A Paradox Between IRT Invariance and Model-Data Fit When Utilizing the One-Parameter and Three-Parameter Models

Michael Custer, Sid Sharairi, Kenji Yamazaki, Diane Signatur, David Swift and Sharon Frey

Riverside Publishing Company

Paper Presented at the Annual Meeting of the
American Educational Research Association
New York City, New York
March 28, 2008

Please send correspondence regarding this paper to Michael_Custer@hmco.com

Abstract

The present study compared item and ability invariance as well as model-data fit between the one-parameter (1PL) and three-parameter (3PL) Item Response Theory (IRT) models utilizing real data across five grades; second through sixth as well as simulated data at second, fourth and sixth grade. At each grade, the 1PL and 3PL IRT models were run with each of three ability groups; low, middle and high utilizing PARSCALE Version 4.1. Results were compared in terms of item fit as well as Pearson and Spearman rank-order correlations between estimated item and ability parameters. At each grade, the 3PL exhibited the best model-data fit. However, the 1PL produced a greater degree of item and ability invariance across the three ability groups.

A Paradox Between IRT Invariance and Model-Data Fit When Utilizing the One-Parameter and Three-Parameter Models

Background

The group invariance property of item and ability parameters is a cornerstone of Item Response Theory (IRT). In IRT, item parameters are postulated to be independent of and invariant across different groups of examinees if those groups of examinees are drawn from the same examinee pool. Likewise, the ability parameter is independent of and invariant across different sets of items when those items are drawn from the same unidimensional pool of items to which an item response model has been fit. Under these conditions, variation in item parameter estimates across different examinee groups is considered to be a result of measurement error only. Likewise, variation in ability estimates across different sets of items is also considered to be due only to measurement error (Baker, 2001; Hambleton & Swaminathan, 1985).

The most widely used IRT models are the one-parameter (1PL), the two-parameter (2PL) and three-parameter (3PL) models. The 1PL utilizes a single item difficulty parameter. The 2PL incorporates an item discrimination parameter as well an item difficulty parameter and the 3PL utilizes an item difficulty, item discrimination and pseudo-guessing parameter. Lord (1980) and Kelkar (2000) suggest that model-data fit improves with the inclusion of each additional model parameter. This is especially important in light of the IRT group invariance property where item and ability parameter invariance are viewed as being dependent upon the closeness of fit between a set of test data and the item response model that is fit to it (Hambleton & Swaminathan, 1985).

A strict interpretation of item parameter invariance implies an identity relationship for which item parameters are identical across different populations when the item parameters are on the same scale. Strict item parameter invariance represents an ideal “errorless” state that can almost never be achieved in practice (Rupp & Zumbo, 2004). A practical and reasonable interpretation of the group invariance property implies that item parameter estimates are invariant to a degree (Rupp & Zumbo, 2004). One method of evaluating the degree of item parameter invariance for a set of unidimensional items is to compare the rank order of calibrated item difficulty estimates across different groups of examinees when these groups are drawn from the same examinee pool. When the rank order of item difficulty estimates is highly similar across groups of examinees, a high degree of item parameter invariance exists and the ability estimates for these groups can be made directly comparable and placed on to the same scale through a linear transformation (Weiss & Yoes, 1991). However, if the rank order in item difficulty measures is inconsistent and less similar across groups then the ability estimates across these groups will also be inconsistent.

Likewise, one would expect that in the case in which the same set of items is administered to two different examinee groups drawn from the same examinee pool, the IRT model with the "greatest" degree of fit would produce the most invariant set of item difficulty estimates across groups. One would expect that the degree to which the rank order of item difficulty estimates is similar across examinee samples is dependent upon model-data fit.

In a similar manner as above, the degree of invariance in the ability parameter can be measured by comparing the rank order of ability estimates when a group of examinees is administered two sets of items drawn from the same pool of items. The greater the similarity in the rank order of ability estimates across different sets of items, the stronger the degree of group invariance. Likewise, we would expect that the best-fitting model would produce the most invariant set of ability estimates.

This matter of consistency and comparability of item and ability estimates is especially important when vertical scales are used and, particularly, in norm-referenced testing when vertical scales are developed across several age groups or grades.

Given the importance of the group invariance property of item and ability parameters within IRT, there has been relatively little empirical investigation conducted. Fan and Ping (1999) examined item parameter invariance utilizing the 1PL, 2PL and 3PL models across different examinee samples drawn from an administration of the *Texas Assessment of Academic Skills* at the eleventh grade in Reading and Math. When utilizing low ability and high ability samples they found that the degree of item parameter invariance was greatest with the 1PL even though the data fit the 3PL best. In an unrelated paper, Kelkar et al (2000) utililized data drawn from a 1994 administration of the *Medical College Admissions Tests* and compared item and ability parameter invariance as well as item fit across the 1PL, 2PL and 3PL models. All three models showed adequate fit as well as stable item and ability parameters.

Wells et al (2002) used simulated data for the two-parameter model to investigate the effects of varying degrees of three different types of parameter drift on ability estimates. Unidirectional drift of the difficulty parameter, of the discrimination parameter, and of both parameters for five to twenty percent of test items were analyzed. They found that drift related to the difficulty parameter had a smaller effect on ability estimates as compared to the discrimination parameter, or when both exhibited drift. In all three cases, however, ability estimates were affected to only a small degree. Rupp and Zumbo (2003) extended this study further by using bias estimates to assess the effects of item difficulty drift using simulated data for the one-, two-, and three-parameter models. No single model emerged as providing the best model fit under all conditions.

Several studies have examined model-data fit utilizing the 1PL, 2PL and 3PL models under different conditions with PARSCALE (Chon, Lee & Ansley, 2007; DeMars, 2005; Kang & Chen, 2007). For item fit, PARSCALE generates a likelihood ratio chi-square statistic, G^2 (Bock, 1972; McKinley & Mills, 1985), for each item. This statistic tests for significant differences between expected and actual response patterns for each item. Significant differences imply an inadequate fit between the item and the IRT model. Chon, Lee and Ansley (2007) examined model-data fit utilizing real data consisting of both dichotomous and polytomous items. This study also examined the PARSCALE G^2 fit statistic relative to Orlando and Thissen's (2000) S-X^2 and S-G^2 fit statistics. For the dichotomous items, item fit was examined across the 1PL, 2PL and 3PL IRT models. The 3PL provided the best model-data fit and had the fewest misfitting items across all three fit indices. The 1PL had the worst fit with largest number of misfitting items. In a comparison of the fit statistics, G^2 tended to exhibit a larger number of misfitting items than the S-X^2 and S-G^2 fit indices.

It has been cited that a shortcoming of the G^2 fit statistic is its sensitivity to test length and sample size and several studies have found that G^2 exhibits inflated Type I error rates with items being incorrectly flagged as misfitting (DeMars, 2005; Kang & Chen, 2007; Stone & Hansen, 2000). DeMars (2005) used PARSCALE with simulated data (N=1000) and found inflated Type I error rates for a short test with 10 polytomous items. Kang and Chen (2007) used PARSCALE and compared the performance of G^2 with the S-X^2 fit statistic utilizing simulated data across varied test lengths of 5, 10 and 20 items with sample sizes of 500, 1000, 2000 and 5000 simulees. Inflated Type I error rates were found with G^2 for tests with 10 items or less across almost all sample sizes and for longer 20 item tests when the largest sample size of 5000 simulees was used.

This study examines the group invariance property and model-data fit utilizing the 1PL and 3PL models. Invariance in the item difficulty parameter is examined by utilizing a common set of items across examinee samples which vary in overall group ability. Invariance in the ability parameter is examined utilizing a common group of examinees across item sets which vary in difficulty. The G^2 item fit statistic in PARSCALE is used to measure model-data fit. This study expands on the work done by Fan and Ping (1999), Kelkar, Wightman and Luecht (2000) and Chon, Lee and Ansley (2007) by evaluating differences across several elementary grades (second thru sixth) , across low-, average- and high-ability samples at each grade, as well as utilizing both real and simulated data.

