**F**	0.698	0.048	0.974	-1.762	0.487	0.067
V26AJG	**F**	0.789	0.029	1.285	-0.896	0.623	0.045
V41AJH	**F**	0.776	0.038	1.230	-1.640	0.602	0.060
V42AJI	**F**	0.424	0.094	0.468	-3.376	0.180	0.080
V43AJJ	**F**	0.720	0.048	1.037	-1.886	0.518	0.069
V44AJL	**F**	0.678	0.057	0.922	-2.079	0.460	0.078
V11ADA	**G**	0.650	0.068	0.855	-2.388	0.423	0.089
V12ADB	**G**	0.678	0.068	0.923	-2.371	0.460	0.093
V31ADD	**G**	0.578	0.062	0.708	-2.187	0.334	0.071
V31ADE	**G**	0.611	0.058	0.773	-2.220	0.374	0.070

V32ADF	**G**	0.709	0.051	1.006	-1.942	0.503	0.073
V33ADG	**G**	0.510	0.050	0.594	-1.530	0.261	0.051
V34ADI	**G**	0.794	0.039	1.306	-1.723	0.630	0.062
V34ADJ	**G**	0.512	0.066	0.596	-2.486	0.262	0.067
V51ADL	**G**	0.831	0.032	1.496	-1.575	0.691	0.054
V52ADM	**G**	0.767	0.035	1.195	-1.508	0.588	0.053
V53ADN	**G**	0.803	0.051	1.347	-2.020	0.645	0.081
V53ADO	**G**	0.703	0.073	0.988	-2.351	0.494	0.103
V23AFD	**H**	0.649	0.049	0.854	-1.710	0.422	0.063
V25AFF	**H**	0.657	0.042	0.870	-1.445	0.431	0.055
V41AFI	**H**	0.783	0.031	1.258	-1.142	0.613	0.049
V44AFL	**H**	0.729	0.034	1.064	-1.087	0.531	0.050
V44AFM	**H**	0.775	0.033	1.225	-1.308	0.600	0.052
AVERAGE		0.697	0.046	1.020	-1.697	0.495	0.063
STD DEV		0.095	0.014	0.255	0.524	0.126	0.014
MIN		0.424	0.024	0.468	-3.376	0.180	0.041
MAX		0.843	0.094	1.566	-0.834	0.710	0.103
MEDIAN		0.711	0.045	1.010	-1.684	0.505	0.061

Form B Weighted Least Squares Normal Ogive Model with Mean and Variance Correction

Item Variable	CASE	Estimate	SE	Discrimination	Difficulty	RSquare	SE
V13AAC	A	0.688	0.041	0.947	1.437	0.473	0.056
V14AAD	A	0.832	0.025	1.498	0.950	0.692	0.042
V31AAE	A	0.673	0.049	0.911	1.751	0.453	0.065
V31AAF	A	0.733	0.034	1.077	1.056	0.537	0.050
V32AAG	A	0.672	0.045	0.908	1.579	0.452	0.060
V33AAH	A	0.604	0.048	0.758	1.574	0.365	0.057
V34AAJ	A	0.621	0.050	0.792	1.732	0.386	0.062
V34AAK	A	0.583	0.056	0.717	1.932	0.340	0.065
V51AAL	A	0.740	0.043	1.099	1.627	0.547	0.064
V52AAM	A	0.706	0.053	0.998	1.946	0.499	0.075
V53AAO	A	0.685	0.053	0.940	1.935	0.469	0.072
V11ABA	B	0.744	0.048	1.114	1.782	0.554	0.071
V12ABB	B	0.784	0.037	1.262	1.456	0.614	0.058
V14ABD	B	0.476	0.075	0.542	2.843	0.227	0.071
V31ABE	B	0.459	0.045	0.517	0.705	0.211	0.041
V32ABG	B	0.510	0.060	0.594	2.107	0.261	0.061
V33ABH	B	0.457	0.047	0.513	0.009	0.208	0.043
V34ABJ	B	0.420	0.050	0.463	1.241	0.177	0.042
V51ABK	B	0.668	0.059	0.899	2.103	0.447	0.079
V51ABL	B	0.748	0.055	1.126	1.969	0.559	0.082
V52ABM	B	0.496	0.046	0.571	1.209	0.246	0.046
V53ABN	B	0.601	0.054	0.753	2.001	0.362	0.065
V21AHA	D	0.691	0.063	0.956	2.199	0.478	0.088
V22AHB	D	0.799	0.048	1.327	1.952	0.638	0.077
V23AHC	D	0.521	0.068	0.611	2.454	0.272	0.071
V23AHD	D	0.541	0.045	0.643	1.171	0.292	0.048
V24AHF	D	0.833	0.026	1.506	0.994	0.694	0.044
V25AHG	D	0.675	0.042	0.914	1.437	0.455	0.057
V26AHH	D	0.718	0.037	1.033	1.019	0.516	0.054
V41AHI	D	0.622	0.041	0.793	1.105	0.386	0.051
V42AHJ	D	0.697	0.049	0.972	1.811	0.486	0.068
V43AHK	D	0.657	0.042	0.871	1.348	0.431	0.056
V44AHL	D	0.836	0.032	1.523	1.312	0.699	0.053
V44AHM	D	0.727	0.038	1.060	1.307	0.529	0.055
V23AIC	E	0.001	0.053	0.001	5.000*	0.000	0.000
V41AII	E	0.743	0.034	1.109	0.965	0.552	0.050
V44AIL	E	0.815	0.026	1.407	0.867	0.664	0.042

