Influence of Item Pool Characteristics on Repeated Measures for Student Growth in Computerized Adaptive Testing

In the Session "Promises and Challenges of Computerized Adaptive Testing in K-12 Assessments"

Liru Zhang
Delaware Department of Education

C. Allen Lau
Pearson

Shudong Wang
Northwest Evaluation Association

Paper presented at the annual meeting of the National Council on Measurement in Education.
April 28-30, 2013, San Francisco, CA.

Introduction

Stimulated by the *NCLB Act* of 2001, the Race to the Top (RTTT) initiatives and the recent Elementary and Secondary Education Act (ESEA), flexibility guidelines have continued to push using student growth data to monitor school performance and educator evaluation. Common Core State Standards (CCSS) and next-generation assessments are central to the RTTT in recent educational reforms, which clearly accentuate the need for new, innovative assessments that provide high-quality measures of student learning gains and sufficient accuracy to support RTTT's use in evaluating the effectiveness of instructional strategies and individual teachers. With wide application of advanced technology and the potential of computerized-adaptive testing (CAT), policy makers and the general public are becoming interested in using CAT for statewide K-12 assessments. The Smarter Balanced Assessment Consortium (SBAC), supported by federal funding, is designed to deliver a computerized-adaptive testing program by 2014-2015 to its over 20 members for high-stakes accountability.

A substantial body of literature has been published about the advantages and disadvantages of CAT in psychological and educational testing. In recent years, CAT's potentials have been recognized for determining student growth in terms of efficient testing, improved precision and test security (Betebenner and Linn, 2010; Ballou, 2008; Yen, 1986). In adaptive testing, each successive item in the test is chosen by a set of constraints that aims to maximize information at test takers' estimated ability levels or to minimize the deviation of the information from a target value at the estimate. In addition to content constraints, constraints on exposure rates of items in the pool is an essential element to maintain test security. Although various approaches and methods of item-exposure control (Sympson and Hetter, 1985; Stocking and Lewis, 1998, 2000; van der Linden and Veldkamp, 2004, 2007; Shin, et al. 2009) have been developed and implemented to guarantee upper bounds of exposure-rate, there is no control for the lower bounds. In "practical experience on adaptive testing, item pools often have surprisingly large subsets of items that are seldom administered" (p. 232); under-exposed items are often overrepresented in the pool or lack desirable characteristics to meet those constraints for item selection on the test (Veldkamp and van der Linden, 2010). The presence of unused items in the pool is an unfortunate waste of resources. Reckase (2007) indicated that the procedures of adaptive testing do not function as expected unless the item pool provides appropriate items that match the criteria of the item selection algorithm. Consequently, good estimates for test takers cannot efficiently be made for test takers on the construct with maximum information and minimum measurement errors. To realize many of the measurement advantages of adaptive testing, the item pool must contain sufficient number of high-quality items over a wide range of proficiency, and these items must have content domain and item difficulty characteristics to provide adequate information (Flaugher, 1990; Wise, 1997). The integrity of adaptive testing depends on the stability of item parameters over time. Over-exposed items not only jeopardize test security, but may result in positive bias of ability estimation, which seriously threaten the validity of high-stakes assessments (Wise, 1997; Iramaneeray and Stahl, 2007).

The current study investigates the relationships between item pool characteristics (e.g., pool size, content balance and item difficulty distributions) and item exposure rate and its influence on student growth measures. The Rasch-based assessment program is a fixed-length, 50-item computerized adaptive test using the joint maximum likelihood approach for item calibration and scoring by Winsteps (3.60.1).

Methods of Study

Purpose of Study

This study investigates the influences of item pool characteristics on student growth measure in mathematics and reading. Use empirical data and simulation, item pools are evaluated in terms of content and statistical characteristics (e.g., pool size, content balance, and item difficulty distributions). The relationship between the item pool characteristics and item exposure rates is analyzed and its influence is estimated on student growth in repeated measures.

Assessment Instrument

The Delaware Comprehensive Assessment System (DCAS) reading and mathematics are Rasch-based, fixed-length online adaptive tests. To measure academic growth, students are required to take DCAS multiple times throughout each school year. Gain scores from fall to spring are used to determine student growth, which serves as an indicator for educator evaluation. Table 1 displays the schedule for the 2010-2011 administrations. Schools are responsible for scheduling assessments based on available technology facilities and arrangement of classroom instruction. Students can accomplish each test in multiple sessions.

The DCAS mathematics and reading are designed to measure the Delaware Prioritized Content Standards with multiple-choice (MC) and machine-scored constructed-response (MSCR) questions of varying score scales from 0-1 to 0-4. In reading, students read a passage and answer all or selected attached items measuring two standards: Informative and Literary. In mathematics, items measure four content categories: Numeric Reasoning, Algebraic Reasoning, Geometric Reasoning and Quantitative Reasoning.

Two primary criteria, matching the test specifications and matching estimated ability for students, are prioritized in the item selection algorithm. Students are not seeing the same item(s) across multiple administrations within a school year, except to satisfy the test specifications. The blended-design assessments derive two scores: the accountability scores, based on 40-50 on-grade items, and the instructional scores, based on both on-grade items and off-grade items ranging from 0-10. Selected off-grade items according to Learning Regression join the on-grade item pool, with two grades above and two grades below the designated grade in reading, and one grade above and one grade below the designated grade in mathematics. Those off-grade items could be selected during the process of adaptive testing. In this study, all analyses focused on on-grade items only. Target test specifications with constraints are presented in Tables 2 and 3, respectively, for each test.

Student performance is reported on the vertical scale, approximately from 300-1200, across grades 2-10, as well as in four performance levels: *Well Below Standard*, *Below Standard*, *Meets Standard* and *Advanced*. Descriptive statistics of student performance are summarized in Tables 4 and 5, respectively, for each test and technical quality, including reliability and standard error of measurement, are presented in Table 6.

Data Source

Two data files were generated from three administrations in the 2010-2011 school year—fall, winter and spring—at grades 3, 8 and 10. Students who received a valid score on each measure were included. Two corresponding item pools were used to analyze each test. Note that two items pools were used for grade 10 mathematics: the original pool was used for 2010 fall and 2011 winter and the extensive pool was for 2011 spring. Summaries of the operational item pools can be found in Tables 7-9.

Methods and Process for Data Analysis

(1) Review of Item Pools

Each operational item pool was reviewed for size and content balance with respect to number and percentage of items by constraints for item selection. The item pool characteristics were compared with the corresponding test specifications for evaluation. In reading, the combination of two measured standards was reviewed by item as well as within passage.

(2) Item Exposure Rate

Item exposure rate is determined by the ratio of number of item responses to the total number of students participating by test, grade and administration. An item with an exposure rate of 5% or less across three administrations is defined as under-exposed; an item with an exposure rate of 90% or more across administrations is defined as over-exposed. Items that were never used were identified as under-exposed.

(3) Influence of Item Pool on Student Growth

Three comparisons were performed by using operational data to evaluate the influence of the item pool characteristics on student growth measures.

a. Gap analysis was conducted for each operational item pool compared with the corresponding test specifications for content balance.
b. Since item parameter (b) and person parameter (θ) are on the same scale in Rasch model, to what extent the item difficulties in the pool match student abilities is found through a mapping analysis.
c. An item exposure analysis was conducted. Identified over- and under-exposed items/passages are then compared with content-related constraints indicated in test specifications.
d. The conditional standard error of measurement (CSEM) was compared in magnitude and distributions, along with the reporting scale across test administrations for each test. The simulation was generated using SIMULATION-2.0.2.e and SIMCATAlgorithm-2.0.2 (Pearson, 2012). The simulation was based on the known item parameters and student responses from the 2011 spring administration for grades

3, 8 and 10 in mathematics, using multiple-choice (MC) items only. Each simulation condition is with two alterations: (a) fix b-parameters and manipulate *θs* in the unit of 0.50 logit; (b) fix *θs* and manipulate b-parameter by 0.50. Simulation results are evaluated using two general criteria:

- Measurement precision (average conditional standard error of measurement)
- Measurement bias (average bias and mean square error)

$$\textbf{Bias} = \frac{1}{m}\sum_{m=1}^{m}(\hat{\theta}_m - \theta_m)$$

$$\textbf{MSE} = \frac{1}{m}\sum_{m=1}^{m}(\hat{\theta}_m - \theta_m)^2$$

Where

$\hat{\theta}_m$ is the m^{th} examinee's estimated theta

θ_m is the m^{th} examinee's true theta

- Item pool utilization (item exposure rate and under-utilization).

Results of Analyses

Test Specifications and Descriptive Statistics

In the 2010-2011 school year, the Delaware Comprehensive Assessment System (DCAS) reading and mathematics were administered in the fall, winter, and spring for all public-school students. Each test window lasted about 2-3 months (Table 1).

The target test specifications for mathematics (Table 2) indicate that four standards are measured: Standard 1: Numeric Reasoning (Number Sense and Operations); Standard 2: Algebraic Reasoning (Patterns and Change; Representation; Symbols); Standard 3: Geometric Reasoning (Classifications, Location and Transformation; Measurement); and Standard 4: Quantitative Reasoning (Collect, Present and Analyze; Probability) with varying weights from grade to grade. Machine-scored constructed-response questions are set to two \items per student across grades. The complexity of items, however, was set as low priority in selecting items. The two standards measured in reading are Standard 2: Construct, examine, and extend the meaning of literary, informative and technical texts through listening, reading, and viewing; and Standard 4: Use literary knowledge accessed through print and visual media to connect the self to society and culture. The target test specifications for reading (Table 3) show the general constraints for passage type, content category in standard and MSCR questions. Among the three constraints, content category at item-level is the priority for item selection.

Descriptive statistics are summarized in Tables 4 and 5 for mathematics and reading, respectively, by grade and test window, based on operational data. Reliability and standard error of measurement are presented by test, grade and test administration (Table 6).

Gap Analysis for Item Pool

Gap analysis is a crucial element in item pool evaluation for computerized adaptive testing**.** In this study, the analyses focused on content balance compared with the corresponding test specifications and the quality of item pool statistical characteristics by inspecting to what extent item difficulties match estimated student abilities. Tables 7a-b show descriptive statistics for mathematics item pool by grade, standard and item type. Tables 9a-b for reading by grade, standard and item type at the item and passage levels.

In mathematics, a significant discrepancy was observed in grades 3 and 10 between test specifications and item pool (Table 8). For instance, grade 3 has a 15% discrepancy between the item pool (45%) and test specifications (60%) in Numeric Reasoning, while Quantitative Reasoning has a discrepancy of 9% more items in the pool (13%), versus 4% actually needed. Similarly, in grade 10, the operational pool for the fall and winter administrations contains 17% more items than expected in the test specifications (40%) for Algebraic Reasoning (57%) and 19% fewer items for Geometric Reasoning (25% vs. 44%). The test specifications set the minimum and the maximum two MSCR questions for all grades, but did not specify either the content or the complexity constraints.

