The Effects of Mixture Distribution of Calibration Sample on the Accuracy of Rasch Item Parameter Estimations in Computerized Adaptive Tests

Shudong Wang
Northwest Evaluation Association

Paper presented at the meeting of the American Educational Research Association (AERA).
April 28, 2013, San Francisco, CA

Send correspondence to:
Shudong Wang
Northwest Evaluation Association (NWEA)
121 NW Everett St.
Portland, OR 97209
Shudong.Wang@NWEA.org

Purpose

The purpose of this study is to investigate the impact of population heterogeneity on the accuracy of Rasch model item parameter estimation.

Perspectives

Computerized adaptive testing (CAT) is currently used in many K-12, licensure, certification and medical tests (ACRE, 2010; DRC, 2007; Reckase, 2011; Way, Twing, Camara, Sweeney, Lazer & Mazzeo, 2010; International Association for Computerized Adaptive Testing, 2012). The majority routinely conduct item parameter estimations using the Rasch model in their CAT operational systems.

The advantages of CAT rely on Rasch model assumptions and conditions that are required for implementing the appropriate Rasch model. For example, one of the most important Rasch model properties (sufficiency, separability, specific objectivity and latent additivity) is specific objectivity (SP), which means that the item parameters (or item characteristic curves) are invariant across samples of examinees from the same population. However, this property assumes that the population contains only one latent ability or trait that accounts for examinee test performance (person unidimensional assumption), which means that examinees in the calibration sample (1) should come from the same population and (2) their achievement abilities are random variables that follow a normal distribution. Many research studies (Harrison, 1986; Kirisci, Hsu & Yu, 2001; Stocking, 1990; Thissen & Wainer, 1982) show a close relationship between the precision of item parameter estimation and the ability distribution of examinees used for calibration. If this assumption is violated—i.e., examinees come from multiple populations and don't follow normal distribution—it is likely that the SP property will no longer hold, calibrated item parameters tend to be biased and a misleading inference can be made. Thus, the claim of the sample-free feature of Rasch model may no longer be valid.

Unlike item calibration in fixed-form tests, the additional complexities of item parameter estimation in CAT field tests, compared to those in fixed-form tests, include (1) that items match examinees' provisional abilities and (2) that different examinees see different sets of items. In the current CAT operational calibration procedure, two potentially major violators of the Rasch

model unidimensional assumption may come from (1) using examinees from multiple grades in the calibration sample and (2) aggregating multiple calibration samples across time.

In some research focused on CAT calibration procedure algorithm and optimal design (Ban, Hanson, Wang, Yi & Harris, 2001; Kingsbury, 2009; Van der Linden & Glas, 2000), a few studies have focused on the impact of population heterogeneity on the accuracy of item parameter estimation. Sampling from multiple populations can lead to heterogeneous samples.

Research Questions

1) What is the actual relationship of distributions between calibration and operational samples used in the calibration procedure of a large-scale K-12 CAT assessment?
2) Is there any mixture distribution in calibration samples of a large-scale K-12 CAT assessment?
3) If there is mixture distribution in the operational sample of a large-scale K-12 CAT assessment, are there any statistically significant differences in accuracy of item parameter estimates between calibration and operational samples?
4) If there is mixture distribution in calibration sample of a large-scale K-12 CAT assessment, are there any practical significant differences in accuracy of item parameter estimates between calibration and operational samples?

Statistical Hypotheses

1) The type of distribution does not affect the accuracy of item parameter estimations when some or all dependent variables (or log transform of them)—correlation, bias, abias, SE and RMSE—are used in different simulation conditions.
2) Sample size does not affect the accuracy of item parameter estimations when some or all of the dependent variables (or log transform of them) —correlation, bias, abias, SE and RMSE—are used in different simulation conditions.
3) There are no interaction effects between the two factors mentioned above when some or all of the dependent variables (or log transform of them) —correlation, bias, abias, SE and RMSE—are used in different simulation conditions.

Methods

Data Sources

Both real and simulated data are used in this study. All simulated data for both person and items are based on real data that include both calibration and operational samples for Reading and Mathematics 2011 spring administration large-scale CAT tests across grades 2 to 12 and more than 20 U.S. states. The operational sample size by subject (Reading and Mathematics) and grade (2 to 12) is over 100,000; the calibration sample size by subject (Reading and Mathematics) and grade (2 to 12) is over 1100. Figures 1 to 13b depict distributions of students' final scale scores (SS) across subject and grade for both calibration and operational samples.

Design of Study

The primary goal of this design is to answer the stated research questions and to maximize the generalizability and replicability of research results. Both descriptive methods and inferential procedures are used in this study. The steps to simulate data involve two major steps:

Step 1: Investigate the empirical relationship of distributions between calibration and operational samples by:

1) Drawing 10 reading and 10 mathematics items randomly from item pools. The pools from which items were drawn contain more than 10,000 reading items and 10,000 mathematics items. Plotting sample distributions of 10 reading and 10 mathematics items across grades.
2) Determining the shapes of sample distributions.

Step 2: Generate item response based on empirical distributions.

1) Select typical mismatch distributions between calibration and operational samples.
2) Choose operational distribution as base samples or true samples.
3) Fit both operational and calibration empirical distributions to theoretical distribution densities.
4) Use operational (true) distribution density to generate items parameters (40 items for each of the samples).
5) Use generated item parameters and operational density to generate operational response samples.

6) Use generated item parameters and calibration density to generate calibration response samples.

Independent Variables

Two independent variables used in this study are distribution type and sample sizes. The distribution of the calibration sample can be a finite mixture, i.e.

$$f(x)=\sum_{i=1}^{k} w_i p_i(x) \quad (1)$$

where $p_i(x)$ is a set of finite standard normal probability density functions (i = 1, 2, …, k) and w_i is a set of weights that $w_i \geq 0$ and $\sum w_i = 1$. In this study, k = 2. There is a total of 3 types of distributions for calibration samples and four sample sizes. Table 1 shows detailed information of two independent variables.

Dependent Variables

Five criterion variables used in this study are: correlations between true and estimated parameters, biases, absolute biases (abias), standard errors (SEs) and root mean square errors (RMSEs). These criteria are used to examine the effects of the manipulated independent variables described in the last subsection to provide complementary evidence. For each i item, the conditional bias (abias), SE and RMSE of an estimator $\hat{b}$ across N (r = 1, 2, … , N) replications can be expressed as following:

$$\text{Bias}(\hat{b}_i) = E(\hat{b}_i) - b_i = \frac{1}{N}\sum_{r=1}^{N}\hat{b}_{ri} - b_i = \frac{1}{N}\sum_{r=1}^{N}\hat{b}_{ri} - \frac{1}{N}\sum_{r=1}^{N} b_i = \frac{1}{N}\sum_{r=1}^{N}(\hat{b}_{ri} - b_i) \quad (2)$$

$$\text{Abias}(\hat{b}_i) = |E(\hat{b}_i) - b_i| = \frac{1}{N}\sum_{r=1}^{N}|\hat{b}_{ri} - b_i| \quad (3)$$

$$SE(\hat{b}_i) = \sqrt{Var(\hat{b}_i)} = \sqrt{E[(\hat{b}_i - E(\hat{b}_i))^2]} = \sqrt{\frac{1}{N}\sum_{r=1}^{N}\left(\hat{b}_{ri} - \frac{1}{N}\sum_{r=1}^{N}\hat{b}_{ri}\right)^2} \quad (4)$$

Where $\hat{b}$ is the estimated item difficulty and b is true difficulty

$$RMSE(\hat{b}_i) = \sqrt{\frac{1}{N}\sum_{r=1}^{N}(\hat{b}_{ri} - b_i)^2} \quad (5)$$

The relationship between MSE (=RMSE^2), SE and bias is:

$$MSE(\hat{b}_i) = E\left[(\hat{b}_i - b_i)^2\right] = E\left[\left(\hat{b}_i - E(\hat{b}_i)\right)^2 + \left(E(\hat{b}_i) - b_i\right)^2\right] = Var(\hat{b}_i) + Bias^2(\hat{b}_i) \quad (6)$$

This relationship can be used to verify the calculation accuracy for each criterion index. The average of bias, Abias, SE, and RMSE across M items (i=1, 2, …, M) can be described as:

$$\text{Bias}(\hat{b}) = \frac{1}{M}\frac{1}{N}\sum_{i=1}^{M}\sum_{r=1}^{N}(\hat{b}_{ri} - b_i) \quad (7)$$

$$\text{Abias}(\hat{b}) = \frac{1}{M}\frac{1}{N}\sum_{i=1}^{M}\sum_{r=1}^{N}|\hat{b}_{ri} - b_i| \quad (8)$$

$$SE(\hat{b}) = \sqrt{\frac{1}{M}\frac{1}{N}\sum_{i=1}^{M}\sum_{r=1}^{N}\left(\hat{b}_{ri} - \frac{1}{N}\sum_{r=1}^{N}\hat{b}_{ri}\right)^2} \quad (9)$$

$$RMSE(\hat{b}) = \sqrt{\frac{1}{M}\frac{1}{N}\sum_{i=1}^{M}\sum_{r=1}^{N}(\hat{b}_{ri} - b_i)^2} \quad (10)$$

The relationship among average bias, SE and RMSE in (6) is no longer true for average bias, SE and RMSE.

Data Analysis

All data are calibrated by using WINSTEPS (Linacre, 2009). For calibration samples, there are 4 (Type of Distribution) x 4 (Sample Sizes) = 16 data sets. For operational samples, one type distribution with 4 different sample sizes are generated so that there are only 4 data sets. Each of the data sets are replicated 99 times so there are a total (4 x 4 + 4) x 99 = 1,980 data sets generated and calibrated in this study. All operational data are treated as true and are matched to different calibration data. Overall indices of dependent variables (or log transformation of dependent variables if necessary) are used in analysis of variance (ANOVA) to test the main effect of independent variables.

Results and Discussions

1. Empirical and Simulated Sample Distributions

1.1 Empirical operational and calibration sample distributions of reading and mathematics items at each of 12 grades.

Figures 1a to 3b present typical sample distributions of mathematics items from grade 2 to 12; Figures 4a to 13b demonstrate the empirical sample distributions of all 10 reading items from grade 2 to 12. Figures 14a to 18c illustrate both operational and calibration sample distributions based on groups of grades. In general, no matter which group samples are used, operational samples are approximately normally distributed and calibration samples are clearly bimodal.

1.2 Simulated operational and calibration samples distributions of reading items

Figures 19a to 19c plot the simulated distributions of operational and calibration samples, and all simulated distribution parameters are listed in Table 1. The distribution parameters listed in the table are for standard normal distribution, or they are in logit units that can be transferred to the scale score SS (also called RIT) by following formula:

$$SS \text{ (or } RIT) = 10 * logit + 200$$

All figures presented so far are based on the scale score.

2. *Descriptive Statistics of Dependent Variables*

2.1 Correlations among true and estimated item parameters.

Tables 2 to 4 present the average correlations of item parameter estimates between operational and calibration samples by type of distribution and sample size, over type of distribution and over sample size. Figure 20 demonstrates average correlations between operational and calibration item parameter estimates by type of distribution and sample size.

2.2 Average Bias and Absolute Bias

Tables 2 to 4 list average bias and abias over replication and independent variables. Figures 21 to 22 plot the bias and abias at different types of sample distributions and sample sizes.

2.3 Average SE

The average SEs at different types of sample distributions and sample sizes are presented in Tables 2 to 4 and plotted in Figure 23.

2.4 Average RMSE

The average RMSEs at different types of sample distributions and sample sizes are presented in Tables 2 to 4 and plotted in Figure 24.

3. Inferential Statistics of Average Dependent Variables

Because the ANOVA assumption requires that variables be normally distributed, three of them (abias, SE and RMSE) are needed to log transformation to satisfy the normal assumption by examining the histograms of five dependent variables.

3.1 Statistical Hypotheses

Based on the research questions proposed in the introduction section, the statistical null hypotheses are tested.

In this study, in order to have adequate power for the statistical tests in the Monte Carlo study to detect effects of interest, each simulated condition has been replicated 99 times. The magnitude of significant effects is estimated using eta-squared η^2 (empirical η^2 as an effect size estimate).

3.2 ANOVA Results

Tables 5 to 9 show the results of the two-way ANOVA of average of correlation, bias, log(abias), log(SE) and log(RMSE) for this study. The two main effects—TD (Type Distribution) and SS (Sample Size) —and one interaction effect—TD x SS—are all statistically significant except for bias.

Figures 25 to 27 plot the log(abias), log(SE) and log(RMSE) under different TD and SS. Among the main effects for correlation, according to Cohen (1988), both sample size and type of distribution effect size are in the small ranges ($0.01 < \eta^2 < 0.0588 \approx 0.06$). For log(abias), log(SE) and log(RMSE), the sample size accounts for most of the variance (32.5%, 39.8% and 32.5%) and the type of distribution accounts for the second highest amount of total variance (8.4%, 12.4% and 8.4%). For the sample size, the effect sizes are in the large ranges ($\eta^2 > 0.14$), and for the type of distribution, the effect sizes are in the medium ranges ($0.06 < \eta^2 < 0.1379 \approx 0.14$).

4. *Relative Increase of Average Dependent Variables of Different Type Distributions Over Base Distribution (Type 4 Distribution)*

In this study, distribution type 4 can be used as base distribution (see Table 1), and any deviations of dependent variable values from base distribution values can be used to compare the relative increase (or decrease) of dependent variables. Table 10 lists ratios of average correlation, abias, SE and RMSE of different distributions over the base distribution, and Figure 28 plots these values.

5. *Summary of Results*

First, there are statistically significant differences in the accuracy of item parameter estimates among different types of calibration sample distributions in terms of correlation, log(abias), log(SE) and log(RMSE). Second, there are statistically significant differences in the accuracy of item parameter estimates among different calibration sample sizes in terms of correlation, log(abias), log(SE) and log(RMSE). Third, there are interaction effects between two factors mentioned above when correlation, log(abias), log(SE) and log(RMSE) are used in different simulation conditions.

In general, the type of distribution accounts for 8% to 12% of total variance, and sample size accounts for 32% to 40% of total variance. The type of distribution has medium ranges of effect sizes and the sample size has large effect sizes. Type 1 distribution has the largest calibration errors and type 3 distribution has the smallest calibration errors. Both calibration errors and relative calibration errors decrease as the sample size increases.

Scientific Significance of the Study

The accuracy of item parameter estimation is the psychometric foundation of any test program that uses an IRT model and intends to provide valid testing results. The results of this study illustrate fairly clearly that the sample size and type of distribution have not only a statistically significant impact, but also a practical implication on the accurate recovery of item parameters. Regardless of the type of distribution, calibration errors decrease as the sample size increases, and the implication here is that CAT users should always try to use large sample sizes for its item calibration. The benefit of doing so is when under certain circumstances where factors that could affect calibration accuracy are not clear, the large sample size could compensate for the effect of these factors. **References**

Accountability and Curriculum Reform Effort (2010). *Computerized adaptive testing: How CAT may be utilized in the next generation of assessments*. Retrieved from

http://www.ncpublicschools.org/docs/acre/publications/2010/publications/20100716-0.1.pdf

Ban, J. C., Hanson, B. H., Wang, T., Yi, Q. & Harris, D. J. (2001). A comparative study of on-line pretest item calibration-scaling methods in computerized adaptive testing. *Journal of Educational Measurement* 38: 191-212.

Buyske, S. G. (1998). Optimal design for item calibration in computerized adaptive testing: The 2PL case. *New Developments and Applications in Experimental Design* 34: 115-125.

Campbell, D. T. & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.

Cohen, J. (1988). *Statistical Power Analysis for the Behavior Sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Data Recognition Corporation (2007). *Study on the feasibility and cost of converting the state assessment program to a computer-based or computer-adaptive format*. Retrieved from http://eoc.sc.gov/NR/rdonlyres/CAEF9136-26CB-421D-80E3-D5B35B72CE76/5535/SCFeasibilityFinalReport.pdf.

Hambleton, R. K. & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Higman, MA: Kluwer Academic Publishers.

Harrison, D. A. (1986). Robustness of IRT parameter estimation to violations of the unidimensionality assumption. *Journal of Educational Statistics* 11: 91–115.

Harwell, M. R. (1997). *Analyzing the results of Monte Carlo studies in item response theory.* Educational and Psychological Measurement 57: 266-279.

Harwell, M. R., Stone, C. A., Hsu, T. C. & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement* 20: 101-125.

Hoaglin, D. C. & Andrews, D. F. (1975). The reporting of computation-based results in statistics. *American Statistician* 29: 122-126.

Holman, R. & Berger, M. P. F. (2001). Optimal calibration designs for tests of polytomously scored items described by item response theory models. *Journal of Educational and Behavioral Statistics* 26: 361-380.

Hsu, Y., Thompson, T. D. & Chen, W.H. (1998). *CAT item calibration*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Diego.

International Association for Computerized Adaptive Testing (2012). *Operational CAT Programs*. Retrieved from http://iacat.org/node/427.

Kingsbury, G. G. (2009). Adaptive item calibration: A process for estimating item parameters

within a computerized adaptive test. In D. J. Weiss (Ed.), *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing*.

Kirisci, L., Hsu, T. & Yu, L. (2001). Robustness of item parameter estimation programs to assumptions of unidimensionality and normality. *Applied Psychological Measurement* 25 (2): 146–162.

Linacre, J. M. (2009). *Winsteps* (Version 3.69) Chicago, IL: Winsteps.com.

Naylar, T. H., Balintfy, J. L., Burdick, D. S. & Chu, K. (1968). *Computer simulation techniques*. New York: Wiley.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.

Rasch, G. (1977). On specific objectivity: An attempt at formalising the request for generality and validity of scientific statements. *Danish Yearbook of Philosophy* 14: 58-94.

Reckase, M. (2011). Computerized adaptive assessment (CAA): The way forward. In Policy Analysis for California Education and Rennie Center for Education Research & Policy (Eds.), *The road ahead for state assessments* (pp. 1–12). Cambridge, MA: Rennie Center for Education Research & Policy.

Spence, I. (1983). Monte Carlo simulation studies. *Applied Psychological Measurement* 7: 405-425.

Stocking, M. L. (1990). Specifying optimum examinees for item parameter estimation in item response theory. *Psychometrika* 55: 461-475.

Thissen, D. & Wainer, H. (1982). Some standard errors in item response theory. *Psychometrika* 47: 397-412.

Timm, N. H. (1975). *Multivariate analysis with applications in education and psychology*. Monterey CA: Brooks/Cole.

Van der Linden, W. J. & Glas, C. A. W. (2000). Cross-validating item parameter estimation in adaptive testing. In A. Boorsma, M. A. J. van Duijn & T. A. B. Snijders (Eds.), *Essays on item response theory*. New York: Springer.

Way, W. D., Twing, J. S., Camara, W., Sweeney, K., Lazer, S. & Mazzeo, J. (2010). *Some considerations related to the use of adaptive testing for the Common Core Assessments*. Retrieved from http://www.ets.org/s/commonassessments/pdf/AdaptiveTesting.pdf

Wingersky, M. S. & Lord, F. M. (1984). An investigation of methods for reducing sampling error in certain IRT procedures. *Applied Psychological Measurement* 8: 347-364.

Wright, B. D. & Stone, M. H. (1979). *Best test design*. MESA Press, Chicago.

