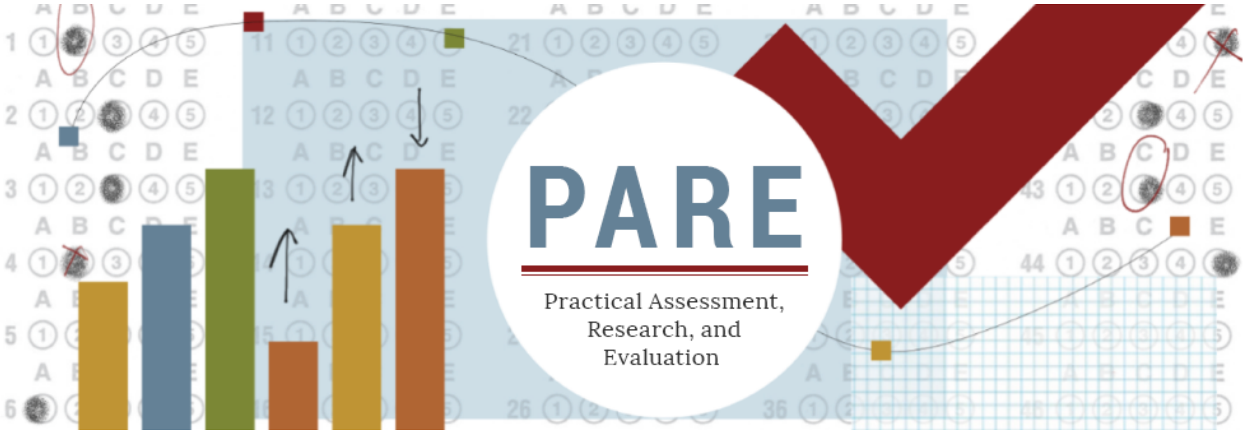


A peer reviewed, open-access electronic journal: ISSN 1531-7714

# Item Parameter Estimation of the 2PL IRT Model With Fixed Ability Estimates: Choices of Ability Estimation Methods and Priors on Slopes

Jianbin Fu, *Educational Testing Service*
TsungHan Ho, *Educational Testing Service*
Xuan Tan, *Educational Testing Service*

**Abstract:** Item parameter estimation using an item response theory (IRT) model with fixed ability estimates is useful in equating with small samples on anchor items. The current study explores the impact of three ability estimation methods (weighted likelihood estimation [WLE], maximum a posteriori [MAP], and posterior ability distribution estimation [PST]) and three types of priors set for slopes (true lognormal prior, alternative lognormal prior less informative than the true prior, and no prior) on the item parameter estimations of the two-parameter logistic (2PL) model under different conditions with varying slope mean, slope standard deviation, and data noise in a simulation study. The model is also applied to a real dataset, and the results from the three ability estimation methods and three priors on slopes are compared. The MAP ability estimation with true prior on slopes is recommended as the best choice, followed by the WLE ability estimation with less informative prior on slopes. In practice, it is recommended to use the MAP ability estimation and empirical lognormal prior on slopes and normal prior on intercept derived from a historical dataset (e.g., item bank). If a testing program prefers the WLE ability estimation, then couple it with a less informative prior on slopes than the empirical one from historical data.

**Keywords:** discrimination/slope prior, two-parameter logistic (2PL) model, item parameter estimation with fixed ability estimates

## Introduction

In item response theory (IRT) models, item and ability parameters are usually estimated simultaneously by the maximum marginal likelihood estimation with an expectation-maximization algorithm (MML-EM;

Bock & Aitkin, 1981; Fu, 2019) or the Markov chain Monte Carlo (MCMC) method (Gilks et al., 1996; Tierney, 1994). Item parameter estimations for IRT models with fixed ability parameters have received little attention, especially for the traditional unidimensional IRT models, for example, the two-parameter logistic (2PL) model (Birnbaum, 1968). There may be two main reasons for that. The first is that IRT models with fixed ability parameters are rarely used in practice. Second, item parameter estimations for IRT models with fixed ability parameters are a straightforward multivariate numerical optimization problem. Without priors on item parameters, they are just the common probit regression or logistic regression models that had extensive research.

However, estimating item parameters of an IRT model with fixed ability estimates as a simple and straightforward linking method was discussed long ago by Stocking (1988; known as Method A). We recently found it beneficial in equating with small samples on anchor items in operational testing (Ho, 2023, 2024). In the assessment program involved in these studies, field test items are separated into small subsets (e.g., six subsets) and embedded into operational forms by a spiral design so that each test taker only sees one subset of field test items. This way, more new items can be field tested to save costs for a testing program. Normally, the scale of field test items can be linked back to the bank scale using operational items as anchors by an equating method such as the test characteristic curve method (TCC; Stocking & Lord, 1983) — the traditional method used in many assessment programs. The number of operational forms administered at a given administration can be very high (e.g., 192 forms) to avoid test exposure. However, nowadays, the sample sizes on many operational items tend to be too small (e.g., smaller than 150 test takers) to get stable estimations due to reduced numbers of test takers taking the assessment overall. One solution is to calibrate the field test items by fixing test takers' ability scores at their estimates from the operational forms using their operational bank parameters. The 2PL model is used to calibrate the operational tests. Ho (2023) evaluated four linking methods using the 2PL model based on simulated data by comparing with the TCC method in terms of item parameter recovery: (a) concurrent calibration of operational and field test items with fixed operational item parameters, (b) IRT simultaneous linking method (Haberman, 2009), (c) item calibration of field test items with fixed ability estimates from operational items and without priors on item parameters, and (d) the same as (c) except using priors on item parameters. The results showed that the parameter estimation of field test items with fixed ability estimates and item parameter priors (i.e., method d above) performed the best in terms of estimation stability and robustness across various simulated conditions, including the small (500) and large (1,000) sample sizes of the field test items (the sample sizes of the operational items varied from below 100 to above 500 in both conditions; see Table 3 in Ho, 2023).

There are many factors to consider in using the 2PL model with fixed ability parameters in equating that may impact the accuracy of item parameter estimation, for example, the sample size for field test items, the ability estimation method on operational forms, the item parameter estimation method (e.g., maximum likelihood vs. MCMC) on field test items, and priors on item parameters. There is little research in the literature providing practical guidance on these issues. Therefore, we conducted some studies to address these practical issues.

In the current paper, we focus on two factors. The first one is the ability estimation method on operational forms. Ho (2023) found that in the item parameter estimations of the 2PL model with fixed ability estimates, the weighted likelihood estimation (WLE; Warm, 1989) performed better than the maximum likelihood estimation (MLE), the MLE-Lord's bias-correction with iteration, and the inverse-TCC (test characteristic curve) scoring method. The WLE provides bias-corrected ability point estimates that are less biased than maximum likelihood estimates. Furthermore, Ho (2024) has shown that the maximum a posteriori (MAP) ability estimation (Baker & Kim, 2004) led to better item parameter recovery than WLE. In this paper, we consider a new type of ability estimation: the posterior ability distribution estimation that, for each test taker, generated a vector of posterior probability estimates at 61 equally spaced scores from -6 to 6 (referred to as PST ability estimates). With the PST ability estimates, we considered measurement errors associated with ability estimates. Some argued (e.g., Lockwood & McCaffrey, 2017) that in the application of ability estimates like this, measurement errors of ability estimates should be considered to achieve more accurate results.

The second factor considered in this paper is the prior on the slope parameter of the 2PL model. Prior information is a concept in Bayesian statistics. Numerous research has been conducted on item response theory (IRT) models under the Bayesian framework (e.g., Baker & Kim, 2004; Fox, 2010; Kim et al., 2004; Mislevy, 1986; Levy, 2009). If prior information is appropriate and informative, model estimations may become more accurate and stable under Bayesian statistics than those under frequentist statistics, especially for small sample sizes (Rupp et al., 2004). The impact of the choice of item parameter priors on the accuracy of Bayesian estimations of IRT models has been studied in many papers (e.g., Chang, 2017; Marcoulides, 2018). All these studies investigated item parameter recovery of IRT models with unknown ability parameters. In operational investigation, we accidentally found that, in item parameter estimation of the 2PL model with fixed WLE ability estimates, for the lognormal prior on slope parameters, using the true mean and standard deviation at the normal scale generally produced a better estimation of item parameters than using the true mean and standard deviation (SD) at the log scale (i.e., in the lognormal distribution, set the mean and SD parameters to be the mean and SD of the true slopes rather than the logarithms of the true slopes). This result is somehow counterintuitive and less known to the educational measurement community. Thus, conducting a simulation study to investigate this issue further is worthwhile. In the current study, we consider three prior conditions: no prior, true prior (for generating data), and alternative lognormal prior on slope parameters, as mentioned previously.

Research on item parameter estimations of the 2PL model with fixed ability estimates is relatively scarce. The current study fills this gap and is valuable to the literature. The current study also provides practical suggestions to practitioners on the 2PL model estimation with fixed ability estimates.

The rest of the paper is organized as follows. First, we introduce the item response function of the 2PL model and its Bayes modal estimation. Second, we describe the simulation study, including the design, data generation, estimation program, evaluation criteria, and results. Third, we apply the model to a real dataset and compare the results. Fourth, we summarize the main findings, provide practical recommendations, discuss the current study's limitations, and suggest further research. Finally, we provide a brief conclusion that includes the major takeaways from this study.

## The 2PL Model and Bayes Modal Estimation

The 2PL response function is

$$P_{ij} = \frac{\exp\,(s_i\theta_i + d_i)}{1 + \exp\,(s_i\theta_i + d_i)} \tag{1}$$

where $P_{ij}$ is the probability of test taker *j* answering item *i* correctly, $s_i$ is the slope parameter of item *i*, $d_i$ is the intercept parameter of item *i*, and $\theta_j$ (theta) is test taker *j*'s latent ability score. The probability of test taker *j* answering item *i* incorrectly is just $1 - P_{ij}$. The traditional discrimination parameter at the normal ogive scale is $a_i = s_i/1.702$, and the difficulty parameter is $b_i = -d_i/s_i$.

The ability parameter $\theta_j$ is often assumed to follow a standard normal distribution to identify the model. However, in the current study, we assume $\theta_j$ is known and estimate the slope parameter $s_i$ and the intercept parameter $d_i$. Because of the local independence of $P_{ij}$ (i.e., a test taker's response to an item is independent of their responses to other items in a test given their latent ability score), item parameter estimation with fixed ability parameters can be done item by item. In the Bayes modal estimation of item *i*, the function to maximize is the sum of the log-likelihood function,

$$l_i = \sum_j^J \left[x_{ij} \log P_{ij} + \left(1 - x_{ij}\right) \log\left(1 - P_{ij}\right)\right], \tag{2}$$

and the logarithm of the densities of the prior distributions of slope and intercept parameters, that is,

$$F_i = \sum_j^J \left[x_{ij} \log P_{ij} + \left(1 - x_{ij}\right) \log\left(1 - P_{ij}\right)\right] + \log \pi(s_i) + \log g(d_i) \tag{3}$$

where $x_{ij}$ is test taker *j*'s binary score (0 = incorrect, or 1 = correct) on item *i*, $\pi(s_i)$ and $g(d_i)$ are the density functions of the prior distributions of the slope and intercept parameters, respectively. For intercept parameters, a normal prior distribution is commonly used. Because slope parameters are often assumed to be positive numbers in achievement tests, a lognormal prior distribution is often used for slope parameters as it only allows positive numbers (Baker & Kim, 2004). A lognormal distribution has two distribution parameters, mean and standard deviation at the log scale. For example, a lognormal prior distribution for a random slope parameter, *s*, has mean $\mu_s^l$ and standard deviation $\sigma_s^l$, where $\mu_s^l$ and $\sigma_s^l$ are the mean and standard deviation of log(*s*), respectively. To obtain the mean $\mu_s$ and standard deviation $\sigma_s$ of *s* at the normal scale, we need to use the following transformation,

$$\mu_s = \exp\left(\mu_s^l + \sigma_s^{l2}/2\right), \tag{4}$$

$$\sigma_s = \exp\left(2\mu_s^l + 2\sigma_s^{l2}\right) - \exp\left(2\mu_s^l + 2\sigma_s^{l2}\right). \tag{5}$$

Equivalently, we can set $s_i = \exp(z_i)$ where $z_i$ follows a normal distribution with mean $\mu_s^l$ and standard deviation $\sigma_s^l$. Then $s_i$ follows a lognormal distribution with mean $\mu_s^l$ and standard deviation $\sigma_s^l$ at the log scale. Some researchers like to work with $z_i$ (e.g., Baker & Kim, 2004; Mislevy, 1986) because it appears convenient to work with the normal distributions as the priors for both slope and intercept in estimation.

Finding the slope and intercept maximizing Equation 2 is referred to as maximum likelihood estimation, and maximizing Equation 3 is Bayes modal estimation. Equations 2 and 3 can be optimized by the Newton–Raphson method (NR; Atkinson, 1991) or a quasi-Newton method (Nocedal & Wright, 2006), for example, the BFGS (Broyden–Fletcher–Goldfarb–Shanno) algorithm (Broyden, 1970).

## Simulation Study

The simulation study followed the equating process with fixed ability estimates obtained from operational forms to scale field test items back to the bank scale. In particular, test takers take both operational forms and field test items in the same testing session. Then, the test takers' ability scores are estimated by the operational items with fixed item parameters at bank values, and the field test items are calibrated with fixed ability scores estimated from the operational items.

### Method

***Data Simulation***. Three factors were manipulated to generate simulated datasets: slopes' mean, SD, and data fit. The first two were considered because one of the study's main purposes was to investigate the appropriateness of the priors on slopes, and we would like to inspect if the mean and SD of the slopes had an impact on the choice of the prior on slopes. The slopes' means had three levels: low (0.7), medium (1.23), and high (1.80), corresponding to the 10th percentile, mean, and 90th percentile of the item slopes, respectively, in a large-scale assessment program. The slopes' SDs also had three levels: low (0.2), medium (0.44), and high (1), with the medium corresponding to the SD of the item slopes in the real assessment program. All the means and SDs in the nine conditions were converted to the logarithm scale (see Table 1). Then, the slope parameters in a condition were drawn from the lognormal distribution with the corresponding mean and SD on the logarithm scale.

A simulated dataset included 1,000 test takers, 40 operational items, and 115 field test items in all conditions. The ability parameters were drawn from a standard normal distribution. The intercept parameters were drawn from a normal distribution with a mean of 0.62 and an SD of 1.19. These setups were also based on the large-scale assessment program. Note that we did not manipulate these parameters because we tried to make the simulation study manageable, and these parameters appeared less relevant to the focus of the current study. The impact of the ability distribution and sample size of field test items on the 2PL model estimation with fixed ability estimates has been explored in Ho (2023, 2024).

