Effect of Missing Data in Computerized Adaptive Testing on Accuracy of Item Parameter Estimation: A comparison of NWEA and WINSTEPS Item Parameter Calibration Procedures

Shudong Wang
Gregg Harris

NWEA

Paper presented at the annual meeting of the American Educational Research Association (AERA) . April 12-16, 2012, Vancouver, British Columbia, Canada.

Send correspondence to:
Shudong Wang
Northwest Evaluation Association (NWEA)
121 NW Everett St.
Portland, OR 97206
Shudong.Wang@NWEA.org

Abstract

The purposes of the study are to investigate effects of missing data in computerized adaptive testing (CAT) on accuracy of item parameter estimation based on Rasch model using Northwest Evaluation Association (NWEA) and WINSTEPS calibration methods, and to provides information on the relationship between two methods. The results illustrate that the test length (or missing rate) and calibration methods have not only a statistically significant impact, but also practical implication on the accuracy of recovery of item parameters. The relationship between true and estimate item parameters from CAT data with large missing rate show that item parameter recovery is not as good as conventional thinking by using WINSTEPS, while the robustness of NWEA calibration methods for CAT data is encouraging.

Purpose of Study

The major purpose of the study is to investigate effects of missing data in computerized adaptive testing (CAT) on accuracy of item parameter estimation using Northwest Evaluation Association (NWEA) calibration method and commercialized software WINSTEPS. The study also provide information on the relationship between NWEA calibration methods and WINSTEP.

Perspective

Item and person parameter estimation are the most important and difficult tasks in item response theory (IRT). NWEA routinely conducts item and person parameter estimations using the Rasch model (Wright & Stone, 1979) in the NWEA computerized adaptive testing (CAT) operational system. The advantages of CAT over traditional paper-pencil testing rely on psychometric assumptions and conditions that are required for implementing appropriate IRT models. Any violation of such assumptions and conditions could lead to less efficient CAT administration and less accurate measurement of CAT results (Hambleton & Swaminathan, 1985). Unfortunately, in real world operational work, the advantages offered by CAT have a variety of challenges that include the qualities of calibration samples used for calibrating new/field test/pre-test items, calibration algorithm, calibration procedures, and CAT administration procedures. Estimation of new/field test/pre-test item parameters is one of the most critical components in a CAT system and a factor that could affect the quality of CAT tests. Many research studies (Stocking ,1990; Thissen & Wainer, 1982; Wingersky & Lord, 1984) show the close relationship between precision of item parameter estimation and the ability distribution of examinees used for calibration. Because CAT selects items that match an examinee's ability, the calibration sample obtained from a CAT administration restricts the range of ability with sparse item response data. This could lead to less than ideal estimation of new/field test/pre-test item parameters (Ban, Hanson, Wang, Yi, & Harris, 2000; Hsu, Thompson, & Chen, 1998). The restricted ability range with sparse matrix responses has caused some concerns about new/field test/pre-test item parameter calibration at NWEA. One approach to overcoming the limited ability range of the CAT calibration procedure is to move the new/field test/pre-test item position close to the start point when student provisional ability estimate has not narrowed to his or her true ability yet, so the person ability range of calibration sample can be close to the situation in linear test. Since the early provisional ability estimate can differ from the final estimate, this provides some variance in the ability range of responses.

In order to use new items in operational CAT, major underlying assumptions used by this procedure include:

1) Calibration procedure used to estimate the new/field test/pre-test items is correct.
2) The construct measured by the item remains the same for both calibration and operational items.
3) The ability distribution of examinees for the calibration sample is close to the ability distribution of examinees for the real operational sample, if they are not the same.
4) Item administration conditions (such as administration modes and procedures) are the same for both calibration and operational items.

Any violation of the above assumptions could lead to differences of CAT results between new/field test/pre-test items and operational items. For example, one of the potential factors that could affect the second and third assumptions is the difference of mixed calibration samples between on-grade and off-grade. On-grade calibration uses item responses from students whose ability matches item difficulties at his or her grade, and off-grade calibration uses item responses without limiting students whose ability matches item difficulties across grades (effect of mixture of distribution). This report is only concerned about the first assumption. The effect of the other assumptions on the quality of item parameters will be studied in other reports.

This report focuses on the accuracy of item calibration procedure used in NWEA. The primary independent variable examined in this study is the calibration program. Two calibration programs investigated in this study are NWEA and WINSTEPS. WINSTEPS use two different estimation procedures, i.e., PROX for initial estimation and Joint Maximum Likelihood Estimation called UCON (Unconditional Maximum Likelihood Estimation) for fine-tuned estimation (Linacre, 2009). For WINSTEPS calibration methods, three different calibration methods used in this study are free estimation for full data (EF), free estimation for sparse data (ES), and fixed sparse observed person parameter estimation for sparse data (FSOS). For NWEA calibration, two calibration procedures are non-filtered (NWEANF) and filtered (NWEAF) methods. There are a total of five calibration methods used in this study.

Other factors, such as the calibration algorithm, have been ruled out as factors of this study. The general item calibration procedure used by NWEA works like this: For replenishing item pools, the new/field test/pre-test items are usually administrated to students along with operational items in a regular test session. After new items have accumulated an adequate response sample, they are then calibrated by fixing student abilities at their final CAT ability estimates based on operational items. This is basically a common person design. This study primarily focuses on answering the following three research questions:

1) What are the relationships among true and estimated item parameters across calibration methods and test lengths?
2) Are there any statistically significant differences among item parameters in terms of calibration accuracy for the five calibration methods (EF, ES, FSOS, NWEANF, and NWEAF) for a given test length (or missing rate)?
3) Are there any statistically significant differences among item parameters in terms of calibration accuracy for the five calibration methods (EF, ES, FSOS, NWEANF, and NWEAF) across different test lengths (or missing rates)?

Methods

The Monte Carlo (MC) technique seems to be an appropriate choice to conduct this research and meets the main purpose of this study because both true parameters of item difficulty and person ability that are difficult to obtain in a real-life experiment must be known in order to obtain the error estimation. The MC indices are necessary and sufficient to evaluate the accuracy of ability estimation methods. The main advantages of MC procedures (Harwell, Stone, Hsu, & Kirisci, 1996; Naylor, Balintfy, Burdick, and Chu, 1968) are (a) they can be used to solve problems in situations where an analytic solution does not exist or is impractical; (b) they can determine system outcomes in a stochastic process and chance parameters can be controlled or manipulated in order to study their effects on system performance; and (c) cost is low.

Design of Monte Carlo Study

The primary goal of this design is to answer the stated research questions and to maximize generalizability and replicability of research results. Both descriptive methods and inferential procedures are used in this MC study. Although the descriptive methods provide global summaries of the study results, the deficiencies of descriptive methods are (a) masking of complex effects, (b) failing to provide estimates of the magnitudes of the effects, and (c) failing to systematically take into account the sampling error associated with the random generation of data (Harwell, 1997; Hoaglin and Andrews, 1975; Timm, 1976). Inferential procedures, on the other hand, can overcome all of these deficiencies by conceptualizing the MC study as a statistical sampling experiment (Harwell, 1997; Spence, 1983). Additional advantages of using this conceptualization are that the threats to internal and external validity can be evaluated and that the similarities and differences among the results of different studies on the same problems can be compared (Campbell & Stanley, 1963). Based on the objectives of this study, one balanced factorial experiment designs are used.

Independent Variables

1. Calibration Program

The primary independent variable examined in this study is the calibration program. There are five calibration methods (EF, ES, FSOS, NWEANF, and NWEAF) used in this study.

2. Data Type

The second independent variable examined in this study is data type. Two types of data investigated in this study are non-missing and missing data. For the missing data, percentage of missing data is measured by the test length of CAT test for a given item pool.

Dependent Variables

There are a variety of statistics that can be used to evaluate how well true parameters are recovered for each of the simulation conditions. 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:

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

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

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

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

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

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

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

This relationship can be used to verify the correctness of calculation of each criterion index. Besides bias, absolute bias (absolute value of bias) is used because the direction of bias (positive or negative) is a function of either person ability or item difficulty. The average of bias, Abias, SE, and RMSE across M items (i=1, 2, …, M) can be described as following:

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

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

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

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

The relationship among average bias, SE, and RMSE in (5) is no longer true for average bias, SE, and RMSE.

Data Sources

All the data used in this study are simulated.

1. Missing Data Issue

Two issues in response data obtained from CAT administration that do not exist in linear tests are the restricted range of ability and sparse item response data. Both issues are frequently dealt with as a missing data issue in the field of statistics. Because for a given item bank, each examinee will only answer a partial sample of items in a test. The rest of the items in the bank remain unanswered, but responses of these remaining items could be imputed based on some theoretical models. However, the true parameters still remain unknown. In this study, a reversal approach is used, i.e., the full data matrices are generated with known person and item parameters. Then for a given CAT test length, responses for the remaining items in the item bank are deleted.

2. Data Generation and Calibration

The design of the data generation procedure is shown in Figure 1. There are three major steps in data generation and calibration in this study and all data generation and calibration are based on the Rasch Model.

For the first step, ability parameters for 20,000 persons and difficulty parameters for 1300 items in an item bank are generated based on standardized normal distribution *N*(0,1). In the second step, item and person parameters are sorted so that person ability optimal matches item difficulty. Then responses for given test lengths are kept and the remaining responses are deleted to create 10 data sets with different missing rates. Table 1 lists the test lengths and missing rates of the generated data. The second Step is replicated 50 times for a total 500 data sets to be generated. In the third step, the previously generated 500 data sets are calibrated using the six different calibration methods mentioned in section 1.3.1.

Because NWEA methods do not apply to full data matrices that have a test length of 1300 items, there are a total 2300 calibrations conducted in this study {[3(WINSTEPS) + 2(NWEA)] x 9 (test length) + 1 (WINSTEPS) x 1 (test length)}.

Results

A. Descriptive Statistics of Conditional Dependent Variables

Parameter recovery is evaluated by comparing the estimates to the true (generated) parameters in terms of five dependent variables: correlations, bias, Abias, SE, and RMSE. The descriptive statistics, such as tabular summaries and graphical presentations, are used to present these dependent variables.

1. Correlations among true and estimated item parameters.

On average, over 50 replications, the Pearson's correlation coefficient between true and estimated item parameters of free estimation for full data (EF) is 0.9999, which indicates almost perfect recovery of at least rank order of true parameters by WINSTEPS. Table 2 shows Pearson's correlation coefficients across different test lengths under different calibration conditions from one replication. Figures 2 to 7 illustrates the plots of true and estimated parameters across different test lengths under different calibration conditions from one replication. Apparently, for WINSTEPS, the parameter recovery is noticeably worse for a short test length or a large missing rate. In contrast with the relationship between true and estimated parameters in which the value of correlation coefficient increases as the value of test length increases, the pattern of relationships among estimated parameters for different test lengths reveals that the values of correlation coefficients do not increase as test length increases within certain test length ranges. However, WINSTEPS recovery improves considerably for longer tests, when test length increases to 50 items or missing rate decreases to .962. For NWEA calibration methods, the item parameter recovery is very robust across test lengths or missing rates.

After test length increases to 70 items, the recovery of correlation coefficient from WINSTEPS shows better results than that from NWEA methods. There is little difference of recovery results between NWEAF and NWEANF calibration methods in this study. Table 3 presents average Pearson's correlation coefficients between true and estimated item parameters over 50 replications across test length and calibration conditions. Overall, NWEA calibration methods are better than WINSTEPS calibration methods for short tests or large missing rates; for longer tests, WINSTEPS recovery is more accurate than that of NWEA's calibration methods.

Test length and missing rate have a more significant impact on item parameter recovery for WINSTEPS than for NWEA calibration methods.

2. Conditional Bias

Figures 8 to 12 depict the conditional biases along the theta scale of true item parameters across different test lengths for each of the given calibration conditions. Figures 13 to 21 plot the conditional biases along the theta scale of true item parameters across different calibration conditions for each of the given test lengths.

For WINSTEPS, across test lengths, the biases of item parameter estimates at the two extremes are larger than that at the middle. For easy items, the biases are over-estimated and for hard items, the biases are under-estimated, i.e., the biases are distributed “inward” along the theta scale for all three WINSTEPS calibration methods (SE, FTS, and FSOS). However, as test length increases, the biases approach but are not equal to zero for both easy and hard items and the degree of “inward” decreases significantly.

For both NWEA calibration methods (NWEANF and NWEAF), the biases are evenly distributed along the theta scale and there is no “inward” trend as there is in WINSTEPS. The values of biases for NWEA calibration methods are smaller than the values of WINSTEPS calibration, and this is especially true at the two extremes of the scale for short tests.

3. Conditional SE

The conditional SEs along the theta scale of true item parameters across different test lengths for each of the given calibration conditions and the conditional SEs along the theta scale of true item parameters across different calibration conditions for each of the given test lengths are presented in Figures 22 to 26 and Figures 27 to 35.

For all three WINSTEPS calibration methods (SE, FTS, and FSOS), if test length is longer than 60, the SEs of item parameter estimates are almost evenly distributed across theta scale; for tests that have less than 70 items, the SEs are slight lower at the middle of the theta scale. The results also show that the SEs for all three methods decrease as test length increases. For both NWEA calibration methods (NWEANF and NWEAF), the variability in SE is dramatically larger compared to the variability of SE from WINSTEPS methods. Although the results show that the SE decreases as test length increase for NWEA calibration methods, the magnitude of the reduction in SEs is less than the magnitude of the reduction of WINSTEPS calibration methods. Overall, the comparison of the SE results across all five calibrations in Figures 27 to 35 show that the WINSTEPS calibration methods are more stable and have smaller SEs than the NWEA calibration methods. The SEs from both WINSTEPS and NWEA calibration methods decrease as test length increases.

4. Conditional RMSE

Both the conditional RMSEs along the theta scale of true item parameters across different test lengths for each of the given calibration conditions and the conditional RMSEs along the theta scale of true item parameters across different calibration conditions for each of the given test lengths are depicted in Figures 36 to 40 and Figures 41 to 49.

Because RMSE is the function of both bias and SE, comparisons of Figures 36 to 40 indicate that the shape of the RMSEs is largely affected by bias and it tends to be lower at the middle of the theta scale (lower bias) and higher at the two extremes (higher bias) for WINSTEPS calibration methods (SE, FTS, and FSOS). From Figures 39 and 40 for the NWEA calibration methods. It is clear that, because the shape of the RMSEs is affected by the SEs, NWEA method RMSEs tend to show larger variability than those from WINSTEPS methods. Comparisons of Figures 41 to 49 also reveal that the combination effects of bias and SE determine that RMSEs are larger at the two extremes for the WINSTEPS calibration methods and show consistent variability across the theta scale for the NWEA calibration methods. In general, as test length increase, the RMSEs for WINSTEPS calibration methods and variability of RMSEs for NWEA calibration methods decrease.

5. Average Dependent Variables of Correlation, Bias, Abias, SE, RMSE, Variance, and MSE

Five average dependent variables or overall indexes (Correlation, Bias, Abias, SE, and RMSE) along two additional overall indexes (Variance, and MSE) at different test lengths and calibration methods over 50 replications are presented in Table 4. The distributions of Correlation, Bias, Abias, SE, and RMSE variables along test length across calibration methods are shown in Figures 50 to 54. Result of correlations in Figure 50 shows that for WINSTEPS, the correlations are function of test length until test length reaches 70. The test length has less effect on correlation for NWEA calibration methods, which means NWEA calibration methods are more robust in term of test length than WINSTEPS calibration methods. Figure 51 demonstrates that NWEA calibration methods have less bias than WINSTEPS and FTS has the largest bias associated with test length. The absolute bias (abias) is shown in Figure 52. In general, short tests have larger abias than longer tests across different calibration conditions, especially for WINSTEPS calibration methods. When test lengths reach 80, there is little difference of abias among calibration methods. Figure 53 depict the SE of calibration methods with test length and WINSTEPS calibration methods have large SE than NWEA calibration methods before test length reaches to 75, and have smaller SE than NWEA calibration methods when test length is longer than 75. Figure 54 shows that for both WINSTEPS and NWEA calibration methods, the RMSE decrease as test length increases and WINSTEPS calibration methods have larger RMSE than NWEA calibration methods. However, test length has more impact on RMSE for WINSTEPS calibration than it does for NWEA calibration methods. When test length is longer than 85, the WINSTEPS calibration methods have smaller RMSE than NWEA calibration methods.

B. Inferential Statistics of Average Dependent Variables

The conditional dependent variables are computed at each theta point of item parameter, and average dependent variables (bias, abias, SE, and RMSE) are computed by using equations 6 to 9 on page 5 and correlation remain the same.

2.2.1   Statistical Hypotheses

For this study, the five dependent variables (correlation, bias, abias, SE, RMSE) are used to evaluate the accuracy of item calibration, based on research questions proposed in the introduction section. The statistical null hypotheses are presented as follows.

1) There are no effects of calibration method on the accuracy of item parameter estimations when some or all of the dependent variables (correlation, bias, abias, SE, and RMSE) are used in different simulation conditions.

2) There are no effects of test length on the accuracy of item parameter estimations when some or all of the dependent variables (correlation, bias, abias, SE, and RMSE) are used in different simulation conditions.

3) There are no interaction effects between two factors mentioned above when some or all of the dependent variables (correlation, bias, abias, SE, and RMSE) are used in different simulation conditions.

Because the Monte Carlo study is really a statistical sampling technique with an underlying model, the number of replications in this Monte Carlo study is the analogue of sample size. 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 50 times. The simulation results of dependent variables from two-way (calibration method and test length) crossed ANOVA are presented. Both test statistics and effect sizes are used to determine levels of significant effects. The magnitude of significant effects is estimated using eta-squared η^2 (empirical η^2 as an effect size estimate). Following the advice of Cohen (Cohen, 1988), the effect size in terms of η^2 had been classified as: (a) no effect ($\eta^2 < 0.0099 \approx 0.01$), (b) small effect ($0.01 < \eta^2 < 0.0588 \approx 0.06$), (c) medium effect ($0.06 < \eta^2 < 0.1379 \approx 0.14$), and (d) large effect ($\eta^2 > 0.14$).

1. ANOVA Results

Tables 5 to 9 show the results of the two-way ANOVA of average correlation, bias, abias, SE, and RMSE for this study. Using $\alpha = 0.01$ for each hypothesis tested (means of each dependent variables are equal across calibration condition and test length), the two main effects: M (method) and L (length), and one interaction effect M x L are all statistically significant. For the two-way interaction, this significance means that for a given dependent variable, mean differences among the levels of any factor are not constant across all levels of the remaining factor. Because the interaction is significant, it is not appropriate to perform multiple comparisons on the main effect means of dependent variable, but it is appropriate to compare simple effects, i.e., to test each factor at each level of the other factor.

In terms of η^2 (the proportion of variance in the dependent variable that is explained by the differences among the groups), among the interactions, the two-way M x L interaction effects accounted for the second largest parts of the variance for correlation (39.8%), abias (35.8%), and RMSE (28.9%). Their effect sizes are in the large ranges ($\eta^2 > 0.14$). For both interaction effects of bias and SE, the effect sizes can be categorized as small ($0.01 < \eta^2 < 0.0588 \approx 0.06$).

Among the main effects, for bias, the method effect accounted for most of the variance (68.65%) and test length account for the least (0.5%). For the rest of the dependent variables (correlation, abias, SE, and RMSE), test length accounts for most of the variance (40.6%, 37.9%, 41.2%, and 39.1%) and method accounts for second or third highest amount of total variance (13.5%, 21.3%, 8.8%, and 27.3%).

For the mean difference between true and estimated parameters under different simulation condition, two sample t-tests show that almost all calibration methods are statistically significant even though most mean differences are at 5 or 6 decimal points because the standard errors are very small.

In general, the results for average dependent variables are consistent with the results of conditional dependent variables. The factor of calibration methods has the most influence on bias because it accounts for 68.7.0% of the total variance of bias. On the other hand, the factor of test length has the most influence on correlation, abias, SE, and EMSE.

C. Summary of Results

The results in Figure 50 and Table 5 indicate that relationships among true and estimated item parameters from both NWEA and WINSTEPS calibration methods are strong, but correlations between true and estimated item parameters from NWEA calibration methods are better than those from WINSTEPS calibration for short tests or large missing rates. Across test lengths, NWEA calibration methods have less bias than WINSTEPS calibration methods. For longer tests (longer than 75 items), the SEs of WINSTEPS calibration methods at least match or are better than NWEA calibration methods (See Figure 53). Although average SE of NWEA calibration methods for short tests are smaller than those from WINSTEPS, a small percent of conditional SEs of NWEA calibration methods shows unusual large values in Figures 25 and 25. In general, as test length increases, the accuracy of all calibration methods increases. The test length (or missing rate) has a larger impact on the WINSTEPS calibration methods than on the NWEA calibration methods in terms of both conditional and average correlation, bias, abias, SE, and RMSE. The NWEA calibration methods are superior to WINSTEPS calibration methods on all dependent variables except conditional SE for the test that has test length less than 70. The WINSTEPS calibration methods show more favorable results of both conditional and average dependent variables when test length is longer than 70. Although NWEA calibration methods show some large variation on SE. On average, NWEA calibration methods have smaller SE than WINSTEPS calibration methods. Both test length and calibration methods have a statistically significant impact on these average dependent variables. It is important to note that simulated data for both person and item parameters used in this study are from normal distributions which satisfies the assumption used in the Rasch model. In practical settings, sample of distributions are seldom normal and future research should be conducted to assess the consequence of violating model assumptions.

Scientific Significance of the Study

Recently, the CAT has been seen as a particularly effective method in measuring an individual student's growth over time in K-12 assessment (Way, Twing, Camara, Sweeney, Lazer, & Mazzeo, 2010). Currently, besides Oregon, Delaware, and Idaho, which are using CAT based on Rasch model in their state assessments, many other states (Georgia, Hawaii, Maryland, North Carolina, South Dakota, Utah, and Virginia) are also in various stages of CAT development. As a matter of fact, one of the two consortia created as part of the *Race to the Top* initiative (Race to the Top Assessment Program, 2010), the SMARTER Balanced Assessment Consortium (SBAC), is committed to a computer adaptive model because it represents a unique opportunity to create a large-scale assessment system that provides maximally accurate achievement results for each student.

However, little study has focused on the effect of CAT missing data on accuracy of item parameter estimation. In particular, WINSTEPS is one of the most widely used commercial Rasch software and no knowledge exists on the missing data impact on WINSTEPS item parameter estimation in CAT. The accurate estimation of item parameters is the psychometric foundation for many Rasch model applications, such as, equating, scaling, and growth modeling. This study provides insights into the accuracy of both NWEA and WINSTEPS calibration methods and results could give guidelines for practitioners to make reasonable decisions on application of calibration results.

References

Ban, J.-C., Hanson, B. H., Wang, T., Yi, Q., & Harris, D. J. (2000). *A comparative study of online pretest item calibration/scaling methods in computerized adaptive testing*. (ACT Research Report 00-11). Iowa City, IA: ACT, Inc. [Available at http://www.b-a-h.com/ papers/paper0003.html].

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.

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

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.

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.*

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

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

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.

Washington State, on behalf of the SMARTER Balanced Assessment Consortium. (2010). *Race* *to the Top Assessment Program Application for New Grants*. Retrieved from http://www.k12.wa.us/SMARTER/pubdocs/SBAC_Narrative.pdf.

Way, W., Twing, J., Camara, W., Sweeney, K., Lazer, S., & Mazzeo, J. (2010). *Some considerations related to the use of adaptive testing for the Common Core Assessments*. Retrieved June 11, 2010, from 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. CAT Test Length and Missing Rate of Generated Data

Number of Person	Bank Size	CAT Test Length and Missing Rate										
20000	1300	Test Length	20	30	40	50	60	70	80	90	100	1300
		Missing Rate (%)	.985	.977	.969	.962	.954	.946	.938	.931	.923	0

Table 2. Pearson Correlation Coefficients* between True and Estimated Item Parameters Among Different Test Lengths for Free Estimation of Sparse Data (ES) Condition with One Replication

	True	L20	L30	L40	L50	L60	L70	L80	L90	L100
True	1.0000									
L20	0.5191	1.0000								
L30	0.8532	0.8432	1.0000							
L40	0.9631	0.6995	0.9536	1.0000						
L50	0.9823	0.6402	0.9260	0.9939	1.0000					
L60	0.9947	0.5806	0.8905	0.9818	0.9945	1.0000				
L70	0.9970	0.5562	0.8767	0.9759	0.9910	0.9990	1.0000			
L80	0.9980	0.5399	0.8665	0.9706	0.9876	0.9978	0.9994	1.0000		
L90	0.9984	0.5312	0.8602	0.9677	0.9855	0.9969	0.9988	0.9997	1.0000	
L100	0.9986	0.5313	0.8607	0.9676	0.9853	0.9967	0.9987	0.9996	0.9998	1.0000

*: N=1300 and p<0.001.