Method

Sample Selection. The data used for this study are from the Form S Vocabulary test of the *Gates-MacGinitie Reading Tests*®, Fourth Edition (GMRT) administered during the spring of 2006. This source data is comprised of a weighted nationally representative sample from second grade through sixth grade. At each grade, three separate ability groups (low, average and high) were derived through a three step process. First, all examinees who were administered the vocabulary test of the GMRT were split by grade and a random number was assigned to each examinee. Since only a sub-sample of the total sample was required at each grade, examinees were selected for possible inclusion in the study if their assigned random number fell within a certain range. At each grade those selected were then assigned to one of six raw score categories or subgroups. For example, the maximum raw score for most vocabulary tests was 45. Examinees with raw scores between 0 and 7 were assigned to group 1, scores between 8 and 15 were assigned to group 2, scores between 16 and 23 to group 3, scores between 24 and 30 to group 4, scores between 31 and 37 to group 5 and scores between 38 and 45 to group 6.

The second step in the process involved the assignment of examinees to either the low-, average- or high-ability group data sets at each grade. This was accomplished through a random selection of examinees by controlling the probability that students would be selected from particular raw score subgroups. As a result, the three "ability group" data sets differed at each grade according to the proportion of students falling within each of the six raw score subgroups. For example, when compared to the low- and average-ability group data sets the higher ability group data included a smaller proportion of students drawn from the low raw score subgroups and a higher proportion drawn from the high raw score subgroups. In this manner, all six raw score subgroups were used in creating each of the three ability group data sets ensuring a full range of scores for each data set.

The final step in the derivation of the three data sets at each grade was to systematically remove duplicate cases thus ensuring that an examinee did not belong to more than one ability group or data set.

Factor Analysis. Unidimensionality of the item pool is an important assumption behind IRT models and was examined with an exploratory factor analysis using tetrachoric correlations. At each grade, the three data sets were combined for the purpose of the factor analysis. The factor analysis utilized an unweighted least squares extraction method and was performed at each grade. The initial eigenvalues and the percent of variance explained was then reviewed to ascertain the number of potential factors.

Item Calibrations and Correlational Analyses to Evaluate Invariance In the Item Difficulty and Ability Parameters. Once the unidimensionality of the item pool had been established, the one- and three-parameter models were each executed utilizing PARSCALE Version 4.1 for each of the three ability groups at each of the five grades. The 1PL and 3PL model were each executed utilizing several of the same program control settings. In both cases, items were calibrated with the partial credit response model utilizing the logistic response function. The number of quadrature points was set to 30. The convergence criterion for EM cycles was set to .0005. Maximum likelihood estimation (MLE) was used for scoring and ability estimates were scaled to have a mean of 0.00 and a standard deviation of 1.00. The convergence setting for MLE estimation was set to .0005. Tighter than default convergence settings were used for both the calibration and scoring phases mentioned above. Custer, Omar and Pomplun (2006) compared WINSTEPS and BILOG-MG utilizing simulated data across eleven grades and found that both programs recaptured simulated parameter estimates more accurately when tighter than default convergence settings were used.

Besides the control settings mentioned above, the 1PL was executed utilizing the Rasch Model variant in which the item discrimination (slope) parameter was fixed (not estimated) and uniformly set to 1.0 for each item. The guessing parameter was not estimated.

Executions of the 3PL utilized the SPRIOR and GPRIOR keywords for the estimation of the item discrimination (slope) and guessing parameters. Use of the SPRIOR option assumes a log-normal prior distribution on the slope parameter and assists in preventing Heywood cases. Utilization of the GPRIOR keyword assumes a normal prior distribution on the guessing parameter which is useful for the estimation of plausible values for easy items which carry little or no information about guessing (Muraki & Bock, 2003).

Implicit to the study design, all of the examinees were administered a vocabulary test that was appropriate for their specific grade. At each grade a total of six calibrations were run. The 1PL was run once with each of the three ability group data sets. Likewise, the 3PL was also run with each of the three ability group data sets. As noted earlier, the mean for the ability estimates in each calibration was set to 0.00 with a standard deviation of 1.00. As a result ability group differences were reflected in the item difficulty estimates. Though all of the items were common across ability groups within each grade, no attempts were made to link the scales because of the unitless nature of both the Pearson and the Spearman rank order correlations which were used to examine both the degree of association and the rank order of these “common” item difficulty measures across ability groups.

At each grade and for each IRT model across the three ability group data sets, Pearson and Spearman rank order correlational analyses were run between associated item difficulty estimates. Specifically, 1PL item difficulty estimates for the average ability group were correlated with both the 1PL estimates from the run using the low ability group as well as those of the high ability group. The 3PL item difficulty measures were correlated across ability groups in the same manner. This process allowed the authors to examine IRT invariance in the item difficulty parameter by measuring both the degree of association between a set of item difficulty measures as well as how consistently the rank order of these measures was maintained across ability groups when run with the 1PL model versus the 3PL.

The degree of invariance in the ability parameter was measured utilizing a common person design across two sets of items, one set composed of easier items and another set of more difficult items. At each grade, the items were first sorted by item difficulty (p-values) and then assigned to the easy or hard test according to their difficulty. For example, at third grade there were 45 items in the Vocabulary test. The 27 least difficult items were assigned to the easy test and the 27 most difficult items were assigned to the hard test. There were nine common items across the two tests leaving 18 items that were unique to that test alone. The common item block was composed of the 9 most difficult items in the easy test and the 9 easiest items in the hard test. At each grade, these two tests were then calibrated with the average ability group utilizing both the 1PL and 3PL IRT models. Hence at each grade there were four scalings: 1PL with the easier test, 1PL with the harder test, 3PL with the easier test and 3PL with the harder test.

In each calibration, the ability estimates were originally set to have a mean of 0.00 and a standard deviation of 1.00. These estimates were then recentered by applying a constant that was equal to the mean difference in item difficulty for the common item sets. Recentering the abilities in this manner allowed the authors to adjust these common person ability estimates for the difference in test difficulty. At each grade and for each IRT model, the ability estimates derived from the scaling with the easy test were correlated with the ability estimates derived in the scaling with the hard test. Pearson and Spearman rank order correlations were used for these analyses. Also, due to the weakening effect of measurement error and the reduction in reliability that occurred with the shortening of the original 45 item test into two 27 item tests, a correction for attenuation was also applied to and reported for the Pearson correlations.

Lastly, simulated data was utilized to study item parameter invariance. Within each grade, the low-, average- and high-ability group "real" data sets were combined. The 3PL was then run utilizing the combined data set. In this manner, target item parameter estimates were generated for each grade. Data was then simulated for different ability groups utilizing the "target" 3PL item parameter estimates. Additional detail as well as the results can be found in Appendix B.

Results

Descriptive statistics for the three ability groups are provided in Table 1. As expected, within each grade the mean raw scores increase and the standard deviations generally decrease across the three ability groups. The effect size differences in mean scores between the lower and average ability groups range from .39 at grades 2, 5 and 6 to .46 at grade 3. Effect size differences between the average and higher ability groups range from .24 at grade 4 to .30 at grade 3. Differences between the lower and higher ability groups range from .66 at grade 5 to .78 at grade 3. Reliabilities for the three ability groups at each grade were above .90.

Table 1. Performance Characteristics of Lower, Average and Higher Ability Groups

Grade	Ability Group	N	# of Items	Mean	SD	Min	Max	Cronbach's Alpha
	Lower	641	43	26.55	9.11	4	43	.913
2	Average	990	43	30.08	8.91	6	43	.919
	Higher	866	43	32.41	8.02	4	43	.910
	Lower	496	45	26.83	9.67	3	44	.919
3	Average	910	45	31.19	9.24	4	45	.922
	Higher	1017	45	33.78	8.17	4	45	.910
	Lower	514	45	25.56	9.70	3	45	.920
4	Average	858	45	29.54	9.40	3	45	.923
	Higher	1324	45	31.69	8.52	5	45	.911
	Lower	640	45	24.43	9.37	3	45	.907
5	Average	939	45	28.12	9.54	4	45	.917
	Higher	1046	45	30.42	8.69	4	45	.904
	Lower	666	45	23.74	9.58	4	45	.913
6	Average	1016	45	27.49	9.43	3	45	.914
	Higher	888	45	29.79	9.03	6	45	.911

Tables 2 and 3 present the minimum, mean and maximum item p-value and corrected point-biserial summary statistics for each ability group by grade. As expected the means for the item p-values increase across ability groups. In addition, the range in item p-values at each grade and ability group seems reasonable. This can also be said for the mean and range of the corrected point-biserals. These item level results provide support for the stability of the items used in this study.

