# Validating American Institutes for Research's Calibration and Scoring Processes for Science Assessments


**June 2019**

# Contents

## Background

Introduction of new three-dimensional science standards have ushered in innovative methods for assessing students' knowledge of science. The new standards also posed challenges in regards to implementation. New psychometric models were needed to account for the fact that item-level scores are clustered due to a common stimulus. Rijmen, Jiang, and Turhan (2018) proposed use of a generalized Rasch testlet model for item clusters that assess the new science standards. This testlet model accounts for local item dependencies by including nuisance dimensions and also allows for items that do not belong to a cluster. In recent technical reports, AIR documented unusually long item calibration times – weeks in the case of grade 8 – due to the estimation of cluster variances. AIR psychometricians addressed this issue by using an internally developed item calibration program (in MATLAB) to reduce run times. In addition, they proposed a marginal maximum likelihood estimator for proficiency, marginalizing out nuisance dimensions. This estimator is not available in any commercial IRT software. AIR conducted a simulation study to demonstrate the performance of the marginal maximum likelihood estimator (M-MLE) under various conditions. Their findings were that the M-MLE proficiency estimator was less biased than other estimators and would be used to estimate science student proficiency scores.

As most of the software AIR used for science item calibration and scoring was developed in-house, state level data quality control checks will be met with logistical challenges. In light of this, the Connecticut State Department of Education (CSDE) conducted a series of studies to validate AIR's calibration and scoring processes.

The purpose of this paper is to describe our validation process, summarize results, and provide recommendations for future science assessment administrations.

## Validating the Item Calibration Process

AIR provided CSDE with item parameters based on two samples: multi-state data and Connecticut only data. Parameters were estimated using AIR's program written in MATLAB. All item calibrations at CSDE were based on Connecticut data using a marginal maximum likelihood procedure in the flexMIRT software. CSDE compared our item calibration results to the two datasets provided by AIR (multi-state calibration and Connecticut only calibration). Overall, differences in item parameter estimates were negligible.

Table 1 of the Appendix shows counts and percentages of differences in item difficulties for specified intervals.

## Validating the M-MLE Estimator

AIR recently published a web-based scoring tool, METRICS, for public use. METRICS uses R routines to compute student proficiency scores. We tested the software by scoring the 2018 field test data and compared the results to those provided by AIR. CSDE was able to match AIR's scores and standard errors exactly, giving us confidence that participating states can use METRICS during data quality control efforts.

Members of CSDE's Research team independently developed a Fortran 90 computer program (Rogers & Swaminathan, 2019) for computing the M-MLE proficiency estimator proposed by AIR. The program was validated by generating 1P testlet data for a sample of 4100 students, with 100 students at each of 41 theta levels between -4 and 4 in increments of 0.2. The simulated response data were scored using both the Fortran M-MLE estimator and AIR's METRICS. Figure 1 shows the relationship between proficiency scores obtained by Rogers and Swaminathan's Fortran M-MLE and AIR's M-MLE. The estimates from the Fortran program agreed exactly with those of AIR except at the extremes. This is due to differences between methods used for handling zero and perfect scores. In sum, Rogers and Swaminathan's Fortran M-MLE program produces the same proficiency scores as AIR's METRICS tool and is appropriate for use in a simulation study comparing proficiency score estimators.

### Simulation Studies

CSDE conducted two simulation studies using the same design as that of Rijmen, Jiang, and Turhan (2018). The design was extended by using a wider range of proficiency values and adding other estimators for comparison. Study 1 assessed the proficiency recovery for six different proficiency estimators under 12 conditions. Study 2 used a single condition based on a more realistic test design.

*Simulation Study I*

The test design consisted of 40 binary assertions – 28 assertions clustered in sets of 4 and 12 standalones. Rijmen, Jiang, and Turhan (2018) explained that four assertions per cluster were

chosen to evaluate model parameter and proficiency recovery as they expected that this number would pose a challenge for recovery of cluster variances. As in their study, we varied the size of the testlet effect for nuisance dimensions: small (0.25), medium (0.50), large (0.75), and very large (1.00). The variance for the general science dimension was set to one (1) for all conditions. Mean test difficulty was varied as well. There were three conditions: low (-0.5), medium (0.0), and high (0.5) difficulty. This resulted in 12 study conditions.

All data were generated under a 1P testlet model. A sample of 10,000 responses was generated under each study condition and calibrated using 3 different IRT models: unidimensional one- (Rasch) and two-parameter logistic IRT models and the Rasch testlet model. All models were calibrated with flexMIRT using a marginal maximum likelihood procedure. After item calibration, 5000 responses were generated for each of 41 true proficiency values ranging from -4 to 4 in increments of 0.2. Item parameters obtained during calibration were used to estimate proficiency. For all conditions, response data were scored using six different proficiency estimators in order to assess the bias and accuracy of recovered proficiency estimates:

- Unidimensional IRT Model Proficiency Estimators (FlexMIRT)_
    - 1P Expected A Posteriori (EAP)
    - 2P EAP
- Multidimensional IRT Model Proficiency Estimators
    - Joint Maximum A Posteriori (J-MAP) (FlexMIRT)
        - General and nuisance dimension proficiencies estimated jointly
        - Normal (0,1) prior placed on the general science dimension
    - Marginal Expected A Posteriori (M-EAP)
        - Marginalizes nuisances dimensions before for computing an EAP estimate for general science dimension
        - Normal (0,3) prior placed on the general science dimension
    - Marginal Maximum A Posteriori (M-MAP)
        - Marginalizes nuisances dimensions before computing Bayesian modal proficiency estimate for general science dimension
        - Normal (0,3) prior placed on the general science dimension
    - Marginal Maximum Likelihood Estimation (M-MLE)

