10-2021

Item Parameter Recovery With and Without the Use of Priors

Paulius Satkus

Christine E. DeMars
demarsce@jmu.edu

Recommended Citation

Satkus, P., & DeMars, C. (2021, October). Item Parameter Recovery With and Without The Use Of Priors [Paper presentation]. Northeastern Educational Research Association 52nd Annual Meeting, virtual.

Item Parameter Recovery With and Without the Use of Priors

Paulius Satkus

Christine E. DeMars

Department of Graduate Psychology and Center for Assessment and Research Studies,

James Madison University

Satkus, P., & DeMars, C. (2021, October). *Item Parameter Recovery With and Without The Use Of Priors* [Paper presentation]. Northeastern Educational Research Association 52nd Annual Meeting, virtual.

Item Parameter Recovery With and Without the Use of Priors

Introduction

Estimating item parameters for the 3-parameter item response theory (IRT) model can be difficult (e.g., Lord 1968, Appendix B; Thissen & Wainer, 1982). Using Bayesian priors on the likelihood functions can reduce estimation problems and can increase the accuracy of item parameter estimates (Harwell & Janosky, 1991; Mislevy, 1986, Swaminathan & Gifford, 1986). In other areas of Bayesian statistics, choosing the prior distribution is a much-debated issue because ill-matched priors can severely bias the items estimates. Ideally, the prior distribution would have high density near the parameter value, because the estimate will be biased toward regions of higher density. If the mean of the prior is not near the true parameter value and the prior distribution has a relatively small variance (highly informative prior), the estimate may be seriously biased unless there is enough information in the data to overcome the information in the prior. In IRT marginal maximum likelihood (MML) estimation, a typical approach is to apply priors to the item discriminations and to the lower asymptotes to avoid implausible estimates. To gain these advantages without biasing the estimates too much, IRT analysts often use relatively diffuse priors (Lord, 1986). For estimating item parameters by MML, it has long been suggested that the specific priors are not that important as long as they are not highly informative. Unlike other areas of Bayesian statistics, within the MML literature there has been little exploration of choosing the specific priors for analyzing a given dataset. The purpose of this study was to assess the sensitivity of item parameter estimation in the 3PL model to different prior distributions. More specifically, we varied the appropriateness (i.e., the mean) and the informativeness (i.e., the variance) of the prior distributions for item parameter estimation.

Priors can be utilized in joint maximum likelihood (JML), marginal maximum likelihood (MML), or fully Bayesian methods such as Monte Carlo Markov Chains (MCMC; Sheng, 2010). When priors are used in MML, the procedure is sometimes called marginal Bayesian estimation (MBE) and the resulting item parameter estimates are sometimes called Bayes model estimates (BME). The focus of this study is the use of priors in MML, but some of the literature from Bayesian JML or MCMC may generalize.

The 3-parameter-logistic (3PL) model is defined as:

$$P(\theta) = c_i + (1 - c_i)\frac{e^{a_i(\theta - b_i)}}{1+e^{a_i(\theta - b_i)}}, \quad (1)$$

where P(θ) is the probability of correctly responding to an item given person's ability level (θ), a_i is the item discrimination, b_i is the item difficulty, and c_i is a lower asymptote/pseudo guessing parameter. In the 2PL model the lower asymptote is fixed to zero. A constant D = 1.7 may be added before the a_i; this will make the *a*-parameters smaller so that they are virtually identical to those from a normal ogive (probit) model. When judging how informative a prior is, it is important to know whether the researcher was using a logistic or normal (D = 1.7) metric, because a variance of 1 in the logistic metric is equivalent to a variance of 0.588^2 in the normal metric (conversely, a variance of 1 in the normal metric is equivalent to a variance of 1.7^2 in the logistic metric). Similarly, judging the magnitude of the bias and SE or RMSE of the estimates of the *a*-parameter depends on the metric.

One important issue is that estimation may not converge or may result in implausible parameter estimates or extremely large estimates of the standard errors. The 3PL model often runs into these problems (Lord, 1975; 1986; Mislevy, 1986; Swaminathan & Gifford, 1986), although such problems are less frequent with the 2PL model. Thissen and Wainer (1982) noted

that "estimation of a lower asymptote can wreak havoc with the accuracy of a location parameter" (p. 398). Applying reasonable priors often reduces the number of estimation problems with the 3PL model. For example, Gao and Chen (2005) reported that a small proportion of replications did not converge without priors, and they reported a number of *a* or *c* estimates that were unreasonably; with priors, there were no such problems. Convergence also required fewer iterations with priors; with N = 100, the average number of iterations to reach their convergence criterion was 52 without priors but 15 with priors. The informativeness of the priors may greatly influence the occurrence of convergence or estimation problems. Sheng (2010) showed that far fewer iterations were needed for MCMC convergence with more informative priors, which may be true for MML as well.

Within the MML literature for 2PL models, Lim and Drasgow (1990) compared not using priors to using relatively diffuse priors, variance = 2^2 for *b*-parameter and variance = 1 for ln(*a*).[1] With a sample size of 250, the SE was smaller with priors, but with sample sizes of 750, priors had little impact on the SE except for items with large *a*'s and extreme *b*'s. A study by Harwell and Janosky (1991) focused on the effect of variance (i.e., the informativeness of the priors) on item discrimination parameters, with the variance of ln(*a*) ranging from 0.1^2 to 0.75^2.[2] The authors found that with small examinee sample sizes the more informative the priors were, the smaller the RMSE of the item estimates. This relationship held for both *a* and *b*, although the priors were only applied to the *a*-parameters, and was more pronounced in small test length

[1] With a log-normal (0,1) distribution, 95% of the density falls between 0.14 and 7.10 after exponentiating. These authors were using the normal metric, so 95% range would be 0.24 to 12.07 in the logistic metric.
[2] After transforming back to the metric of the *a*-parameters, 95% of the density falls between 0.82 and 1.22 for a log-normal (0, 0.1^2) distribution or between 0.23 and 4.35 for a log-normal (0, 0.75^2).

conditions (i.e., 15 vs. 25 item tests). With a sample size of 250 or more and 25 items, the most informative prior yielded slightly higher RMSE than the more diffuse priors or no prior. Presumably, the more informative prior led to greater bias for some items which outweighed the decrease in SE, although bias was not reported separately. When the sample sizes were greater than 250 examinees, the authors concluded that the effect of prior informativeness was reduced. The authors explained that with greater sample sizes and longer tests, the likelihood functions draw enough information from the data and thus the priors are no longer necessary for item estimation. In other words, the final estimates are produced primarily from the data and not from the prior distribution. The prior neither harms (biases) or helps (reduction in SE) if there is enough information in the data.

Zeng (1997) compared three sets of priors for the 3PL model using MML, but did not include a condition without priors. Zeng used 4-parameter beta distributions for all three parameters, once with the parameters of the prior fixed and again with the mean of the prior updated after each cycle to match the mean of the item parameter estimates. The third condition also updated the mean of the prior, but used a lognormal distribution for the *a*-prior, a normal distribution for the *b*-prior and a 2-parameter beta distribution for the *c*-prior. Zeng did not vary the spread of the prior distributions; the priors were moderately informative and the sample size was also moderate, 500 or 1000. When the item parameters were centered near the center of the initial prior density, RMSEs were slightly smaller when the mean of the prior was fixed, presumably because estimating the mean added more random error than the systematic error caused by the mismatch between prior mean and true parameter. Otherwise, the RMSEs were generally somewhat smaller when the means were estimated. Overall, with these moderately

informative priors and moderate sample sizes, it did not make a great difference which priors were used.

Gao and Chen (2005) also used 4-parameter beta distributions for all three parameters in MML estimation and compared these priors to MML without priors. They used three sets of 4-parameter beta distributions, one set with a mode well-matched to the true item parameters, and two other sets centered above or below the mode of the true parameters. The priors were moderately informative, with standard deviations of 0.46-0.49 for the *a*-parameters after transforming to the logistic metric (reported as 0.27-0.29 in the normal metric), 0.87-0.90 for the *b*-parameters, and 0.04 for the *c*-parameters. Priors increased the correlation between true and estimated parameters and decreased the RMSE, particularly for the smallest sample (N = 100), for which using the wrong prior often yielded lower RMSE than no prior. For larger samples (N = 500 or 1000), the benefits of priors were smaller, but they still made a noticeable decrease in RMSE. For item sets with mostly moderate discrimination and moderate difficulty, the center of the prior made little difference, except for the *c*-parameters or the smallest sample. For a set of items that were easy and not very discriminating, the correlations were smaller and the RMSEs were larger than for the other datasets. Estimation for this set of items was better when the prior was a better match for the true values, but a mismatched prior still yielded more accurate estimates than no prior.