Table 3. Pearson Correlation Coefficients between True and Estimated Item Parameters Among Different Test Length and Condition Over 50 Replications

Test Length	Calibration Condition					
	ES	FTS	FSOS	NWEAF	NWEANF	EF
20	0.6074	0.4836	0.6264	0.9892	0.9892	
30	0.8095	0.7744	0.8213	0.9913	0.9913	
40	0.9215	0.9071	0.9253	0.9927	0.9927	
50	0.9691	0.9717	0.9702	0.9938	0.9938	
60	0.9882	0.9922	0.9883	0.9952	0.9952	
70	0.9951	0.9959	0.9951	0.9957	0.9957	
80	0.9972	0.9972	0.9971	0.9961	0.9961	
90	0.9979	0.9979	0.9978	0.9963	0.9963	
100	0.9983	0.9983	0.9982	0.9967	0.9967	
1300						0.9999

Table 4. Average Observe Mean* of Item Parameter Estimates, Correlation, Bias, Abias, SE, RMSE, Variance, and MSE for Different Test Length and Calibration Conditions Over 50 Replications

Calibration Method	Test Length	Correlation	Mean	Bias	Abias	SE	RMSE	VAR	MSE
ES	20	0.6027	0.0000	0.0154	0.7262	0.1876	0.8380	0.0359	0.7022
	30	0.8121	0.0000	0.0154	0.6130	0.1393	0.6918	0.0198	0.4786
	40	0.9242	0.0000	0.0154	0.4744	0.1205	0.5357	0.0148	0.2869
	50	0.9710	0.0000	0.0154	0.3228	0.1335	0.3779	0.0190	0.1428
	60	0.9894	0.0000	0.0154	0.1855	0.1259	0.2355	0.0170	0.0555
	70	0.9954	0.0000	0.0154	0.1136	0.0991	0.1492	0.0104	0.0223
	80	0.9973	-0.0593	0.0154	0.0790	0.0771	0.1023	0.0061	0.0105
	90	0.9979	0.0000	0.0151	0.0625	0.0679	0.0802	0.0048	0.0064
	100	0.9983	0.0000	0.0154	0.0542	0.0583	0.0682	0.0035	0.0047
FTS	20	0.4812	0.0292	0.0447	0.7683	0.1845	0.8991	0.0347	0.8084
	30	0.7765	0.0392	0.0547	0.6442	0.1393	0.7335	0.0197	0.5380
	40	0.9096	0.0270	0.0583	0.4989	0.1253	0.5742	0.0162	0.3297
	50	0.9734	0.0562	0.0716	0.3125	0.1376	0.3748	0.0202	0.1405
	60	0.9926	0.0671	0.0825	0.1770	0.1049	0.2214	0.0116	0.0490
	70	0.9960	0.0645	0.0799	0.1282	0.0857	0.1618	0.0076	0.0262
	80	0.9972	0.0580	0.0734	0.1019	0.0742	0.1287	0.0057	0.0166
	90	0.9979	0.0516	0.0670	0.0862	0.0637	0.1065	0.0041	0.0113
	100	0.9983	0.0448	0.0602	0.0756	0.0579	0.0933	0.0034	0.0087

FSOS	20	0.6179	0.0000	0.0154	0.7014	0.2354	0.8220	0.0583	0.6756
	30	0.8216	0.0000	0.0153	0.5902	0.1826	0.6758	0.0350	0.4567
	40	0.9274	0.0000	0.0153	0.4575	0.1501	0.5238	0.0236	0.2743
	50	0.9719	0.0000	0.0153	0.3123	0.1455	0.3705	0.0228	0.1372
	60	0.9895	0.0000	0.0153	0.1815	0.1296	0.2327	0.0181	0.0541
	70	0.9954	0.0000	0.0154	0.1112	0.1004	0.1471	0.0106	0.0216
	80	0.9972	0.0000	0.0154	0.0789	0.0780	0.1023	0.0063	0.0105
	90	0.9979	0.0000	0.0154	0.0642	0.0669	0.0817	0.0046	0.0067
	100	0.9982	0.0000	0.0151	0.0551	0.0598	0.0698	0.0036	0.0049
NWEANF	20	0.9891	-0.0161	-0.0007	0.1137	0.1393	0.1502	0.0220	0.0226
	30	0.9913	-0.0151	0.0004	0.0991	0.1221	0.1338	0.0175	0.0179
	40	0.9927	-0.0154	0.0001	0.0907	0.1110	0.1225	0.0146	0.0150
	50	0.9939	-0.0149	0.0005	0.0837	0.1022	0.1125	0.0123	0.0127
	60	0.9952	-0.0151	0.0003	0.0751	0.0919	0.0991	0.0096	0.0098
	70	0.9957	-0.0155	-0.0001	0.0707	0.0866	0.0939	0.0086	0.0088
	80	0.9961	-0.0153	0.0002	0.0675	0.0830	0.0899	0.0079	0.0081
	90	0.9963	-0.0157	-0.0003	0.0649	0.0799	0.0869	0.0073	0.0076
	100	0.9967	-0.0155	-0.0001	0.0617	0.0763	0.0824	0.0066	0.0068
NWEAF	20	0.9891	-0.0161	-0.0007	0.1137	0.1393	0.1502	0.0220	0.0226
	30	0.9913	-0.0151	0.0004	0.0991	0.1221	0.1338	0.0175	0.0179
	40	0.9927	-0.0154	0.0001	0.0907	0.1111	0.1226	0.0146	0.0150
	50	0.9939	-0.0149	0.0005	0.0837	0.1022	0.1125	0.0123	0.0127
	60	0.9952	-0.0151	0.0003	0.0751	0.0919	0.0991	0.0096	0.0098

70	0.9957	-0.0155	-0.0001	0.0707	0.0866	0.0939	0.0086	0.0088
80	0.9961	-0.0153	0.0001	0.0675	0.0830	0.0899	0.0079	0.0081
90	0.9963	-0.0157	-0.0003	0.0649	0.0799	0.0869	0.0073	0.0076
100	0.9967	-0.0155	-0.0001	0.0617	0.0763	0.0824	0.0066	0.0068

*: Observed Mean of True Item Parameters is -0.01543

Table 5. Results of ANOVA of Correlation

Source	DF	Type I SS	Mean Square	*F*-Value	Pr > F	η^2
Method	5	4.3694	0.8739	1171.2700	<.0001	0.1354
Length	8	13.0846	1.6356	2192.1500	<.0001	0.4055
Method x Length	40	12.8429	0.3211	430.3300	<.0001	0.3980

Table 6. Results of ANOVA of Bias

Source	DF	Type I SS	Mean Square	*F*-Value	Pr > F	η^2
Method	5	1.3852	0.2770	1278.6000	<.0001	0.6865
Length	8	0.0105	0.0013	6.0500	<.0001	0.0052
Method x Length	40	0.0488	0.0012	5.6200	<.0001	0.0242

Table 7. Results of ANOVA of Abias

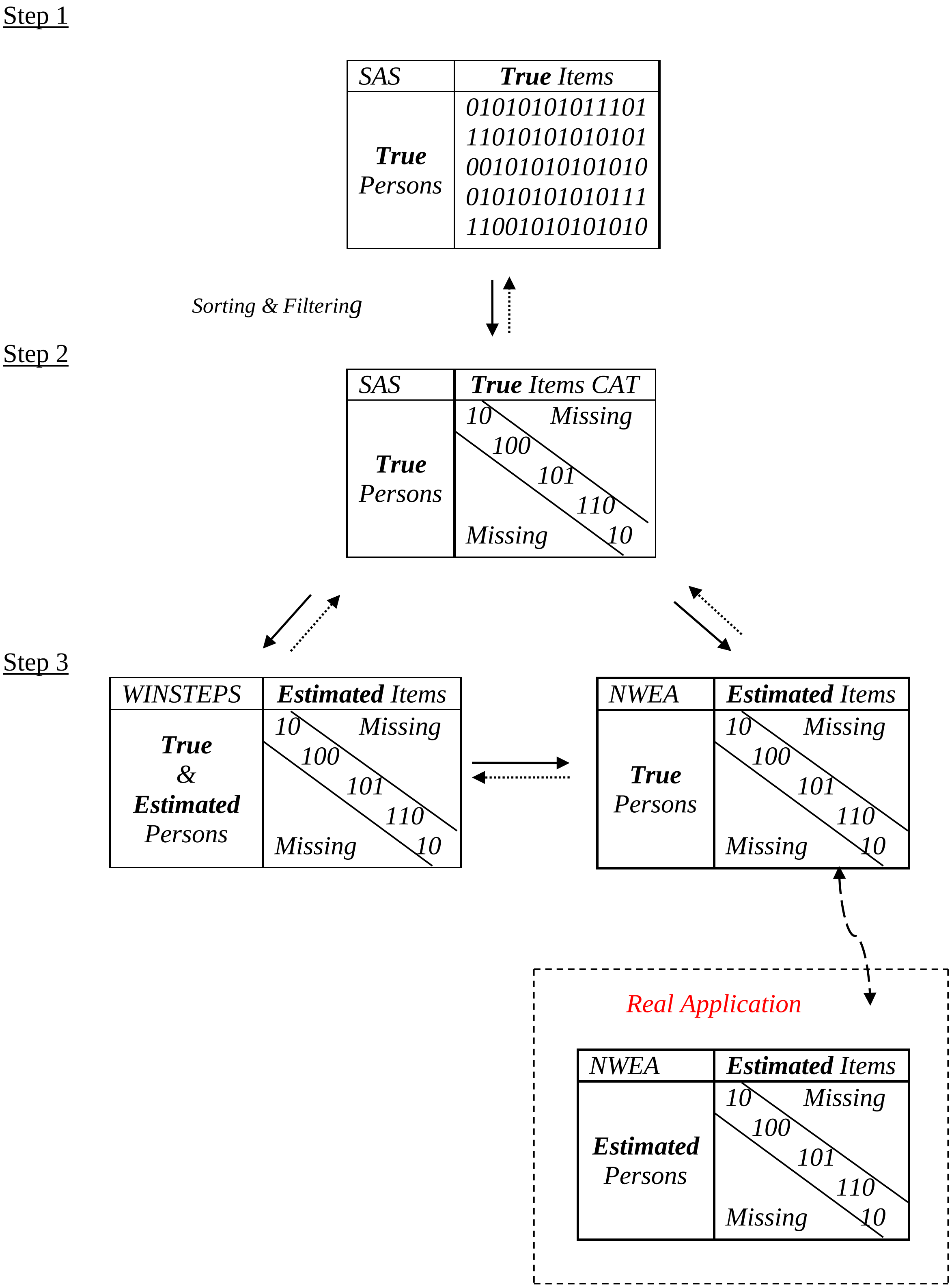
Source	DF	Type I SS	Mean Square	*F*-Value	Pr > F	η^2
Method	5	23.5944	4.7189	2315.8100	<.0001	0.2134
Length	8	41.9832	5.2479	2575.4300	<.0001	0.3797
Method x Length	40	39.5994	0.9900	485.8400	<.0001	0.3581

Table 8. Results of ANOVA of SE

Source	DF	Type I SS	Mean Square	*F*-Value	Pr > F	η^2
Method	5	0.5601	0.1120	105.3700	<.0001	0.0881
Length	8	2.6232	0.3279	308.4100	<.0001	0.4126
Method x Length	40	0.3612	0.0090	8.4900	<.0001	0.0568

Table 9. Results of ANOVA of RMSE

Source	DF	Type I SS	Mean Square	*F*-Value	Pr > F	η^2
Method	5	40.7460	8.1492	3051.1700	<.0001	0.2733
Length	8	58.2233	7.2779	2724.9500	<.0001	0.3906
Method x Length	40	43.0321	1.0758	402.8000	<.0001	0.2887



SAS	***True*** *Items*
True *Persons*	*01010101011101* *11010101010101* *00101010101010* *01010101010111* *11001010101010*

Figure 1. Design of Calibrations Procedure

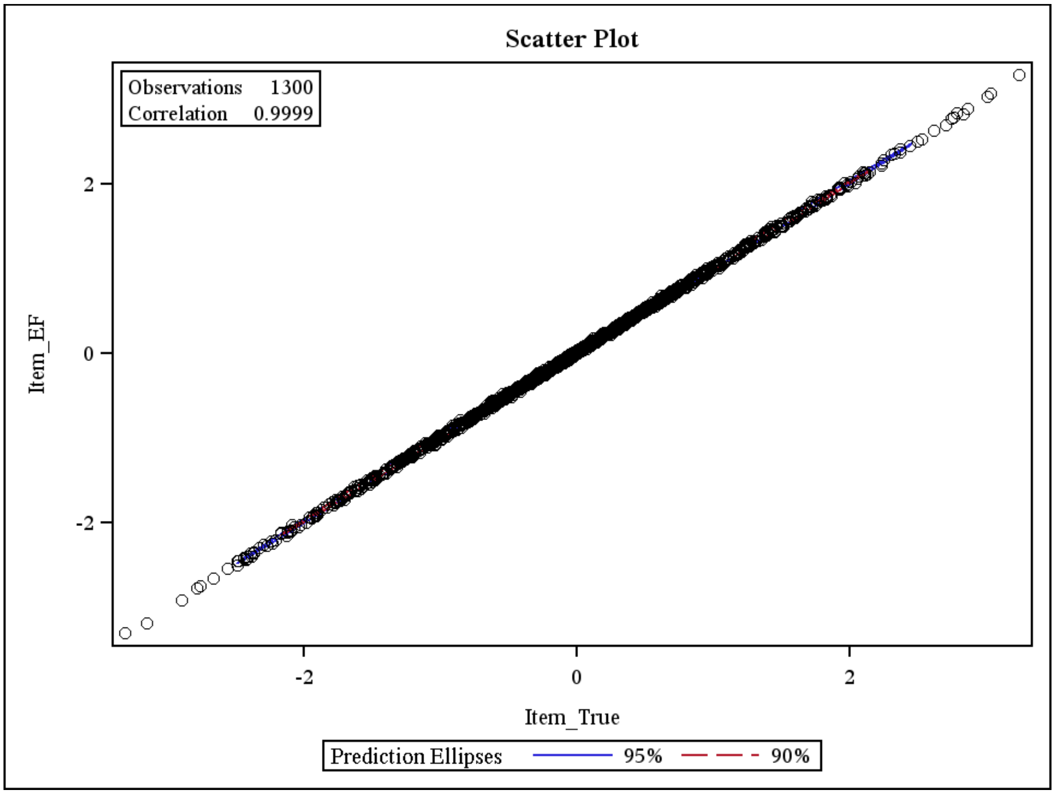


Figure 2. Plot of True and Estimated Item Parameters free estimation for full data (EF) with One Replication

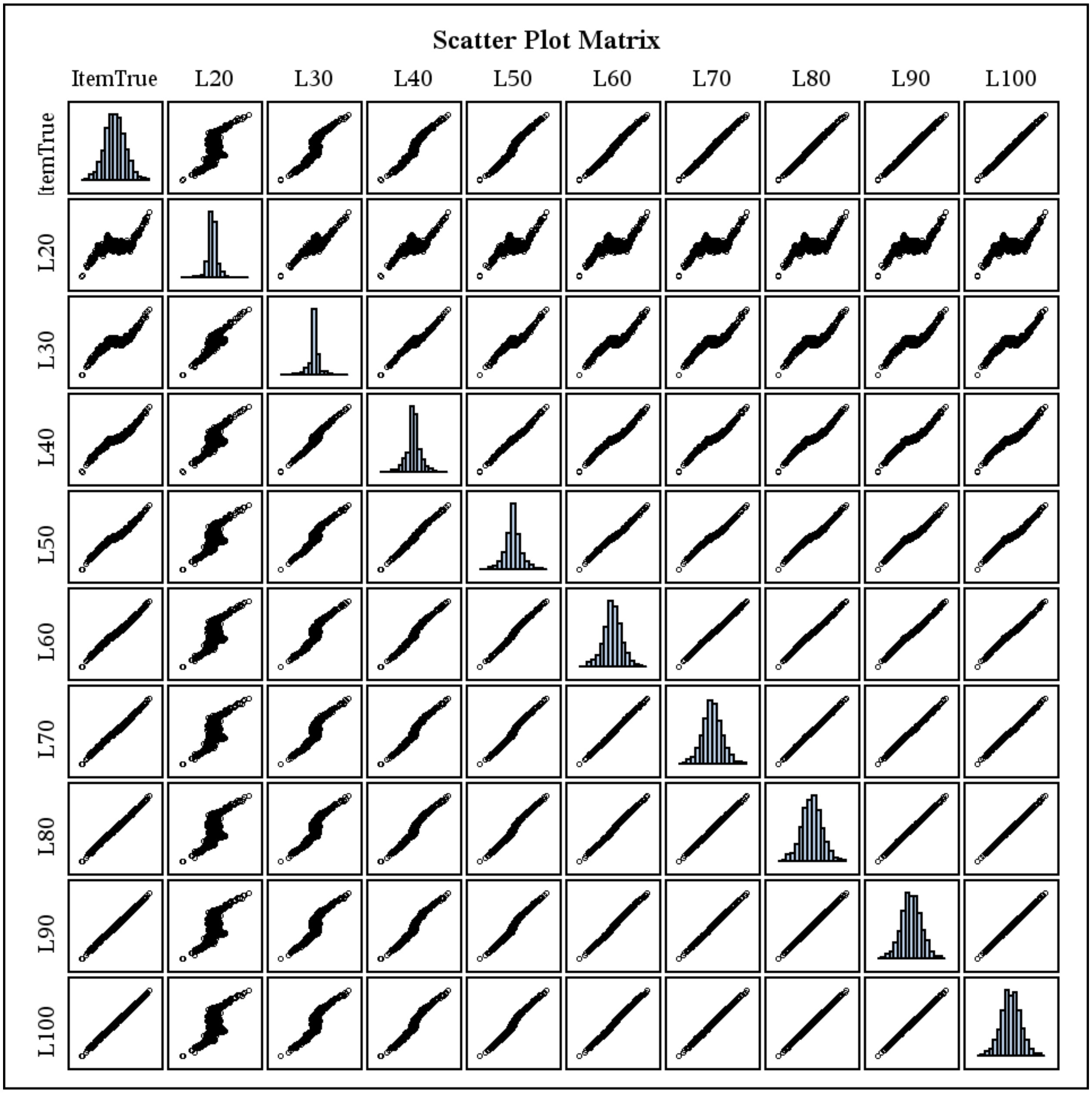


Figure 3. Plot of True and Estimated Item Parameters Across Test Length for Free Estimation of Sparse Data (ES) Condition with One Replication

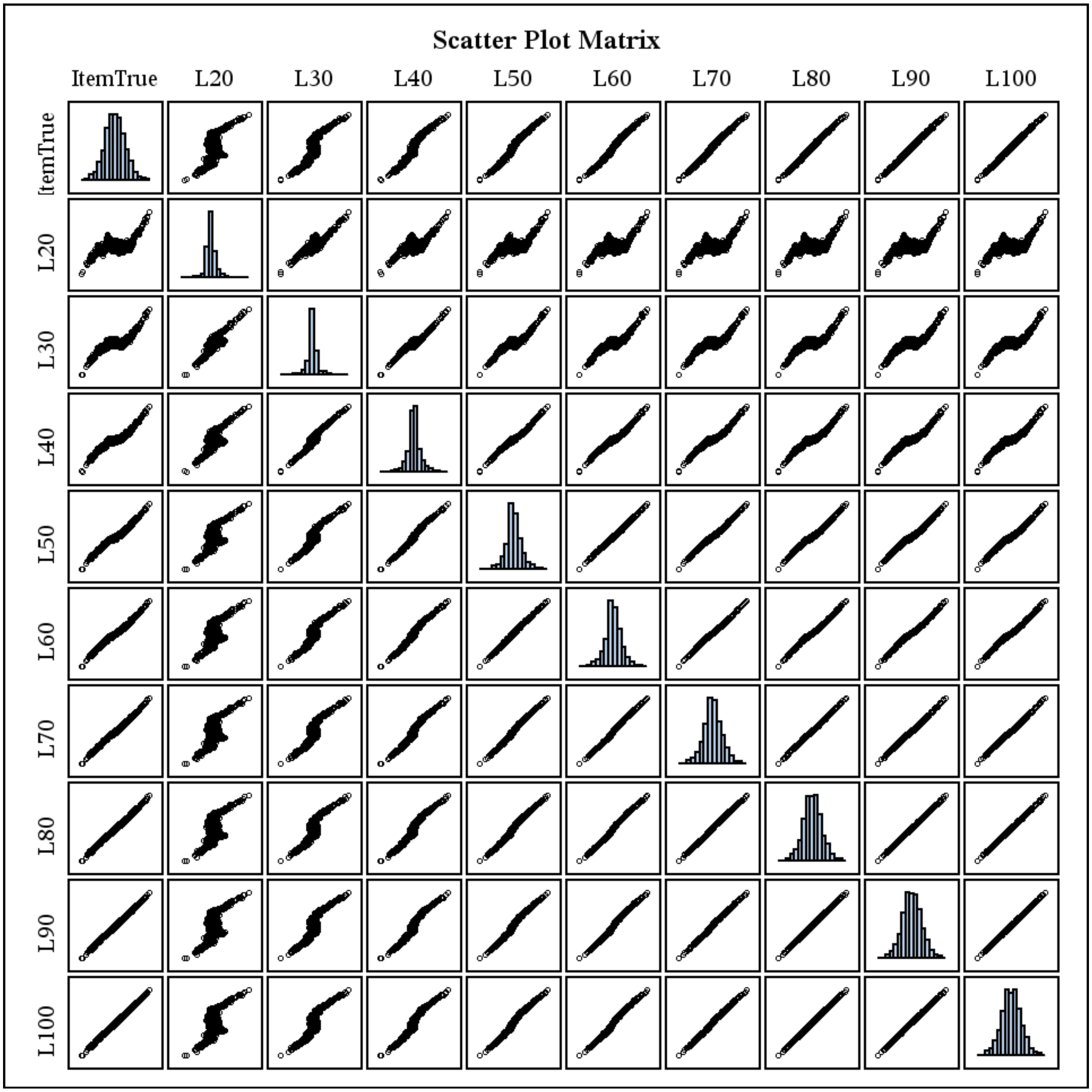


Figure 4. Plot of True and Estimated Item Parameters Across Test Length for Fixed True Person Parameter Estimation of Sparse Data (FTS) Condition with One Replication

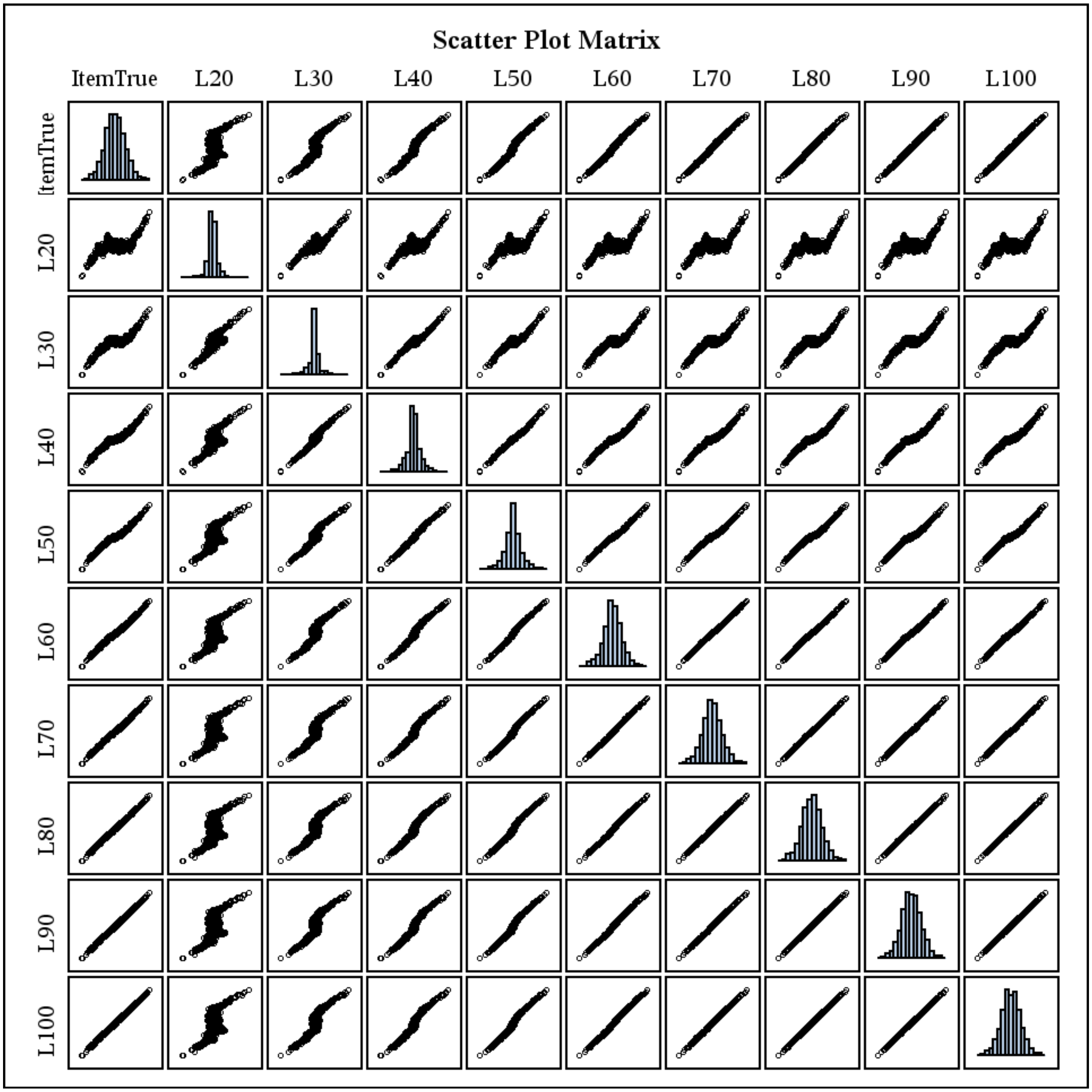


Figure 5. Plot of True and Estimated Item Parameters Across Test Length for Fixed Sparse Observed Person Parameter Estimation of Sparse Data (FSOS) Condition with One Replication

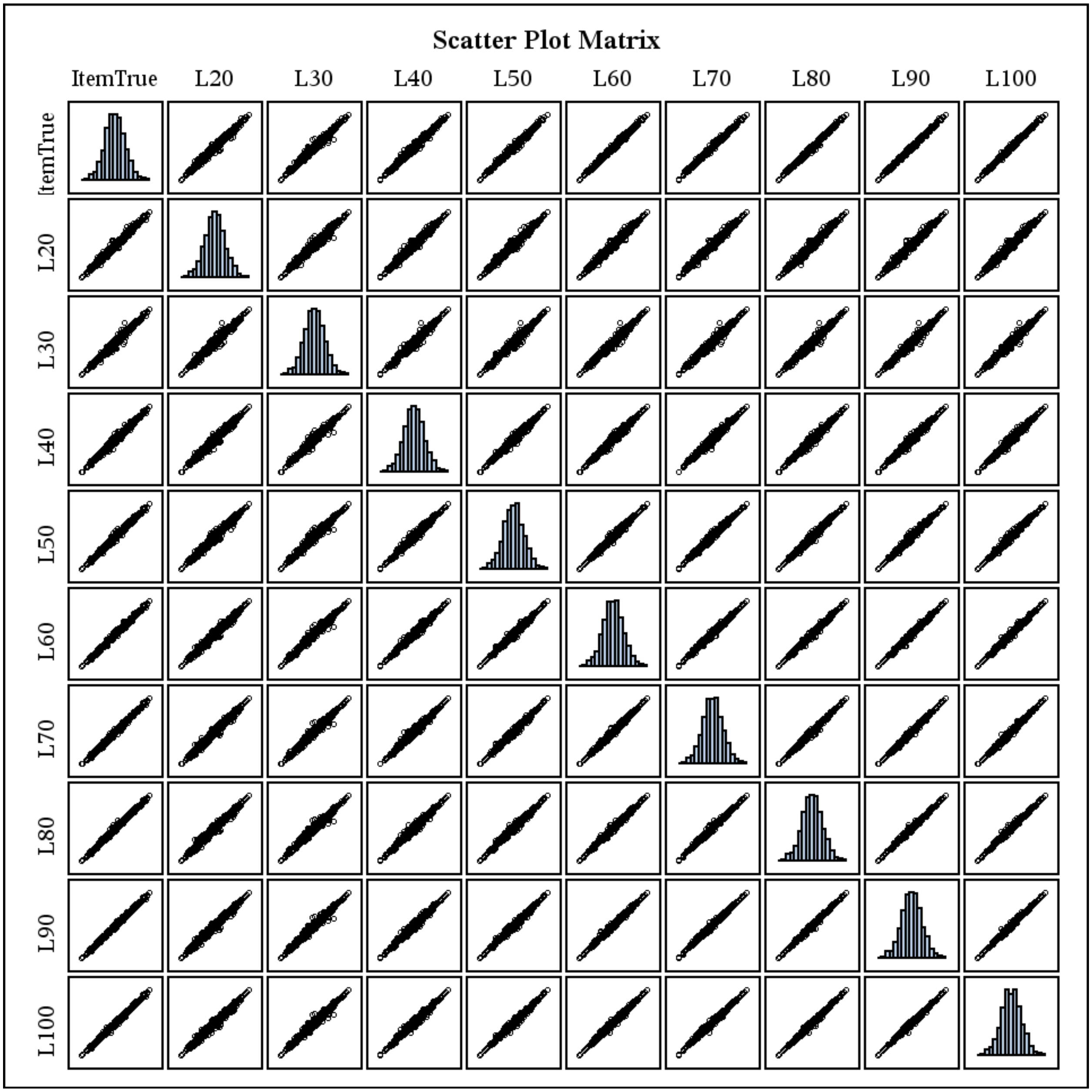


Figure 6. Plot of True and Estimated Item Parameters Across Test Length for NWEA Non-Filtered (NWEANF) Calibration Condition with One Replication

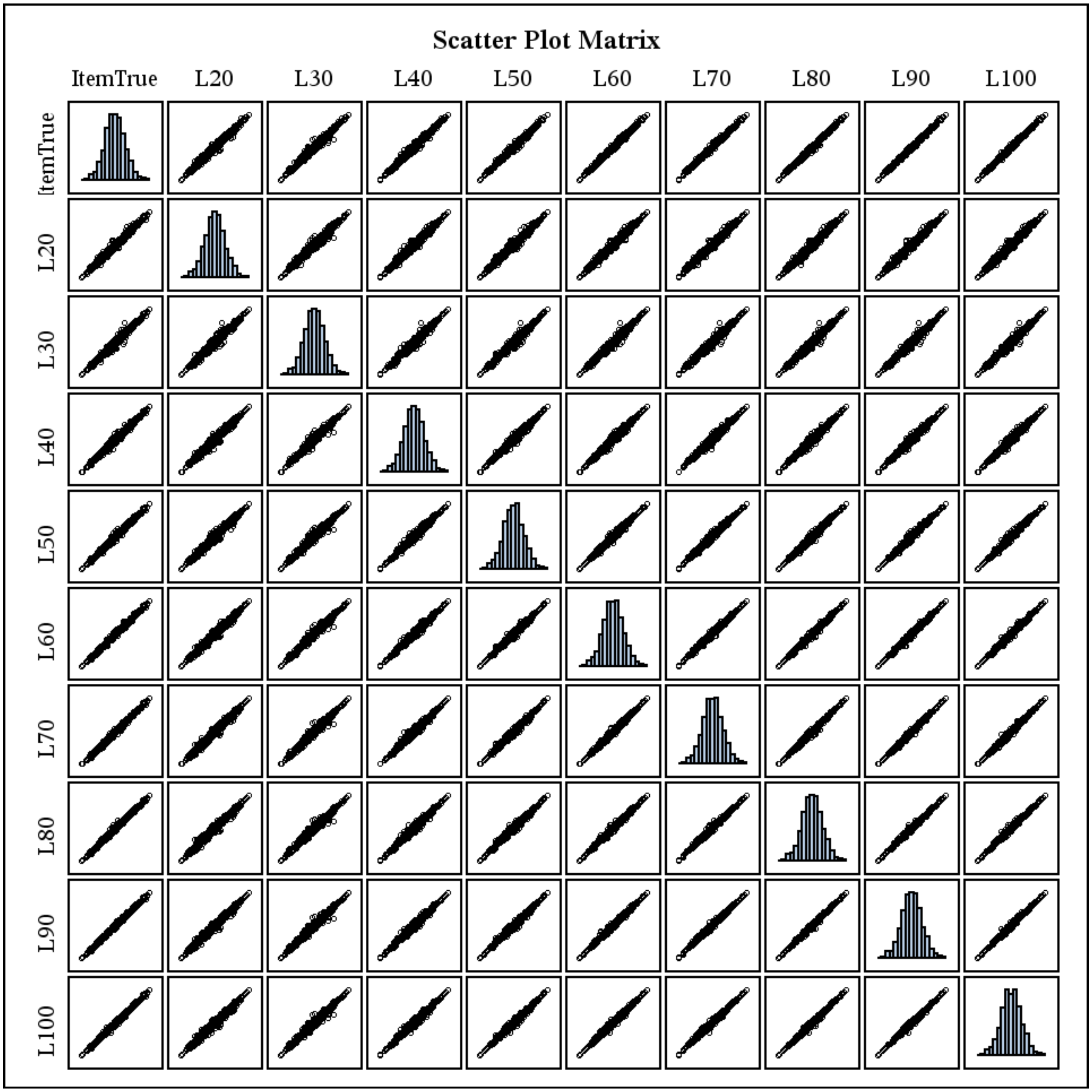


Figure 7. Plot of True and Estimated Item Parameters Across Test Length for NWEA Filtered (NWEAF) Calibration Condition with One Replication