**Table 1.** Simulation Conditions on Slopes

| *SD* | *M* | | |
|---|---|---|---|
| | Low (0.7) | Medium (1.23) | High (1.8) |
| Low (0.2) | *M* =-0.40, *SD* = 0.28 | *M* = 0.19, *SD* = 0.16 | *M* = 0.58, *SD* = 0.11 |
| Medium (0.44) | *M* =-0.52, *SD* = 0.58 | *M* = 0.15, *SD* = 0.35 | *M* = 0.56, *SD* = 0.24 |
| High (1) | *M* =-0.91, *SD* = 1.05 | *M* = -0.05, *SD* = 0.71 | *M* = 0.45, *SD* = 0.52 |

*Note*. The means and SDs in the table headings are at the normal scale; the means and SDs in the table cells are at the logarithm scale.

We also manipulated the data-fit factor because data noise commonly exists in operational test data. This factor had three levels: no misfit, difficulty misfit, and discrimination misfit. In the no-misfit condition, the true item parameters described previously were used to generate simulated data. In the difficulty-misfit condition, a uniform random number between 0 and 0.5 was subtracted from the true difficulty parameter ($b_i$) of each of the first 20 operational items. Those modified difficulty parameters were transformed into the intercepts ($d_i$). The modified intercepts and the other true item parameters were used to generate simulated data. In the discrimination-misfit condition, a uniform random number between 0 and 0.5 was added to the true discrimination parameter ($a_i$) of each of the first 20 operational items. Those modified discrimination parameters were transformed into the slopes ($s_i$). The modified slopes and the other true item parameters were used to generate simulated data. Table 1 lists the true parameter distributions in the three data generation conditions.

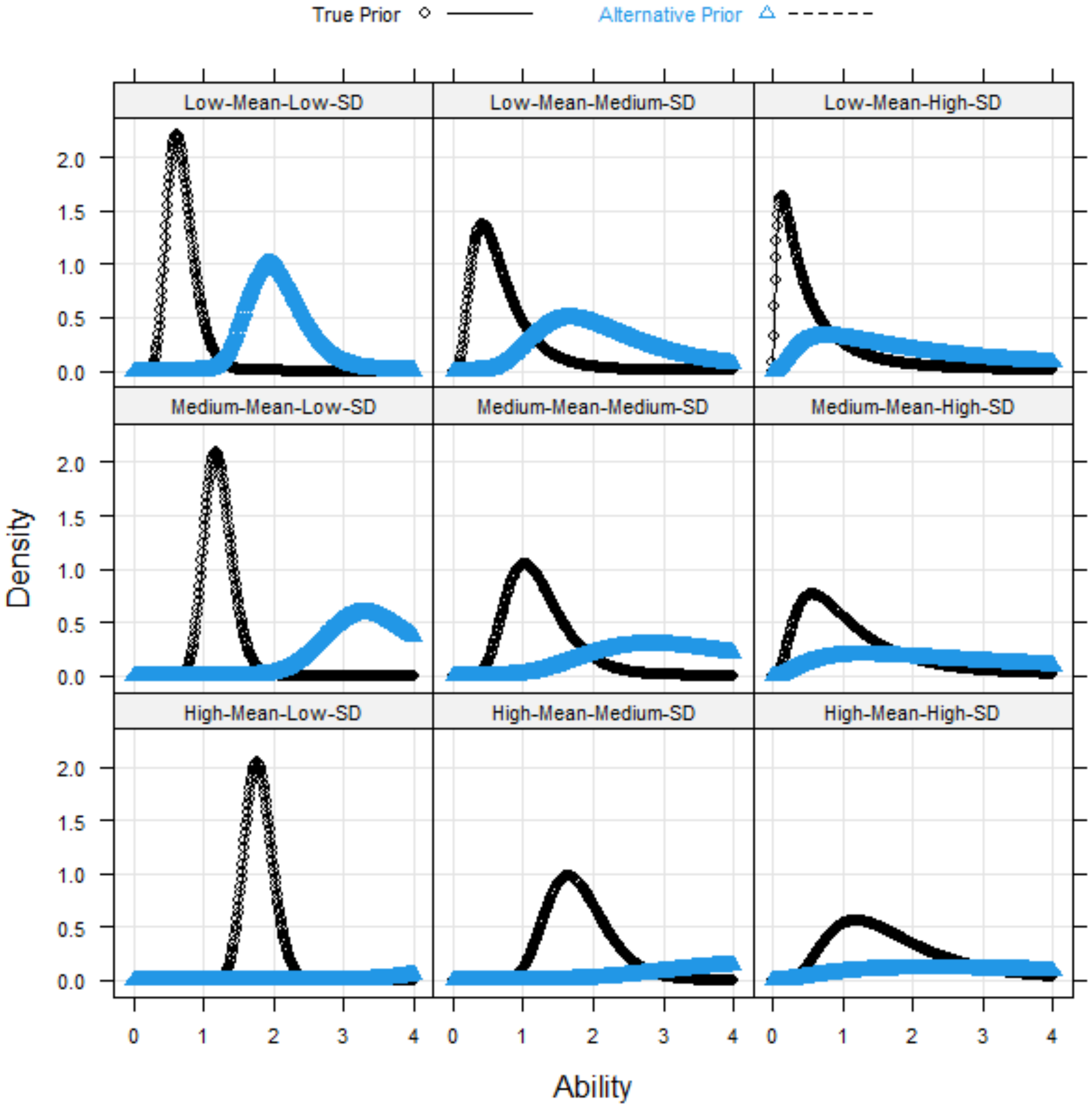
Under each condition, 100 simulated datasets were generated; for each simulated dataset, a new set of model parameters was drawn.

***Model Estimations***. The 40 operational items' parameters were assumed to be known and fixed to their true values. These 40 operational items were used to calibrate test takers' latent ability scores with fixed item parameters. Three types of ability estimations were conducted as described previously: WLE, MAP, and PST. In addition, the true ability parameters were used as fixed ability estimates as a baseline model.

With fixed test takers' ability estimates, the item parameters of the 115 field test items were calibrated with three different priors on slopes. The first one did not impose any prior on item parameters. The second one imposed the true prior distribution on item parameters, that is, for the slope parameters, a lognormal distribution with the mean and SD in each condition as listed in Table 1, and for the intercept parameters, a normal distribution with a mean of 0.62 and an SD of 1.19. The third one was like the second one, except that the alternative distribution parameters are used for the lognormal distribution. In particular, the mean and SD at the normal scale in each condition were used directly in the lognormal function, leading to a much flatter (noninformative) slope distribution than the true one. For example, in the medium-mean-medium-SD condition, using the mean, 1.23, and the SD, 0.44, at the normal scale in the lognormal function led to the slope parameters' mean and SD being 3.80 and 1.78, respectively. Figure 1 shows the density curves of the true and alternative lognormal prior distributions on the slope parameters in each condition. The alternative prior had a much flatter density curve than the true prior in each condition. Even for some alternative density curves with moderate slopes (e.g., in the low-mean-low-SD and medium-mean-low-SD conditions), the peaks of the densities were much lower and located on higher ability scores than those on the true density curves. The no-prior estimation of item parameters with fixed ability parameters employed the maximum likelihood estimation, while the estimations with the alternative or true prior used the Bayes modal estimation.

In combination, there were 12 estimation conditions: four ability estimates (WLE, MAP, PST, and True) by three prior distributions on slopes (alternative prior, true prior, and no prior). The 12 estimation conditions were crossed with the 27 data generation conditions (i.e., three slope means by three slope SDs

**Figure 1.** Density Curves of True and Alternative Lognormal Prior Distributions



***Estimation Program.*** All analyses were conducted in the 64-bit R 4.2.1 program (R Core Team, 2022). The WLE and MAP ability estimations with fixed item parameters were done using the `fscores` function in the `mirt` 1.37 package (Chalmers, 2012). The first author modified the `fscores` function to obtain PST ability estimates. PST ability estimates are a by-product of the expected a posteriori (EAP) score estimation (Baker & Kim, 2004) at an intermediate step. The revised `fscores` function outputs PST ability estimates before EAP score estimates are calculated at the final step.

**Table 2.** All Simulation Conditions

| Factor | Number of levels | Levels |
|---|---|---|
| Data generation | 27 | |
| Slope mean | 3 | low (0.7), medium (1.23), high (1.8) |
| Slope SD | 3 | low (0.2), medium (0.44), high (1) |
| Data fit | 3 | no misfit, difficulty misfit, discrimination misfit |
| Estimation | 12 | |
| Types of ability estimates | 4 | WLE, MAP, PST, True |
| Priors on item parameters | 3 | alternative prior on slopes, true prior, no prior |

*Note.* All factors were fully crossed, leading to a total of 324 conditions. In each of the 27 data generation conditions, 100 datasets were simulated.

We also modified the relevant functions in `mirt` to estimate item parameters with fixed ability parameters. This estimation can be treated as one cycle of the MML-EM estimation of item parameters. In particular, the fixed ability scores are used in the expectation step (E-step) to calculate the expected number of test takers in an item score category by ability score table. The expected count table is then used in the maximization step (M-step) to find the item parameters that maximize Equation 2 or 3. For a 2PL model without priors, the optimization method is 'BFGS' in the R function, `optim`, and with priors, the optimizer is the R function, `nlminb`. Both functions use the quasi-Newton method. Within the modified functions, item parameter estimation is done for one item at a time. The maximum number of iterations in the M-step was set to 200; all the other default settings in the `mirt` function were kept. All estimation functions used in this study are available from the first author.

***Evaluation Criteria.*** We estimated item slopes $s_i$ and intercepts $d_i$ because, in `mirt`, the 2PL model was implemented based on Equation 1. The traditional item parameters, discrimination at the normal ogive scale ($a_i = s_i/1.702$) and the difficulty parameter ($b_i = -d_i/s_i$), are more commonly used in practice. Thus, we transformed the item slopes and intercepts to the discrimination and difficulty parameters and evaluated their recovery. The following two criteria were used:

$$\text{1. Mean Bias} = \sum_{r=1}^{R}\left[\sum_{i=1}^{I}(\hat{t}_{ri} - t_{ri})/I\right]/R, \tag{6}$$

where $\hat{t}_{ri}$ refers to the estimated item parameter (discrimination or difficulty) of item *i* in simulated dataset *r*, $t_{ri}$ is the true item parameter, *I* (=115) is the total number of field test items in a simulated dataset, and *R* (=100) is the total number of datasets.

$$\text{2. Root Mean Square Error (RMSE)} = \sum_{r=1}^{R}\left[\sum_{i=1}^{I}(\hat{t}_{ri} - t_{ri})^2/I\right]^{\frac{1}{2}}/R\,. \tag{7}$$

We also calculated the mean biases for the WLE and PST ability estimates and SDs of ability estimates. For the WLE estimates, the formulas for the mean biases were similar to Equation 6. For the PST estimates, each test taker's ability estimates are represented by 61 equally spaced points between -6 and 6 with weights at each point. Thus, the formula to calculate the mean bias of the PST estimates is

$$\text{Mean Bias} = \sum_{r=1}^{100}\left[\sum_{j=1}^{1000}\sum_{h=1}^{61}\left(\theta_h - \theta_{rj}\right)\widehat{w}_{rjh}/1000\right]/100, \tag{8}$$

where $\theta_h$ is the $h^{\text{th}}$ (=1,…,61) ability point between -6 and 6, $\theta_{rj}$ is the true ability of test taker *j* in dataset *r*, and $\widehat{w}_{rjh}$ is the estimated weight at $\theta_h$ for test taker *j* in dataset *r*. The formula for the mean bias of the SDs of the PST estimates is

$$\text{Mean Bias} = \sum_{r=1}^{100}\left(\widehat{\text{PST}}\_\text{SD}_r - \theta\_\text{SD}_r\right)/100 \tag{9}$$

$$\widehat{\text{PST}}\_\text{SD}_r = \sqrt{\sum_{j=1}^{100}\sum_{h=1}^{61}\left(\theta_h - \widehat{\text{EAP}}_r\right)^2\widehat{w}_{rjh}/1000,} \tag{10}$$

$$\text{EAP}_r = \sum_{j=1}^{1000}\sum_{h=1}^{61}\theta_h\widehat{w}_{rjh}/1000 \tag{11}$$

where $\widehat{\text{PST}}\_\text{SD}_r$ is the SD of the PST ability estimates in dataset *r*, $\theta\_\text{SD}_r$ is the SD of the true ability scores in dataset *r*, and $\widehat{\text{EAP}}_r$ is the estimated mean of the PST ability estimates in dataset *r*.

**Results**

***Discrimination and Standard Deviations of Ability Estimates.*** As mentioned previously, the simulation study had 324 conditions. To make the comparisons of the discrimination estimates manageable and meaningful, we compared the discrimination estimates across the 12 combinations of the priors by ability estimation methods within each of the 27 combinations of the slope means by slope SDs by data fits. We drew multipanel plots for the comparisons. Each panel represented one combination of the slope means by slope SDs by data fits, and within each panel, we drew the mean biases or RMSEs of the three ability estimations against the three prior conditions. Because we focused on the comparisons within each panel rather than across panels, we used different scales on the y-axis in each panel to make the within-panel comparisons easier. In a rare case, an ability estimate might not converge: the maximum nonconvergence rates in a panel were 0.025% and 0.008% for WLE and MAP ability estimates, respectively. These cases were removed from the analyses and item parameter estimations. All item parameter estimations converged.

Figures 2 and 3 show the comparisons of mean biases and RMSEs, respectively, of discrimination estimates under the medium-mean slope condition. We have the following observations for both plots.

1. MAP generally performed better than WLE, which outperformed PST across all conditions. The discrimination estimates from MAP were the closest to those estimated from the true ability scores. There were a few exceptions: (a) in the low-SD slope condition, using the alternative

priors, WLE and PST had more accurate discrimination estimates than MAP in the no-misfit and difficulty-misfit conditions, and similar estimates in the discrimination-misfit condition; and (b) WLE and PST performed similarly under the discrimination-misfit and low- or medium-SD slope condition or the no-misfit, low-SD, and alternative-prior condition.

2. For WLE and PST, the discrimination parameters were underestimated in all conditions except for the low-SD, alternative-prior, and no-misfit or difficulty-misfit conditions. For MAP, the mean biases of the discrimination estimates were close to those from the true ability scores in the no-misfit and difficulty-misfit conditions. In these conditions, the mean biases of the discrimination estimates from the MAP and true ability scores were close to 0 in the true-prior and no-prior or high-SD conditions, while the discrimination parameters from the MAP and true ability scores were overestimated in the alternative-prior and low- or medium-SD conditions. In the discrimination-misfit condition, the discrimination parameters based on MAP were underestimated except for the alternative-prior and low-SD condition.