V44AIM	E	0.571	0.039	0.696	0.559	0.327	0.045
V14ADC	G	0.417	0.052	0.458	1.601	0.174	0.043
V33ADH	G	0.369	0.050	0.398	0.106	0.136	0.037
V52ADM	G	0.740	0.039	1.100	1.426	0.547	0.058
V53ADN	G	0.781	0.052	1.252	1.978	0.611	0.082
V53ADO	G	0.726	0.063	1.055	2.129	0.527	0.092
V22AGB	I	0.800	0.038	1.332	1.485	0.639	0.061
V23AGD	I	0.704	0.043	0.990	1.529	0.495	0.061
V24AGE	I	0.745	0.038	1.116	1.336	0.554	0.057
V25AGF	I	0.776	0.031	1.232	0.701	0.603	0.049
V26AGG	I	0.660	0.061	0.879	2.113	0.436	0.081
V41AGH	I	0.843	0.025	1.570	1.098	0.711	0.043
V43AGJ	I	0.772	0.039	1.213	1.459	0.596	0.060
AVERAGE		0.654	0.046	0.940	13.278	0.451	0.058
STD DEV		0.154	0.011	0.330	83.440	0.167	0.016
MIN		0.001	0.025	0.001	0.009	0.000	0.000
MAX		0.843	0.075	1.570	5.00*	0.711	0.092
MEDIAN		0.690	0.046	0.952	1.472	0.476	0.058

***numerical problems in estimation set to value of 5.00**

Form A Weighted Least Squares with mean and Variance Correction
Bi-Factor Solution (Primary Factor Plus Independent Factor for each Case)