Table 2. Item P-Value Descriptive Statistics

	Lower Ability Group			Average Ability Group			Higher Ability Group		
Grade	Min	Mean	Max	Min	Mean	Max	Min	Mean	Max
2	0.35	0.62	0.90	0.44	0.70	0.94	0.47	0.75	0.96
3	0.27	0.60	0.95	0.33	0.69	0.98	0.38	0.75	0.99
4	0.18	0.57	0.94	0.25	0.66	0.96	0.32	0.70	0.98
5	0.20	0.54	0.93	0.24	0.62	0.96	0.27	0.68	0.97
6	0.23	0.53	0.90	0.32	0.61	0.93	0.35	0.66	0.97

Table 3. Item Corrected Point-Biserial Descriptive Statistics

	Lower Ability Group			Average Ability Group			Higher Ability Group		
Grade	Min	Mean	Max	Min	Mean	Max	Min	Mean	Max
2	0.26	0.43	0.57	0.31	0.44	0.56	0.29	0.42	0.59
3	0.20	0.43	0.60	0.24	0.44	0.61	0.22	0.42	0.56
4	0.28	0.43	0.56	0.26	0.45	0.60	0.21	0.42	0.58
5	0.20	0.40	0.54	0.25	0.43	0.56	0.25	0.40	0.55
6	0.23	0.42	0.56	0.27	0.42	0.54	0.22	0.42	0.57

The initial set of eigenvalues for a factor analysis utilizing an unweighted least squares (ULS) extraction is presented in Table 4. At each grade, evidence of unidimensionality and the existence of one dominant factor is provided by the percent of variance explained by the first factor as well as a the noticeable drop in eigenvalues

Table 4. Initial Eigenvalues for a Factor Analysis Utilizing ULS Extraction

	Grade				
Factor	2	3	4	5	6
1	17.02/82.0%*	18.45/74.4%*	17.81/77.5%*	15.63/78.4%*	16.25/77.6%*
2	1.40	1.76	1.90	1.54	1.45
3	0.70	0.68	0.79	0.71	0.66
4	0.57	0.58	0.53	0.52	0.60

Note: *The percent of variance explained is reported for the first factor

between the first and second factor. At all grades, a substantial percent of variance is explained by the first factor and ranges from a low of 74.4% at grade 3 to a high of 82.0% at grade 2. This is coupled with a large decline in the initial eigenvalues between the first and second factors. Based on these results, the assumption of unidimensionality appears to hold for the data used in this study.

A summary of the item-fit statistics for the one-parameter and three-parameter models by ability group and grade are presented in Table 5. The number of misfitting items across grade and ability group ranged from 13 to 28 for the 1PL and between 0 and 10 for the 3PL. As noted, several studies (Chon, Lee & Ansley, 2007; DeMars, 2005; Kang and Chen, 2007; Stone & Hansen, 2000) have linked the G^2 fit statistic used in PARSCALE with the over-identification of misfitting items.

Table 5. Number of Items Identified* as Misfitting with the 1PL and 3PL IRT Models by Grade and Ability Group

Grade	Ability Group	# of Items	# of Misfitting items 1PL	3PL
	Lower	43	14	0
2	Average	43	13	1
	Higher	43	14	0
	Lower	45	17	6
3	Average	45	18	2
	Higher	45	15	3
	Lower	45	16	2
4	Average	45	18	10
	Higher	45	25	4
	Lower	45	20	5
5	Average	45	22	8
	Higher	45	28	10
	Lower	45	18	0
6	Average	45	22	5
	Higher	45	17	9

***Note**: Items are considered misfitting at alpha=.01 level.

Given the large sample sizes used in this study, a large number of misfitting items is not wholly unexpected. However, even with this taken into consideration, the results indicate that across all grades and ability groups the 3PL demonstrated the strongest model-data fit.

The results of correlational analyses between item difficulty estimates across examinee samples and by grade are presented in Tables 6 thru 8. Each table presents Pearson and Spearman rank order (Rho) correlations for both the 1PL and the 3PL. Table 6 presents correlations between item difficulty estimates when each model was run with the lower and average ability group. In all grades the item measures correlated most highly with the one-parameter model. Across grades the average of the Pearson correlations was .991 for the 1PL and .963 for the 3PL. The average for the Spearman rank order correlations was .985 for the 1PL and .967 for the 3PL.

Table 6. Correlations of IRT Item Difficulty Parameter Estimates Between Lower and Average Ability Groups

	1PL	1PL	3PL	3PL
Grade	Pearson	Spearman's Rho	Pearson	Spearman's Rho
2	.993	.991	.985	.988
3	.989	.979	.962	.971
4	.993	.987	.960	.967
5	.989	.981	.935	.932
6	.990	.987	.974	.975
Mean	.991	.985	.963	.967

Table 7 presents item measure correlations between the average and higher ability groups. In each grade the item measures correlated most highly with the 1PL. Across grades the average of the Pearson correlations was .993 for the 1PL and .953 for the 3PL. The average for the Spearman rank order correlations was .990 for the 1PL and .963 for the 3PL

Table 7. Correlations of IRT Item Difficulty Parameter Estimates between Average and Higher Ability Groups

	1PL	1PL	3PL	3PL
Grade	Pearson	Spearman's Rho	Pearson	Spearman's Rho
2	.992	.990	.954	.961
3	.995	.987	.958	.965
4	.994	.991	.953	.963
5	.994	.991	.932	.953
6	.989	.989	.968	.972
Mean	.993	.990	.953	.963

Item measure correlations between the lower and higher ability groups are presented in Table 8. As with the other comparisons, item measure correlations were strongest with the 1PL. Across grades the average of the Pearson correlations was .989 for the 1PL and .954 for the 3PL. Likewise, the average of the Spearman rank order correlations was .982 for the 1PL and .958 for the 3PL. Similar results as those reported in Table 8 can be found in Appendix B for the comparison between lower and higher ability groups when utilizing simulated data.

Table 8. Correlations of IRT Item Difficulty Parameter Estimates between Lower and Higher Ability Groups

	1PL	1PL	3PL	3PL
Grade	Pearson	Spearman's Rho	Pearson	Spearman's Rho
2	.989	.985	.975	.977
3	.987	.967	.952	.949
4	.992	.988	.978	.972
5	.988	.980	.910	.936
6	.989	.990	.955	.958
Mean	.989	.982	.954	.958

A pictorial representation of the changes in the rank order of the item difficulty estimates by IRT model and across examinee samples can be seen in Tables 9 and 10. For illustrative purposes, these tables present the results for grade 5. Results for the other grades can be found in Appendix A. Specifically, for each 1PL and 3PL execution and for each examinee sample the item difficulty estimates were sorted in ascending order. For both the 1PL and 3PL IRT models, the change in rank order was observed by tracing an item's rank difficulty position from the average ability group run to that same item's rank order position with the high ability group. Table 9 presents this information for Grade 5 with the 1PL and Table 10 presents this information for Grade 5 with the 3PL.

Review of Tables 9 and 10 reveals a surprising degree of switching in the rank order of item difficulty estimates between the average and high ability groups at grade 5. The tables themselves are a pictorial representation of the reported Spearman rank order correlations of .991 for the 1PL and .953 for the 3PL. A greater degree of rank order switching with the 3PL is evident in Table 10.

Table 9 shows that with the 1PL thirteen items maintained their same rank order across the two runs, 16 items shifted only 1 position and the remaining 16 shifted two or more positions. There were no items that shifted more than 5 positions. As revealed in Table 10 with the 3PL, only 7 items maintained their identical position across the two runs, 12 shifted only 1 position and 26 shifted two or more positions. There were 5 items that shifted more than 5 positions.