Table 1. Type of Distribution and Sample Size

Sample Type	Sample Size	Type of Distribution	Distribution Density $p_1 \sim N(\mu_1, \sigma_1^2)$			Distribution Density $p_2 \sim N(\mu_2, \sigma_2^2)$		
			w_1	μ_1	σ_1	w_2	μ_2	σ_2
Calibration	300	1 (Grades 2-5)	0.70	-0.50	1.00	0.3	2.30	1.10
	500		0.70	-0.50	1.00	0.3	2.30	1.10
	1000		0.70	-0.50	1.00	0.3	2.30	1.10
	2000		0.70	-0.50	1.00	0.3	2.30	1.10
	300	2 (Grades 6-12)	0.45	-0.70	1.20	0.55	2.70	1.30
	500		0.45	-0.70	1.20	0.55	2.70	1.30
	1000		0.45	-0.70	1.20	0.55	2.70	1.30
	2000		0.45	-0.70	1.20	0.55	2.70	1.30
	300	3 (Grades 2-12)	0.70	-0.90	1.40	0.3	2.70	1.10
	500		0.70	-0.90	1.40	0.3	2.70	1.10
	1000		0.70	-0.90	1.40	0.3	2.70	1.10
	2000		0.70	-0.90	1.40	0.3	2.70	1.10
	300	4 (True)	1.00	3.40	1.00			
	500		1.00	3.40	1.00			
	1000		1.00	3.40	1.00			
	2000		1.00	3.40	1.00			
Operation	300	True	1.00	3.40	1.00			
	500		1.00	3.40	1.00			
	1000		1.00	3.40	1.00			
	2000		1.00	3.40	1.00			

Table 2. Average Correlation, Bias, Abias, SE and RMSE for Different Types of Distributions and Sample Size Over Replications

Type of Distribution	Sample Size	N-counts	Correlation	Bias	Abias	SE	RMSE
1	300	198	0.955	-0.059	0.237	0.251	0.320
	500	198	0.978	-0.063	0.072	0.130	0.188
	1000	198	0.977	-0.096	0.145	0.119	0.166
	2000	198	0.988	-0.076	0.031	0.069	0.124
2	300	198	0.981	-0.080	0.141	0.160	0.224
	500	198	0.984	-0.070	0.056	0.105	0.155
	1000	198	0.994	-0.076	0.023	0.068	0.124
	2000	198	0.990	-0.083	0.058	0.072	0.127
3	300	198	0.983	-0.026	0.059	0.130	0.195
	500	198	0.991	-0.069	0.022	0.089	0.144
	1000	198	0.995	-0.077	0.016	0.062	0.124
	2000	198	0.988	-0.026	0.011	0.044	0.103
4	300	198	0.989	-0.067	0.023	0.082	0.149
	500	198	0.995	-0.070	0.014	0.063	0.116
	1000	198	0.997	-0.074	0.010	0.045	0.101
	2000	198	0.999	-0.079	0.009	0.032	0.093

Table 3. Average Correlation, Bias, Abias, SE and RMSE Over Types of Distributions

Sample Size	N-counts	Correlation	Bias	Abias	SE	RMSE
300	792	0.977	-0.058	0.115	0.156	0.222
500	792	0.987	-0.068	0.041	0.097	0.151
1000	792	0.991	-0.084	0.043	0.074	0.128
2000	792	0.994	-0.079	0.027	0.054	0.112

Table 4. Average Correlation, Bias, Abias, SE and RMSE Over Sample Sizes

Type of Distribution	N-counts	Correlation	Bias	Abias	SE	RMSE
1	792	0.975	-0.076	0.116	0.142	0.199
2	792	0.987	-0.077	0.070	0.101	0.158
3	792	0.992	-0.063	0.027	0.081	0.141
4	792	0.995	-0.072	0.014	0.056	0.115

Table 5. Results of ANOVA of Correlation

Source	DF	Type I SS	Mean Square	*F*-Value	Pr > F	η^2
Distribution Type	3	0.118	0.039	15.540	<.0001	0.014
Sample Size	3	0.167	0.056	22.000	<.0001	0.020
Type x Size	9	0.069	0.008	3.010	0.0014	0.008

Table 6. Results of ANOVA of Bias

Source	DF	Type I SS	Mean Square	*F*-Value	Pr > F	η^2
Distribution Type	3	0.089	0.030	1.050	0.368	0.001
Sample Size	3	0.275	0.092	3.260	0.021	0.003
Type x Size	9	0.302	0.034	1.190	0.296	0.003

Table 7. Results of ANOVA of log(Abias)

Source	DF	Type I SS	Mean Square	*F*-Value	Pr > F	η^2
Distribution Type	3	169.706	56.5688	152.39	<.0001	0.084
Sample Size	3	659.103	219.701	591.83	<.0001	0.325
Type x Size	9	29.7912	3.31013	8.92	<.0001	0.015

Table 8. Results of ANOVA of log(SE)

Source	DF	Type I SS	Mean Square	*F*-Value	Pr > F	η^2
Distribution Type	3	137.615	45.872	274.570	<.0001	0.124
Sample Size	3	443.017	147.672	883.900	<.0001	0.398
Type x Size	9	5.552	0.617	3.690	0.000	0.005

Table 9. Results of ANOVA of log(RMSE)

Source	DF	Type I SS	Mean Square	*F*-Value	Pr > F	η^2
Distribution Type	3	42.427	14.142	152.390	<.0001	0.084
Sample Size	3	164.776	54.925	591.830	<.0001	0.325
Type x Size	9	7.448	0.828	8.920	<.0001	0.015

Table 10. Ratios of Average Correlation, Bias, Abias, SE and RMSE of Different Distributions Over Base Distribution (4)

Type of Distribution	Sample Size	Correlation	Abias	SE	RMSE
1	300	0.97	10.30	3.06	2.15
	500	0.98	5.14	2.06	1.62
	1000	0.98	14.50	2.64	1.64
	2000	0.99	3.44	2.16	1.33
2	300	0.99	6.13	1.95	1.50
	500	0.99	4.00	1.67	1.34
	1000	1.00	2.30	1.51	1.23
	2000	0.99	6.44	2.25	1.37
3	300	0.99	2.57	1.59	1.31
	500	1.00	1.57	1.41	1.24
	1000	1.00	1.60	1.38	1.23
	2000	0.99	1.22	1.38	1.11
4	300	1	1	1	1
	500	1	1	1	1
	1000	1	1	1	1
	2000	1	1	1	1

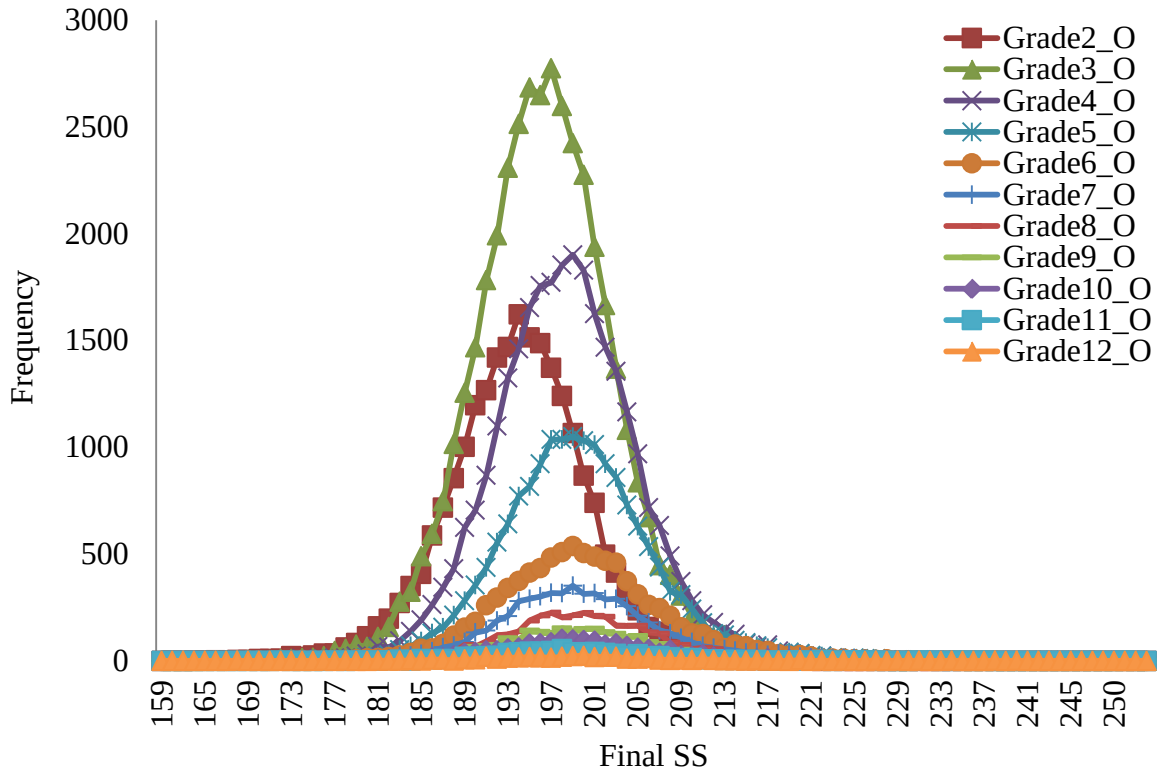

Figure 1a. Frequency Distributions of Mathematics Operational Samples across Grades (Item 1)

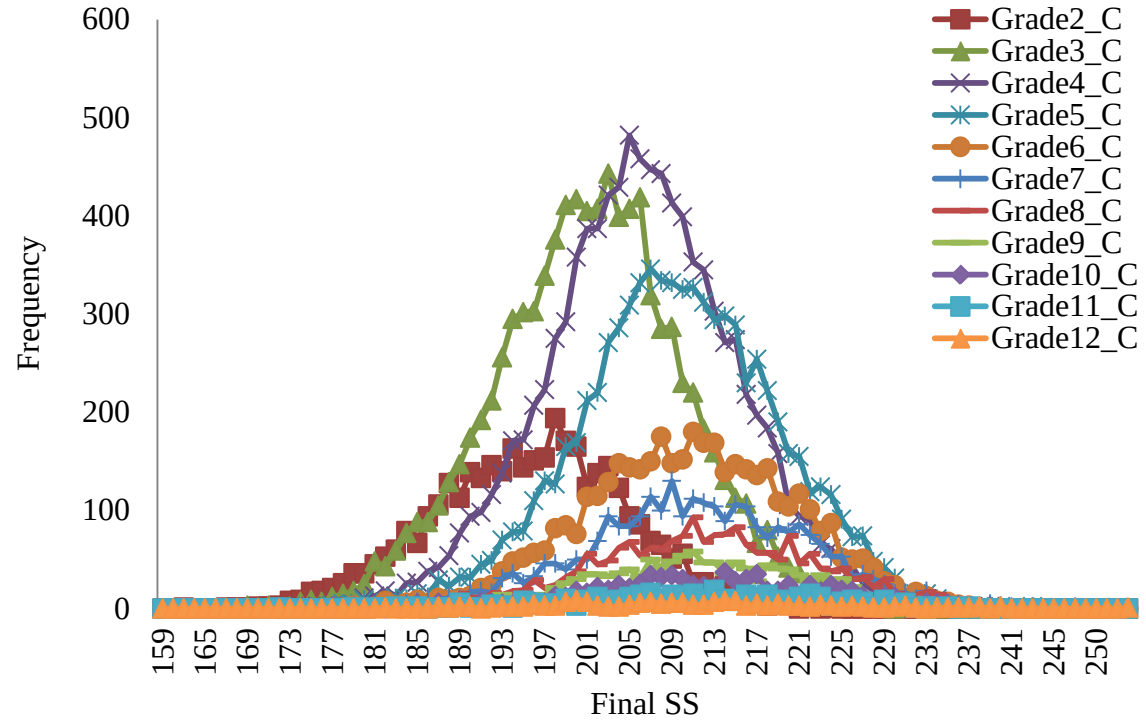

Figure 1b. Frequency Distributions of Mathematics Calibration Samples across Grades (Item 1)

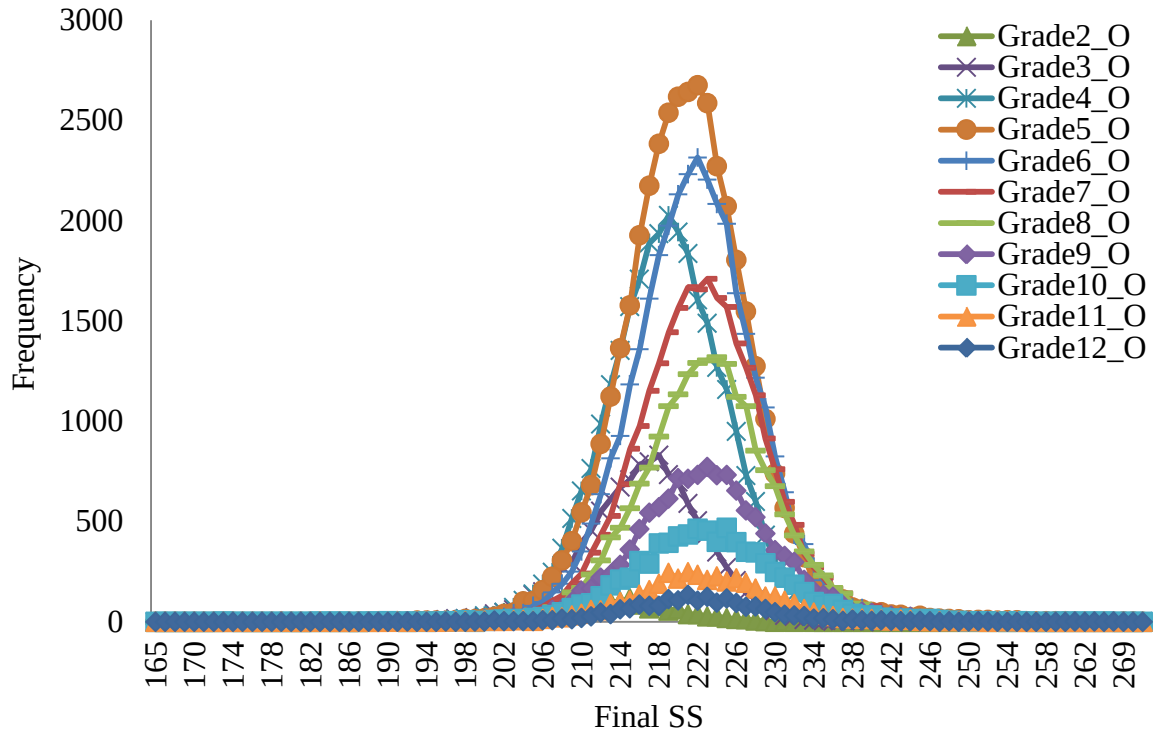

Figure 2a. Frequency Distributions of Mathematics Operational Samples across Grades (Item 2)

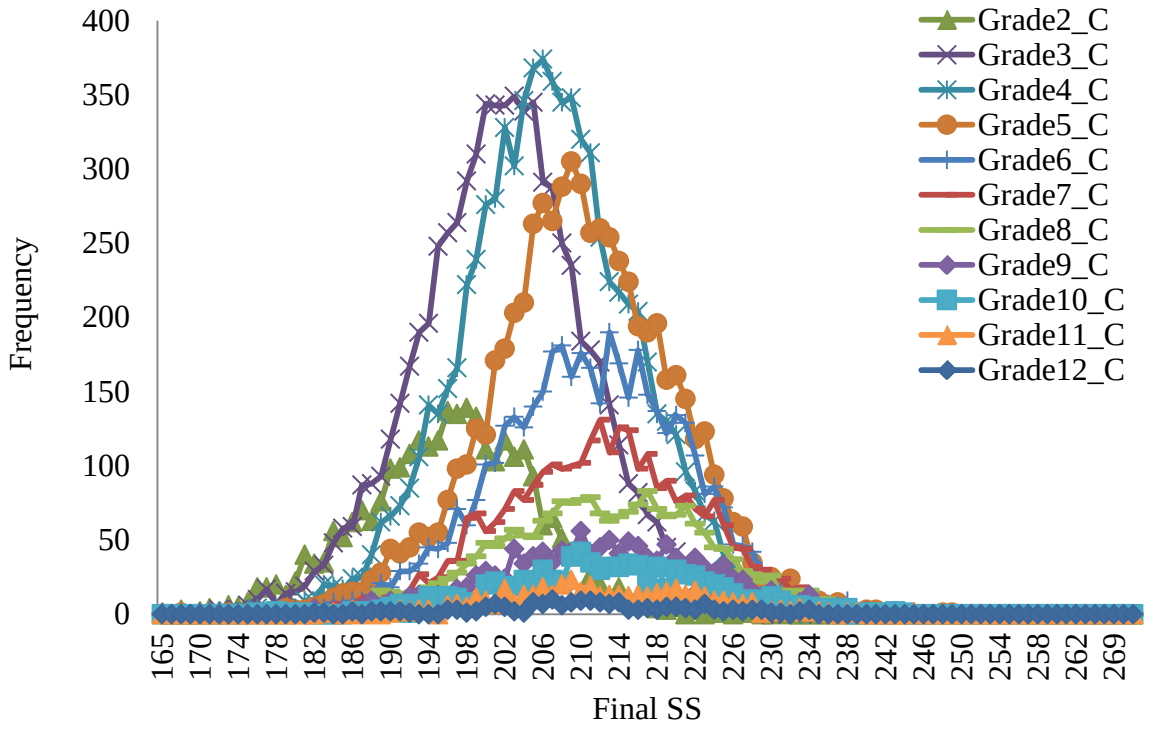

Figure 2b. Frequency Distributions of Mathematics Calibration Samples across Grades (Item 2)

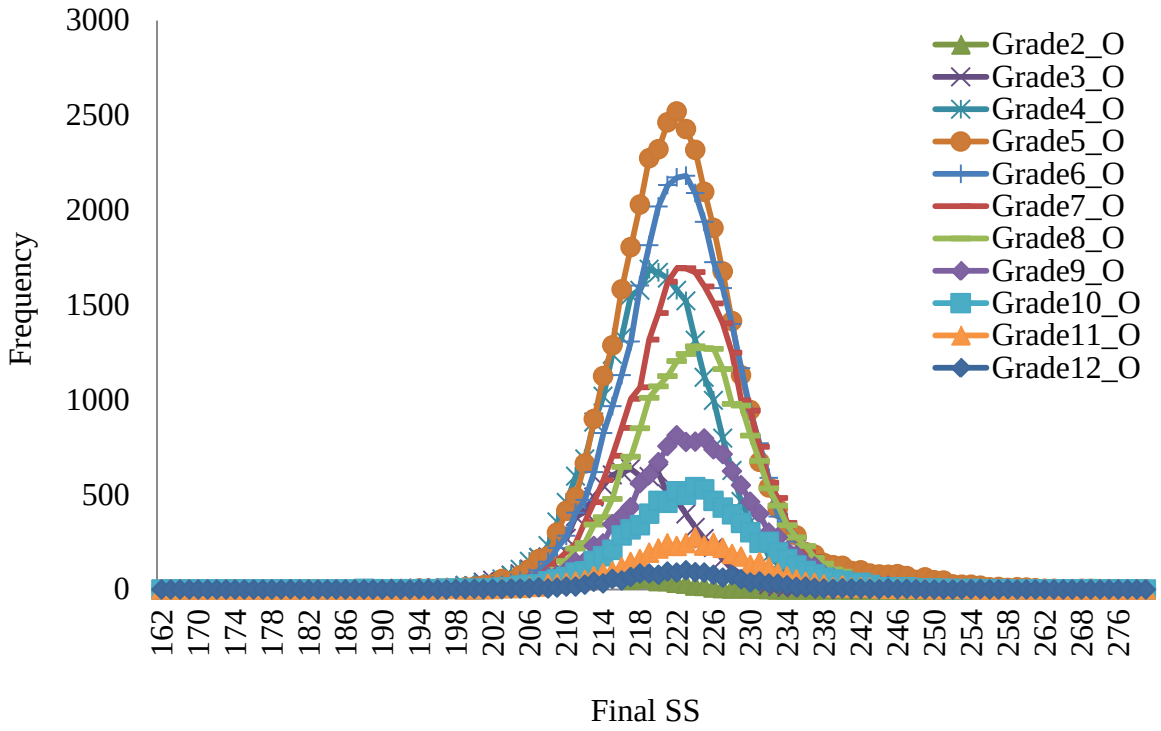

Figure 3a. Frequency Distributions of Mathematics Operational Samples across Grades (Item 3)

