Sample Size and Item Parameter Estimation Precision When Utilizing the One-Parameter “Rasch” Model

Michael Custer
Houghton Mifflin Harcourt - Riverside

Paper Presented at the Annual Meeting of the
Mid-Western Educational Research Association
Evanston, Illinois
October 21-24
2015

Please send correspondence regarding this paper to Michael.Custer@hmhco.com

Abstract

This study examines the relationship between sample size and item parameter estimation precision when utilizing the one-parameter model. Item parameter estimates are examined relative to “true” values by evaluating the decline in root mean squared deviation (RMSD) and the number of outliers as sample size increases. This occurs across three conditions. The first condition utilizes one test level of 40 items without missing data. The second incorporates missing data and the third utilizes a vertical scale across three levels. When test levels were calibrated individually, the majority of RMSD and outlier-reduction improvement was achieved once a sample size of 400 examinees had been reached. Incremental increases to sample size beyond a threshold of 500 examinees added little gain to estimation precision.

With the vertical-scaling condition, the majority of RMSD and outlier-reduction improvement was seemingly reached with a sample size of approximately 300 examinees per level with 500-600 examinees per common item linking set. However, when compared to conditions one and two, a similar level of estimation precision was not reached until a sample size of approximately 400-450 examinees per each of the three levels.

Background

There exists a positive relationship between incremental increases in sample size and estimation precision when utilizing Item Response Theory (IRT) models to estimate item parameters. An inadequate sample size can lead to increased estimation error with negative implications for the analysis of item and test data including IRT based test construction (Hambleton, Jones & Rogers, 1993; Swaminathan et al, 2003; He & Wheadon, 2012). Sample size considerations are especially important with activities typically associated with high stakes assessments. Development of a scale score reporting metric, the use of IRT-based ability estimates to derive cut scores, IRT true score equating and item calibrations associated with "item banking" for a computer adaptive test (CAT) are examples of activities where estimation precision is critical and larger sample sizes are generally recommended. (Reeve & Fayers, 2005)

From a theoretical perspective, IRT models of greater complexity require larger sample sizes to achieve a similar level of estimation precision than do less complex models. In this manner, multi-dimensional IRT models generally require a greater sample size than do unidimensional models. Likewise, IRT models with more parameters require a greater sample size than do models with fewer parameters. Research and practice has generally shown that smaller sample sizes are required with the Rasch Model, a one-parameter model, than is the case with the two-parameter or three-parameter models (Edelen & Reeve, 2007).

Though theoretical underpinnings exist, there has been little empirical research focused on the relationship between sample size and estimation precision. Reise and Yu (1990) in a study using MULTILOG and the graded response model (a generalization of the two-parameter model) to estimate item and ability parameters recommended a sample size of at least 500 examinees to achieve an adequate level of estimation precision. They also recommended that as many as 1000 to 2000 examinees may be required for each level or form when a linking design is utilized to calibrate multiple levels or forms to the same underlying scale.

Sample size and estimation precision has also been linked through the standard error of estimate. Embretson and Reise (2000) recommended that, with regard to sample size, researchers should focus on the size of the standard error of item parameter estimates to ensure that these are reasonable. Further, Wasserman and Bracken (2003) noted that the standard error of item parameter estimates becomes smaller as sample size increases up to a point after which diminishing returns are expected. It is this notion of diminishing returns that will be the central focus for the remainder of this paper.

Method

The focus of this study is an examination of the recovery of estimated IRT item parameters relative to their "true" values by evaluating the decline in root mean squared deviation (RMSD) and the number of outliers as sample size increases. This evaluation occurs across three conditions. For the first condition, WINGEN2 (Han, 2007) was used to simulate set of 40 dichotomously scored items with no missing data and an overall item difficulty mean of 0.00 with a standard deviation (SD) of 1.00. This condition utilized a set of 3,000 examinees simulated to have an ability mean of 0.00 with an SD of 1.00. "True" item parameter values were derived by calibrating the 40 items with the full set of 3000 examinees. In order to derive item parameter measures across 10 levels of sample size (10 levels ranging from 100 to 1000 examinees in increments of 100), random samples of 100 thru 1000 cases were drawn from the simulated set of 3000 examinees. Ten replications were executed for each of the 10 levels of sample size. This process created 100 data sets (10 levels of sample size x 10 replications) to be used for the item calibrations. WINSTEPS 3.65 (Linacre, 1991-2002) was used to calibrate each data set with the Rasch Model. Item parameter estimates were evaluated relative to their “true” base calibration values. For each level of sample size the average RMSD was computed as well as the average number of outlier items (defined as the number of item difficulty measures that differ from “true” by .20 or more) across the 10 runs. The selection of the threshold value of .20 reflects the notion that item difficulty measures that differ from their base calibration measure by less than .20 or .30 have no practical impact on person measurement (Wright, 1977). Also, a threshold of .20 is more than 4 times the average of the base calibration item standard errors of estimate which was equal to .043.