Both Zeng (1997) and Gao and Chen (2005) used priors on all parameters. A limited small-scale study compared priors on only the *c* to priors on both *a* and *c* for MML (DeMars, 2019, footnote 5). For a small sample of 400, bias and RMSE for the *a*-parameters were considerably lower with priors on both *a* and *c* compared to priors on only *c*.

In addition to these studies using MML estimation, a study by Sheng (2010) used MCMC estimation, more specifically the Gibbs sampler. Sheng's results replicate the findings found in Harwell and Janovsky's (1991) and in Gao and Chen's (2005) studies in that larger sample sizes are necessary to reduce the effect of the priors on item estimates. However, the recommendations differ in the suggested number of examinees. With the 2PL model, Harwell and Janovsky (1991) recommended sample sizes greater than 250, whereas with the 3PL model Sheng recommended sample sizes greater than 1000, and Gao and Chen noted that for samples of 7500 or more there was no difference between priors and no priors for the 3PL model. These recommendations refer to the accuracy (bias and/or RMSE) of the item estimates.

Sheng also showed that using non-informative or informative but ill-matched priors had less impact on the accuracy of estimates for 2-parameter models than 3-parameter models. Marcoulides (2018) showed that ill-matched informative priors **could** have a large impact on 2PL item parameter estimates using MCMC with a small sample (N = 150) and short test, but the bad informative priors were quite far from most of the true item parameters. After transforming back to the metric of the *a*-parameters, 95% of the density fell between 1.13 and 6.53, but the mean of the true *a*-parameters was 0.63. Consequently, *a*-parameters were positively biased. Similarly, the *b*-parameters, with a prior mean of −2 but a true mean of 0.9, were quite negatively biased. Not using priors is not an option for MCMC, but diffuse parameters yielded less bias than the poorly matched informative priors and comparable bias to informative priors centered closer to the means of the true parameters. Sheng (2010) suggested choosing prior distributions such that most of the prior's density is within a plausible range, effectively excluding extreme values. Thus, the key message, whether MCMC or MML estimation, is to specify a prior that is moderately informative and contains the feasible item parameter values.

Such values should be "sufficiently mild to affect most item parameters minimally when the data supply information about them, but keep all parameters within a 'reasonable' range" (Mislevy, 1986, p. 190) and "not too vague and at the same time not too precise" (Swaminathan & Gifford, 1986, p. 597).

Ideally, moderately-informative priors would provide enough information to avoid estimation problems while not biasing estimates very much. Particularly when the mean of the prior is not well-matched to an item's true parameter, the prior distribution needs to be diffuse enough not to seriously bias the estimate, yet informative enough to prevent unrealistic estimates. Thus, two key aspects of the prior distribution were of interest in this study: the match of the prior to true parameters, and the spread of the prior density. Although two other studies (Gao & Chan, 2005; Zeng, 1997) have explored multiple specifications of priors in MML estimation of the 3PL model, both used moderately informative priors and did not modify the spread. We will examine the impact of different variances, similar to Harwell and Janosky's (1991) study of the 2PL model. Further, both 3PL studies changed the match of the prior by changing the mean of the prior; moving the mean of the prior further from the mean of the true parameters would make the prior a worse match on average, but it would be a better match for a few items. In this study, we instead varied the match of the prior by examining the accuracy of recovering each item parameter, some well-matched and some not well-matched. In some contexts, such as equating or adaptive testing, the items are not viewed as exchangeable and the bias and RMSE of each, not just the average across items, is important.

The following research questions in the context of estimating 3PL item parameters using MML estimation were explored in this study:

1) How do priors impact estimation problems, including standard errors that can not be computed and unreasonable values for the item parameter estimates? If these problems seldom occur without priors, there may be little need to use priors. However, if these problems occur without priors, it is expected that using priors will reduce the frequency of problems.
2) How sensitive are the item parameter estimates to the mean and variance of the prior? Clearly, item parameters mismatched to the mean of the prior will be biased toward the prior, and the degree of bias will increase as the prior becomes more informative. However, the standard error of the estimates will decrease as the prior becomes more informative. The purpose was to explore the magnitude of these effects and whether a balance can be achieved.
3) How does sample size impact Research Questions 1 and 2? More estimation problems are expected with small samples, which would indicate priors are needed more with small samples than large samples. However, a given prior distribution will be more informative for a small sample, because the informativeness of the prior is relative to the information in the data. Thus, if the prior is informative enough to mitigate estimation problems, it may also be informative enough to create non-negligible bias, making it difficult to find a balance.

Method

To answer the research questions, we conducted a simulated study with a hypothetical 45-item test. Three levels of item discrimination (1.02, 1.53, 2.04) were crossed by five levels of item difficulty (−2, −1, 0, 1, 2) by three levels of lower asymptote (.05, .15, .25). The test in this study was longer than simulated tests in some other studies (e.g., Harwell & Janosky, 1991,

Sheng, 2010), but resemble tests often used in educational assessment settings. The values of the item parameters were fixed across replications so that we could assess if or how the difficulty, discrimination, and guessing were related to the accuracy of parameter recovery.

We manipulated sample size, use of priors, variance of prior (i.e., informativeness), and the match between the mean of the prior and the true parameter (appropriateness). Sample size had two levels (100 and 500). Use of priors had three levels: none, priors on *c*'s, priors on both *a*'s and *c*'s. Priors were not applied to the *b*'s because Mislevy (1986) noted they were generally better determined by the data than the *a*'s and *c*'s. Harwell and Janosky (1991) noted that priors are generally not needed for the *b*'s, Gao and Chen (2005) suggested a non-informative prior would generally be acceptable for the *b*'s, and Kim, Cohen, Baker, Subkoviak, and Leonard (1994) found that adding a prior to the *b*'s made little difference in RMSE, bias, or correlation with true parameters. Lord (1986) also suggested priors for the *a*'s and *c*'s but not for the *b*'s. The use of priors only on the *c*'s was prompted by Harwell and Janosky's finding that, with moderately large samples (500 or more), priors on the *a*'s made little difference in the 2PL model. This result motivated the idea that it might be adequate to apply priors only to the *c*-parameters in the 3PL model with moderately large samples. When priors were used (for the latter two conditions), the prior for *a* was normal, mean = 0, and the prior for *c* was logit-normal, mean = −1.73. The variance of the prior for *a* had three levels (0.3, 1.0, 1.5) as did the variance of the prior for the logit of *c* (0.46, 0.59, 0.72). The match of the prior mean was manipulated within each test form by using the same prior mean for all items. For one third of the items, the prior mean was too high, for one third it was too low, and for one-third it was well matched.

For each sample size, data were drawn for 500 replications. θ was distributed ~N(0,1), and was sampled for each replication. Probability of correct response for each response was calculated as a function of the item parameters and θ and compared to a random draw from U[0,1]; if the probability was greater than the random draw, the response was coded correct. The simulated data was then used for item parameter estimation in each of the various use of prior and variance of prior conditions. The *mirt* package (Chalmers, 2012) in R was used for model estimation.

Results

Research Question 1

Research Question 1 was assessed by counting the proportion of replications in each condition which had a) no reported standard errors; b) at least one *a*-parameter estimate less than -2 or greater than 7, c) at least one *b*-parameter estimate less than -6 or greater than 6, or d) at least one *c*-parameter estimate greater than .7. These problems were noted at the replication (within condition) level because this represents estimating data from one test administration. With real data, if the analysts noticed problems with any item, it is likely they would then start tweaking the estimation, either by discarding items or applying priors to the likelihood. The absence of standard errors in the *mirt* package indicates that the information matrix during the estimation could not be inverted. Implausibly high discrimination, or discrimination estimates less than zero when the item-total correlation was positive, indicate an anomaly that most analysts would notice and investigate. Similarly, very large or small item difficulties, or very large lower asymptotes would clue the analyst that there were estimation problems. The information relevant to Research Question 1 is presented in Table 1. Not specifying priors led to

the highest frequency of replications that contained at least one issue. When the sample size was equal to 500, 60.8% of the replications had at least one problem. This number increased to 100% when the sample size was equal to 100. Thus, in small or moderate sample sizes, not specifying priors is vital. Applying priors only on the *c* parameter helped estimation for the larger sample size condition (n= 500) but not for the smaller sample size condition as about 60% of replications still had at least one of the problems. Specifying priors for both *a* and *c* parameters reduced the number of estimation issues even further for both sample size conditions. For the larger sample size condition, having priors on both *a* and *c* parameters resulted in almost no estimation problems. However, that was not the case for the smaller sample size condition as the percent of replications with problems still ranged from 16.2% to 39.8% even when both priors were applied. As the prior on *a* became less informative (i.e., the variance became larger), more estimation problems were observed. As the prior on *c* became less informative, more estimation problems were observed as well; however, the effect was smaller. Interestingly, most of the estimation problems in the conditions where priors were applied (either on *c* or on both *a* and *c*) were extreme *b* values. This shows the interrelatedness of the parameter estimates; although the true parameters were independent, the estimation errors were not (Mislevy, 1986).