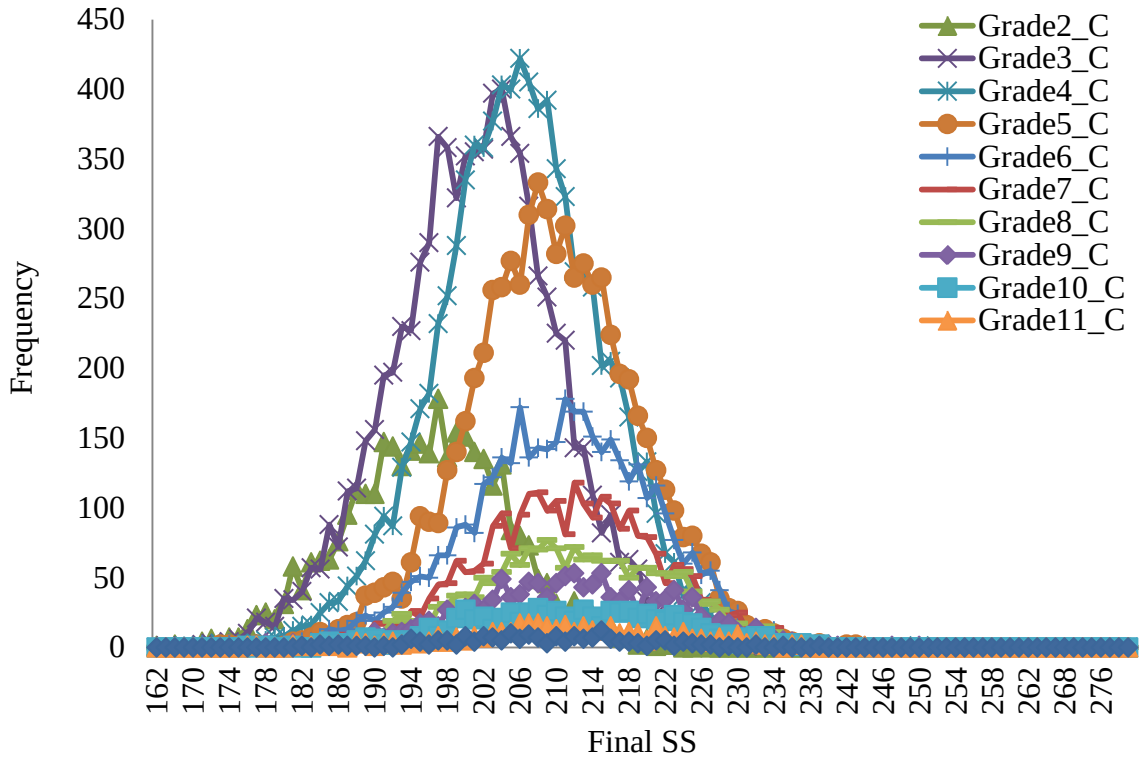

Figure 3b. Frequency Distributions of Mathematics Calibration Samples across Grades (Item 3)

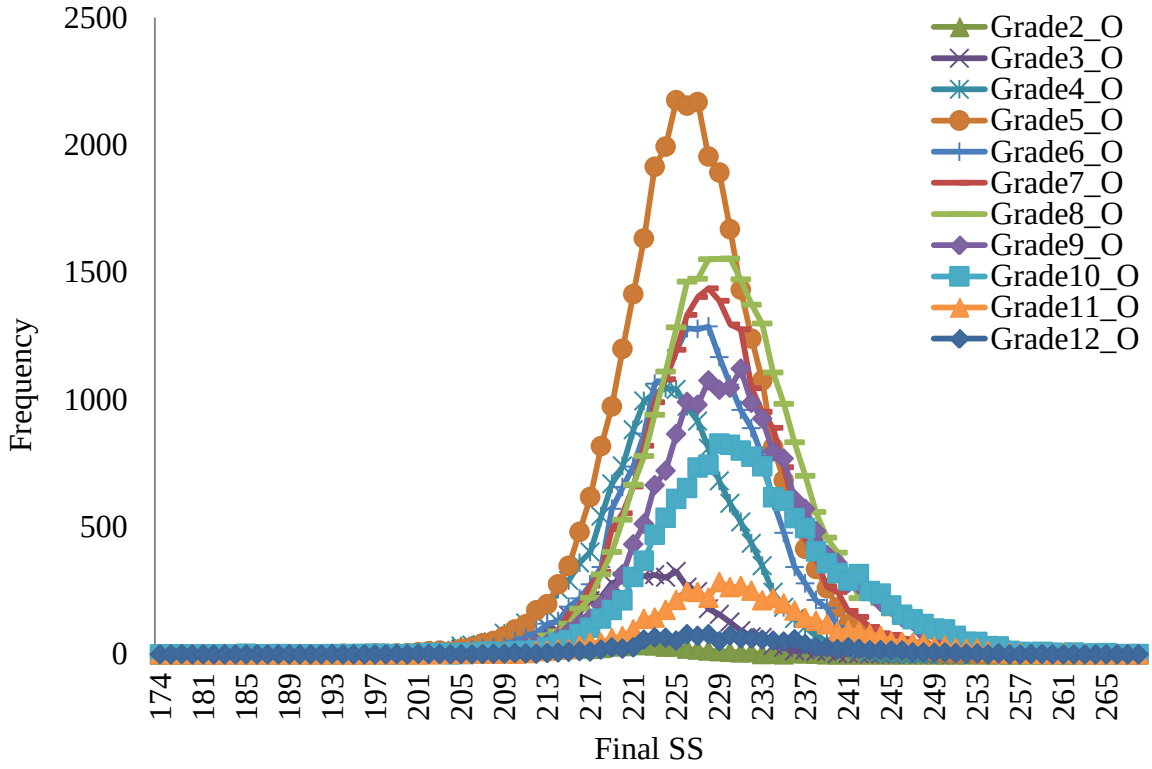

Figure 4a. Frequency Distributions of Reading Operational Samples across Grades (Item 1)

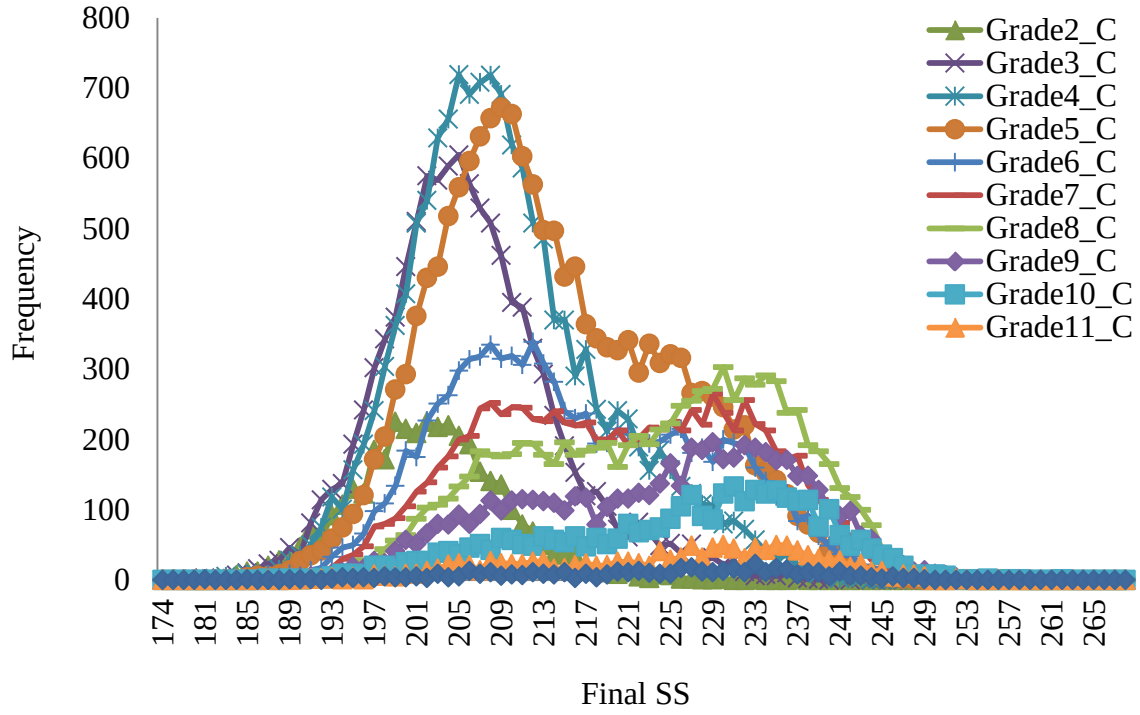

Figure 4b. Frequency Distributions of Reading Calibration Samples across Grades (Item 1)

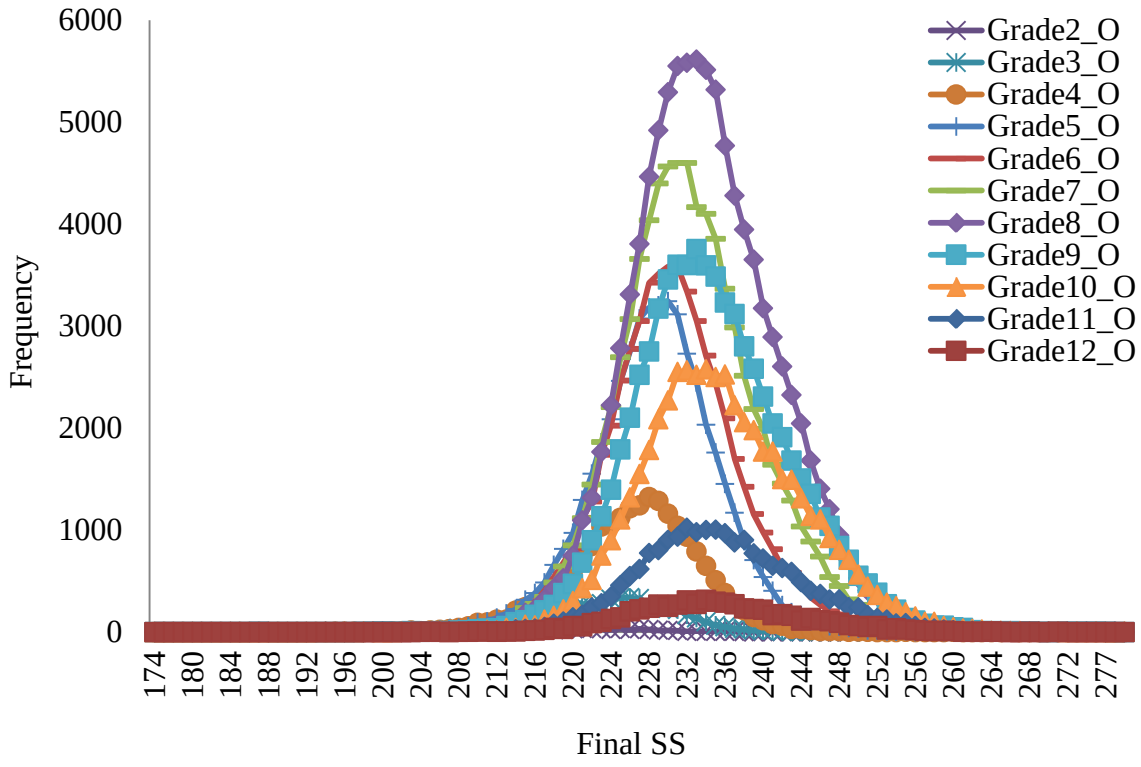

Figure 5a. Frequency Distributions of Reading Operational Samples across Grades (Item 2)

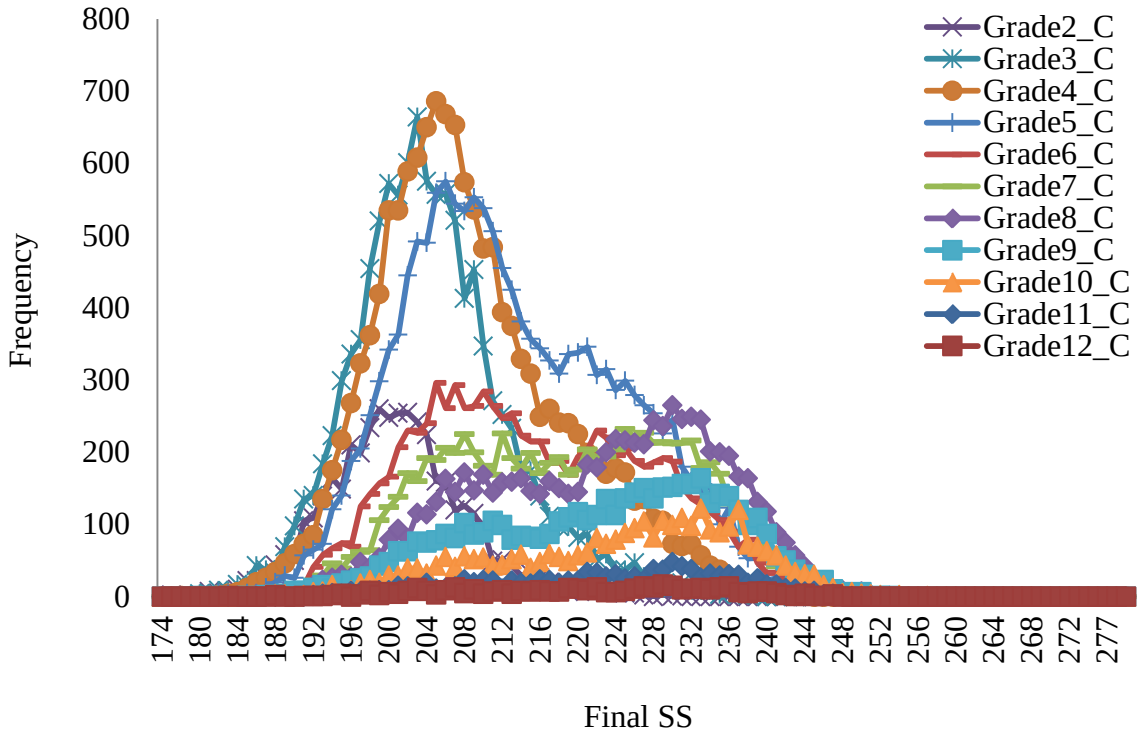

Figure 5b. Frequency Distributions of Reading Calibration Samples across Grades (Item 2)

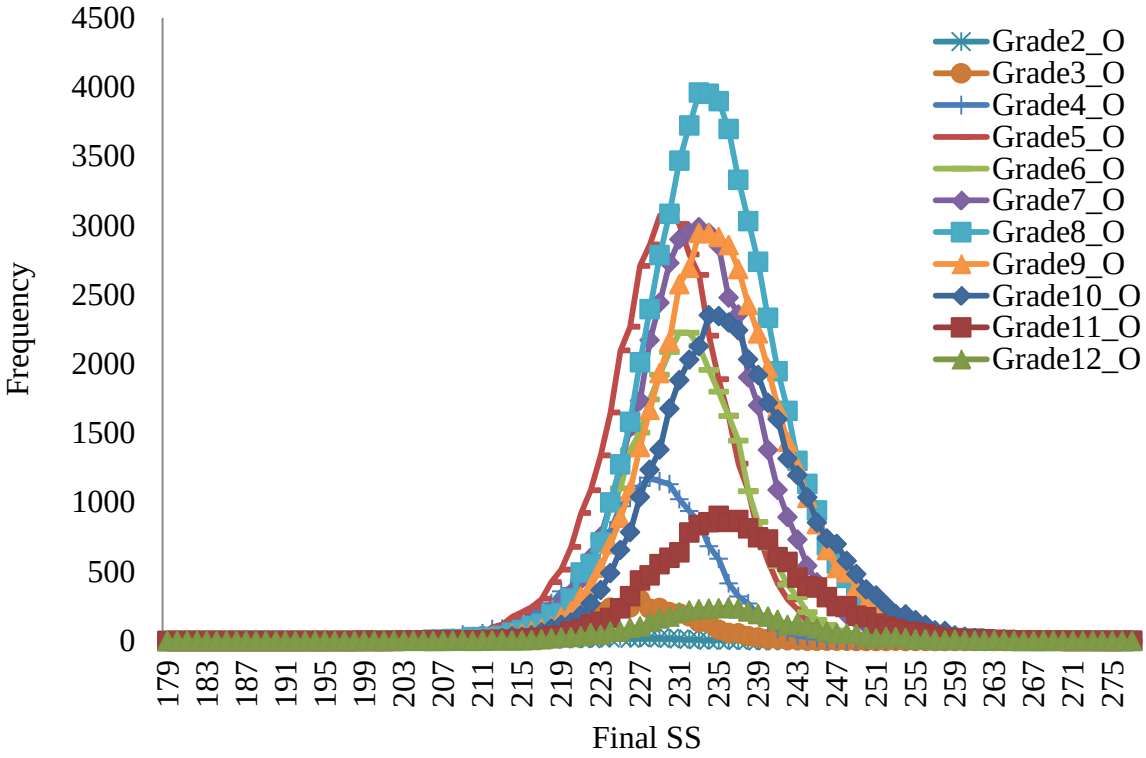

Figure 6a. Frequency Distributions of Reading Operational Samples across Grades (Item 3)