In reading, a set of items is attached to a common stimulus or a reading passage. For DCAS, passages are categorized as informational or literary and items are coded with Standards 2 and 4. Note that literary passages could include both literary and informational items, whereas informational passages can only have informational items. Table 9a provides a brief description of item pools by grade. Although the test specifications denote both passage-level and item-level constraints, the algorithm actually selects passages and items based on content standard rather than passage type. The comparison between the test specifications and the associated item pool is summarized in Table 10. At the passage-level, there are 3-6% more literary passages, but in short informational passages in the pool of grades 3 and 10, whereas there are more informational passages in grade 8. The results indicate a significant disparity between test specifications and availability in the pool for all three grades. Although the content constraint was set in a wide range of 20% by standard in the test specifications, there is obvious shortage of items measuring standard 4 in the pool. In other words, there are many informational passages in the pool that measure standard 2 only. Item type constraint was set loosely (0-1) since MSCR questions are very limited in the pool (4%) and across passages.

To examine the extent to which item difficulties in the pool match student abilities, mapping analysis was conducted. Since item parameter (b_i) and person parameter (θ_j) are calibrated on the same scale under Rasch model, the values of item difficulty and estimated theta are divided into sub-intervals for appropriate distributions. Figures 1-3 are for mathematics and Figures 4-6 are for reading by grade and test administration. Red represents b-parameter and blue represents person parameter.

In grade 3 mathematics, item difficulties seem to match student abilities with slight disparity at the high end of the scale in the first measure of fall (Figures 1a-1c). The disparity between b-parameters and estimated theta becomes significant in the second (winter administration) and the third (spring administration) measures. In the grade 8 pool, item parameters are heavily distributed in the middle without adequate hard and easy items to fit the needs for high- and low-achieving students (Figures 2a-2c). For the first test administration in the fall, item parameters seem to be slightly higher than most estimated theta. When the distribution of persons' parameters is shifted to the right in the following two test administrations, the fact that the pool lacks hard items is evident. A similar pattern is observed in grade 10 (Figures 3a-3c). The enlarged item pool for the spring administration improved the content balance for Standards 3 (57 new items) and 4 (19 new items) but did not help improve the distribution of item parameters along with the scale. The average item difficulty from the first item pool to the second pool remains the same (from 2.14 to 2.15) for Standard 3 and stays very similar (from 1.95 to 1.91) for Standard 4 (Table 7a).

The mappings for reading between item parameter and person parameter show a similar pattern as those for mathematics across grades (Figures 4-6). As the majority of item difficulties are overly distributed in the middle, the item pool is apparently lacking hard and easy items for the needs of high- and low-achieving students. For the winter and spring test administrations, the discrepancy becomes significant between estimated students' abilities and the item difficulties in the corresponding item pool.

Item Exposure

Item exposure is an important issue in computerized adaptive testing for its efficiency and security. The appropriate item exposure rate is usually determined depending upon the purpose of the test and the availability of the item pool. In this study, a cumulated exposure rate was calculated across three repeated measures. The exposure rate of 90% or above for an item or a passage is considered over-exposed, while 5% or below is considered as under-exposed.

The percentage of under-exposed and over-exposed items is presented in Table 11 by standard for mathematics. It is not surprising to notice that the size and content balance of the items pool strongly impact item exposure rate. The deficiency of items in grade 3 (-15% for Standard 1) and in grade 10 (-19% for Standard 3; -3% for Standard 1) is certainly one reason to cause a higher percentage of over-exposed items (28% for grade 3; 84% and 43% for grade 10). On the other hand, over-supplied items in the pool—for instance in grade 3 (+9% for Standard 4) and grade 10 (+20% for Standard 2) —seem to be the source for under-exposed items (43% for grade 3; 8% for grade 10).

The analysis results for reading summarized in Table 12 provide additional evidence about the relationship between item pool structure and item exposure rate. Although the test specifications denote the target expectations for standards 2 and 4 with a great flexibility (5-35%) for the reading test, the imbalanced item pool is certainly responsible for the under- and over-exposure rates. The excess number of items measuring Standard 2 yield a higher percentage of under-exposed items (41% for grade 3; 43% for grade 8, 17% for grade 10) than for Standard 4, whereas the insufficient number of items measuring Standard 4 create a higher percentage of over-exposed items (34% for grade 3; 39% for grade 8; 45% for grade 10) than Standard 2. Grade 10 demonstrates a smaller difference of item exposure rate between the two standards, which is perhaps due to the smaller discrepancy between the test specifications and the item pool. However, the smallest item pool among the three grades yields the highest percentage of over-exposed items (45%).

The plots of the conditional standard error of measurement (CSEM) in the unit of reporting scores vs. scale scores are displayed for mathematics (Figures 7-9) and for reading (Figures 10-12). These plots reveal a significantly larger SEM for the high-achieving students in both tests across grades and test administrations, indicating that the item pool is considerably lacking of hard items. The range of the minimum CSEM is 18-19 for mathematics and 20-21 for reading; the range of the maximum of CSEM is 65-92 for mathematics and 72-102 for reading. In grade 3, the average SEM increases from 20 to 28 in mathematics and from 25 to 28 from the fall to the spring administration.

Simulation Study and Results

To further this exploration, a simulation study was designed and performed. Using the item parameters and student response data from the 2011 spring test administration, simulations were executed for grades 3, 8, and 10 in mathematics. The 3x3x3x1 Crossed Factorial Design described in Chart 1 and associated tables indicate that two alternative manipulations were used: (a) fix b-parameters and manipulate *θs* in the unit of 0.50 logit (plus or minus 0.5 logit); (b) fix *θs* and manipulate b-parameter by 0.50 (plus or minus 0.5 logt). Thus, there are a total of nine combined simulations per test per grade with 50,000 simulated examinees per condition. The simulation results are generated in terms of the percentage of examinees whose test forms match the test specifications, measurement precision (average conditional standard error of measurement), measurement bias (average bias and mean square error) and item pool utilization (item exposure rate and item pool usage). Simulation results can be found in Tables 13-15 by grade.

Figure 13 is a plot of the average value of bias ($\hat{\theta} - \theta$), the discrepancy between estimated theta and true theta by grade and simulation condition. The smallest bias can be achieved if the average item parameter increased by 0.50, but the average person parameter decreased by 0.50 in grade 3 (-.0002), if the average item parameter increased by 0.50 for the current average person parameter in grades 8 (-.0012) and 10 (-.0003). The negative value of bias suggest that the current item pool does not match the need of students across grades (bias = -.0076, -.0214, and -.0100 for grades 3, 8, and 10) with under-estimated thetas. Similarly, the mean-square error (MSE) and conditional standard error of measurement (CSEM) show a consistent pattern as bias in grades 3 and 8, with trivial difference in grade 10. The item exposure rate is generally high for all grades due to the small item pool, particularly for easy and hard items due to the shortage of items in these areas. It is clear that item exposure rate always follows the changes of person parameters.

Discussion

The current study is a primary investigation of the influence of item pools' characteristics on student growth measure using computerized adaptive tests in K-12 education. First, the results of gap analysis reveal some major issues. The item pool size, 152-225 items for mathematics (Table 7a) and 30-35 passages with 189-258 attached items for reading (Table 9a) across grades, is far from sufficient to support multiple measures in adaptive testing, especially with the constraint of not seeing the same item over test administrations. The discrepancy between target test specifications and available items by test (Tables 8 and 10) indicates the content imbalanced structure of the item pool. The results of mapping visually illustrate the disparity of person parameters (θ_j) from item parameters (b_i) along with the scale in both mathematics (Figures 1-3) and reading (Figures 4-6). As the majority of item difficulties are overly distributed in the middle, the item pool is lacking items in the extreme ends, particularly hard items to match the needs of high-achieving students in adaptive testing. This issue became more obvious when student performance improved from the fall administration to the winter and spring test administrations. The deficit and excess of items for certain assessed standard(s) consequently created the issue of under- and over-exposure of test items in the pool (Tables 11 and 12). In reading, for example, a larger percentage of items that measure Standard 2 are under-exposed due to the excess of items in the pool, whereas a larger percentage of items that measure Standard 4 are over-exposed because of the lack of such items in the pool.

Using the criteria to evaluate measurement precision and measurement bias, the simulation derived consistent results as those from empirical analyses (Figures 13-15; Tables 13-15). The data indicate that for a more accurate measure of student achievement with minimum bias, the difficulty level of the item pool must match student ability level. Otherwise, the disparity between the two, as in the current case, would introduce bias or measurement error for student growth.

Computerized adaptive testing (CAT) has received considerable attention in recent years for K-12 assessments because of its attractive features that may improve learning particularly for low-achieving students. The procedures of adaptive testing, however, do not automatically function as expected with all advantages. Many operational issues and problems that have occurred and been attributed to CAT are the result of it having been used under inappropriate circumstances (Davey & Nering, 2002). Among the many particular requirements for CAT, a sizeable and well- balanced item pool with regard to content and psychometrics characteristics is a fundamental condition for success. For K-12 assessment programs, a large population, wide range of proficiency level, broader content coverage and high-stakes nature introduces additional technical challenges in the development and implementation of CAT. As indicated earlier, the current study is a primary investigation. Factors such as the process of item selection, testlet effect in reading and usage of machine-scores constructed-response items, should be taken into consideration for future investigations.

References

Ballou, D. (2008, April). *Test scaling and value-added measurement*. Paper presented at the National Conference on Value-Added Modeling, Madison, WI.

Betebenner, D. W. & Linn, R. L. (2010). Growth in student achievement: Issues of measurement, longitudinal data analysis, and accountability. Exploratory Seminar: *Measurement Challenges within the Race to the Top Agenda*. Center for K-12 Assessment & Performance Management.

Iramaneerat, C. & Stahl, J. (2007). Optimizing item pool characteristics to control item exposure in a computerized adaptive test. Paper presented at the AREA Annual Meeting, Chicago, IL, 2007.

Davey, T. & Pitoniak, M. J. (2006). Designing Computerized Adaptive Tests in S. W. Downing & T. M. Haladyna (Eds). *Handbook of Test Development*. Lawrence Erlbaum Associates.

Leung, C.K., Chang, H. H. & Hau, K. T. (2003). Computerized adaptive testing: A comparison of three content balancing methods. *The Journal of Technology, Learning, and Assessment* 2 (5): December 2003.

Li, X., Becker, K., Gorham, J. & Woo, A. (2009). Limiting item exposure for target difficulty ranges in a high-stakes CAT. Paper presented at the 2009 GMAC Conference on Computerized Adaptive Testing.

Parshall, C. G., Spray, J. A., Kalohn, J. C. & Davey, T. (2002). *Practical Considerations in Computer-Based Testing*. Springer-Verlag New York, Inc.

Reckase, M. D. (2007). The design of p-Optimal item pools for computerized adaptive tests. Keynote Address presented at the 2007 GMAC® Conference on Computerized Adaptive Testing. June 7, 2007.

Reckase, M. D. (2010). Designing item pools to optimize the functioning of a computerized adaptive test. *Psychological Test and Assessment Modeling*, Volume 52, 2010 (2), 127-141.

Shin, C. D., Chien, Y., Way, W. D. & Swanson, L. (2009). Weighted penalty model for content balancing in CATS. Pearson, April, 2009.

Technical Report for the Delaware Comprehensive Assessment System (DCAS). Delaware Department of Education. 2012.

Veldkamp, B. P. & van der Linden, W. J. (2000). Designing item pools for computerized adaptive testing. In W. J. van der Linden & C. A. Glas (Eds.) *Computerized Adaptive Testing: Theory and Practice*. Dordrecht, The Netherlands: Kluwer.