Research Question 2

Parameter Estimates

To address Research Question 2, bias across 500 replications was calculated. Before calculating bias, estimates for an item (within a single replication) were removed if the estimated *a* was negative or > 7, if the estimated *b* was < −6 or >6, or if the estimated *c* was > .7. Since a substantial proportion of replications were plagued by estimation issues when no priors were specified or priors were used only on *c* in the small sample (Table 1), the bias for those

conditions is not presented. To assess the relative importance of the factors, the variance was partitioned and η^2 was computed.

Priors Applied Only on the Pseudo-guessing Parameter (*c*). Figure 1 presents bias for all item parameters when priors only on *c* were specified and sample size was equal to 500. The mean of the prior on *c* was set to be equal to .15, which leads to the prior being appropriate for one-third of the items (items with true *c* values of .15). For the other two-thirds of the items (items with true *c* values of .05 and .25), the prior was not appropriate and thus bias in opposite directions was expected. The bias in *a* estimates ranged from about -0.16 to .68 (top panel of Figure 1), and depended on how well *c* matched the prior (η^2 = .51) and *b* (η^2 = .26) and their interaction (η^2 = .10). The *a*-parameter was positively biased when *c* was below the mean of its prior distribution but negatively biased when *c* was above the mean of its prior distribution. This effect was accentuated for difficult items. Items that were easier (difficulties of -2 and -1) had almost no bias, whereas items that were more difficult (difficulties of 1 and 2) had noticeable bias. To a small extent, the strength of the prior on *c* had the expected effect as well (η^2 = .02 each for *c* x *c*-prior-variance and *b* x *c* x *c*-prior-variance). The less informative the prior on *c* was (e.g., variance of the prior on the logit of *c* increasing from .46 to .59 to 72), the less biased the *a* parameter estimates were (for items with high-*b* combined with high or low *c*). The pattern was similar for the mean bias in *b* and *c* in that the bias was positive when the mean of the *c*-prior was too high and negative when the mean of the prior was too low (middle and bottom panels of Figure 1). However, the variance of the prior for *c* made little difference in the bias for *b* or *c*. Item difficulty parameters were most biased when the prior on *c* was not matched well, especially for the easier items (η^2 = .03 for main effect of *b* and .30 for *b* x *c*). This contrasted

with the bias in discrimination, which was greatest for difficult items. Only when the true difficulties were equal to 2 were the estimated *b* parameters unbiased, regardless of the true *c* values. More discriminating items had bias closer to zero for *b* (η^2 = .11 for *a* x *c*).

For *c*, the main effect of how well the *c* matched the prior explained 84% of the variance. Item pseudo-guessing parameters were the most biased (in expected directions) when the items were easy (for *b* x *c*, η^2 = .12), regardless of the strength of the prior on *c*. The items with true *c* values of .05 were almost always overestimated to be .15 for the easiest items, which lead to the bias being .10. Conversely, the easy items with true *c* values of .25 were almost always underestimated to be .15, which lead to the bias being -.10. Thus, the appropriateness of the prior had expected effects, while the strength of the prior did not seem to affect the accuracy of *c* parameters (the main effect and interactions involving the strength of the prior each accounted for < 1% of the variance in mean difference).

Priors Applied on Item Discrimination (a) and the Pseudo-guessing Parameters (c). Figure 2 presents the bias for *a* parameters when priors were applied on *a* and *c* parameters and sample size was equal to 500 (see Appendix for N = 100). The variance in the bias was partitioned with a model that included sample size, but the main effects and interactions with sample size will be discussed separately—in this section, trends across both sample sizes will be discussed. The prior on *c* had a mean of .15 and three different variance levels, as before. The prior on *a* had a mean of 1.53 and three different variance levels (i.e., .3, 1, and 1.5). Thus, the prior on *a* was appropriate for one third of the items (items that had true *a* value of 1.53) and not appropriate for the other two thirds of the items (items that had true *a* values of 1.02 and 2.04). When the prior on *a* was the most informative (i.e., the variance of prior on *a* equaled .3, top

panels of Figure 2), the bias in *a* parameters was in the expected directions. That is, the most discriminating items (i.e., items with true *a* values of 2.04) were underestimated, whereas the least discriminating items (i.e., items with true *a* values of 1.02) were overestimated. This pattern replicated regardless of the informativeness of the prior on *c*. However, when the prior on *a* parameter was less informative (i.e., the variance of prior on *a* equaled to 1 or 1.5, middle and bottom panel of Figure 2), the above-mentioned pattern changed (η^2 = .19 for interaction between match of *a* to prior mean and strength of *a*-prior). Instead, the magnitude of bias depended more on the match between the *c* parameter and its prior mean (η^2 = .20 for the main effect of *c* and η^2 = .03 for interaction between *c* and variance of *a*-prior), although within each level of *c* the bias was most negative/least positive for the higher values of *a* (η^2 = .38 for the main effect of *a*). That is, items with *c* below the prior mean (.05) had overestimated *a* parameters, whereas items with levels of *c* above the prior mean (.25) had underestimated *a* parameters. Moreover, the bias increased as difficulty increased (η^2 = .04 for both *a* x *b* and *b* x *c*), especially when the prior on *c* was informative (i.e., variance of prior on the logit of *c* equaled to .46). This "funnel" effect was most pronounced when the prior for *c* was informative and dissipated as the prior on *c* became less informative.

Figure 3 presents the bias for *b* parameters when priors were applied on *a* and *c* parameters and sample size was equal to 500 (see Appendix for N = 100). The greatest magnitude of bias was observed in conditions where *c* was not well-matched to the prior mean (η^2 = .40), especially for the easiest and least-discriminating items (η^2 = .16 for b, .11 for *a* x *b*, .03 for *a* x

c, and .04 for *b* x *c*).[3] The strength of the prior on *a* also made a difference for the easiest and least-discriminating items (η^2 = .02 for *b* x prior variance, .03 for *a* x prior variance, and .06 for the 3-way interaction). The strength of the prior on *c* did not affect the bias in *b*.

Figure 4 presents the bias for *c* parameters when priors were applied on *a* and *c* parameters and sample size was equal to 500 (see Appendix for N = 100). The same pattern of bias replicated with *c* parameters as before with priors only on *c*. That is, the bias depended on the item's match to the mean of the prior of the *c* parameter (η^2 = .84). Easy items (difficulties of -2 and -1) that had true *c* below the prior mean were always overestimated, whereas easy items that had true *c* above the prior mean were always underestimated (η^2 = .09 for *b* x *c*). Thus, overall, the results confirm our hypothesis about the appropriateness of the mean of the priors. However, the effect of informativeness of the prior often depended on the item difficulties.

Standard Error Estimates

The precision of parameter estimates was evaluated by computing empirical standard error across the 500 replications (after discarding items with extreme values, as described for bias). Ideally, parameter estimates would not be biased and they would have small empirical standard error. Generally, the RMSE (the total error) followed the same pattern as the standard error; exceptions where the RMSE was influenced more by the bias will be noted. Figures with RMSE are available in the appendix. Before partitioning the variance, the standard error was log-transformed to reduce the skew.

[3] This interaction appears to be due to the value of *a*, not the match of the *a* to the prior mean; the highest *a*'s are not well-matched, but they have the smallest magnitude of bias for the *b*'s.