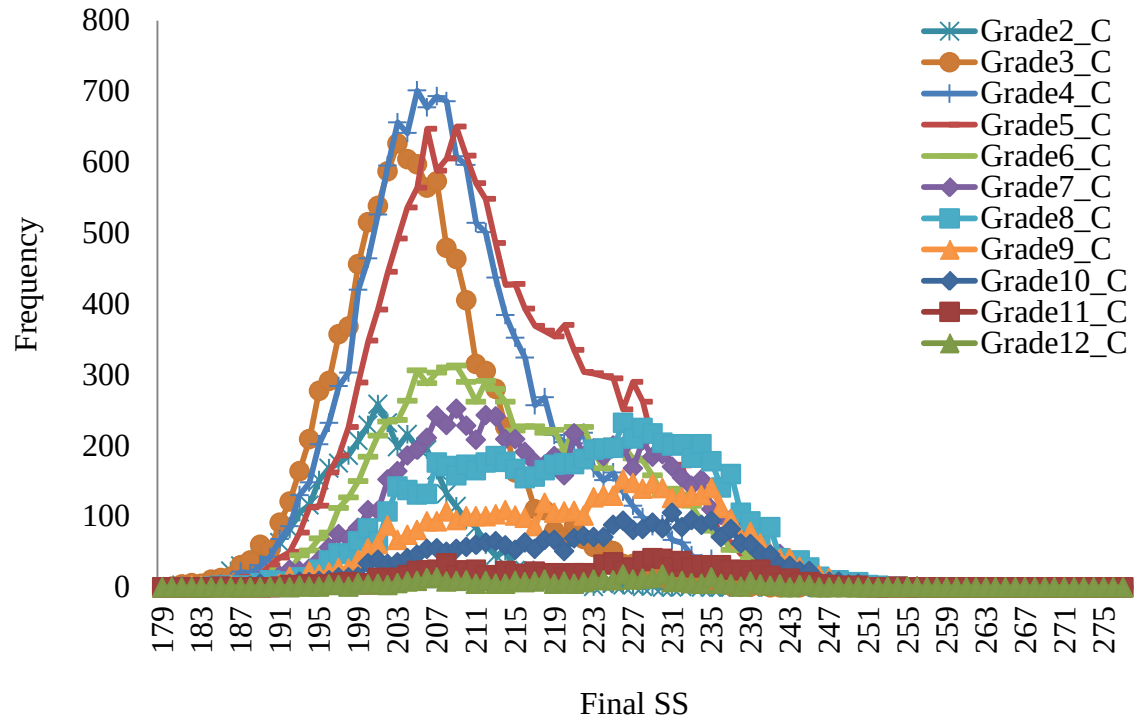

Figure 6b. Frequency Distributions of Reading Calibration Samples across Grades (Item 3)

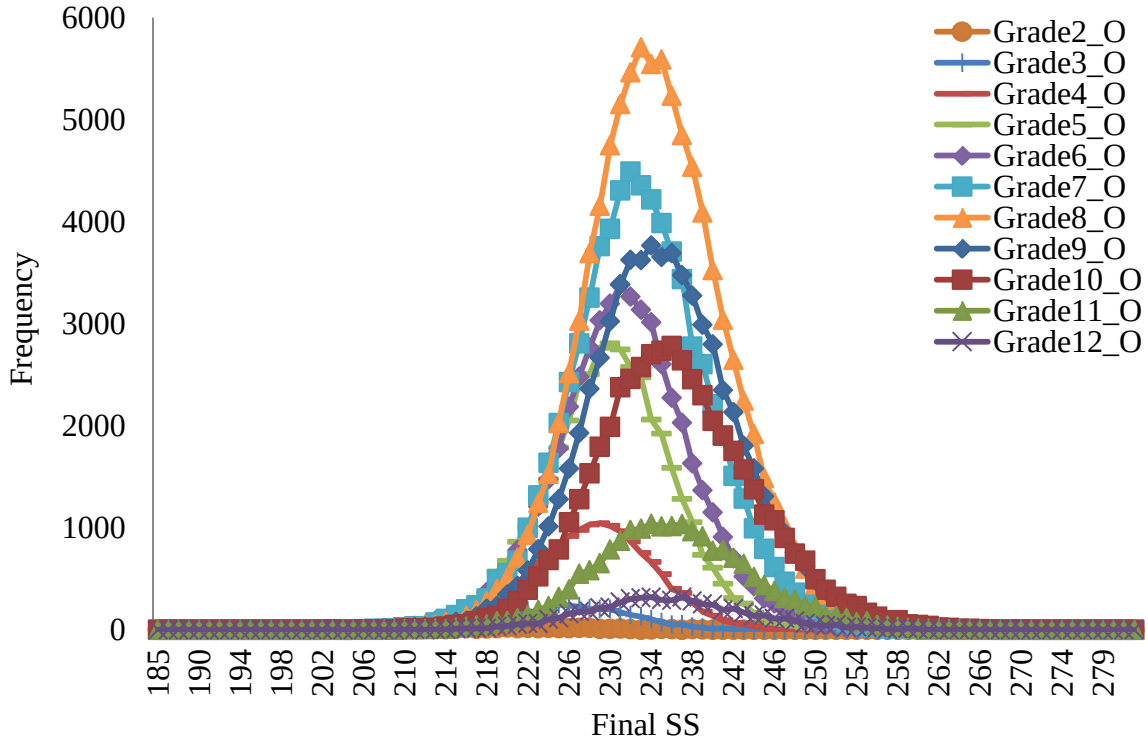

Figure 7a. Frequency Distributions of Reading Operational Samples across Grades (Item 4)

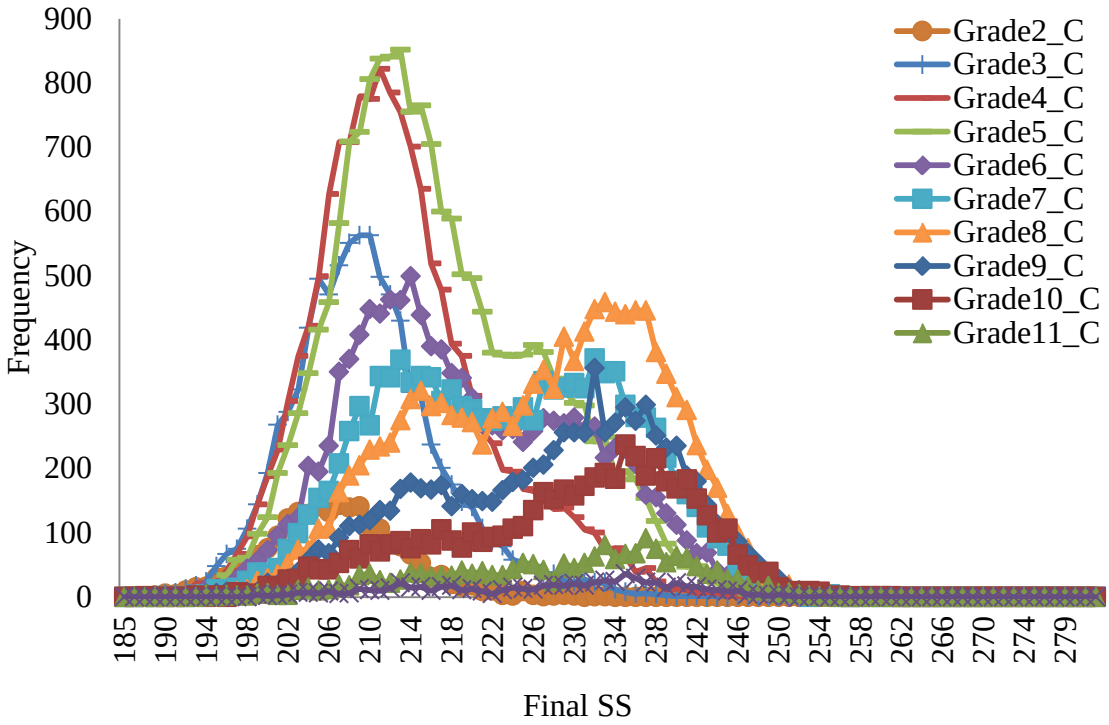

Figure 7b. Frequency Distributions of Reading Calibration Samples across Grades (Item 4)

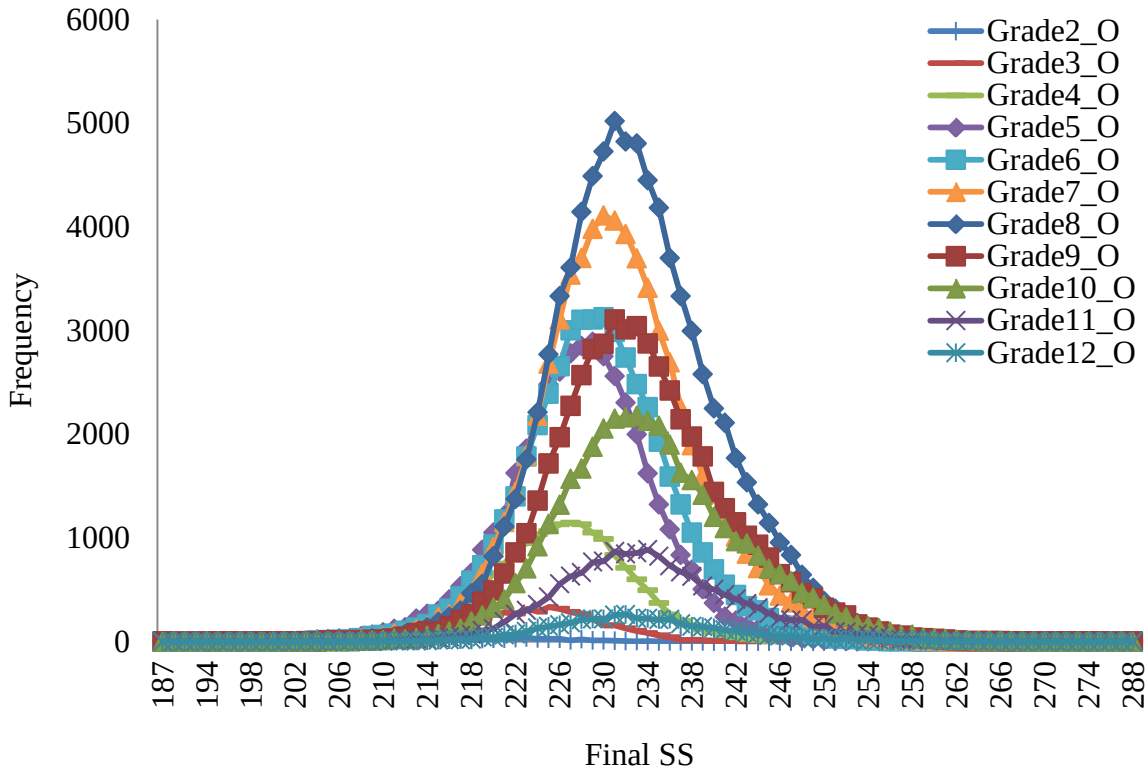

Figure 8a. Frequency Distributions of Reading Operational Samples across Grades (Item 5)

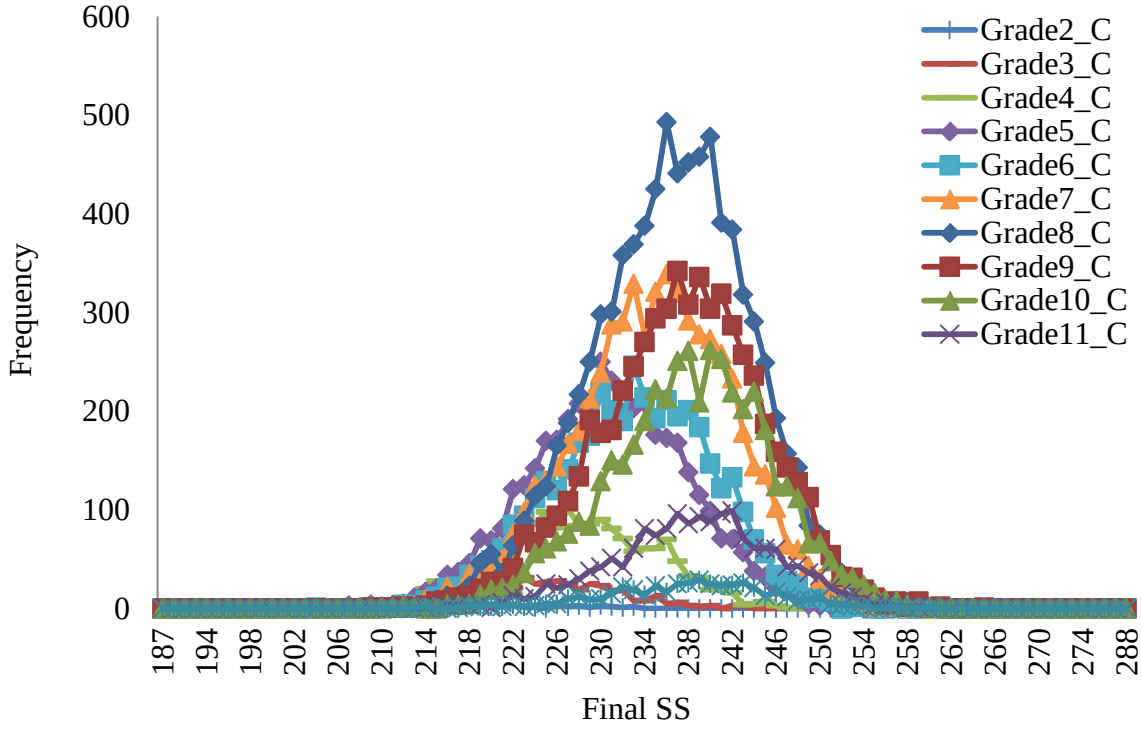

Figure 8b. Frequency Distributions of Reading Calibration Samples across Grades (Item 5)

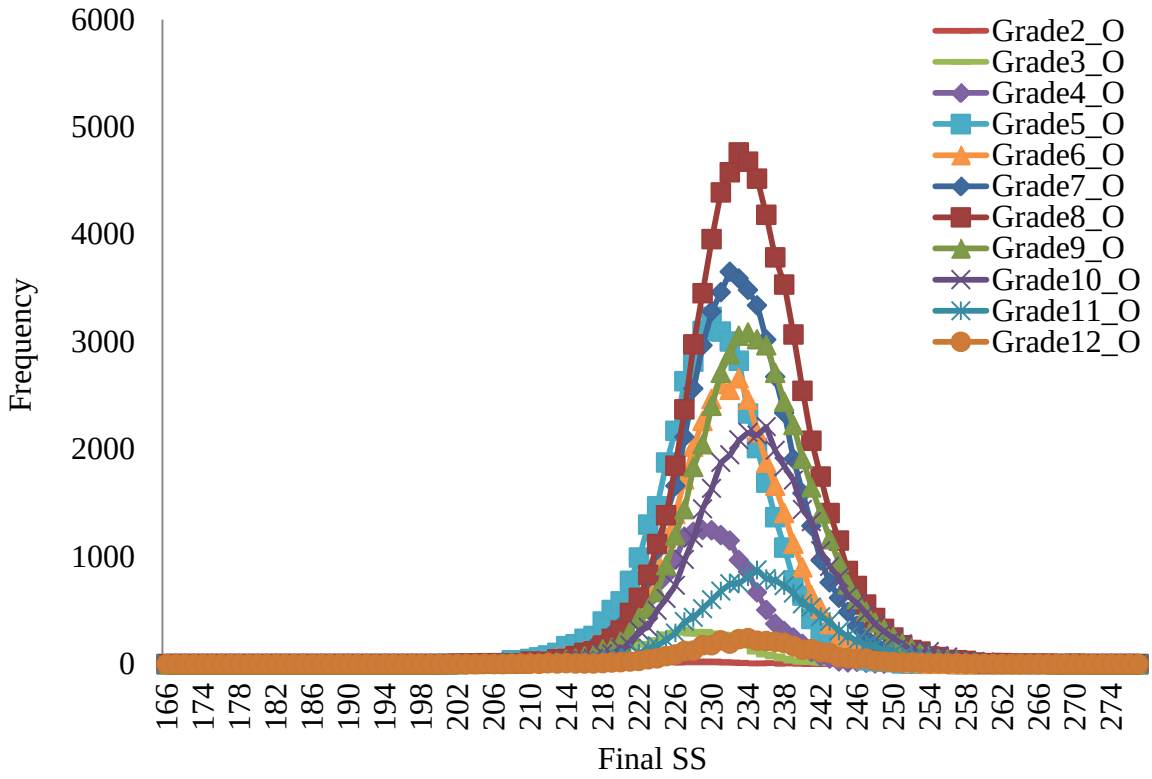

Figure 9a. Frequency Distributions of Reading Operational Samples across Grades (Item 6)

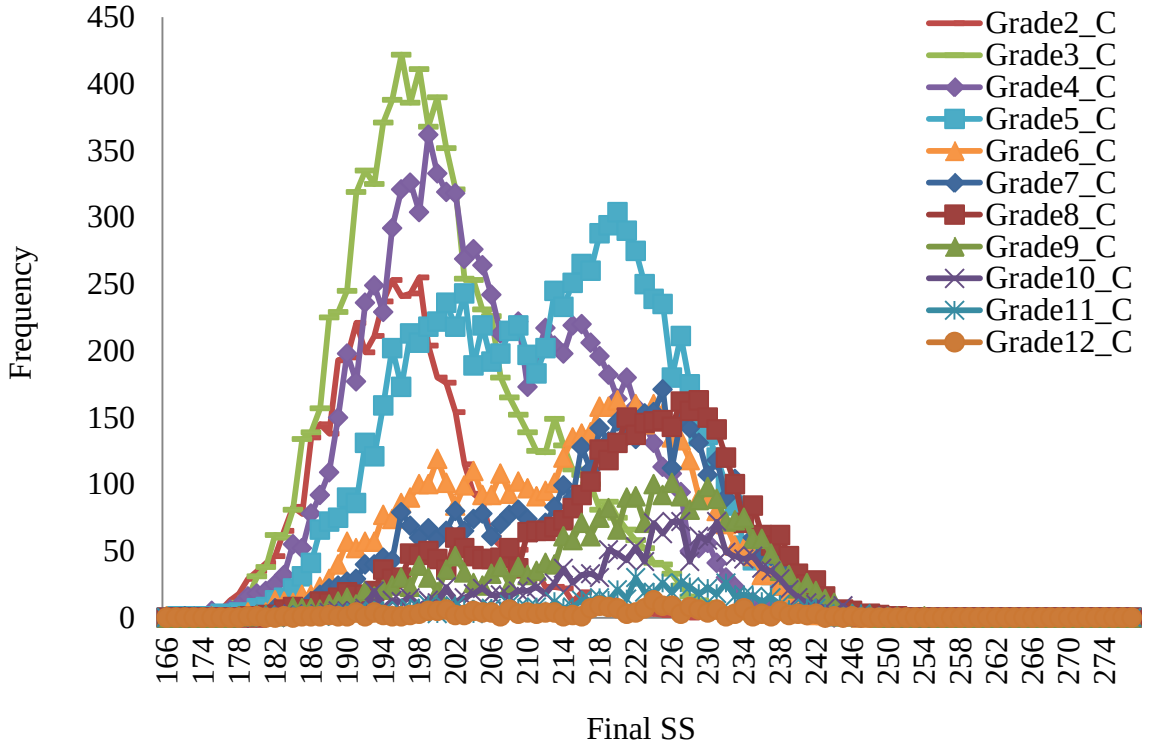

Figure 9b. Frequency Distributions of Reading Calibration Samples across Grades (Item 6)