Table 9. Switching in the Rank Order of Item Difficulty Estimates Across Average and High Ability Groups Utilizing the 1PL IRT Model: Grade 5

Average Ability Group		Higher Ability Group	
Item Name	Item Diff	Item Name	Item Diff
MC01	-2.24672	MC01	-2.47836
MC02	-1.82047	MC02	-1.94103
MC03	-1.31117	MC03	-1.51855
MC04	-1.26415	MC04	-1.44207
MC05	-1.20788	MC05	-1.37670
MC17	-1.10753	MC06	-1.25169
MC06	-1.00006	MC09	-1.23637
MC09	-1.00006	MC17	-1.21625
MC10	-0.91325	MC21	-1.17220
MC12	-0.89997	MC12	-1.16263
MC21	-0.86079	MC10	-1.14370
MC11	-0.84366	MC11	-0.99317
MC13	-0.78502	MC13	-0.91966
MC14	-0.73622	MC08	-0.91568
MC08	-0.63057	MC23	-0.88034
MC20	-0.61914	MC14	-0.86873
MC19	-0.60777	MC20	-0.81937
MC23	-0.59646	MC19	-0.76416
MC27	-0.45730	MC25	-0.67915
MC26	-0.45016	MC18	-0.67569
MC18	-0.42884	MC15	-0.63798
MC07	-0.41824	MC27	-0.63459
MC25	-0.41471	MC26	-0.57106
MC15	-0.36916	MC07	-0.52209
MC44	-0.35873	MC33	-0.47720
MC31	-0.27961	MC31	-0.47403
MC33	-0.27961	MC44	-0.46451
MC28	-0.26258	MC24	-0.46135
MC34	-0.20172	MC34	-0.43302
MC24	-0.19163	MC22	-0.31295
MC36	-0.09807	MC36	-0.31295
MC22	-0.07812	MC30	-0.26472
MC37	-0.06483	MC28	-0.25873
MC42	-0.06151	MC37	-0.24377
MC30	-0.05819	MC42	-0.19024
MC16	-0.01508	MC38	-0.15184
MC38	0.05455	MC16	-0.10480
MC40	0.06450	MC35	-0.07843
MC35	0.10437	MC40	-0.04624
MC29	0.22146	MC32	-0.00825
MC32	0.22821	MC29	0.10582
MC41	0.34416	MC43	0.21793
MC43	0.45633	MC41	0.25671
MC39	0.63669	MC39	0.47171
MC45	0.88432	MC45	0.77193

Table 10. Switching in the Rank Order of Item Difficulty Estimates Across Average and High Ability Groups Utilizing the 3PL IRT Model: Grade 5

Average Ability Group		Higher Ability Group	
Item Name	Item Diff	Item Name	Item Diff
MC02	-2.47690	MC01	-3.56059
MC01	-2.44913	MC03	-2.94953
MC03	-1.99764	MC02	-2.91391
MC11	-1.89269	MC18	-2.79449
MC04	-1.76256	MC05	-2.20601
MC05	-1.70880	MC09	-2.15165
MC08	-1.29791	MC04	-1.98612
MC09	-1.24366	MC17	-1.89547
MC13	-1.24318	MC11	-1.67686
MC17	-1.22108	MC06	-1.57838
MC06	-0.99162	MC21	-1.49065
MC21	-0.97267	MC12	-1.26031
MC10	-0.90464	MC13	-1.22512
MC14	-0.83740	MC08	-1.21212
MC12	-0.78467	MC10	-1.20298
MC16	-0.71027	MC14	-1.16625
MC19	-0.59248	MC33	-1.08143
MC20	-0.59048	MC23	-1.03113
MC23	-0.58024	MC20	-0.99514
MC18	-0.53447	MC34	-0.93501
MC34	-0.48717	MC19	-0.82687
MC07	-0.46522	MC27	-0.74758
MC26	-0.37502	MC25	-0.65794
MC27	-0.30855	MC16	-0.56948
MC44	-0.30461	MC15	-0.50204
MC25	-0.29021	MC31	-0.47467
MC33	-0.21624	MC07	-0.46771
MC31	-0.18784	MC44	-0.46576
MC15	-0.11069	MC26	-0.44924
MC36	-0.04547	MC36	-0.39594
MC22	0.05474	MC22	-0.32921
MC24	0.17680	MC24	-0.12355
MC30	0.19493	MC38	-0.08653
MC42	0.21569	MC42	-0.03257
MC38	0.21604	MC28	0.06238
MC40	0.30455	MC40	0.11257
MC28	0.31001	MC32	0.13789
MC37	0.46595	MC30	0.20425
MC32	0.47269	MC35	0.28533
MC35	0.53418	MC43	0.42556
MC43	0.69299	MC37	0.45286
MC29	0.79140	MC29	0.56085
MC41	0.79425	MC39	0.62111
MC39	1.06617	MC41	0.67962
MC45	1.32782	MC45	1.12801

Table 11 presents classical item statistics for each of the items that shifted more than five positions across the two runs with the 3PL. Specifically, the p-value (p) and corrected point-biserial (pb) by ability group as well as the ETS classification for the Mantel-Haenszel differential item functioning (DIF) statistic are displayed for each item. It also presents the mean p-values and corrected point-biserials for the remaining 40 items.

Table 11. Classical Item Statistics for Items with the Greatest Degree of Rank Order Switching with the Three Parameter Model at Grade 5.

Average Ability Group			High Ability Group		Difference	MH-Dif Average Ability Reference Group
Item #	P-Value	Pt.-Biserial	P-Value	Pt.-Biserial	p / pb	ETS Classification
8	.70	.37	.77	.36	.07 / -.01	A
16	.51	.26	.54	.25	.03 / -.01	A
18	.64	.51	.71	.40	.07 / -.11	A
26	.64	.39	.68	.41	.04 / .02	A
33	.59	.52	.65	.43	.06 / -.09	A
Mean 40	.62	.43	.68	.40	.06 / -.03	NA

It is interesting to note that four of the five items for which statistics are presented above demonstrate some degree of functioning that is outside of expectations as defined by the remaining 40 items for which information is presented in the bottom row. Across ability groups the 40 item set had an increase in the mean p-value from .62 to .68 and a drop in the mean pont-biserial from .43 to .40. Relative to this 40 item group, items 18 and 33 had much larger differences in their corrected point-biserials (a drop of .11 and .09 respectively). Likewise, the p-value for item 16 increased by only .03 across groups while the p-value for item 26 increased by .04 with a positive change in point-biserial from the average to the high ability group (the only positive point-biserial change for the five items). Each of these items were accompanied by an ETS classification of “A” indicating a negligible level of differential item functioning across ability groups. It is also interesting to note that the p-values for these five items indicate that they fell near the middle of the distribution in terms of item difficulty. A review of the “sort in descending order” of the item p-values on the original 45 items revealed that four of the five items fell between the 15th position and the 26th position while item 16 fell at the 36th position. Appendix A includes the results for second, third, fourth and sixth grade.

In order to evaluate invariance in the ability parameter, the original 43 items at second grade and the original 45 items at third through sixth grade were split according to item difficulty into an easy and hard test at each grade. Descriptive statistics for these two tests are provided for the average ability group in Table 12. As expected, at each grade the mean raw score for the hard test is lower than that of the easy test. Also, the standard deviation for the hard test is generally higher. For all grades the test reliabilities for the shortened tests are reasonable and range from .865 to .898.

Table 12. Average Ability Group Performance Characteristics by Grade For The Easier And Harder Item Sets

Grade	Item Test Type	N	# of Items	Mean	SD	Cronbach's Alpha
2	Easier Item Set	990	26	20.79	5.10	.882
	Harder Item Set	990	26	15.53	6.41	.887
3	Easier Item Set	910	27	21.71	5.44	.895
	Harder Item Set	910	27	15.97	6.42	.883
4	Easier Item Set	858	27	21.33	5.62	.898
	Harder Item Set	858	27	14.50	6.44	.883
5	Easier Item Set	939	27	19.68	6.02	.893
	Harder Item Set	939	27	14.05	6.30	.869
6	Easier Item Set	1016	27	19.56	5.65	.874
	Harder Item Set	1016	27	13.29	6.24	.865

The results of the correlational analyses between the ability parameter estimates across the easy and hard test by grade and IRT model are presented in Table 13. This table reports Pearson, Pearson corrected for attenuation and Spearman rank order correlations for both the 1PL and the 3PL by grade.