Yen, W. M. (1986). The choice of scale for educational measurement: An IRT perspective. *Journal of Educational Measurement* 23 (4): 299–325.

Table 1. 2010-2011 Delaware Assessment Calendar

Measure	Start Day	End Day	Duration in Weeks
2010 Fall	October 11, 2010	December 14, 2010	Over 9 weeks
2011 Winter	January 5, 2011	April 11, 2011	Over 13 weeks
2011 Spring	April 18, 2011	June 3, 2011	7 weeks

Table 2. Target Test Specifications for Mathematics

Content Standard	Item Complexity						MSCR Items	Total	
	Low		Moderate		High				
	N	%	N	%	N	%		N.	%
Grade 3									
Mathematics-3.1	7	14	15	30	8	16		30	60
Mathematics-3.2	2	4	4	8	2	4		8	16
Mathematics-3.3	3	6	5	10	2	4		10	20
Mathematics-3.4	0	0	1	2	1	2		2	4
Total	12	24	25	50	13	26	2 to 2	50	100
Grade 8									
Mathematics-8.1	3	6	6	12	3	6		12	24
Mathematics-8.2	7	14	12	24	6	12		25	50
Mathematics-8.3	2	4	4	8	2	4		8	16
Mathematics-8.4	1	2	2	4	2	4		5	10
Total	13	26	24	48	13	26	2 to 2	50	100
Grade 10									
Mathematics-10.1	1	2	2	4	0	0		3	6
Mathematics-10.2	5	10	10	20	5	10		20	40
Mathematics-10.3	6	12	11	22	5	10		22	44
Mathematics-10.4	1	2	2	4	2	4		5	10
Total	13	26	25	50	12	24	2 to 2	50	100

Std. 1 - Numeric Reasoning measures Number Sense and Operations

Std. 2 - Algebraic Reasoning measures Patterns and Change; Representation; Symbol

Std. 3 - Geometric Reasoning measures Classifications, Location and Transformation; Measurement

Std. 4 - Quantitative Reasoning measures Collect, Present and Analyze; Probability

Table 3. Target Test Specifications for DCAS Reading

Grade	Standard	Passages			Items		N. MSCR Items	
		N.	%	Total	N.	%	Min	Max
3	Informational	4 to 5	50		20 - 30	40 - 60		
	Literary	4 to 5	50		20 - 30	40 - 60		
	Recall				15	30		
	Interpret				25	50		
	Evaluate				10	20		
	Total			8 to 10	50		0	1
8	Informational	5 to 6	60		23 - 33	46 - 66		
	Literary	3 to 4	40		17 - 27	34 - 54		
	Recall				10	20		
	Interpret				25	50		
	Evaluate				15	30		
	Total			8 to 10	50		0	1
10	Informational	6 to 7	70		30 - 40	60 - 80		
	Literary	3 to 3	30		10 - 20	20 - 40		
	Recall				10	20		
	Interpret				22	44		
	Evaluate				18	36		
	Total			8 to 10	50		0	1

Informational passages measure standard 2; literary passages measure both Standards 2 and 4.
The number of passages is estimated based on the target percentage by passage type.
The percentage is accurate. The number of passages is tentative.

Table 4. Descriptive Statistics of Student Performance in Mathematics

Test	Grade	Variable	N	Minimum	Maximum	Mean	SD	Skewness	Kurtosis
Fall	3	Scale Score	9681	371	916	617.9477	71.7184	.193	.177
		Theta	9681	-5.5006	2.9999	-1.6472	1.1196		
	8	Scale Score	9334	596	1137	790.8982	57.3305	.820	1.794
		Theta	9334	-1.9882	6.4541	1.0528	0.8951		
	10	Scale Score	9607	539	1142	823.3206	52.2088	.815	2.465
		Theta	9607	-2.8859	6.5400	1.5590	0.8150		
Winter	3	Scale Score	9696	369	959	656.2486	77.8900	.198	.353
		Theta	9696	-5.5410	3.6711	-1.0492	1.2160		
	8	Scale Score	9323	621	1124	807.4674	61.5297	.847	1.681
		Theta	9323	-1.5923	6.2573	1.3116	0.9606		
	10	Scale Score	9368	539	1171	836.3289	62.3281	.787	1.232
		Theta	9368	-2.8859	6.9897	1.7621	0.9731		
Spring	3	Scale Score	9664	440	930	695.5313	80.4653	.176	-.086
		Theta	9664	-4.4190	3.2194	-0.4359	1.2562		
	8	Scale Score	9273	486	1106	826.7953	69.0672	.620	.596
		Theta	9273	-3.7131	5.9660	1.6127	1.0797		
	10	Scale Score	9123	527	1187	848.7793	64.6278	.836	1.859
		Theta	9123	-3.0709	7.2393	1.9565	1.0090		

Table 5. Descriptive Statistics of Student Performance in Reading

Test	Grade	Variable	N	Minimum	Maximum	Mean	SD	Skewness	Kurtosis
Fall	3	Scale Score	9671	447	943	657.5921	75.7552	.329	-.284
		Theta	9671	-3.6227	3.3849	-0.6462	1.0694		
	8	Scale Score	9328	407	1051	788.6856	67.5962	-.013	.012
		Theta	9328	-4.1832	4.9055	1.2046	0.9537		
	10	Scale Score	9596	435	1113	818.5474	67.4613	.256	.245
		Theta	9596	-3.7853	5.7741	1.6251	0.9525		
Winter	3	Scale Score	9658	313	993	686.8792	78.2238	.146	-.164
		Theta	9658	-5.5152	4.0881	-0.2331	1.1038		
	8	Scale Score	9286	569	1088	804.4207	71.7661	.107	.079
		Theta	9286	-1.8911	5.4311	1.4255	1.0126		
	10	Scale Score	9350	435	1141	820.7068	63.3066	-.043	-.069
		Theta	9350	-3.7853	6.1792	1.6552	0.8930		
Spring	3	Scale Score	9631	315	997	710.5930	77.0937	.089	.196
		Theta	9631	-5.4811	4.1375	0.1015	1.0880		
	8	Scale Score	9236	412	1099	817.5035	74.1457	-.083	.509
		Theta	9236	-4.1183	5.5876	1.6102	1.0462		
	10	Scale Score	9113	440	1084	836.7948	65.8943	-.305	.172
		Theta	9113	-3.7155	5.3690	1.8823	0.9299		

Table 6. Reliability and Standard Error of Measurement by Grade and Test

Grade	Test	Mathematics				Reading			
		Reliability*	Standard Error of Measurement			Reliability	Standard Error of Measurement		
			Minimum	Maximum	Mean		Minimum	Maximum	Mean
	Fall	0.92	21.290	72.283	24.643	0.89	18.101	65.135	20.006
3	Winter	0.91	21.096	101.648	25.876	0.89	18.207	92.012	22.436
	Spring	0.86	20.413	101.768	27.680	0.86	18.364	91.852	27.825
	Fall	0.88	21.268	101.078	24.450	0.87	18.558	91.296	20.017
8	Winter	0.87	21.268	101.218	26.943	0.86	18.603	91.316	21.564
	Spring	0.88	21.059	101.149	26.915	0.87	18.737	91.255	22.719
	Fall	0.84	21.019	101.055	25.837	0.85	19.053	91.370	20.818
10	Winter	0.88	21.019	102.030	24.246	0.85	18.690	91.602	21.514
	Spring	0.89	20.671	100.914	25.613	0.84	18.565	91.604	20.869

Reliability is marginal reliability coefficient

Table 7a. Summary of Mathematics Item Pool by Grade and Standard

Grade	Standard	N	%	Range	Minimum	Maximum	Mean	SD	Skewness	Kurtosis
3	Mathematics-3.1	101	45	4.8727	-4.7216	0.1511	-1.9056	1.1253	-0.2960	-0.5238
	Mathematics-3.2	47	21	4.5265	-4.1993	0.3272	-1.8007	1.0591	-0.3875	-0.2399
	Mathematics-3.3	46	21	4.6924	-4.2270	0.4654	-1.6364	0.9580	-0.3180	0.2690
	Mathematics-3.4	30	13	5.4753	-5.1894	0.2860	-2.1845	1.3396	-0.4217	-0.0013
	Total	224								
8	Mathematics-8.1	54	26	2.7838	-0.0066	2.7771	1.4824	0.6681	-0.0563	-0.6884
	Mathematics-8.2	98	48	3.8670	-0.8489	3.0182	1.1275	0.7366	0.0670	0.1163
	Mathematics-8.3	33	16	2.1438	0.0948	2.2385	1.3339	0.6216	-0.4202	-0.8707
	Mathematics-8.4	21	10	3.8532	-1.2791	2.5741	1.2798	0.9887	-0.6721	0.5400
	Total	206								
10_Pool1	Mathematics-10.1	7	5	1.7745	1.0750	2.8495	1.6287	0.5734	1.9880	4.6477
	Mathematics-10.2	91	60	3.7685	0.0397	3.8082	1.8902	0.7346	0.0359	0.5342
	Mathematics-10.3	38	25	3.6606	0.3661	4.0267	2.1780	0.8004	0.0526	-0.0670
	Mathematics-10.4	16	11	3.3015	0.8442	4.1457	1.9795	1.0016	0.8105	0.1907
	Total	152								
10_Pool2	Mathematics-10.1	7	3	1.7745	1.0750	2.8495	1.6287	0.5734	1.9880	4.6477
	Mathematics-10.2	91	40	3.7685	0.0397	3.8082	1.8902	0.7346	0.0359	0.5342
	Mathematics-10.3	97	43	4.1990	-0.1722	4.0267	2.1539	0.7305	-0.1623	0.7389
	Mathematics-10.4	30	13	4.4457	-0.3000	4.1457	1.9080	1.0794	0.1086	-0.1757
	Total	225								

Table 7b. Summary of Mathematics Item Pool by Grade, Standard and Item Type

Grade	Item Type	Standard	N	Range	Minimum	Maximum	Mean	SD
3	GI	Mathematics-3.1	5	4.8727	-4.7216	0.1511	-2.2420	1.7704
		Mathematics-3.2	3	2.0435	-2.6283	-0.5849	-1.4781	1.0457
		Mathematics-3.3	1	0.0000	-1.7388	-1.7388	-1.7388	.
		Mathematics-3.4	3	1.5770	-3.1584	-1.5814	-2.5038	0.8219
	MC	Mathematics-3.1	96	4.6192	-4.4935	0.1257	-1.8881	1.0930
		Mathematics-3.2	44	4.5265	-4.1993	0.3272	-1.8227	1.0684
		Mathematics-3.3	45	4.6924	-4.2270	0.4654	-1.6341	0.9687
		Mathematics-3.4	27	5.4753	-5.1894	0.2860	-2.1490	1.3916
8	GI	Mathematics-8.1	2	1.1019	0.9017	2.0036	1.4526	0.7792
		Mathematics-8.2	3	0.6593	1.5300	2.1893	1.8296	0.3337
		Mathematics-8.3	5	1.4397	0.7989	2.2385	1.6048	0.5855
		Mathematics-8.4	3	0.2932	0.7836	1.0767	0.8868	0.1647
	MC	Mathematics-8.1	52	2.7838	-0.0066	2.7771	1.4835	0.6722
		Mathematics-8.2	95	3.8670	-0.8489	3.0182	1.1054	0.7357
		Mathematics-8.3	28	2.1424	0.0948	2.2371	1.2855	0.6255
		Mathematics-8.4	18	3.8532	-1.2791	2.5741	1.3453	1.0559
10_Pool1	GI	Mathematics-10.1	2	1.7745	1.0750	2.8495	1.9622	1.2548
		Mathematics-10.2	4	2.5645	0.6720	3.2365	2.1903	1.0939
		Mathematics-10.3	3	1.4877	1.9691	3.4568	2.8086	0.7621
		Mathematics-10.4	2	1.9755	2.1702	4.1457	3.1579	1.3969
	MC	Mathematics-	5	0.3746	1.3212	1.6958	1.4953	0.1471