- Marginalizes nuisances dimension before computing maximum likelihood proficiency estimate for general science dimension
- No priors placed on the general science dimension

For each estimator, the bias in the estimates was calculated under each condition for 5000 replications at each of the 41 trait values. Bias is defined as the average difference between estimated values and true values. The M-MLE estimator demonstrated the least amount of bias in estimating proficiency followed by M-MAP and M-EAP under all 12 study conditions. The marginal MAP and EAP estimators overestimated true proficiency at the lower end and underestimated true proficiency at the higher end of the proficiency distribution. However, the unidimensional 1P EAP and 2P EAP demonstrated the greatest amount of bias under all conditions. As in the case of M-MAP and M-EAP, true proficiency was consistently overestimated at the lower end and underestimated at the higher end of the proficiency distribution, but to a much greater extent than the M-MAP and M-EAP estimators. Refer to Figures 2 – 5 for bias plots. Plots are for medium (0.0) test difficulty at each level of cluster variance. Plots for easy and hard test conditions are available upon request.

In addition to bias, the root mean squared error (RMSE) was calculated under each condition to assess the accuracy of proficiency estimates. RMSE is the square root of the average squared difference between the estimate and the true value and may be interpreted as the average amount of error in the estimates. A low RMSE value is preferred. Except at the extremes of the proficiency distribution, the M-MLE estimator was consistently the least accurate estimator of proficiency under all study conditions. The unidimensional 1P EAP and J-MAP estimators were most accurate in the average proficiency range when the testlet effect is lowest (0.25) (i.e., the data are close to unidimensional). However, as the testlet effect increases, the accuracy of these estimators begins to decrease. Refer to Figures 6 - 9 for accuracy plots. Plots are for medium (0.0) test difficulty at each level of cluster variance. Plots for easy and hard test conditions are available upon request.

*Simulation Study II*

As in Rijmen, Jiang, and Turhan's (2018) study, the test design for the second simulation study consisted of 53 binary assertions. The number of assertions per clusters were varied from 4 to 11

for a total 41 assertions across 6 clusters. The number of standalones was twelve. Cluster variance for the second study was varied from 0.31 to 2.06. These conditions were based on observations from the 2018 field test placing the cluster variances in for the second study closer to actual test design. Results for Study II were very similar to those of Study 1. Once again, the M-MLE estimator was the least biased across the proficiency continuum but was less accurate than the estimators that employed a prior distribution. Refer to Figures 10 and 11 for bias and accuracy plots.

## Recommendations

This study examined item calibration and proficiency recovery across select proficiency estimators for data of the type collected during the science assessment. Results of this study provide support for AIR's findings that use of a testlet model and the M-MLE estimator is optimal for minimizing bias in proficiency estimates. However, the M-MLE estimator does not provide the most accurate proficiency estimates. In choosing a proficiency estimator, there is a tradeoff: reduce bias in estimates or increase accuracy. The answer to this issue depends on how test results will be used. Wang and Vispoel (1998) list three conditions where minimizing bias is most important: (1) comparing group means, (2) comparing proficiency estimates based on different tests, and (3) score classification. The authors explain that bias could cause systematic shifts in group means, individual proficiency estimates, and classification cut points, particularly in the case of extreme proficiency scores. Given that the science assessments are used to obtain scores at the group level and classify scores into proficiency levels, minimally biased proficiency estimates are preferable.

## References

Rijmen, F., Jiang, T., & Turhan, A. (2018). An item response theory model for new science assessments. Paper presented at the National Council on Measurement in Education, New York, NY.

Rogers, H. J. & Swaminathan, H. (2019). Marginal maximum likelihood proficiency estimator.

Wang, T. & Vispoel, W. P. (1998). Properties of ability estimation methods in computerized adaptive testing. *Journal of Educational Measurement*, 35(2), 109-135.

# Appendix A: Tables and Figures

Table 1. Differences in assertion difficulties between CT Calibration and AIR Calibration: Counts and percentages of assertions

| | Grade 5, n=660 | | Grade 8, n=1036 | | Grade 11, n=854 | |
|---|---|---|---|---|---|---|
| **Difference in Assertion Difficulties (on theta scale)** | **Differences between CT Calibration of CT Data and AIR Calibration of All States Data** | **Differences between CT Calibration of CT Data and AIR Calibration of CT Only Data** | **Differences between CT Calibration of CT Data and AIR Calibration of All States Data** | **Differences between CT Calibration of CT Data and AIR Calibration of CT Only Data** | **Differences between CT Calibration of CT Data and AIR Calibration of All States Data** | **Differences between CT Calibration of CT Data and AIR Calibration of CT Only Data** |
| **<-0.3** | 16 (2%) | 7 (1%) | 34 (3%) | 3 (0%) | 32 (4%) | 7 (1%) |
| **from -0.3 to -0.21** | 23 (3%) | 3 (0%) | 57 (6%) | 7 (1%) | 70 (8%) | 8 (1%) |
| **from -0.22 to -0.11** | 81 (12%) | 32 (5%) | 157 (15%) | 49 (5%) | 141 (17%) | 18 (2%) |
| **from -0.1 to 0.1** | 420 (64%) | 544 (82%) | 561 (54%) | 891 (86%) | 512 (60%) | 679 (80%) |
| **from 0.11 to 0.2** | 87 (13%) | 55 (8%) | 150 (14%) | 69 (7%) | 67 (8%) | 114 (13%) |
| **from 0.21 to 0.3** | 21 (3%) | 13 (2%) | 44 (4%) | 16 (2%) | 13 (2%) | 24 (3%) |
| **>0.3** | 11 (2%) | 6 (1%) | 33 (3%) | 1 (0%) | 19 (2%) | 4 (0%) |

Figure 1. Scatter plot of proficiency scores obtained by Rogers and Swaminathan's (Fortran) M-MLE versus AIR's M-MLE.