Table 13. Correlations of IRT Ability Parameter Estimates between Easier and Harder Item Sets Utilizing the Average Ability Group By Grade

	1 Parameter Model			3 Parameter Model		
Grade	Pearson	Corrected For Attenuation	Spearman's Rho	Pearson	Corrected For Attenuation	Spearman's Rho
2	.826	.934	.852	.791	.894	.845
3	.828	.931	.850	.738	.830	.821
4	.831	.933	.864	.742	.833	.798
5	.855	.971	.886	.817	.927	.889
6	.861	.990	.888	.838	.964	.865
Mean	.840	.952	.868	.785	.890	.844

In all grades the ability estimates between the two tests correlated most highly with the one parameter model. Across grades the average of the Pearson correlations was .840 for the 1PL and .785 for the 3PL. When corrected for attenuation the mean correlations were .952 for the 1PL and .890 for the 3PL with the Spearman rank order correlations for these .868 and .844, respectively.

Discussion

Overall the results of this study concur with Fan and Ping's (1999) earlier findings and point to a paradox between group invariance in the item and ability parameters and model-data fit. Though model-data fit is best achieved with the 3PL, item and ability parameter invariance seems to be strongest with the relatively worse fitting 1PL. A possible explanation as to why might be found in Tables 10 and 11. These results seem to indicate that when items behave outside of expectations relative to the item set as a whole and exhibit a reasonable or even negligibly small degree of differential item functioning across ability groups, large shifts in the rank order of the item difficulty estimates may occur across these groups with the 3PL. This effect may be more pronounced if the items in question are near the middle of the distribution in terms of item difficulty. This instability in the item difficulty estimates manifests in the ability estimates. An unintended consequence of the higher level of sensitivity or "fit" between the 3PL and the data that it is meant to model may be a less invariant set of parameter estimates. The 3PL with its greater sensitivity to subtle changes in item behavior is more likely to capture this behavior and as a result may produce a less invariant set of parameter estimates.

Limitations

Some of the nuances and subtleties associated with real data are difficult to replicate within a data simulation. There are several ways that "real" data can be used to extract different ability groups from a sample of examinees. This study uses one method that may have inadvertently influenced the results. Further research should extend this investigation by using alternative methods for deriving different ability groups from the same pool of examinees. Likewise, results utilizing simulated data were presented in Appendix B of this study. This simulated data was in part derived with "known" IRT parameter estimates. Data simulated in this manner may have inadvertently impacted the results. Future research should investigate non-IRT methods for simulating data. In addition, of the several IRT programs available, PARSCALE was used for this study. Future research should utilize other IRT programs in addition to PARSCALE. Lastly, this paper attempts to identify some of the reasons that may explain the paradox between model-data fit and parameter invariance. The authors recognize that this list is not exhaustive and additional research is needed.

References

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

Baker, F. (2001). The Basics of Item Response Theory 2nd Edition. ERIC Clearinghouse on Assessment and Evaluation.

Chon, K., Lee, W., & Ansley, T. (2007). Assessing IRT Model-Data Fit for Mixed Format Tests. Center for Advanced Studies in Measurement and Assessment (CASMA) Research Report Number 26.

Custer, M., Omar, M.H., & Pomplun, M. (2006) Vertical Scaling With the Rasch Model Utilizing Default and Tight Convergence Settings With WINSTEPS and BILOG-MG. Applied Measurement In Education , 19(2), 131-147

DeMars, C.E. (2005). Type I error rates for PARSCALE's fit index. Educational and Psychological Measurement, 65, 42-50

Fan, X., & Ping, Y. (1999). Assessing the Effect of Model-Data Misfit on the Invariance Property of IRT Parameter Estimates. Paper presented at the annual meeting of the American Educational Association, Montreal, Canada, April 19-23,1999.

Hambleton, R.K., & Swaminathan, H. (1985). *Item Response Theory: Principles and Applications.* Boston, MA: Kluwer Academic Publishers

Han, K. T. (2007). WinGen2: Windows software that generates IRT parameters and item responses [computer program]. Amherst, MA: University of Massachusetts, Center for Educational Assessment. Retrieved May 13, 2007, from http://www.umass.edu/remp/software/wingen/

Han, K. T., & Hambleton, R. K. (2007). User's Manual: WinGen (*Center for Educational Assessment Report No. 642).* Amherst, MA: University of Massachusetts, School of Education.

Kang, T., & Chen, T. (2007). An Investigation of the Performance of the Generalized S-X^2 Item-Fit Index for Polytomous IRT Models. ACT Research Report Series 2007-1.

Kelkar, V., Wightman, L., & Leucht, R. (2000). Evaluation of the IRT Parameter Invariance Property for the MCAT. Paper presented at the annual meeting of the National Councail on Measurement in Education, New Orleans, LA, April 25 - 27, 2000.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum.

MacGinitie, W., MacGinitie, R., Maria, K., & Dreyer, L., (2000). Gates-MacGinitie Reading Tests Fourth Edition. Rolling Meadows, IL: Riverside Publishing.

McKinley, R., & Mills, C.N. (1985). A comparison of several goodness-of-fit statistics. Applied Psychological Measurement, 9, 49-57.

Muraki, E., & Boch, R.D. (2003). Parscale Version 4.1 IRT Item Analysis and Test Scoring for Rating-Scale Data. Chicago: Scientific Software International.

Orlando, M., & Thissen, D. (2000). New item fit indices for dichotomous item response theory models. Applied Psychological Measurement, 24, 50-64.

Rupp, A.A. & Zumbo, B.D. (2003). Which model is best? Robustness properties to justify model choice among unidimensional IRT models under item parameter drift. *Alberta Journal of Educational Research*, 49, 264-276.

Rupp, A.A. & Zumbo, B.D. (2004). A note on how to quantify and report whether IRT parameter invariance holds: when pearson correlations are not enough. *Educational and Psychological Measurement*, 65(4), 588-599.

Stone, C.A., & Hansen, M.A. (2000). The effect of errors in estimating ability on goodness-of-fit tests for IRT models. Journal of Educational Measurement, 37, 58-75

Weiss, D.J., & Yoes, M.E. (1991). Item Response Theory in R.K. Hambleton and J.N. Zeal (eds.), *Advances in educational and psychological testing*, pp. 69-96. Boston, MA: Kluwer Academic Publishers

Wells, C., Subkoviak, M., & Serlin, R. (2002). The effect of item parameter drift on examinee ability estimates. *Applied Psychological Measurement*, 26, 77 - 87.

Appendix A

Table A1. Switching in the Rank Order of Item Difficulty Estimates Across Average and High Ability Groups Utilizing the 1PL IRT Model: Grade 2

Average Ability Group		Higher Ability Group	
Item Name	Item Diff	Item Name	Item Diff
MC02	-2.08039	MC02	-2.40979
MC01	-2.04513	MC08	-2.29099
MC08	-1.92727	MC01	-2.25519
MC05	-1.68333	MC04	-2.04179
MC04	-1.65312	MC05	-1.97729
MC16	-1.63104	MC11	-1.96495
MC03	-1.62379	MC16	-1.96495
MC11	-1.60944	MC07	-1.82908
MC07	-1.52072	MC03	-1.81864
MC15	-1.39109	MC06	-1.67449
MC06	-1.37371	MC15	-1.64807
MC17	-1.14778	MC13	-1.40833
MC10	-1.09085	MC18	-1.32131
MC13	-1.01365	MC10	-1.29575
MC18	-1.00923	MC17	-1.26453
MC21	-1.00923	MC19	-1.15793
MC19	-0.92723	MC21	-1.12418
MC23	-0.89381	MC28	-1.06429
MC28	-0.87322	MC23	-1.00670
MC14	-0.67031	MC14	-0.94616
MC22	-0.67031	MC22	-0.91670
MC12	-0.64429	MC20	-0.90699
MC20	-0.61851	MC12	-0.86868
MC24	-0.59660	MC09	-0.81727
MC09	-0.58208	MC25	-0.78979
MC25	-0.52114	MC24	-0.76269
MC29	-0.50696	MC43	-0.65330
MC36	-0.47877	MC29	-0.61094
MC30	-0.40917	MC26	-0.58585
MC34	-0.37136	MC36	-0.55688
MC26	-0.33723	MC34	-0.52413
MC43	-0.33383	MC30	-0.48770
MC42	-0.23259	MC42	-0.48368
MC32	-0.19912	MC35	-0.41998
MC31	-0.17909	MC32	-0.38072
MC27	-0.16909	MC27	-0.35346
MC38	-0.11584	MC38	-0.30712
MC35	-0.10587	MC41	-0.27264
MC41	-0.01288	MC33	-0.26119
MC33	0.11045	MC31	-0.22697
MC40	0.12386	MC40	-0.12149
MC39	0.19123	MC39	-0.00941
MC37	0.19800	MC37	0.10261