**Figure 2.** Medium-Mean Slope Condition: Comparisons of Mean Biases of Discrimination Parameters

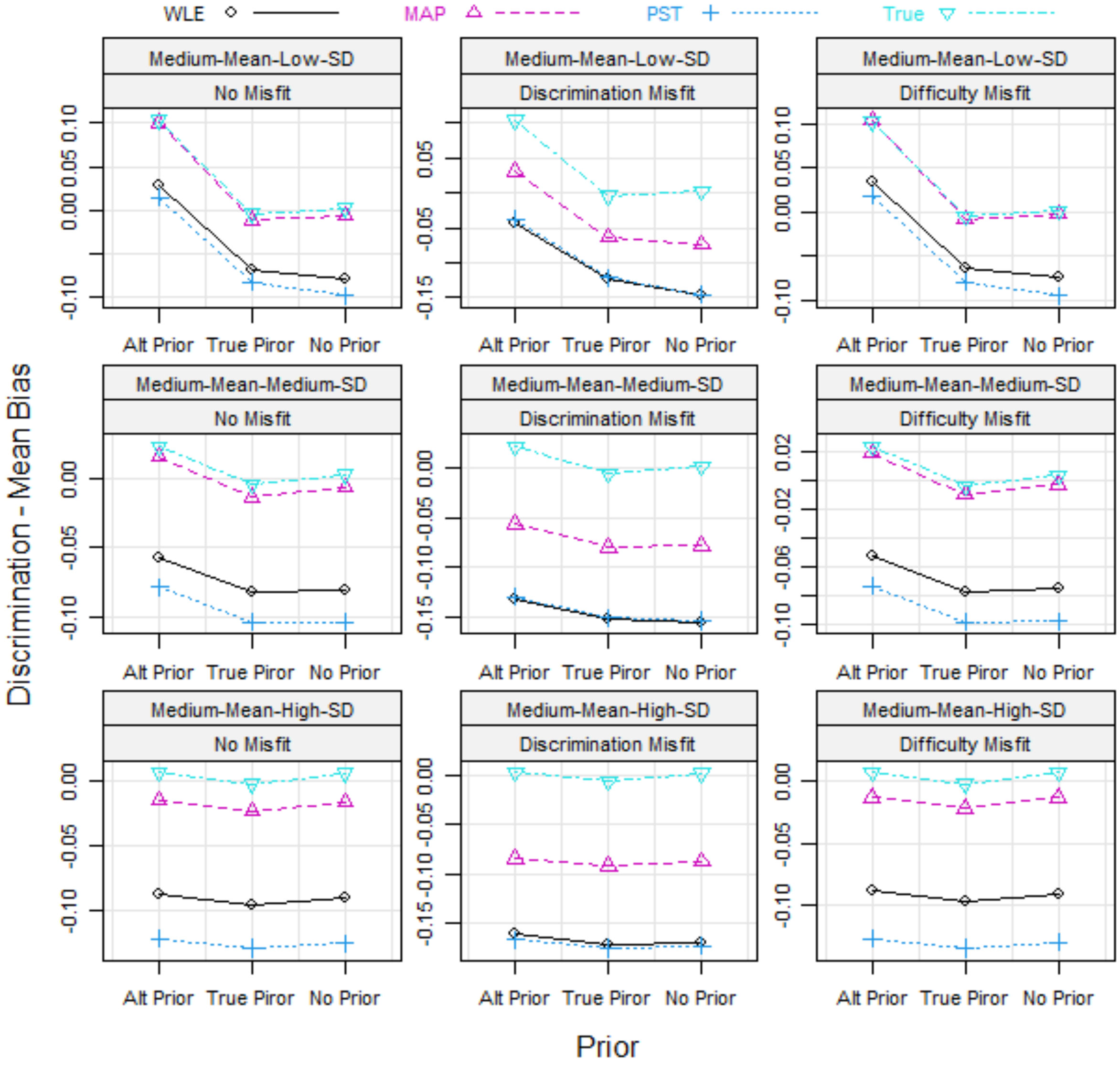
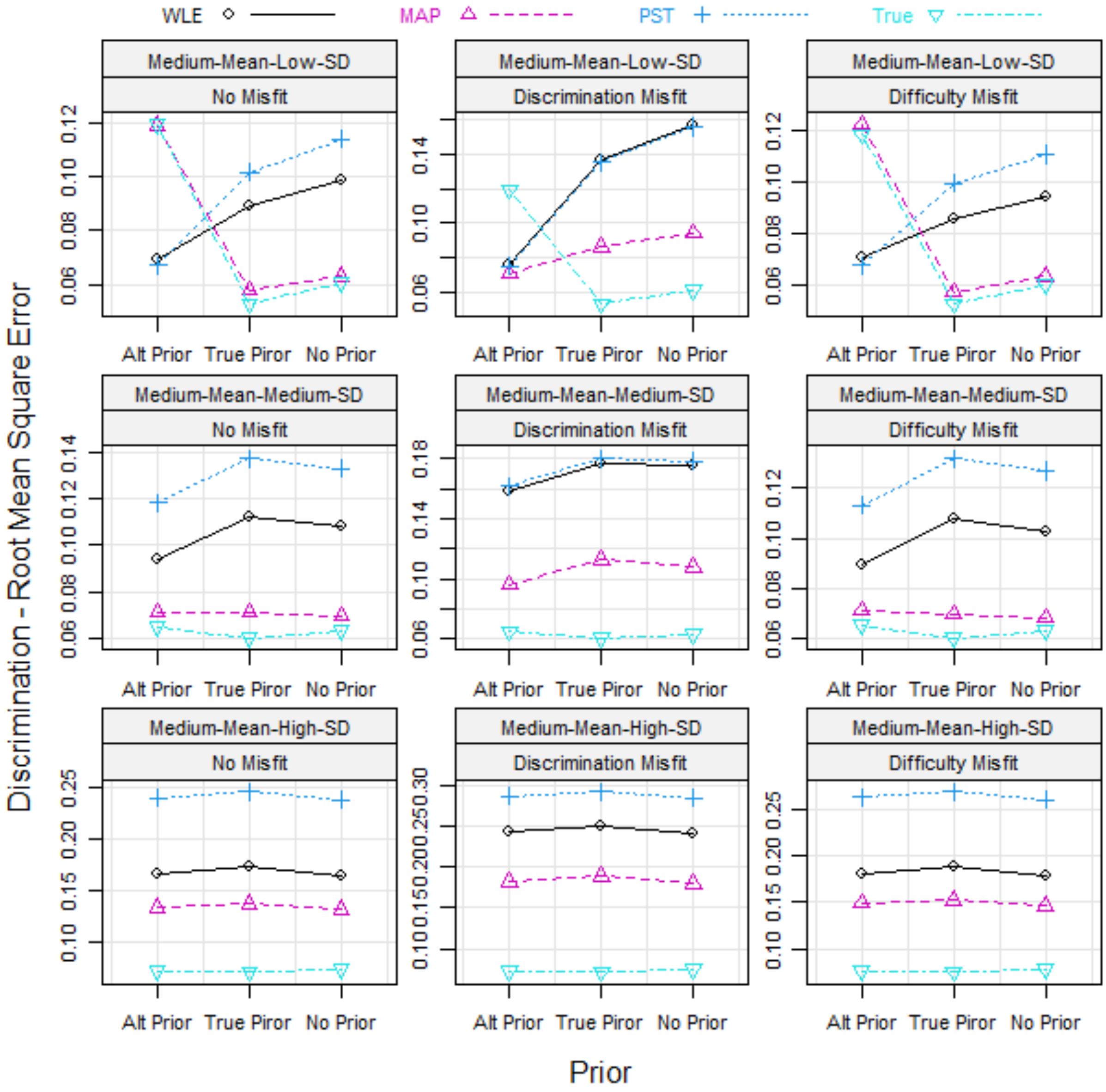


**Figure 3.** Medium-Mean Slope Condition: Comparisons of RMSEs of Discrimination Parameters



3. For WLE and PST, the discrimination parameters were underestimated in all conditions except for the low-SD, alternative-prior, and no-misfit or difficulty-misfit conditions. For MAP, the mean biases of the discrimination estimates were close to those from the true ability scores in the no-misfit and difficulty-misfit conditions. In these conditions, the mean biases of the discrimination estimates from the MAP and true ability scores were close to 0 in the true-prior and no-prior or high-SD conditions, while the discrimination parameters from the MAP and true ability scores were overestimated in the alternative-prior and low- or medium-SD conditions. In the discrimination-misfit condition, the discrimination parameters based on MAP were underestimated except for the alternative-prior and low-SD condition.

4. The accuracy of the discrimination estimations was similar in the no-misfit and difficulty-misfit conditions and much worse in the discrimination-misfit conditions.

5. In the high-SD conditions, the different priors did not significantly impact the discrimination estimations. In the low- and medium-SD conditions, the alternative priors were better than the true priors for WLE and PST. For WLE and PST, the true priors were better than no prior in the low-SD conditions and slightly worse in the medium-SD conditions. In the medium-SD conditions, the three prior conditions did not have noticeable effects on the discrimination estimations based on the MAP and true ability scores, except that in the discrimination-misfit condition, for MAP, the alternative prior was slightly better than no prior that was, in turn, slightly better than the true prior. In the low-SD conditions, for the MAP and true ability scores, the true priors had the most accurate discrimination estimations, and the alternative priors had the worst, except that in the discrimination-misfit condition, for MAP, the alternative priors performed best while no prior performed the worst.

The comparison results for the low- and high-mean slope conditions were generally similar to those for the medium-mean slope conditions; we provide the figures for these two groups in the Appendix. Some noticeable differences are listed below.

1. In the low-mean conditions, for MAP, the alternative prior was worse than the true prior and no prior in the low-SD and discrimination-misfit condition and the medium-SD and no-misfit or difficulty-misfit condition.
2. In the high-mean and low-SD conditions, for WLE, the alternative prior was worse than the true prior and no prior in the no-misfit and difficulty-misfit conditions.

Based on the 2PL response function (Equation 1), the discrimination estimation is closely related to the SD of the ability estimates: an overestimated SD of the ability estimates will cause underestimated discrimination parameters on average, and vice versa. This rule is generally true in our simulation study. Figure 4 is a multipanel plot comparing the mean biases of the SDs of ability estimates. The panels are the nine combinations of the slope mean and SD conditions. Across all panels, the patterns are similar: (a) the SDs of the WLE estimates were overestimated the most across the three data fit conditions; (b) the SDs of the PST estimates matched the true ones in the no-misfit and difficulty-misfit conditions and were overestimated in the discrimination-misfit conditions; (c) the SDs of the MAP estimates were close to the true ones in the discrimination-misfit conditions and were overestimated in the no-misfit and difficulty-misfit conditions; and (d) across the three ability estimation methods, the discrimination-misfit conditions had the largest SDs of ability estimates comparing to the no-misfit and difficulty-misfit conditions that had similar SD estimates. The last pattern (d) is because the inflated discriminate parameters in generating the simulated data in the discrimination-misfit condition led to the overestimated SDs of the ability estimates.

Comparing the mean biases between the discrimination estimates and SDs of the ability estimates, we found that they were negatively correlated in almost all conditions: a higher mean bias of the SDs of the ability estimates corresponded to a lower mean bias of discrimination estimates. For example, compared to the no-misfit and difficulty-misfit conditions, the discrimination-misfit conditions had the largest mean biases of SDs of ability estimates (see Figure 4) that corresponded to the most deeply underestimated discrimination parameters (see Figures 2, A1, and A3). However, the strength of the correlations varied across conditions. For example, with the true priors, the correlations ranged from -0.04 to -0.84 across slope mean, slope SD, data fit, and ability estimation method conditions, except for one condition with a positive correlation of 0.04 (see Figure 5). From Figure 5, we observed that (a) the correlation strength decreased when slope SDs increased; (b) in general, WLE and PST had the strongest and weakest correlations, respectively, and the gaps increased when the slope means decreased; (c) the two data misfit conditions had stronger correlations than the no-misfit condition. The correlations with no prior were very similar to those with true priors.

The alternative priors on slopes help increase the discrimination estimates; the magnitudes of enlargements increase when the slope means and SDs decrease because, with small slope means or SDs, the true prior density curves concentrate more on low ability scores or are very steep, while the alternative priors have flatter density curves (see Figure 1). Thus, the alternative priors helped lift the discrimination estimates and made them closer to the true ones in some conditions, for example, (a) for WLE and PST in the low- and medium-SD conditions (except for WLE in the high-mean, low-SD, and no-misfit or difficulty-misfit conditions); (b) for MAP in the discrimination-misfit and low- or medium-SD conditions (except for the low-mean, low-SD, and discrimination-misfit condition). In the high-SD slope condition, the alternative priors did not have a significant impact and performed similarly to the true priors and no prior. In the other conditions, the alternative priors lifted the discrimination estimates too much, deviating them more from the true ones than the true priors and no prior ones.