As with condition one, the second condition utilized a simulated set of 40 dichotomously scored items simulated to have an overall item difficulty mean of 0.00 with an SD of 1.00. Likewise, this condition utilized a set of 3,000 examinees simulated to have an ability mean of 0.00 with an SD of 1.00. However, with condition two missing data was applied to the simulated data set of 3000 examinees. For this condition, each item was omitted an average of 48 times (1.62% of all examinees) with 20% of the examinees not reaching the end of the test. Item responses were selected for omission randomly and were programmatically embedded according to the notion that omits could also be guessed at. Hence, item data that had originally been scored as 1 was selected for omission 25% of the time and item data that had been originally scored as 0 was selected for omission 75% of the time. Not-reached items were embedded by utilizing a random selection of examinees to sequentially embed not-reached item strings, which varied in length, between items 33 and 40 by programmatically assigning blanks to item response strings extending from a given item within the 33-40 range to the end of the test. “True” item parameter values were then derived by calibrating the 40 items with the full set of 3000 examinees. Random sample draws for each of 10 levels of sample size were derived in a similar manner as was done with condition one. This process was replicated 10 times, for each of 10 levels, resulting in 100 data sets. Each data set was then calibrated with WINSTEPS. The average RMSD, between estimated and “true” parameters, was computed as well as the average number of outliers (defined as the number of item difficulty measures that differ from “true” by .20 or more) across the ten runs for each level of sample size.

The third condition utilized three levels of data with levels 1 and 3 containing 40 items and level 2 containing 50 items. The vertical scaling design is presented in Chart A.1 of the Appendix. The levels were vertically linked using common items with 10 items linking between level 1 and 2 and 10 items linking between levels 2 and 3. As a result this third ”vertical scaling” condition incorporated three test levels with 30 operational items and 10 common items for levels 1 and 3 and 20 common items for level 2 (10 in common with level 1 and 10 in common with level 3). All of the items were dichotomously scored.

Under the vertical scaling design, items were concurrently calibrated across levels. Prior to this calibration, the level 1 items were initially simulated to have a target difficulty mean of -.20 and SD of .50. The level 2 items were simulated to have a mean of .00 and SD of .50 and the level 3 items were simulated to have a target difficulty mean of .20 and SD of .50. Condition three utilized simulated data sets across each of the three levels with each level containing 3,000 examinees. Level 1 examinees were initially simulated to have a target ability mean of -.20 and SD of .50. The level 2 examinees were simulated to have a mean of .00 and SD of .50 and the level 3 examinees were simulated to have a target ability mean of .20 and SD of .50. Once the data had been simulated, p-values were computed for each of the linking items to ensure that the differences in linking item difficulty across levels demonstrated reasonable growth. A comparison of the linking item p-values is presented in Table A.2 of the Appendix.

"True" item parameter values were derived through a concurrent calibration of the 110 items with the full set of 9000 examinees (3 levels * 3000 examinees). The post-concurrent calibration "true" means and SDs are reported in the Table A.3 of the Appendix. In order to derive item parameter measures across 15 levels of sample size (15 levels ranging from 100 to 1500 examinees in increments of 100), random samples of 100 thru 1500 cases were drawn from the simulated set of 9000 examinees. Ten replications were executed for each of the 15 levels of sample size. This process created 150 data sets (15 levels of sample size x 10 replications) to be used for the item calibrations. As was the case with conditions one and two, item parameter estimates were evaluated relative to their "true" base calibration values. For each level of sample size and across the 10 runs, the average RMSD was computed as well as the average number of outlier items.

Results

For each of the three conditions respectively, the mean RMSD across the 10 runs for each level of sample size is presented in Table 1. A decrease in RMSD is indicative of improved retrieval of the "true" item parameters. This improvement in "true" parameter recovery as sample size increases is presented in Figures 1-3. Likewise, the number of outlier items relative to their "true" parameter values is presented in Table 2 with improvement as sample size increases plotted in Figures 4-6. For all three conditions, the points where the curves flatten are suggestive of a threshold where gains in estimation precision from incremental increases in sample size appear to become consistent and small.