		10.1						
		Mathematics-10.2	87	3.7685	0.0397	3.8082	1.8764	0.7202
		Mathematics-10.3	35	3.6606	0.3661	4.0267	2.1240	0.7905
		Mathematics-10.4	14	2.9260	0.8442	3.7702	1.8111	0.8736
10_Pool2	GI	Mathematics-10.1	2	1.7745	1.0750	2.8495	1.9622	1.2548
		Mathematics-10.2	4	2.5645	0.6720	3.2365	2.1903	1.0939
		Mathematics-10.3	3	1.4877	1.9691	3.4568	2.8086	0.7621
		Mathematics-10.4	2	1.9755	2.1702	4.1457	3.1579	1.3969
	MC	Mathematics-10.1	5	0.3746	1.3212	1.6958	1.4953	0.1471
		Mathematics-10.2	87	3.7685	0.0397	3.8082	1.8764	0.7202
		Mathematics-10.3	94	4.1990	-0.1722	4.0267	2.1330	0.7239
		Mathematics-10.4	28	4.0702	-0.3000	3.7702	1.8187	1.0272

Table 8. Comparison of Test Specifications and Item Pool for Mathematics

	Test Spec.		Item Pool		Discrepancy	Item Type	Test Spec.	Item Pool	Discrepancy
	N.	%	N	%	%				%
Grade 3									
Mathematics-3.1	30	60	101	45	-15	MSCR - N.	2	12	
Mathematics-3.2	8	16	47	21	5	MSCR - %	4	5	1
Mathematics-3.3	10	20	46	21	1	MC - N.	48	212	
Mathematics-3.4	2	4	30	13	9	MC - %	96	95	
Total	50		224			Total	50	224	
Grade 8									
Mathematics-8.1	12	24	54	26	2	MSCR - N.	2	13	
Mathematics-8.2	25	50	98	48	-2	MSCR - %	4	6	2
Mathematics-8.3	8	16	33	16	0	MC - N.	48	193	
Mathematics-8.4	5	10	21	10	0	MC - %	96	94	
Total	50		206			Total	50	206	
Grade 10_1									
Mathematics-10.1	3	6	7	5	-1	MSCR - N.	2	11	
Mathematics-10.2	20	40	91	60	20	MSCR - %	4	7	3
Mathematics-10.3	22	44	38	25	-19	MC - N.	48	141	
Mathematics-10.4	5	10	16	11	1	MC - %	96	93	
Total	50		152			Total	50	152	
Grade 10_2									
Mathematics-10.1	3	6	7	3	-3	MSCR - N.	2	11	
Mathematics-10.2	20	40	91	40	0	MSCR - %	4	5	1
Mathematics-10.3	22	44	97	43	-1	MC - N.	48	214	
Mathematics-10.4	5	10	30	13	3	MC - %	96	95	
Total	50		225			Total	50	225	

Table 9a. Summary of Reading Item Pool by Grade and Item Type

Grade	Standard	Item						
		N	%	Range	Minimum	Maximum	Mean	SD
3	Standard 2	191	74	4.4381	-2.9520	1.4860	-0.8794	0.7988
	MSCR	1			-0.9213	-0.9213	-0.9213	
	MC	190		4.4381	-2.9520	1.4860	-0.8792	0.8009
	Standard 4	67	26	4.3762	-2.8032	1.5731	-0.8076	0.9306
	MSCR	1			-0.4614	-0.4614	-0.4614	
	MC	66		4.3762	-2.8032	1.5731	-0.8128	0.9367
	Total	258						
8	Standard 2	204	81	3.6430	-0.8323	2.8108	0.7037	0.7764
	MSCR	1		0.0000	1.0891	1.0891	1.0891	
	MC	203		3.6430	-0.8323	2.8108	0.7018	0.7778
	Standard 4	49	19	3.3518	-1.0299	2.3219	0.7386	0.7970
	MSCR	0						
	MC	49		3.3518	-1.0299	2.3219	0.7386	0.7970
	Total	253						
10	Standard 2	160	85	3.0226	-0.4342	2.5884	1.1157	0.6910
	MSCR	5		1.0192	0.9509	1.9702	1.4675	0.4445
	MC	155		3.0226	-0.4342	2.5884	1.1044	0.6955
	Standard 4	29	15	3.0449	-0.4907	2.5542	0.9150	0.6800
	MSCR	1			1.4853	1.4853	1.4853	
	MC	28		3.0449	-0.4907	2.5542	0.8946	0.6834
	Total	189						

Table 9b. Summary of Reading Item Pool by Passage

Grade	Passage Type	Standard	Reading Passage		
			N	Average Attached Items	Range
	Informational	2 only	15		
3	Literary	2 and 4	19		
	Total		34	7.6	5 - 11
	Informational	2 only	22		
8	Literary	2 and 4	13		
	Total		35	7.2	5 - 11
	Informational	2 only	20		
10	Literary	2 and 4	10		
	Total		30	6.3	5 - 9

Table 10. Comparison between Test Specifications and Item Pool for Reading

Grade	Passage	Test Specs.		Item Pool		Discrepancy	Standard	Test Specs.		Item Pool		Discrepancy
		N.	%	N.	%	%		N.	%	N.	%	%
3	Informational	4 to 5	50	15	44	6	2	20 - 30	40 - 60	191	74	14 - 34
	Literary	4 to 5	50	19	56		4	20 - 30	40 - 60	67	26	
	Total	8 to 10		34			Total	50		258		
8	Informational	5 to 6	60	22	63	3	2	23 - 33	46 - 66	204	81	15 - 35
	Literary	3 to 4	40	13	37		4	17 - 27	34 - 54	49	19	
	Total	8 to 10		35			Total	50		253		
10	Informational	6 to 7	70	20	67	3	2	30 - 40	60 - 80	160	85	5 - 25
	Literary	3 to 3	30	10	33		4	10 - 20	20 - 40	29	15	
	Total	8 to 10		30			Total	50		189		