Priors Applied Only on the Pseudo-guessing Parameter (*c*). Figure 5 presents empirical standard errors for all item parameters when priors only on *c* were specified and sample size was equal to 500; as was done for bias, the standard error for the smaller sample was were not reported for this condition because most replications had one or more estimation problems. The standard errors for item discrimination or item difficulty remained approximately the same regardless of the strength of the prior on *c* ($\eta^2 \leq .01$ for all interactions involving the variance of the prior). The only parameter that was affected by the variance of the prior on *c* was the pseudo-guessing parameter (bottom panel in Figure 5) ($\eta^2 = .15$ for the main effect and .05 for the interaction between *b* and strength of *c*-prior). The less informative the prior on *c* was, the higher were the standard errors of *c*. However, the difference in magnitude between the standard errors from the least informative condition (i.e., where the variance of prior on *c* was equal to .72) and the most informative condition (i.e., where the variance of prior on *c* was equal to .46) was likely negligible. When this effect was combined with no effect of the strength of the prior on bias, the total effect of the prior variance on RMSE (Appendix) was very small. Larger standard errors were observed for items with higher true levels of pseudo-guessing parameter (, η^2 due to *c* = .03 for SE of *a*, .17 for SE of *b*, and .11 for SE of *c*), particularly for the hardest items (η^2 = .03, .02, .12 for the *b* x *c* interaction when the outcome was SE of *a*, SE of *b*, SE of *c*, respectively). In the case of the SE, in contrast to the bias, this effect was seemly not due to the match of the *c* to the prior mean but literally to the value of *c*, because low values of *c* (poor match) had the lowest SE.

Priors Applied on Item Discrimination (a) and the Pseudo-guessing Parameters (c).

Figure 6 presents the standard errors for *a* parameter when priors were applied on *a* and *c*

parameters and sample size was equal to 500 (N =100 available in the Appendix). The variance in the log of SE was partitioned with a model that included sample size, but the main effects and interactions with sample size will be discussed separately—in this section, trends across both sample sizes will be discussed. As expected, the smallest standard errors were observed when the prior on *a* was the most informative (i.e., the variance of prior on *a* was equal to .3) (η^2 = .74). There were interactions between the strength of the prior on a and the level of difficulty (η^2 = .05) and discrimination (η^2 = .03); the latter interaction appears to be due to the value of *a*, not the match of I to the mean of the prior. Varying the strength of the prior on *c* did not seem to affect the standard errors of item discrimination.

Figure 7 presents the standard errors for *b* parameter when priors were applied on *a* and *c* parameters and sample size was equal to 500 (N =100 available in the Appendix). Varying the information in the prior on *c* did not seem to affect the standard errors of *b*. However, varying the strength of the prior on *a* had an effect (η^2 = .03). The highest standard errors for item difficulty were observed for the conditions in which the prior on *a* was the least informative (i.e., variance of prior on *a* equal to 1.5). Lastly, items that had the highest difficulty (η^2 = .34) or highest level of *c* parameter, resulted in the highest standard errors for the *b* parameters.

Lastly, Figure 8 presents the standard errors for the *c* parameter when priors were applied on *a* and *c* parameters and sample size was equal to 500 (N =100 available in the Appendix). As expected, the more informative the prior on *c* was, the lower were the standard errors for the c parameter (η^2 = .14). Varying the information in the prior on *a* did not affect results, aside from a small interaction (η^2 = .02) between the value of *a* and the variance in the prior for *a*. The estimation of items with highest true values of *c* resulted in highest estimated *c* standard errors.

Research Question 3

To address Research Question 3, we compared the bias across the two sample size conditions. Because most of the replications for N = 100 had estimation issues if there was no prior distribution on the *a*'s, only the conditions with priors on both *a*'s and *c*'s were considered. When the priors were applied to both *a* and *c* parameters, sample size had little impact on bias. For the bias in the *a* parameter, the interaction between sample size and the match of the *a* to the prior mean explained 4% of the variance; when *a* was poorly matched to the mean of the prior and thus biased, the absolute value of the bias was greater for the smaller sample. The same pattern was observed for N = 100 and N = 500, however in the *n* = 100 condition it is clearer, because a prior with a given variance is relatively more informative with a smaller sample.

Similar conclusions can be drawn about the impact of sample size on the bias on the *b* parameter when priors on both *a* and *c* parameter were specified. The interaction between sample size and *b* explained 5% of the variance, and the interaction between sample size, *b* and the variance of the prior for *a* explained 2% of the variance —essentially, in conditions where the bias was further from zero, the absolute value of the bias was greater for the smaller sample. Thus, the difference between bias on *b* parameter across the sample size conditions is the magnitude. When the prior on *a* was most informative (i.e., the variance of prior on *a* equaled to .3), the range of mean bias in *b* for all smaller sample size conditions was almost twice as wide as the range of mean bias in *b* for all larger sample size conditions. However, when the prior on *a* was the least informative (the variance of prior on *a* equaled to 1.5), the range of mean bias was approximately the same. Thus, the effect of sample size on mean bias in *b* was somewhat moderated by the informativeness of the prior on *a*.

Lastly, different sample size conditions did not affect the bias in *c* when both priors on *a* and *c* were applied. Not only the patterns of bias were the same (primarily dictated by the true levels of *c*), but also the magnitude of bias was the same across the sample size conditions.

In general, one would expect SEs to be larger with smaller samples, but a given prior is also relatively more informative for smaller sample sizes, reducing the variance. No priors were applied to the *b*'s, so the SEs for the *b*'s were substantially higher with the small sample (η^2 = .41). This main effect was smaller for the *a*'s (η^2 = .03), and there was an interaction between sample size and the strength of the prior for *a* (η^2 = .07); the decrease in the SE of it as the variance in the prior on *a* decreased was somewhat greater for smaller N. For the SE of c, again the main effect of sample size was small (η^2 = .02), with a small interaction between sample size and item difficulty (η^2 = .05).

Summary and Discussion

Table 2 and Table 3 summarize the effects due to either the strength of the prior or the match of the item parameter to the mean. In short, the mean of the prior impacted the bias, but the variance of the prior impacted the standard error. In Figure 1 and Figure 4, it is clear that mis-specifying the prior mean leads to sizeable bias in the *c*-parameter for the easy items. Unless much larger samples are available, there is almost no information in the data to estimate the *c*-parameter. This bias in the *c*-parameter leads to predictable bias in the *b*-parameter for easy items. There is no easy way to avoid this issue; fixing the *c*-parameter to zero (using a 2PL model) would yield greater bias for items with non-zero *c*-parameters. The specification of the mean of the prior for the *a*-parameter impacted the bias in the *a*-parameter when the variance of

the *a*-parameter was small; otherwise, the difference between the bias for *a*-parameters above and below the prior mean was relatively small.

The standard errors of the *a*-parameters noticeably increased as the variance of the *a*-parameter increased. The variance of the c-parameter had a less sizeable impact on standard errors; perhaps the effect would have been greater if a wider range of variances had been included.

One limitation in this study was that the match of the parameter to the mean of the prior was manipulated by varying the item parameters. When the bias was similar in magnitude but with opposite signs for items above vs. below the mean of the prior, it could reasonably be inferred that the bias was due to how well the item parameter matched the mean. However, the SE (and in some cases the bias) appeared to depend on the value of the parameter, not its matched to the prior. This could be inferred when, for example, instead of similar SE for parameter values above and below the mean, the lowest (or highest) parameters had the lowest SE, followed by middle values of SE for the values best matched to the mean, followed by highest SE for the highest (or lowest) parameter values. It might have been clearer to keep constant values of the *a* and *c* parameters across items while varying the mean of the prior for different items.

Overall, these findings reinforce the earlier suggestions that it makes little difference what priors are specified, as long as they are not too informative. For the prior of the *a*-parameter, variances of 1.0 or 1.5 seem to give similar levels of bias for these sample sizes. For the prior of the *c*-parameter, the examined variances were perhaps too strong for the easiest items. However, we do not recommend weaker priors because, with so little information in the data for easy items, weaker priors for the *c*-parameter would likely lead to estimation problems.

References

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of statistical Software*, *48*, 1-29.

DeMars, C. E. (2019). Revised Parallel Analysis with non-normal ability and a guessing parameter. *Educational and Psychological Measurement, 79* (1), 151-169. doi: 10.1177/0013164418767009

Gao, F., & Chen, L. (2005). Bayesian or Non-Bayesian: A comparison study of item parameter estimation in the three-parameter logistic model. Applied Measurement in Education, 18, 351-380.

Harwell, M. R., & Janosky, J. E. (1991). An empirical study of the effects of small datasets and varying prior variances on item parameter estimation in BILOG. *Applied Psychological Measurement, 15* (3), 279-291.

Kim, S.-H., 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.

Lim, R. G., & Drasgow, F. (1990). Evaluation of two methods for estimating item response theory parameters when assessing differential item functioning. Journal of Applied Psychology, 5 (2), 164-174.

Lord, F. M. (1968). An analysis of the Verbal Scholastic Aptitude Test using Birnbaum's three-parameter logistic model. *Educational and Psychological Measurement, 28* (4), 989-1020.