**Figure 4.** Comparisons of Mean Biases of SDs of Ability Estimates

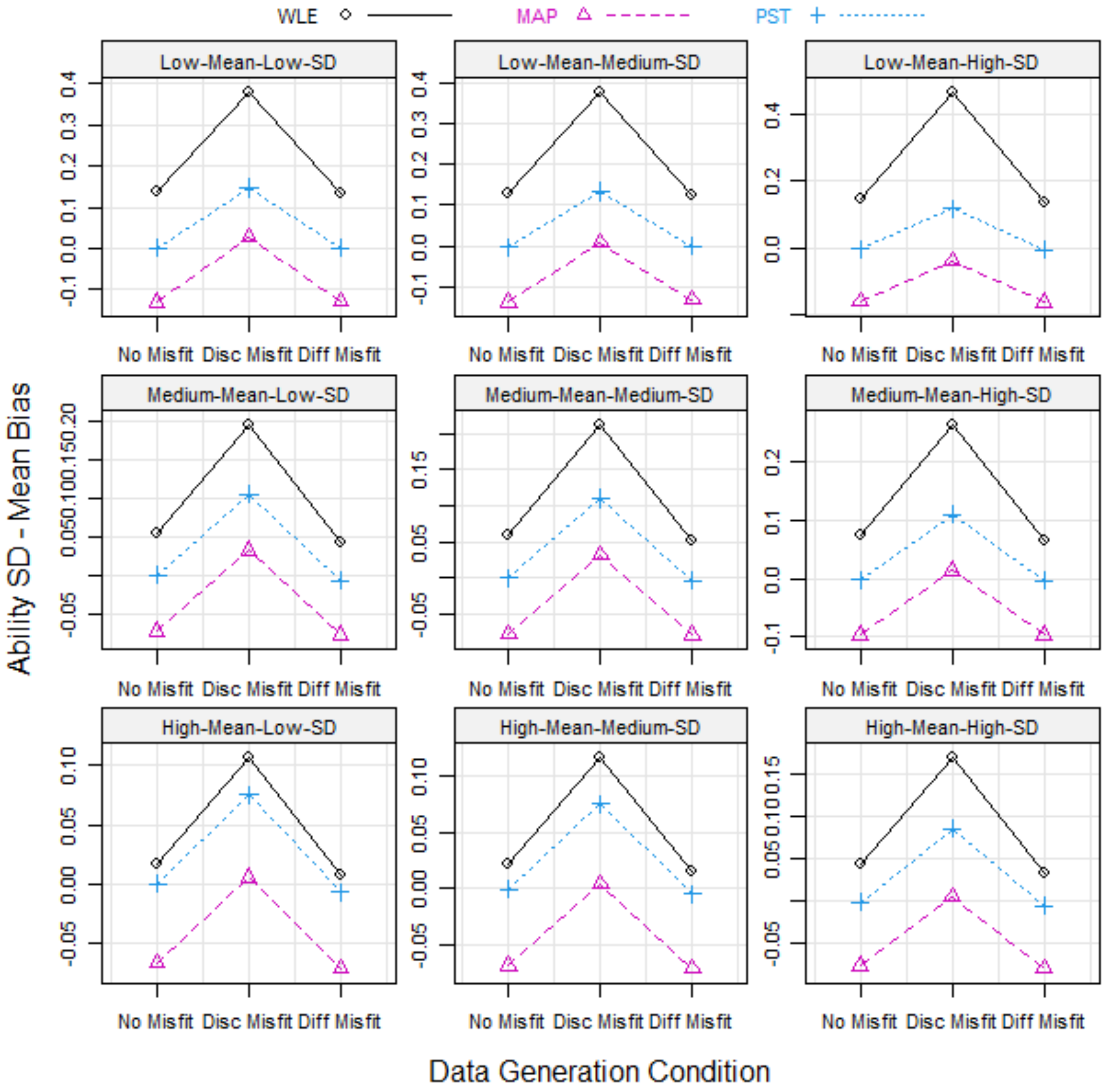
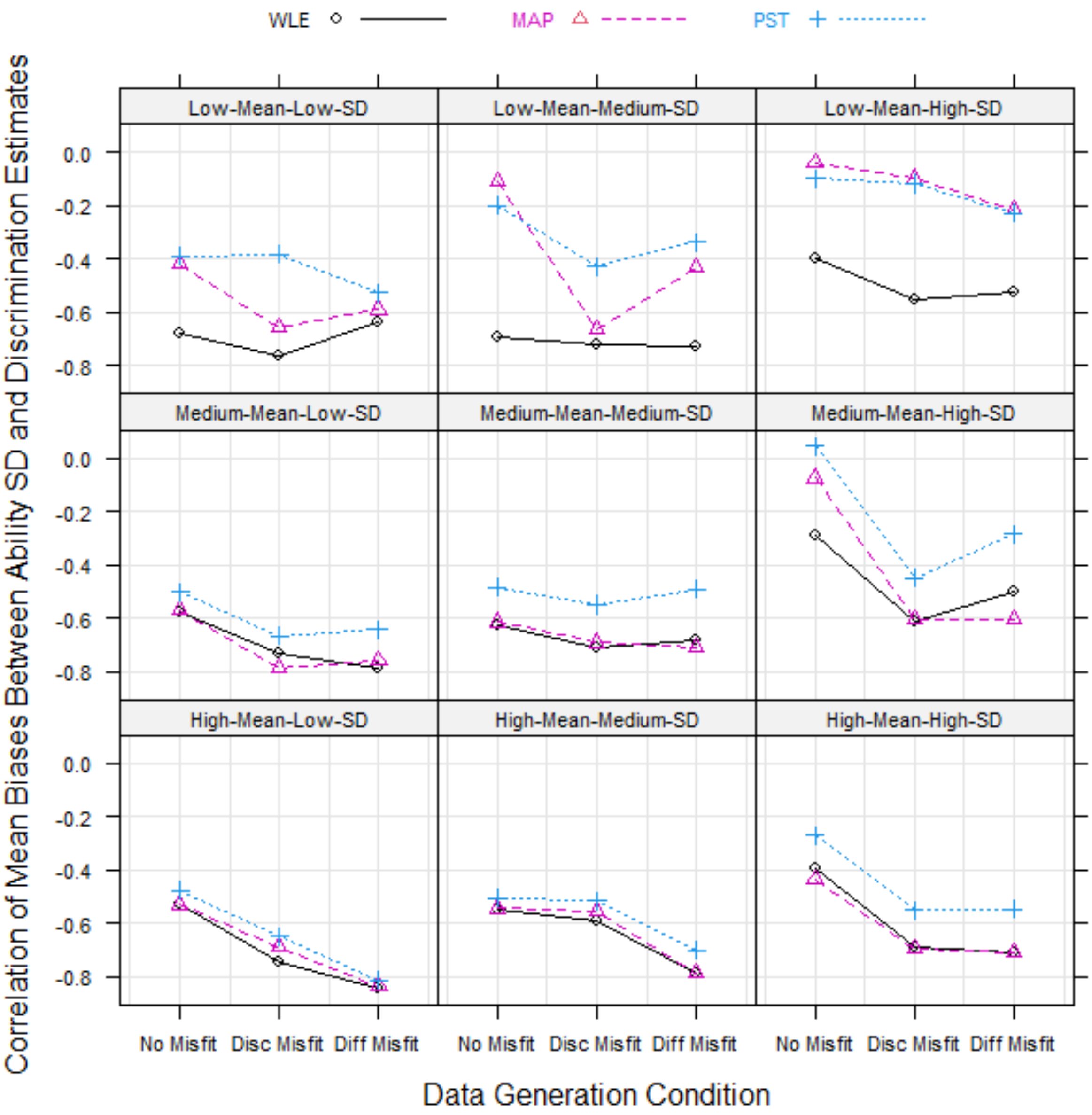


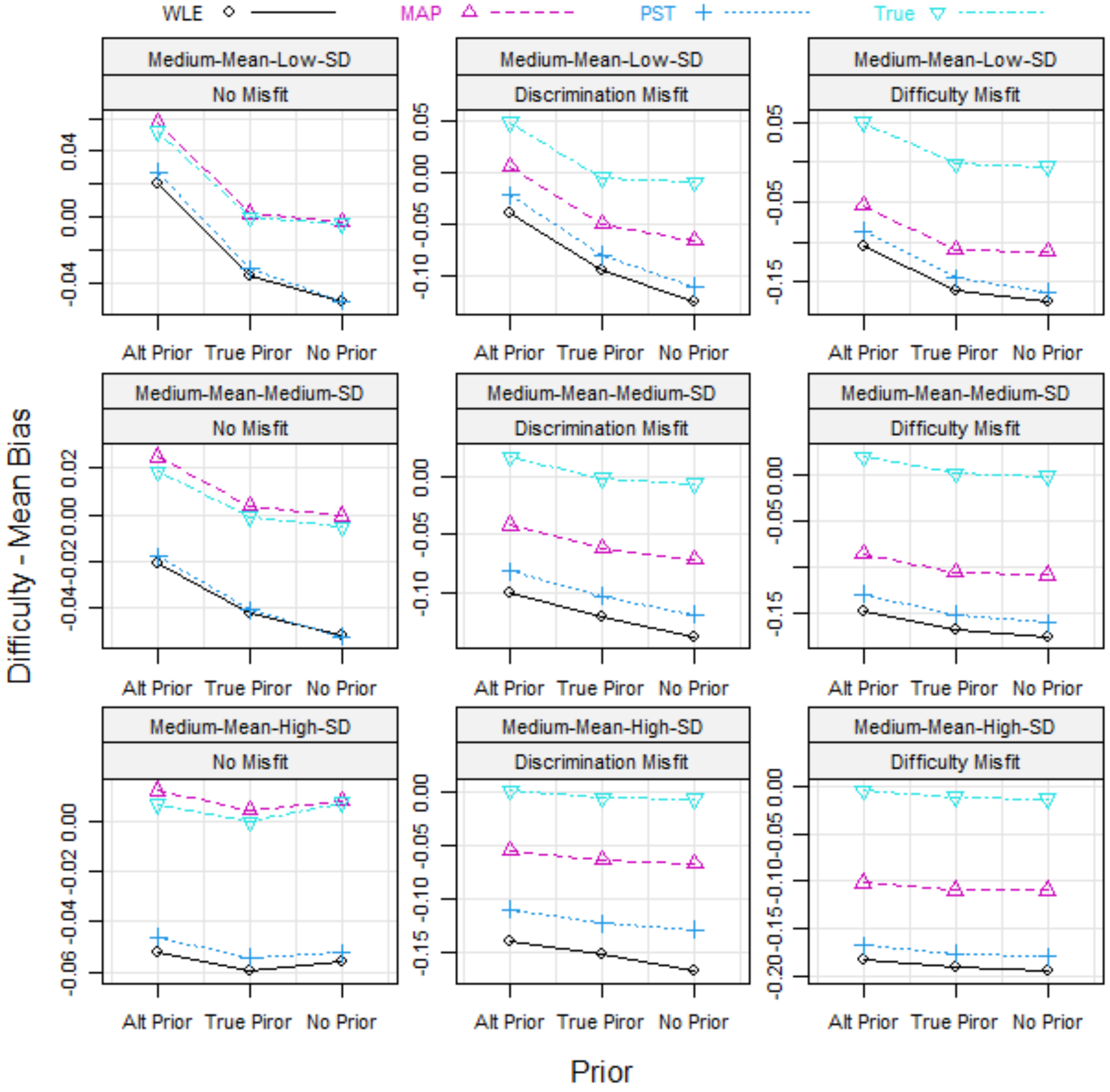
**Figure 5.** Comparisons of Correlations of Mean Biases Between Ability SD and Discrimination Estimates with True Priors



***Difficulty and Ability Estimates.*** The difficulty parameters with the true values beyond the range between -4 and 4 were removed from the analyses in this section. Extreme difficulty values often have extreme mean biases and RMSEs that distort the results and make the patterns in a plot difficult to observe. Extreme difficulty values appeared more often in the low-mean condition — the average numbers of items with extreme difficulty parameters per dataset were 8, 17, and 36 for the low-, medium-, and high-SD conditions, respectively — and the medium-mean and high-SD condition with the average of 9 items; for all the other conditions, the average number of items with extreme difficulty parameters per dataset was below 1.

Figures 6 and 7 compare the mean biases and RMSEs, respectively, of difficulty estimates in the medium-mean slope conditions. Based on the two multipanel plots, some patterns are obvious.

**Figure 6.** Medium-Mean Slope Condition: Comparisons of Mean Biases of Difficulty Parameters



1. The true ability scores led to the best difficulty estimation, followed by MAP, across all conditions except in the low-SD and alternative-prior conditions, where PST and MLE performed better. PST performed better or closely to WLE in all conditions.

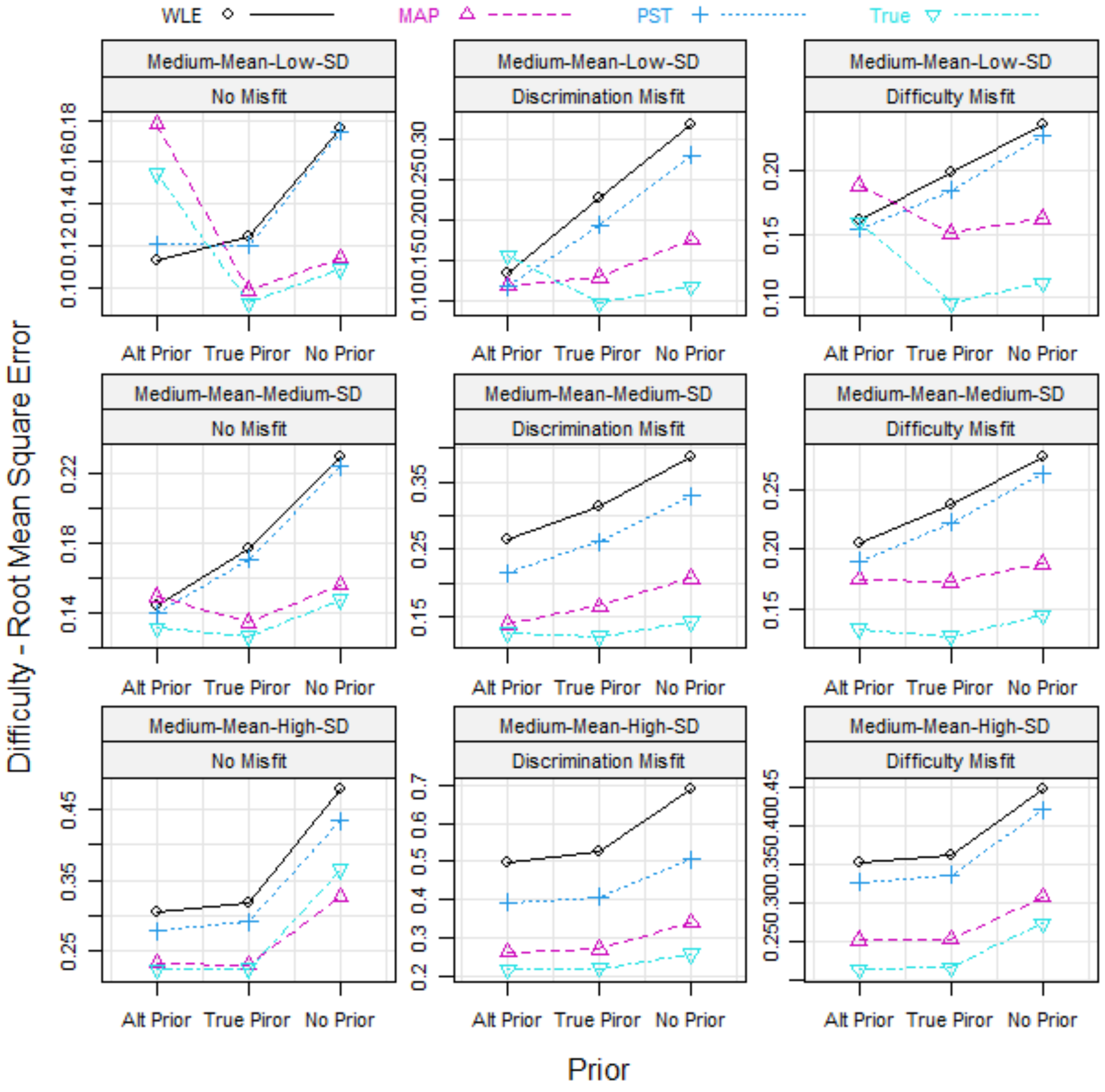
2. For all types of ability scores, the true priors performed better than no prior in all conditions. For WLE and PST, the alternative priors performed better than the true priors across all conditions. For MAP, the true priors performed better than the alternative priors in the low-SD and no-misfit or difficulty-misfit condition or the medium-SD and no-misfit condition; in all the other conditions, the alternative priors performed similarly or better than the true priors. For the true ability scores, the true priors performed better than the alternative priors in the low-SD conditions and similarly in the medium- and high-SD conditions.