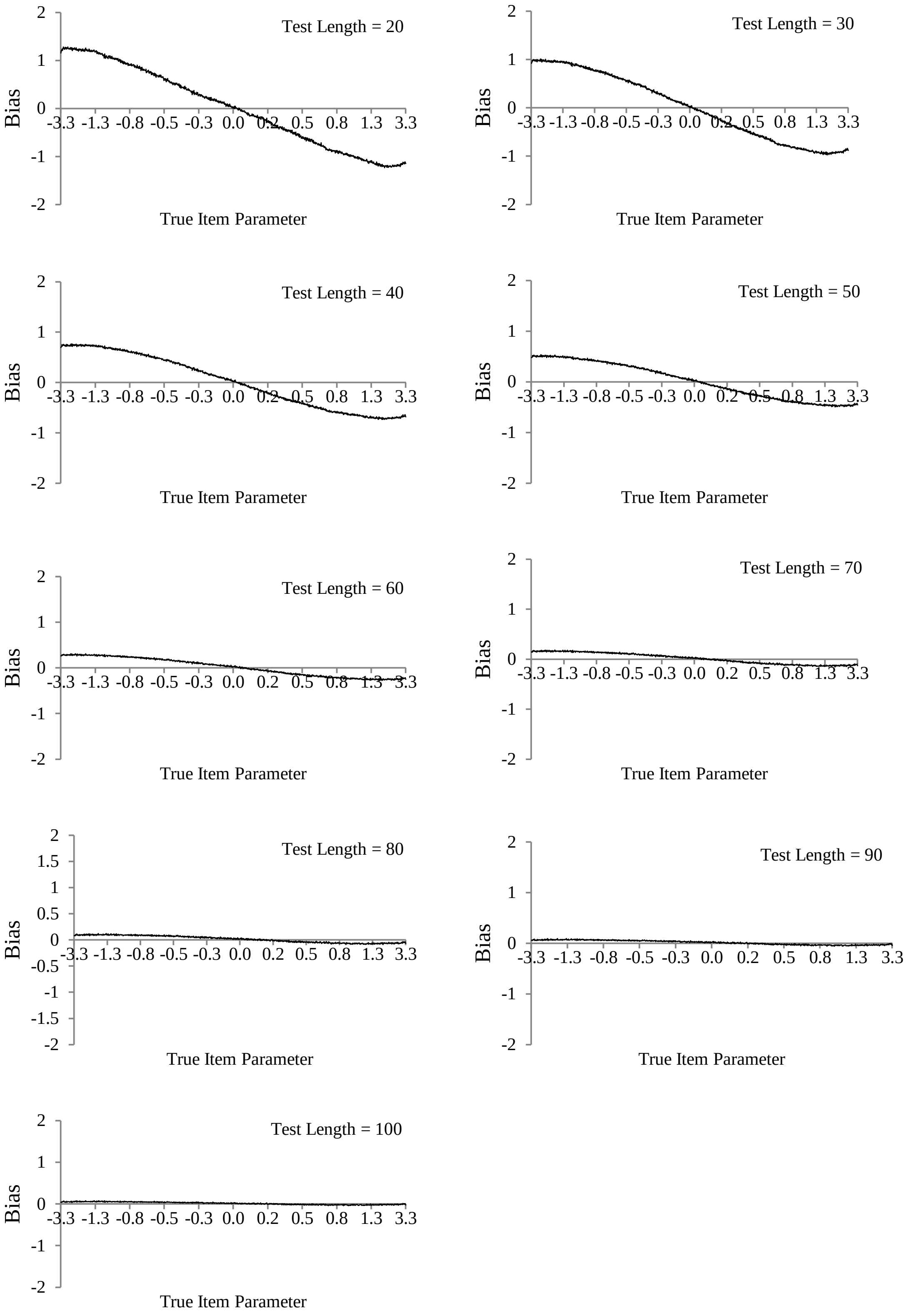


Figure 8. Distribution of Conditional Bias Along Theta Scale for Different Test Length Under Free Estimation of Sparse Data (ES) Condition with One Replication

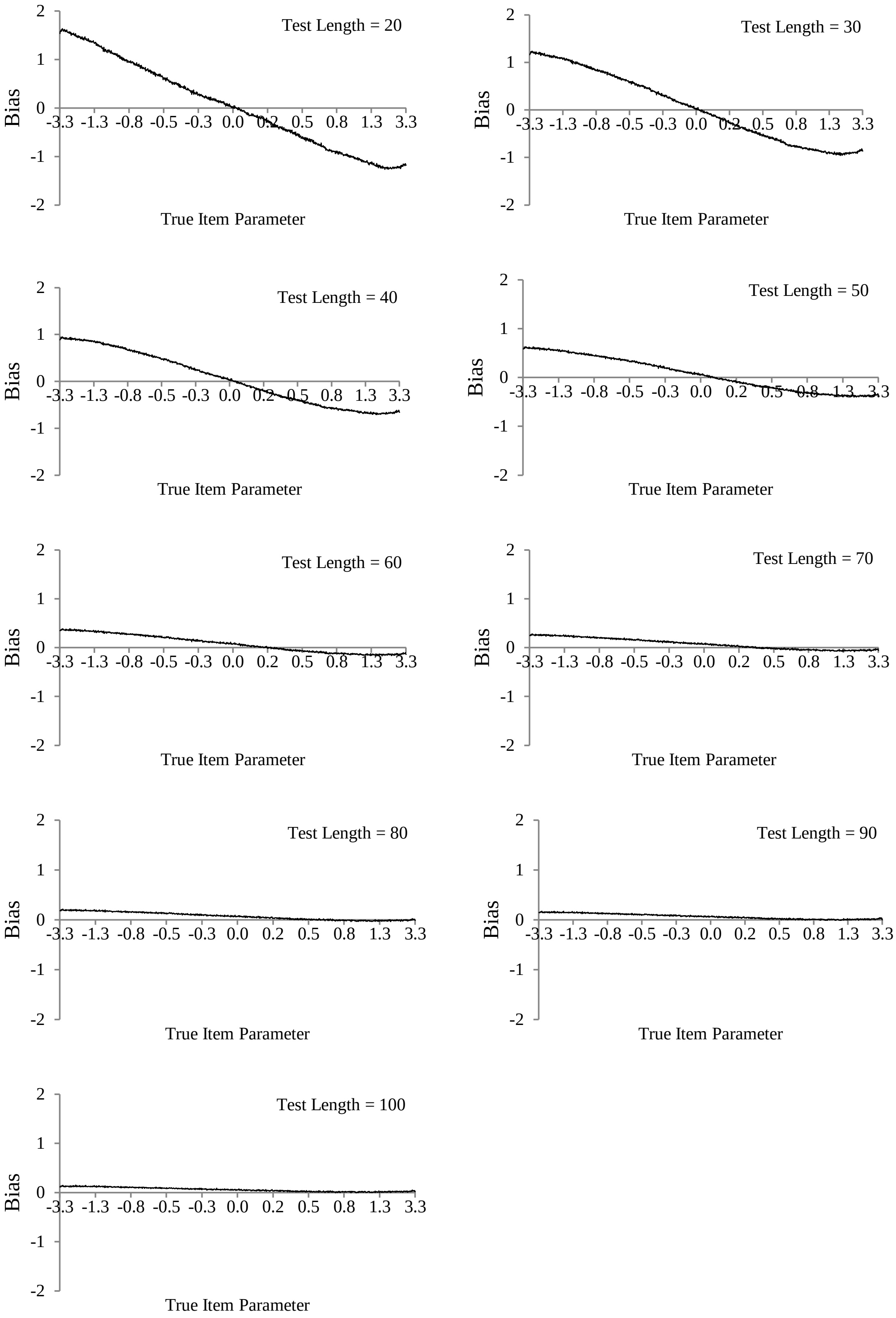


Figure 9. Distribution of Conditional Bias Along Theta Scale for Different Test Length Under Fixed True Person Parameter Estimation of Sparse Data (FTS) Condition with One Replication

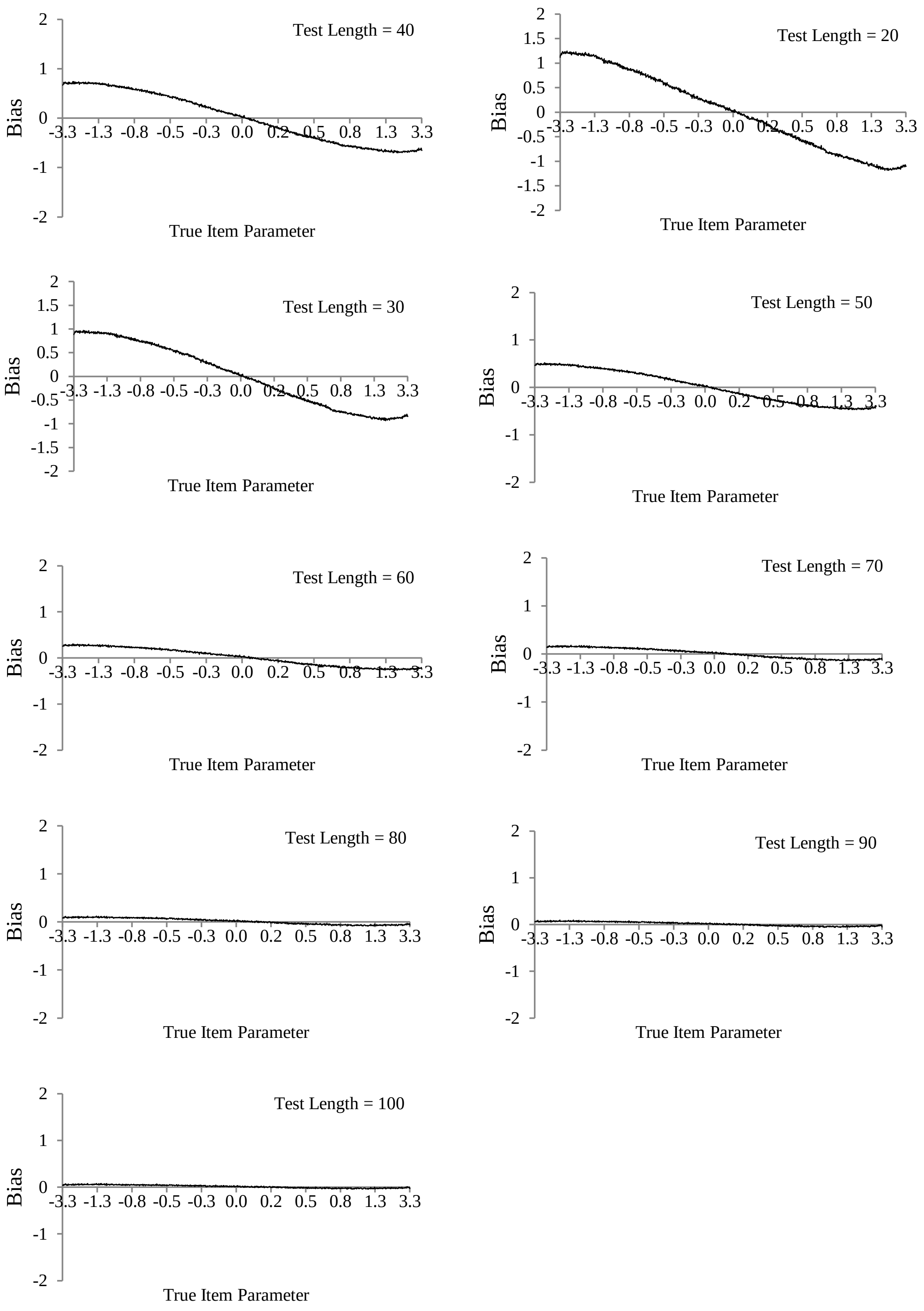


Figure 10. Distribution of Conditional Bias Along Theta Scale for Different Test Length Under Fixed Sparse Observed Person Parameter Estimation of Sparse Data (FSOS) Condition with One Replication

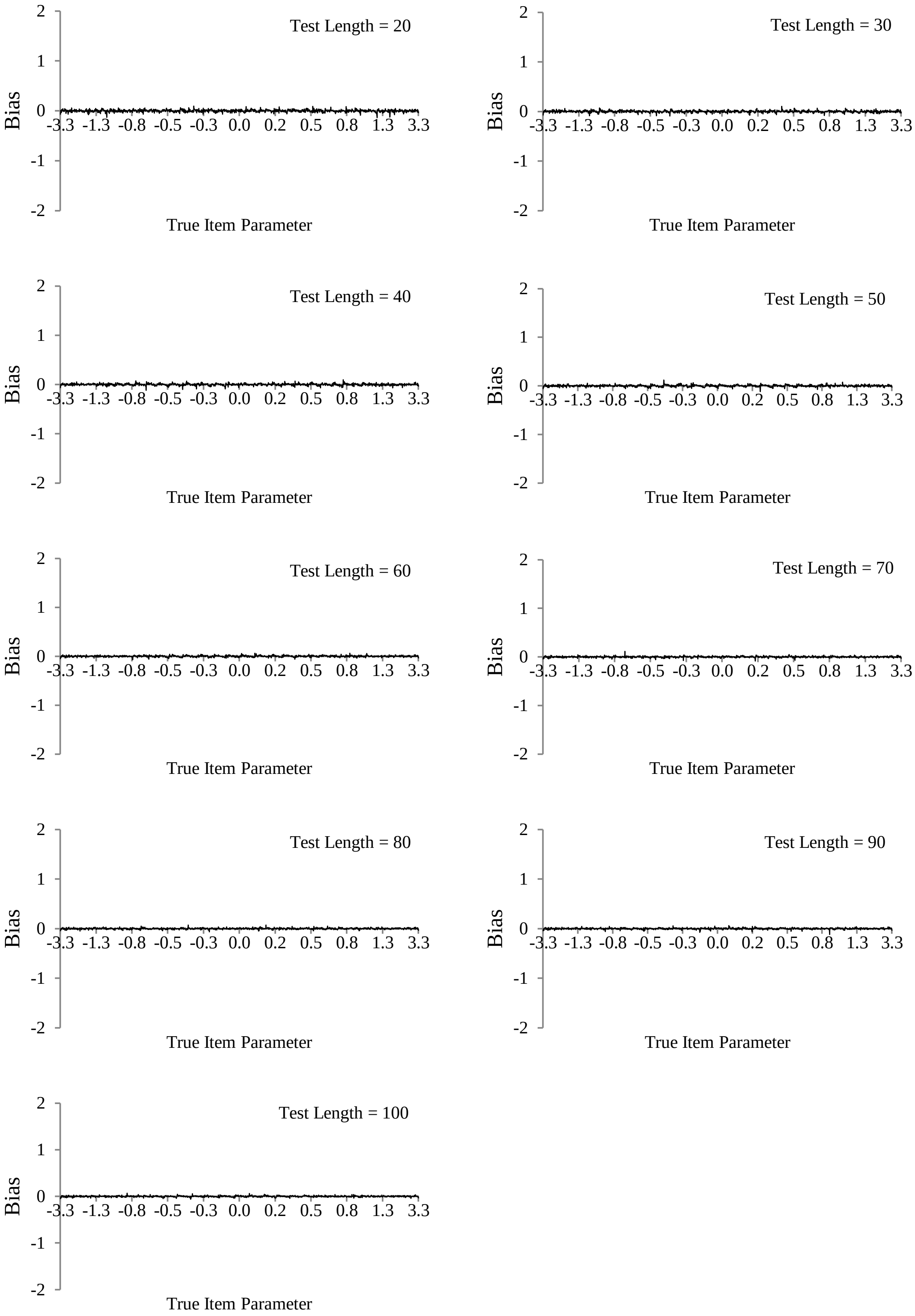


Figure 11. Distribution of Conditional Bias Along Theta Scale for Different Test Length Under NWEA Non-Filtered (NWEANF) Calibration Condition with One Replication

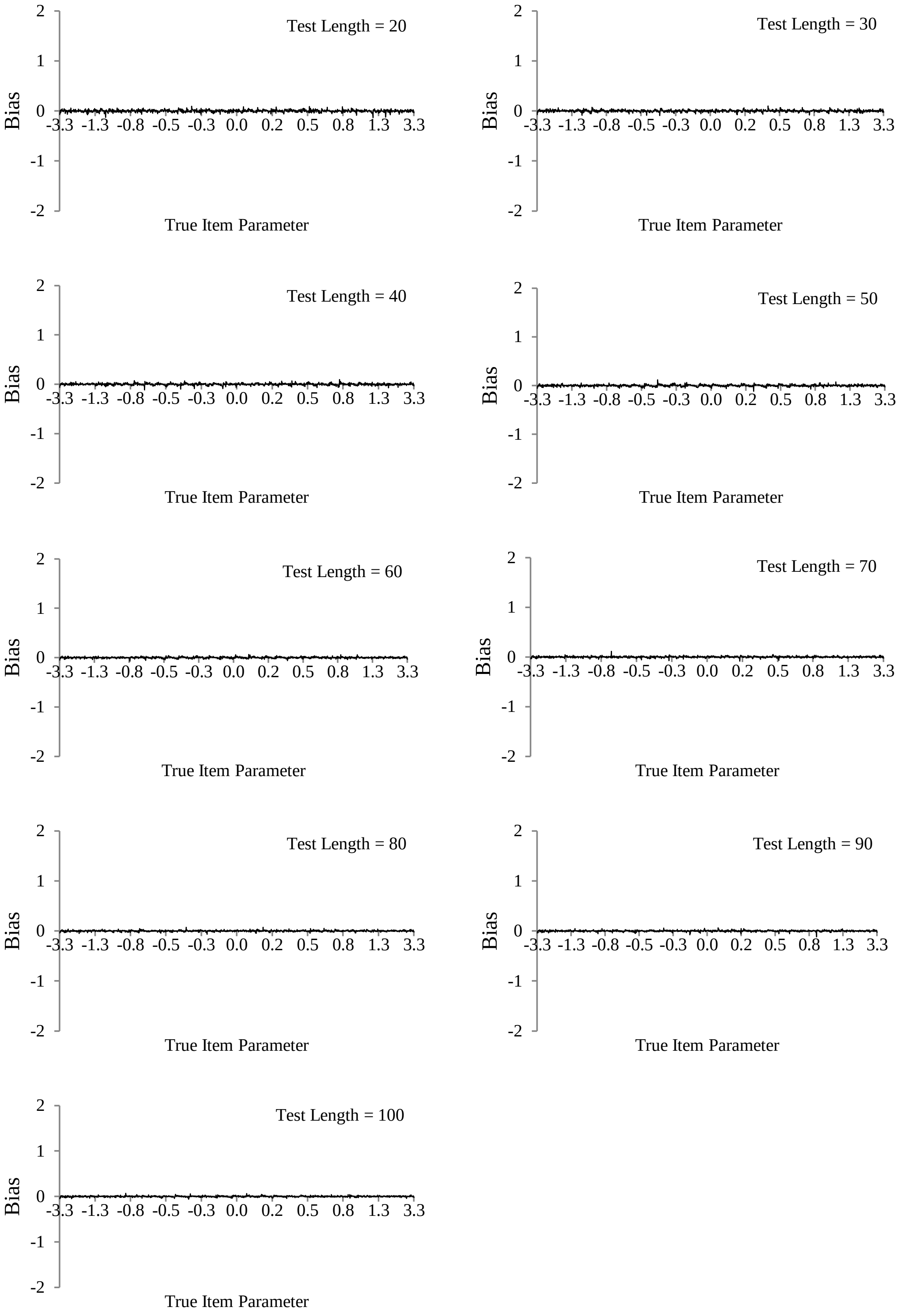


Figure 12. Distribution of Conditional Bias Along Theta Scale for Different Test Length Under NWEA Filtered (NWEAF) Calibration Condition with One Replication

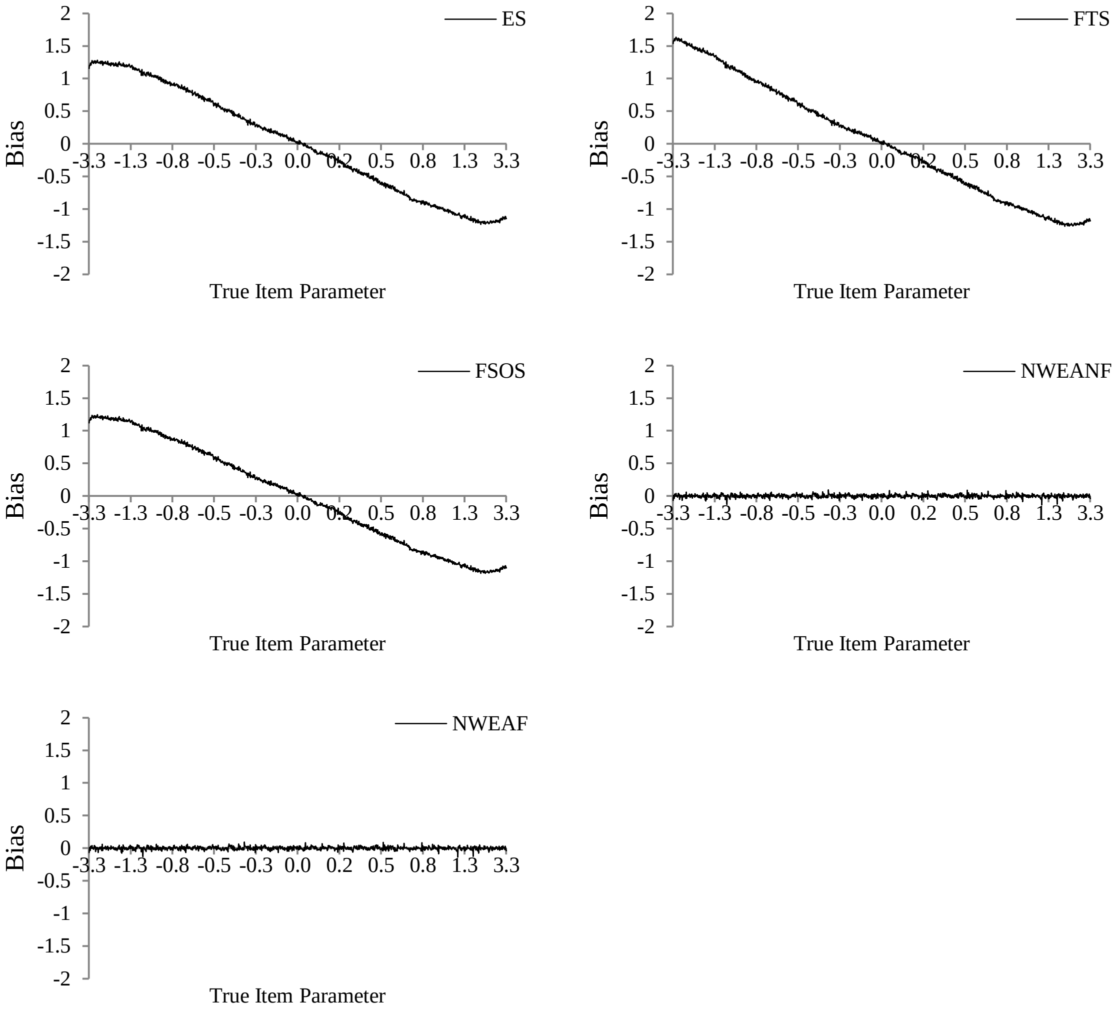


Figure 13. Distribution of Conditional Bias Along Theta Scale for Different Test Calibration Conditions for Test Length 20 with One Replication

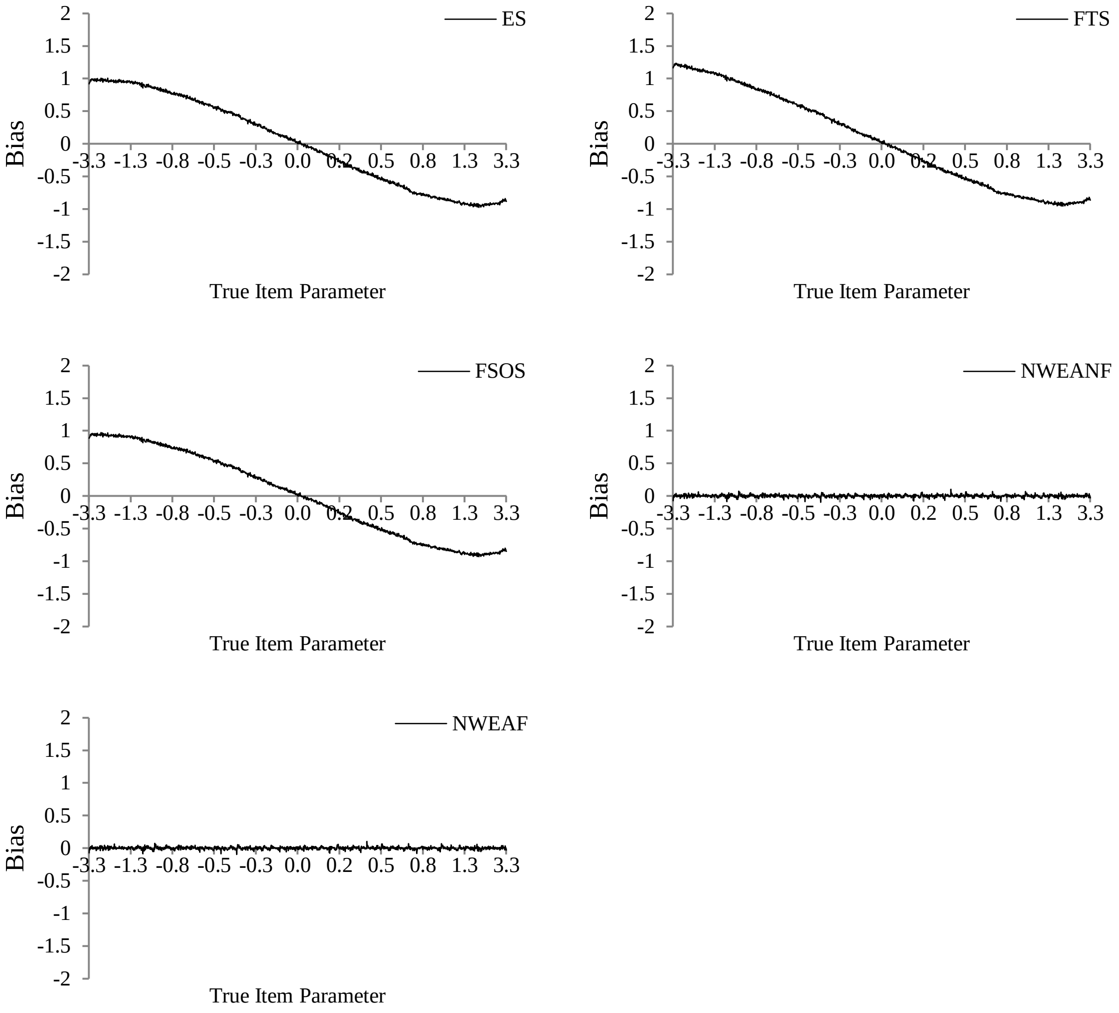


Figure 14. Distribution of Conditional Bias Along Theta Scale for Different Test Calibration Conditions for Test Length 30 with One Replication