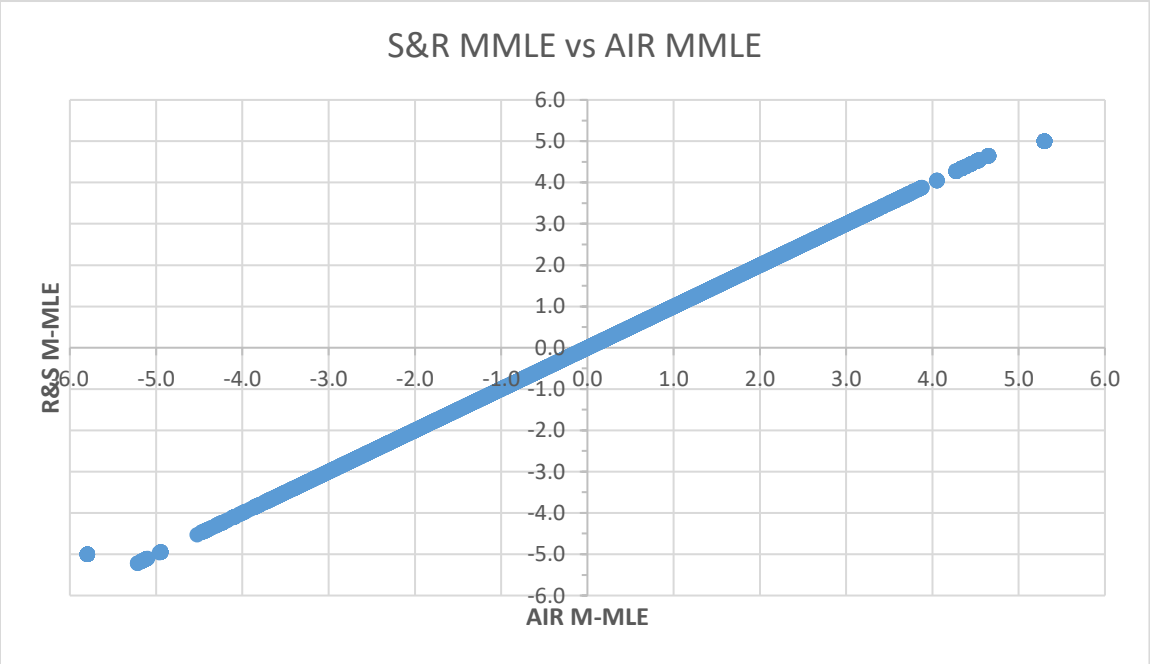

Figure 2. Bias plots for each proficiency estimator under small (0.25) cluster variance.