With respect to condition one (one level with no-missing data), depicted in Figures 1 and 4, as sample size increases the majority of RMSD and outlier-reduction improvement is reached by an n-count of 400 and the shape of the curves begin to flatten at an n-count of 500 (RMSD = 0.087). This flattening effect is especially noticeable in Figure 4 which plots the number of outlier items. As shown in Table 2, at an n-count of 500 the number of outlier items is 2.75%.

With condition two (one level with missing data), depicted in Figures 2 and 5, the majority of RMSD and outlier-reduction improvement is reached by an n-count of 400 and the shape of the curves begin to flatten at an n-count of between 500-600 examinees (RMSD ~= 0.088). This flattening effect is especially noticeable in Figure 5 which plots the number of outlier items. As presented in Table 2, at an n-count of 500 the number of outlier items is at 6.00% but falls to 3.00% at an n-count of 600.

With respect to condition three ("vertical scale" across three levels) depicted in Figures 3 and 6, this flattening seemingly occurs at an overall sample size of between 800-900 examinees (RMSD ~= 0.119). At this sample size there were 267-300 examinees per level with between 534 and 600 examinees being administered the common items between each adjacent level. As shown in Table 2, at an n-count of between 800 and 900, the number of outlier items is roughly 10%. It is important to note that with the vertical scaling condition, similar levels of estimation precision as encountered with conditions one and two were not reached until an overall sample size of between 1,200 and 1,400 examinees (roughly 400-450 per level with approximately 800 or more examinees being administered the common items between each adjacent level) was reached.

Discussion

Calibration objectives are important when practitioners consider the issue of estimation precision and sample size. For example, a high degree of estimation precision is typically required for purposes such as scale construction or the calibration of items for an item bank. In contrast, less estimation precision is typically required for a pilot study item tryout.

When test levels were calibrated individually, the majority of RMSD and outlier-reduction improvement was achieved once a sample size of 400 examinees had been reached which, depending upon the practitioner's calibration objectives, may be a reasonable starting point when considering the issue of appropriate sample size. Support can also be found for the Reise and Yu (1990) study which recommended a sample size of at least 500 examinees when utilizing the graded response model. Though this study utilizes the one-parameter model, the results suggest a similar sample size of 500 examinees as the point at which the gains from incremental increases in sample size become consistent and small.

A concurrently calibrated vertical scaling across three levels was also studied. As might be expected an overall sample size of 500 (on average 167 examinees per each of three levels) was inadequate as far as estimation precision is concerned (21.73% of the items were outliers). An overall sample size of between 800-900 examinees, with approximately 300 examinees per level and between 500 and 600 examinees administered common items between each adjacent level, seemed to be the point by which a majority of the RMSD and outlier-reduction improvement had been reached. In this case, 10% of the items differed from their "true" difficulty measure by .20 or more. This level of sample size (roughly 300 examinees per level and between 500 and 600 examinees administered common items between each adjacent level) might be thought of as a reasonable starting point when thinking about vertical scaling. However, it may not provide an adequate level of estimation precision for all vertical scaling objectives. This is evident when one considers that similar levels of estimation precision as were reached with conditions one and two were not reached with the vertical scaling condition until an overall sample size of between 1200-1400 examinees (approximately 400-450 examinees per level with 800 or more examinees being administered the adjacent level common items).

Practitioners are accustomed to the trade-off that exists between the costs of obtaining a sample and estimation precision. When test levels were calibrated individually, the majority of RMSD and outlier-reduction improvement was achieved once a sample size of 400 examinees had been reached. Incremental increases to sample size beyond a threshold of 500 examinees added little to estimation precision. This threshold was slightly larger in the presence of missing data.

With the vertical scaling condition, the concurrent calibration of just three test levels limits the generalizability of the results. However, the study of diminishing returns as utilized in this paper can be employed to help define the sample size required to obtain desired levels of estimation precision for a prospective scaling across just one level or across multiple levels. Given a known number of tests and levels to be scaled, practitioners might be encouraged to first simulate data and vary the level of sample size to evaluate estimation precision across the scale as part of the vertical scaling design process. This evaluation would aid practitioner's in their selection of an appropriate sample size.

Limitations

The results from this study should be viewed in light of the fairly strict restrictions imposed by the study design that directly affect the generalizability of the results. First, this study utilized the Rasch Model as implemented by WINSTEPS. If instead, the two parameter or three parameter logistic models were used, it seems likely that the sample size needed for satisfactory estimation precision would be greater simply on account of the additional parameters being estimated.

Second, the number of items does not vary for conditions one and two. In both cases 40 items were scaled. Future research might utilize item sets with fewer items to evaluate the impact on RMSD and the number of outlier items.