Item		**General**		**Cluster**				**Residual**
Variable	**CASE**	**Loading**	**SE**	**Loading**	**SE**	**RSquare**	**SE**	**Variance**
V11ABA	B	0.703	0.051	0.455	0.107	0.701	0.100	0.299
V12ABB	B	0.836	0.030	-0.361	0.112	0.829	0.100	0.171
V51ABK	**B**	0.699	0.067	0.345	0.098	0.608	0.101	0.392
V51ABL	**B**	0.776	0.046	0.196	0.101	0.641	0.075	0.359
V52ABM	**B**	0.551	0.045	0.128	0.092	0.320	0.052	0.680
V53ABN	**B**	0.582	0.060	0.540	0.107	0.630	0.126	0.370
V12AEB	**C**	0.747	0.034	0.142	0.076	0.579	0.055	0.421
V13AEC	**C**	0.719	0.036	-0.277	0.110	0.593	0.076	0.407
V21AEA	**C**	0.798	0.029	0.099	0.073	0.647	0.049	0.353
V31AEE	**C**	0.603	0.053	0.032	0.098	0.365	0.064	0.635
V32AEG	**C**	0.770	0.031	0.396	0.078	0.749	0.066	0.251
V33AEH	**C**	0.724	0.042	0.121	0.086	0.539	0.061	0.461
V33AEI	**C**	0.754	0.039	0.330	0.080	0.677	0.067	0.323
V34AEK	**C**	0.682	0.048	0.027	0.098	0.466	0.065	0.534
V51AEL	**C**	0.782	0.032	0.316	0.071	0.712	0.060	0.288
V52AEO	**C**	0.836	0.025	0.191	0.068	0.735	0.046	0.265
V21AIA	**E**	0.681	0.049	0.234	0.107	0.518	0.076	0.482
V23AID	**E**	0.715	0.041	0.232	0.095	0.565	0.070	0.435
V24AIE	**E**	0.729	0.058	0.019	0.112	0.533	0.084	0.467
V26AIH	**E**	0.513	0.050	-0.470	0.149	0.484	0.149	0.516
V41AII	**E**	0.698	0.037	0.263	0.093	0.556	0.066	0.444
V43AIK	**E**	0.686	0.060	-0.159	0.122	0.496	0.094	0.504
V44AIL	**E**	0.796	0.029	0.268	0.087	0.706	0.059	0.294
V22AJA	**F**	0.533	0.064	0.630	0.097	0.680	0.107	0.320
V23AJB	**F**	0.586	0.047	0.585	0.085	0.685	0.094	0.315
V24AJD	**F**	0.703	0.035	0.400	0.075	0.655	0.059	0.345
V25AJE	**F**	0.747	0.043	-0.199	0.091	0.598	0.072	0.402
V26AJF	**F**	0.700	0.049	0.034	0.081	0.492	0.068	0.508
V26AJG	**F**	0.800	0.029	-0.138	0.072	0.659	0.053	0.341
V41AJH	**F**	0.770	0.040	0.224	0.075	0.643	0.058	0.357
V42AJI	**F**	0.427	0.094	-0.023	0.078	0.183	0.079	0.817
V43AJJ	**F**	0.713	0.050	0.235	0.079	0.563	0.071	0.437
V44AJL	**F**	0.677	0.058	0.117	0.080	0.471	0.079	0.529
V11ADA	**G**	0.633	0.070	0.300	0.110	0.491	0.104	0.509
V12ADB	**G**	0.639	0.072	0.530	0.095	0.690	0.110	0.310
V31ADD	**G**	0.556	0.063	0.371	0.090	0.447	0.086	0.553

V31ADE	**G**	0.587	0.059	0.387	0.090	0.495	0.087	0.505
V32ADF	**G**	0.680	0.055	0.456	0.086	0.671	0.077	0.329
V33ADG	**G**	0.512	0.051	0.039	0.100	0.263	0.051	0.737
V34ADI	**G**	0.796	0.040	0.049	0.100	0.636	0.063	0.364
V34ADJ	**G**	0.501	0.066	0.222	0.097	0.300	0.081	0.700
V51ADL	**G**	0.839	0.032	-0.044	0.102	0.705	0.056	0.295
V52ADM	**G**	0.771	0.035	0.013	0.098	0.594	0.054	0.406
V53ADN	**G**	0.791	0.052	0.255	0.093	0.691	0.088	0.309
V53ADO	**G**	0.666	0.077	0.499	0.085	0.692	0.111	0.308
V23AFD	**H**	0.655	0.049	-0.051	0.097	0.432	0.065	0.568
V25AFF	**H**	0.671	0.042	-0.301	0.117	0.541	0.088	0.459
V41AFI	**H**	0.773	0.033	0.468	0.100	0.817	0.090	0.183
V44AFL	**H**	0.724	0.035	0.243	0.077	0.584	0.057	0.416
V44AFM	**H**	0.765	0.035	0.366	0.087	0.719	0.071	0.281
AVERAGE		0.692	0.047	0.175	0.093	0.581	0.077	0.419
STD DEV		0.097	0.015	0.252	0.015	0.137	0.021	0.137
MIN		0.427	0.025	-0.470	0.068	0.183	0.046	0.171
MAX		0.839	0.094	0.630	0.149	0.829	0.149	0.817
MEDIAN		0.703	0.047	0.223	0.093	0.596	0.072	0.404