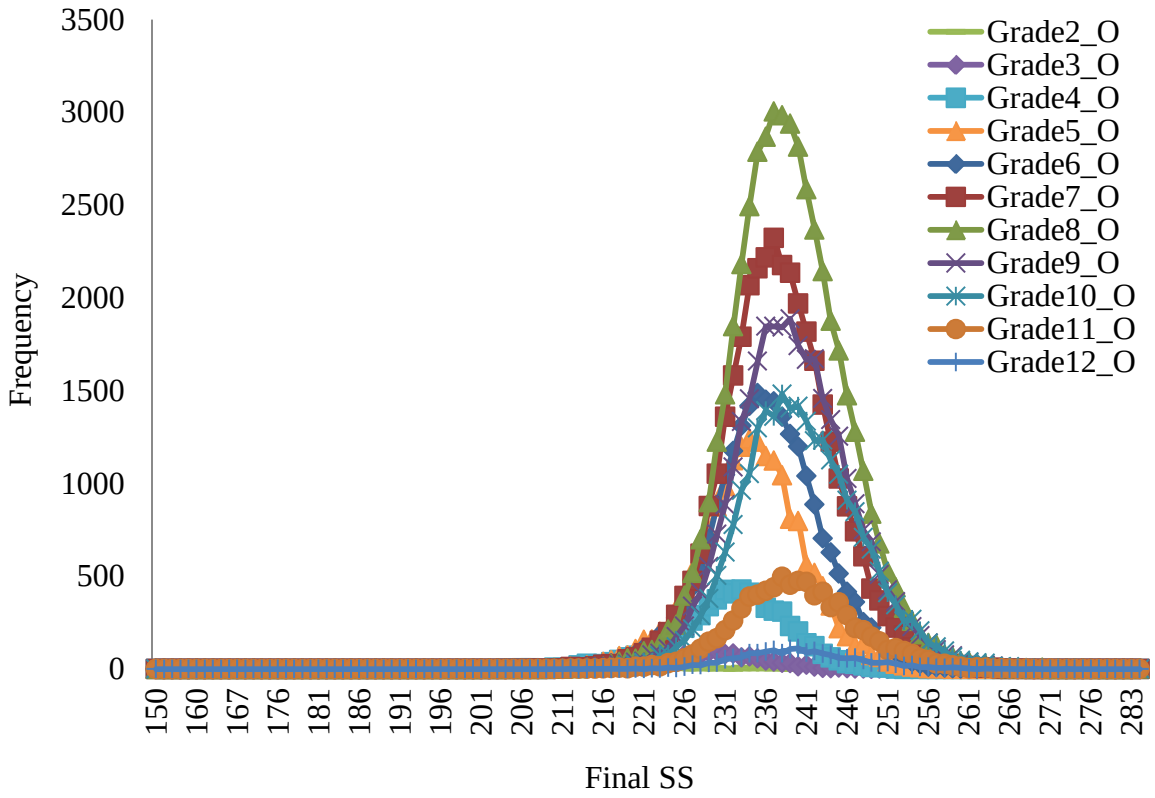

Figure 10a. Frequency Distributions of Reading Operational Samples across Grades (Item 7)

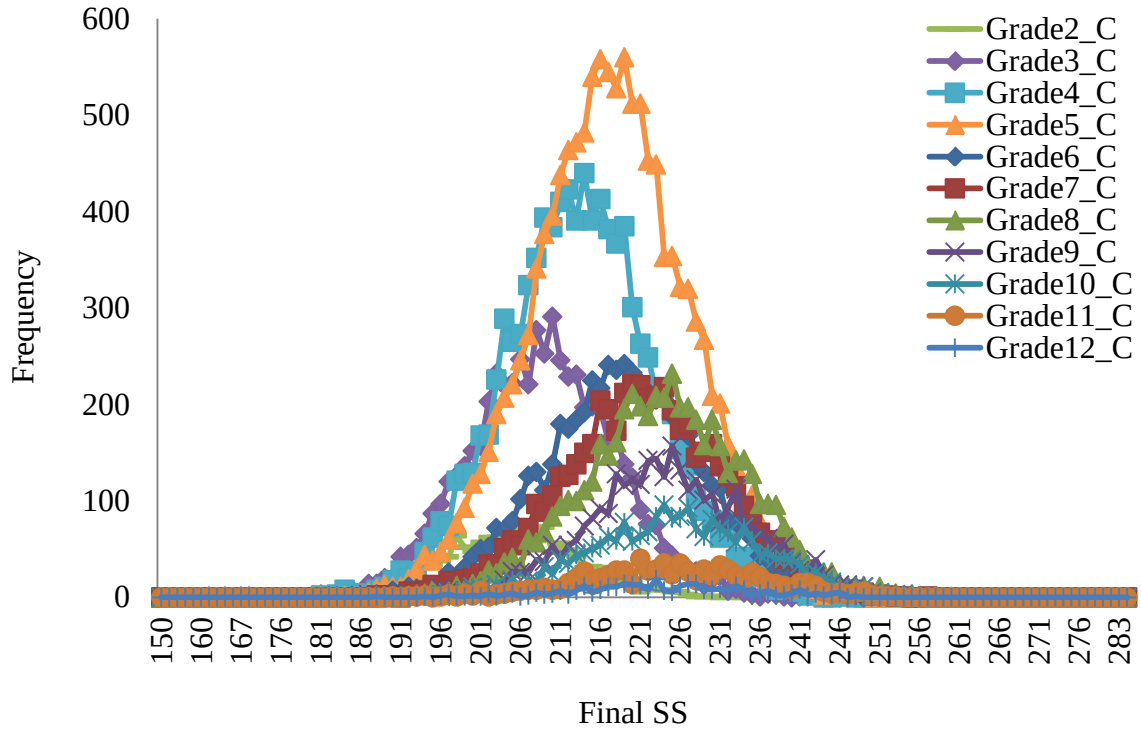

Figure 10b. Frequency Distributions of Reading Calibration Samples across Grades (Item 7)

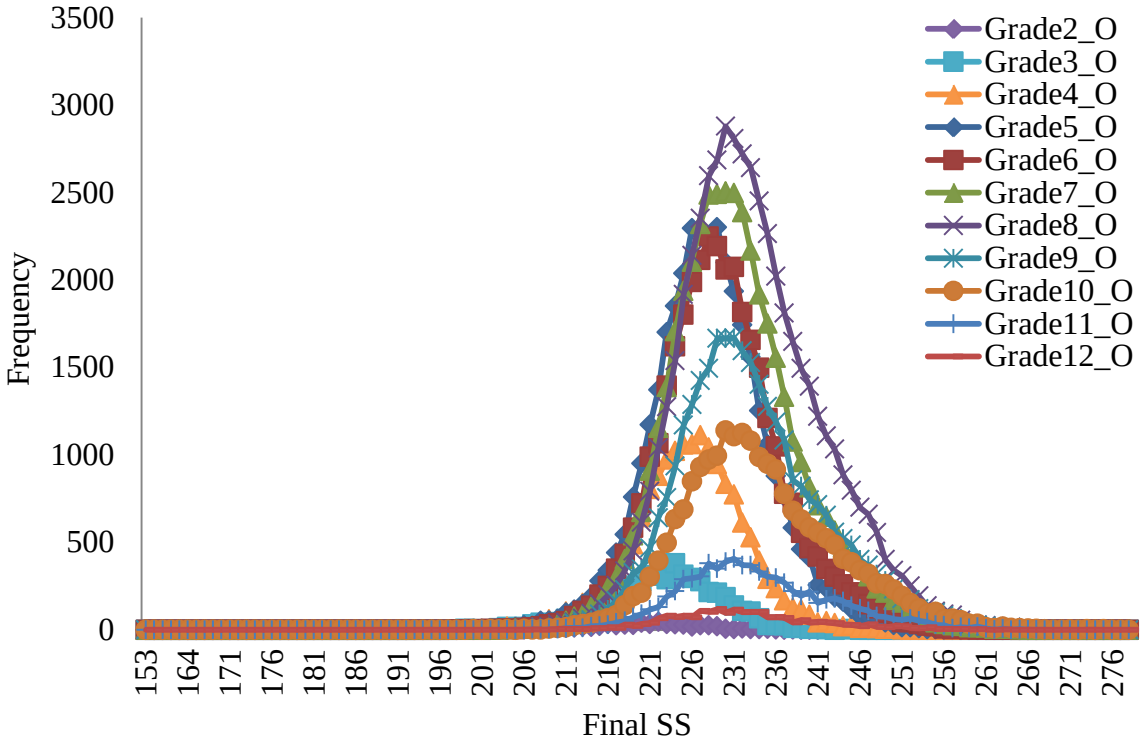

Figure 11a. Frequency Distributions of Reading Operational Samples across Grades (Item 8)

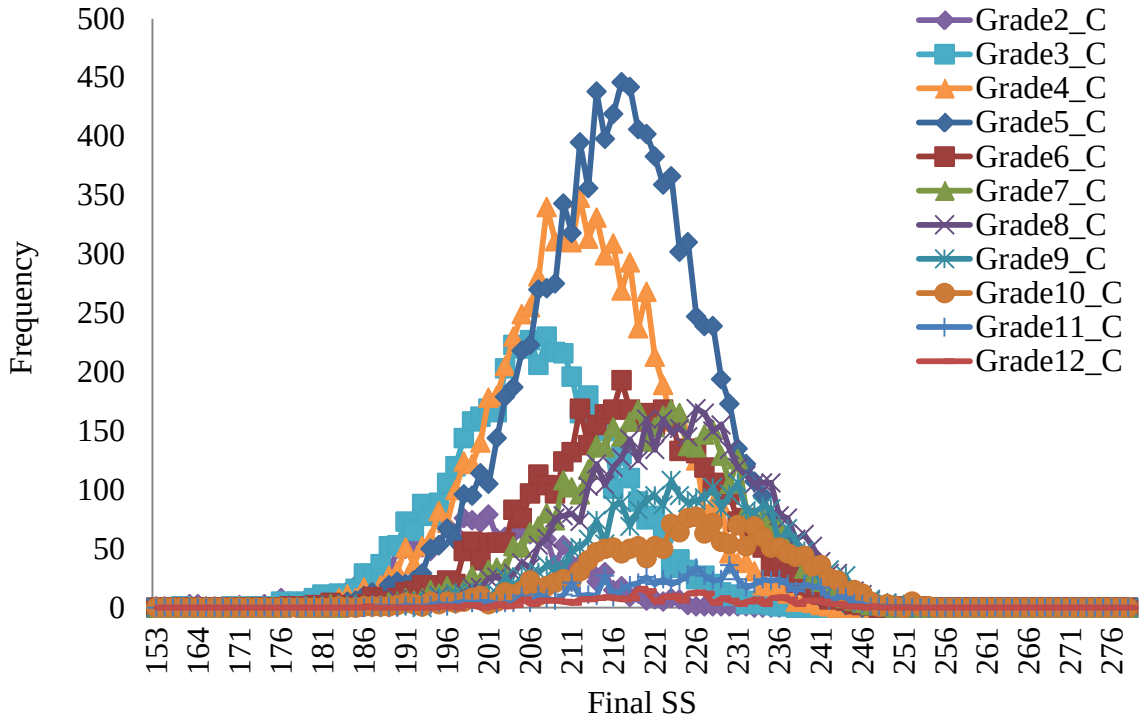

Figure 11b. Frequency Distributions of Reading Calibration Samples across Grades (Item 8)

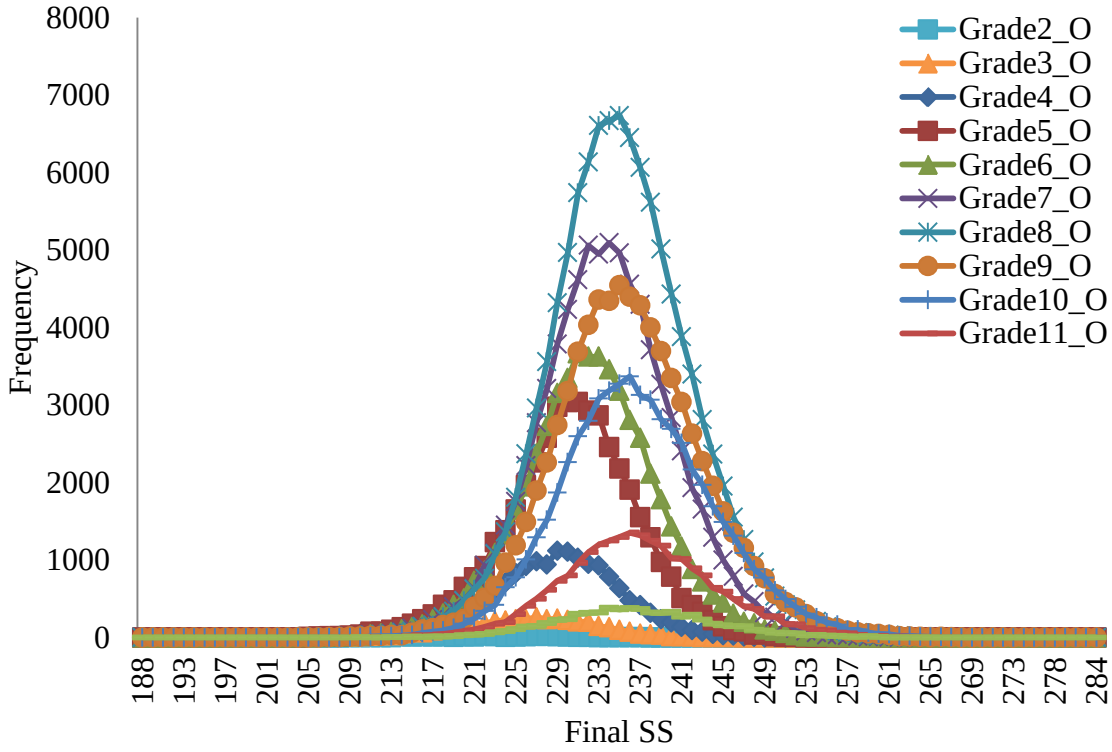

Figure 12a. Frequency Distributions of Reading Operational Samples across Grades (Item 9)

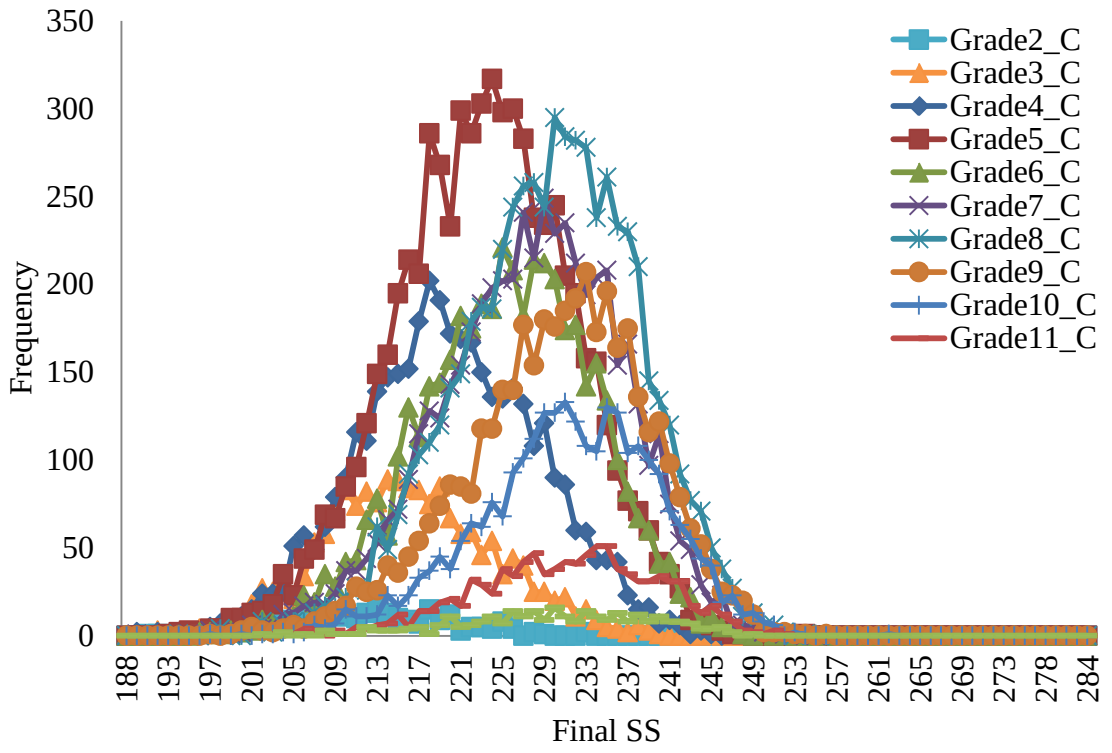

Figure 12b. Frequency Distributions of Reading Calibration Samples across Grades (Item 9)

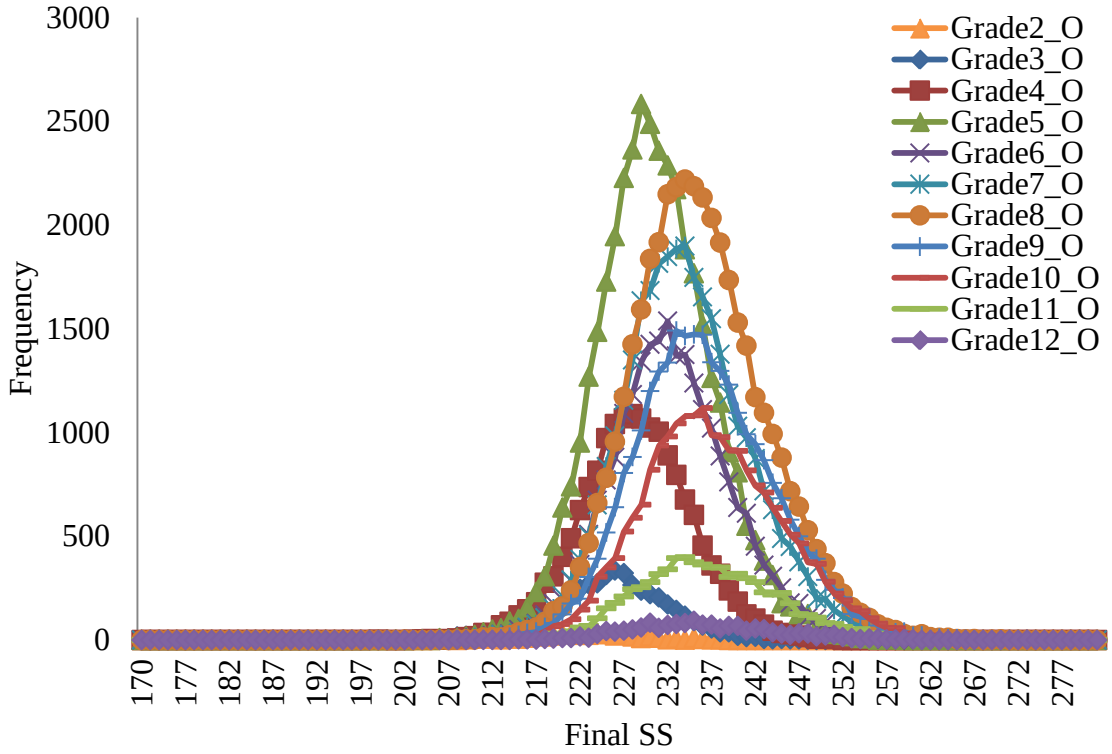

Figure 13a. Frequency Distributions of Reading Operational Samples across Grades (Item 10)