3. For the true ability scores, the alternative priors led to overestimated difficulty parameters in the low- or medium-SD conditions; in all the other conditions, the mean biases of difficulty estimates

were close to 0. MAP performed similarly to the true ability scores in the no-misfit conditions and underestimated difficulty parameters in the two misfit conditions. For WLE and PST, the difficulty parameters were underestimated in all conditions except for the low-SD, no-misfit, and alternative-prior condition. In terms of the magnitude of underestimation, among the three ability estimation methods across all conditions, WLE had the largest, PST was close to WLE in most of the conditions, and MAP had the smallest; among the three prior conditions, the order from the low to high was the alternative priors, true priors, and no prior; among the three data misfit conditions, the order from the low to high was no misfit, discrimination misfit, and difficulty misfit.

For the low- and high-mean slope conditions, the patterns were generally similar to those for the medium-mean slope conditions; we also provide the figures for these two groups in the Appendix. One exception exists in the high-mean and high-SD condition: the true priors performed the best, while the

**Figure 7.** Medium-Mean Slope Condition: Comparisons of RMSEs of Difficulty Parameters



exception exists in the high-mean and high-SD condition: the true priors performed the best, while the alternative priors and no prior performed similarly. In this condition, the alternative priors were so flat that they did not differ much from no prior.

Difficulty estimations are closely related to ability estimations. Theoretically, ability estimation errors are positively correlated with difficulty estimation errors. Figure 8 compares the mean biases of ability estimates across all conditions in multiple panels. The patterns were consistent across all panels. The three ability estimation methods generally had similar mean biases of ability estimates. For the no-misfit conditions, the mean biases were close to 0; for the discrimination-misfit conditions, the mean biases were close to 0 or slightly differed from 0; and for the difficulty-misfit conditions, the ability scores were deeply underestimated due to the reduced difficulty parameters of the first 20 operational items during the data generations. This result is consistent with the magnitudes of underestimating the difficulty parameters by the three data fit conditions that were presented previously.

**Figure 8.** Comparisons of Mean Biases of Ability Estimates

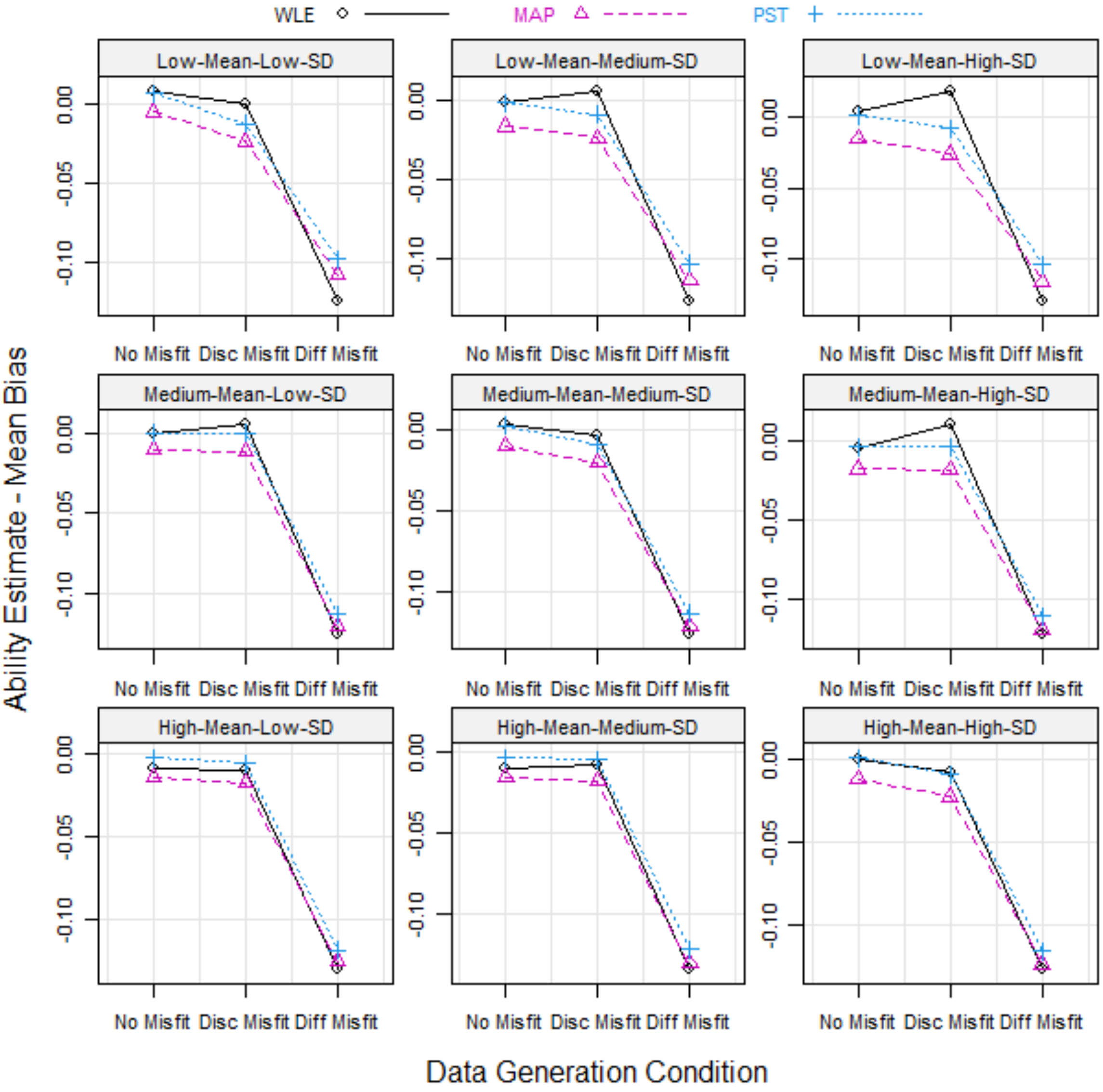
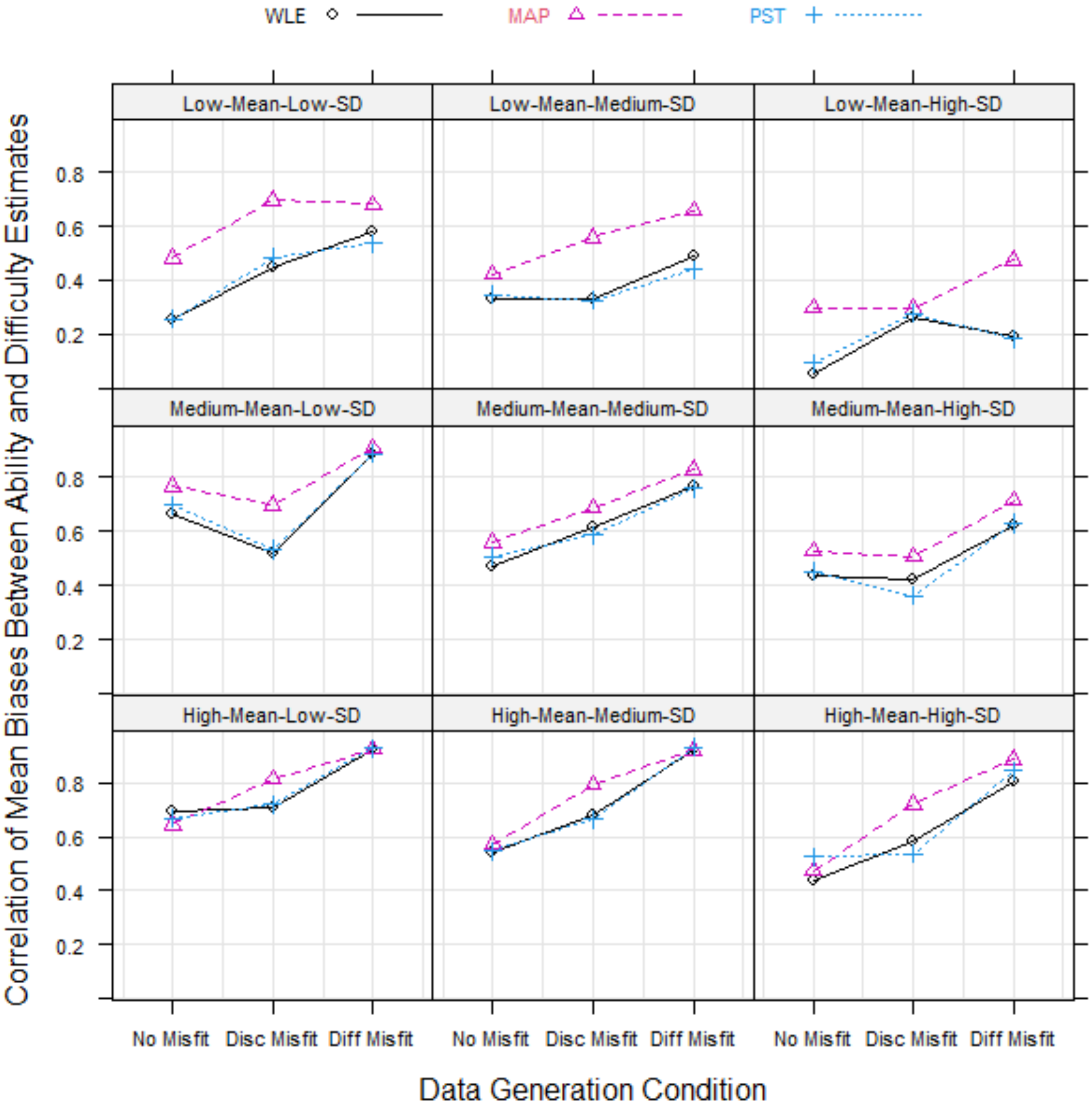


Figure 9 compares the correlations of mean biases between ability and difficulty estimates across all conditions with true priors. The correlations are all positive, ranging from 0.06 to 0.93. MAP had higher correlations than WLE and PST, which had similar correlations. The difficulty-misfit condition had the highest correlations among the three data fit conditions. The correlations increased with the increasing slope means and decreasing slope SDs. For no prior, the correlations were lower than those with true priors to various degrees across simulation conditions.

**Figure 9.** Comparisons of Correlations of Mean Biases Between Ability and Difficulty Estimates with True Priors



## Real Data Study

We used the 2PL model with fixed ability estimates to estimate the item parameters of the field test items in an operational dataset. The operational test was a two-stage adaptive test. Each assessment unit group (AUG) had one Stage 1 assessment unit (AU) and three Stage 2 AUs. Each AU contained 20 dichotomous items. A test taker assigned an AUG took Stage 1 AU first and then, based on their score on Stage 1 AU, was assigned to one of the easy, medium, and hard Stage 2 AUs. Thus, an operational AUG contained three complete test forms, each with 40 items. The real dataset included 64 operational AUGs (192 operational forms) and six field test forms, each with 20 dichotomous items.

The operational AUGs and field test forms were randomly assigned to 8,527 test takers, so each test taker took only one operational form and one field test form. Five field test items were dropped due to content issues. Among the 115 field test items, the sample size for each item ranged from 1,376 to 2,850. Each test taker's ability was estimated based on their operational items with fixed item parameters from the item bank. Like the simulation study, three ability estimation methods were applied: WLE, MAP, and PST. All ability estimations converged. Then, the 115 field test items were calibrated by the 2PL model with fixed ability estimates under the three prior conditions: alternative prior on slopes, true prior derived from the item bank, and no prior. The prior conditions were the same as those in the medium-mean and medium-SD slope condition in the simulation study. All field test item estimations converged.

In the simulation study, the estimations based on MAP ability estimates and true item priors appeared to perform the best. Thus, in the real data study, we used the estimation in the MAP-true prior condition as the baseline to compare the estimations from the other conditions. Table 3 lists the mean differences (MDs) and root mean square differences (RMSDs) of discrimination and difficulty estimates with respect to the MAP-true prior condition for each estimation condition. Based on the MDs of the discrimination estimates, (a) WLE and PST had similar means while MAP had the highest, and (b) the alternative prior generated a little higher means than the other two priors that led to the same means. This pattern is similar to the simulation study's medium-mean, medium-SD, and discrimination-misfit condition. The MDs of the difficulty estimates show that the means of the difficulty estimates were almost the same for all types of ability estimates and priors. This result is consistent with the means and SDs of the ability estimates of the three types: all means were almost the same (-0.14 or -0.15), while MAP had a lower SD (0.94) than WLE (1.04) and PST (1.00).

**Table 3.** Real Data: Comparisons of Mean Differences and Root Mean Square Differences of Item Parameter Estimates with Those Based on MAP Ability Estimates and True Item Priors

| Prior | Discrimination | | | Difficulty | | |
|---|---|---|---|---|---|---|
| | WLE | MAP | PST | WLE | MAP | PST |
| Mean difference | | | | | | |
| Alternative prior | -0.04 | 0.02 | -0.05 | 0.00 | 0.00 | 0.00 |
| True prior | -0.05 | 0 | -0.07 | 0.00 | 0 | 0.00 |
| No prior | -0.05 | 0.00 | -0.07 | -0.01 | -0.01 | -0.01 |
| Root mean square difference | | | | | | |
| Alternative prior | 0.04 | 0.02 | 0.06 | 0.06 | 0.03 | 0.06 |
| True prior | 0.06 | 0 | 0.08 | 0.09 | 0 | 0.09 |
| No prior | 0.05 | 0.01 | 0.07 | 0.19 | 0.12 | 0.19 |

The RMSDs on discrimination parameters show a similar pattern as the MDs; however, the RMSDs on difficulty parameters indicate that there were much more differences at the individual item level among the estimation conditions than those indicated by the MDs at the aggregate level. For example, for the MAP-no prior condition, the MD on difficulty was -0.01, while the RMSD was 0.12. Thus, no estimation in the other conditions was similar to that in the MAP-true prior condition.