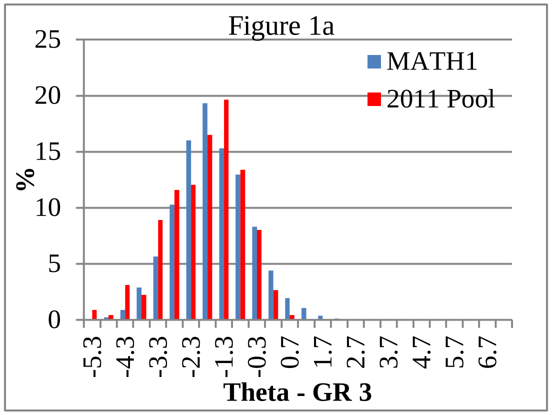

Figure 1a

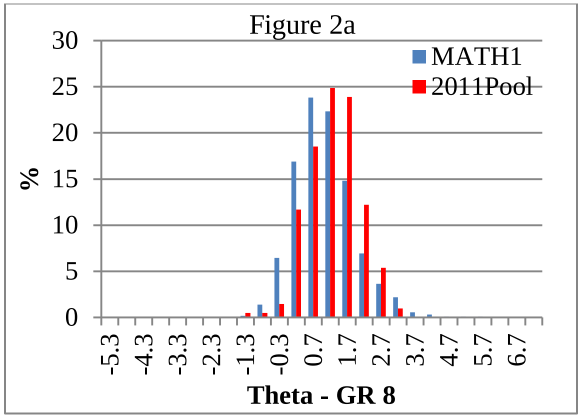

Figure 2a

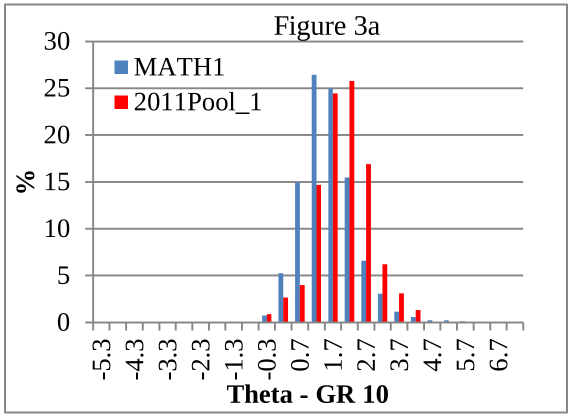

Figure 3a

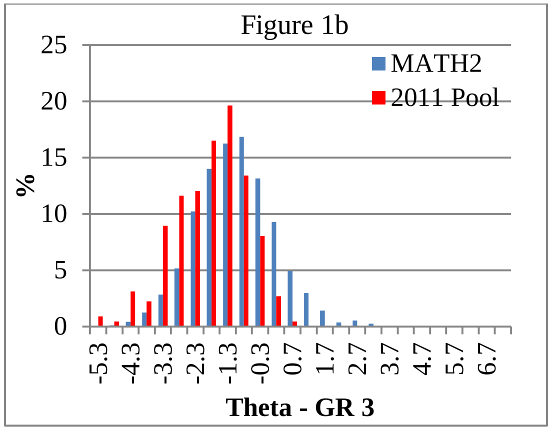

Figure 1b

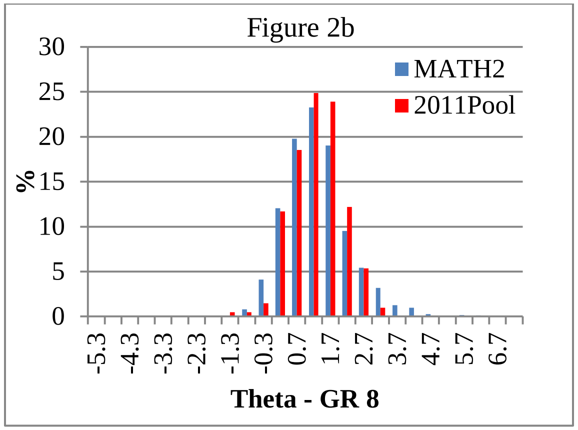

Figure 2b

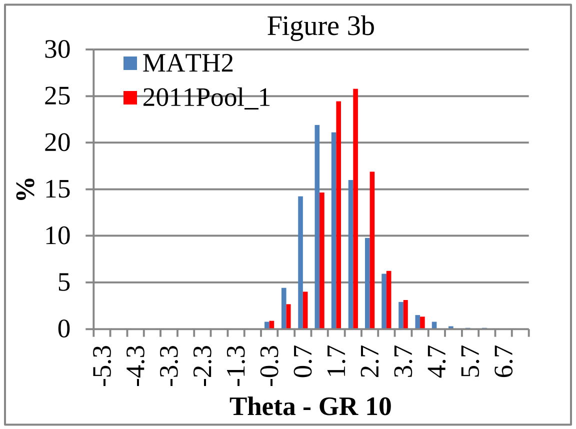

Figure 3b

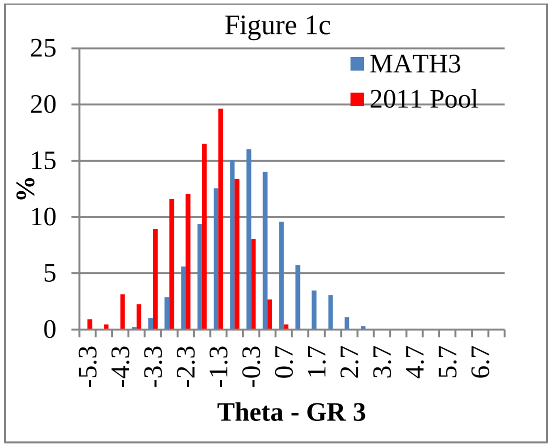

Figure 1c

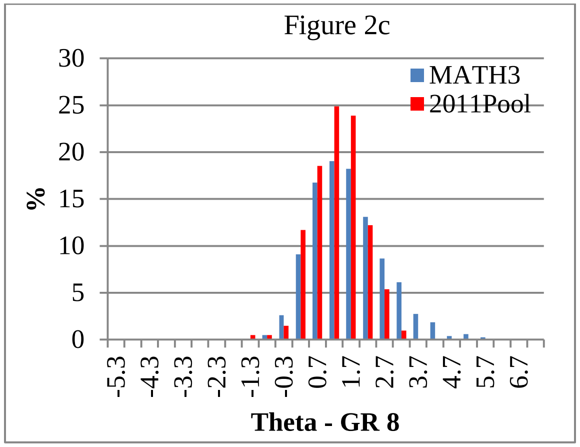

Figure 2c

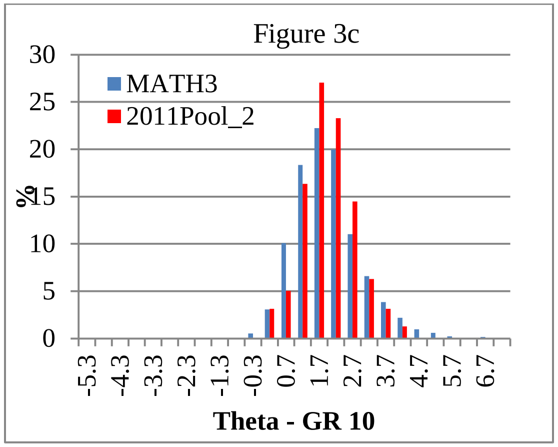

Figure 3c

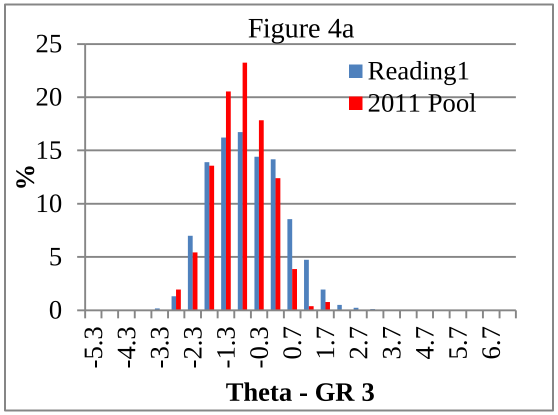
Figure 4a

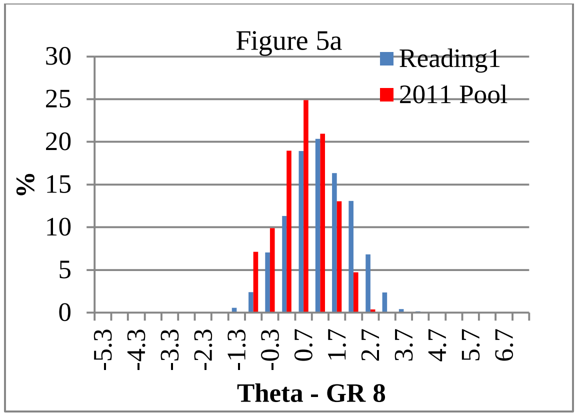
Figure 5a

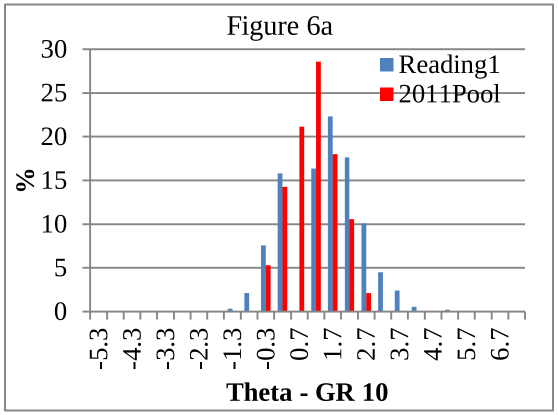
Figure 6a

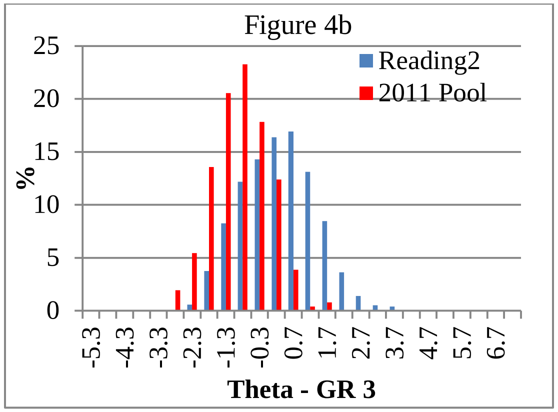
Figure 4b

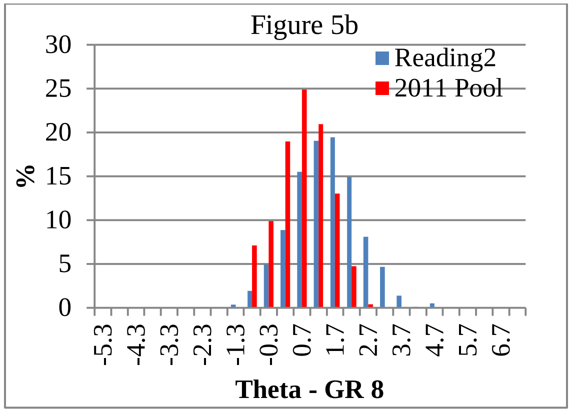
Figure 5b

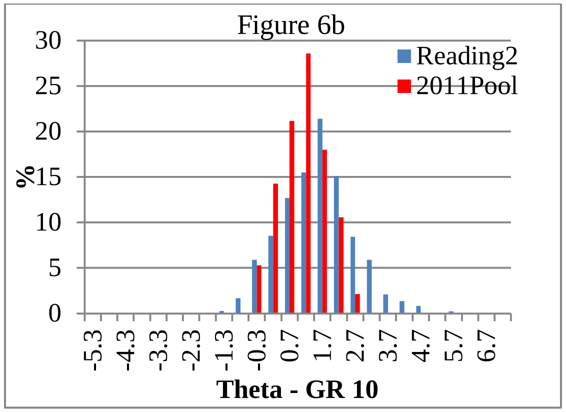
Figure 6b

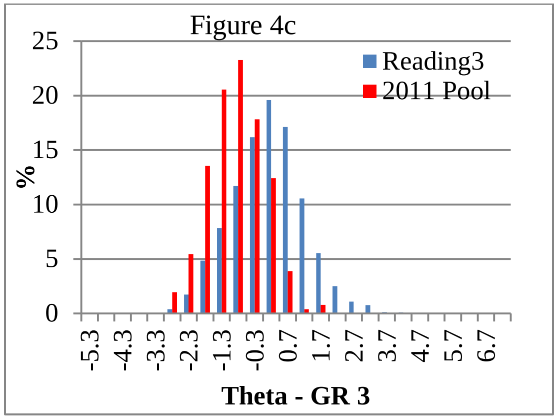
Figure 4c

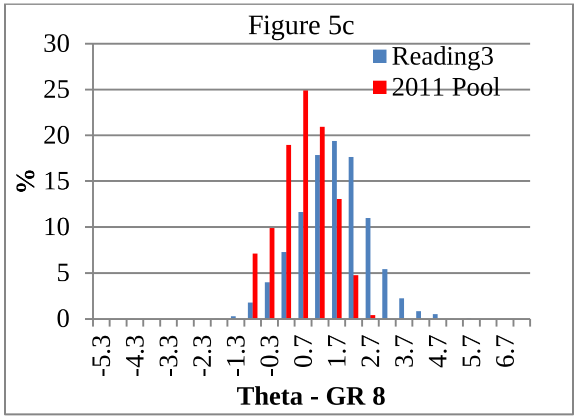
Figure 5c

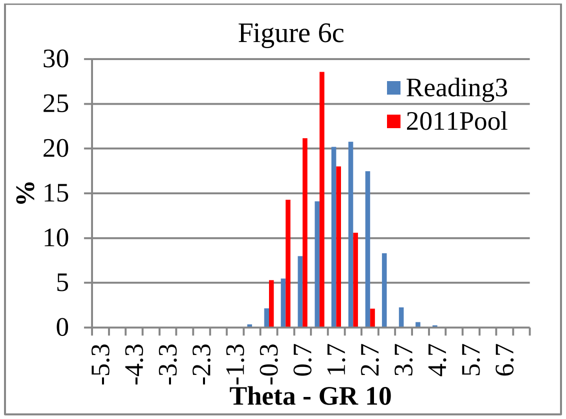
Figure 6c

Table 11. Comparison of Item Pool and Item Exposure Rate for Mathematics

Standard	Test Spec.		Item Pool		Discrepancy	Percent of Items	
	N.	%	N	%	%	Under-Exposed	Over-Exposed
Grade 3							
Mathematics-3.1	30	60	101	45	-15	1	28
Mathematics-3.2	8	16	47	21	5	9	6
Mathematics-3.3	10	20	46	21	1	4	4
Mathematics-3.4	2	4	30	13	9	43	0
Total	50		224				
Grade 8							
Mathematics-8.1	12	24	54	26	2	9	6
Mathematics-8.2	25	50	98	48	-2	1	18
Mathematics-8.3	8	16	33	16	0	6	15
Mathematics-8.4	5	10	21	10	0	0	67
Total	50		206				
Grade 10_Pool 1[1]							
Mathematics-10.1	3	6	7	5	-1	0	43
Mathematics-10.2	20	40	91	60	20	8	10
Mathematics-10.3	22	44	38	25	-19	0	84
Mathematics-10.4	5	10	16	11	1	0	50
Total	50		152				
Grade 10_ Pool 2[2]							
Mathematics-10.1	3	6	7	3	-3	0	43
Mathematics-10.2	20	40	91	40	0	2	10
Mathematics-10.3	22	44	97	43	-1	15	30
Mathematics-10.4	5	10	30	13	3	7	30
Total	50		225				