Form B Weighted Least Squares with mean and Variance Correction
Bi-Factor Solution (Primary Factor Plus Independent Factor for each Case)

Item		General		Cluster			Residual
Variable	CASE	Loading	SE	Loading	SE	RSquare	Variance
V13AAC	A	0.689	0.041	0.072	0.085	0.479	0.521
V14AAD	A	0.832	0.026	0.444	0.078	0.889	0.111
V31AAE	A	0.676	0.049	0.228	0.087	0.509	0.491
V31AAF	A	0.730	0.034	0.208	0.068	0.576	0.424
V32AAG	A	0.669	0.045	0.274	0.096	0.522	0.478
V33AAH	A	0.604	0.048	0.118	0.090	0.378	0.622
V34AAJ	A	0.626	0.049	0.387	0.086	0.542	0.458
V34AAK	A	0.587	0.056	0.311	0.082	0.441	0.559
V51AAL	A	0.741	0.043	0.036	0.082	0.550	0.450
V52AAM	A	0.709	0.053	0.205	0.084	0.544	0.456
V53AAO	A	0.688	0.053	0.245	0.083	0.534	0.466
V11ABA	B	0.748	0.047	0.087	0.066	0.568	0.432
V12ABB	B	0.787	0.037	0.076	0.056	0.626	0.374
V14ABD	B	0.458	0.077	0.357	0.076	0.337	0.663
V31ABE	B	0.418	0.048	0.664	0.064	0.615	0.385
V32ABG	B	0.474	0.064	0.584	0.064	0.565	0.435
V33ABH	B	0.444	0.048	0.240	0.068	0.255	0.745
V34ABJ	B	0.370	0.053	0.727	0.065	0.666	0.334
V51ABK	B	0.674	0.059	0.100	0.070	0.464	0.536
V51ABL	B	0.749	0.054	0.012	0.067	0.562	0.438
V52ABM	B	0.485	0.048	0.238	0.068	0.292	0.708
V53ABN	B	0.611	0.053	0.183	0.074	0.406	0.594
V21AHA	D	0.693	0.063	0.371	0.088	0.618	0.382
V22AHB	D	0.797	0.048	0.146	0.092	0.657	0.343
V23AHC	D	0.524	0.068	0.463	0.097	0.488	0.512
V23AHD	D	0.545	0.045	0.313	0.088	0.395	0.605
V24AHF	D	0.831	0.027	0.302	0.068	0.783	0.217
V25AHG	D	0.679	0.042	0.268	0.081	0.533	0.467
V26AHH	D	0.717	0.038	0.219	0.077	0.562	0.438
V41AHI	D	0.623	0.041	0.079	0.079	0.394	0.606
V42AHJ	D	0.700	0.049	0.204	0.083	0.532	0.468
V43AHK	D	0.657	0.042	0.058	0.079	0.435	0.565
V44AHL	D	0.834	0.032	0.209	0.070	0.740	0.260
V44AHM	D	0.727	0.038	0.136	0.076	0.548	0.452
V23AIC	E	0.026	0.053	0.794	0.241	0.632	0.368
V41AII	E	0.739	0.034	0.204	0.076	0.587	0.413
V44AIL	E	0.811	0.027	0.260	0.081	0.726	0.274