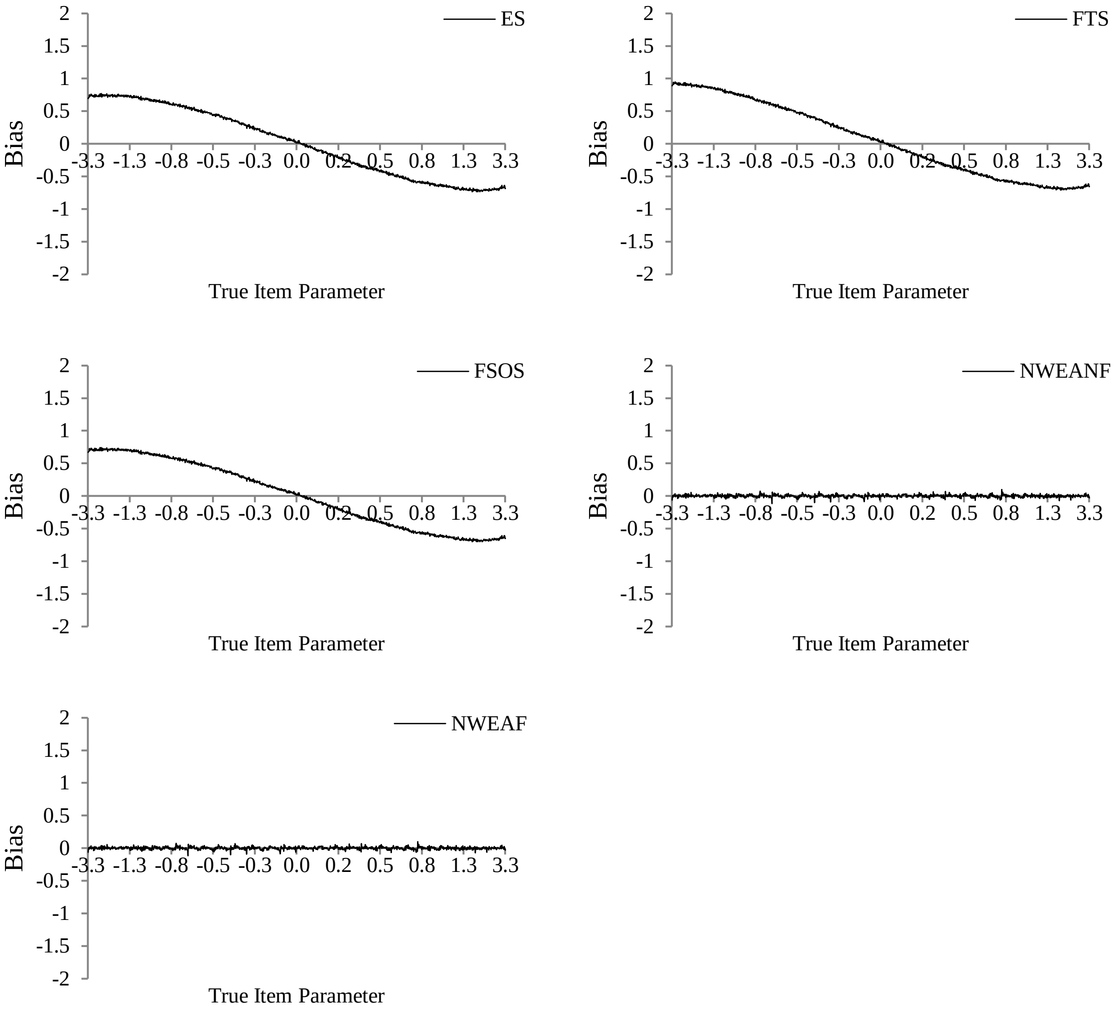


Figure 15. Distribution of Conditional Bias Along Theta Scale for Different Test Calibration Conditions for Test Length 40 with One Replication

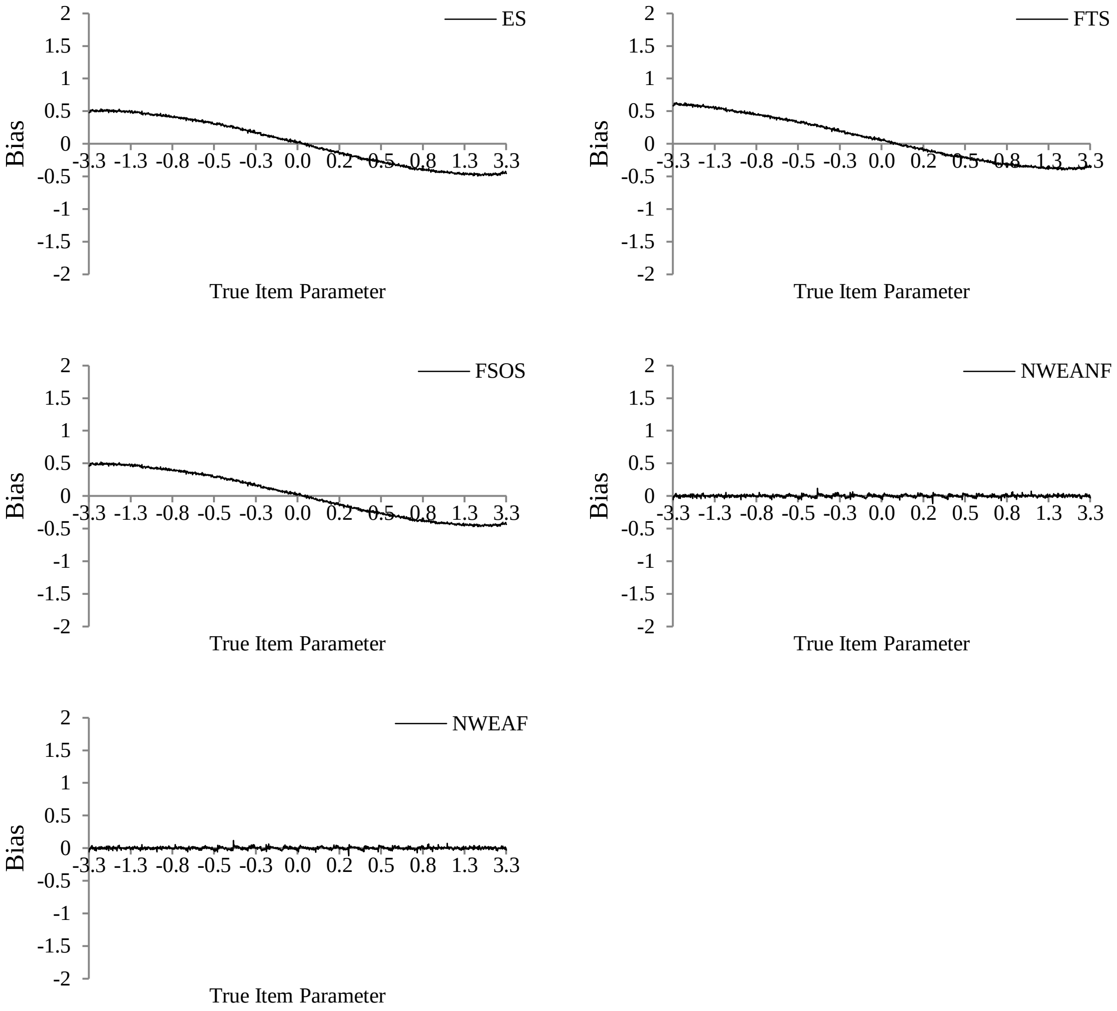


Figure 16. Distribution of Conditional Bias Along Theta Scale for Different Test Calibration Conditions for Test Length 50 with One Replication

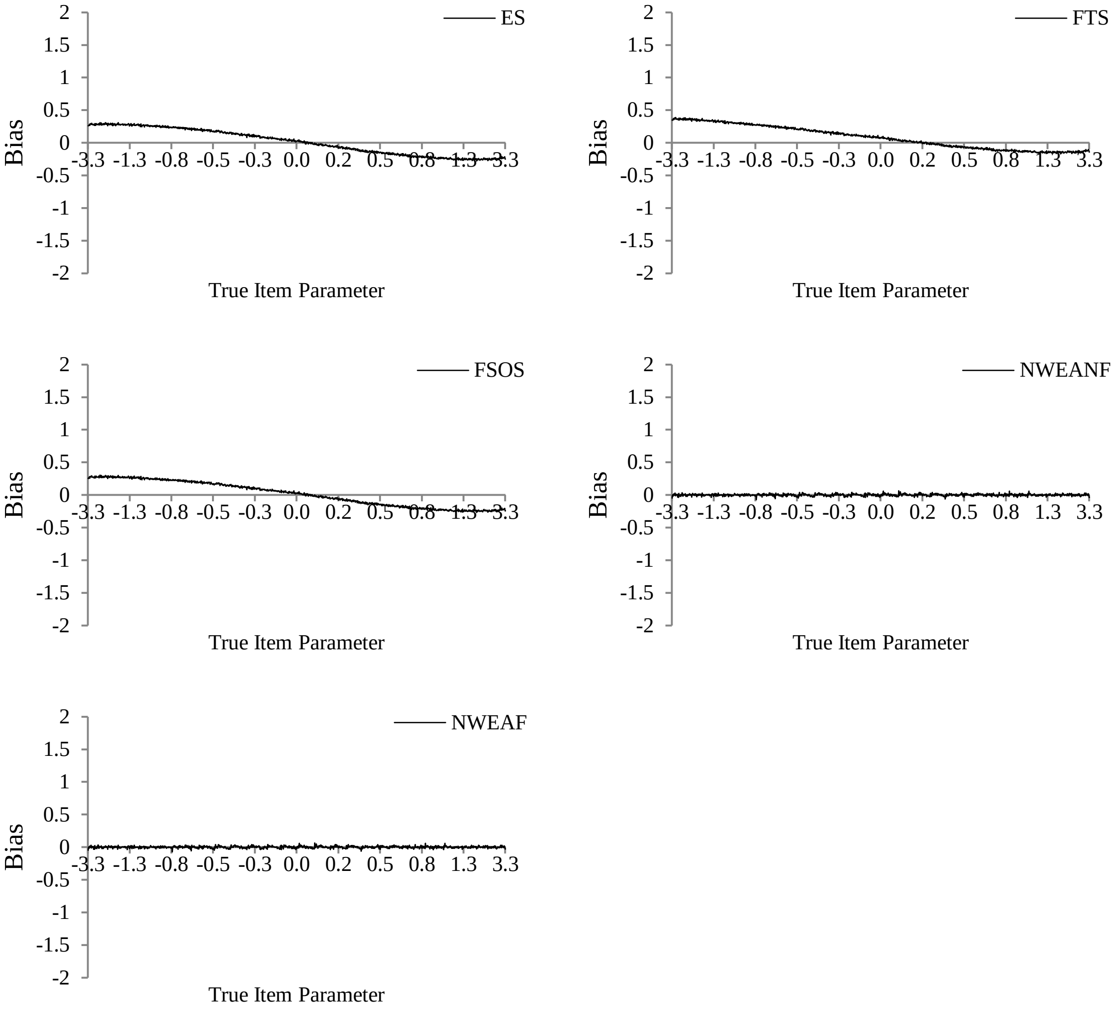


Figure 17. Distribution of Conditional Bias Along Theta Scale for Different Test Calibration Conditions for Test Length 60 with One Replication

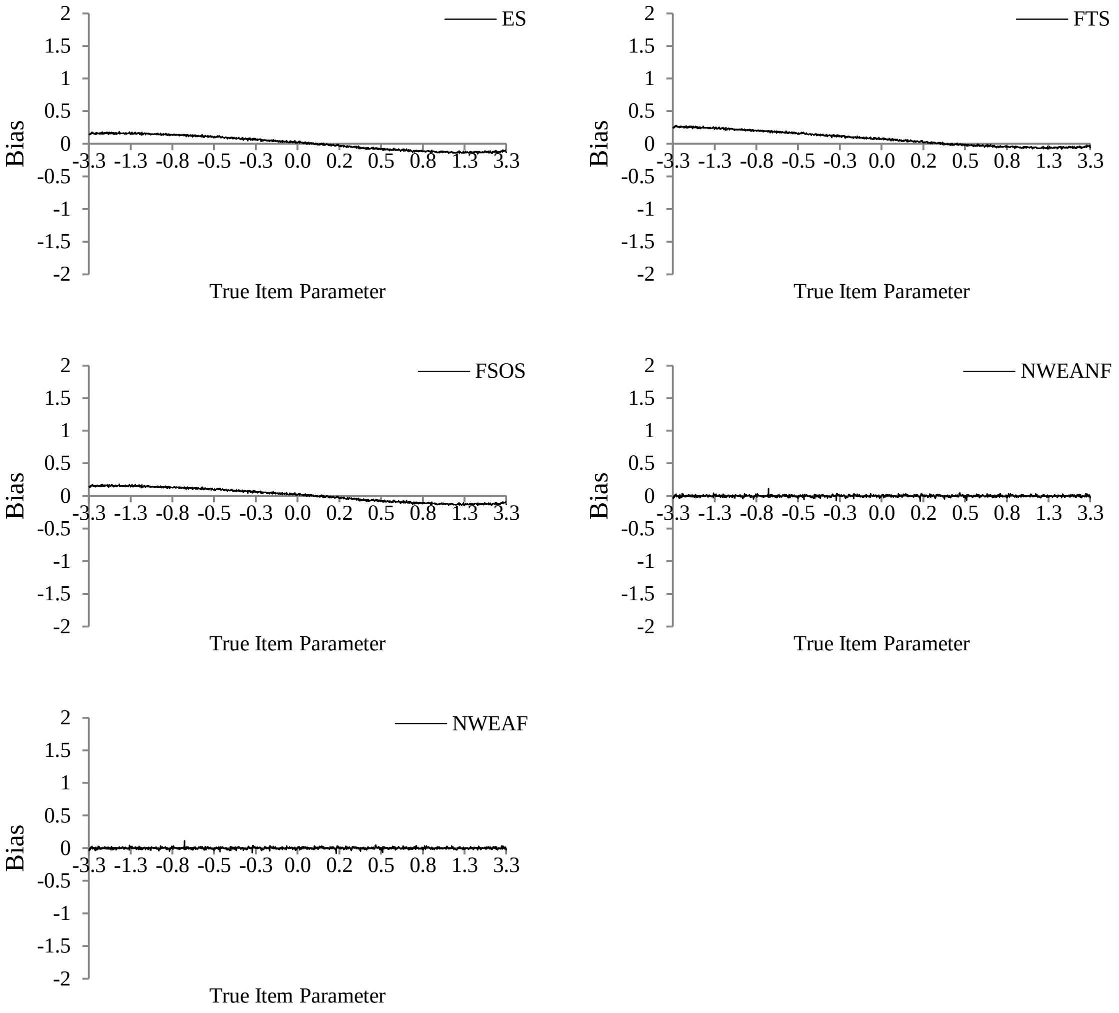


Figure 18. Distribution of Conditional Bias Along Theta Scale for Different Test Calibration Conditions for Test Length 70 with One Replication

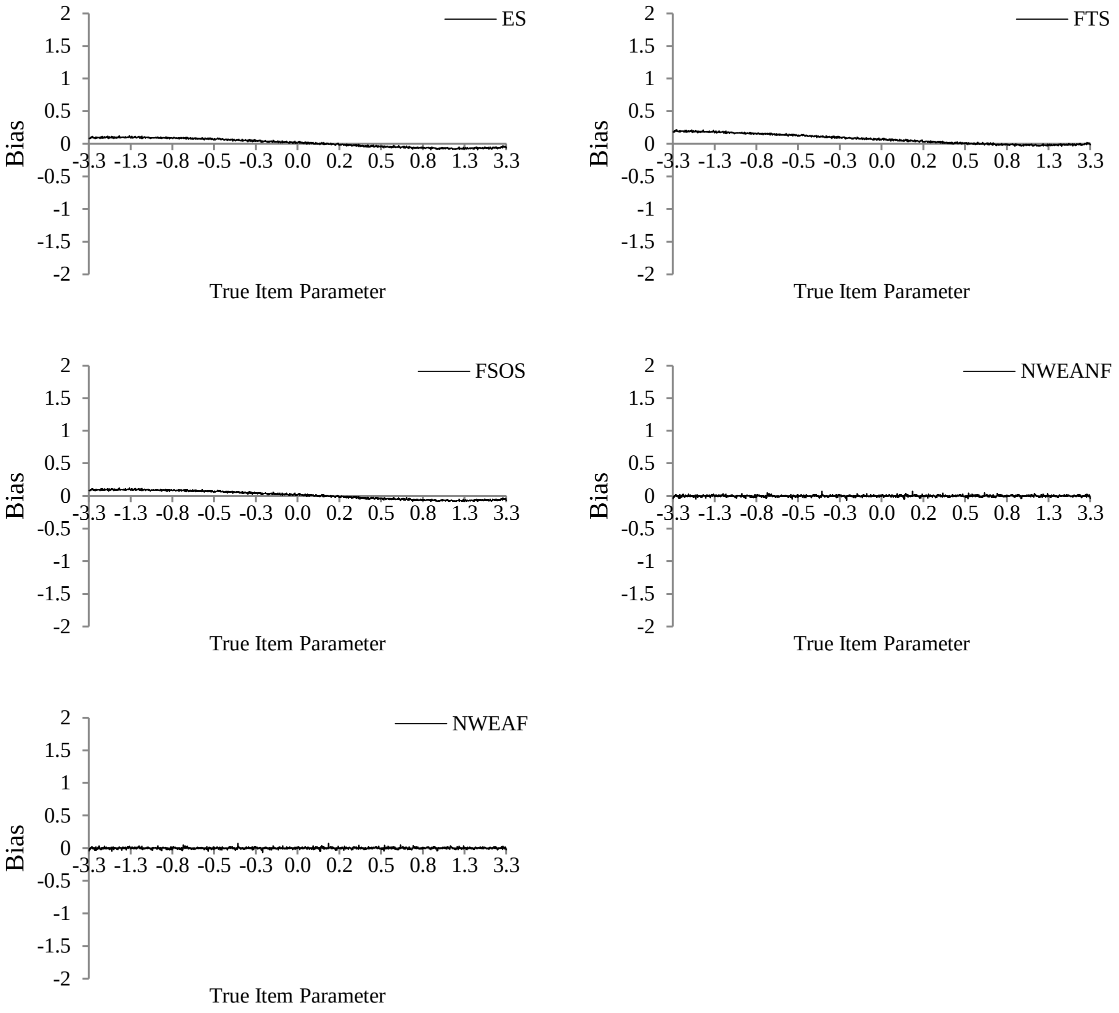


Figure 19. Distribution of Conditional Bias Along Theta Scale for Different Test Calibration Conditions for Test Length 80 with One Replication

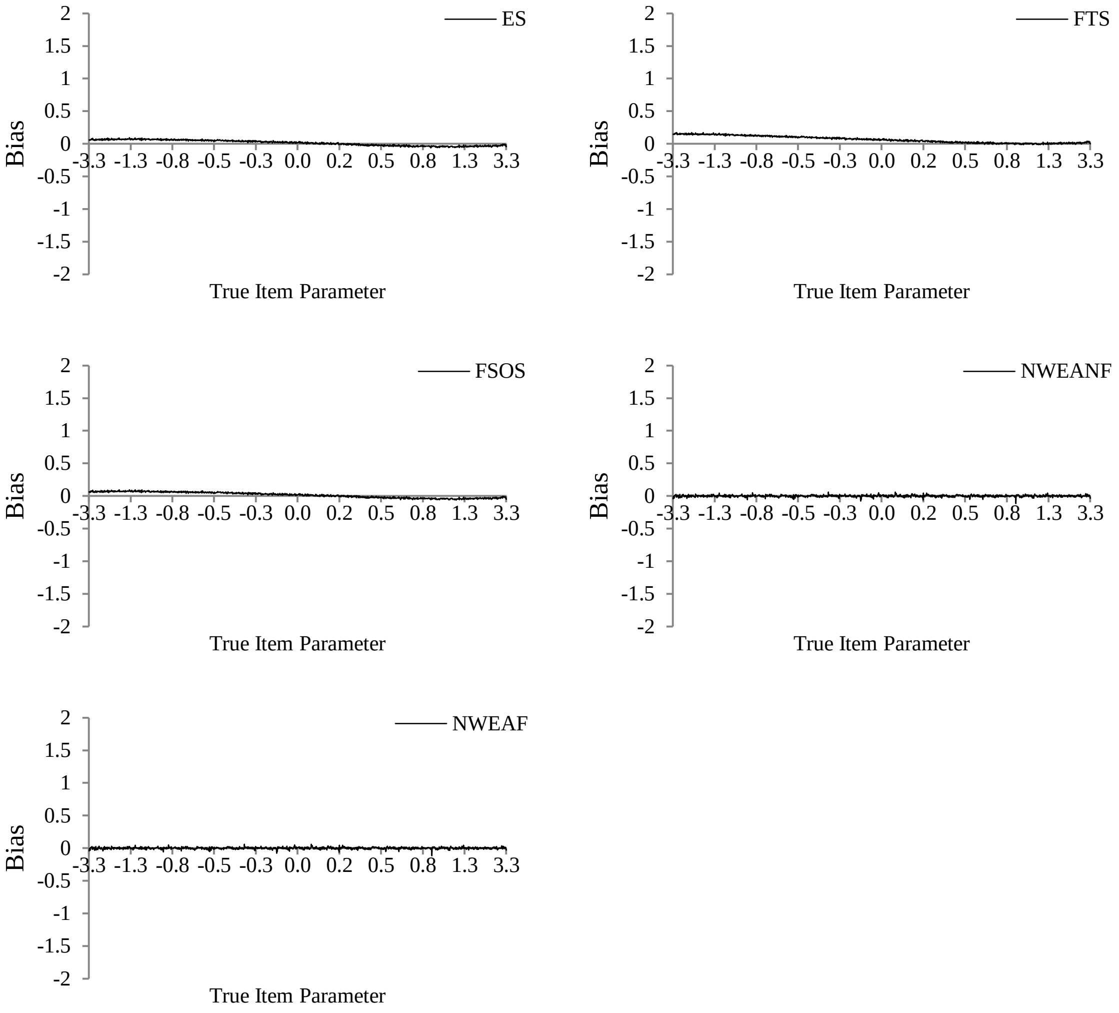


Figure 20. Distribution of Conditional Bias Along Theta Scale for Different Test Calibration Conditions for Test Length 90 with One Replication

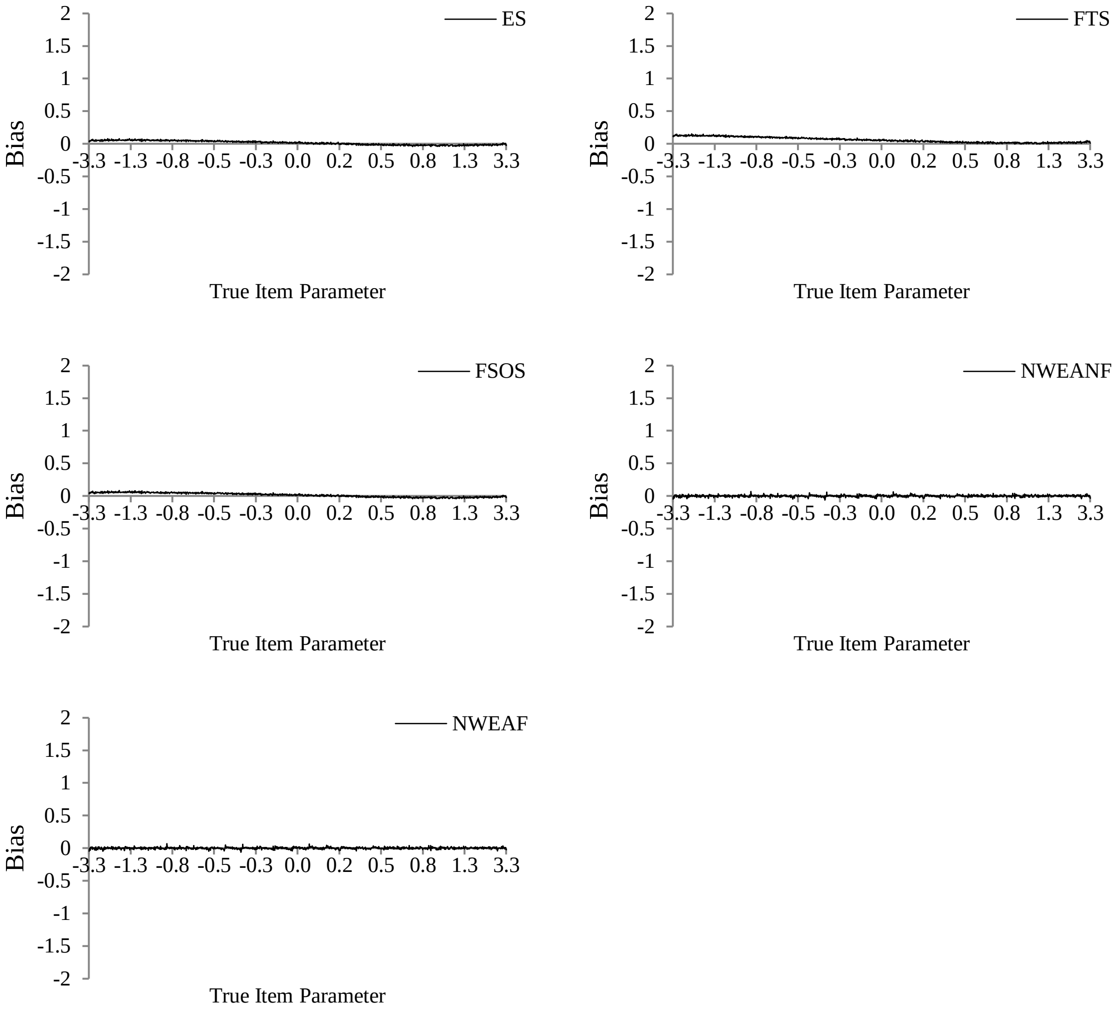


Figure 21. Distribution of Conditional Bias Along Theta Scale for Different Test Calibration Conditions for Test Length 100 with One Replication

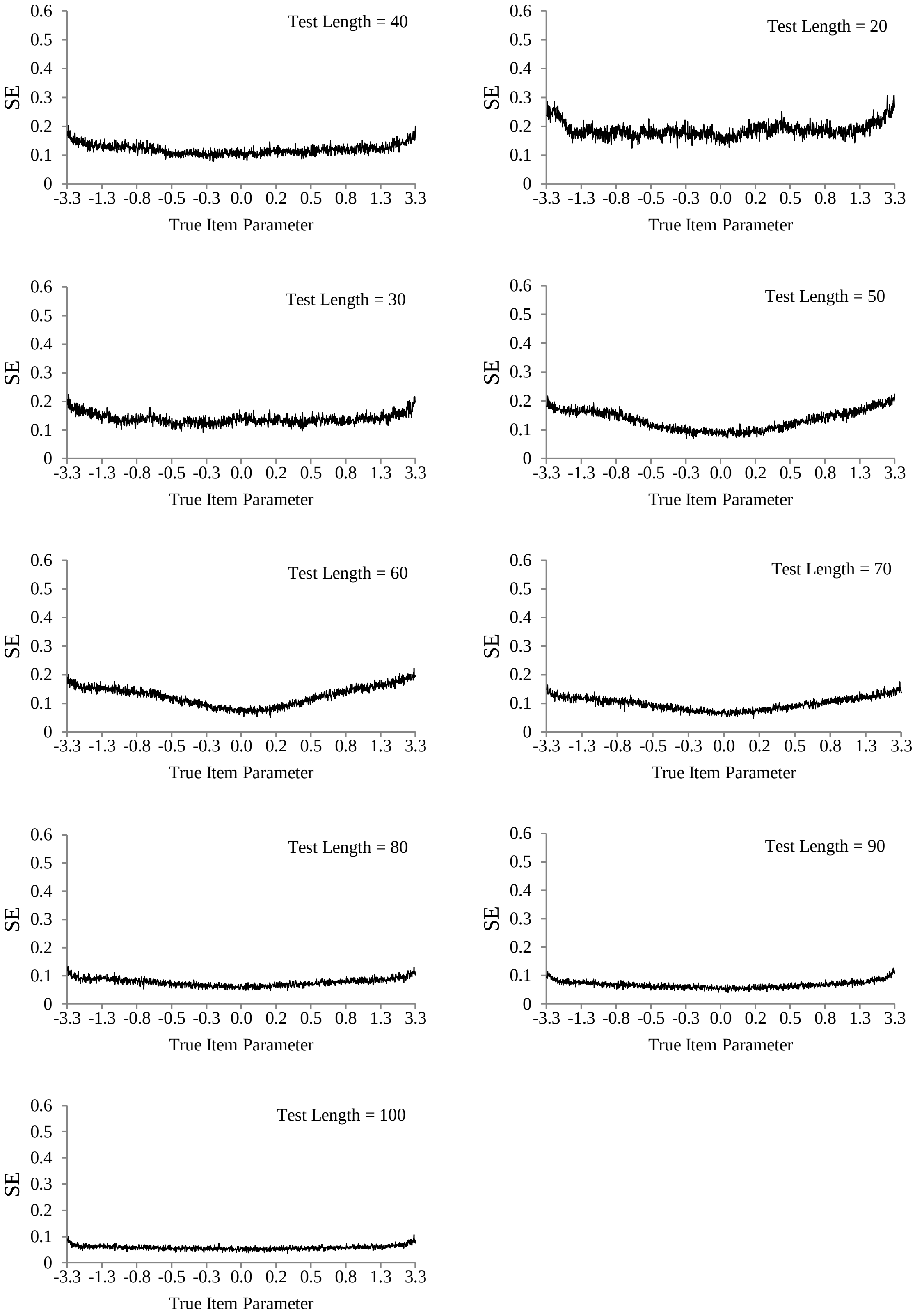


Figure 22. Distribution of Conditional SE Along Theta Scale for Different Test Length Under Free Estimation of Sparse Data (ES) Condition with One Replication