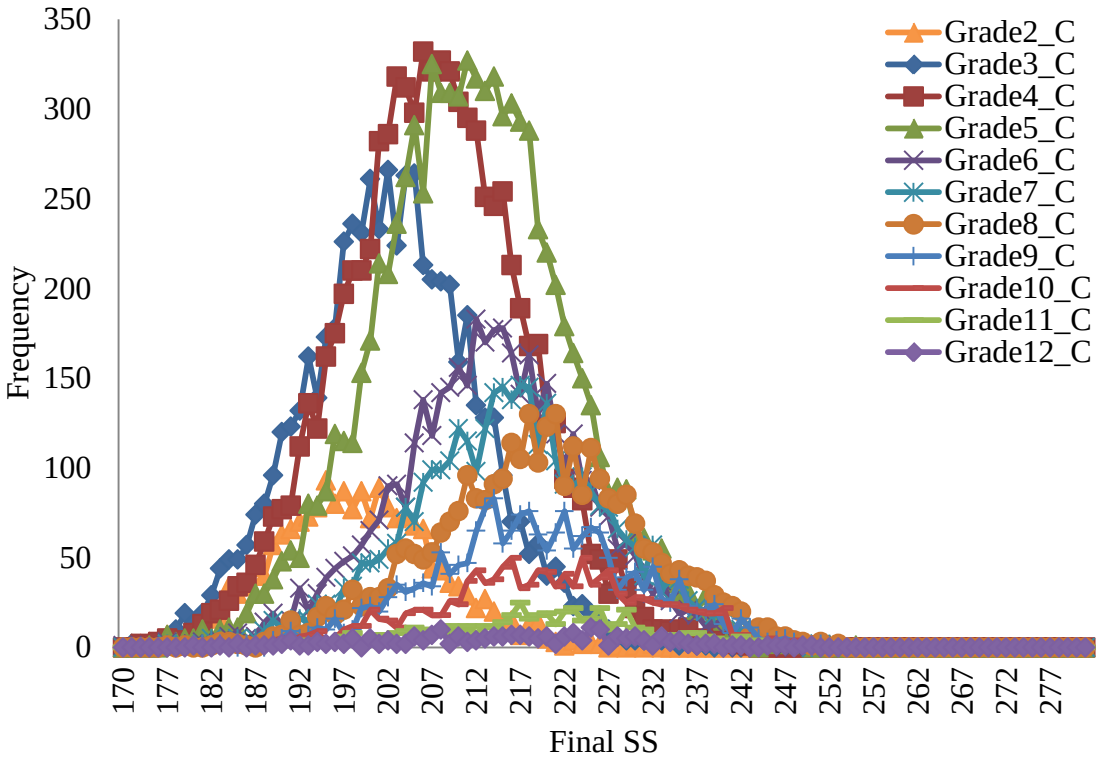

Figure 13b. Frequency Distributions of Reading Calibration Samples across Grades (Item 10)

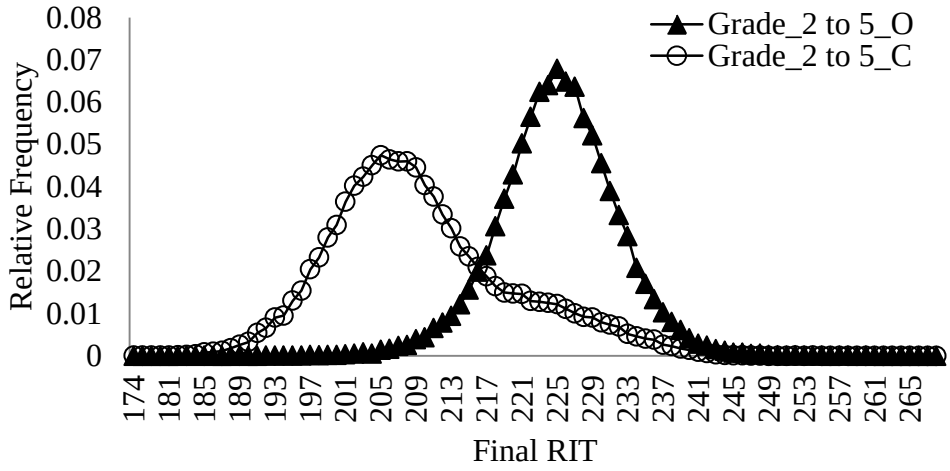

Figure 14a. Empirical Distributions of Reading Operational and Calibration Samples of Grades 2 to 5 for Item 1

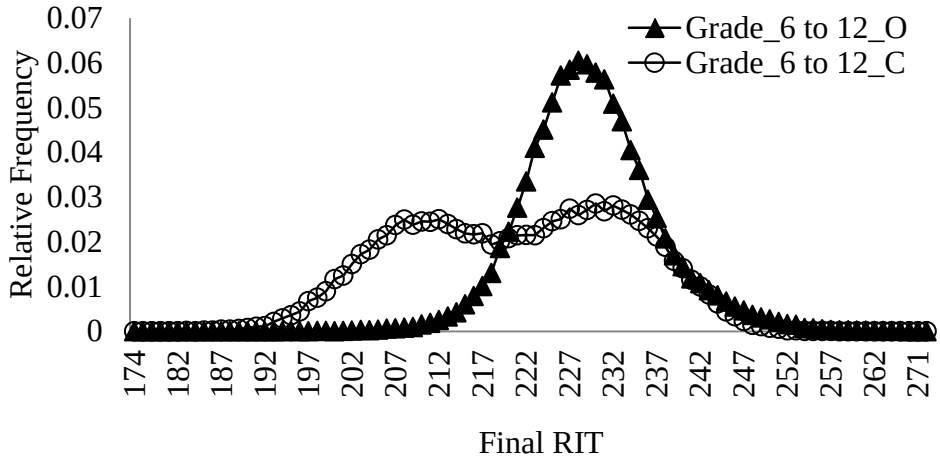

Figure 14b. Empirical Distributions of Reading Operational and Calibration Samples of Grades 6 to 12 for Item 1

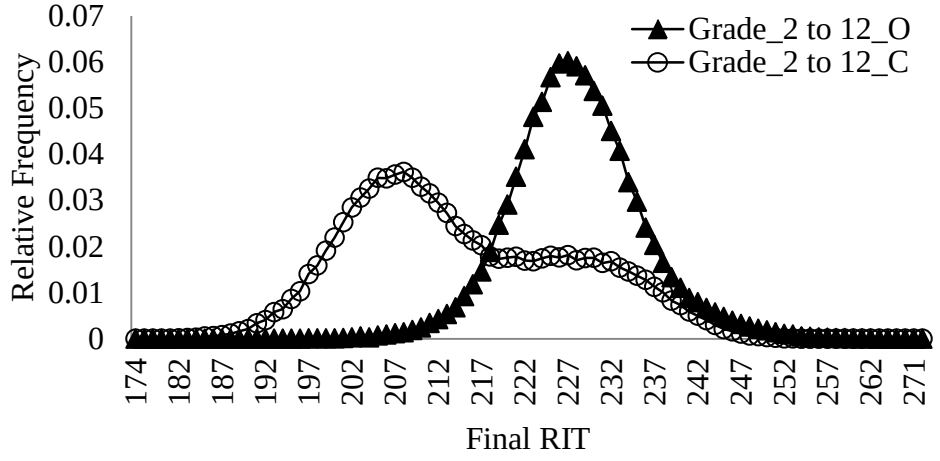

Figure 14c. Empirical Distributions of Reading Operational and Calibration Samples of Grades 2 to 12 for Item 1

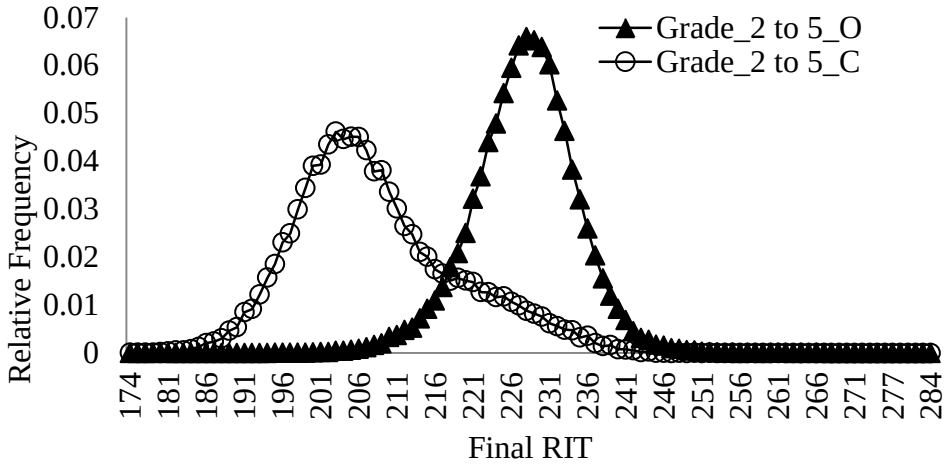

Figure 15a. Empirical Distributions of Reading Operational and Calibration Samples of Grades 2 to 5 for Item 2

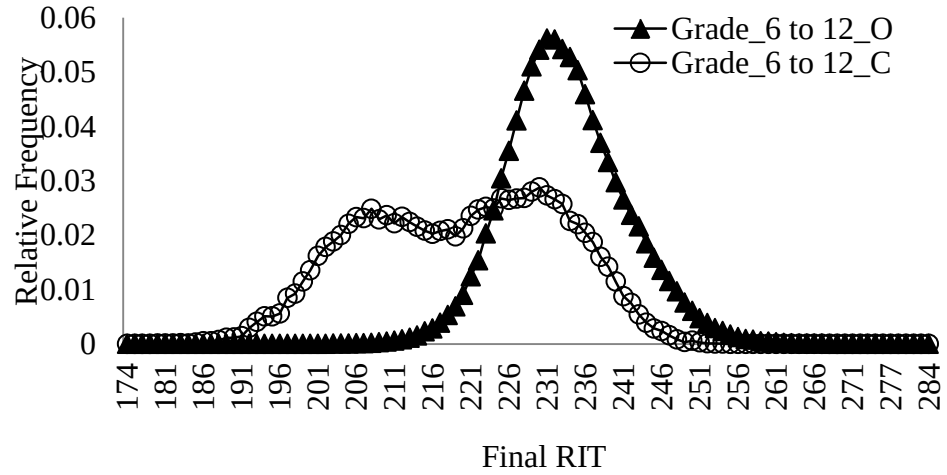

Figure 15b. Empirical Distributions of Reading Operational and Calibration Samples of Grades 6 to 12 for Item 2

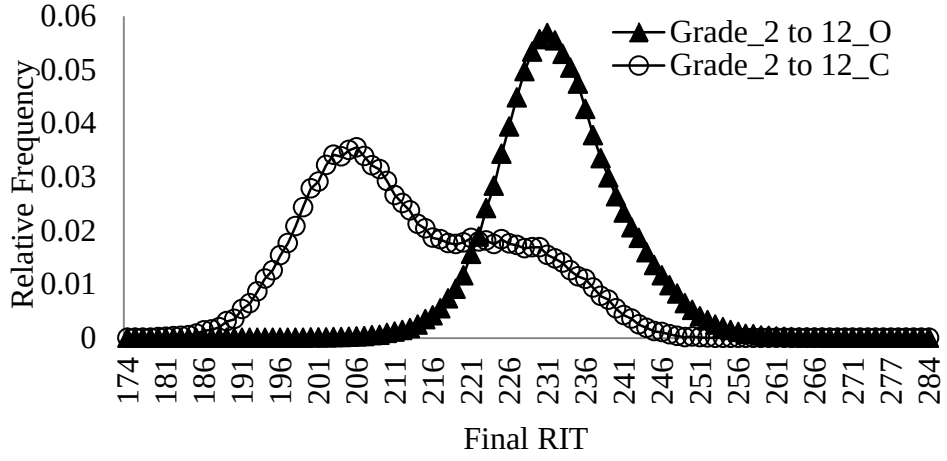

Figure 15c. Empirical Distributions of Reading Operational and Calibration Samples of Grades 2 to 12 for Item 2

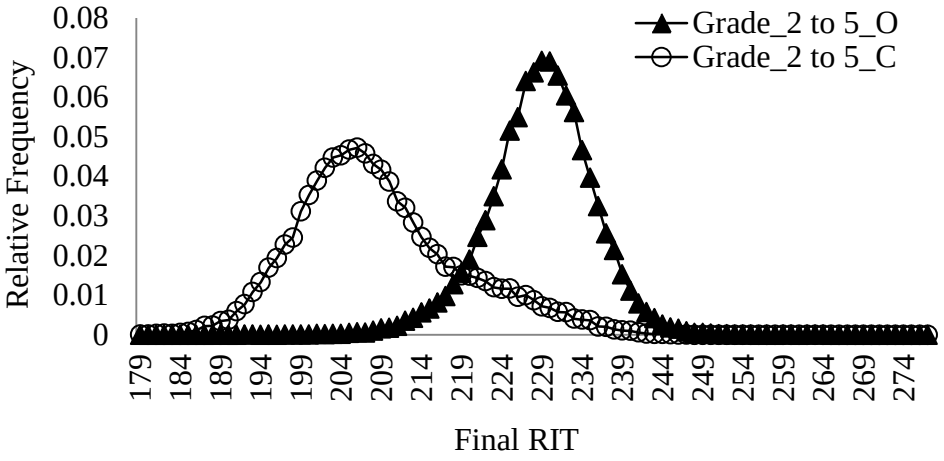

Figure 16a. Empirical Distributions of Reading Operational and Calibration Samples of Grades 2 to 5 for Item 3

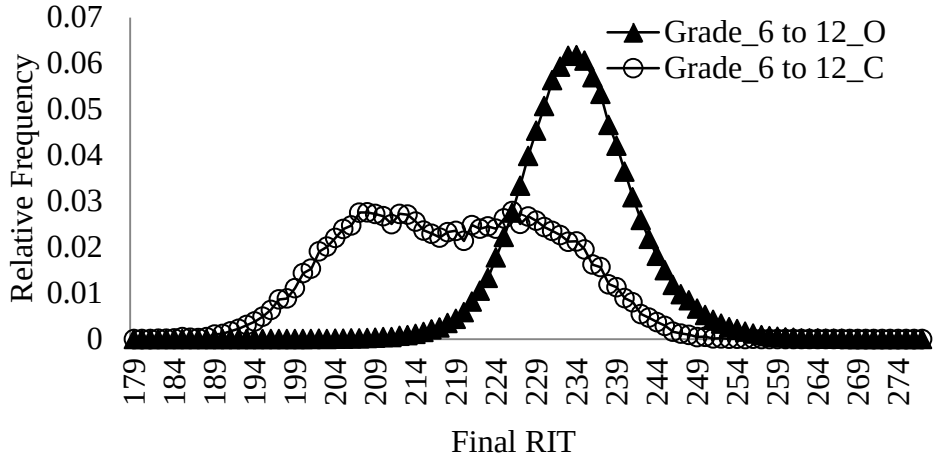

Figure 16b. Empirical Distributions of Reading Operational and Calibration Samples of Grades 6 to 12 for Item 3

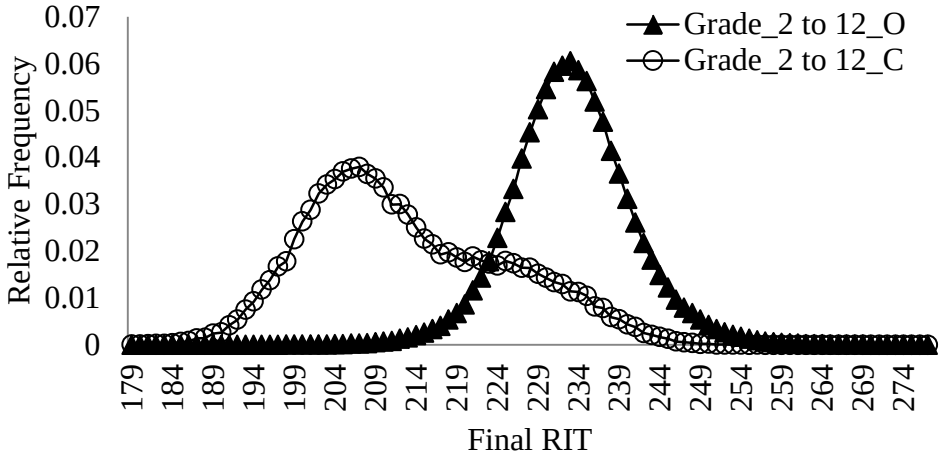

Figure 16c. Empirical Distributions of Reading Operation*al* and Calibration Samples of Grades 2 to 12 for Item 3

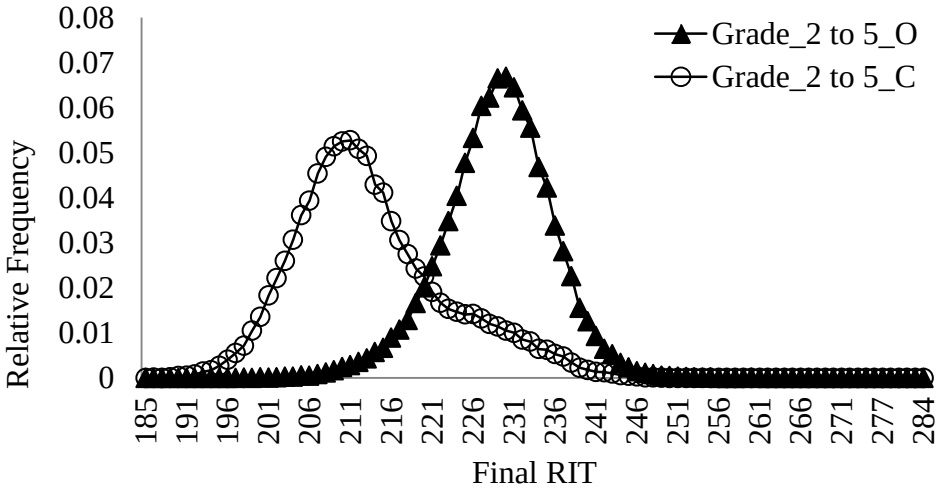

Figure 17a. Empirical Distributions of Reading Operational and Calibration Samples of Grades 2 to 5 for Item 4

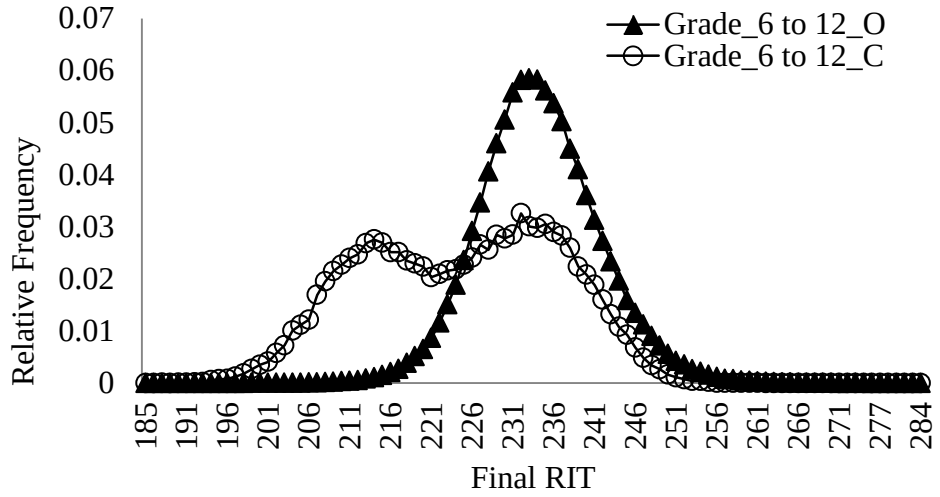

Figure 17b. Empirical Distributions of Reading Operational and Calibration Samples of Grades 6 to 12 for Item 4