V44AIM	**E**	0.565	0.040	0.226	0.088	0.370	0.630
V14ADC	**G**	0.406	0.053	0.576	0.202	0.497	0.503
V33ADH	**G**	0.363	0.051	0.416	0.147	0.305	0.695
V52ADM	**G**	0.746	0.039	0.235	0.102	0.611	0.389
V53ADN	**G**	0.782	0.052	0.063	0.092	0.615	0.385
V53ADO	**G**	0.725	0.063	0.143	0.100	0.547	0.453
V22AGB	**I**	0.802	0.038	0.179	0.089	0.675	0.325
V23AGD	**I**	0.705	0.043	0.055	0.091	0.500	0.500
V24AGE	**I**	0.746	0.038	0.099	0.091	0.566	0.434
V25AGF	**I**	0.781	0.031	0.392	0.135	0.764	0.236
V26AGG	**I**	0.664	0.062	0.393	0.134	0.595	0.405
V41AGH	**I**	0.844	0.025	0.093	0.076	0.721	0.279
V43AGJ	**I**	0.774	0.039	0.215	0.095	0.646	0.354
AVERAGE		0.652	0.046	0.258	0.090	0.547	0.453
STD DEV		0.157	0.011	0.180	0.033	0.131	0.131
MIN		0.026	0.025	0.012	0.056	0.255	0.111
MAX		0.844	0.077	0.794	0.241	0.889	0.745
MEDIAN		0.691	0.048	0.223	0.083	0.549	0.451

Form A Weighted Least Squares One Factor Logistic Regression Model (2 Parameter IRT)

Item Variable	CASE	Estimate	SE	Discrimination	Difficulty	RSquare	SE
V13AAC	A	1.879	0.236	1.105	-1.960	0.518	0.063
V14AAD	A	2.735	0.336	1.609	-1.272	0.695	0.052
V31AAE	A	1.671	0.305	0.983	-2.239	0.459	0.091
V31AAF	A	2.339	0.379	1.376	-1.867	0.624	0.076
V32AAG	A	1.042	0.141	0.613	-1.239	0.248	0.050
V33AAH	A	1.354	0.179	0.797	-2.251	0.358	0.061
V34AAJ	A	2.174	0.259	1.279	-1.176	0.590	0.058
V34AAK	A	1.809	0.195	1.064	-1.074	0.499	0.054
V51AAL	A	2.599	0.290	1.529	-0.881	0.673	0.049
V52AAM	A	1.297	0.191	0.763	-1.838	0.338	0.066
V53AAO	A	2.225	0.245	1.309	-0.819	0.601	0.053
V11ABA	B	2.078	0.265	1.222	-1.541	0.568	0.063
V12ABB	B	2.245	0.308	1.321	-1.405	0.605	0.066
V14ABD	B	1.651	0.237	0.971	-1.722	0.453	0.071
V31ABE	B	2.487	0.312	1.463	-1.181	0.653	0.057
V32ABG	B	3.088	0.342	1.817	-0.953	0.744	0.042
V33ABH	B	1.781	0.232	1.048	-1.753	0.491	0.065
V34ABJ	B	1.858	0.237	1.093	-1.415	0.512	0.064
V51ABK	B	1.876	0.293	1.103	-2.085	0.517	0.078
V51ABL	B	0.950	0.135	0.559	-1.582	0.215	0.048
V52ABM	B	1.816	0.206	1.068	-1.087	0.501	0.057
V53ABN	B	1.675	0.261	0.985	-2.175	0.460	0.077
V21AHA	D	2.338	0.258	1.375	-0.854	0.624	0.052
V22AHB	D	1.139	0.168	0.670	-2.355	0.283	0.060
V23AHC	D	1.380	0.167	0.812	-1.441	0.367	0.056
V23AHD	D	1.888	0.202	1.110	-1.039	0.520	0.053
V24AHF	D	2.149	0.288	1.264	-1.690	0.584	0.065
V25AHG	D	1.793	0.254	1.055	-1.685	0.494	0.071
V26AHH	D	2.318	0.253	1.364	-0.869	0.620	0.051
V41AHI	D	2.360	0.310	1.388	-1.548	0.629	0.061
V42AHJ	D	0.579	0.218	0.341	-4.545	0.093	0.063
V43AHK	D	1.853	0.257	1.090	-1.837	0.511	0.069
V44AHL	D	1.621	0.255	0.953	-2.052	0.444	0.078
V44AHM	D	1.559	0.248	0.917	-2.330	0.425	0.078
V23AIC	E	1.576	0.272	0.927	-2.407	0.430	0.085
V41AII	E	1.204	0.190	0.708	-2.199	0.306	0.067
V44AIL	E	1.417	0.198	0.834	-2.135	0.379	0.066