In the real study, we cannot judge which estimation method is best because we do not know the true item parameters of the field test items. We may, however, evaluate whether the estimations from the different conditions have meaningful differences in terms of external criteria. For example, we may check the impact of different item parameter estimates on test takers' ability score estimates, such as score differences, ranking order, and cut scores. We did not conduct such evaluations in the current study because these field test items needed further inspections before adding them into the item bank and including them in operational forms in test assembly. (We also limited the number of new items in an operational form.) Nevertheless, the findings from a simulation study, especially one mimicking the real conditions, can provide guidance on selecting a preferable estimation method.

## Discussion

### Summary of Main Findings

The main questions of the current study are, in the 2PL item parameter estimation with fixed ability estimates, (a) which prior on slopes should be used (alternative prior, true prior, or no prior), and (b) which ability estimation method should be chosen (MLE, MAP, or PST). We conducted a simulation study to investigate the questions under various conditions. We also compared the item parameter estimates on a real dataset using different ability estimation methods and priors on item parameters. We have the following main findings.

1. In general, MAP ability estimates with true priors performed best except when the operational items used to estimate ability scores had item drift on discrimination parameters and the slopes of the field test items had medium SDs or low SDs (excluding the low-mean and low-SD condition).
2. In general, WLE performed better or similarly to PST. For WLE, the alternative priors were the best choices unless the slopes of the field test items had a high mean and SD.
3. PST performed better than WLE on difficulty estimates to various degrees in the discrimination-misfit conditions. However, in general, PST was not a better choice than WLE. Considering the variances of individuals' ability estimates did not seem beneficial in this application. In addition, as an estimation of a test taker's ability distribution, PST is hard for measurement practitioners to understand.
4. In general, the mean biases of the discrimination estimates were negatively correlated with the mean biases of the SDs of the ability estimates, and the mean biases of the difficulty estimates were positively correlated with the mean biases of the ability estimates. However, most of the correlations were low to medium. For example, in the medium-mean and medium-SD condition, the mean biases of the SDs of WLE, MAP, and PST estimates were 0.06, -0.08, and 0.00, respectively, while their mean biases on discrimination estimates with true priors were - 0.08, -0.01, and -0.10, respectively; the mean biases of WLE, MAP, and PST estimates were 0.00, -0.01, and 0.00, respectively, while their mean biases on difficulty estimates with true priors were -0.04, 0.01, and -0.04, respectively.
5. In the low- and medium-SD conditions, alternative priors increased the discrimination estimates compared to the true priors. In contrast, with the true priors, WLE and PST had underestimated discrimination estimates in all conditions, and MAP had them in the discrimination-misfit conditions. Thus, under these conditions, the alternative priors were generally better choices than the true priors and no prior because they lifted the discrimination estimates closely to the true ones. The alternative priors also led to better or similar difficulty estimates for WLE and PST under all conditions and for MAP under the high-SD or discrimination-misfit conditions.

### Practical Recommendations

Based on the simulation study, we have general recommendations for measurement practitioners when applying the 2PL model estimation with fixed ability estimates in equating.

The first choice is the MAP ability estimation and empirical lognormal prior on slopes and normal prior on intercept derived from a historical dataset (e.g., item bank). In general, MAP ability estimates with correct priors on item parameters lead to the most accurate item parameter estimations in all conditions.

One may argue that an issue with the MAP ability estimation is that the prior distribution of ability scores may be unknown. For a testing program using the 2PL model with the MML-EM or MCMC estimation, its scale (i.e., the population distribution of ability scores) is defined when it starts because the 2PL model estimation requires it. For example, a testing program commonly assumes the population distribution of ability scores as the standard normal. However, for a particular sample of test takers who take field test forms, their ability distribution may be unknown and different from the population distribution, especially if the sample size is small. Nevertheless, this is not an issue per se because the testing program's scale has been fixed by the operational items' parameters in the bank, and it is correct to use the program's scale as the prior distribution to estimate MAP ability scores regardless of the distribution of the field test sample (see Ho, 2024, for empirical results).

If a testing program selects the WLE ability estimation, then couple it with a less informative prior on slopes than the empirical one from historical data, like the alternative priors used in this study. This study has shown that for a typical testing program (i.e., with the medium mean and SD of slope parameters), a less informative prior on slopes led to a better item parameter estimation than the informative true prior when using the WLE ability estimation.

Finally, when should we apply the linking method with fixed ability estimates rather than the traditional TCC method? A basic consideration is that if we worry that the anchor items in the TCC method cannot be calibrated well due to, for example, the small sample size issue, then the fixed ability linking method is a valuable option. By fixing anchor items' parameters to their bank values, we preserve the original scale of the testing program. The drawback is that it does not account for possible item drifts on anchor items. That is why we investigated the impact of data fit on the fixed ability linking method in our study. The TCC method alleviates the impact of anchor item drifts by removing drift items from the anchor set. However, if the calibration of anchor items is not stable, then the theoretical principle of the TCC method collapses.

### Limitations and Future Research

We have demonstrated that the alternative priors improve the 2PL item parameter estimations in the low- and medium-SD slope conditions for WLE and PST (and for MAP in the discrimination-misfit conditions), and the reason is under these conditions, the alternative priors increase the discrimination estimates compared to the true priors. In contrast, the true priors underestimate the discrimination estimates. However, we cannot quantify the increase of discrimination estimates by the alternative priors and the underestimations by true priors in a more mathematical way. The same is true for the correlations of estimation biases between discrimination parameters and SDs of ability scores and between difficulty parameters and ability scores. It is helpful to study further these issues from the theoretical and empirical perspectives.

The alternative priors on slopes in this study were chosen by accident: The alternative prior was mistakenly treated as the true prior in the equating study we conducted using the WLE ability estimates. After applying the true prior, we found that the alternative prior performed better than the true prior. If the underlying mathematical process is more understood, better alternative priors may be proposed.

Finally, we focus on the 2PL model in this study. For the other two unidimensional IRT models commonly used in practice, the Rasch model and the three-parameter logistic (3PL) model, although their model complexity and sample size requirements differ (Yen & Fitzpatrick, 2006), we think the findings from the current study also apply to the two models as long as a population distribution of ability is predefined in the model estimations, because the rationales underlying these findings apply to them. However, a testing program using the Rasch model may employ conditional likelihood estimation (CLE) and does not assume a population distribution of ability. Sometimes, a small testing program's low testing volume makes it hard to estimate a population distribution of ability. It causes difficulty in defining the prior ability distribution for estimating MAP scores. Even if the prior ability distribution can be correctly defined, it is unsure if the current findings apply to this case because the population distribution of ability is not used in item parameter estimation. Thus, further research is needed to extend the current study to the Rasch and 3PL models to verify whether the current findings are still valid, especially for the Rasch model with CLE, where both correct and incorrect prior ability distributions can be explored.

## Conclusion

The linking method with fixed ability estimates was proposed long ago with solid statistical ground (Stocking, 1988). It is a valid alternative to the popular TCC method when the sample sizes for anchor items are too small to validate the TCC method. Our current study showed that, for estimating the 2PL model with fixed ability estimates, the MAP ability estimation with true prior on slopes performed the best, followed by the WLE ability estimation with less informative prior on slopes. In practice, we recommend using the MAP ability estimation and empirical priors on slopes and intercepts derived from a historical dataset (e.g., item bank) for this fixed ability linking method with the 2PL model. If a testing program prefers the WLE ability estimation, use the alternative less informative prior on slopes than the empirical one from historical data.

**Received:** 12/27/2023. **Accepted:** 1/13/2025. **Published:** 1/22/2025.

**Citation:** Fu, J., Ho, T., & Tan, X. (2025). Item parameter estimation of the 2PL IRT model with fixed ability estimates: Choices of ability estimation methods and priors on slopes. *Practical Assessment, Research, & Evaluation*, 30(1). Available online: https://doi.org/10.7275/pare.2104

**Corresponding Author:** Jianbin Fu, Educational Testing Service. Email: jfu@ets.org

## References

Atkinson, K. E. (1991). *An introduction to numerical analysis* (2nd ed.). John Wiley & Sons.

Baker, F. B., & Kim, S. H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). CRC Press. https://doi.org/10.1201/9781482276725

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores.* Addison-Wesley Pub. Co.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 443–459. https://doi.org/10.1007/BF02293801

Broyden, C. G. (1970). The convergence of a class of double-rank minimization algorithms 1. general considerations. *IMA Journal of Applied Mathematics, 6*, 76–90. https://doi.org/10.1093/imamat/6.1.76

Chalmers, R., P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1–29. https://doi.org/1.18637/jss.v048.i06

Chang, M. (2017). *A comparison of two MCMC algorithms for estimating the 2PL IRT models* [Unpublished doctoral dissertation]. Southern Illinois University Carbondale.

Fox, J. (2010). *Bayesian item response modeling: Theory and applications*. Springer. https://doi.org/10.1007/978-1-4419-0742-4

Fu, J. (2019). *Maximum marginal likelihood estimation with an expectation-maximization algorithm for multigroup/mixture multidimensional item response theory models*. (ETS Research Report No. 19-35). Educational Testing Service. https://doi.org/10.1002/ets2.12272

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (1996). Introducing Markov Chain Monte Carlo. In W. R. Gilks, S. Richardson & D. J. Spiegelhalter (Eds.), *Markov Chain Monte Carlo in practice*. (pp. 1–20). Chapman and Hall. https://doi.org/10.1201/b14835

Haberman, S. J. (2009). *Linking parameter estimates derived from an item response model through separate calibrations* (ETS Research Report No. 09-39). Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2009.tb02197.x

Ho, T. (2023). Bayesian logistic regression: A new method to calibrate pretest items in multistage adaptive testing. *Applied Measurement in Education, 36*(4), 355–371. https://doi.org/10.1080/08957347.2023.2274572

Ho, T. (2024). *Calibrating pretest items when responses are sparse in multistage adaptive testing* [Manuscript submitted for publication]. Educational Testing Service.

Kim, S., Cohen, A. S., Baker, F. B., Subkoviak, M. J., & Leonard, T. (1994). An investigation of hierarchical Bayes procedures in item response theory. *Psychometrika, 59*(3), 405–421. https://doi.org/10.1007/BF02296133

Levy, R. (2009). The rise of Markov chain Monte Carlo estimation for psychometric modeling. *Journal of Probability and Statistics,* 1–18. https://doi.org/10.1155/2009/537139

Lockwood, J.R., & McCaffrey, D.F. (2017). Simulation-extrapolation with latent heteroskedastic error variance. *Psychometrika, 82*(3), 717–736. https://doi.org/10.1007/s11336-017-9556-y

Marcoulides, K. M. (2018). Careful with those priors: A note on Bayesian estimation in two-parameter logistic item response theory models. *Measurement: Interdisciplinary Research and Perspectives, 16*(2), 92–99. https://doi.org/10.1080/15366367.2018.1437305

Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika, 51*(2), 177–195. https://doi.org/10.1007/BF02293979

Nocedal, J., & Wright, S. (2006). *Numerical optimization*. Springer.

R Core Team (2022). *R: A language and environment for statistical computing* (Version 4.2.1) [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/.

Rupp, A. A., Dey, D. K., & Zumbo, B. D. (2004). To Bayes or not to Bayes, from whether to when: Applications of Bayesian methodology to modeling. *Structural Equation Modeling, 11*(3), 424–451. https://doi.org/10.1207/s15328007sem1103_7

Stocking, M. L. (1988). *Scale drift in online calibration*. (ETS Research Report. No. RR-88-28-ONR). Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1988.tb00284.x

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*(2), 201–210. https://doi.org/10.1177/014662168300700208

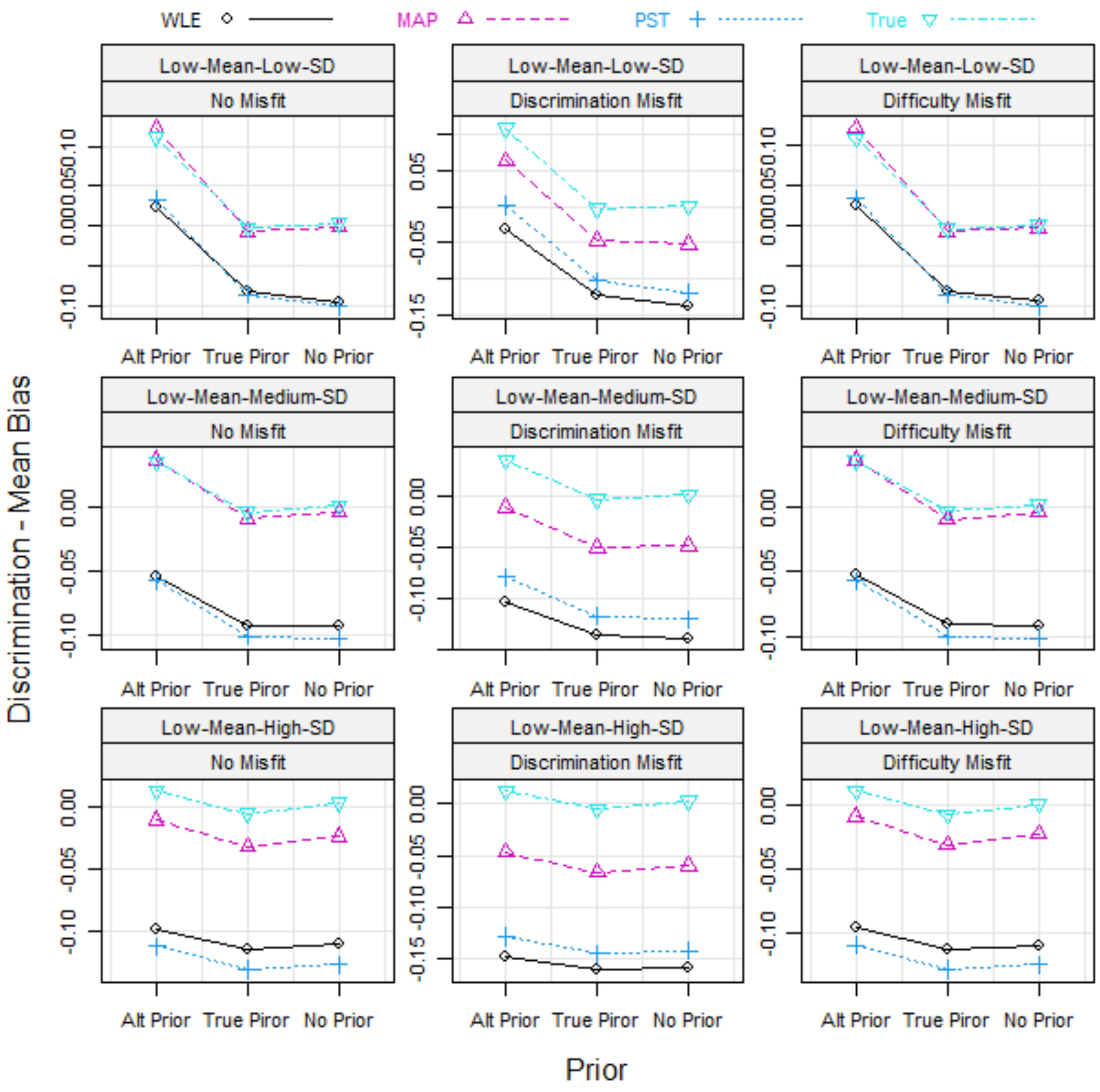
Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussion). *The Annals of Statistics, 22*(4), 1701–1762. https://doi.org/10.1214/aos/1176325750

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*(3), 427–450. https://doi.org/10.1007/BF02294627

Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 111–153). Rowman & Littlefield Publishers.