Third, the vertical scaling portion of this study incorporated a concurrent vertical scaling utilizing only three test levels. It seems plausible that if concurrently scaling across four or more levels, that larger sample sizes at each level would be necessary in order to achieve similar levels of estimation precision.

Each of the above limitations restrict the generalizability of these results. However, despite these limitations, this study incorporates a practical methodology that provides useful guidelines for the study of the relationship between sample size and estimation precision.

References

Edelen, M. O.; & Reeve, B. B. (2007). Applying Item Response Theory (IRT) Modeling to Questionnaire Development, Evaluation, and Refinement. *Quality of Life Research, 16(Supp 1),* 5-18.

Embretson, S.E. and Reise, S.P (2000) Item Response Theory For Psychologists. Mahwah, NJ: Lawrence Erlbaum Associate.

Hambleton, R., Jones, R and Rogers, H. (1993) Influence of Item Parameter Estimation Errors in Test Development. *Journal of Educational Measurement* 30, 143-155.

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/

He, O and Wheadon, C. (2012) The Effect of Sample Size on Item Parameter Estimation For the Partial Credit Model. Centre for Education Research and Policy, www.cerp.or.uk

Linacre, J.M (1991-2002). *WINSTEPS 3.65* [Computer Software]. Chicago, IL.: John M. Linacre www.WINSTEPS.com.

Linacre, J.M (1991-2002). *A User's Guide to WINSTEPS MINISTEP Rasch-Model Computer Program*. Chicago, IL.: Winsteps.com.

Swaminathan, H., Hambleton, R., Sireci, S., Xing, D. and Rizavi, S. (2003) Small Sample Estimation in Dichotomous Item Response Models: Effect of priors Based on Judgemental Information on the Accuracy of Item Parameter Estimates. *Applied Psychological Measurement* 27, 27-51.

Stahl, J., Muckle, T., (2007) Investigating Drift Displacement in Rasch Item Calibrations. *Rasch Measurement Transactions*, 2007, 21:3 p.1126-1127.

Reeve, B.B. and Fayers, P. (2005) Applying Item Response Theory Modeling for Evaluating Questionnaire Items and Scale Properties. In P.M. Fayers & R.D. Hays (Eds.) Assessing Quality of Life in Clinical Trials: Methods and Practice (2nd Edition pp. 55-73). New York, NY: Oxford University.

Reise, Steve P., Yu, Jiayuan, (1990) Parameter Recovery in the Graded Response Model Using Multilog. *Journal of Educational Measurement* 27, 133-144.

Wasserman, J. D., & Bracken, B. A. (2003). Psychometric Considerations of Assessment Procedures. In J. Graham and J. Naglieri (Eds*). Handbook of Assessment Psychology (pp. 43 – 66).* New York: Wiley.

Wright, B. (1977) Solving Measurement Problems with the Rasch Model. *Journal of Educational Measurement* 14 (2) pp. 97-116, Summer 1977 (and MESA Memo 42)

Table 1. RMSD and Change As Sample Size Increases

Sample Size	Complete-No Missing Data (One Level 40 Items)		Missing Data–20% Not Reached (One Level 40 Items)		Vertical Scale Across 3 Levels	
	Item RMSD	Diff from Prev. Level	Item RMSD	Diff from Prev. Level	Item RMSD	Diff from Prev. Level
100	0.222		0.232		0.388	
200	0.164	-0.058	0.167	-0.065	0.275	-0.063
300	0.124	-0.040	0.127	-0.040	0.210	-0.065
400	0.097	-0.027	0.105	-0.022	0.185	-0.025
500	0.087	-0.010	0.102	-0.003	0.160	-0.025
600	0.084	-0.003	0.088	-0.014	0.147	-0.013
700	0.079	-0.005	0.081	-0.007	0.135	-0.012
800	0.071	-0.008	0.074	-0.007	0.120	-0.015
900	0.066	-0.005	0.067	-0.007	0.119	-0.001
1000	0.060	-0.006	0.064	-0.003	0.111	-0.008
1100					0.102	-0.009
1200					0.094	-0.008
1300					0.092	-0.002
1400					0.088	-0.004
1500					0.085	-0.003