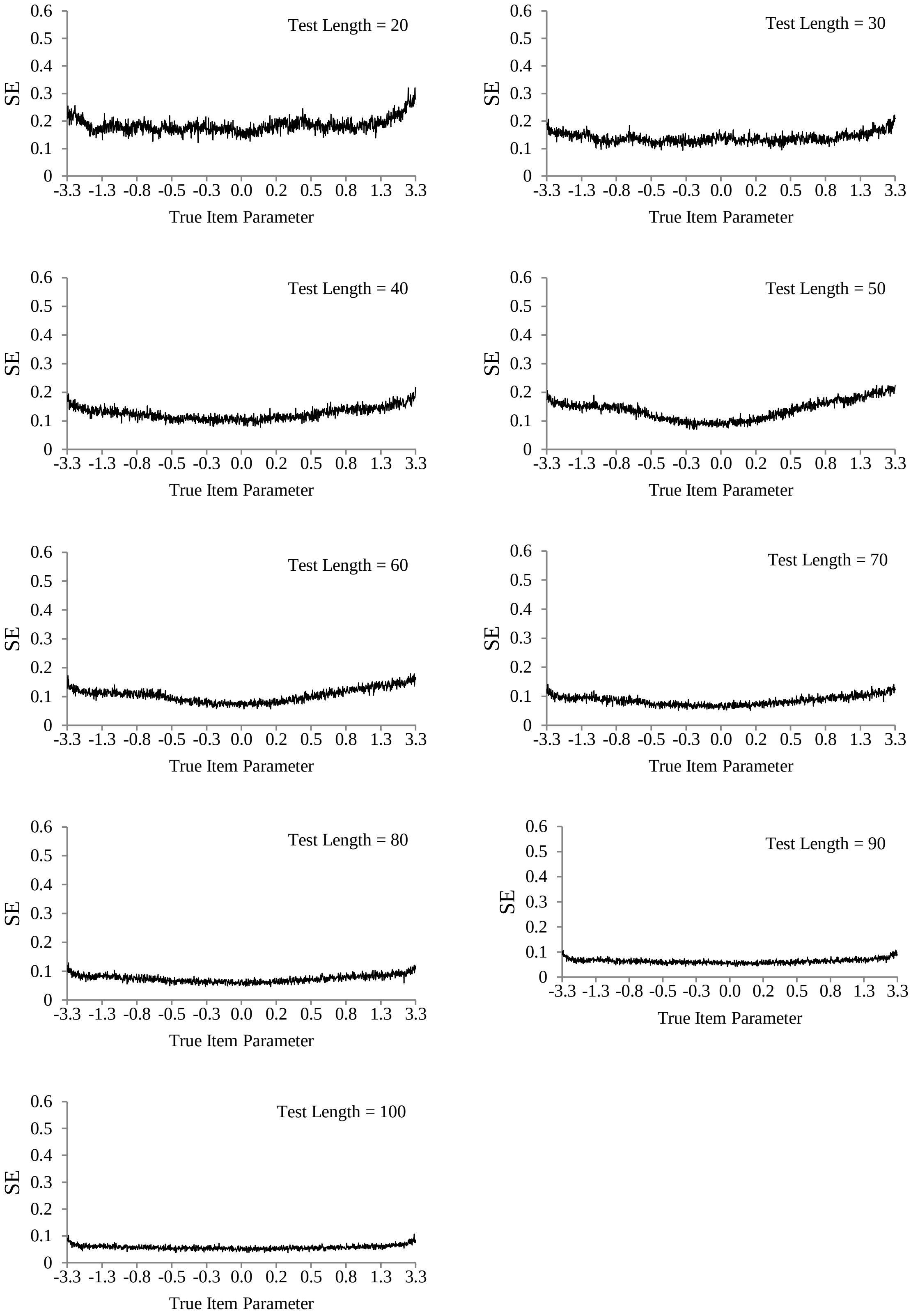


Figure 23. Distribution of Conditional SE Along Theta Scale for Different Test Length Under Fixed True Person Parameter Estimation of Sparse Data (FTS) Condition with One Replication

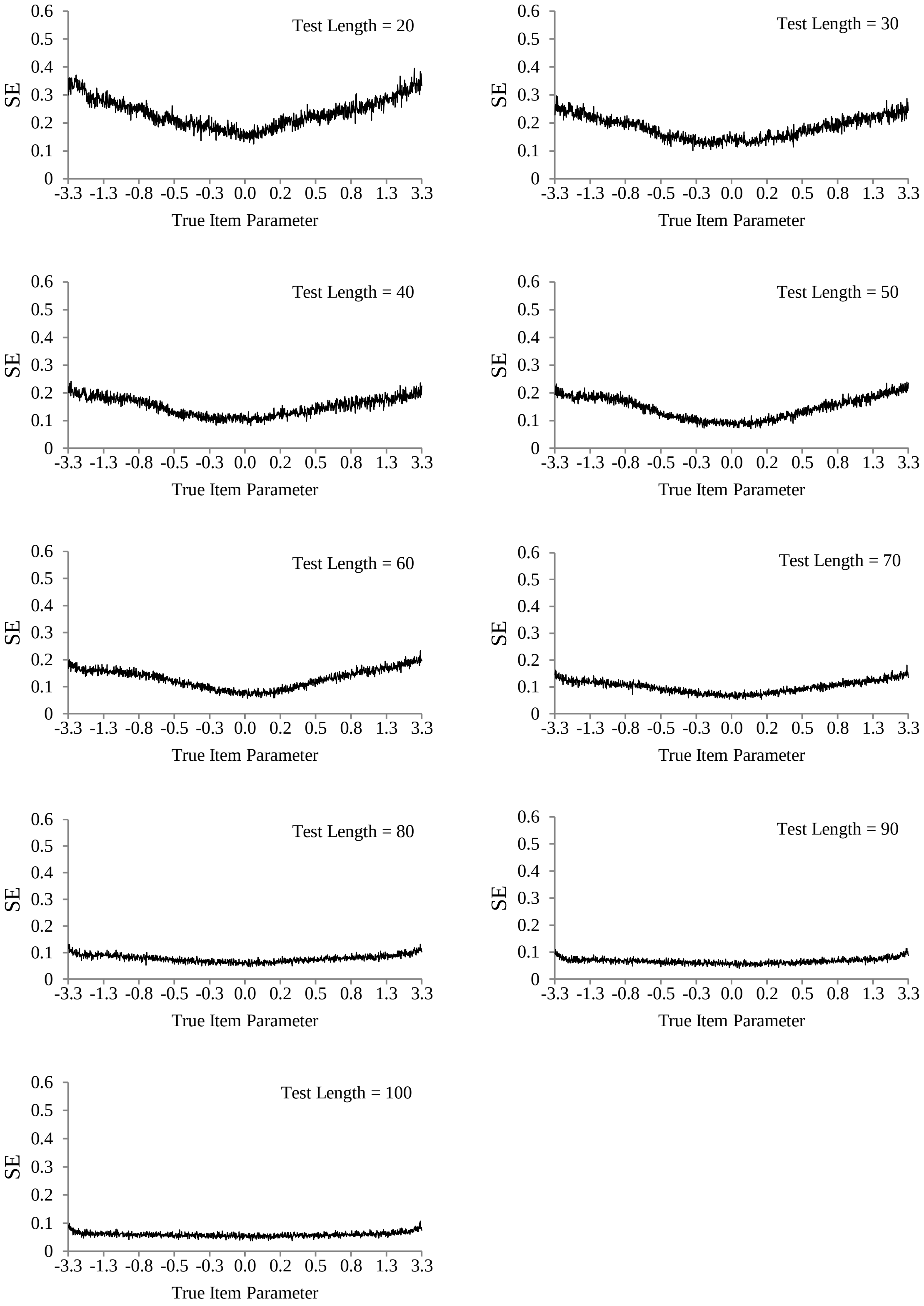


Figure 24. Distribution of Conditional SE Along Theta Scale for Different Test Length Under Fixed Sparse Observed Person Parameter Estimation of Sparse Data (FSOS) Condition with One Replication

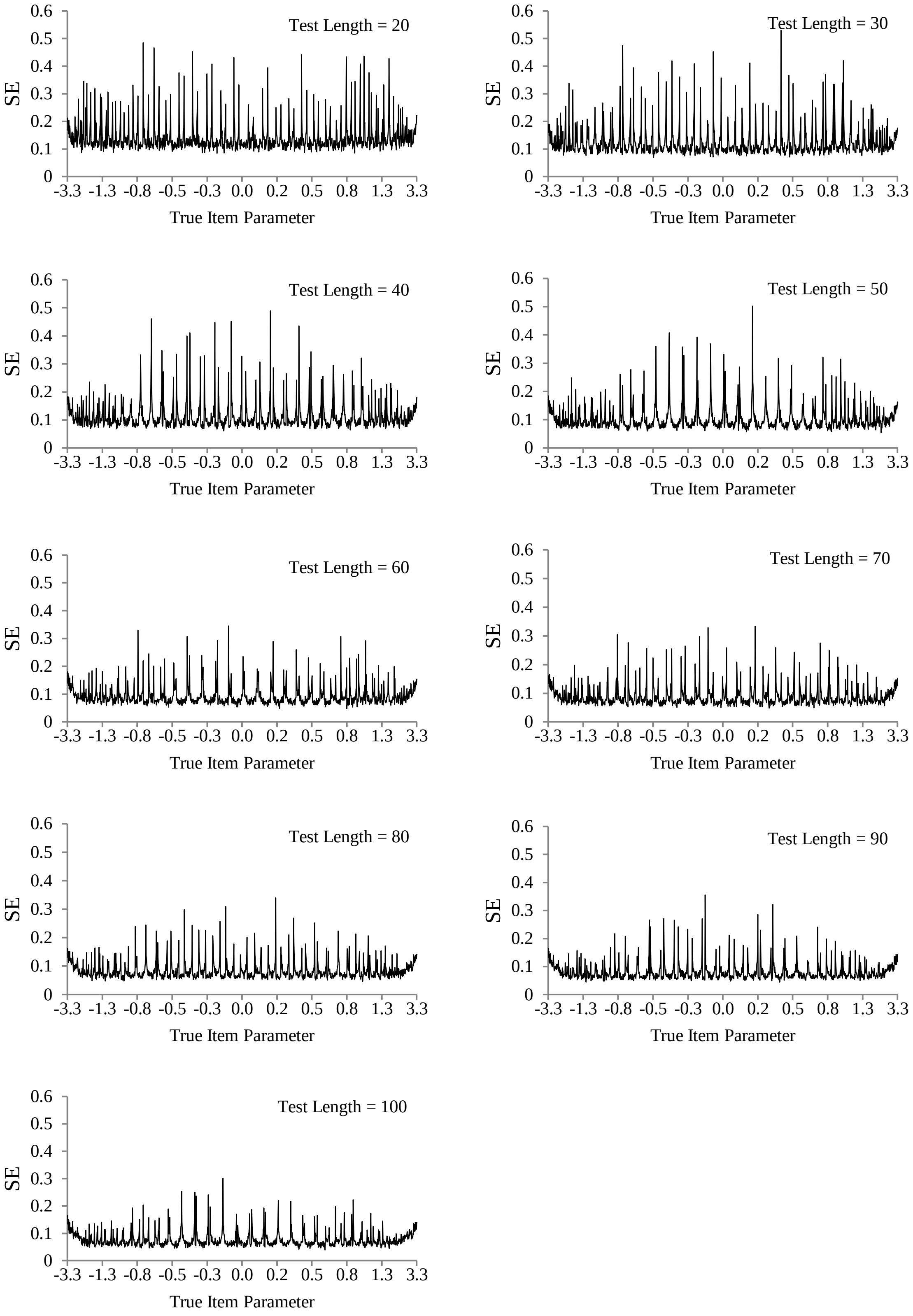


Figure 25. Distribution of Conditional SE Along Theta Scale for Different Test Length Under NWEA Non-Filtered (NWEANF) Calibration Condition with One Replication

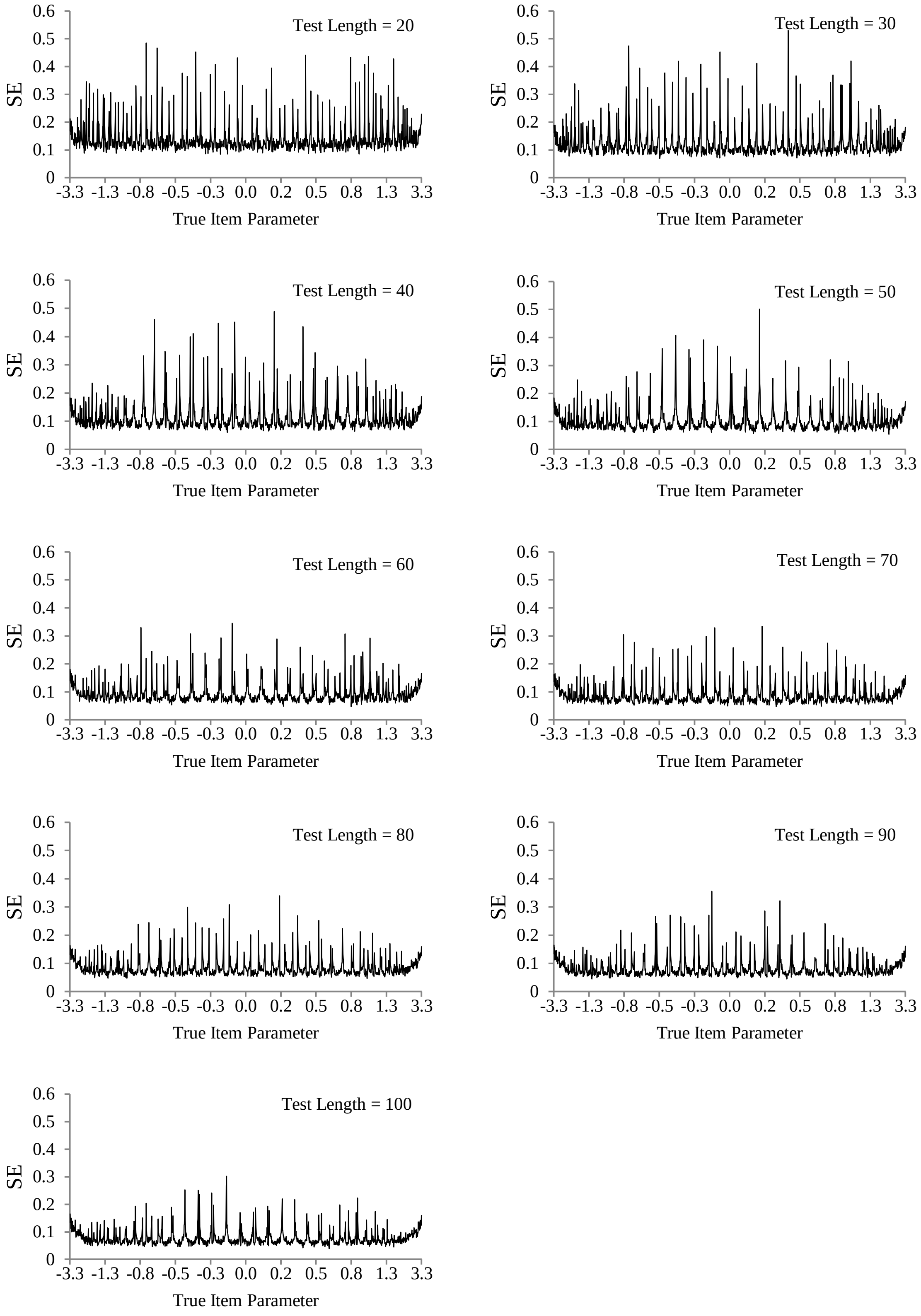


Figure 26. Distribution of Conditional SE Along Theta Scale for Different Test Length Under NWEA Filtered (NWEAF) Calibration Condition with One Replication

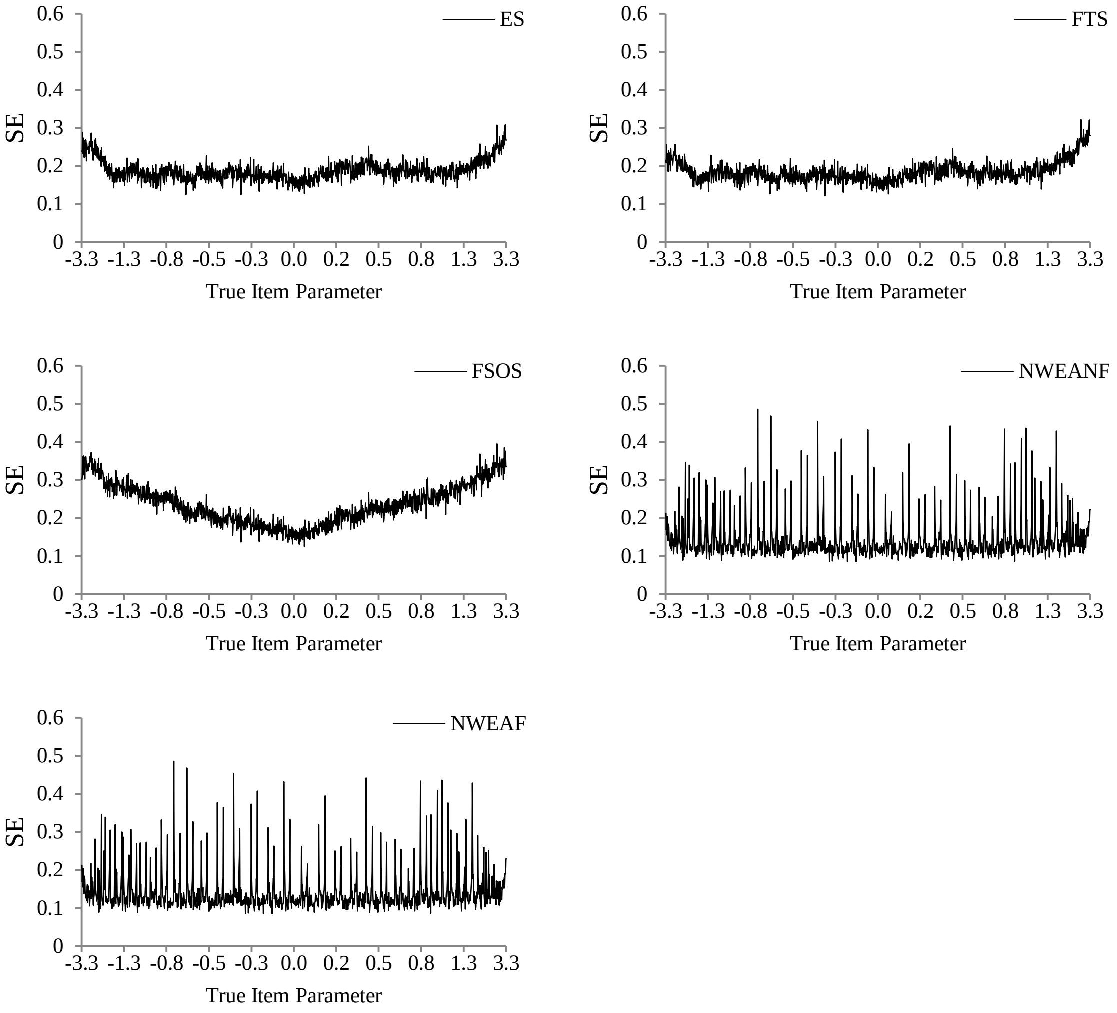


Figure 27. Distribution of Conditional SE Along Theta Scale for Different Test Calibration Conditions for Test Length 20 with One Replication

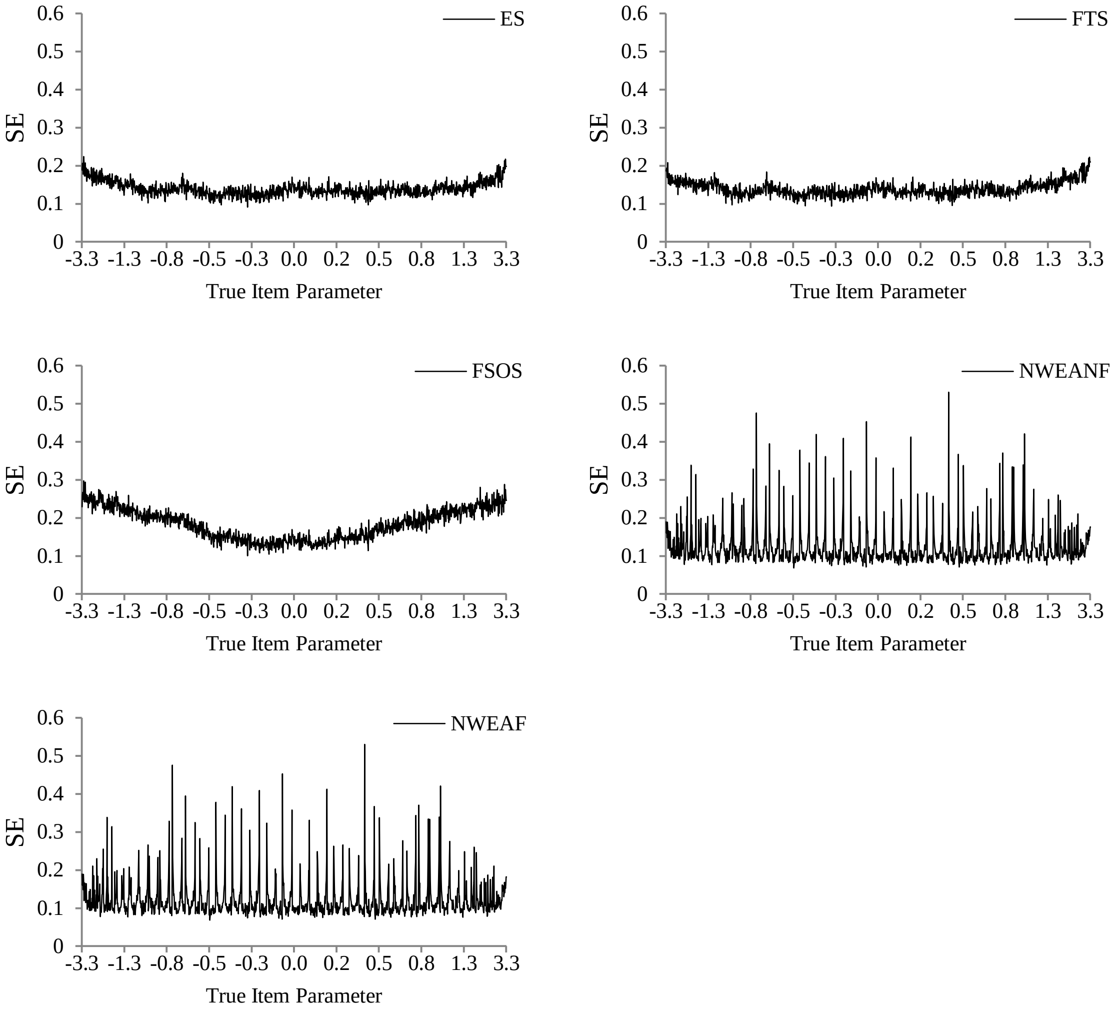


Figure 28. Distribution of Conditional SE Along Theta Scale for Different Test Calibration Conditions for Test Length 30 with One Replication

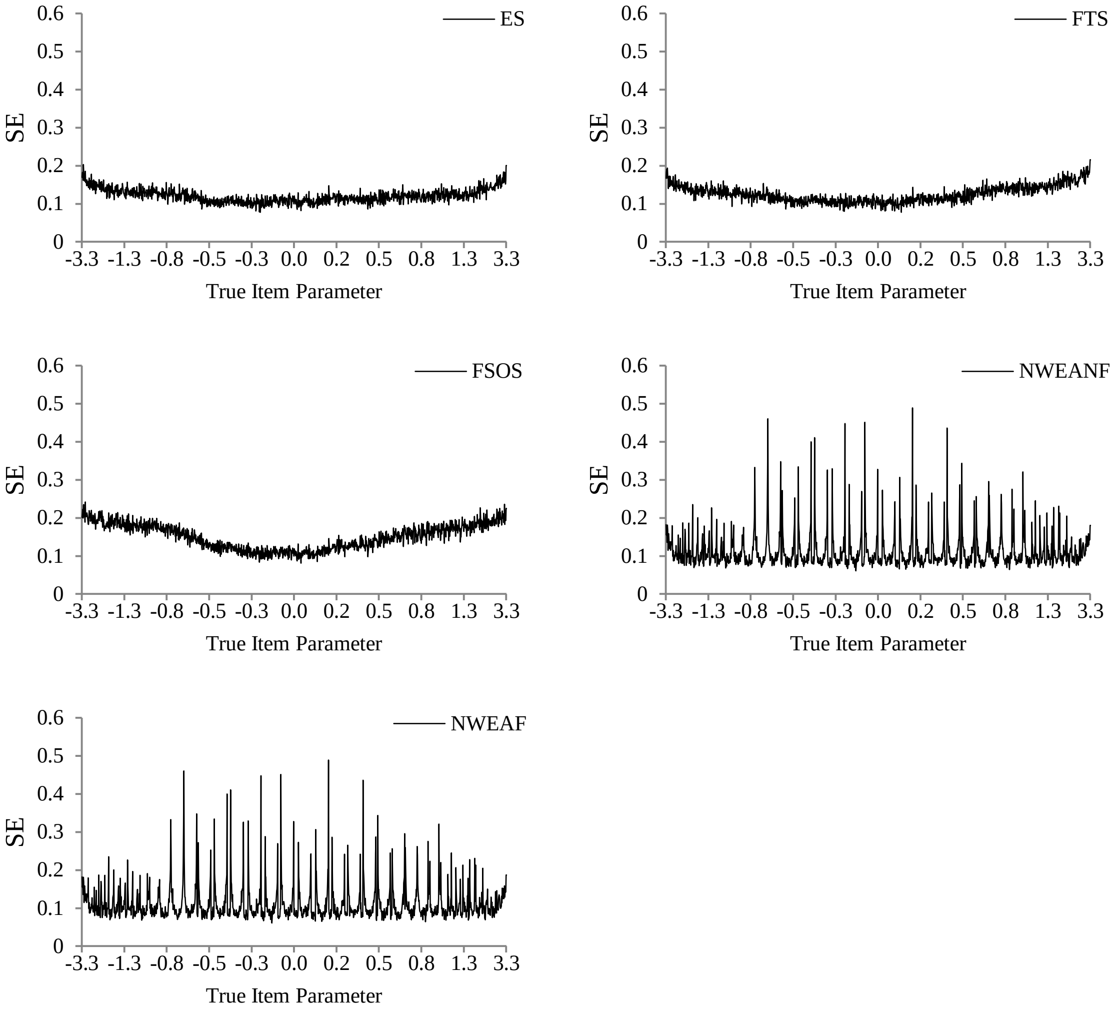


Figure 29. Distribution of Conditional SE Along Theta Scale for Different Test Calibration Conditions for Test Length 40 with One Replication

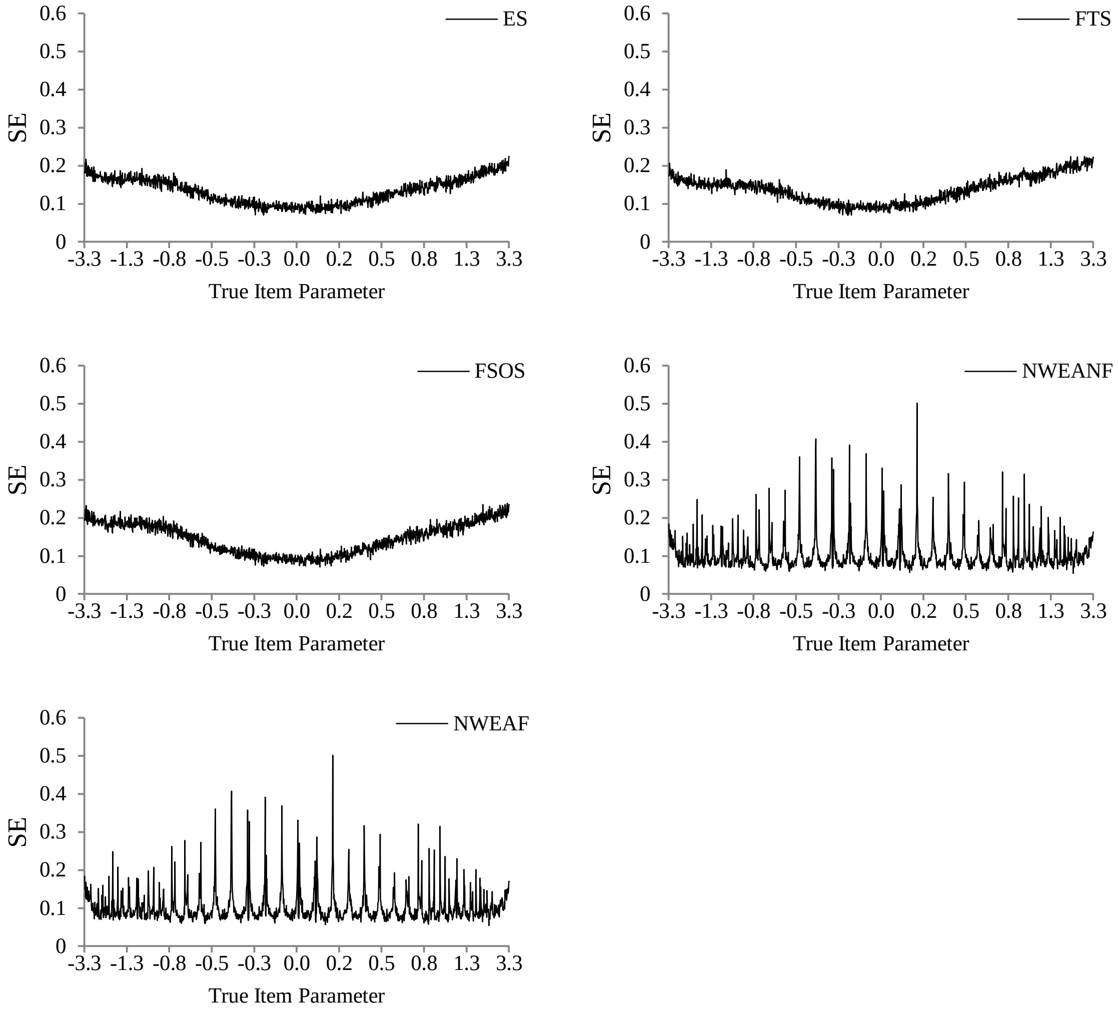


Figure 30. Distribution of Conditional SE Along Theta Scale for Different Test Calibration Conditions for Test Length 50 with One Replication