V44AIM	E	1.649	0.261	0.970	-1.984	0.453	0.078
V14ADC	G	1.025	0.139	0.603	-1.485	0.242	0.050
V33ADH	G	2.587	0.390	1.521	-1.609	0.670	0.067
V52ADM	G	1.045	0.170	0.615	-2.449	0.249	0.061
V53ADN	G	2.999	0.441	1.764	-1.473	0.732	0.058
V53ADO	G	2.379	0.275	1.399	-1.403	0.632	0.054
V22AGB	I	2.579	0.412	1.517	-1.915	0.669	0.071
V23AGD	I	1.484	0.304	0.873	-2.574	0.401	0.098
V24AGE	I	1.503	0.211	0.884	-1.669	0.407	0.068
V25AGF	I	1.565	0.186	0.921	-1.387	0.427	0.058
V26AGG	I	2.243	0.284	1.319	-1.110	0.605	0.060
V41AGH	I	1.968	0.227	1.158	-1.035	0.541	0.057
V43AGJ	I	2.221	0.292	1.307	-1.261	0.600	0.063
AVERAGE		1.861	0.254	1.089	-1.677	0.493	0.064
STD DEV		0.549	0.068	0.323	0.633	0.147	0.011
MIN		0.579	0.135	0.341	-4.545	0.093	0.042
MAX		3.088	0.441	1.817	-0.819	0.744	0.098
MEDIAN		1.835	0.255	1.068	-1.596	0.506	0.063

Form B Weighted Least Squares One Factor Logistic Regression Model (2 Parameter IRT)

Item Variable	CASE	Estimate	SE	Discrimination	Difficulty	RSquare	SE
V13AAC	A	1.788	0.209	1.052	1.354	0.493	0.058
V14AAD	A	2.868	0.331	1.687	0.897	0.714	0.047
V31AAE	A	1.590	0.208	0.936	1.729	0.435	0.064
V31AAF	A	1.952	0.227	1.148	1.013	0.537	0.058
V32AAG	A	1.640	0.208	0.965	1.524	0.450	0.063
V33AAH	A	1.348	0.172	0.793	1.523	0.356	0.058
V34AAJ	A	1.328	0.174	0.781	1.747	0.349	0.06
V34AAK	A	1.117	0.177	0.657	2.058	0.275	0.063
V51AAL	A	1.878	0.257	1.105	1.620	0.517	0.068
V52AAM	A	1.747	0.233	1.028	1.933	0.481	0.067
V53AAO	A	1.622	0.222	0.954	1.935	0.444	0.068
V11ABA	B	1.955	0.265	1.150	1.759	0.537	0.067
V12ABB	B	2.246	0.272	1.321	1.417	0.605	0.058
V14ABD	B	0.958	0.165	0.563	2.812	0.218	0.059
V31ABE	B	0.843	0.117	0.496	0.709	0.178	0.041
V32ABG	B	0.937	0.143	0.551	2.229	0.211	0.051
V33ABH	B	1.037	0.130	0.610	0.004	0.246	0.047
V34ABJ	B	0.735	0.114	0.433	1.278	0.141	0.038
V51ABK	B	1.534	0.232	0.903	2.127	0.417	0.073
V51ABL	B	1.916	0.293	1.127	1.986	0.527	0.076
V52ABM	B	0.944	0.125	0.556	1.211	0.213	0.044
V53ABN	B	1.249	0.174	0.735	2.046	0.322	0.061
V21AHA	D	1.579	0.248	0.929	2.278	0.431	0.077
V22AHB	D	2.344	0.290	1.379	1.933	0.626	0.058
V23AHC	D	0.986	0.173	0.580	2.584	0.228	0.062
V23AHD	D	1.111	0.137	0.653	1.139	0.273	0.049
V24AHF	D	2.829	0.345	1.664	0.944	0.709	0.05
V25AHG	D	1.565	0.195	0.921	1.423	0.427	0.061
V26AHH	D	1.896	0.231	1.116	0.970	0.522	0.061
V41AHI	D	1.253	0.160	0.737	1.142	0.323	0.056
V42AHJ	D	1.690	0.216	0.994	1.796	0.465	0.064
V43AHK	D	1.427	0.180	0.839	1.366	0.382	0.059
V44AHL	D	2.679	0.358	1.576	1.277	0.686	0.058
V44AHM	D	1.820	0.233	1.070	1.290	0.502	0.064
V23AIC	E	0.144	0.086	0.084	4.596	0.006	0.007
V41AII	E	2.034	0.238	1.196	0.921	0.557	0.058
V44AIL	E	2.547	0.282	1.498	0.833	0.663	0.05
V44AIM	E	1.251	0.142	0.736	0.531	0.322	0.05