Figure 1

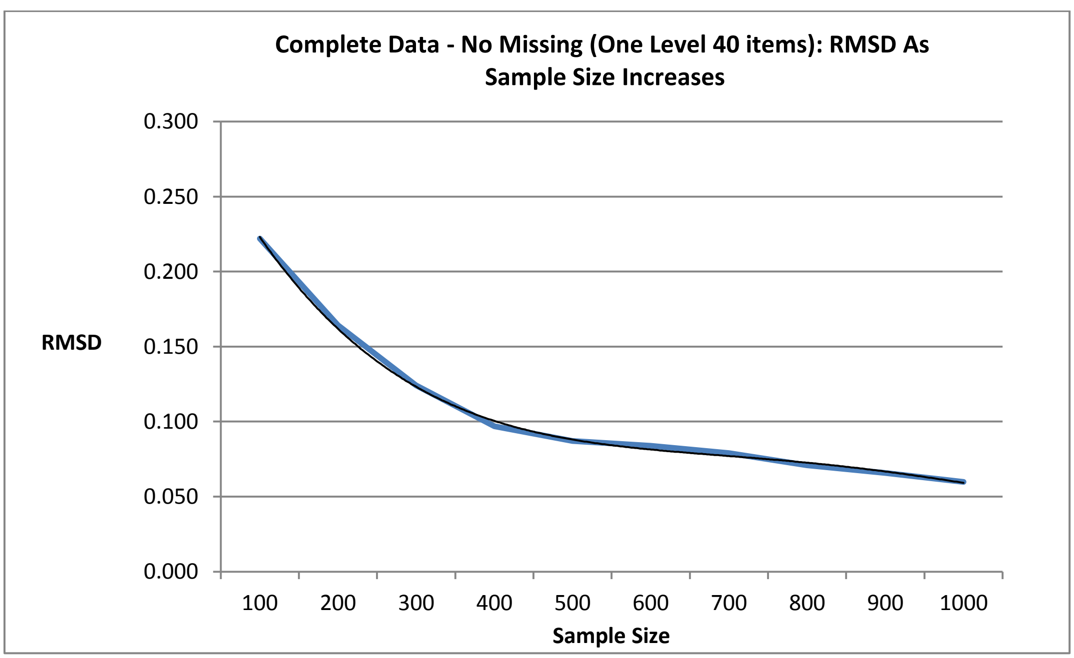


Figure 2

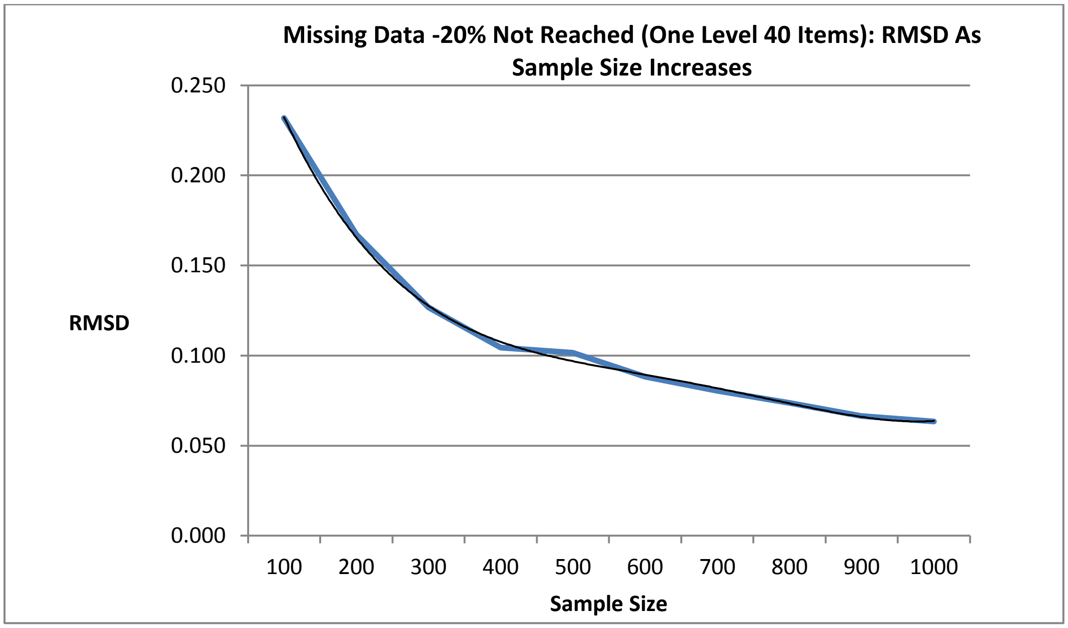


Figure 3

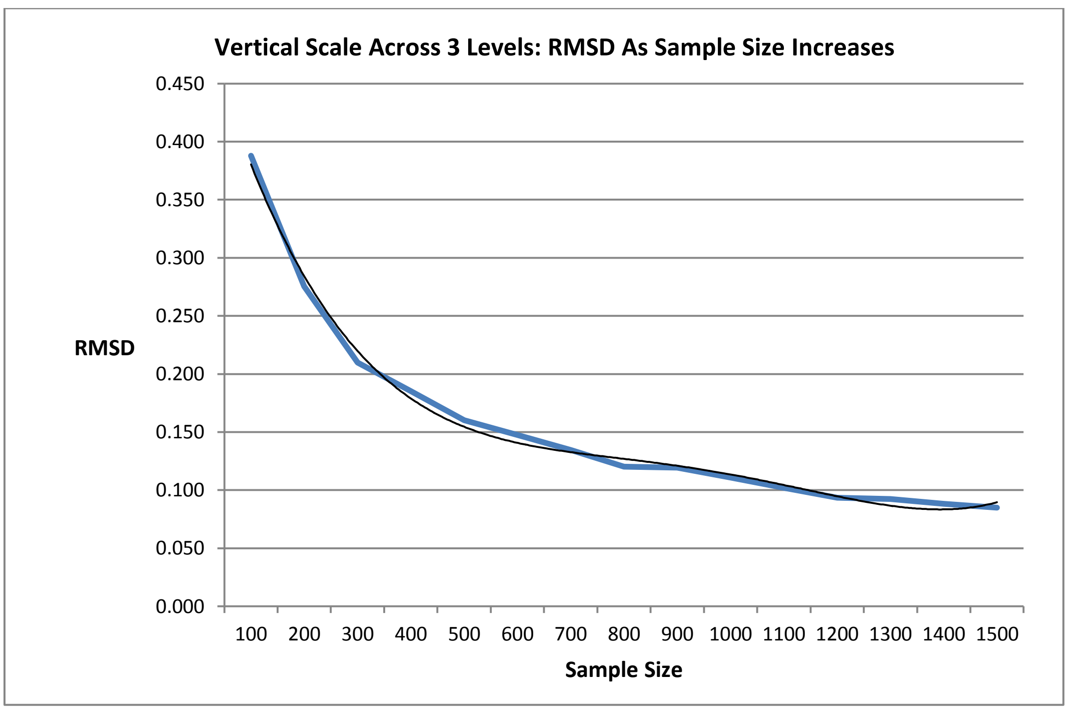


Table 2. The Number of Items with Item Difficulty Measures that Differ From "True Parameter Value" By .20 or More As Sample Size Increases

Sample Size	Complete-No Missing Data (One Level 40 Items)		Missing Data–20% Not Reached (One Level 40 Items)		Vertical Scale Across 3 Levels	
	# of Items Diff > .20	Percent n / 40	# of Items Diff > .20	Percent n / 40	# of Items Diff > .20	Percent n / 110
100	16.00	40.00%	15.90	39.75%	66.60	60.55%
200	10.40	26.00%	9.30	23.25%	52.00	47.27%
300	4.30	10.75%	5.40	13.50%	38.20	34.73%
400	2.00	5.00%	2.30	5.75%	32.20	29.27%
500	1.10	2.75%	2.40	6.00%	23.90	21.73%
600	0.90	2.25%	1.20	3.00%	19.00	17.27%
700	0.70	1.75%	1.00	2.50%	16.60	15.09%
800	0.40	1.00%	0.80	2.00%	11.30	10.27%
900	0.50	1.25%	0.20	0.50%	11.20	10.18%
1000	0.40	1.00%	0.00	0.00%	9.20	8.36%
1100					6.20	5.64%
1200					4.80	4.36%
1300					3.30	3.00%
1400					3.50	3.18%
1500					1.90	1.73%