Table A2. Switching in the Rank Order of Item Difficulty Estimates Across Average and High Ability Groups Utilizing the 3PL IRT Model: Grade 2

Average Ability Group		Higher Ability Group	
Item Name	Item Diff	Item Name	Item Diff
MC02	-2.18520	MC02	-2.57849
MC01	-2.12885	MC08	-2.54400
MC21	-2.11567	MC04	-2.43555
MC04	-1.94591	MC11	-2.24768
MC08	-1.87437	MC01	-2.24267
MC07	-1.79394	MC07	-2.24065
MC03	-1.74321	MC05	-2.10109
MC11	-1.65702	MC06	-2.03102
MC05	-1.54045	MC16	-1.79227
MC16	-1.38324	MC15	-1.67543
MC06	-1.33143	MC03	-1.64822
MC15	-1.25869	MC10	-1.56987
MC18	-0.90132	MC13	-1.43313
MC13	-0.84985	MC28	-1.35987
MC10	-0.81840	MC18	-1.30691
MC17	-0.72820	MC21	-1.30123
MC19	-0.62354	MC17	-1.18876
MC22	-0.57882	MC19	-0.94049
MC23	-0.55398	MC22	-0.90798
MC14	-0.53115	MC14	-0.85149
MC28	-0.49378	MC09	-0.75157
MC12	-0.47775	MC23	-0.69873
MC20	-0.36336	MC20	-0.67641
MC30	-0.25729	MC25	-0.61568
MC24	-0.25209	MC12	-0.61275
MC36	-0.22432	MC24	-0.58772
MC29	-0.19867	MC29	-0.50099
MC25	-0.19066	MC43	-0.46633
MC09	-0.18002	MC36	-0.46563
MC26	-0.08597	MC30	-0.36295
MC34	-0.05474	MC26	-0.35670
MC43	0.02043	MC34	-0.35230
MC42	0.07184	MC42	-0.28366
MC31	0.07288	MC35	-0.26205
MC27	0.11072	MC32	-0.24481
MC32	0.11623	MC38	-0.15906
MC38	0.14638	MC31	-0.11704
MC35	0.21506	MC41	0.01131
MC41	0.38531	MC27	0.01256
MC33	0.38875	MC33	0.02930
MC40	0.50740	MC40	0.21388
MC39	0.72039	MC39	0.54853
MC37	0.90239	MC37	0.69851

Table A3. Switching in the Rank Order of Item Difficulty Estimates Across Average and High Ability Groups Utilizing the 1PL IRT Model: Grade 3

Average Ability Group		Higher Ability Group	
Item Name	Item Diff	Item Name	Item Diff
MC01	-2.76928	MC01	-3.12395
MC02	-2.15175	MC02	-2.63051
MC03	-1.93446	MC03	-2.35972
MC04	-1.73283	MC09	-2.20217
MC09	-1.73283	MC04	-2.12076
MC06	-1.52849	MC05	-1.79253
MC05	-1.51375	MC06	-1.77499
MC14	-1.43607	MC14	-1.69986
MC12	-1.36336	MC10	-1.61530
MC10	-1.30091	MC12	-1.60066
MC08	-1.18463	MC08	-1.55799
MC20	-1.11943	MC20	-1.40869
MC17	-1.11413	MC17	-1.34957
MC07	-0.99237	MC07	-1.33810
MC13	-0.93960	MC25	-1.18291
MC15	-0.90225	MC34	-1.16292
MC16	-0.86116	MC13	-1.14812
MC25	-0.84768	MC23	-1.14323
MC18	-0.81218	MC18	-1.13835
MC34	-0.79033	MC26	-1.07182
MC26	-0.78599	MC15	-1.06718
MC23	-0.78166	MC11	-1.05796
MC11	-0.76872	MC21	-1.05337
MC21	-0.76872	MC16	-1.02164
MC24	-0.74733	MC28	-1.01717
MC19	-0.68446	MC24	-0.94726
MC28	-0.67621	MC19	-0.94300
MC27	-0.55969	MC27	-0.83977
MC22	-0.41405	MC22	-0.69340
MC29	-0.38418	MC32	-0.63118
MC33	-0.36193	MC35	-0.61321
MC30	-0.29950	MC29	-0.60605
MC35	-0.29220	MC33	-0.49772
MC32	-0.24505	MC30	-0.38345
MC41	-0.17319	MC31	-0.38014
MC38	-0.11614	MC41	-0.33747
MC31	-0.08416	MC38	-0.28556
MC42	0.00441	MC40	-0.13911
MC43	0.02565	MC42	-0.12024
MC36	0.08944	MC43	-0.11081
MC44	0.12142	MC44	-0.05441
MC40	0.12854	MC36	0.04237
MC37	0.21434	MC37	0.05485
MC39	0.28664	MC39	0.08295
MC45	0.58912	MC45	0.39106

Table A4. Switching in the Rank Order of Item Difficulty Estimates Across Average and High Ability Groups Utilizing the 3PL IRT Model: Grade 3

Average Ability Group		Higher Ability Group	
Item Name	Item Diff	Item Name	Item Diff
MC01	-2.78146	MC01	-3.51655
MC04	-2.22626	MC06	-3.27199
MC09	-2.08735	MC02	-2.73697
MC02	-2.02318	MC09	-2.48421
MC03	-1.87499	MC04	-2.38171
MC06	-1.68757	MC03	-2.28692
MC12	-1.59790	MC05	-2.14487
MC14	-1.55692	MC14	-1.76868
MC05	-1.44440	MC10	-1.58290
MC15	-1.32923	MC08	-1.53651
MC17	-1.32795	MC13	-1.52108
MC10	-1.17420	MC17	-1.45784
MC20	-1.12247	MC12	-1.44855
MC16	-1.06759	MC20	-1.43669
MC08	-0.99519	MC15	-1.38685
MC07	-0.93913	MC16	-1.35101
MC13	-0.85968	MC27	-1.33854
MC18	-0.82281	MC18	-1.29572
MC25	-0.68260	MC19	-1.19902
MC23	-0.66913	MC34	-1.18553
MC19	-0.66874	MC07	-1.17536
MC21	-0.66211	MC25	-1.17429
MC26	-0.65112	MC24	-1.07241
MC34	-0.61520	MC23	-1.06940
MC24	-0.61313	MC21	-1.01861
MC11	-0.56255	MC26	-0.97891
MC27	-0.53572	MC28	-0.86386
MC28	-0.53341	MC35	-0.84465
MC22	-0.33206	MC22	-0.68576
MC35	-0.15908	MC29	-0.43536
MC33	-0.11825	MC11	-0.41542
MC29	-0.09721	MC33	-0.38002
MC30	-0.09365	MC32	-0.37722
MC41	0.04491	MC31	-0.32106
MC32	0.11361	MC41	-0.26307
MC38	0.15999	MC36	-0.23276
MC43	0.30668	MC38	-0.22093
MC42	0.35022	MC30	-0.21357
MC44	0.36834	MC40	0.01290
MC31	0.38765	MC42	0.01950
MC36	0.40689	MC44	0.14014
MC40	0.46563	MC43	0.14366
MC39	0.58363	MC39	0.23872
MC45	1.04547	MC45	0.74275
MC37	1.09319	MC37	0.92074

Table A5. Switching in the Rank Order of Item Difficulty Estimates Across Average and High Ability Groups Utilizing the 1PL IRT Model: Grade 4