V14ADC	**G**	0.717	0.114	0.421	1.677	0.135	0.037
V33ADH	**G**	0.716	0.109	0.421	0.100	0.135	0.035
V52ADM	**G**	1.968	0.245	1.158	1.382	0.541	0.062
V53ADN	**G**	2.191	0.287	1.289	1.969	0.593	0.063
V53ADO	**G**	1.770	0.288	1.041	2.179	0.488	0.081
V22AGB	**I**	2.247	0.316	1.322	1.476	0.605	0.067
V23AGD	**I**	1.734	0.225	1.020	1.500	0.477	0.065
V24AGE	**I**	1.944	0.254	1.144	1.309	0.535	0.065
V25AGF	**I**	2.374	0.280	1.396	0.658	0.631	0.055
V26AGG	**I**	1.440	0.235	0.847	2.191	0.386	0.077
V41AGH	**I**	3.020	0.334	1.776	1.038	0.735	0.043
V43AGJ	**I**	2.020	0.280	1.188	1.461	0.554	0.068
AVERAGE		1.651	0.219	0.971	1.537	0.432	0.058
STD DEV		0.618	0.068	0.364	0.729	0.174	0.013
MIN		0.144	0.086	0.084	0.004	0.006	0.007
MAX		3.020	0.358	1.776	4.596	0.735	0.081
MEDIAN		1.665	0.226	0.980	1.469	0.458	0.060

References

Adams, R. A., & Wilson, M. (1992). *A random coefficients multinomial logit: Generalizing Rasch models*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Bejar, I.I., Lawless, R.R., Morley, M.E., Wagner, M.E., Bennett, R.E., and Revuelta, J. (2002). *A feasibility study of on-the-fly adaptive testing* (Research Rep. No. 02–23). Princeton, NJ: Educational Testing Service.

Bookkooi-Timminga, E. (1990). A cluster-based method for test construction. *Applied Psychological Measurement*, *14*(4), 341-354.

Bormuth, J. R. (1970). *On the theory of achievement test items*. Chicago, IL: University of Chicago Press.

Burdsal, C. A., & Vaughn, D. S. (1974). A contrast of the personality structure of college students found in the questionnaire medium by items as compared to parcels. *Journal of Genetic Psychology*, 135, 219-224.

Bejar, I. I. (1996). *Generative response modeling: Leveraging the computer as a test delivery medium.* (ETS-RR-96-13). Princeton, NJ: ETS.

Bejar, I. I. (2002). Generative testing: From conception to implementation. In S. H. Irvine & P. C. Kyllonen (Eds.) *Generating items from cognitive tests: Theory* (pp. 199-217). Mahwah, NJ: Lawrence Erlbaum associates.

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika*, 64, 153-168.

Cattell, R. B. (1973). *Personality and mood by questionnaire*. San Francisco, CA: Jossey-Bass.

Cattell, R. B. (1978). *The scientific use of factor analysis in behavioral and life sciences*. New York: Plenum.

Comrey, A. L. (1967). Tandem criteria for analytic rotation in factor analysis. *Psychometrika, 32*, 143-154.

Comrey, A. L. (1984). Comparison of two methods to identify major personality factors. *Applied Psychological Measurement, 8*, 397-408.

Glas, C. A. W., & van der Linden, W. J. (2003). Computerized adaptive testing with item cloning. *Applied Psychological Measurement*, 27, 247-261.