Figure 4

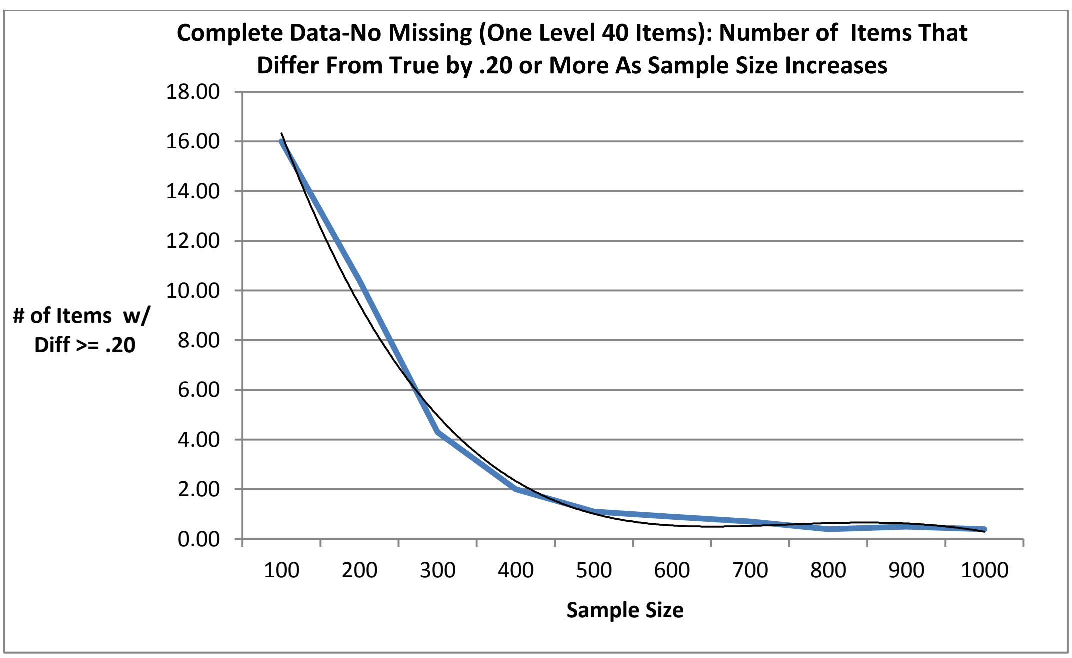


Figure 5

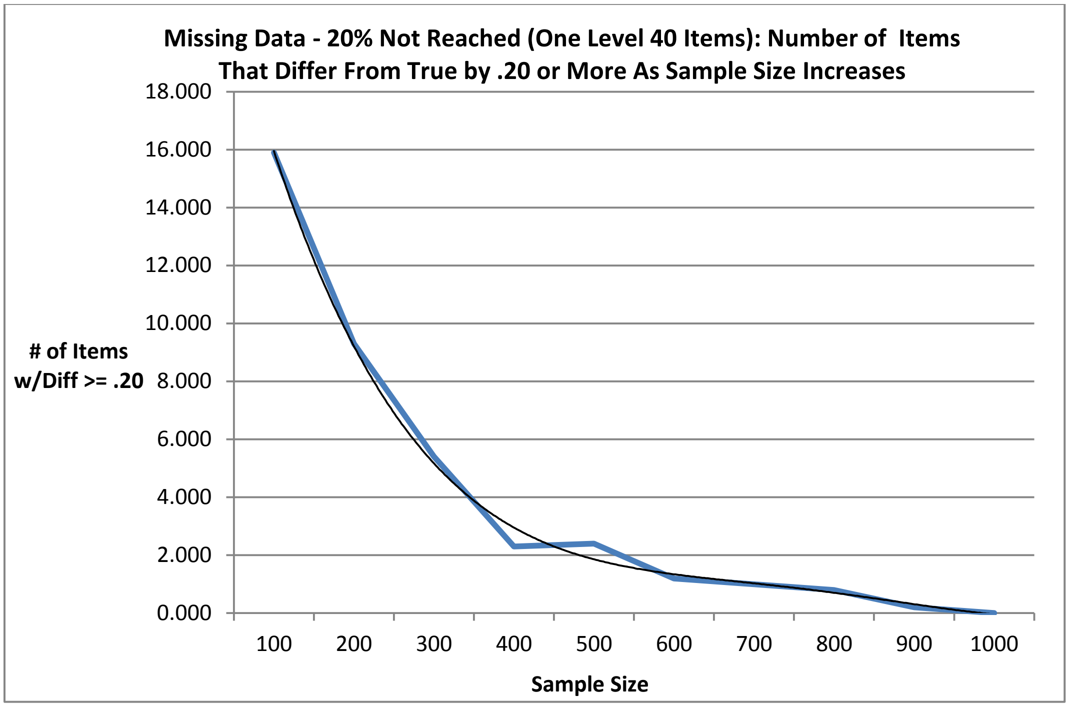
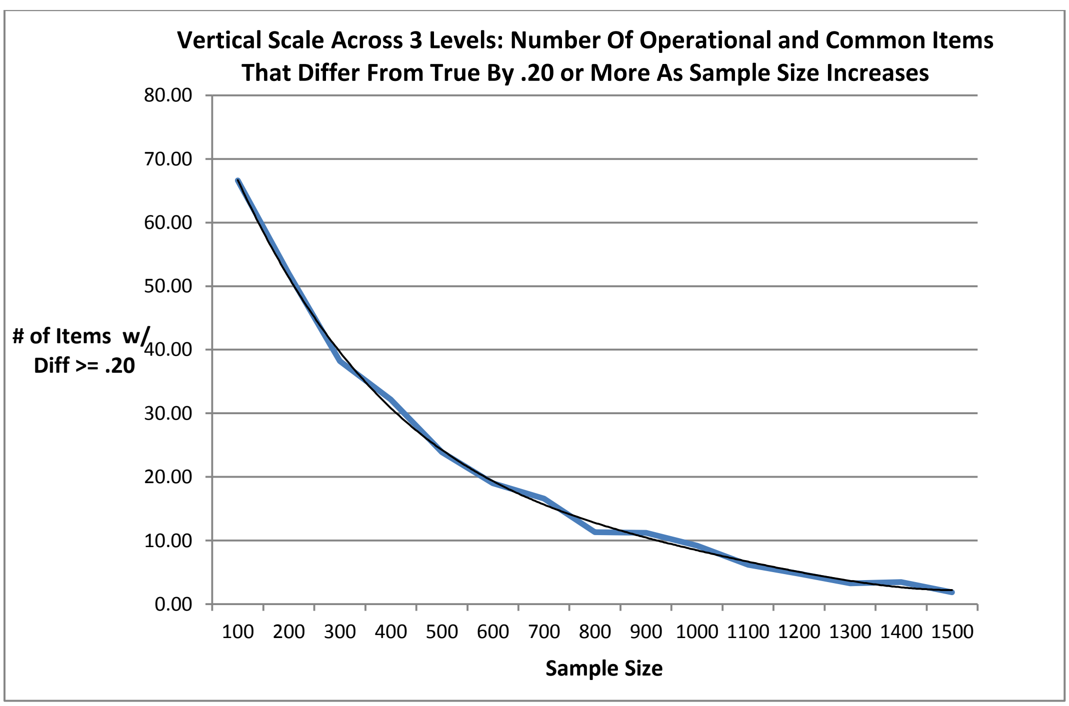
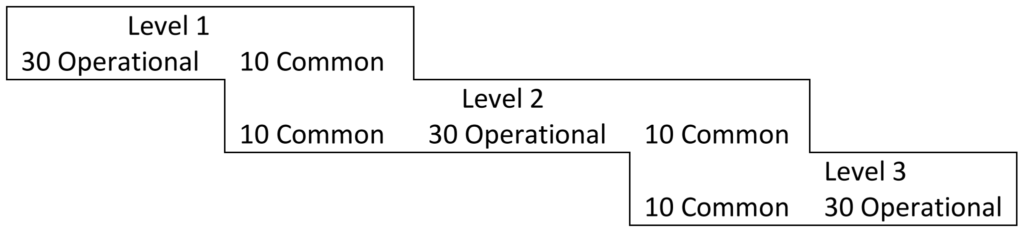