Average Ability Group		Higher Ability Group	
Item Name	Item Diff	Item Name	Item Diff
MC01	-2.36851	MC01	-2.89253
MC02	-1.87901	MC02	-2.13625
MC03	-1.77183	MC03	-2.06797
MC04	-1.71282	MC04	-2.00490
MC06	-1.65710	MC06	-1.95434
MC07	-1.48263	MC07	-1.60004
MC14	-1.39407	MC14	-1.56170
MC05	-1.33854	MC18	-1.54042
MC09	-1.28544	MC05	-1.51957
MC18	-1.23449	MC09	-1.48404
MC17	-1.21589	MC17	-1.46917
MC20	-1.01029	MC21	-1.19890
MC21	-0.95276	MC27	-1.16044
MC08	-0.94763	MC12	-1.14912
MC11	-0.92223	MC20	-1.10843
MC27	-0.91217	MC13	-1.06891
MC15	-0.90717	MC08	-1.05833
MC13	-0.89224	MC15	-0.99977
MC12	-0.87744	MC22	-0.99977
MC10	-0.81471	MC11	-0.99303
MC22	-0.80998	MC10	-0.98632
MC25	-0.69516	MC29	-0.88275
MC31	-0.69516	MC31	-0.87649
MC29	-0.60759	MC25	-0.82724
MC19	-0.57347	MC36	-0.71537
MC36	-0.54817	MC19	-0.67291
MC26	-0.36918	MC40	-0.61758
MC40	-0.34550	MC26	-0.56086
MC37	-0.21754	MC37	-0.47161
MC43	-0.11873	MC41	-0.35240
MC41	-0.08857	MC43	-0.34244
MC28	-0.05472	MC33	-0.27823
MC30	-0.04721	MC30	-0.23185
MC35	0.01277	MC35	-0.19550
MC33	0.01652	MC28	-0.12334
MC23	0.09891	MC16	-0.11855
MC24	0.13268	MC23	-0.08030
MC16	0.15146	MC24	-0.04215
MC32	0.24208	MC32	0.02451
MC38	0.24967	MC38	0.06499
MC39	0.31843	MC34	0.14138
MC34	0.34153	MC39	0.26439
MC44	0.56380	MC44	0.42286
MC45	0.84201	MC45	0.55975
MC42	0.87958	MC42	0.60244

Table A6. Switching in the Rank Order of Item Difficulty Estimates Across Average and High Ability Groups Utilizing the 3PL IRT Model: Grade 4

Average Ability Group		Higher Ability Group	
Item Name	Item Diff	Item Name	Item Diff
MC01	-2.84602	MC01	-3.46277
MC02	-2.51358	MC06	-2.74002
MC20	-2.48328	MC02	-2.68144
MC06	-1.98546	MC03	-2.66214
MC07	-1.93880	MC07	-2.49271
MC03	-1.85978	MC05	-2.30803
MC04	-1.81807	MC11	-2.13639
MC17	-1.62223	MC04	-1.95747
MC21	-1.61280	MC21	-1.81943
MC05	-1.52082	MC14	-1.69927
MC14	-1.52002	MC17	-1.64209
MC11	-1.47334	MC09	-1.62499
MC09	-1.40904	MC18	-1.54767
MC15	-1.10948	MC27	-1.42614
MC18	-1.07894	MC10	-1.35922
MC12	-1.07213	MC29	-1.35516
MC08	-1.05106	MC12	-1.21748
MC29	-0.98997	MC08	-1.20522
MC27	-0.95724	MC13	-1.18484
MC10	-0.93369	MC20	-1.16979
MC22	-0.90617	MC15	-1.12116
MC25	-0.82549	MC22	-1.04823
MC13	-0.76900	MC25	-0.85311
MC36	-0.65503	MC36	-0.80492
MC31	-0.57447	MC31	-0.79867
MC28	-0.51762	MC40	-0.58118
MC19	-0.36514	MC26	-0.53209
MC26	-0.31303	MC19	-0.41660
MC40	-0.29723	MC28	-0.38109
MC41	-0.06418	MC37	-0.23347
MC37	-0.04971	MC41	-0.22129
MC43	0.03505	MC43	-0.18939
MC30	0.11269	MC33	-0.17761
MC35	0.11572	MC30	-0.10352
MC33	0.11587	MC23	0.08544
MC23	0.32113	MC35	0.17135
MC39	0.32455	MC24	0.17658
MC16	0.37585	MC16	0.18052
MC24	0.40835	MC38	0.33467
MC38	0.47026	MC34	0.41338
MC32	0.61812	MC32	0.55348
MC34	0.71063	MC42	0.75560
MC42	0.98647	MC39	0.78946
MC45	1.07041	MC44	0.85879
MC44	1.16079	MC45	1.00537

Table A7. Switching in the Rank Order of Item Difficulty Estimates Across Average and High Ability Groups Utilizing the 1PL IRT Model: Grade 6

Average Ability Group		Higher Ability Group	
Item Name	Item Diff	Item Name	Item Diff
MC01	-1.87490	MC01	-2.38059
MC16	-1.71116	MC13	-1.86617
MC13	-1.51670	MC02	-1.85510
MC02	-1.50994	MC16	-1.65035
MC03	-1.36565	MC03	-1.46107
MC04	-1.31407	MC04	-1.41886
MC06	-1.16694	MC06	-1.33278
MC14	-0.96974	MC14	-1.33278
MC23	-0.93174	MC10	-1.27084
MC07	-0.89874	MC07	-1.17806
MC10	-0.89874	MC15	-1.15045
MC05	-0.85050	MC05	-1.08131
MC15	-0.82689	MC23	-1.06073
MC11	-0.67702	MC20	-0.99101
MC20	-0.67702	MC11	-0.89217
MC30	-0.58182	MC30	-0.75622
MC12	-0.56120	MC08	-0.75199
MC08	-0.54755	MC21	-0.74778
MC21	-0.50702	MC12	-0.71439
MC17	-0.41142	MC17	-0.58572
MC09	-0.40817	MC09	-0.51237
MC35	-0.30902	MC35	-0.48203
MC22	-0.29010	MC36	-0.47450
MC36	-0.25563	MC22	-0.43334
MC18	-0.19658	MC18	-0.41109
MC24	-0.13499	MC24	-0.40370
MC19	-0.12272	MC29	-0.29438
MC29	-0.11659	MC19	-0.27639
MC28	-0.01279	MC39	-0.22627
MC44	0.01463	MC31	-0.20846
MC31	0.03595	MC38	-0.19068
MC41	0.03900	MC28	-0.14459
MC38	0.05729	MC33	-0.13398
MC39	0.09696	MC41	-0.09865
MC25	0.10002	MC37	-0.09159
MC27	0.12448	MC44	-0.06689
MC37	0.12755	MC27	-0.05277
MC33	0.14900	MC25	-0.04572
MC26	0.24781	MC26	-0.01749
MC32	0.32301	MC45	0.11354
MC45	0.37066	MC42	0.11710
MC40	0.41567	MC32	0.18510
MC42	0.41891	MC40	0.22842
MC43	0.57855	MC34	0.45186
MC34	0.58542	MC43	0.46343

Table A8. Switching in the Rank Order of Item Difficulty Estimates Across Average and High Ability Groups Utilizing the 3PL IRT Model: Grade 6

Average Ability Group		Higher Ability Group	
Item Name	Item Diff	Item Name	Item Diff
MC06	-2.62364	MC03	-3.25770
MC01	-2.39407	MC01	-2.69891
MC16	-2.25990	MC06	-2.55661
MC03	-2.11289	MC16	-2.36438
MC13	-1.93953	MC13	-2.09658
MC02	-1.65308	MC02	-2.08281
MC23	-1.42935	MC20	-1.84415
MC04	-1.38916	MC21	-1.45972
MC14	-1.01819	MC10	-1.45717
MC05	-0.97167	MC04	-1.38835
MC10	-0.94455	MC14	-1.35838
MC21	-0.93143	MC05	-1.34498
MC15	-0.92013	MC07	-1.26716
MC11	-0.80920	MC15	-1.19521
MC07	-0.75188	MC11	-1.18273
MC20	-0.72020	MC23	-1.17143
MC30	-0.68073	MC36	-1.12909
MC36	-0.59392	MC08	-0.81178
MC08	-0.47543	MC12	-0.76474
MC12	-0.42971	MC30	-0.69486
MC17	-0.30453	MC35	-0.56655
MC09	-0.29185	MC17	-0.50378
MC29	-0.07033	MC22	-0.47949
MC35	-0.05780	MC09	-0.45168
MC22	0.08916	MC29	-0.27804
MC24	0.11929	MC24	-0.19563
MC31	0.15009	MC38	-0.06988
MC28	0.21966	MC37	-0.06188
MC41	0.22860	MC41	-0.05100
MC37	0.24781	MC18	-0.04350
MC18	0.30502	MC31	0.01062
MC38	0.33646	MC44	0.06043
MC44	0.34362	MC42	0.13283
MC25	0.36398	MC25	0.14562
MC19	0.37999	MC28	0.15257
MC32	0.43826	MC33	0.16108
MC33	0.43899	MC40	0.25217
MC42	0.53698	MC32	0.33826
MC27	0.57342	MC19	0.38846
MC40	0.60334	MC45	0.50140
MC39	0.62937	MC26	0.50602
MC45	0.74221	MC27	0.54652
MC26	0.78861	MC43	0.56028
MC43	0.85181	MC39	0.59584
MC34	1.22380	MC34	1.20846