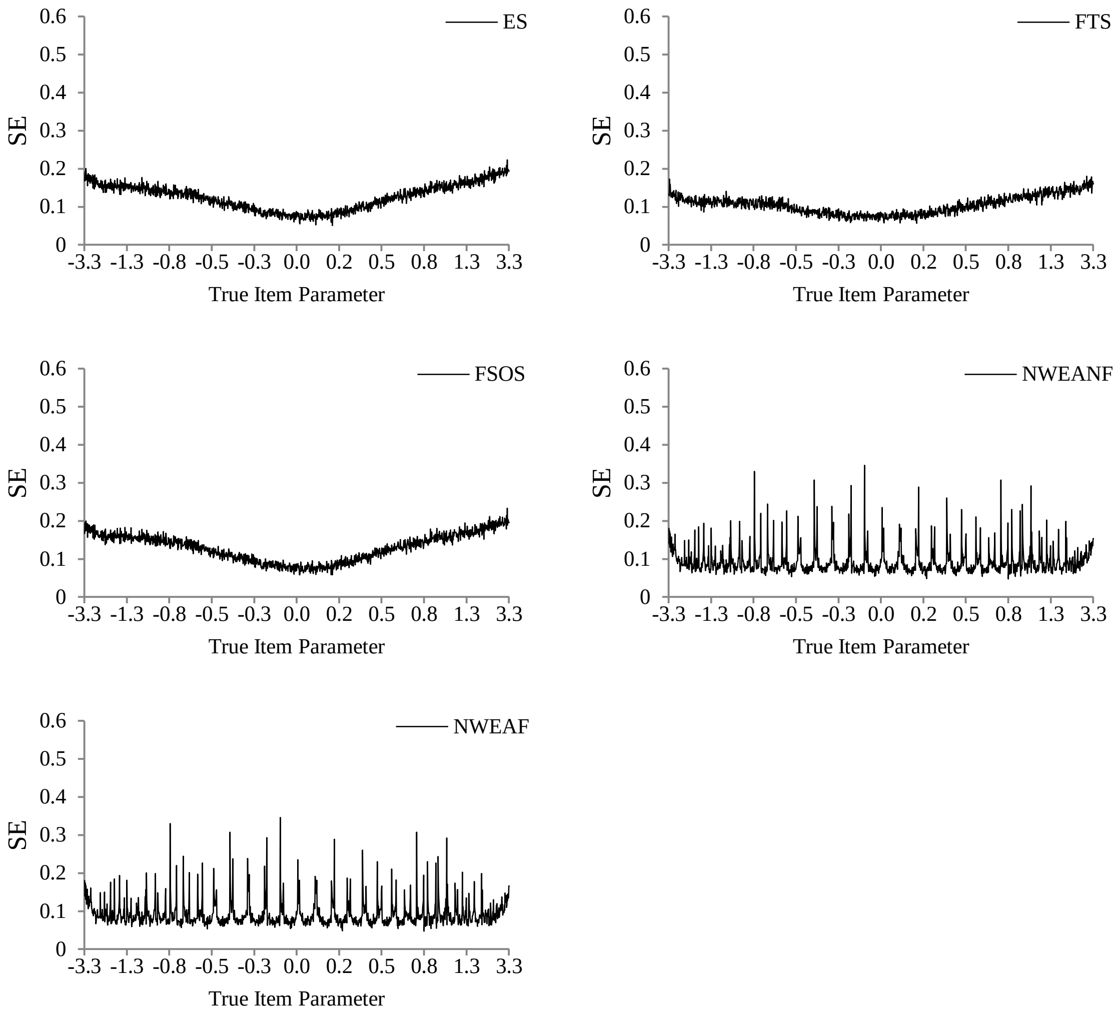


Figure 31. Distribution of Conditional SE Along Theta Scale for Different Test Calibration Conditions for Test Length 60 with One Replication

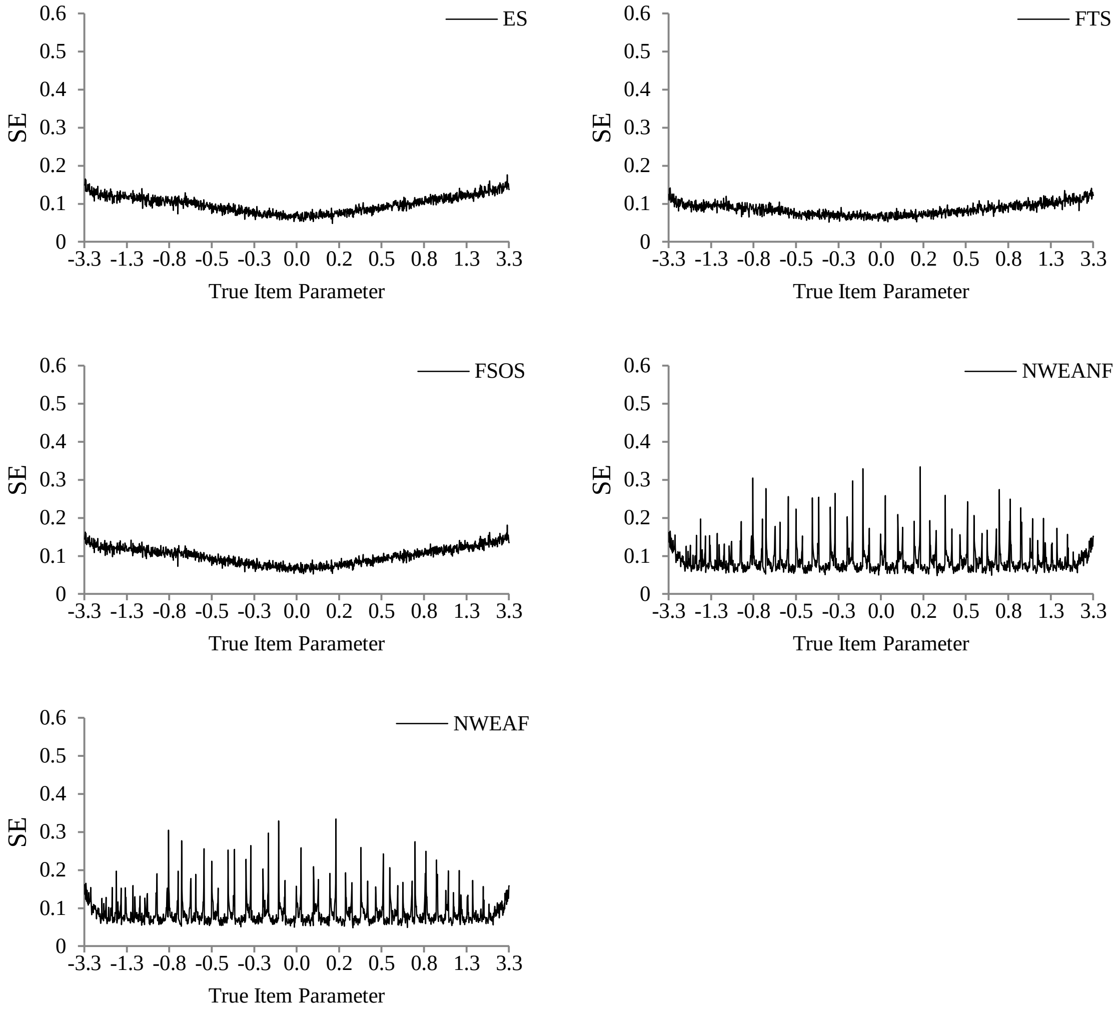


Figure 32. Distribution of Conditional SE Along Theta Scale for Different Test Calibration Conditions for Test Length 70 with One Replication

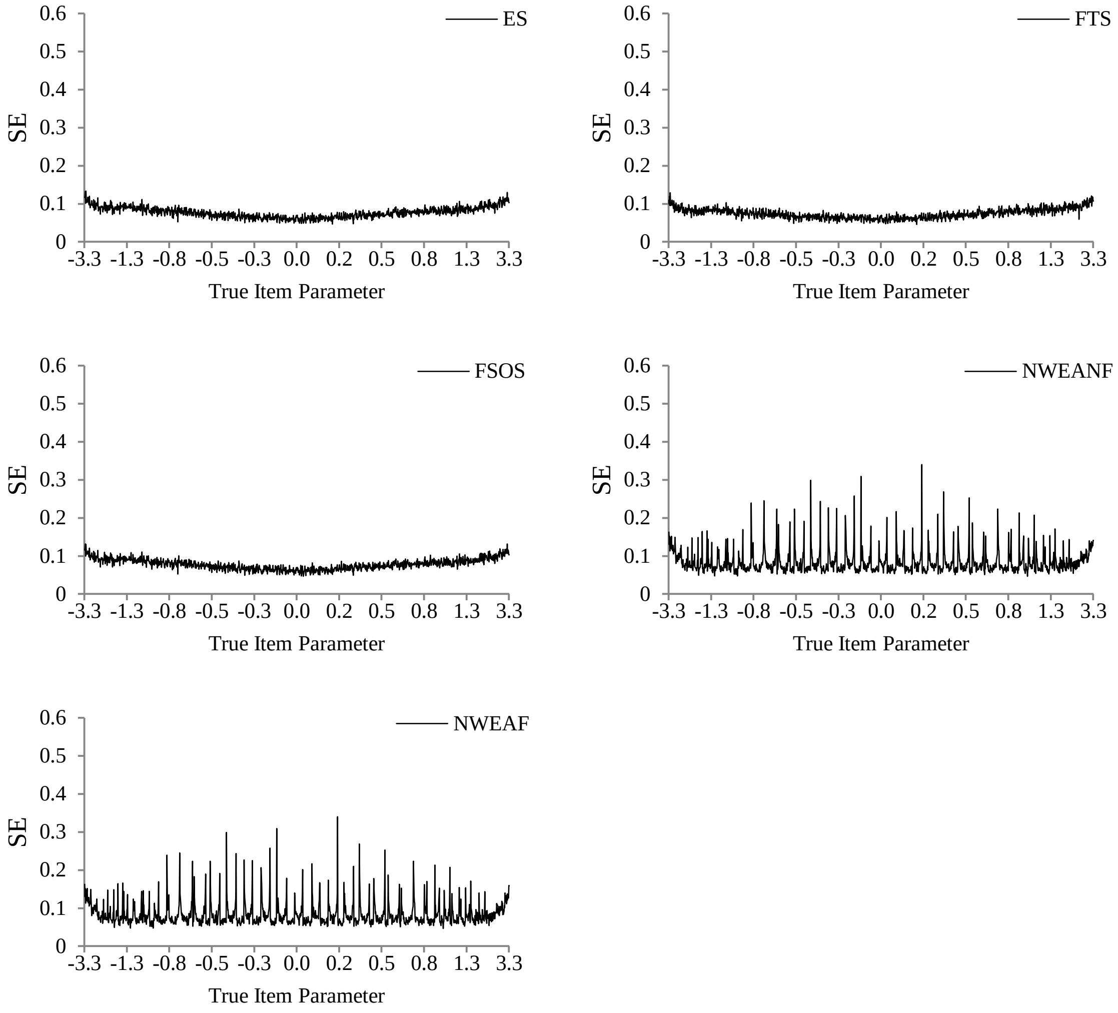


Figure 33. Distribution of Conditional SE Along Theta Scale for Different Test Calibration Conditions for Test Length 80 with One Replication

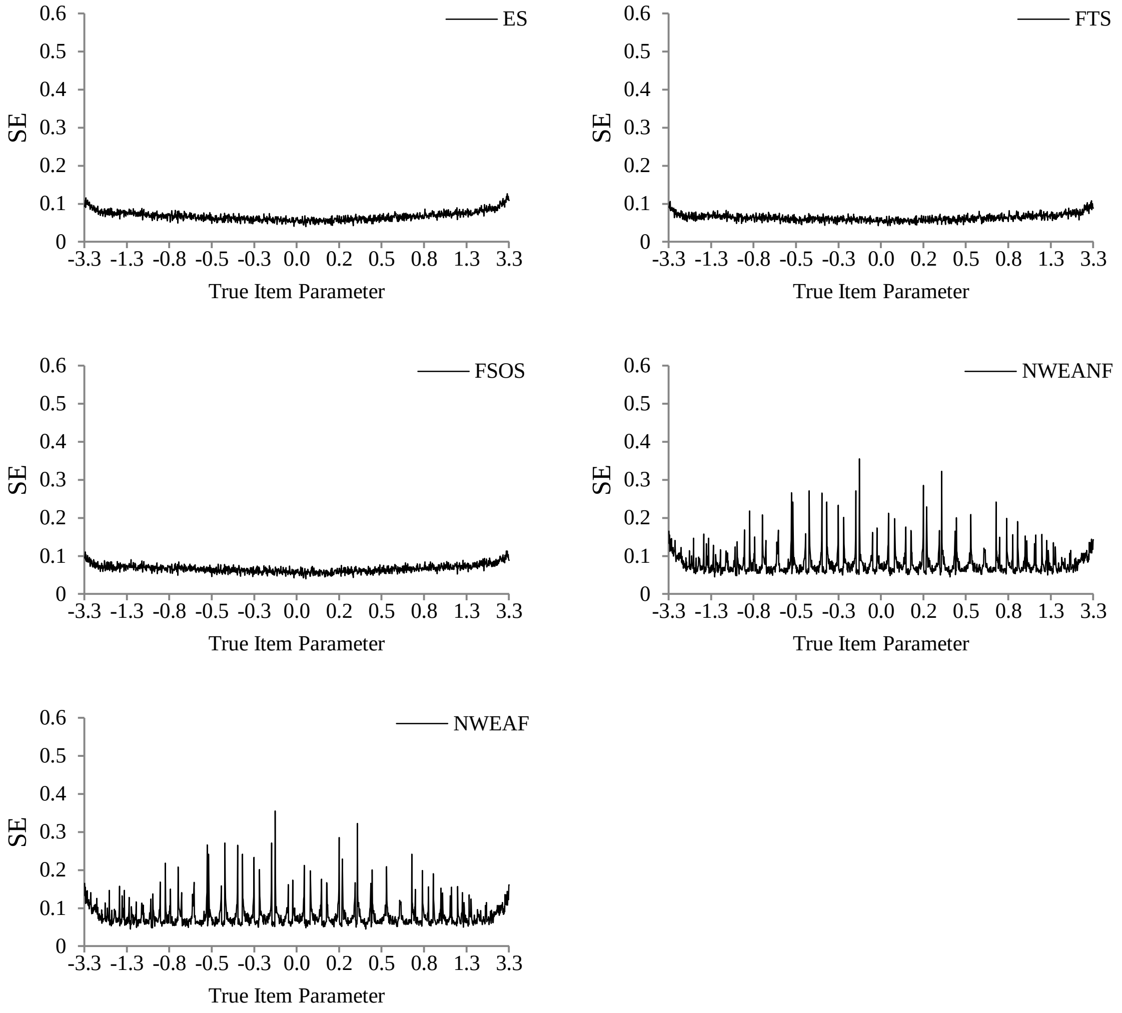


Figure 34. Distribution of Conditional SE Along Theta Scale for Different Test Calibration Conditions for Test Length 90 with One Replication

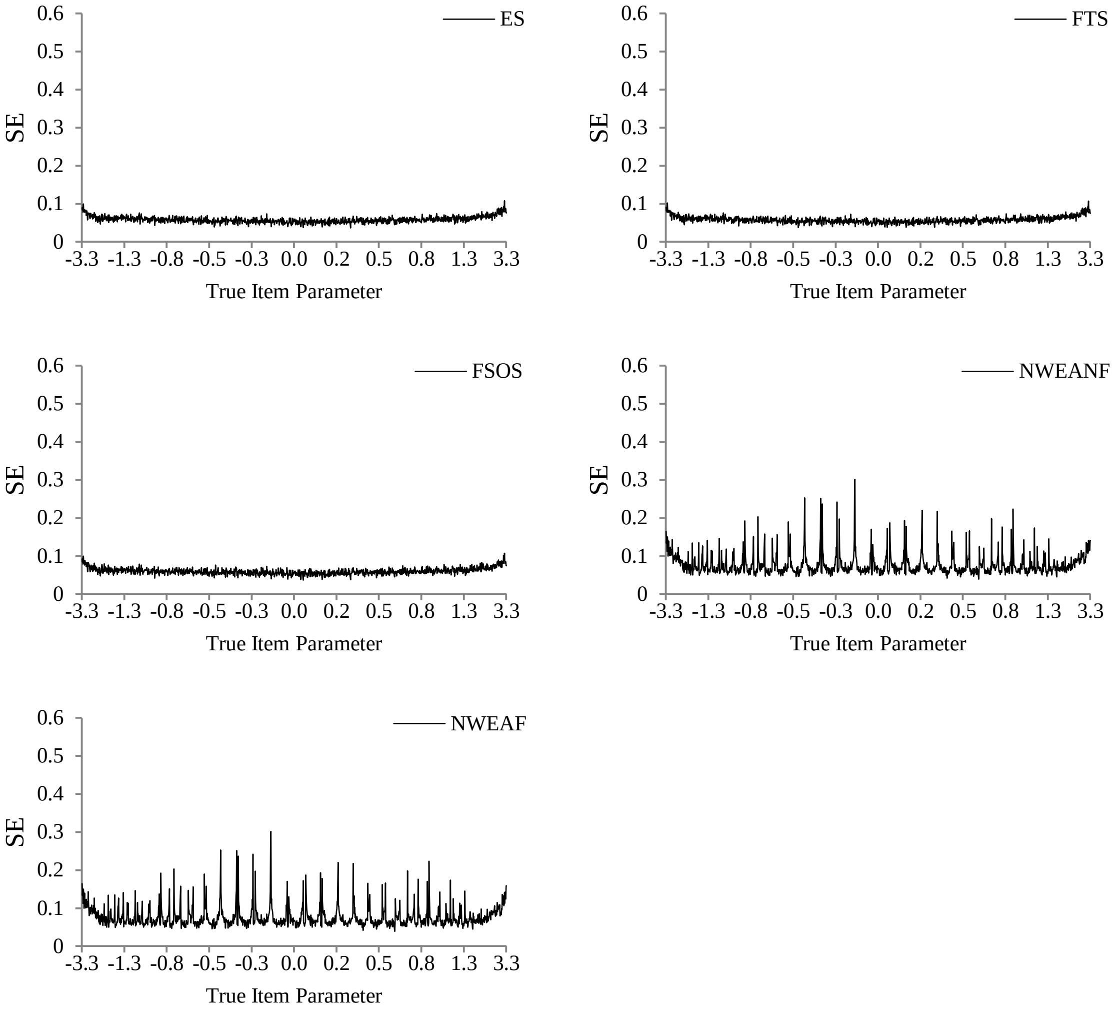


Figure 35. Distribution of Conditional SE Along Theta Scale for Different Test Calibration Conditions for Test Length 100 with One Replication

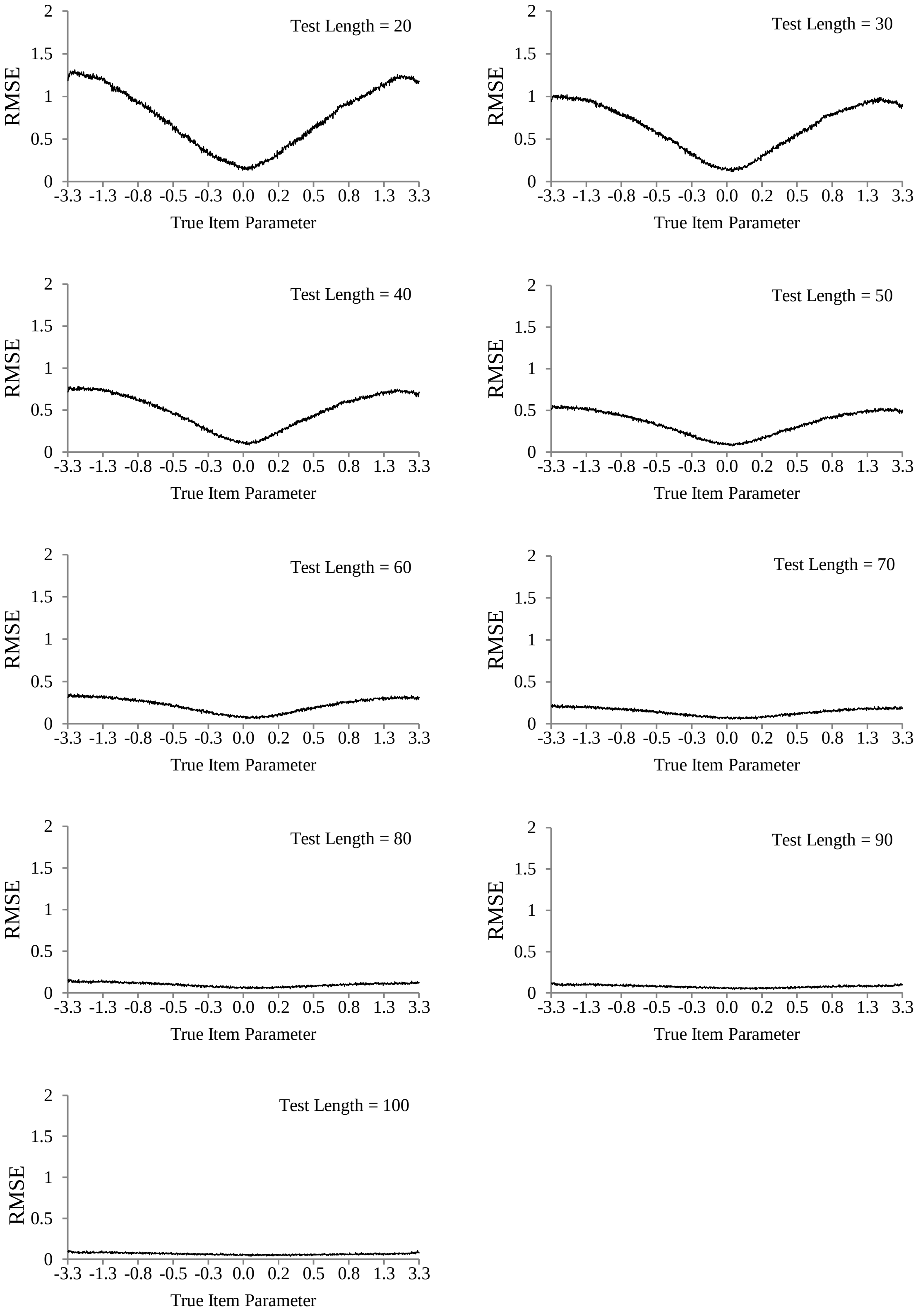


Figure 36. Distribution of Conditional RMSE Along Theta Scale for Different Test Length Under Free Estimation of Sparse Data (ES) Condition with One Replication

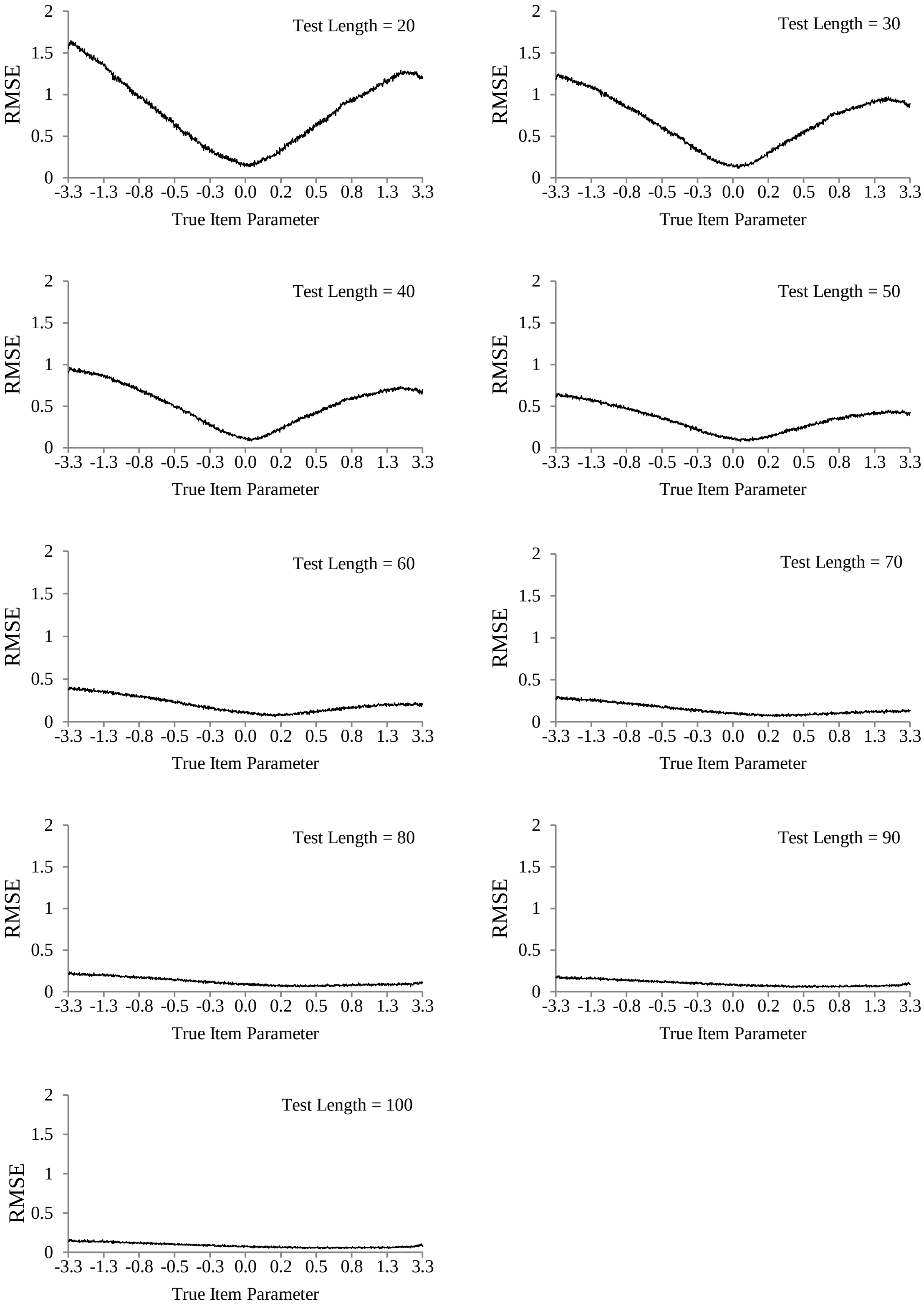


Figure 37. Distribution of Conditional RMSE Along Theta Scale for Different Test Length Under Fixed True Person Parameter Estimation of Sparse Data (FTS) Condition with One Replication