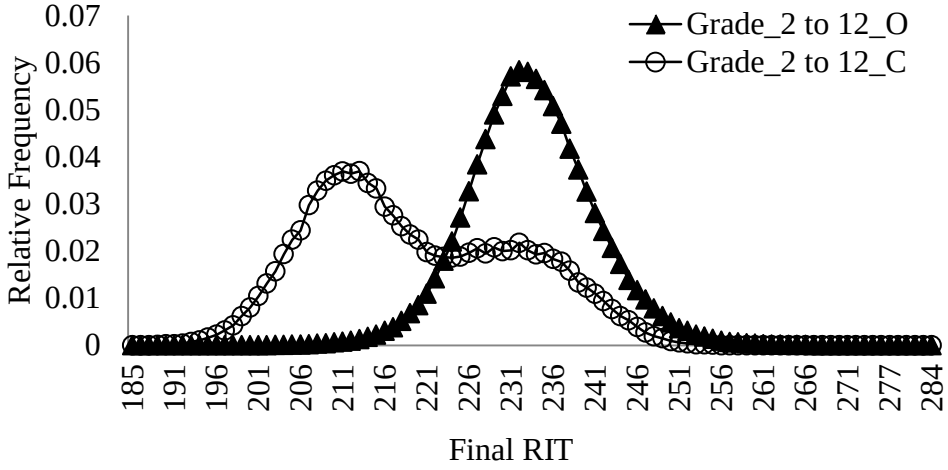

Figure 17c. Empirical Distributions of Reading Operational and Calibration Samples of Grades 2 to 12 for Item 4

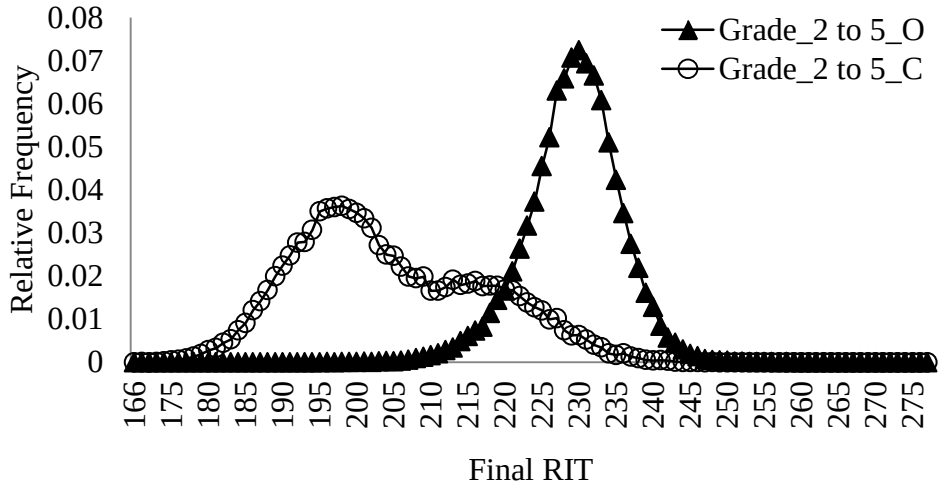

Figure 18a. Empirical Distributions of Reading Operational and Calibration Samples of Grades 2 to 5 for Item 6

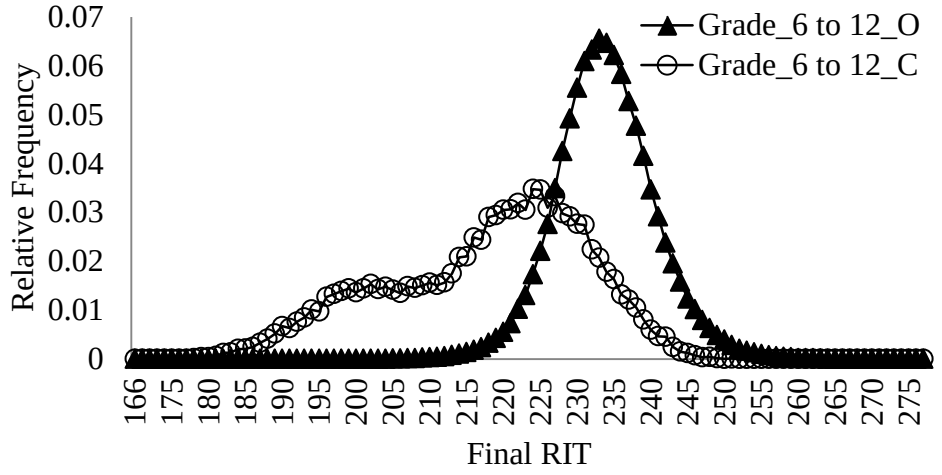

Figure 18b. Empirical Distributions of Reading Operational and Calibration Samples of Grades 6 to 12 for Item 6

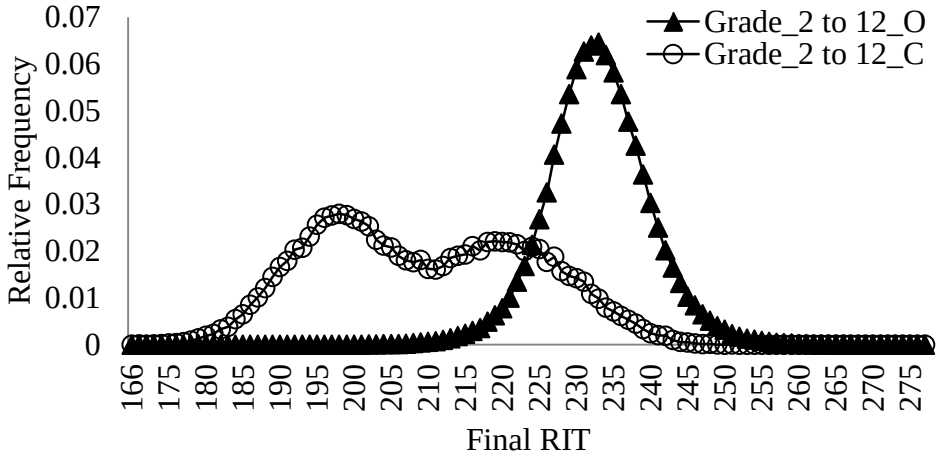

Figure 18c. Empirical Distributions of Reading Operational and Calibration Samples of Grades 2 to 12 for Item 6

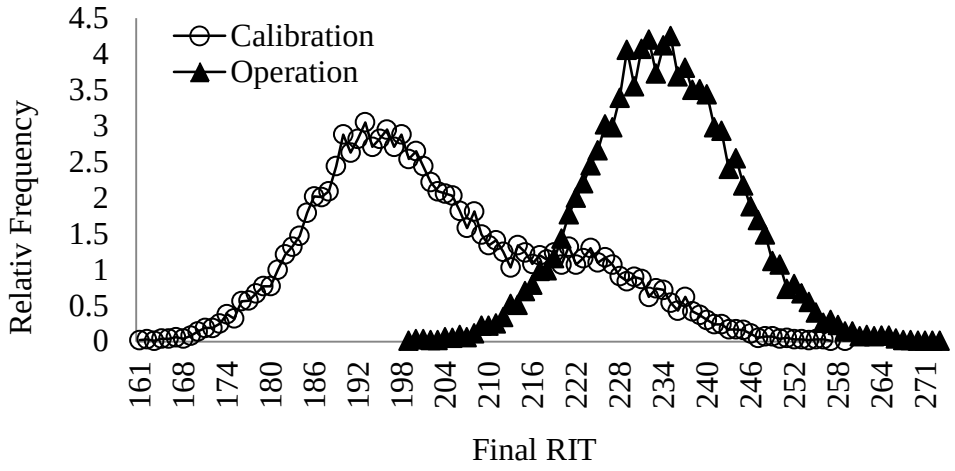

Figure 19a. Simulated Type 1 (Grades 2 to 5) Distributions of Reading Operational and Calibration Samples

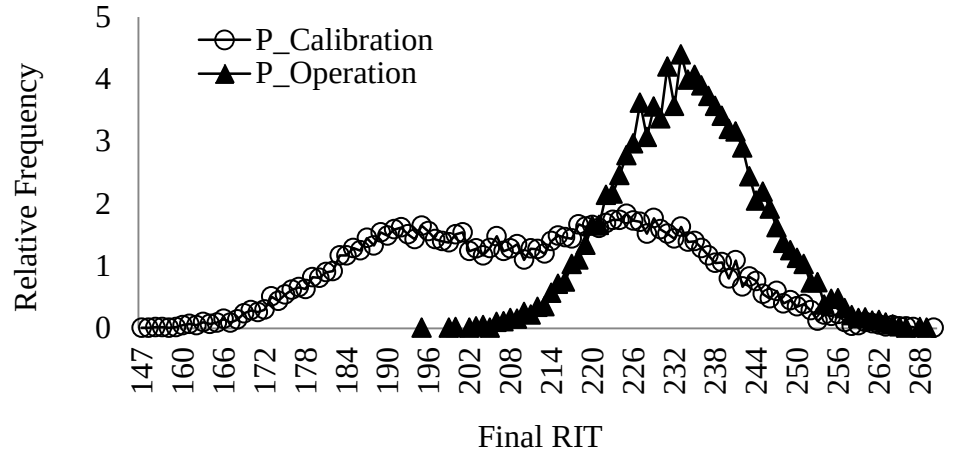

Figure 19b. Simulated Type 2 (Grades 6 to 12) Distributions of Reading Operational and Calibration Samples

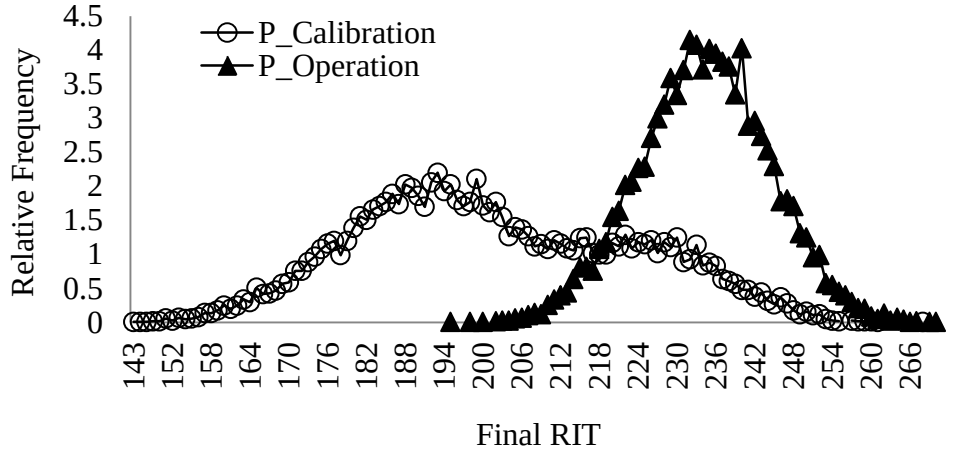

Figure 19c. Simulated Type 3 (Grades 2 to 12) Distributions of Reading Operational and Calibration Samples

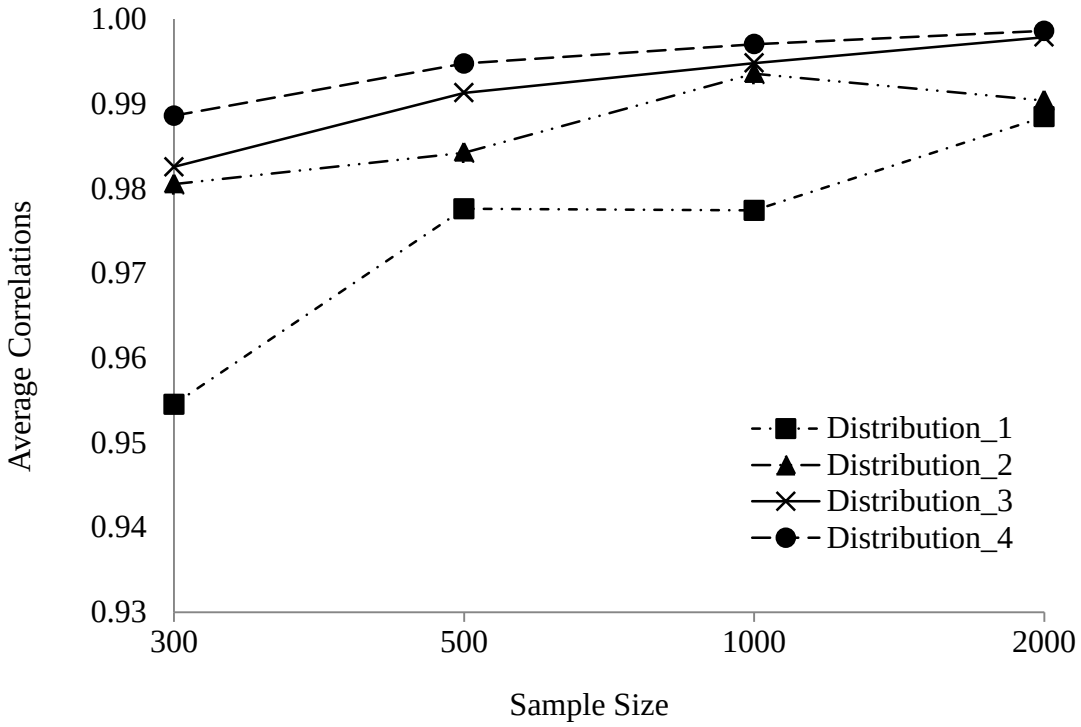

Figure 20. Average Correlations of Item Parameter Estimates between Operational and Calibration Samples

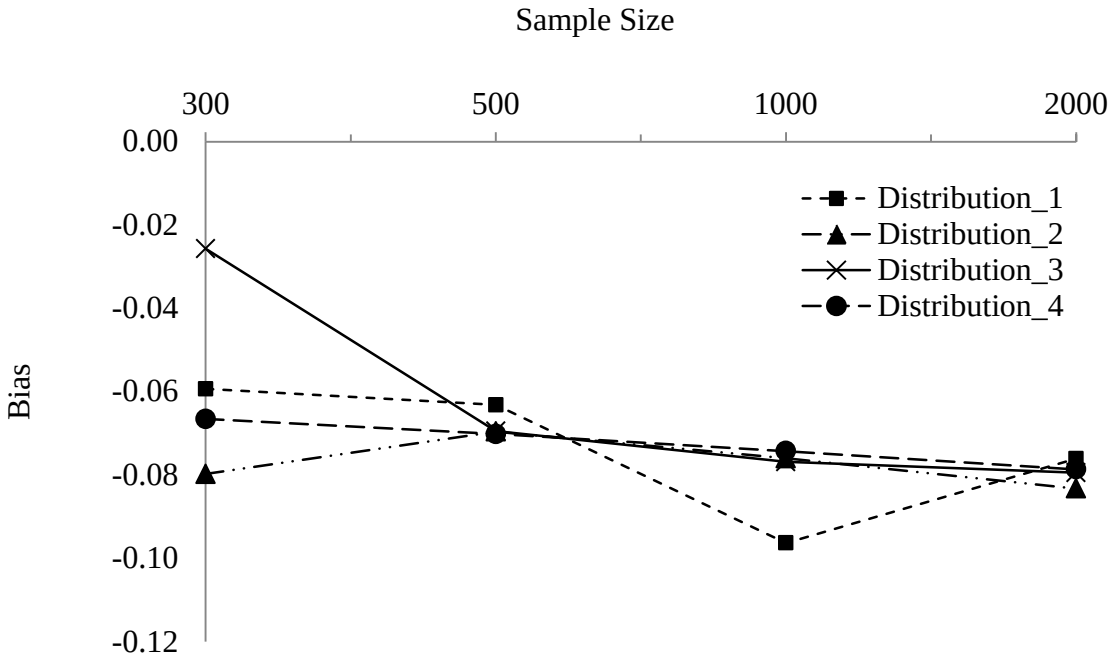

Figure 21. Average Bias of Item Parameter Estimates By Type of Distribution and Sample Size

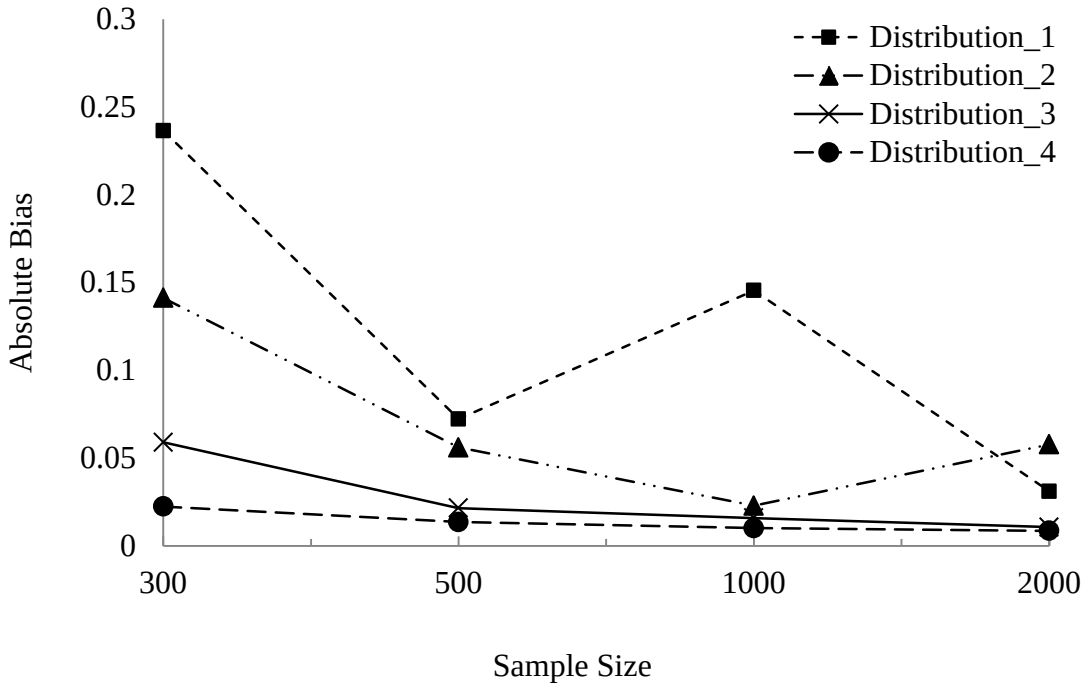

Figure 22. Average Abias of Item Parameter Estimates By Type of Distribution and Sample Size