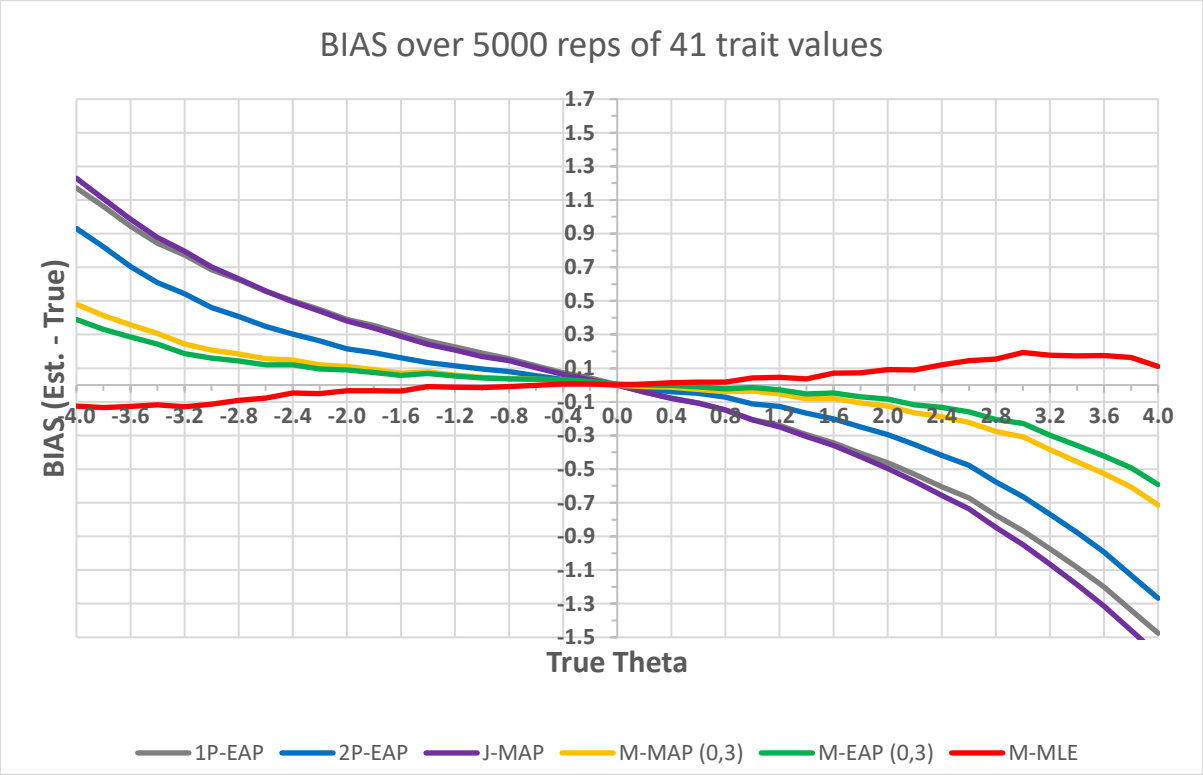


Figure 3. Bias plots for each proficiency estimator with medium (0.50) cluster variance.

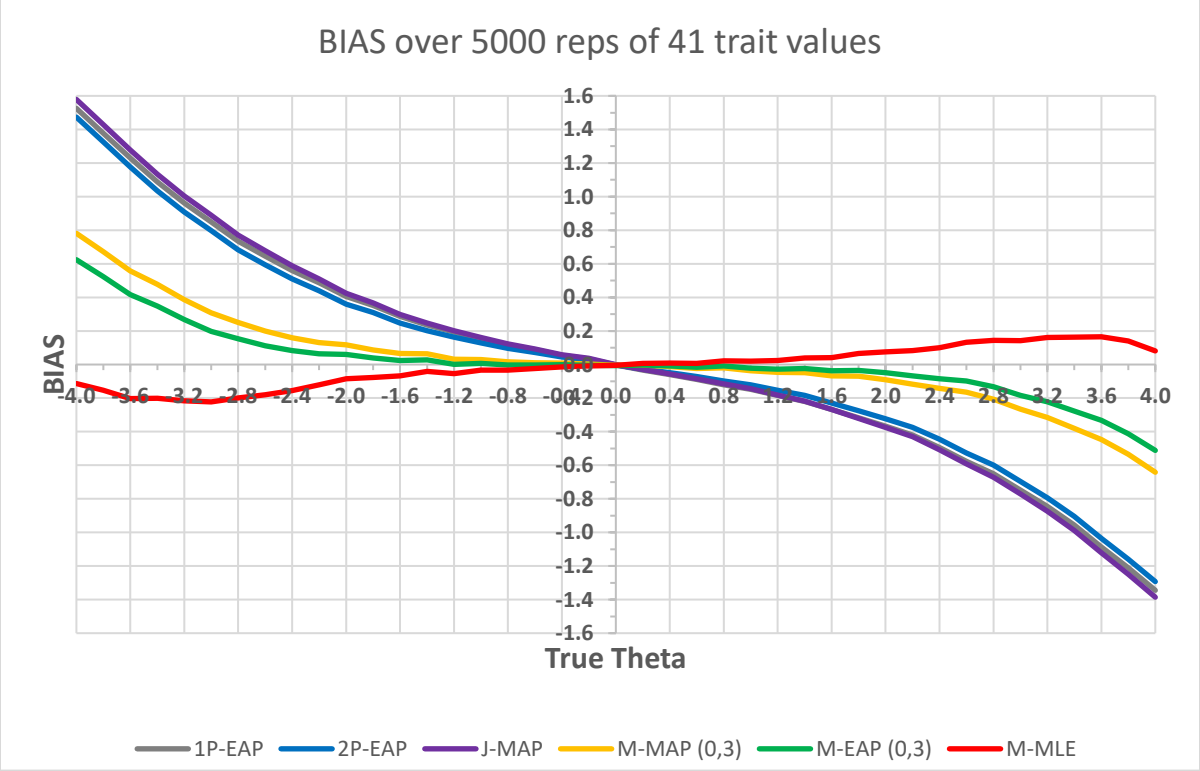

Figure 4. Bias plots for each proficiency estimator with large (0.75) cluster variance.