Figure 6



Appendix

A.1) Vertical Scaling Design for Condition Three



A.2) Vertical Scaling: Linking Item p-Values Across Levels

Linking Item p-Values											
	Linking Items - Level 1 Item Position/Level 2 Item Position										
	31/1	**32/2**	**33/3**	**34/4**	**35/5**	**36/6**	**37/7**	**38/8**	**39/9**	**40/10**	**Mean**
Level 1: p-Values	0.44	0.40	0.38	0.39	0.37	0.42	0.41	0.38	0.39	0.38	**0.40**
Level 2: p-Values	0.74	0.71	0.70	0.69	0.67	0.67	0.65	0.63	0.61	0.59	**0.67**
	Linking Items - Level 2 Item Position/Level 3 Item Position										
	41/1	**42/2**	**43/3**	**44/4**	**45/5**	**46/6**	**47/7**	**48/8**	**49/9**	**50/10**	**Mean**
Level 2: p-Values	0.43	0.42	0.41	0.40	0.41	0.40	0.40	0.39	0.39	0.37	**0.40**
Level 3: p-Values	0.71	0.68	0.68	0.65	0.66	0.64	0.63	0.62	0.62	0.59	**0.65**

A.3) Vertical Scaling: Concurrent Calibration Descriptive Statistics for "True" Items and Abilities

Vertical Scale	**Item**		**Abilities**	
Level	**Mean**	**SD**	**Mean**	**SD**
1	-1.211	.445	-1.17	.644
2	-.028	.442	.028	.589
3	1.183	.565	1.309	.628
Overall	0.000	1.122	0.056	1.188