Lord, F. M. (1975, September). Evaluation with artificial data of a procedure for estimating ability and item characteristic curve parameters (Report Bulletin RB-75-33). Educational

Testing Service. https://onlinelibrary.wiley.com/doi/epdf/10.1002/j.2333-8504.1975.tb01073.x

Lord, F. M. (1986). Maximum likelihood and Bayesian parameter estimation in item response theory. *Journal of Educational Measurement, 23* (2), 157-162.

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.

Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika, 51*(2), 177-195.

Sheng, Y. (2010). A sensitivity analysis of Gibbs sampling for 3PNO IRT models: Effects of prior specifications on parameter estimates. *Behaviormetrika, 37*(2), 87-110.

Thissen, D., & Wainer, H. (1982). Some standard errors in item response theory. *Psychometrika, 47* (4), 397-412.

Zeng, L. (1997). Implementation of marginal Bayesian estimation with four-parameter beta prior distributions. *Applied Psychological Measurement, 21*(2), 143-156.

Tables

Table 1

Proportion of iterations in which at least one item estimate was unreasonable in all conditions for the two sample size conditions

		N= 500					N = 100				
Var of a prior	Var of logit-c prior	Extreme *a*	Extreme *b*	Extreme *c*	No SEs	At least one problem	Extreme a	Extreme b	Extreme c	No SEs	At least one problem
No prior	No prior	0.266	0.004	0.398	0.130	0.608	1.000	0.288	0.902	0.364	1.000
No prior	0.46	0.142	0.048	0	0	0.184	0.944	0.520	0	0.036	0.980
No prior	0.59	0.166	0.044	0	0	0.204	0.952	0.514	0	0.040	0.984
No prior	0.72	0.170	0.036	0	0.004	0.204	0.964	0.496	0	0.054	0.986
0.3	0.46	0	0	0	0	0	0	0.162	0	0	0.162
0.3	0.59	0	0	0	0	0	0	0.188	0	0.004	0.188
0.3	0.72	0	0	0	0	0	0	0.190	0	0.002	0.190
1.0	0.46	0	0.006	0	0.004	0.008	0	0.274	0	0.002	0.274
1.0	0.59	0	0.004	0	0	0.004	0	0.218	0	0	0.218
1.0	0.72	0	0	0	0	0	0	0.210	0	0	0.210
1.5	0.46	0	0.012	0	0.004	0.012	0	0.398	0	0	0.398
1.5	0.59	0	0.012	0	0.004	0.012	0	0.348	0	0	0.348
1.5	0.72	0	0.006	0	0.002	0.006	0	0.288	0	0	0.288

Note. Extreme item estimates were defined as a > 7 or < -2 , $b > 6$ & < -6, $c > .7$.

Table 2

Summary of Effects of Prior Specification on Bias

Priors on *c*'s (N = 500 only)	η^2
Bias for *a*-parameter	
b x match of *c* to prior mean	.10
match of *c* to prior mean x variance of prior for *c*	.02
b x match of *c* to prior mean x variance of prior for *c*	.02
Bias for *b*-parameter	
match of *c* to prior mean	.53
a x match of *c* to prior mean	.11
b x match of *c* to prior mean	.30
Bias for *c*-parameter	
match of *c* to prior mean	.84
a x match of *c* to prior mean	.02
b x match of *c* to prior mean	.12
Priors on *a*'s and *c*'s	
Bias for *a*-parameter	
match of *a* to prior mean	.38
match of *c* to prior mean	.20
match of *a* x *b*	.04
match of *a* x *N*	.04
match of *a* x variance of prior for *a*	.19
match of *a* x match of *c*	.03
Bias for *b*-parameter	
match of *c* to prior mean	.40
match of *a* x *b*	.11
match of *a* x match of *c*	.03
match of *c* x *b*	.04
match of *a* x variance of prior for *a*	.03
b x variance of prior for a	.02

N x *b* x variance of prior for a	.02
match of *a* x *b* x variance of prior for *a*	.06
Bias for *c*-parameter	
match of *c* to prior mean	.84
match of *c* x *b*	.09

Note: Only factors which accounted for at least 2% of the variance are included.

Table 3

Summary of Effects of Prior Specification on SE

Priors on *c*'s (N = 500 only)	η^2
SE for *c*-parameter	
variance of prior for *c*	.15
b x variance of prior for *c*	.05
Priors on *a*'s and *c*'s	
SE for *a*-parameter	
variance of prior for *a*	.74
N x variance of prior for *a*	.07
a x variance of prior for *a*	.03
b x variance of prior for *a*	.05
SE for *b*-parameter	
variance of prior for *a*	.03
SE for *c*-parameter	
variance of prior for *c*	.14
b x variance of prior for *c*	.03
a x variance of prior for *a*	.02

Note: Only factors which accounted for at least 2% of the variance are included.