Tables A9 through A12 present information for the items that switched more than 5 places in rank order across the two ability groups. All of the rank order switches of this magnitude occurred with the 3PL calibrations with the exception of item 16 at grade 3 which switched more than 5 positions with the 1PL run. Since this was the only instance of a rank order switch of more than 5 places with the 1PL, information is not presented for item 16 at grade 3.

Table A9. Classical Item Statistics for Items with the Greatest Degree of Rank Order Switching with the Three Parameter Model at Grade 2.

Average Ability Group			High Ability Group		Difference	MH-Dif Average Ability Reference Group
Item #	P-Value	Pt.-Biserial	P-Value	Pt.-Biserial	p / pb	ETS Classification
9	.67	.45	.74	.42	.07 / -.03	A
21	.78	.34	.81	.34	.03 / .00	A
28	.75	.47	.79	.52	.04 / .05	A
Mean 40	.70	.44	.75	.42	.05 / -.02	NA

Note: When the items were sorted in descending order by item difficulty (p-values) for the average ability group; item 21 ranked 16th out of the total set of 43 items, item 28 ranked 19th , and item 9 ranked 25th.

Table A10. Classical Item Statistics for Items with the Greatest Degree of Rank Order Switching with the Three Parameter Model at Grade 3.

Average Ability Group			High Ability Group		Difference	MH-Dif Average Ability Reference Group
Item #	P-Value	Pt.-Biserial	P-Value	Pt.-Biserial	p / pb	ETS Classification
12	.85	.49	.89	.45	.04 / -.04	A
13	.77	.47	.82	.42	.05 / -.05	A
27	.67	.46	.75	.48	.08 / .02	A
31	.53	.35	.62	.32	.09 / -.03	A
Mean 41	.69	.44	.75	.42	.06 / -.02	NA

Note: When the items were sorted in descending order by item difficulty (p-values) for the average ability group; item 12 ranked 9th out of the total set of 45 items, item 13 was 15th , item 27 was 28th , and item 31 ranked 37th.

Table A11. Classical Item Statistics for Items with the Greatest Degree of Rank Order Switching with the Three Parameter Model at Grade 4.

Average Ability Group			High Ability Group		Difference	MH-Dif Average Ability Reference Group
Item #	P-Value	Pt.-Biserial	P-Value	Pt.-Biserial	p / pb	ETS Classification
15	.76	.43	.79	.43	.03 / .00	A
20	.78	.40	.81	.45	.03 / .05	A
39	.41	.37	.42	.31	.01 / -.06	A
Mean 42	.66	.45	.70	.42	.04 / -.03	NA

Note: When the items were sorted in descending order by item difficulty (p-values) for the average ability group; item 20 ranked 12th out of the total set of 45 items, item 15 was 17th and item 39 ranked 41st .

Table A12. Classical Item Statistics for Items with the Greatest Degree of Rank Order Switching with the Three Parameter Model at Grade 6.

Average Ability Group			High Ability Group		Difference	MH-Dif Average Ability Reference Group
Item #	P-Value	Pt.-Biserial	P-Value	Pt.-Biserial	p / pb	ETS Classification
20	.71	.44	.78	.43	.07 / .-01	A
23	.77	.47	.80	.49	.03 / .02	A
28	.51	.48	.54	.47	.03 / -.01	A
Mean 42	.61	.42	.66	.42	.05 / .00	NA

Note: When the items were sorted in descending order by item difficulty (p-values) for the average ability group; item 23 ranked 9th out of the total set of 45 items, item 20 was 14th and item 28 ranked 29th.

Appendix B

To investigate item parameter invariance based on simulated data, the authors generated item response data using WinGen2 (Han, 2007). This program can generate dichotomous and polytomous item response data for several IRT models such as parametric and non-parametric models and for many ability distribution conditions that resemble reality (e.g., normal and skewed ability distributions) (Han & Hambleton, 2007).

The present study utilized WinGen2 (Han, 2007) to generate dichotomous item response data. Specifically, item parameters and ability distributions were independently specified. The first step was to combine the ”real” data for the low, average and high ability groups into one data set at each grade. A 3PL item calibration was then conducted separately for each grade: 2, 4, and 6. The item parameter estimates derived from these runs served as “target” item parameter estimates for the data simulations that were to follow.

As for ability distributions, each grade was postulated to have two ability groups, one of lower- and one of higher ability. The ability distribution for the lower ability group was derived from a normal distribution with a mean of -0.50 and a standard deviation of 1.00, whereas the ability distribution for the higher ability group was originated from a normal distribution with a mean of 0.50 and a standard deviation of 1.00. Both ability groups contained 1000 simulees.

Given these specifications for item parameters and ability distributions, dichotomous item response datum was generated for every item by examinee for three grades (i.e., 2nd, 4th, and 6th grades) by two ability groups (i.e., lower and higher ability groups), which resulted in six simulated item response data sets.

Table B1 presents descriptive statistics and test reliability for the simulated data sets by grade and ability. The mean raw scores were consistently higher for the higher ability group than for the lower ability group across grades. Likewise, the standard deviations were consistently lower for the higher ability group. The reliabilities were all above .80.

Table B1. Performance Characteristics of Lower and Higher Ability Groups

Grade	Ability Group	N	# of Items	Mean	SD	Cronbach's Alpha
2	Lower	1000	43	25.66	7.59	.85
	Higher	1000	43	32.16	6.61	.85
4	Lower	1000	45	26.56	7.33	.84
	Higher	1000	45	32.88	6.58	.83
6	Lower	1000	45	24.30	7.48	.83
	Higher	1000	45	30.81	7.14	.84

Each of the six simulated data sets was then analyzed using PARSCALE under the 1PL and 3PL models utilizing the same program control settings as were used for the "real" data. Table B2 presents the item fit results at each grade for the simulated data. The number of misfitting items across grade and ability group ranged from 40 to 44 for the 1PL and between 0 and 3 for the 3PL. Given that the data was simulated utilizing 3PL item parameter estimates, it is not surprising that the 3PL fit the data much better that the 1PL. However the number of misfitting items with the 1PL is surprising.

Table B2. Number of Items Identified* as Misfitting with the 1PL and 3PL IRT Models by Grade and Ability Group for the Data Simulation

Grade	Ability Group	# of Items	# of Misfitting items 1PL	3PL
2	Lower	43	42	0
	Higher	43	40	0
4	Lower	45	44	0
	Higher	45	42	0
6	Lower	45	44	3
	Higher	45	42	2

*****Note: Items are considered misfitting at alpha=.01 level.

Table B3 presents the Pearson and Spearman rank order correlations between the IRT item difficulty estimates for the low and higher ability groups for each grade and IRT model. In all grades the item measures correlated most highly with the one-parameter model. Across grades the average of the Pearson correlations was .970 for the 1PL and .927 for the 3PL. The average for the Spearman rank order correlations was .967 for the 1PL and .922 for the 3PL.

Table B3. Correlations of IRT Item Difficulty Parameter Estimates between Lower and Higher Ability Groups

	1PL	1PL	3PL	3PL
Grade	Pearson	Spearman’s Rho	Pearson	Spearman’s Rho
2	.972	.971	.919	.911
4	.971	.969	.957	.951
6	.968	.961	.906	.904
Mean	.970	.967	.927	.922