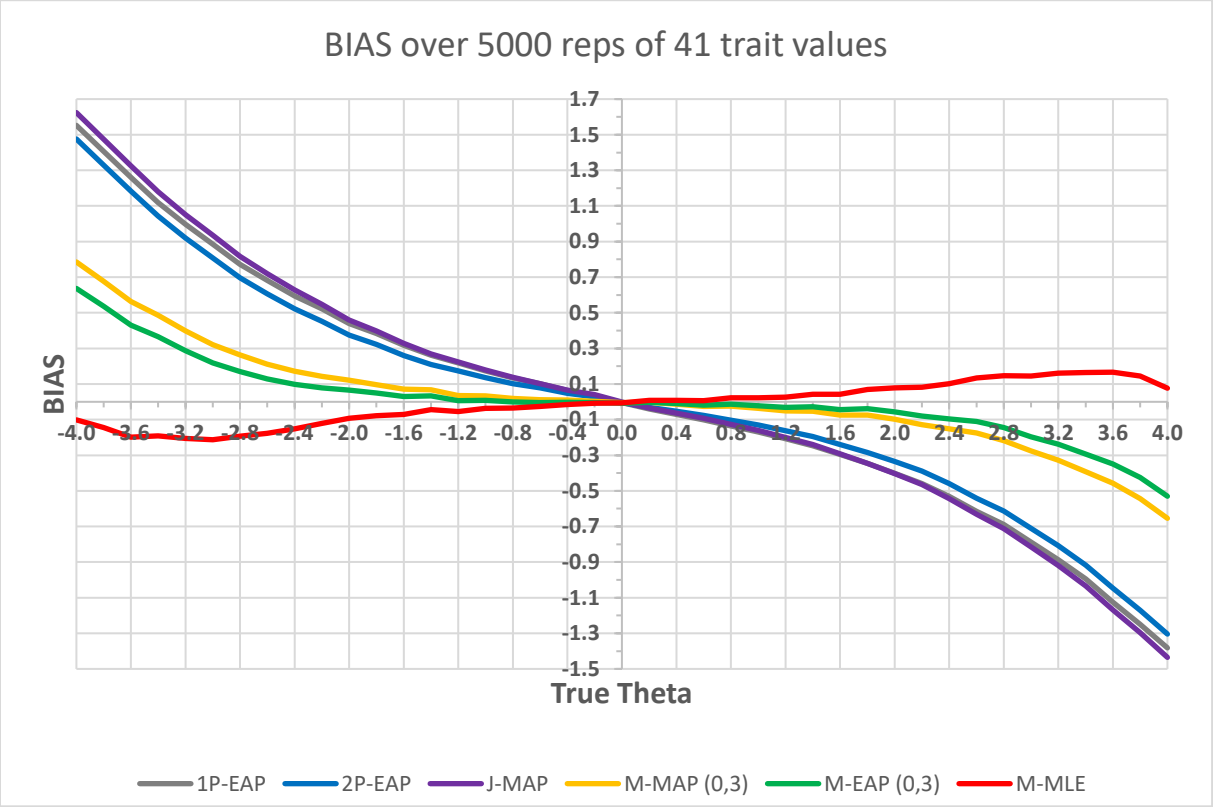


Figure 5. Bias plots for each proficiency estimator with very large (1.00) cluster variance.

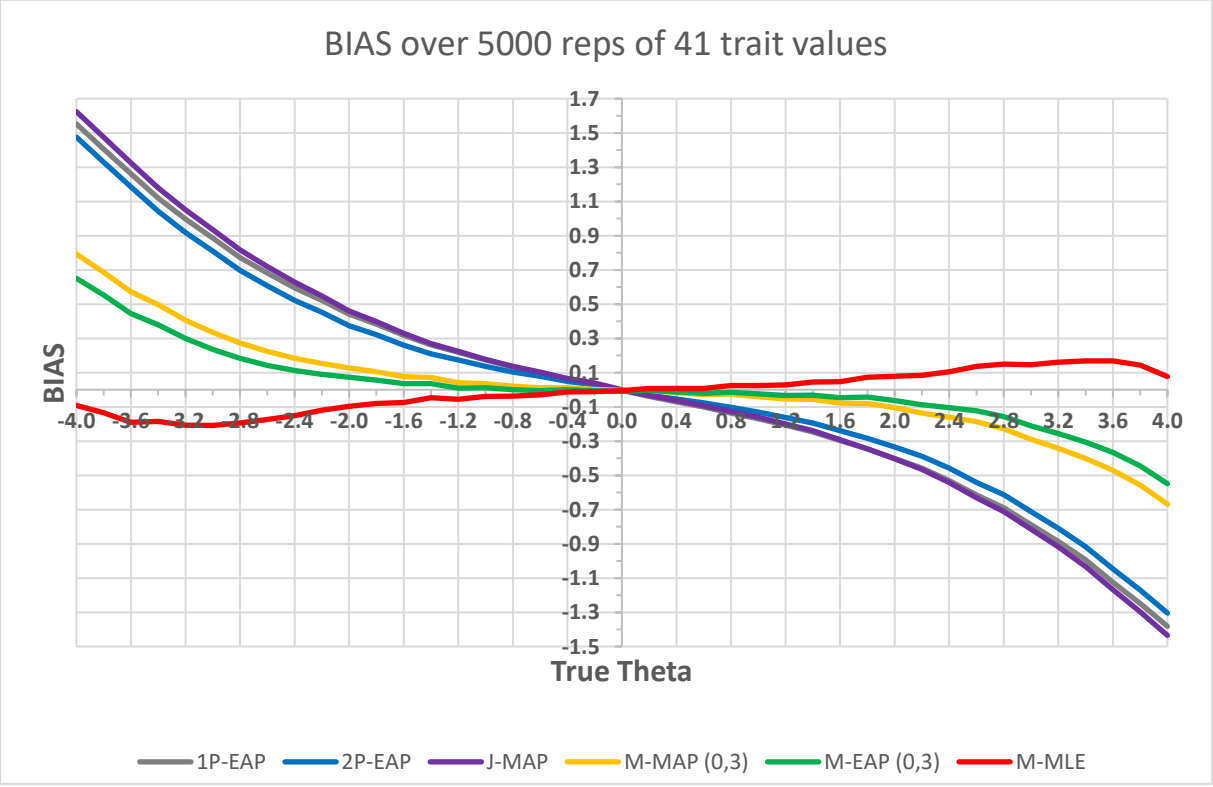

Figure 6. Accuracy (RMSE) plots for each proficiency estimator with small (0.25) cluster variance.