Figures

Figure 1

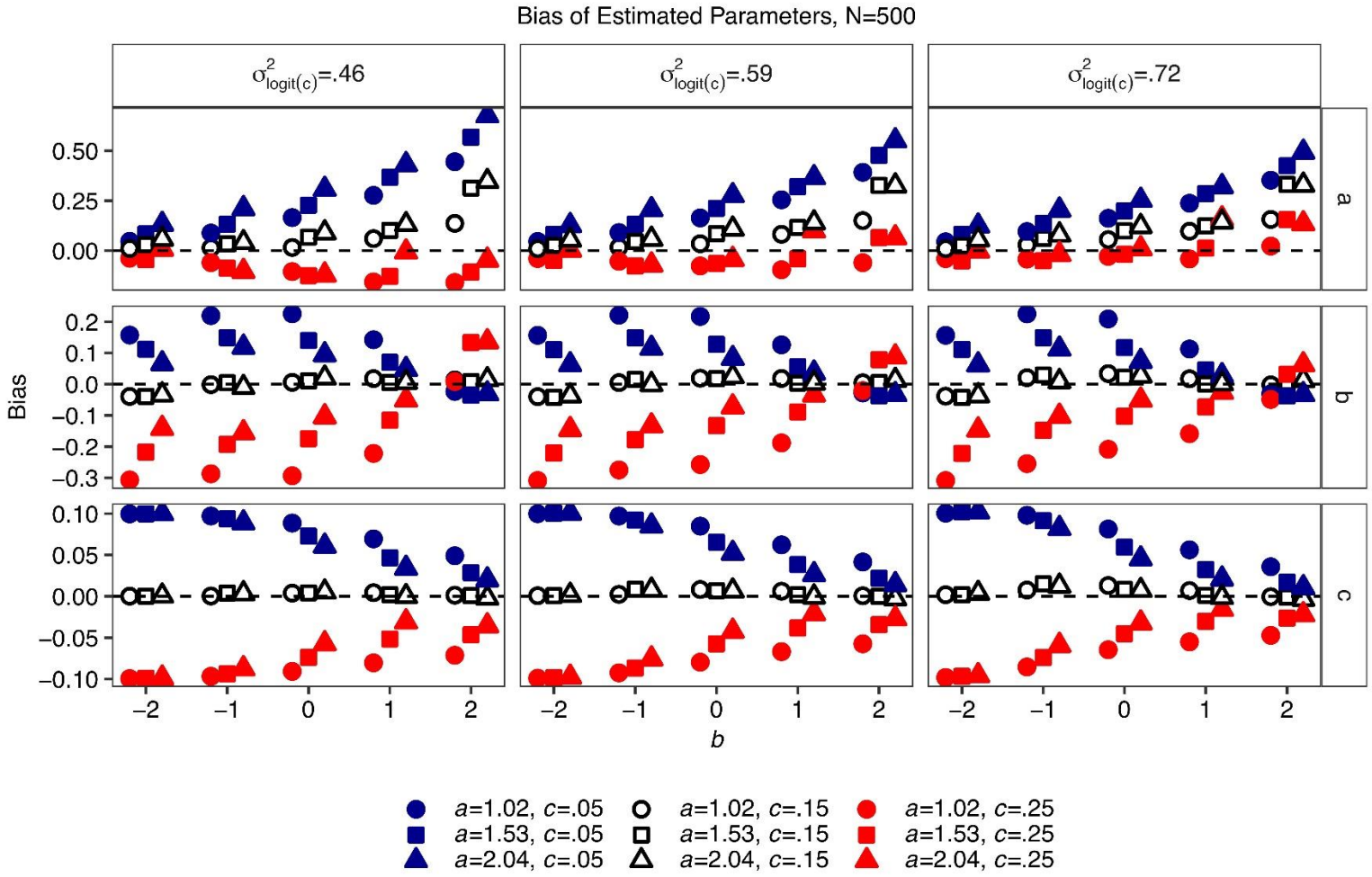


Figure 2

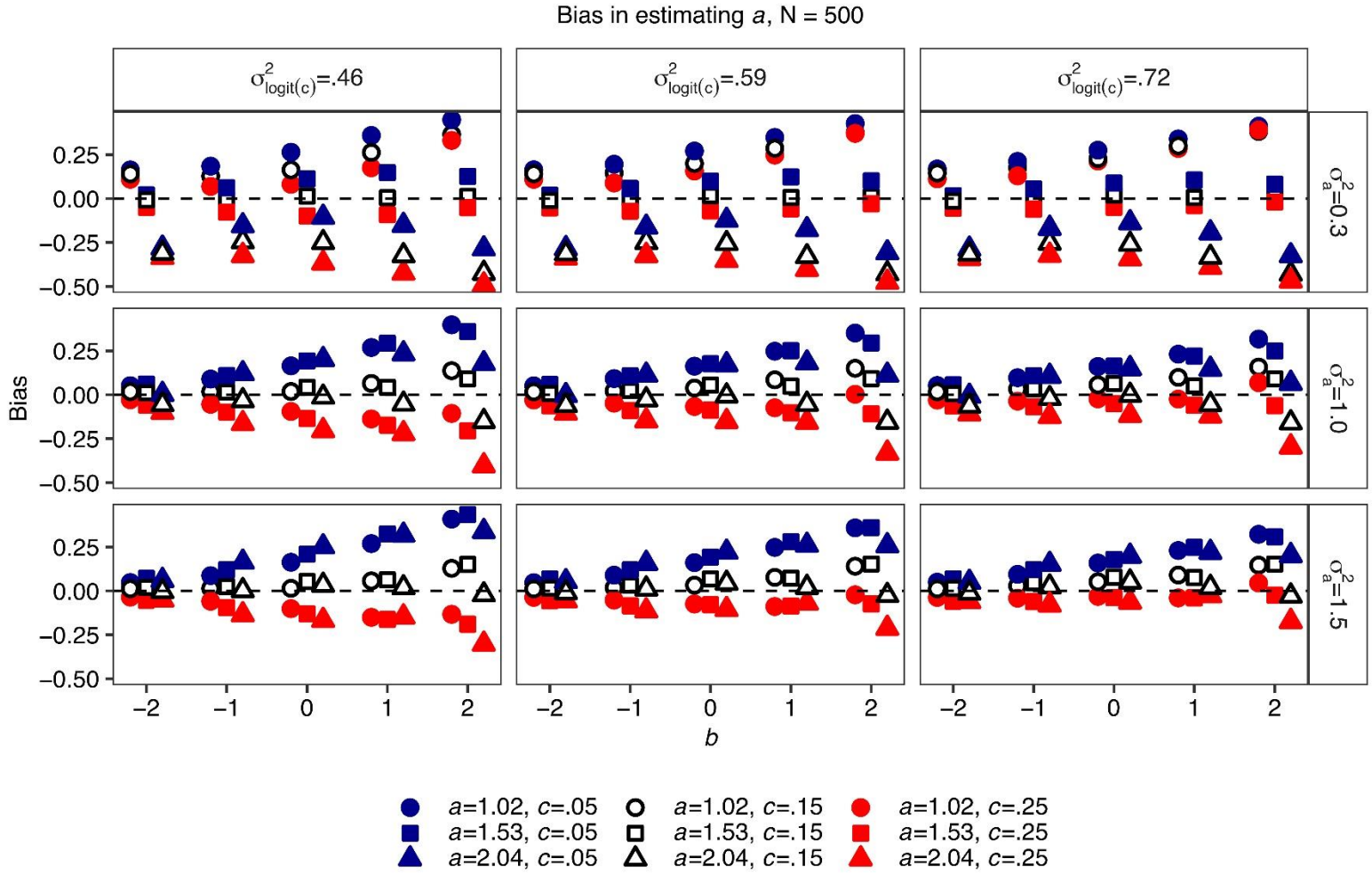

Bias in estimating a, N = 500
σ²logit(c)=.46
σ²logit(c)=.59
σ²logit(c)=.72
σ²a=0.3
σ²a=1.0
σ²a=1.5
Bias
b
a=1.02, c=.05
a=1.53, c=.05
a=2.04, c=.05
a=1.02, c=.15
a=1.53, c=.15
a=2.04, c=.15
a=1.02, c=.25
a=1.53, c=.25
a=2.04, c=.25

Figure 3

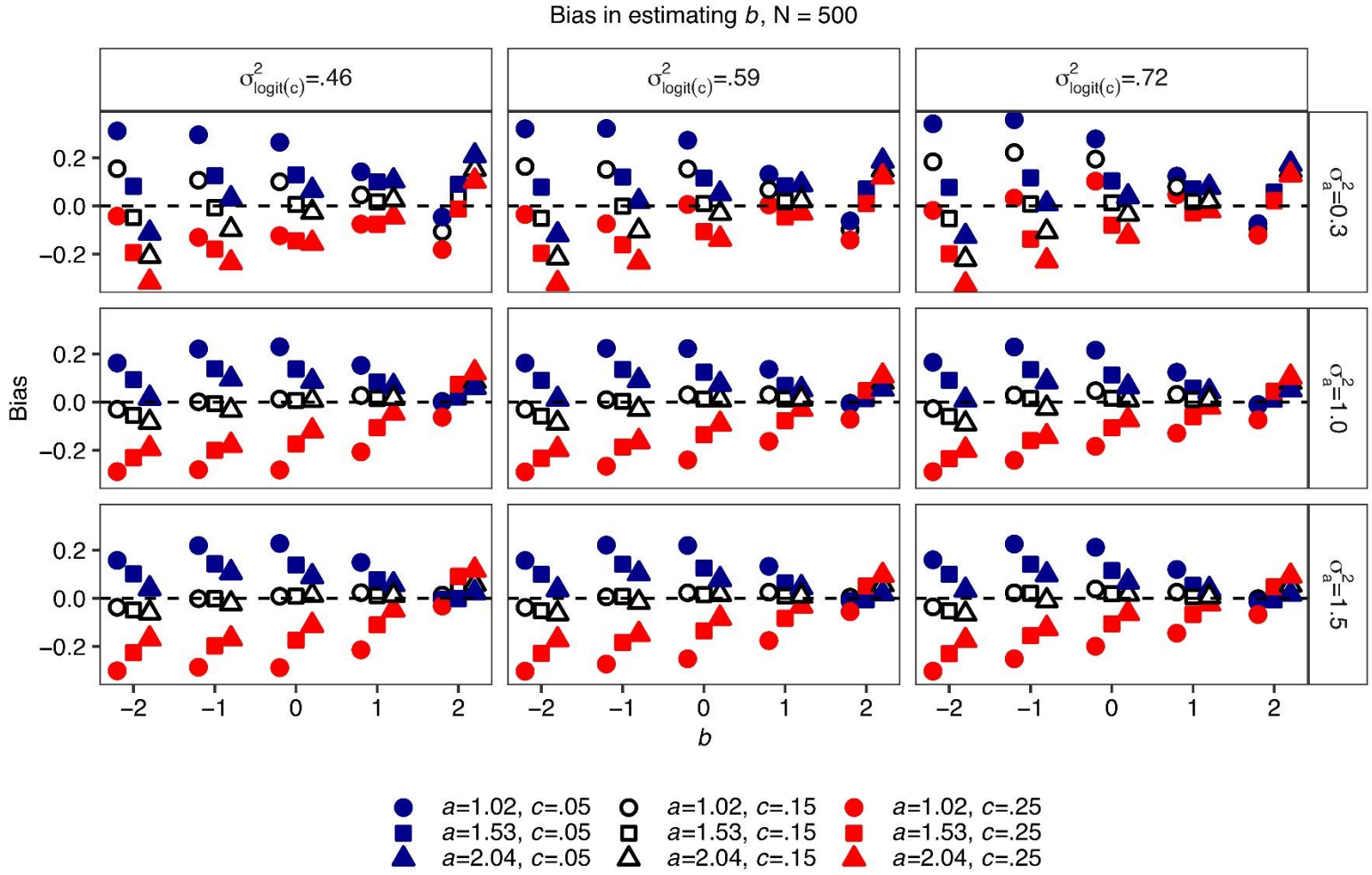

Bias in estimating b, N = 500
σ²logit(c)=.46
σ²logit(c)=.59
σ²logit(c)=.72
σ²a=0.3
σ²a=1.0
σ²a=1.5
Bias
b
a=1.02, c=.05
a=1.53, c=.05
a=2.04, c=.05
a=1.02, c=.15
a=1.53, c=.15
a=2.04, c=.15
a=1.02, c=.25
a=1.53, c=.25
a=2.04, c=.25