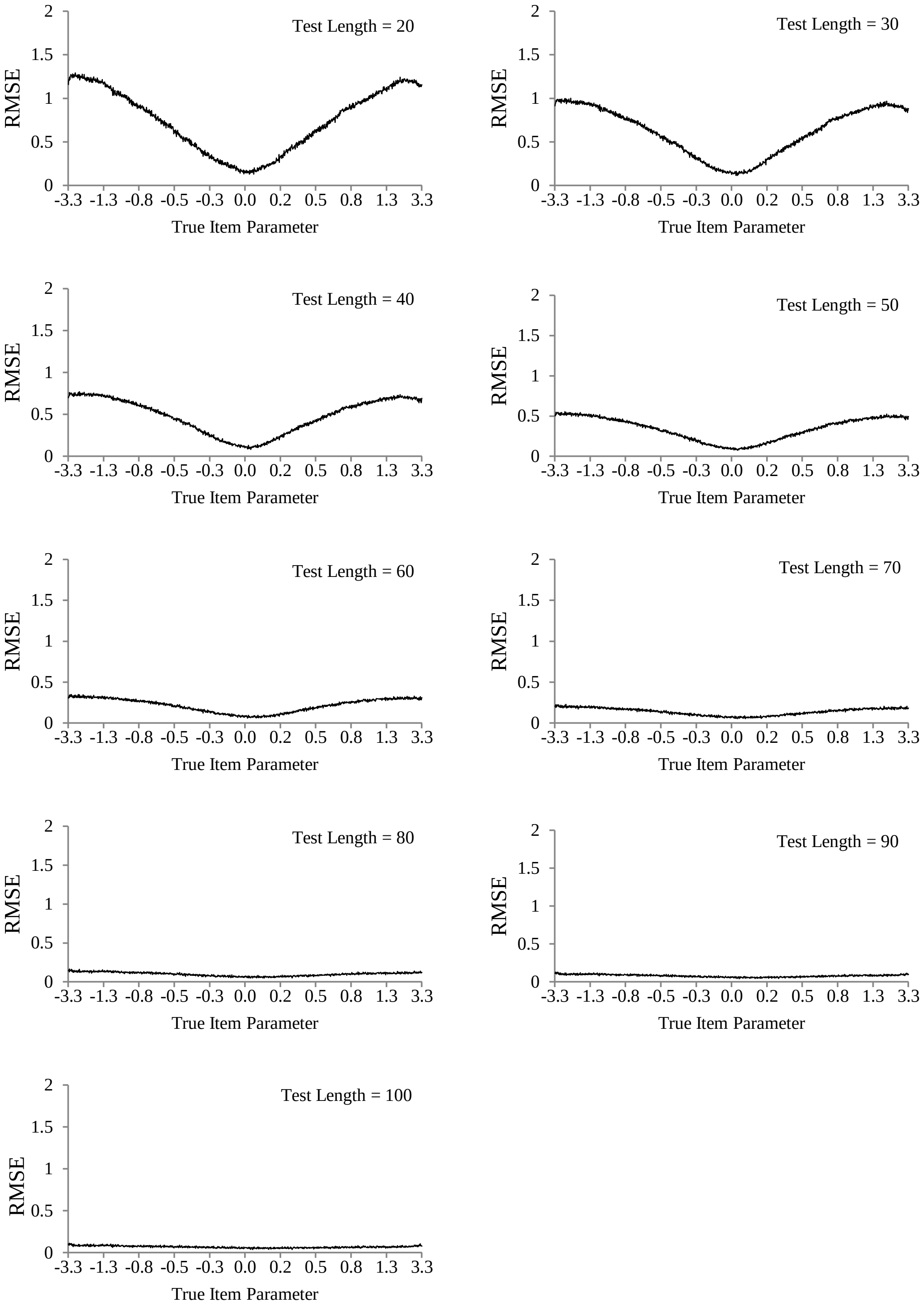


Figure 38. Distribution of Conditional RMSE Along Theta Scale for Different Test Length Under Fixed Sparse Observed Person Parameter Estimation of Sparse Data (FSOS) Condition with One Replication

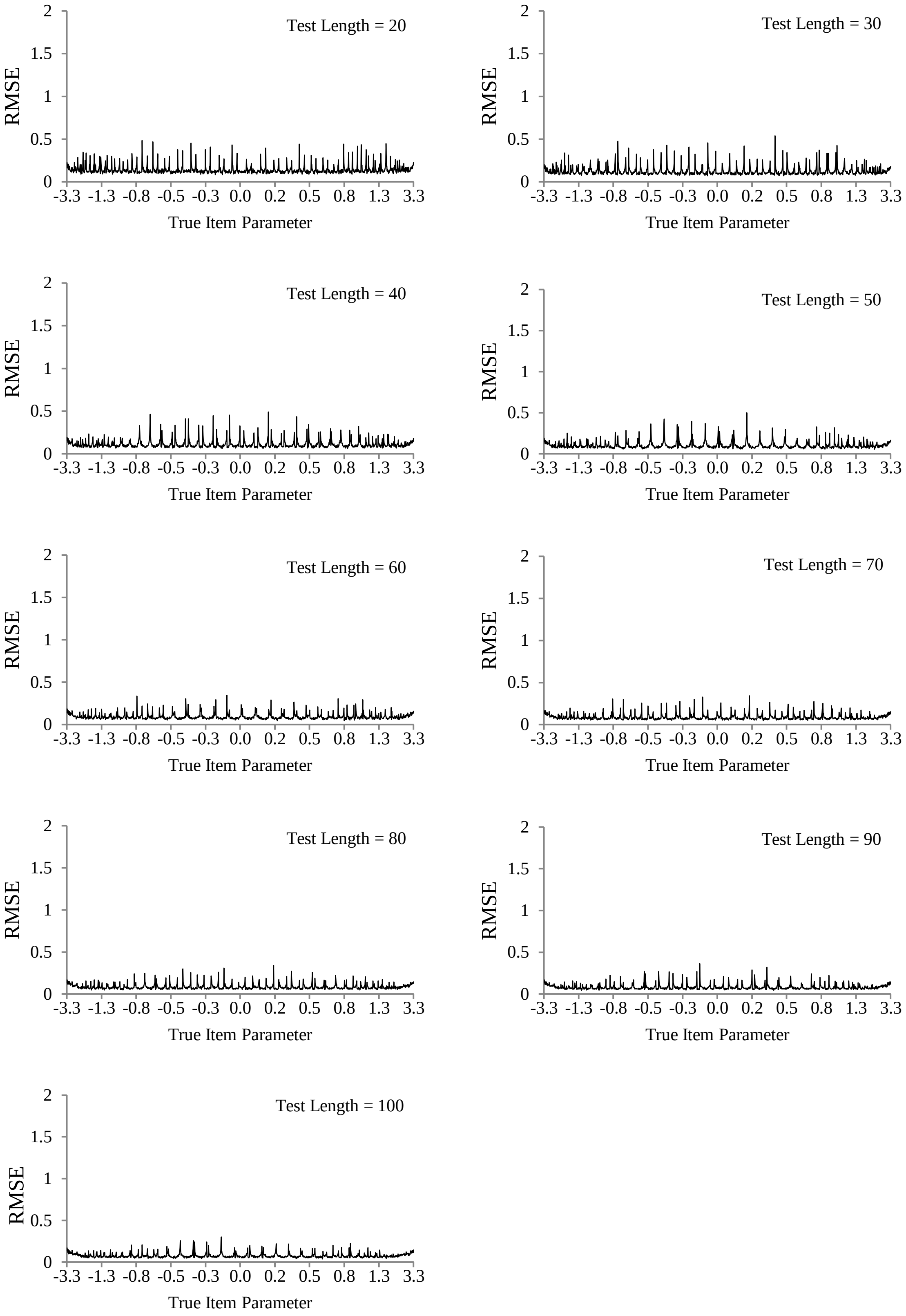


Figure 39. Distribution of Conditional RMSE Along Theta Scale for Different Test Length Under NWEA Non-Filtered (NWEANF) Calibration Condition with One Replication

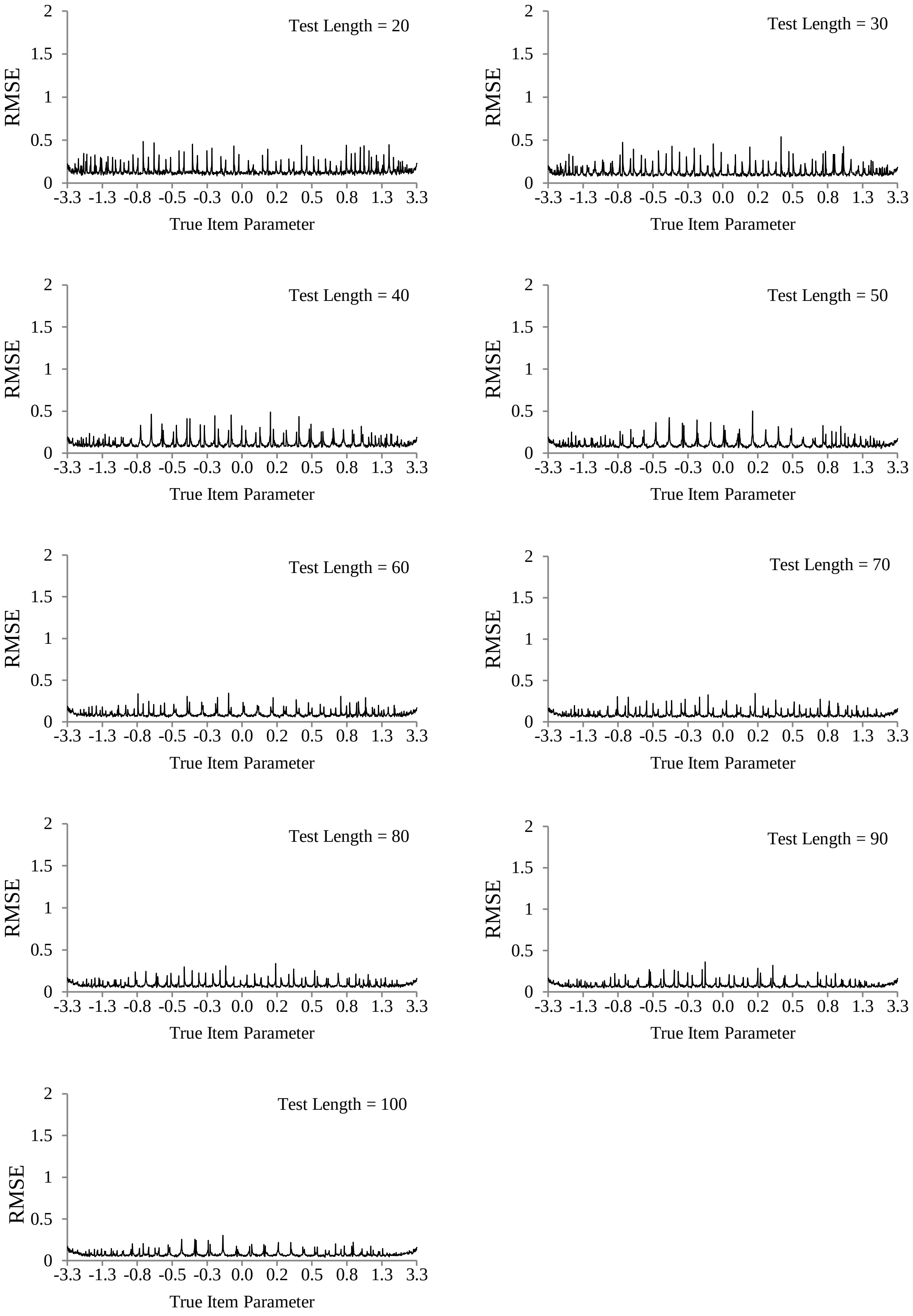


Figure 40. Distribution of Conditional RMSE Along Theta Scale for Different Test Length Under NWEA Filtered (NWEAF) Calibration Condition with One Replication

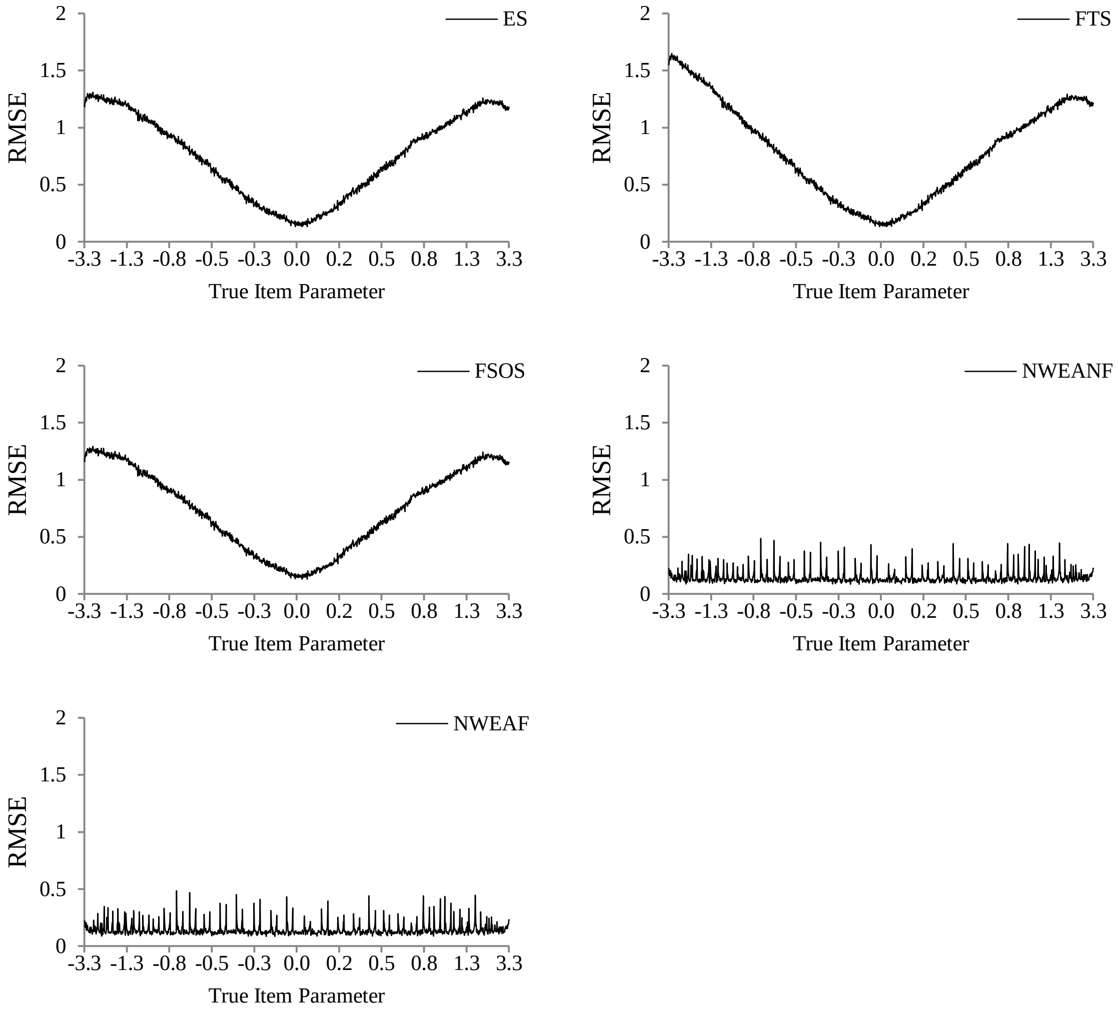


Figure 41. Distribution of Conditional RMSE Along Theta Scale for Different Test Calibration Conditions for Test Length 20 with One Replication

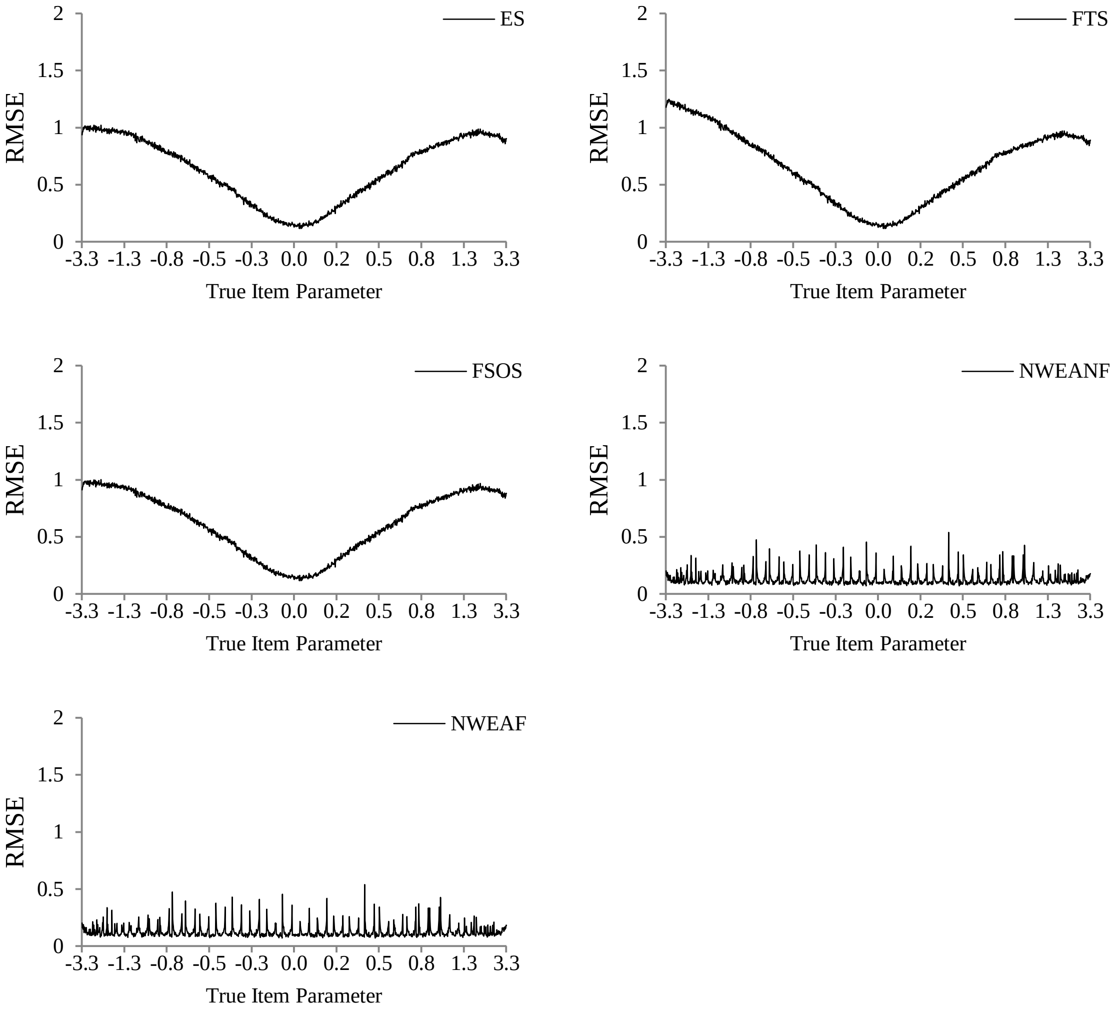


Figure 42. Distribution of Conditional RMSE Along Theta Scale for Different Test Calibration Conditions for Test Length 30 with One Replication

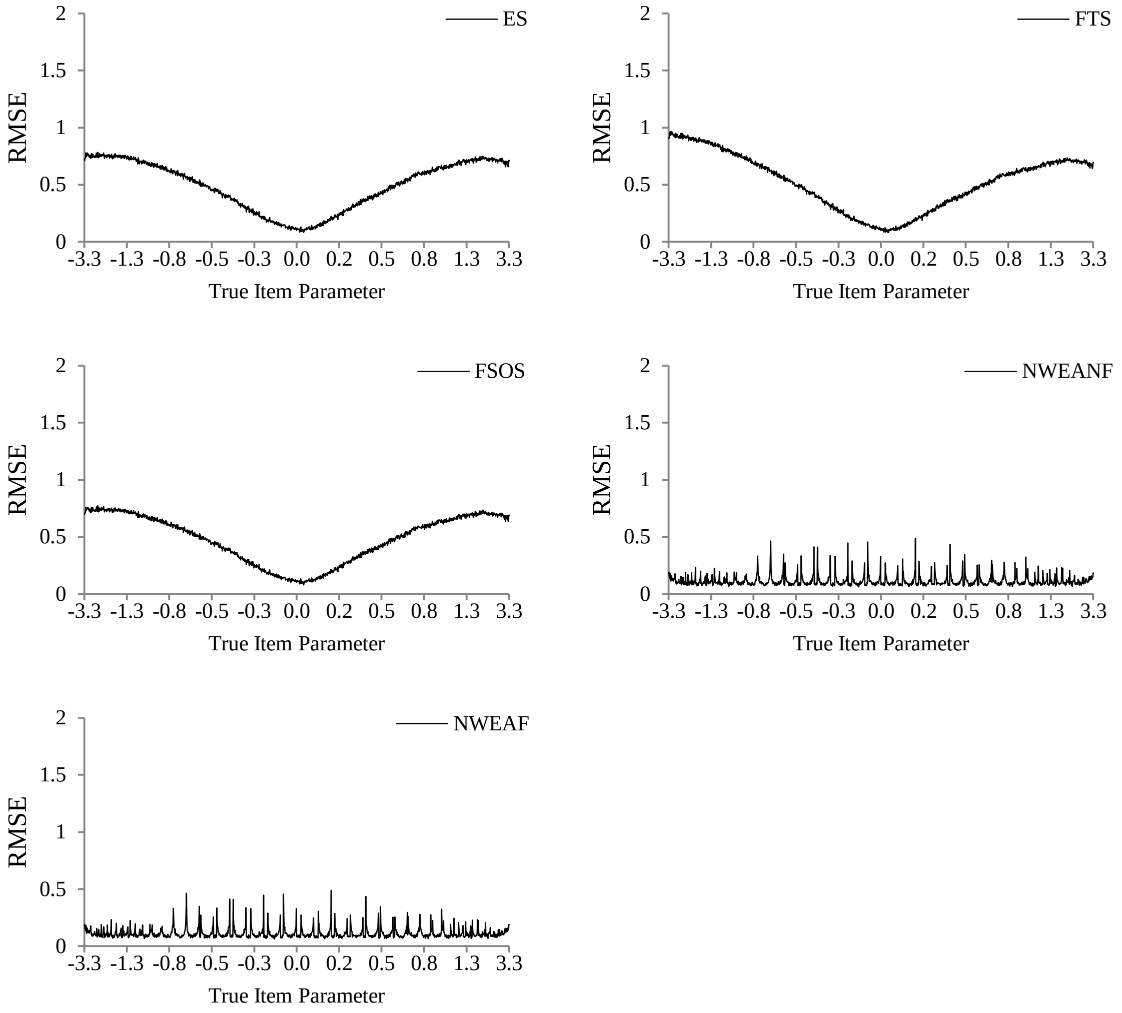


Figure 43. Distribution of Conditional RMSE Along Theta Scale for Different Test Calibration Conditions for Test Length 40 with One Replication

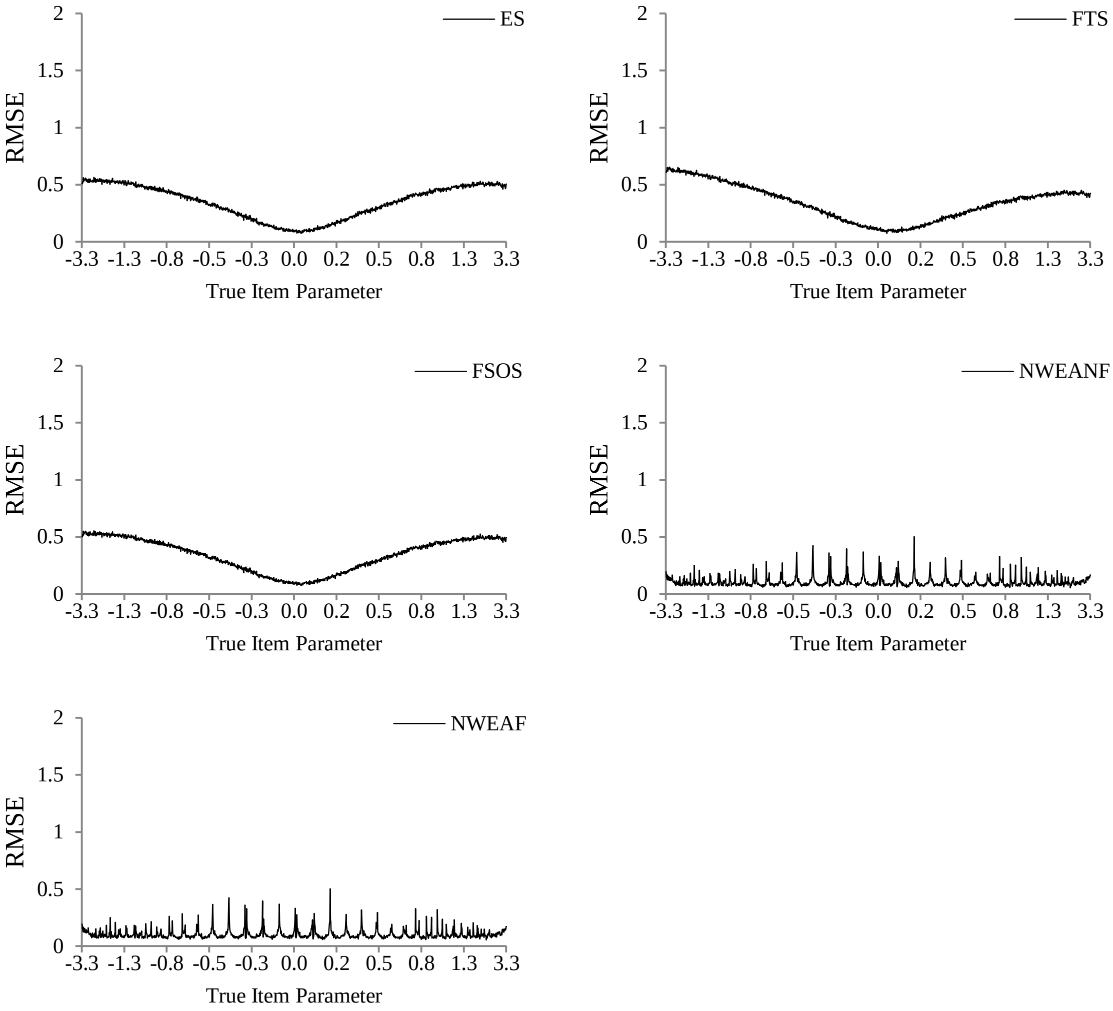


Figure 44. Distribution of Conditional RMSE Along Theta Scale for Different Test Calibration Conditions for Test Length 50 with One Replication

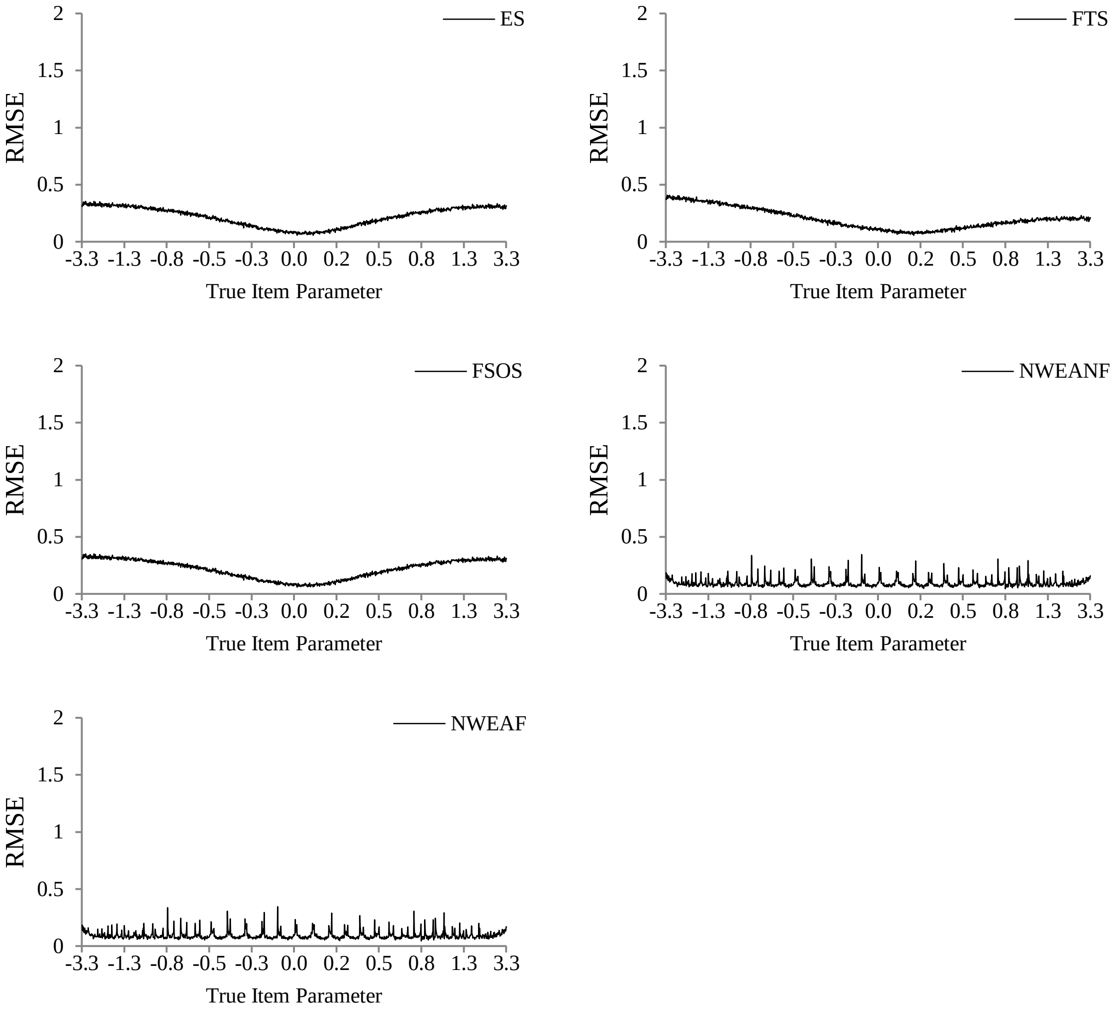


Figure 45. Distribution of Conditional RMSE Along Theta Scale for Different Test Calibration Conditions for Test Length 60 with One Replication