1. The exposure rate for Grade 10_Pool 1 is a summary for the fall and winter administrations based on Pool 1.

2. The exposure rate for Grade 10_Pool 2 is a summary for all three test administrations based on Pool 2.

Table 12. Comparison between Test Specifications and Item Pool for Reading

Grade	Standard	Test Specs.		Item Pool		Dis. %	% Exposed Items		Passage	Pool	Exposed (N)	
		N.	%	N.	%		Under	Over		N.	Under	Over
3	2	20 - 30	40 - 60	191	74	±14 - 34	41	25	Informational	15	5	5
	4	20 - 30	40 - 60	67	26		25	34	Literary	19	7	5
	Total	50		258					Total	34	12	10
8	2	23 - 33	46 - 66	204	81	±15 - 35	43	26	Informational	22	9	6
	4	17 - 27	34 - 54	49	19		16	39	Literary	13	4	4
	Total	50		253					Total	35	13	10
10	2	30 - 40	60 - 80	160	85	±5 - 25	17	40	Informational	20	4	8
	4	10 - 20	20 - 40	29	15		0	45	Literary	10	0	5
	Total	50		189					Total	30	4	13

Dis. - Discrepancy

Figures 7. GR 3 Math

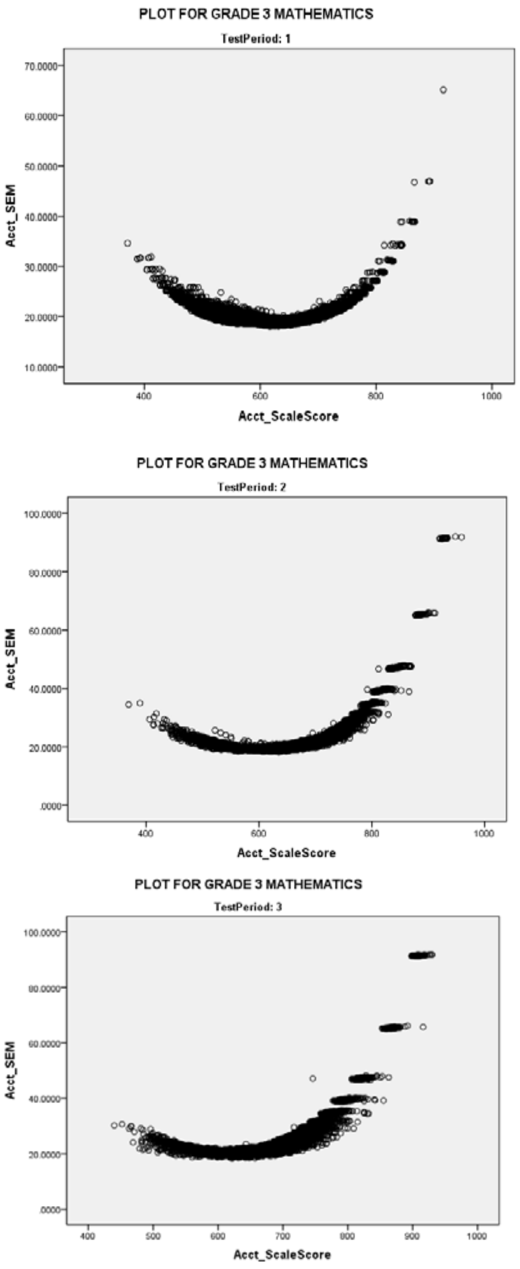

Figure 8. GR 8 Math

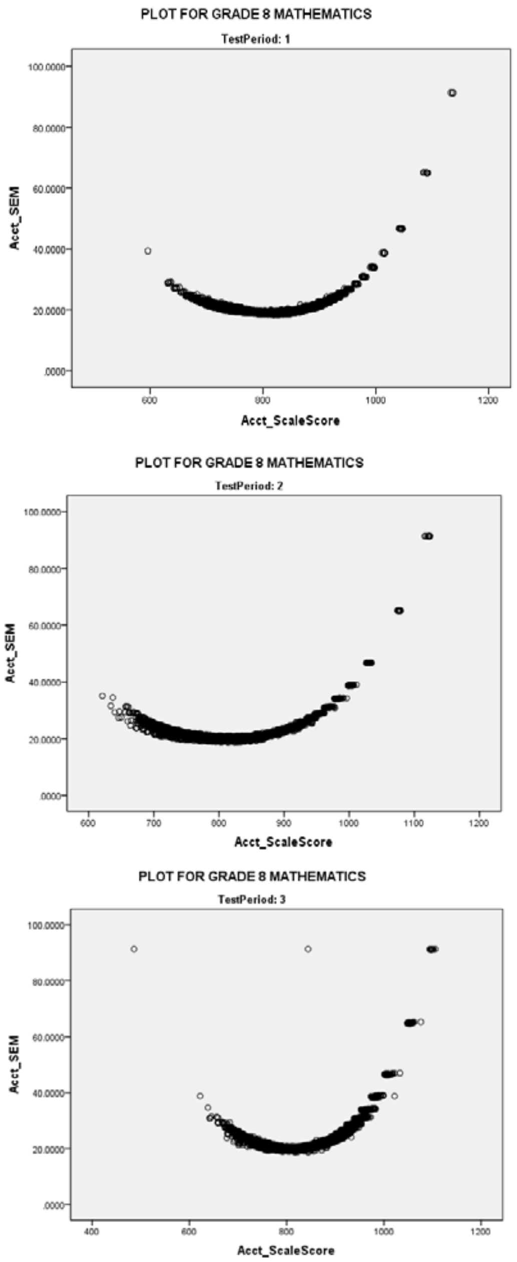

Figure 9. GR 10 Math

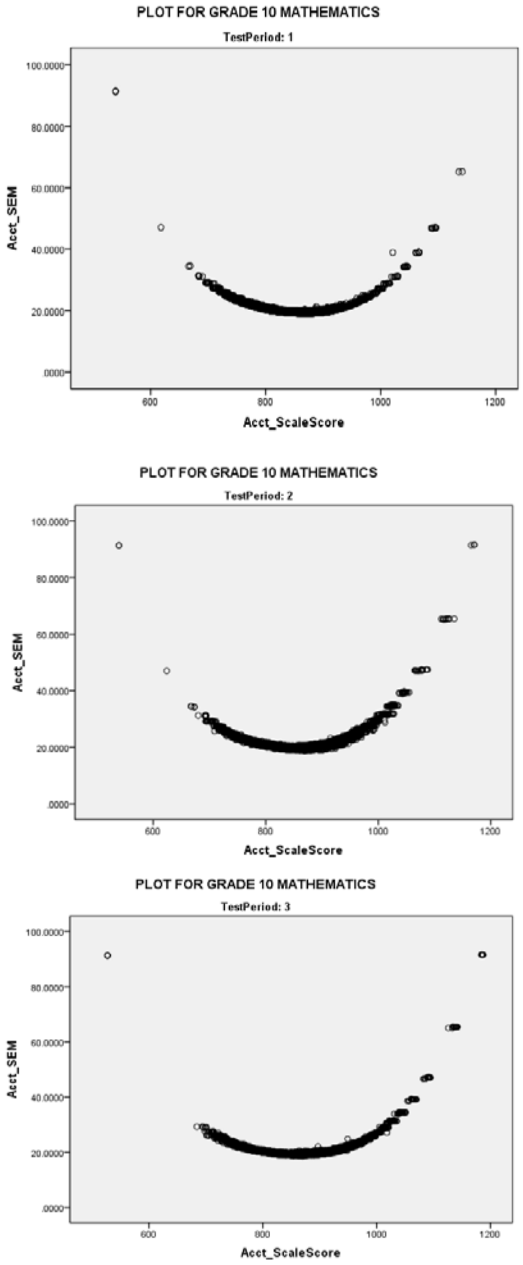

Figure 10. GR 3 Reading

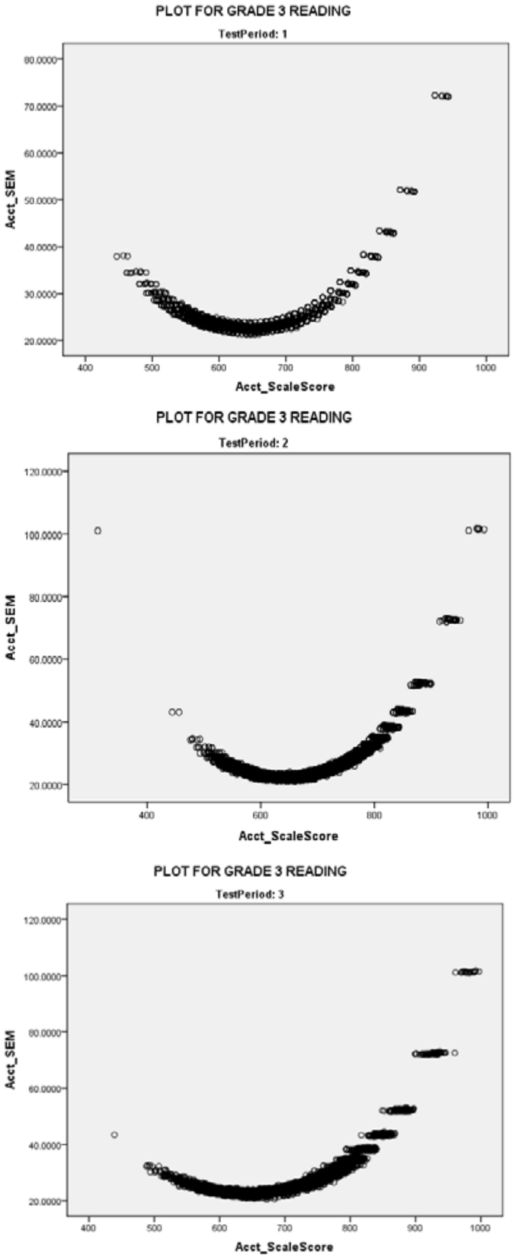

Figure 11. GR 8 Reading

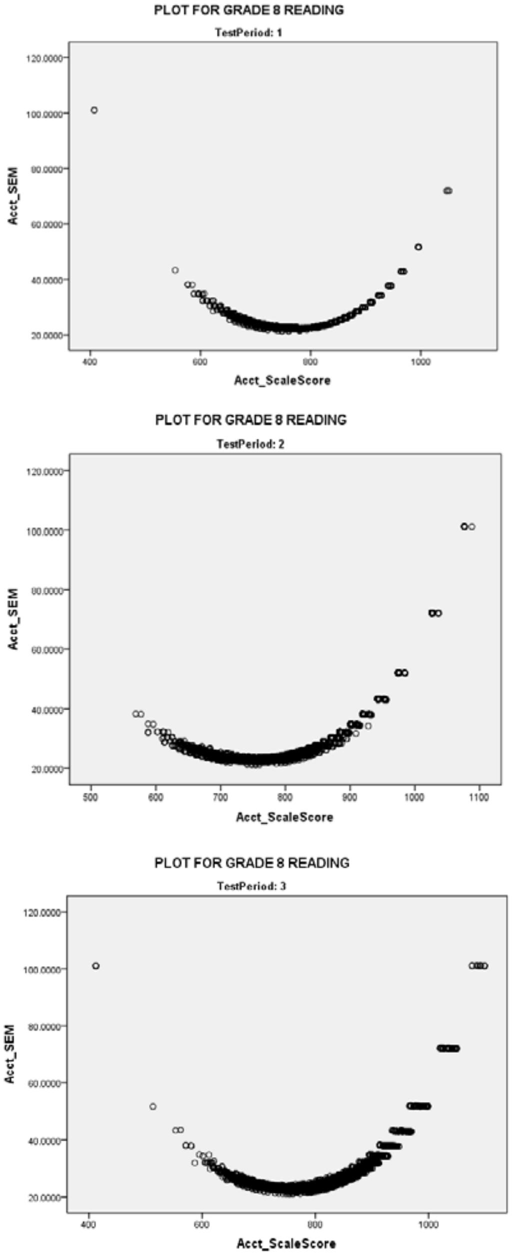

Figure 12. GR 10 Reading

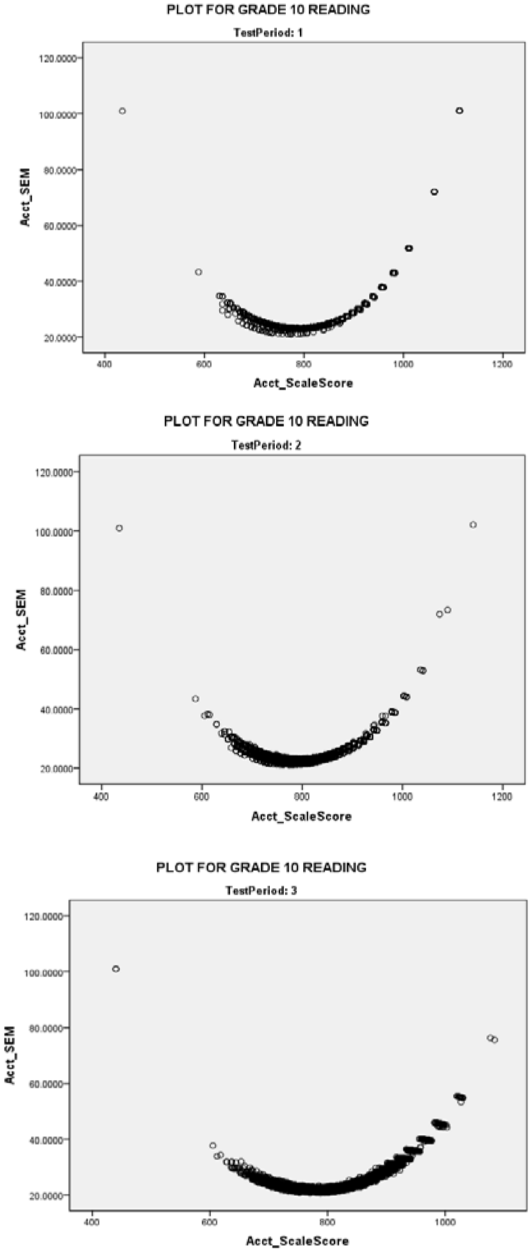

Chart 1. Simulation Design and Conditions

Data Source and Description		Simulation Condition and Index
Item Parameter	Mean, Minimum, Maximum	See Table A below
Person Parameter	Mean, Standard Deviation, Minimum, Maximum	See the Table B below
Test Items	Dichotomous Items	
Test Length	40 - 50	
Design	Independent Variables	
	Grade	3, 8, 10
	Item Pool Characteristics	H, M, L (M is the empirical mean, H=M+0.5, L=M-0.5)
	Examinees' Characteristics (theta distributions)	H, M, L (M is the empirical mean, H=M+0.5, L=M-0.5)
	Content Balance Method	CCAT*
	3x3x3x1 Crossed Factorial Design	
	50,000 Examinees Per Condition	
	Dependent Variables	
	% of Meeting the Test Specifications	
	Measurement Precision	Conditional Standard Error of Measurement
	Measurement Bias	Average Bias and Mean Square Error
	Item Pool Utilization	Item Exposure Rate

*CCAT – Constraint computerized adaptive testing.