Figure 4

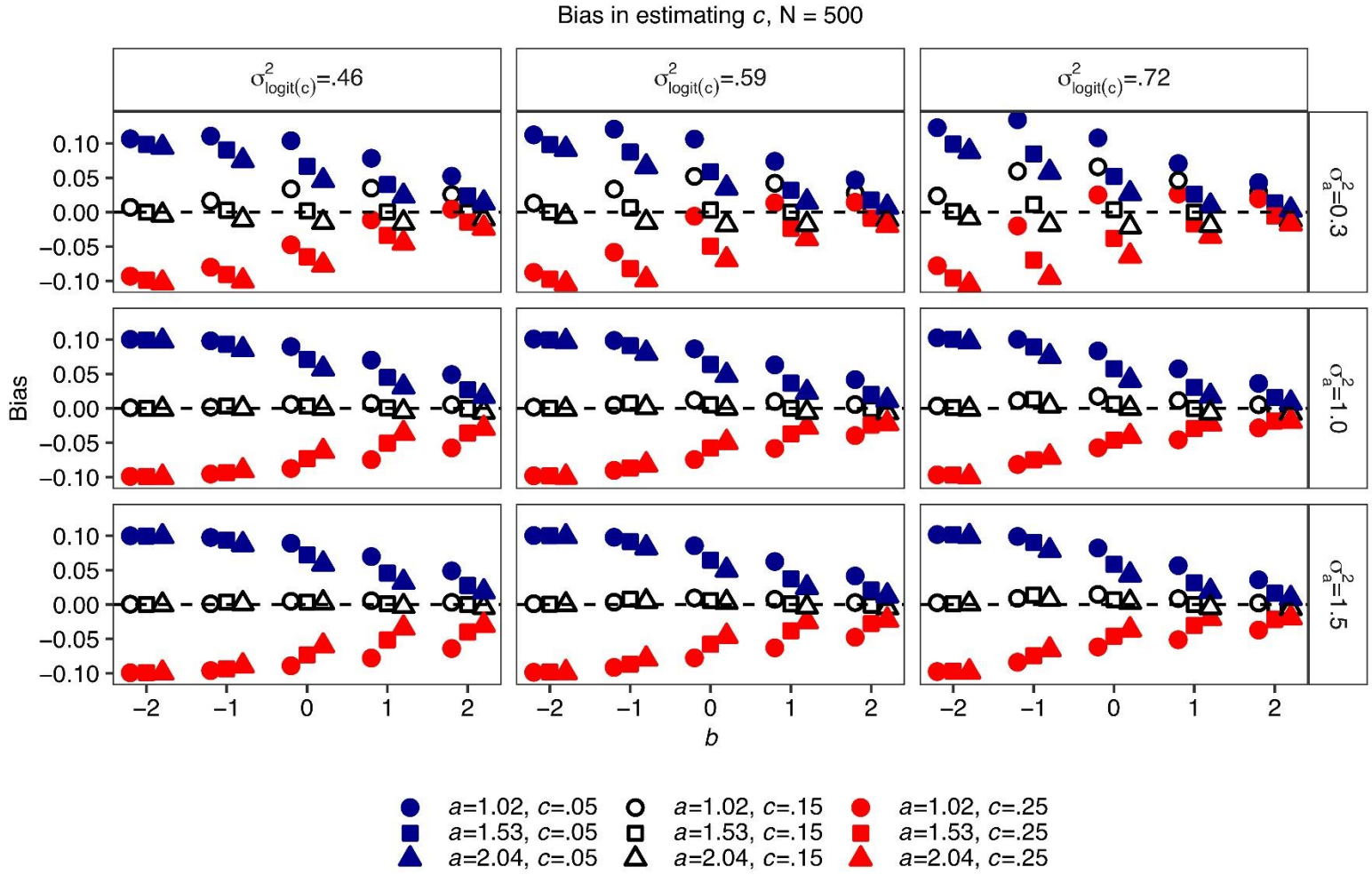

Bias in estimating c, N = 500
σ²logit(c)=.46
σ²logit(c)=.59
σ²logit(c)=.72
σ²a=0.3
σ²a=1.0
σ²a=1.5
Bias
b
0.10
0.05
0.00
−0.05
−0.10
−2
−1
0
1
2
a=1.02, c=.05
a=1.53, c=.05
a=2.04, c=.05
a=1.02, c=.15
a=1.53, c=.15
a=2.04, c=.15
a=1.02, c=.25
a=1.53, c=.25
a=2.04, c=.25

Figure 5

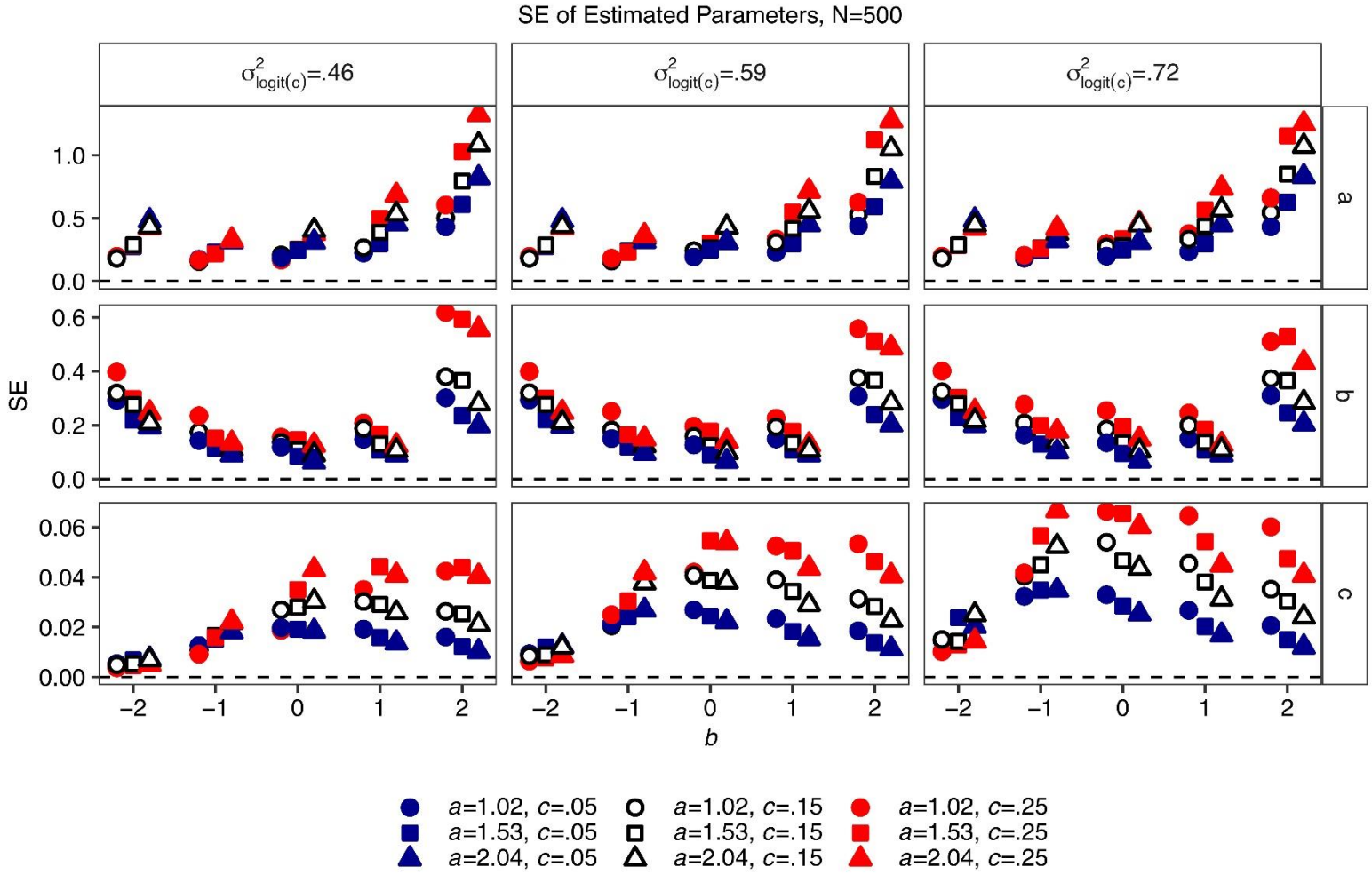

SE of Estimated Parameters, N=500
$\sigma^2_{logit(c)}$=.46
$\sigma^2_{logit(c)}$=.59
$\sigma^2_{logit(c)}$=.72
a
b
c
SE
b
a=1.02, c=.05
a=1.53, c=.05
a=2.04, c=.05
a=1.02, c=.15
a=1.53, c=.15
a=2.04, c=.15
a=1.02, c=.25
a=1.53, c=.25
a=2.04, c=.25