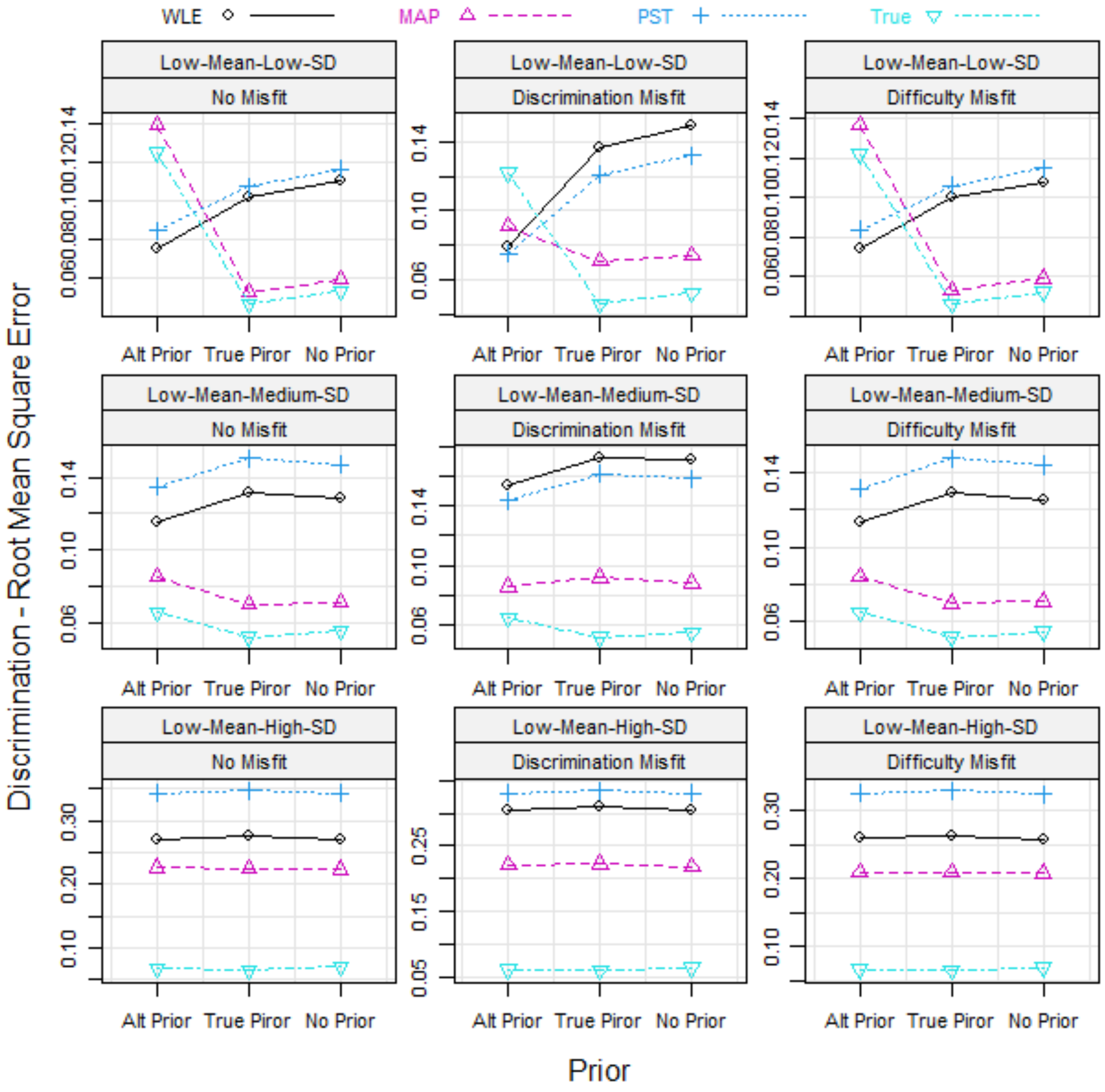
## Appendix A. Plots of Simulation Results in the Low- and High-Mean Slope Conditions

**Figure A1.** Low-Mean Slope Condition: Comparisons of Mean Biases of Discrimination Parameters



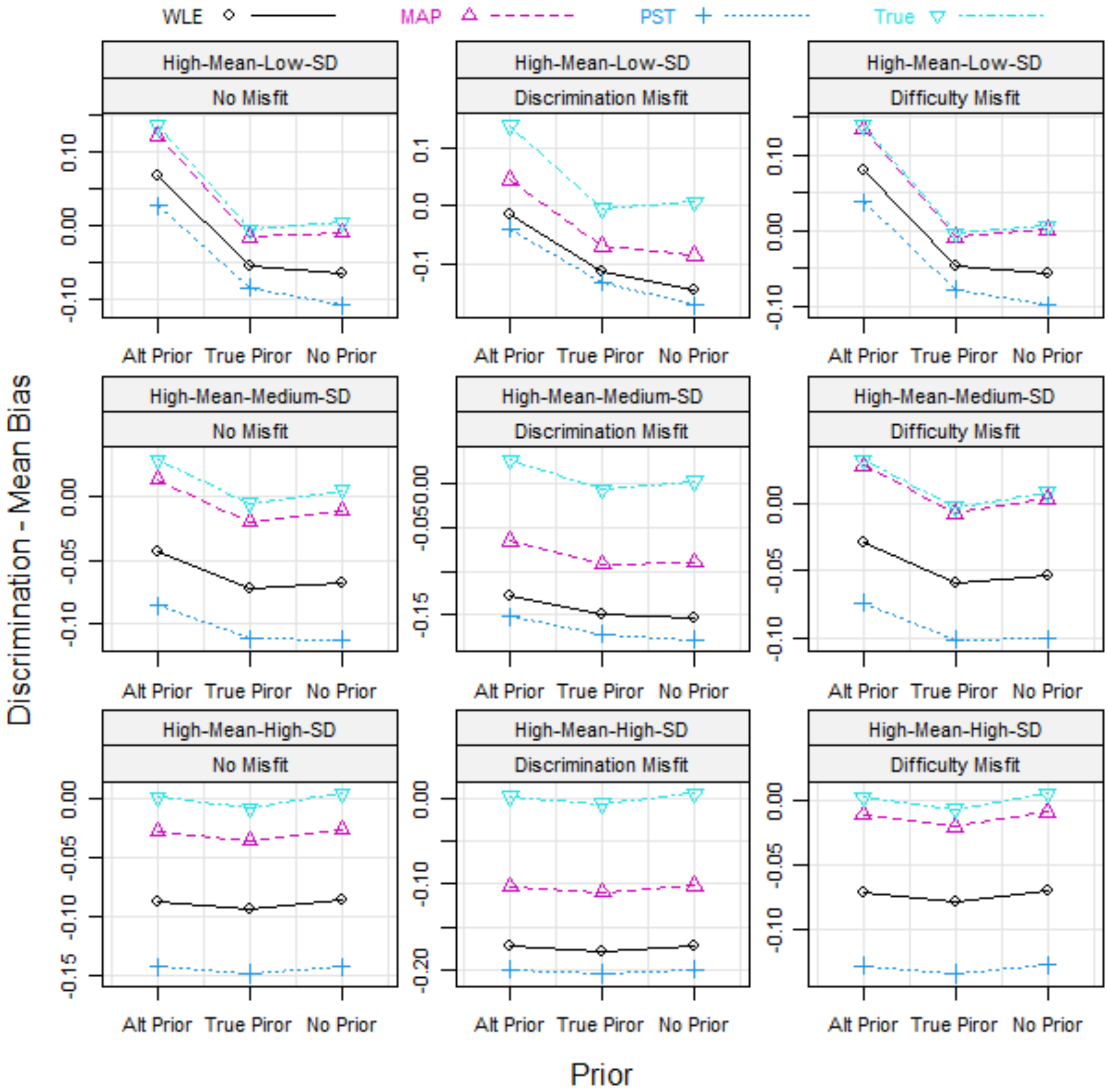
*Note.* Alt Prior = Alternative Prior.

**Figure A2.** Low-Mean Slope Condition: Comparisons of RMSEs of Discrimination Parameters



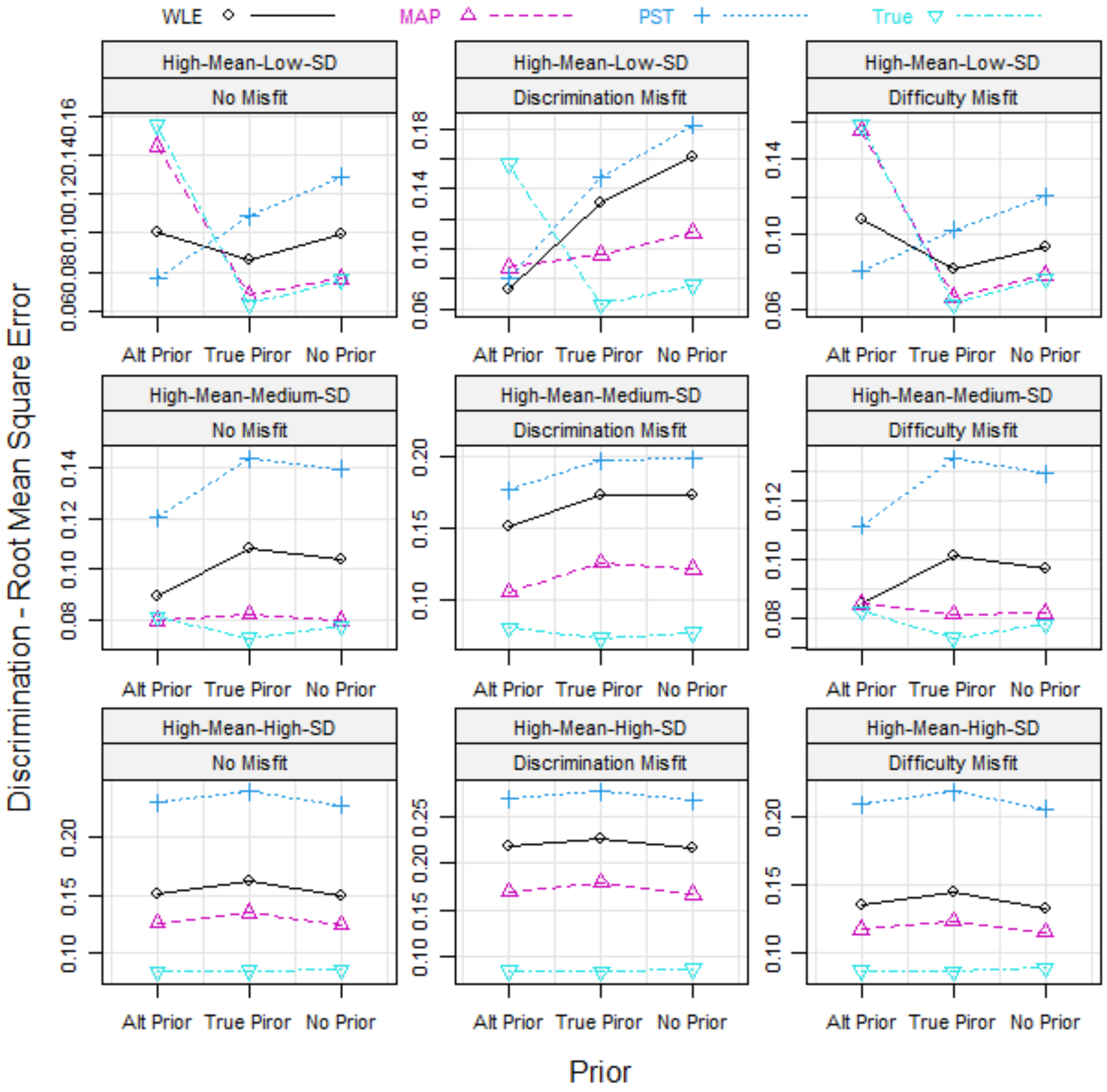
*Note.* Alt Prior = Alternative Prior.