Glas, C. A. W., & van der Linden. W. J. (2001) *Modeling variability in item parameters in item response models*. (Research Report 01-11). Eschende: University of Twente.

Hively, Wells (1974) *Domain-referenced testing*. Englewood Cliffs, NJ: Educational Technology Publications.

Hively, W., Patterson, H. L., & Page, S. H. (1968). A "universe-defined" systems of arithmetic achievement items. *Journal of Educational Measurement*, *5*, 275-290.

Humbo, C. & Dresher, A. (2001, April). *A simulation study of the impact of automatic item generation under NAEP-Like data conditions*. Paper presented at the annual meeting of the National Council on Measurement in Education, Seattle, WA.

Johnson, M. S., & Sinharay, S. (2005). Calibration of polytomous item families using Bayesian hierarchical modeling. *Applied Psychological Measurement*, 29, 369-399.

Irvine, S. H. & Kyllonen, P.C. (2002). *Item generation for test development*. Mahwah, NJ: Lawrence Erlbaum.

Luecht, R. M. (2009a). *Adaptive computer-based tasks under an assessment engineering paradigm.* Paper presented at the Graduate Management Admissions Council (GMAC) Conference on Computerized Adaptive Testing. Minneapolis, MN.

Luecht, R. M., Burke, M., & Devore, R. (2009, April). *Task modeling of complex computer-based performance exercises*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Diego, CA.

Luecht, R. M. (2006a, May). *Engineering the test: From principled item design to automated test assembly*. Paper presented at the annual meeting of the Society for Industrial and Organizational Psychology, Dallas, TX.

Luecht, R. M. (2006b, September). *Assessment engineering: An emerging discipline.* Paper presented in the Centre for Research in Applied Measurement and Evaluation, University of Alberta, Edmonton, AB, Canada.

Luecht, R. M. (2007, February). *Assessment engineering workshop*. Presented at Association of Test Publishers Conference. Palm Spring, CA.

Luecht, R. M., Gierl, M. J., Tan, X., & Huff, K. (2006, April). *Scalability and the development of useful diagnostic scales.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Masters G.N. (1982) A Rasch model for partial credit scoring. *Psychometrika, 47,* 149-174

Millman, J. & Westman, R. S. (1989). Computer-assisted writing of achievement test items: Toward a future technology. *Journal of Educational Measurement*, *26*, 177-190.

Osburn, H. G. (1958). Item sampling for achievement testing. *Educational and Psychological Measurement*, 28, 95-104.

Rosenbaum, P. R. (1988) Item bundles, *Psychometrika 53*(3), 349-359.

Sandene, B., Horkay, N., Bennett, R., Allen, N., Braswell, J., Kaplan, B., and Oranje, A. (2005). *Online Assessment in Mathematics and Writing: Reports From the NAEP Technology-Based Assessment Project, Research and Development Series* (NCES 2005–457). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

Singley, M. K., and Bennett, R.E. (2002). Item Generation and Beyond: Applications of Schema Theory to Mathematics Assessment. In S. Irvine and P. Kyllonen (Eds.), *Item generation for test development* (pp.361–384). Hillsdale, NJ: Erlbaum.

Sinharay, S., & Johnson, M. S. (2008). Use of item models in a large-scale admissions test: A case study. *International Journal of Testing*, 8: 209-236).

Sinharay, S., Johnson, M. S., and Williamson, D. M. (2003). Calibrating item families and summarizing the results using family expected response functions. *Journal of Educational and Behavioral Statistics*, 28(4), 295-313.

Wainer, H., Bradlow, E. T., and Wang, X. (2007). *Testlet response theory and its applications*. New York: Cambridge University Press.

Wainer, H. and Kiely, G. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement*, 24, 185-202.

Williamson, D. M., Johnson, M. S., Sinharay, S., and Bejar, I. I (2002, April). *Applying hierarchical model calibrations to automatically generated items*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Wilson, M., & Adams, R. J. (1995). Rasch models for item bundles. *Psychometrika*, 60(2), 181-198.