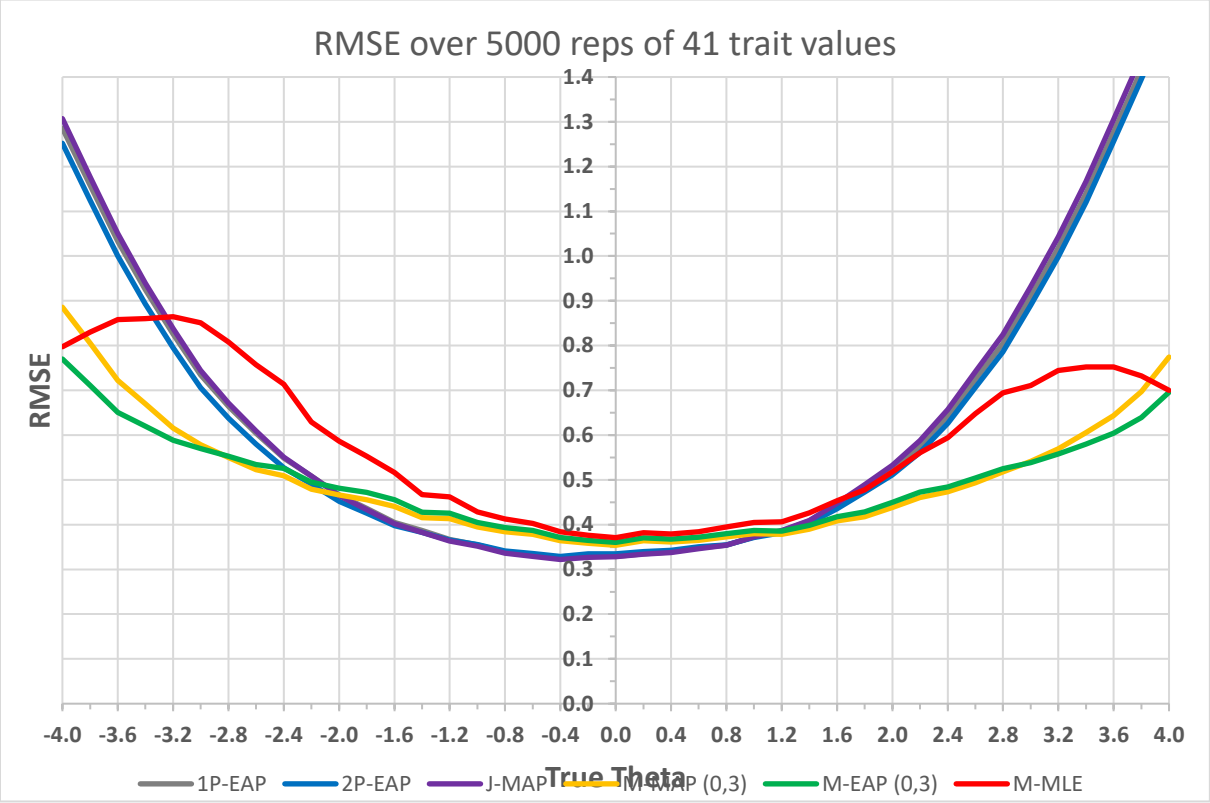


Figure 7. RMSE plots for each proficiency estimator with medium (0.50) cluster variance.