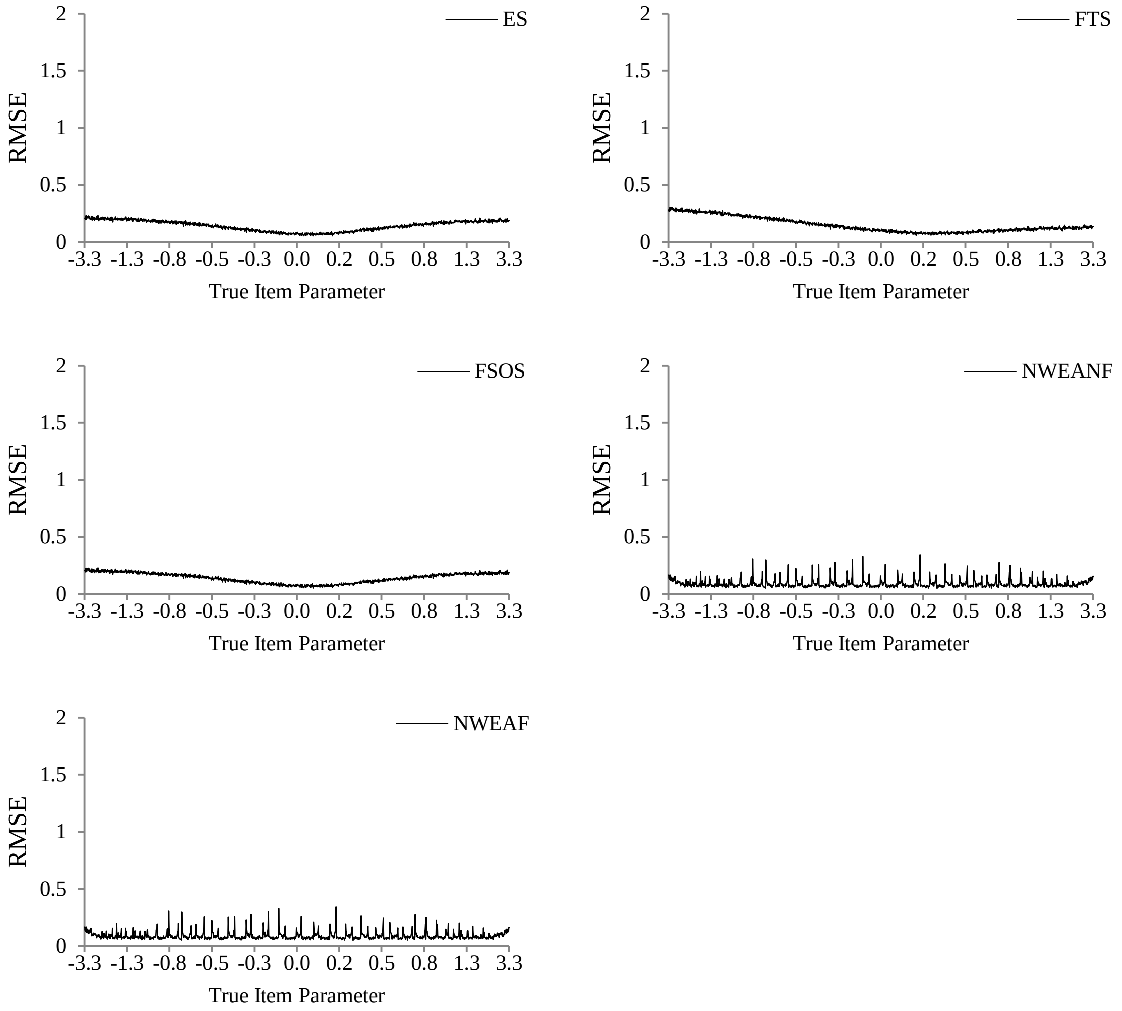


Figure 46. Distribution of Conditional RMSE Along Theta Scale for Different Test Calibration Conditions for Test Length 70 with One Replication

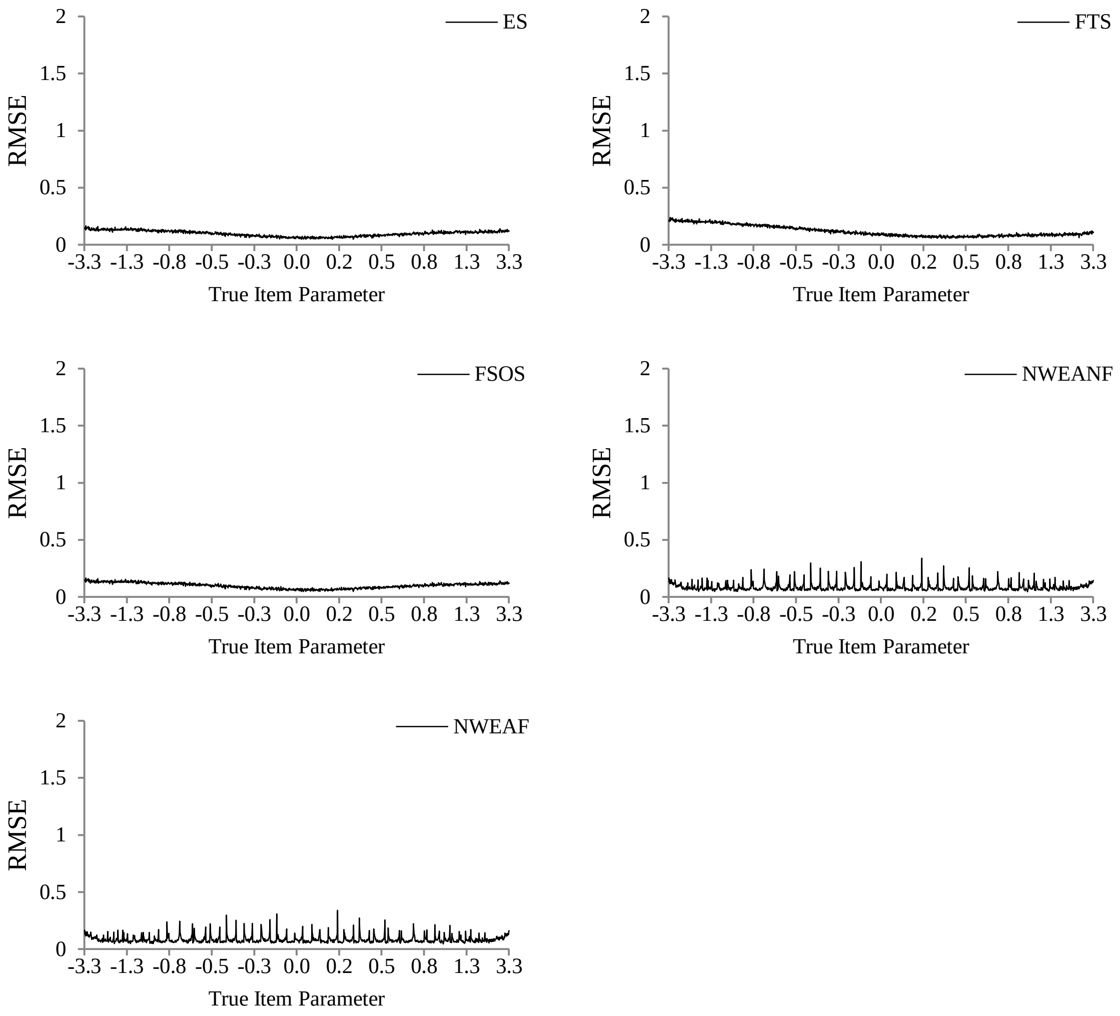


Figure 47. Distribution of Conditional RMSE Along Theta Scale for Different Test Calibration Conditions for Test Length 80 with One Replication

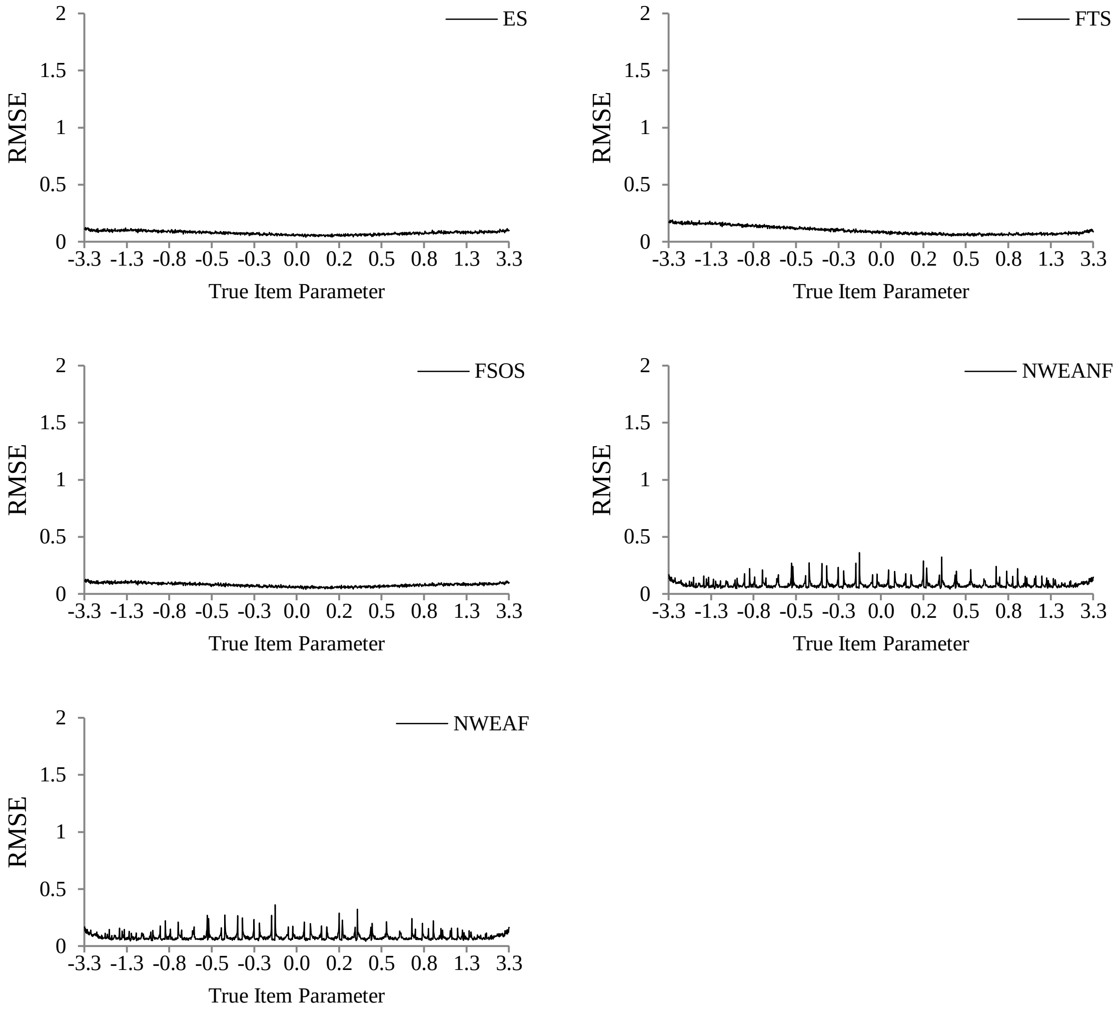


Figure 48. Distribution of Conditional RMSE Along Theta Scale for Different Test Calibration Conditions for Test Length 90 with One Replication

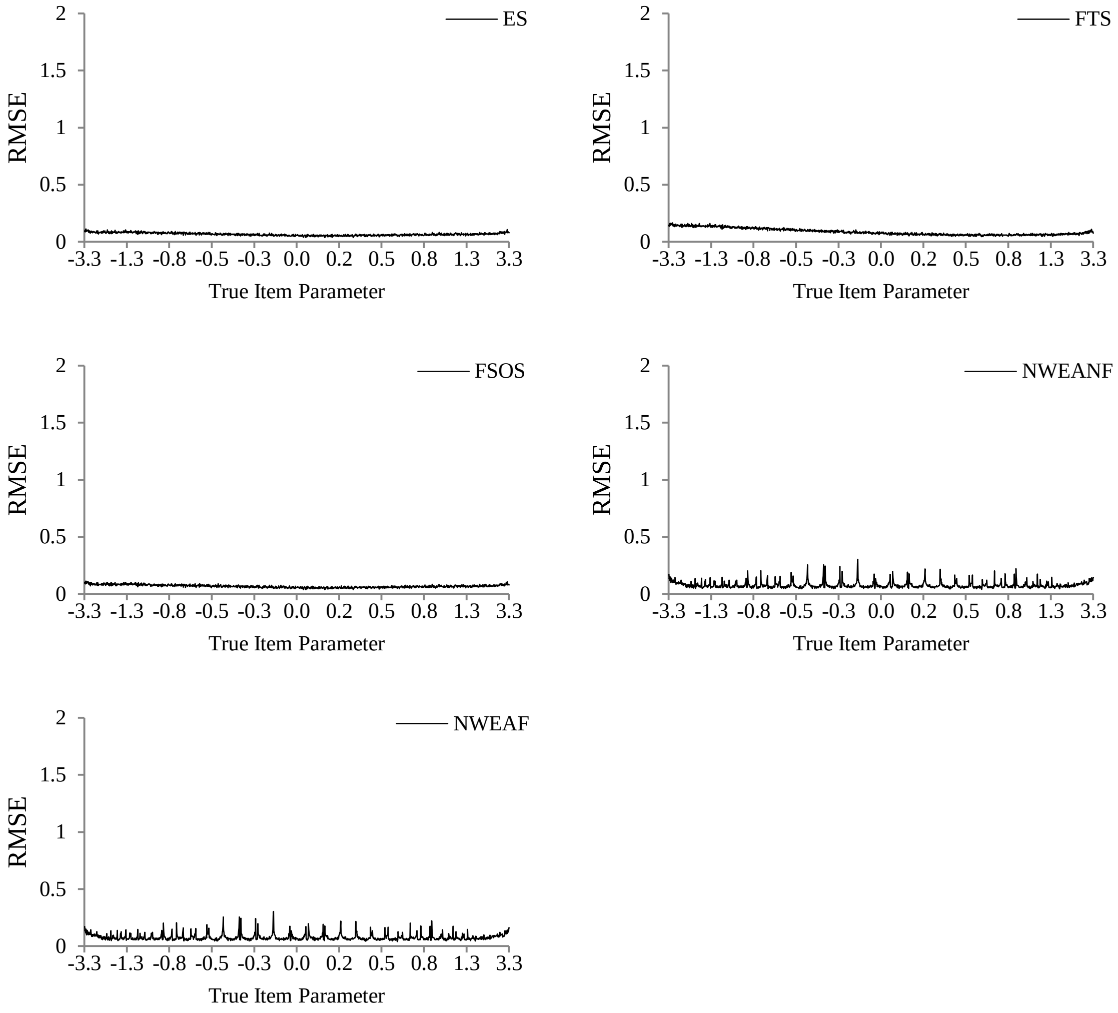


Figure 49. Distribution of Conditional RMSE Along Theta Scale for Different Test Calibration Conditions for Test Length 100 with One Replication

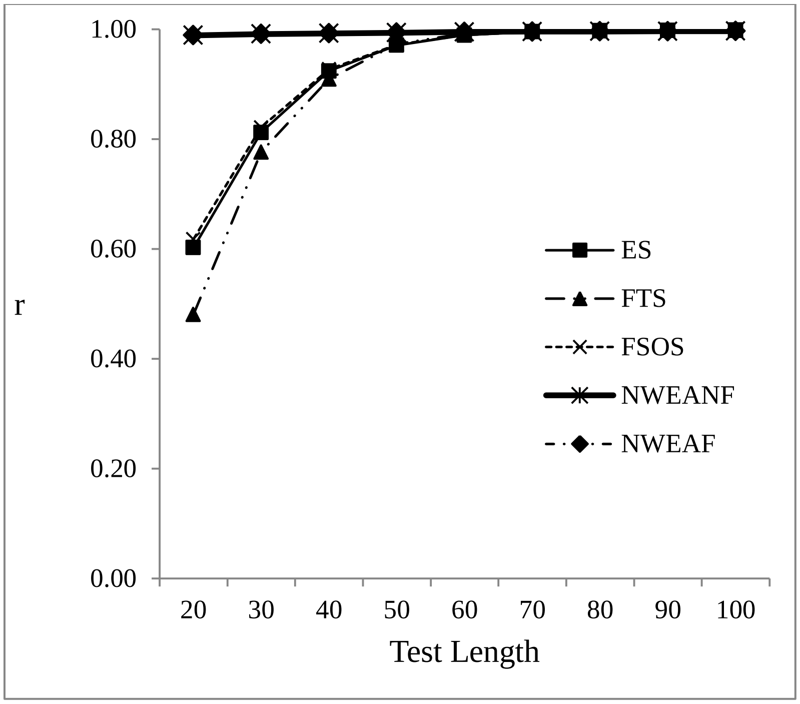


Figure 50. Average Correlation of Different Calibration Methods across Test Lengths over 50 Replications

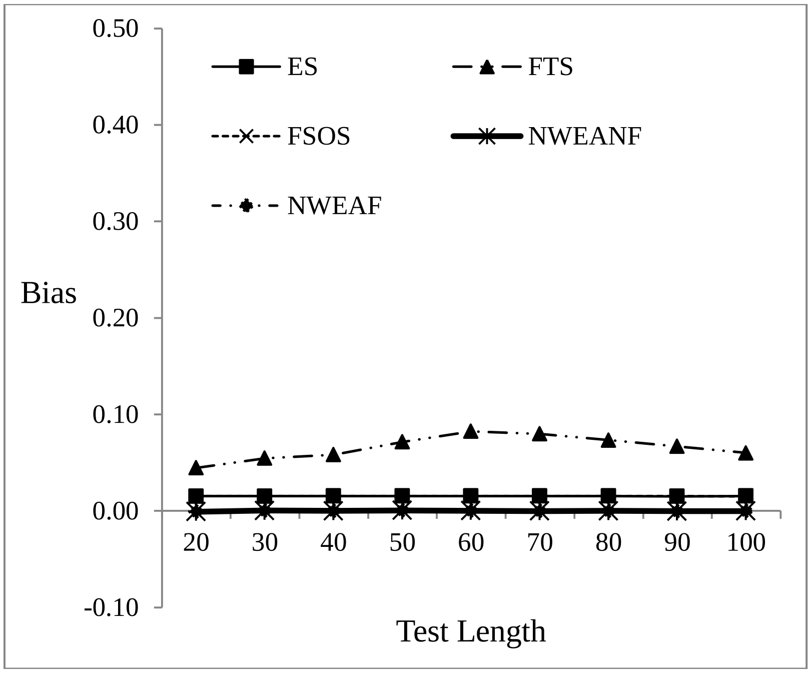


Figure 51. Average Bias of Different Calibration Methods across Test Lengths over 50 Replications

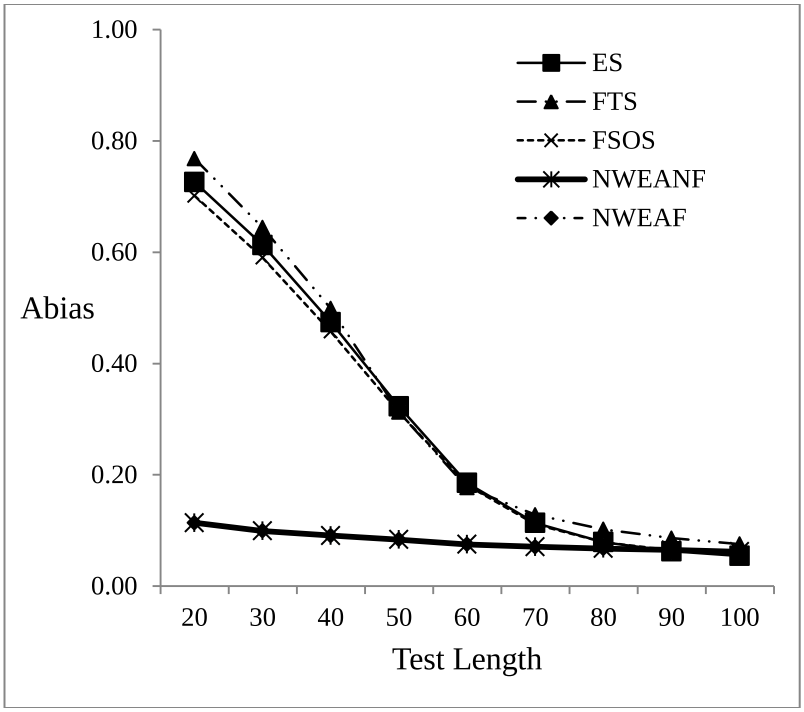


Figure 52. Average Abiases of Different Calibration Methods across Test Lengths over 50 Replications

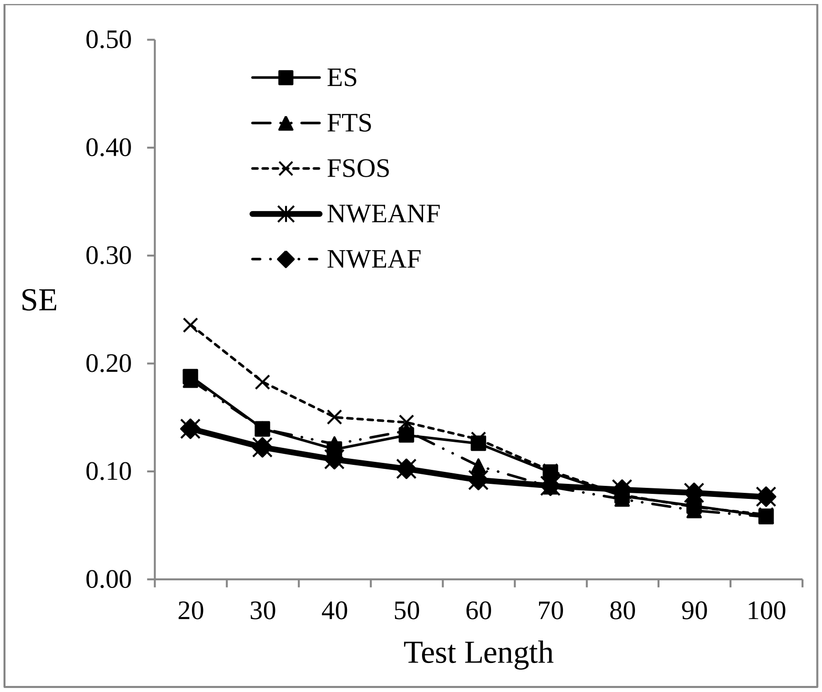


Figure 53. Average SE of Different Calibration Methods across Test Lengths over 50 Replications

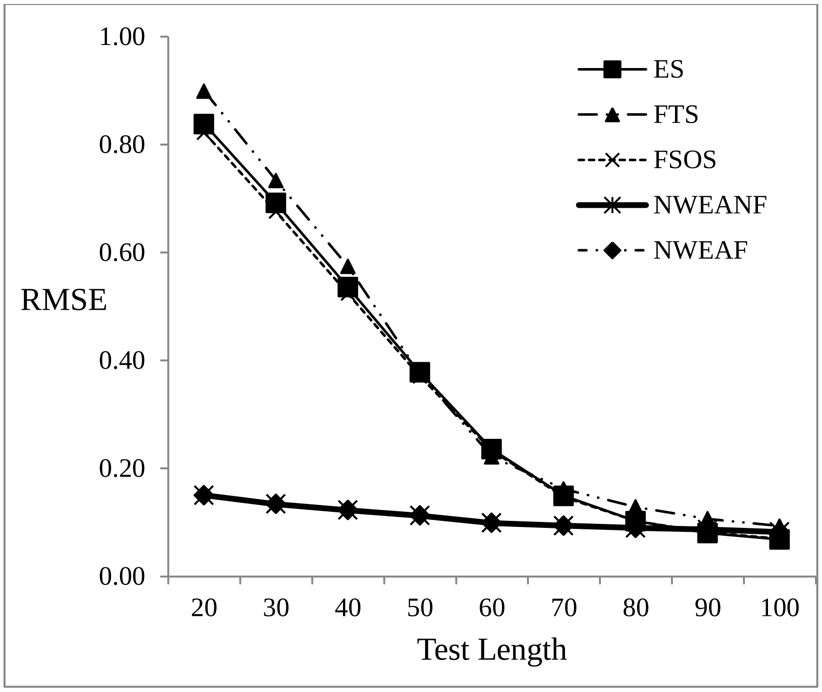


Figure 54. Average RMSE of Different Calibration Methods across Test Lengths over 50 Replications

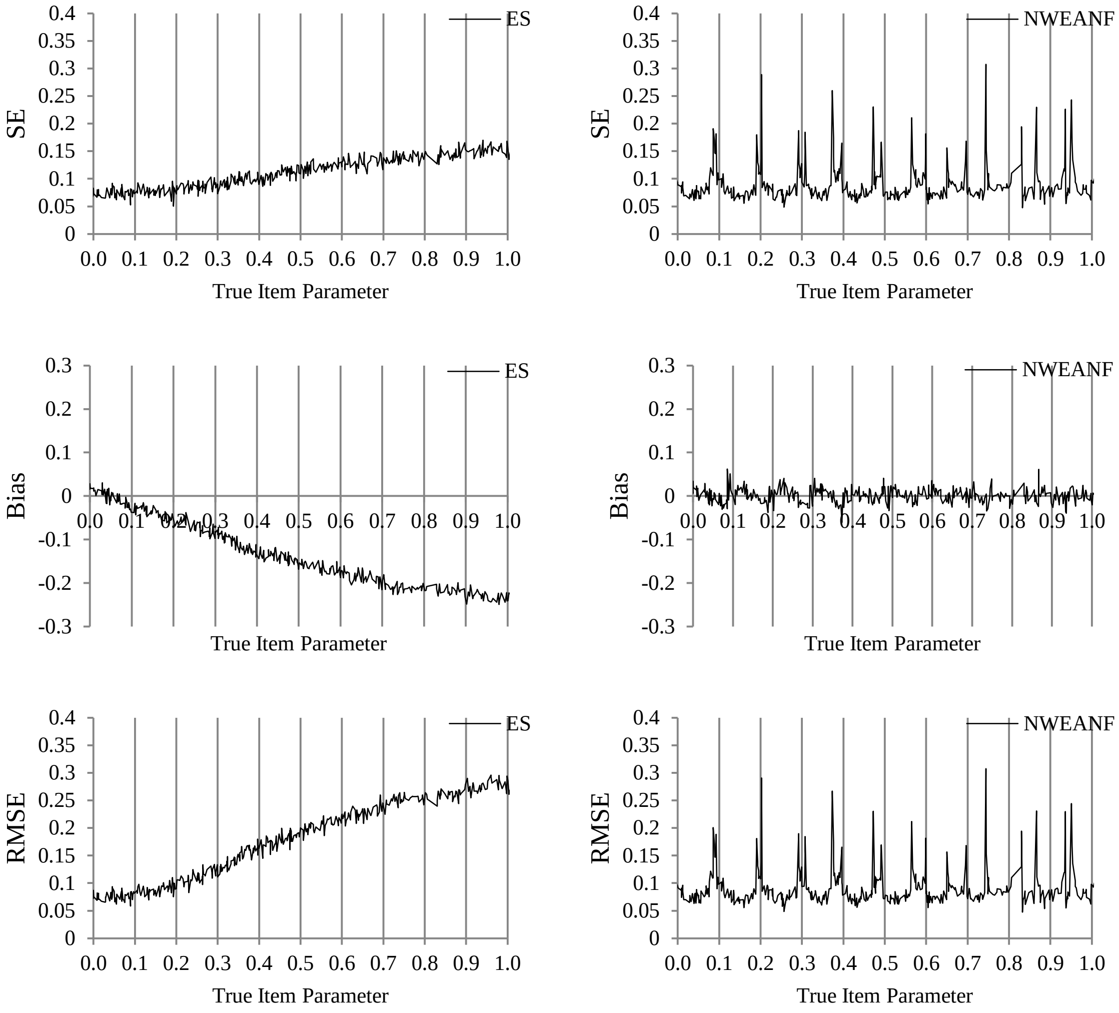


Figure 55. Conditional SE, Bias, and RMSE of Different Calibration Methods of Test Length 50 with One Replication