IL_PL - Both item and person parameters are 0.5 lower than the average (M).

IL_PM - Item parameter is 0.5 lower than the average, while person parameter is at the average.

IL_PH - Item parameter is 0.5 lower, but person parameter is 0.5 higher than the average.

IM_PL - Item parameter is at the average, while the person parameter is 0.5 lower than the average.

IM_PM - Both item and person parameters are at the average.

IM_PH -Item parameter is at the average, but person parameter is 0.5 higher than the average.

IH_PL - Item parameter is 0.5 higher, but person parameter is 0.5 lower than the average.

IH_PM - Item parameter is 0.5 higher than the average, but person parameter is at the average.

IH_PH - Both item and person parameters are 0.5 higher than the average.

Chart 1 - Table A. Descriptive Statistics of Empirical Data

Item Parameter	Grade 3				Grade 8				Grade 10			
	Mean	Min.	Max	SD	Mean	Min.	Max	SD	Mean	Min.	Max	SD
Standard 1	-1.6974	-5.4591	0.2308		1.4128	-0.0066	2.9782		1.4253	1.0750	1.6958	
Standard 2	-1.4956	-4.1993	0.3272		1.2843	-0.2312	3.0182		1.5461	0.0397	3.8082	
Standard 3	-1.2209	-2.4933	0.4654		1.2872	0.0948	2.2428		2.1330	-0.1722	4.0267	
Standard 4	-1.9017	-5.1482	0.2860		0.8928	-1.0856	2.5163		1.7102	-0.3000	4.1457	
Total	-1.5885	-5.4591	0.4654		1.2883	-1.0856	3.0182		1.8390	-0.3000	4.1457	
Person Parameter	-0.4359	-4.4190	3.2194	1.2562	1.6112	-3.7131	5.9660	1.0803	1.9564	-3.0709	7.2393	1.0090

Chart 1 - Table B. Statistics by Simulation Conditions

Item	**Grade**			Person	**Grade**		
	3	8	10		3	8	10
I_L	-2.0885	0.7883	1.3390	P_L	-0.9359	1.1112	1.4564
I_L	-2.0885	0.7883	1.3390	P_M	-0.4359	1.6112	1.9564
I_L	-2.0885	0.7883	1.3390	P_H	0.0641	2.1112	2.4564
I_M	-1.5885	1.2883	1.8390	P_L	-0.9359	1.1112	1.4564
I_M	-1.5885	1.2883	1.8390	P_M	-0.4359	1.6112	1.9564
I_M	-1.5885	1.2883	1.8390	P_H	0.0641	2.1112	2.4564
I_H	-1.0885	1.7883	2.3390	P_L	-0.9359	1.1112	1.4564
I_H	-1.0885	1.7883	2.3390	P_M	-0.4359	1.6112	1.9564
I_H	-1.0885	1.7883	2.3390	P_H	0.0641	2.1112	2.4564

I: Item.
P: Person
L: The original mean minus 0.5 in logit
M: The original mean
H: The original mean plus 0.5 in logit