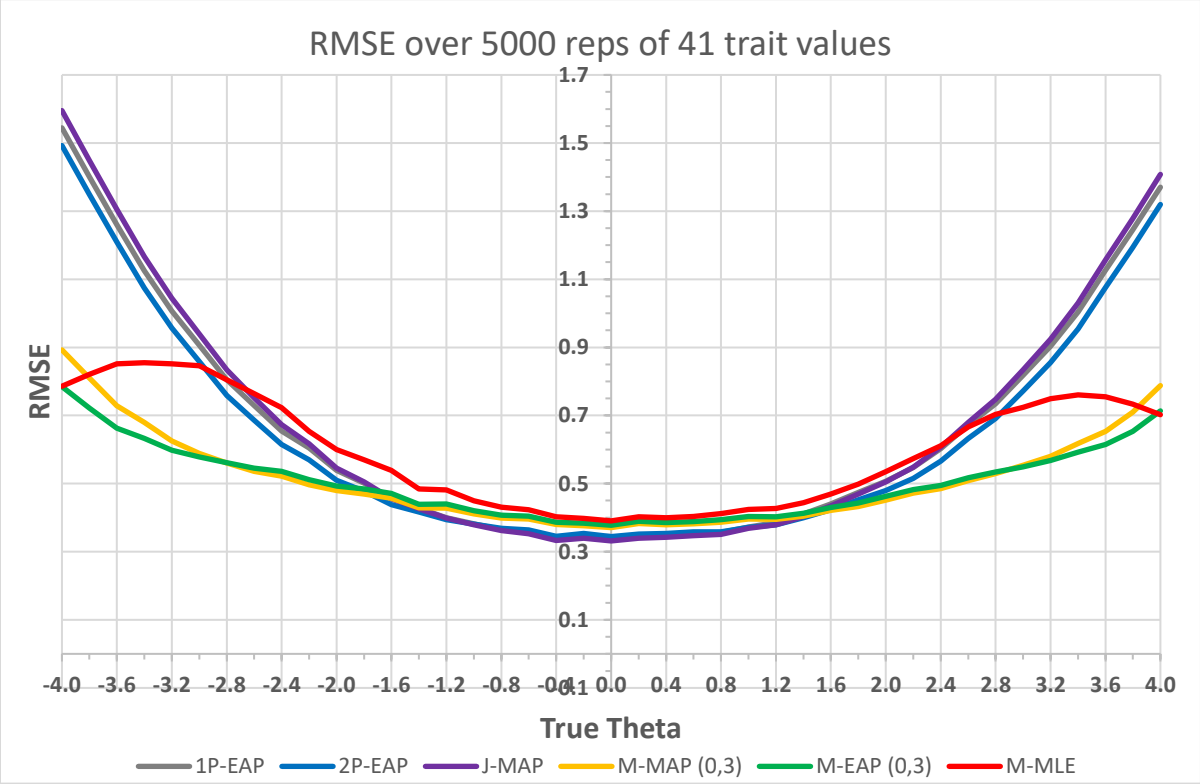

Figure 8. Accuracy (RMSE) plots for each proficiency estimator with large (0.75) cluster variance.

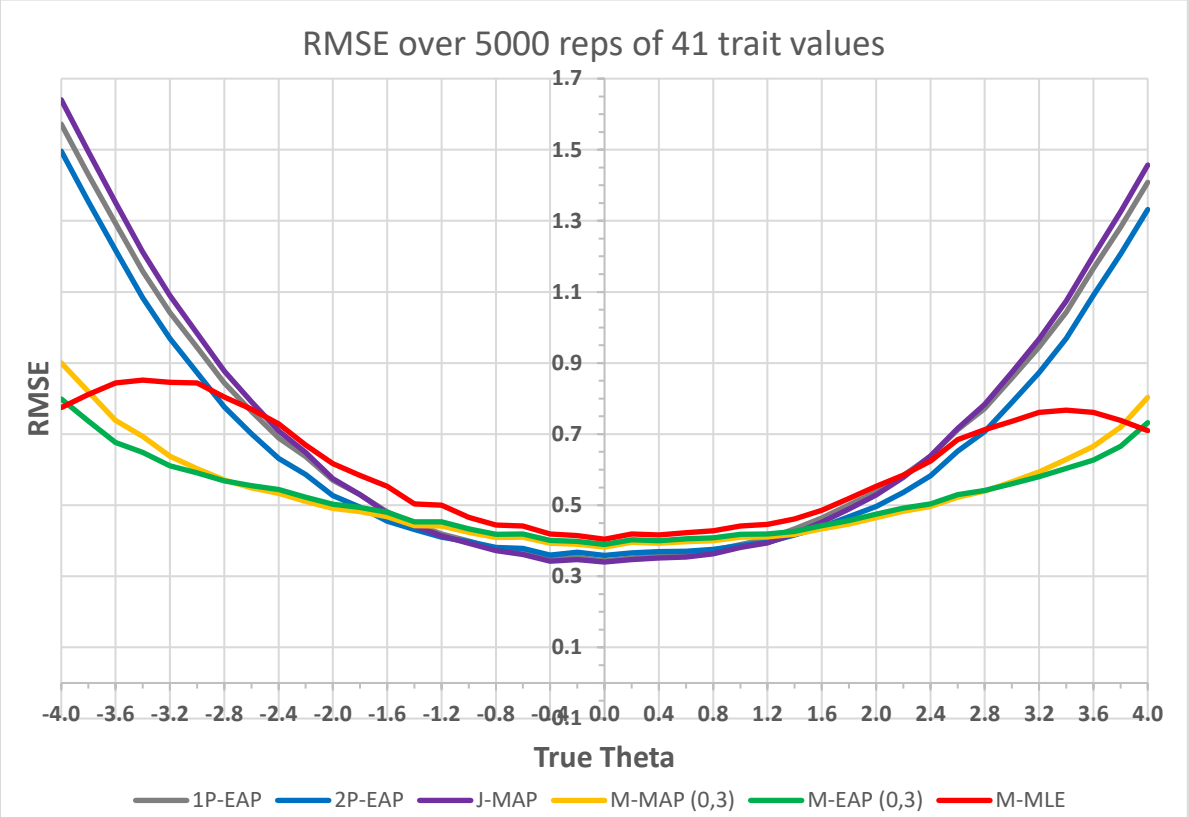


Figure 9. RMSE plots for each proficiency estimator with very large (1.00) cluster variance.