**Figure A3.** High-Mean Slope Condition: Comparisons of Mean Biases of Discrimination Parameters



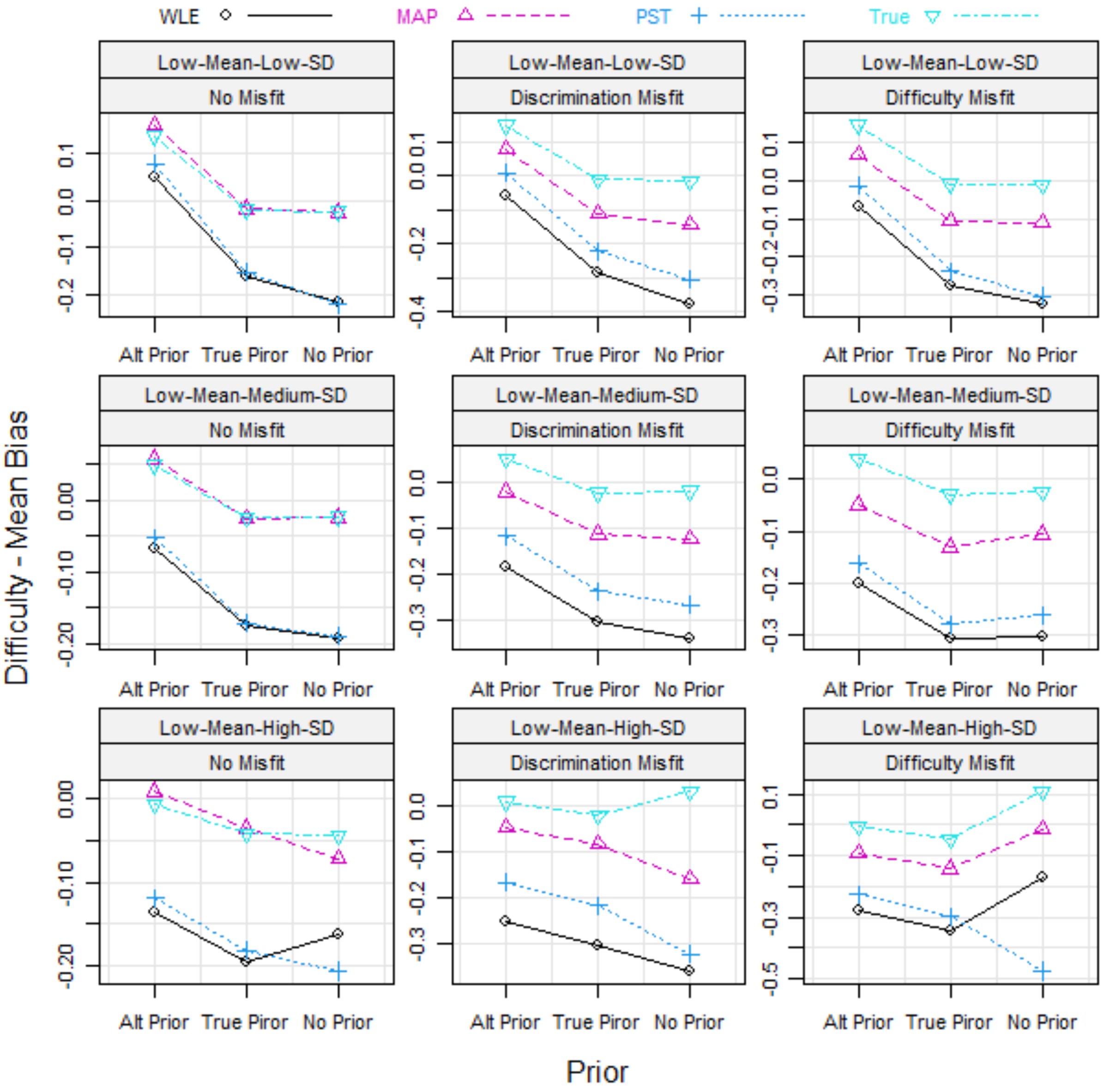
*Note.* Alt Prior = Alternative Prior.

**Figure A4.** High-Mean Slope Condition: Comparisons of RMSEs of Discrimination Parameters



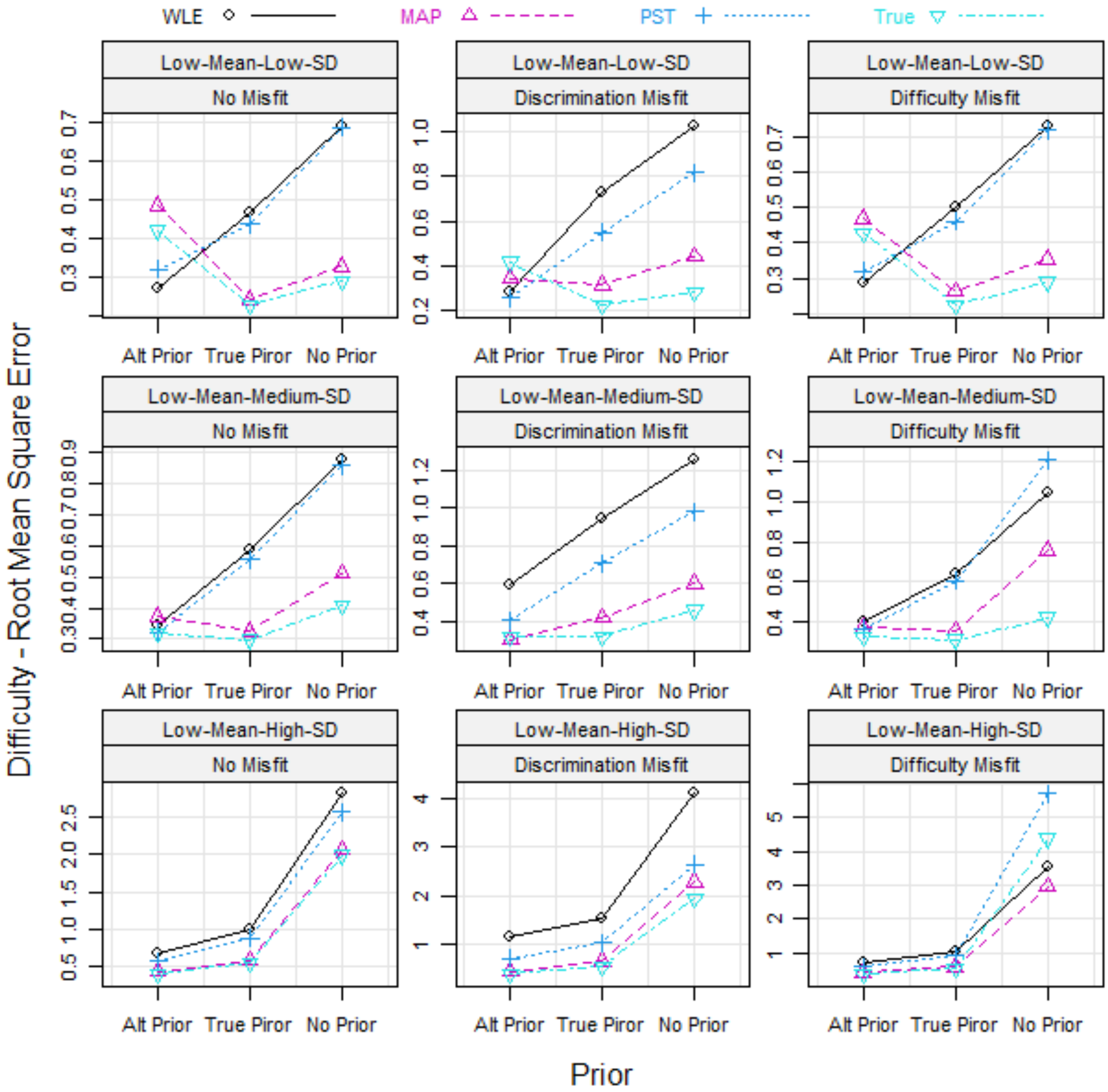
*Note.* Alt Prior = Alternative Prior.

**Figure A5.** Low-Mean Slope Condition: Comparisons of Mean Biases of Difficulty Parameters



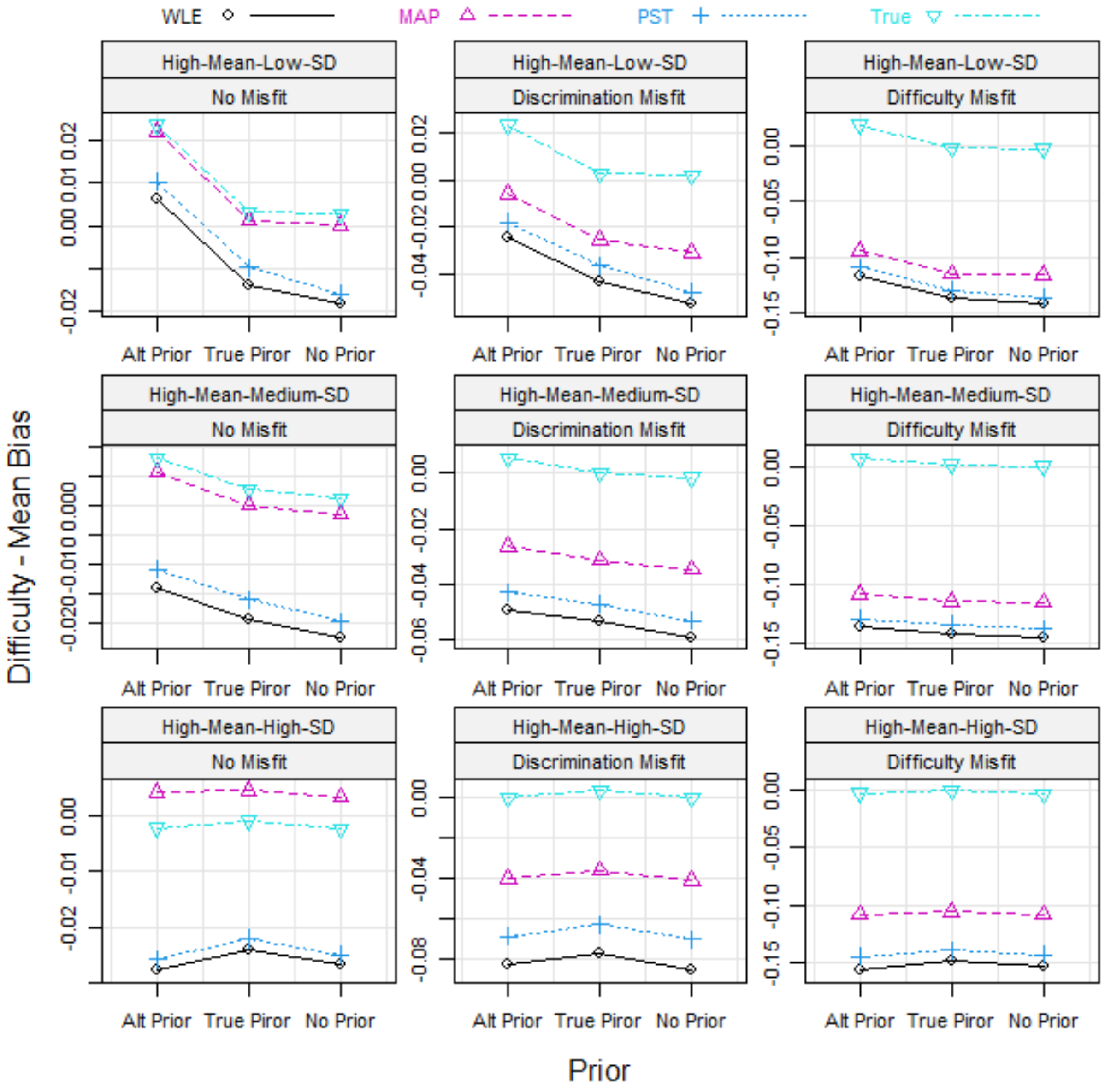
*Note.* Alt Prior = Alternative Prior.

**Figure A6.** Low-Mean Slope Condition: Comparisons of RMSEs of Difficulty Parameters



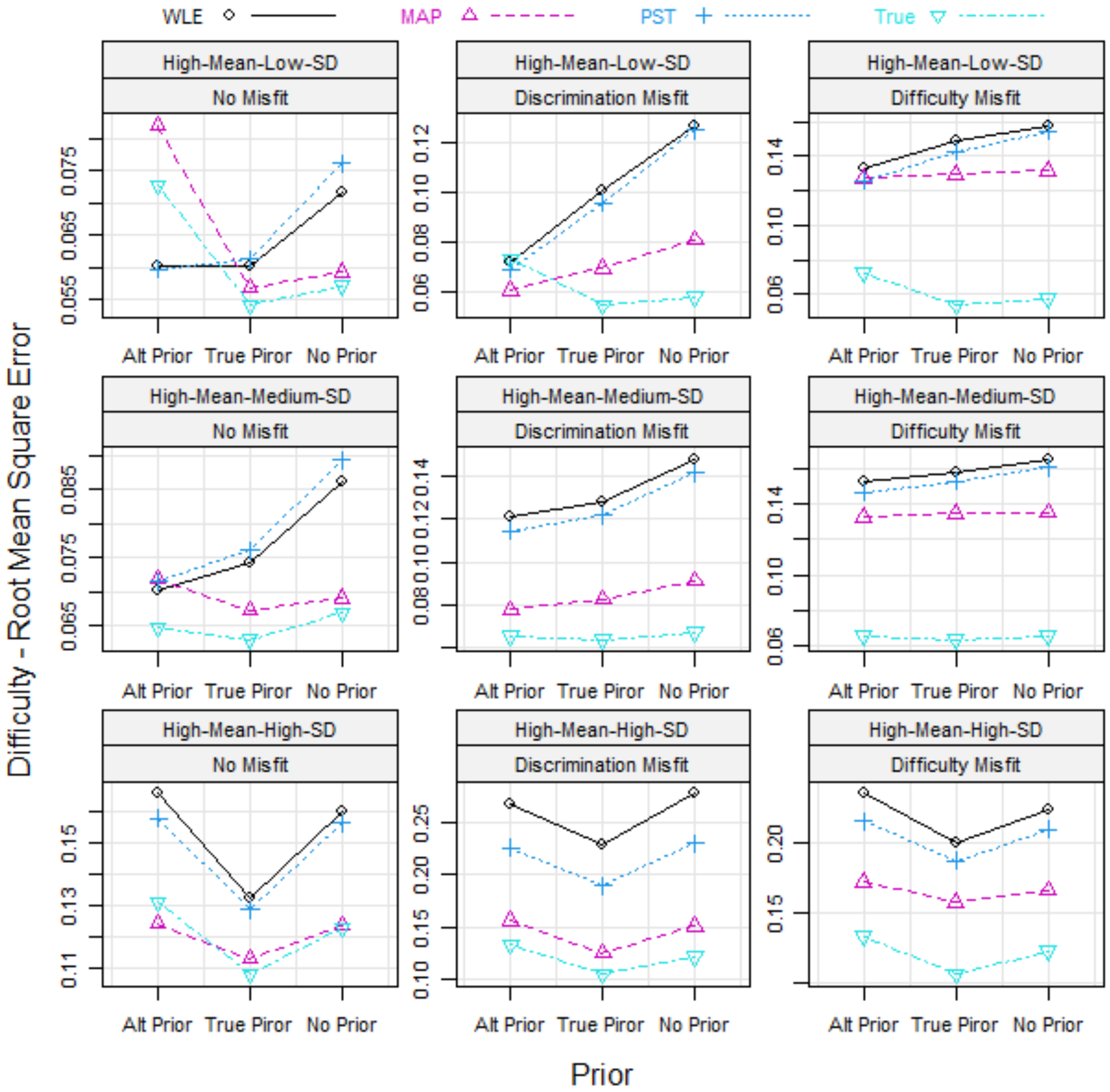
*Note.* Alt Prior = Alternative Prior.

**Figure A7.** High-Mean Slope Condition: Comparison of Mean Biases of Difficulty Parameters



*Note.* Alt Prior = Alternative Prior.

**Figure A8.** High-Mean Slope Condition: Comparisons of RMSEs of Difficulty Parameters



*Note.* Alt Prior = Alternative Prior.