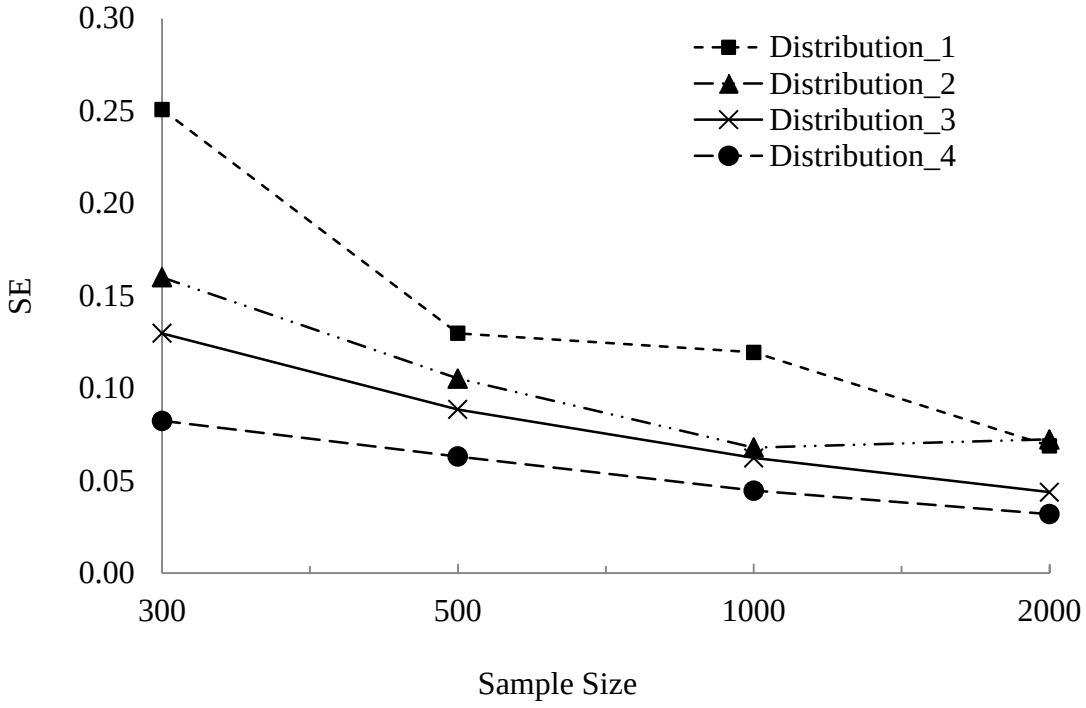

Figure 23. Average SE of Item Parameter Estimates By Type of Distribution and Sample Size

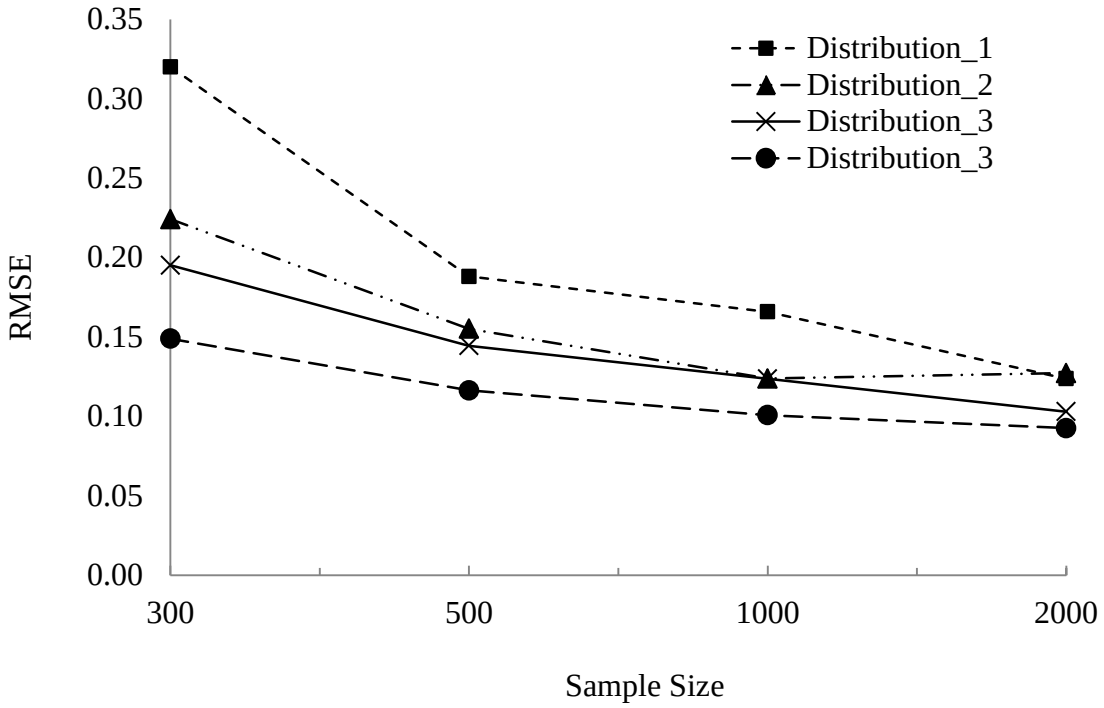

Figure 24. Average RMSE of Item Parameter Estimates By Type of Distribution and Sample Size

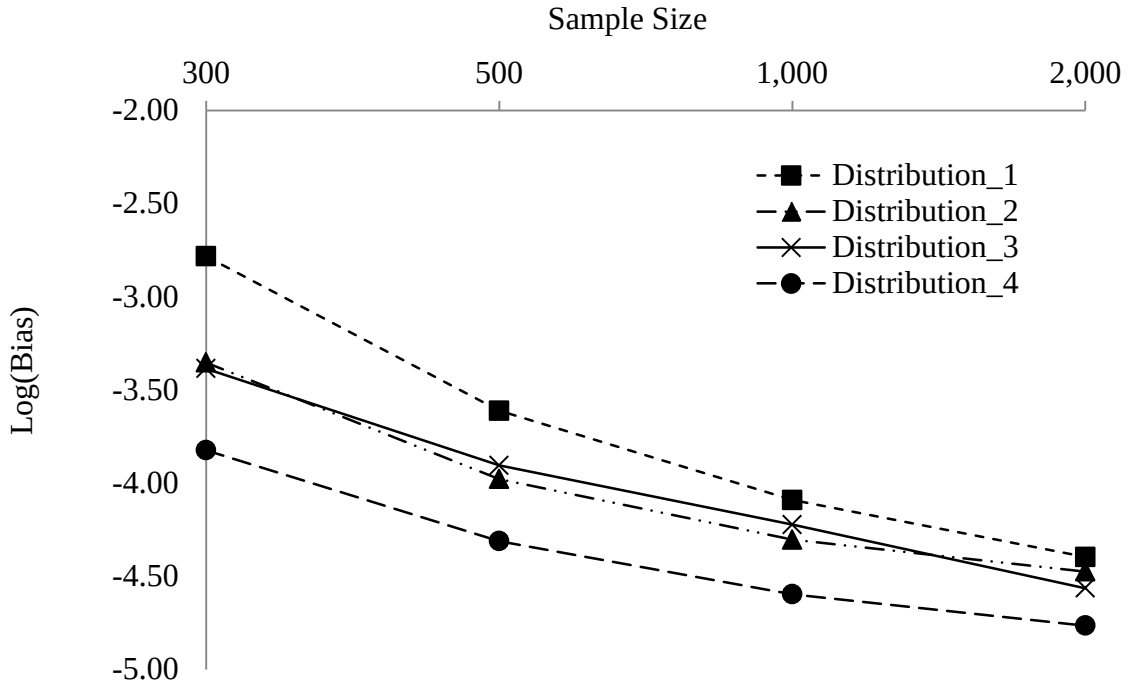

Figure 25. Log(Abias) of Item Parameter Estimates By Type of Distribution and Sample Size

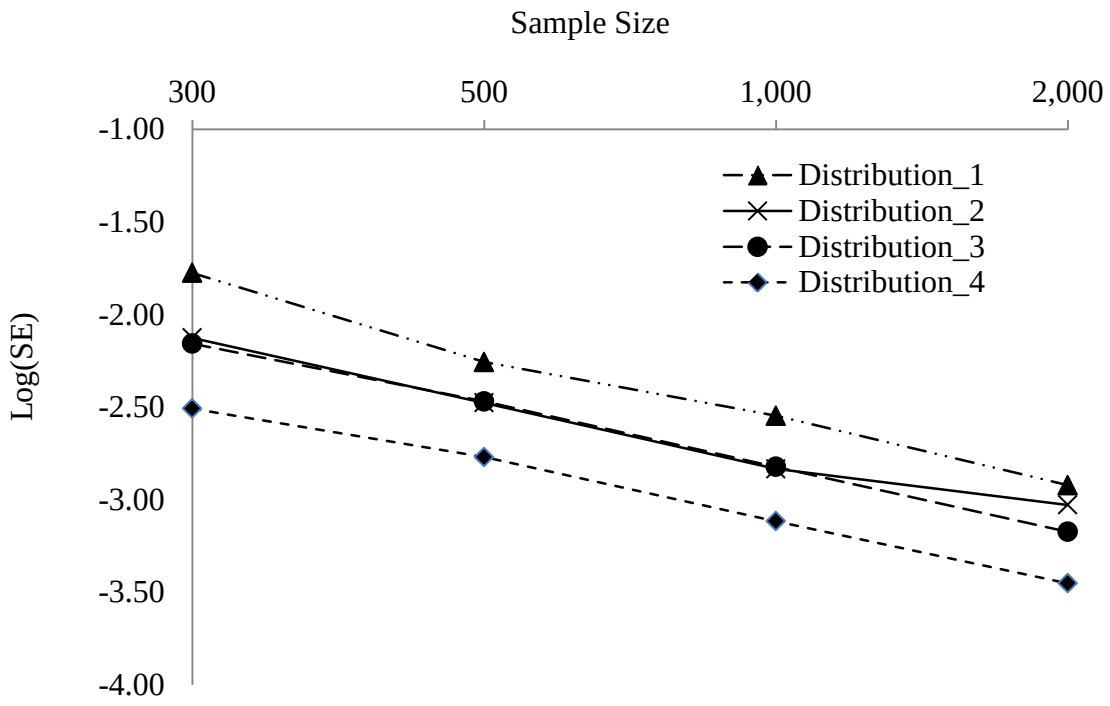

Figure 26. Log(SE) of Item Parameter Estimates By Type of Distribution and Sample Size

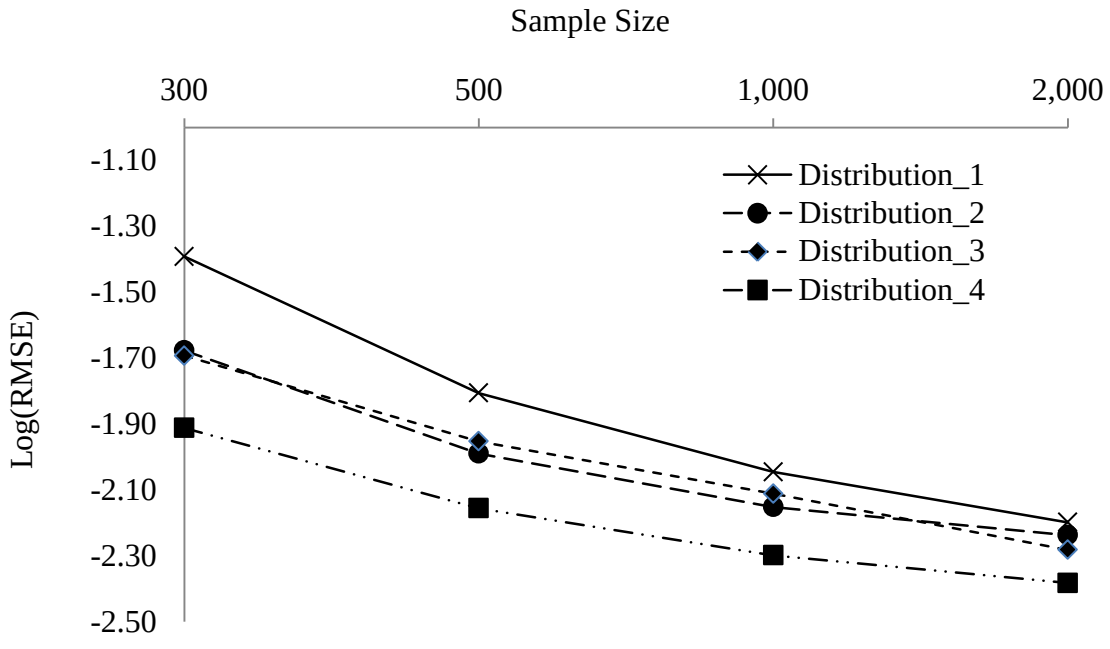

Figure 27. Log(RMSE) of Item Parameter Estimates By Type of Distribution and Sample Size

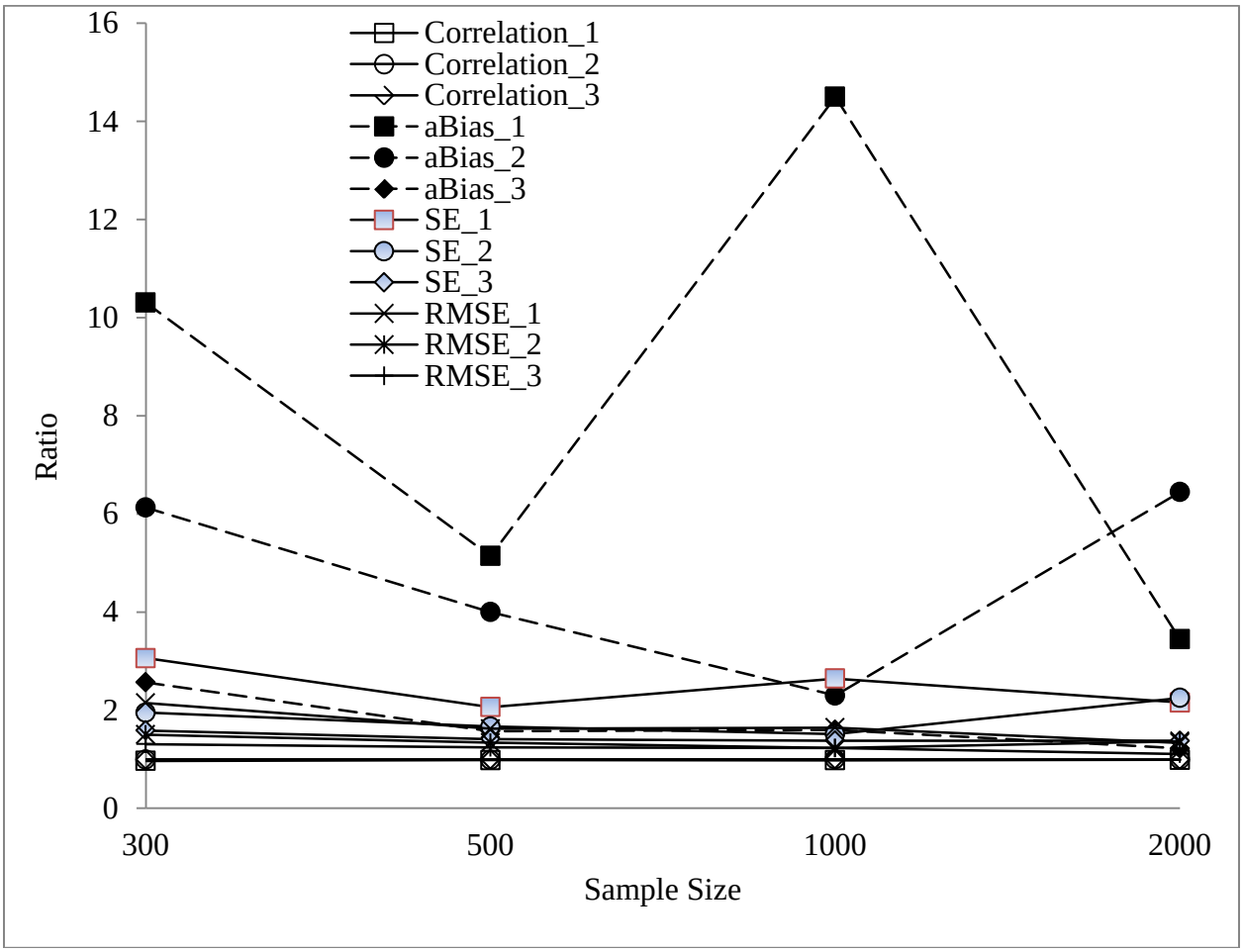

Figure 28. Ratio of Dependent Variables with Different Type Distributions (1, 2, 3) Over Base Distribution (4) Across Different Sample Sizes

Appendix A

Dependent Variables

These criteria are used to examine the effects of the manipulated independent variables described in the last subsection to provide complementary evidence. For each i item, the conditional bias (abias), SE and RMSE of an estimator $\hat{b}$ across N (r=1, 2, … , N) replications can be expressed as following:

$$\text{Bias}(\hat{b}_i) = E(\hat{b}_i) - b_i = \frac{1}{N}\sum_{r=1}^{N}\hat{b}_{ri} - b_i = \frac{1}{N}\sum_{r=1}^{N}\hat{b}_{ri} - \frac{1}{N}\sum_{r=1}^{N} b_i = \frac{1}{N}\sum_{r=1}^{N}(\hat{b}_{ri} - b_i) \quad (1)$$

$$\text{Abias}(\hat{b}_i) = |E(\hat{b}_i) - b_i| = \frac{1}{N}\sum_{r=1}^{N}|\hat{b}_{ri} - b_i| \quad (2)$$

$$SE(\hat{b}_i) = \sqrt{Var(\hat{b}_i)} = \sqrt{E[(\hat{b}_i - E(\hat{b}_i))^2]} = \sqrt{\frac{1}{N}\sum_{r=1}^{N}\left(\hat{b}_{ri} - \frac{1}{N}\sum_{r=1}^{N}\hat{b}_{ri}\right)^2} \quad (3)$$

where $\hat{b}$ is the estimated item difficulty and b is true difficulty

$$RMSE(\hat{b}_i) = \sqrt{\frac{1}{N}\sum_{r=1}^{N}(\hat{b}_{ri} - b_i)^2} \quad (4)$$

The relationship between MSE (=RMSE^2), SE and bias is:

$$MSE(\hat{b}_i) = E\left[(\hat{b}_i - b_i)^2\right] = E\left[(\hat{b}_i - E(\hat{b}_i))^2 + (E(\hat{b}_i) - b_i)^2\right] = Var(\hat{b}_i) + Bias^2(\hat{b}_i) \quad (5)$$

This relationship can be used to verify the calculation accuracy of each criterion index. The average of bias, Abias, SE, and RMSE across M items (i=1, 2, …, M) can be described as:

$$\text{Bias}(\hat{b}) = \frac{1}{M}\frac{1}{N}\sum_{i=1}^{M}\sum_{r=1}^{N}(\hat{b}_{ri} - b_i) \quad (6)$$

$$\text{Abias}(\hat{b}) = \frac{1}{M}\frac{1}{N}\sum_{i=1}^{M}\sum_{r=1}^{N}|\hat{b}_{ri} - b_i| \quad (7)$$

$$SE(\hat{b}) = \sqrt{\frac{1}{M}\frac{1}{N}\sum_{i=1}^{M}\sum_{r=1}^{N}\left(\hat{b}_{ri} - \frac{1}{N}\sum_{r=1}^{N}\hat{b}_{ri}\right)^2} \quad (8)$$

$$RMSE(\hat{b}) = \sqrt{\frac{1}{M}\frac{1}{N}\sum_{i=1}^{M}\sum_{r=1}^{N}(\hat{b}_{ri} - b_i)^2} \quad (9)$$

The relationship among average bias, SE and RMSE in (6) is no longer true for average bias, SE and RMSE.