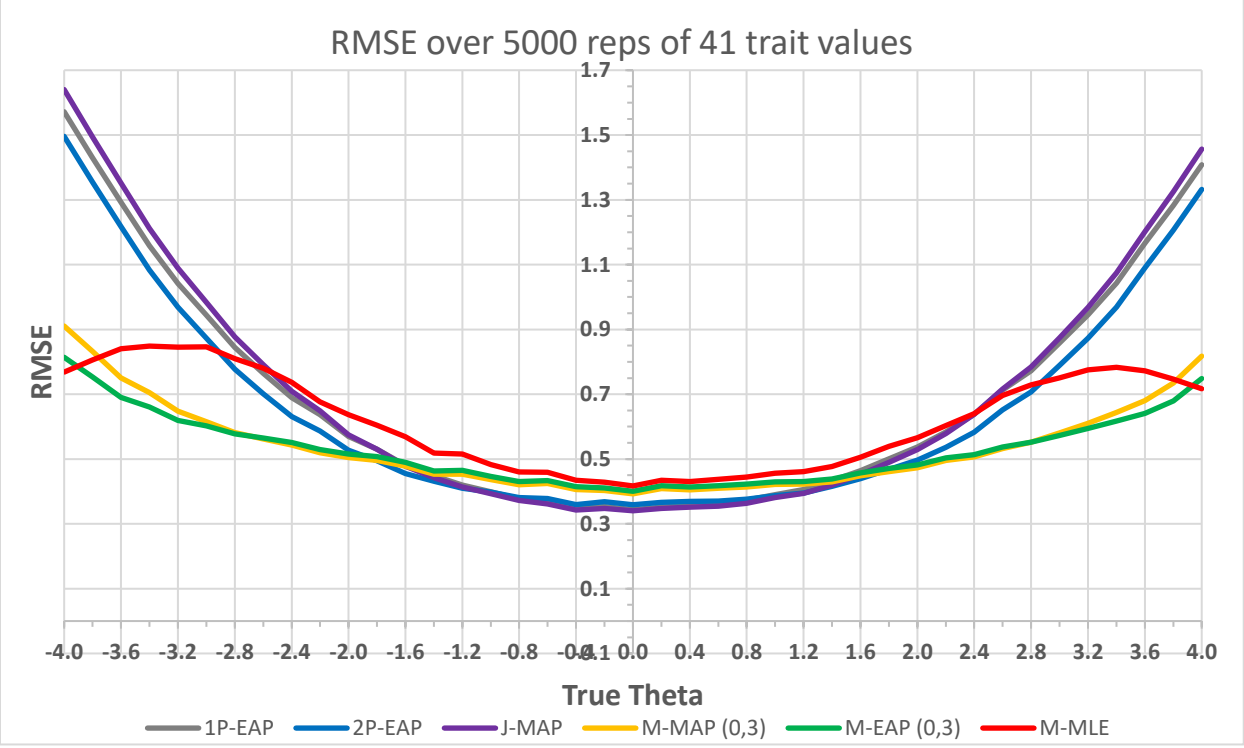

Figure 10. Bias plots for each proficiency estimator with increased number of assertions and differing cluster variances.

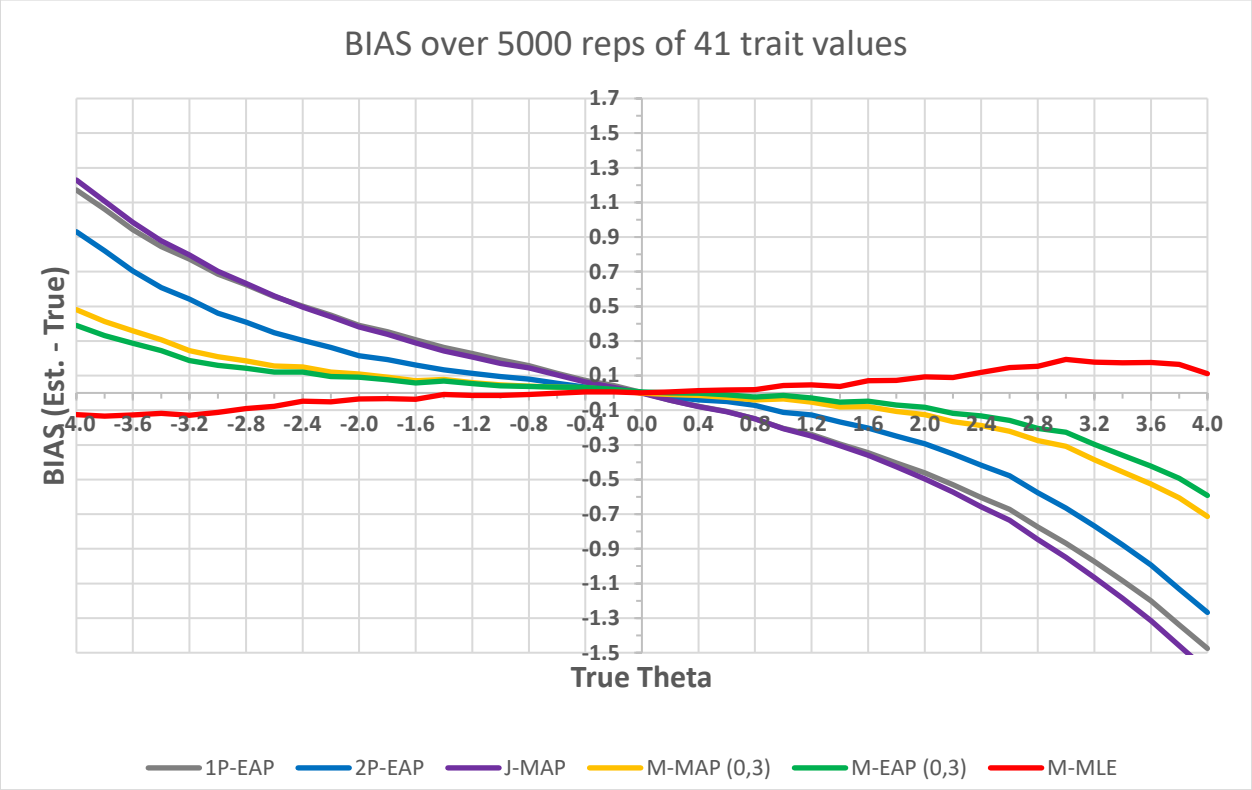


Figure 11. RMSE for each proficiency estimator with increased number of assertions and differing cluster variances.