Figure 13. Plot of Bias by Grade

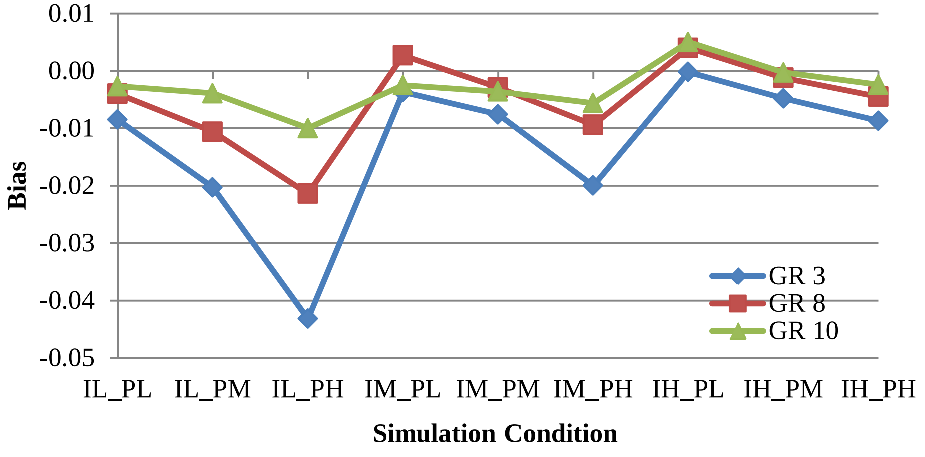

Figure 14. Plot of Mean Square Error by Grade

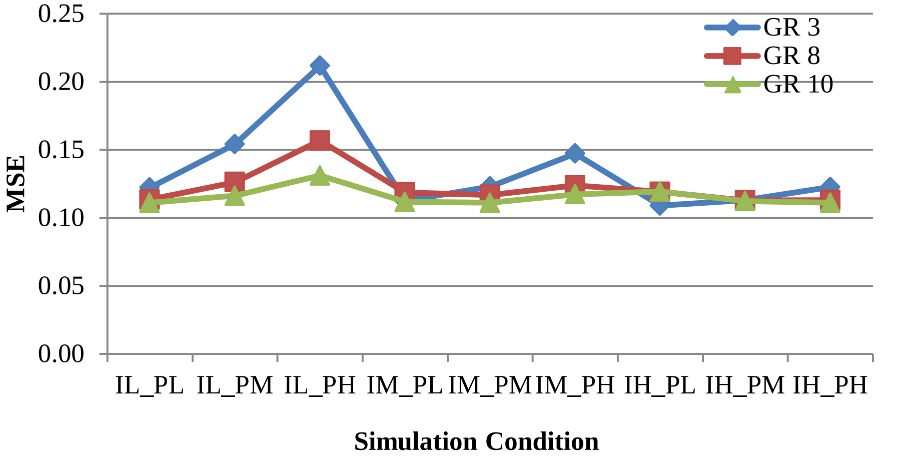

Figure 15. Plot of Conditional SEM

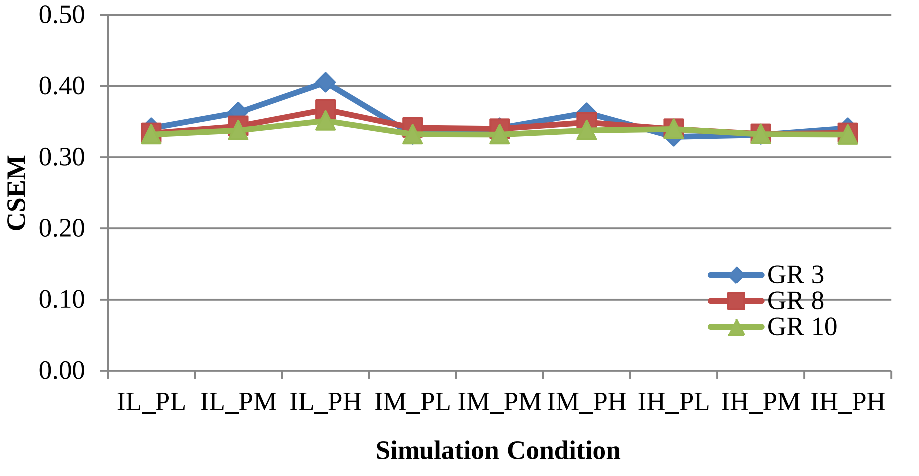

Table 13. Simulation Results for Grade 3

Overall Simulation Result	IL_PL	IL_PM	IL_PH	IM_PL	IM_PM	IM_PH	IH_PL	IH_PM	IH_PH
Bias	-0.0085	-0.0203	-0.0432	-0.0037	-0.0076	-0.0200	-0.0002	-0.0048	-0.0087
MSE (Mean Square Error)	0.1222	0.1542	0.2117	0.1125	0.1230	0.1472	0.1089	0.1130	0.1226
Correlation	0.9454	0.9357	0.9185	0.9486	0.9448	0.9375	0.9499	0.9481	0.9452
Test Length	41.5105	42.8332	44.6272	40.7557	41.5161	42.8237	40.5020	40.7545	41.5117
CSEM	0.3409	0.3630	0.4051	0.3317	0.3409	0.3624	0.3289	0.3316	0.3409
Reach Max Test Length	0.1098	0.2250	0.3917	0.0485	0.1101	0.2222	0.0293	0.0496	0.1086
Item Exposure Result									
Max_IE[-6.00 ~ -5.00]	1.0000			1.0000			1.0000		
Max_IE[-5.00 ~ -4.00]	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Max_IE[-4.00 ~ -3.00]	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Max_IE[-3.00 ~ -2.00]	0.7540	0.7560	0.7794	0.9857	0.9862	0.9797	0.9995	0.9997	1.0000
Max_IE[-2.00 ~ -1.00]	0.8801	0.8228	0.8372	0.7836	0.7017	0.7035	0.8682	0.7602	0.7562
Max_IE[-1.00 ~ 0.00]	0.9976	0.9970	0.9973	0.9671	0.9821	0.9736	0.8049	0.8979	0.8785
Max_IE[0.00 ~ 1.00]	1.0000	1.0000	1.0000	1.0000	0.9998	0.9999	0.9927	0.9944	0.9985
Max_IE[1.00 ~ 2.00]	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Max_IE[2.00 ~ 3.00]	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Max_IE[3.00 ~ 4.00]	1.0000	0.0000	0.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Max_IE[4.00 ~ 5.00]	0.0000	1.0000	1.0000		1.0000	1.0000		1.0000	1.0000
Max_IE[5.00 ~ 6.00]	0.0000								
Max_IE[6.00 ~ 7.00]	0.0000								
Overall_Max_IE	0.8971	0.9193	0.9601	0.8539	0.8963	0.9213	0.7624	0.8530	0.8966
Pool Usage Rate Distribution									
Item Never Used	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
IE_Rate[0.000 ~ 0.005]	0.0328	0.0656	0.1148	0.0109	0.0383	0.0765	0.0000	0.0055	0.0383
IE_Rate[0.005 ~ 0.100]	0.4153	0.5628	0.6011	0.2404	0.4044	0.5574	0.1639	0.2514	0.4044
IE_Rate[0.100 ~ 0.200]	0.2295	0.0929	0.0109	0.3005	0.2295	0.0874	0.2514	0.3005	0.2295
IE_Rate[0.200 ~ 0.300]	0.0546	0.0055	0.0000	0.1858	0.0601	0.0055	0.3497	0.1803	0.0601
IE_Rate[0.300 ~ 0.400]	0.0328	0.0109	0.0000	0.0984	0.0273	0.0109	0.1694	0.0984	0.0328
IE_Rate[0.400 ~ 0.500]	0.0328	0.0328	0.0109	0.0765	0.0383	0.0328	0.0546	0.0765	0.0328

IE_Rate[0.500 ~ 0.600]	0.0656	0.0109	0.0273	0.0546	0.0656	0.0109	0.0055	0.0492	0.0656
IE_Rate[0.600 ~ 0.700]	0.0710	0.0164	0.0164	0.0273	0.0710	0.0164	0.0000	0.0328	0.0710
IE_Rate[0.700 ~ 1.000]	0.0656	0.2022	0.2186	0.0055	0.0656	0.2022	0.0055	0.0055	0.0656

Table 14. Simulation Results for Grade 8

Overall Simulation Result	IL_PL	IL_PM	IL_PH	IM_PL	IM_PM	IM_PH	IH_PL	IH_PM	IH_PH
Bias	-0.0040	-0.0106	-0.0214	0.0027	-0.0029	-0.0094	0.0040	-0.0012	-0.0045
MSE (Mean Square Error)	0.1134	0.1263	0.1567	0.1187	0.1168	0.1238	0.1191	0.1129	0.1131
Correlation	0.9480	0.9440	0.9348	0.9464	0.9470	0.9444	0.9464	0.9485	0.9490
Test Length	40.9907	41.7365	43.0928	40.1154	40.1073	40.2570	41.4801	40.9175	41.0064
CSEM	0.3337	0.3437	0.3669	0.3415	0.3398	0.3489	0.3395	0.3327	0.3340
Reach Max Test Length	0.0625	0.1236	0.2431	0.0094	0.0089	0.0216	0.1026	0.0569	0.0641
Item Exposure Result									
Max_IE[-6.00 ~ -5.00]									
Max_IE[-5.00 ~ -4.00]									
Max_IE[-4.00 ~ -3.00]	1.0000			1.0000			1.0000		
Max_IE[-3.00 ~ -2.00]	1.0000	1.0000	1.0000	0.8889	1.0000	1.0000	1.0000	1.0000	1.0000
Max_IE[-2.00 ~ -1.00]	1.0000	1.0000	1.0000	0.8053	1.0000	0.0000	1.0000	1.0000	1.0000
Max_IE[-1.00 ~ 0.00]	0.9994	1.0000	1.0000	0.5836	0.8696	1.0000	1.0000	1.0000	1.0000
Max_IE[0.00 ~ 1.00]	0.7600	0.7681	0.7782	0.5152	0.5947	0.7155	0.9998	0.9996	0.9997
Max_IE[1.00 ~ 2.00]	0.9880	0.9880	0.9840	0.5039	0.5197	0.5324	0.8066	0.7471	0.7559
Max_IE[2.00 ~ 3.00]	1.0000	1.0000	1.0000	0.5052	0.4966	0.4967	0.9723	0.9729	0.9880
Max_IE[3.00 ~ 4.00]	1.0000	1.0000	1.0000	0.9713	0.5083	0.6937	1.0000	1.0000	1.0000
Max_IE[4.00 ~ 5.00]	1.0000	1.0000	1.0000	1.0000	0.9801	0.9830	1.0000	1.0000	1.0000
Max_IE[5.00 ~ 6.00]	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Max_IE[6.00 ~ 7.00]		1.0000	1.0000		1.0000	1.0000		1.0000	1.0000
Overall_Max_IE	0.8057	0.8746	0.9330	0.4031	0.5454	0.8099	0.8339	0.5652	0.8044
Pool Usage Rate Distribution									
Item Never Used	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

IE_Rate[0.000 ~ 0.005]	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
IE_Rate[0.005 ~ 0.100]	0.0823	0.2848	0.6139	0.0063	0.0127	0.1899	0.2595	0.1013	0.0823
IE_Rate[0.100 ~ 0.200]	0.4114	0.3671	0.0696	0.2848	0.4367	0.5127	0.3228	0.3418	0.4114
IE_Rate[0.200 ~ 0.300]	0.2025	0.0633	0.0063	0.3797	0.2911	0.0316	0.1076	0.1582	0.1962
IE_Rate[0.300 ~ 0.400]	0.1139	0.0253	0.0253	0.3228	0.0823	0.0127	0.0506	0.1899	0.1203
IE_Rate[0.400 ~ 0.500]	0.0759	0.0190	0.0253	0.0063	0.1456	0.0127	0.0443	0.1582	0.0759
IE_Rate[0.500 ~ 0.600]	0.0696	0.0759	0.0063	0.0000	0.0316	0.0759	0.0886	0.0506	0.0696
IE_Rate[0.600 ~ 0.700]	0.0316	0.0633	0.0190	0.0000	0.0000	0.1519	0.1139	0.0000	0.0316
IE_Rate[0.700 ~ 1.000]	0.0127	0.1013	0.2342	0.0000	0.0000	0.0127	0.0127	0.0000	0.0127

Table 15. Simulation Results for Grade 10

Overall Simulation Result	IL_PL	IL_PM	IL_PH	IM_PL	IM_PM	IM_PH	IH_PL	IH_PM	IH_PH
Bias	-0.0027	-0.0039	-0.0100	-0.0025	-0.0036	-0.0056	0.0050	-0.0003	-0.0024
MSE (mean square error)	0.1111	0.1162	0.1311	0.1118	0.1111	0.1173	0.1193	0.1123	0.1111
Correlation	0.9492	0.9480	0.9423	0.9488	0.9487	0.9467	0.9459	0.9486	0.9489
Test Length	40.8058	41.3881	42.5115	40.8375	40.7900	41.3601	41.5065	40.8433	40.8053
CSEM	0.3320	0.3377	0.3514	0.3322	0.3319	0.3377	0.3396	0.3324	0.3318
Reach Max Test Length	0.0473	0.0902	0.1820	0.0485	0.0465	0.0872	0.1161	0.0548	0.0468
Item Exposure Result									
Max_IE[-6.00 ~ -5.00]									
Max_IE[-5.00 ~ -4.00]									
Max_IE[-4.00 ~ -3.00]	1.0000			1.0000			1.0000		
Max_IE[-3.00 ~ -2.00]	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Max_IE[-2.00 ~ -1.00]	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Max_IE[-1.00 ~ 0.00]	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Max_IE[0.00 ~ 1.00]	0.9648	0.9579	0.9539	0.9979	0.9971	0.9973	1.0000	1.0000	1.0000
Max_IE[1.00 ~ 2.00]	0.8986	0.9330	0.9425	0.7401	0.6710	0.7265	0.9786	0.9691	0.9622
Max_IE[2.00 ~ 3.00]	1.0000	1.0000	1.0000	0.9707	0.9854	0.9935	0.7854	0.8416	0.8997
Max_IE[3.00 ~ 4.00]	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Max_IE[4.00 ~ 5.00]	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Max_IE[5.00 ~ 6.00]	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Max_IE[6.00 ~ 7.00]		1.0000	1.0000		1.0000	1.0000		1.0000	1.0000
Overall_Max_IE	0.7347	0.8772	0.9576	0.6737	0.7365	0.8783	0.8348	0.7216	0.7378
Pool Usage Rate Distribution									
Item Never Used	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
IE_Rate[0.000 ~ 0.005]	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
IE_Rate[0.005 ~ 0.100]	0.2736	0.3483	0.6020	0.2587	0.2687	0.3483	0.5174	0.3333	0.2736
IE_Rate[0.100 ~ 0.200]	0.3085	0.3433	0.1393	0.3284	0.3085	0.3433	0.1841	0.3035	0.3085
IE_Rate[0.200 ~ 0.300]	0.2139	0.0945	0.0299	0.2040	0.2239	0.0945	0.0448	0.1144	0.2189
IE_Rate[0.300 ~ 0.400]	0.1045	0.0547	0.0199	0.1144	0.0995	0.0547	0.0597	0.1144	0.1045
IE_Rate[0.400 ~ 0.500]	0.0448	0.0647	0.0199	0.0547	0.0448	0.0597	0.0348	0.0647	0.0398
IE_Rate[0.500 ~ 0.600]	0.0398	0.0149	0.0348	0.0249	0.0398	0.0199	0.0398	0.0597	0.0398
IE_Rate[0.600 ~ 0.700]	0.0100	0.0547	0.0647	0.0149	0.0100	0.0547	0.0697	0.0050	0.0100
IE_Rate[0.700 ~ 1.000]	0.0050	0.0249	0.0896	0.0000	0.0050	0.0249	0.0498	0.0050	0.0050