Figure 6

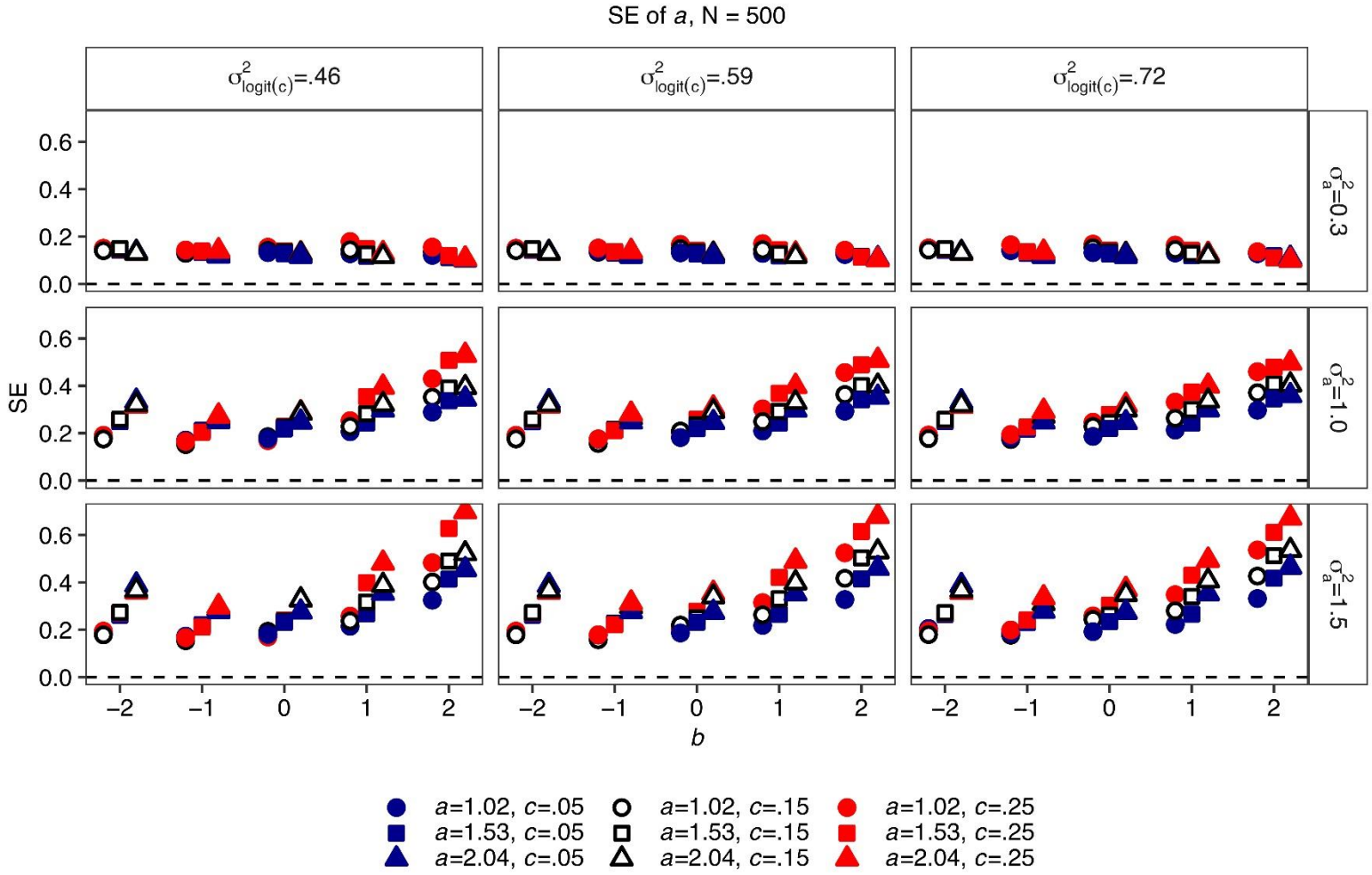

SE of a, N = 500
σ²logit(c)=.46
σ²logit(c)=.59
σ²logit(c)=.72
σ²a=0.3
σ²a=1.0
σ²a=1.5
SE
b
a=1.02, c=.05
a=1.53, c=.05
a=2.04, c=.05
a=1.02, c=.15
a=1.53, c=.15
a=2.04, c=.15
a=1.02, c=.25
a=1.53, c=.25
a=2.04, c=.25

Figure 7

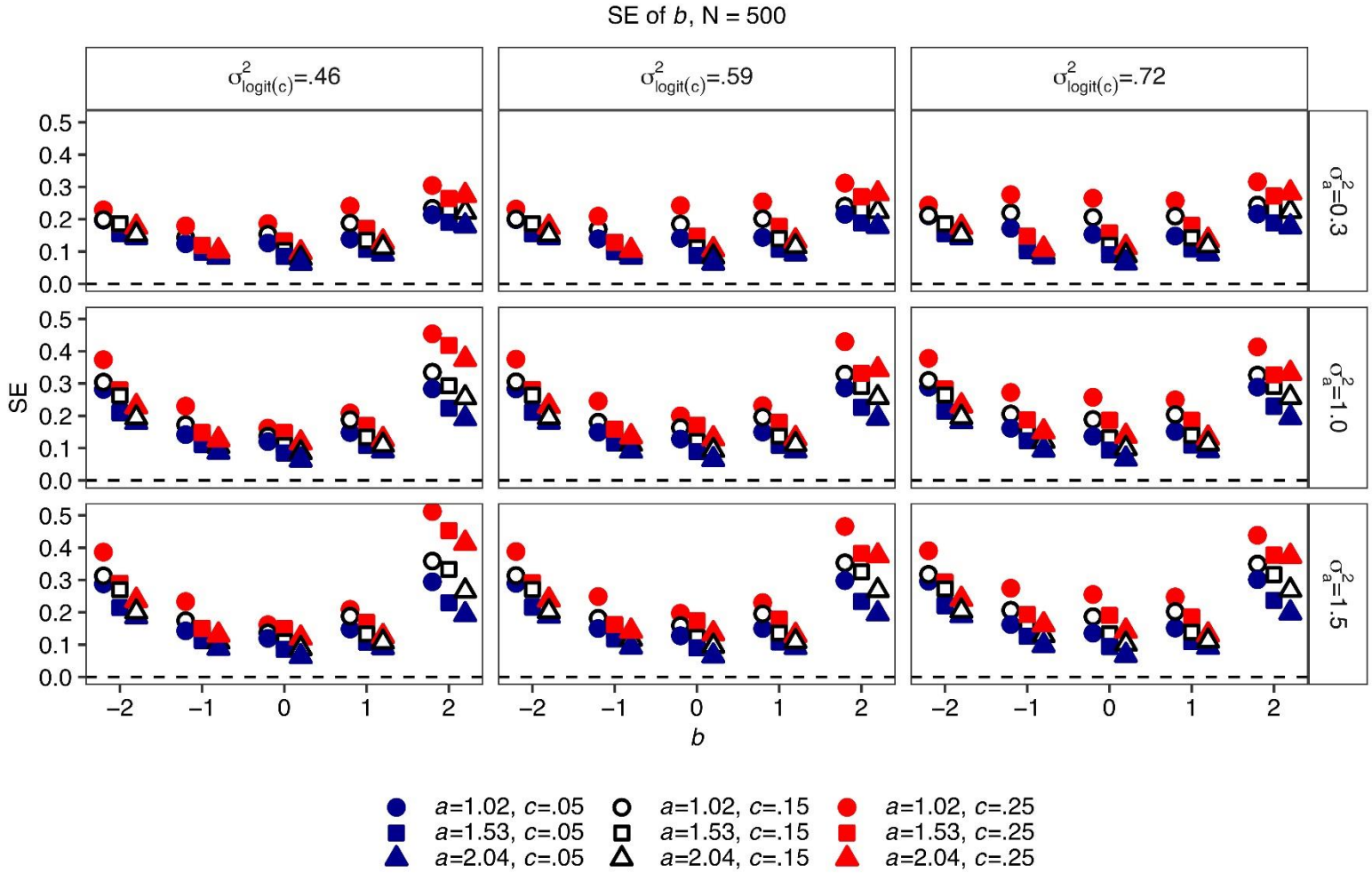
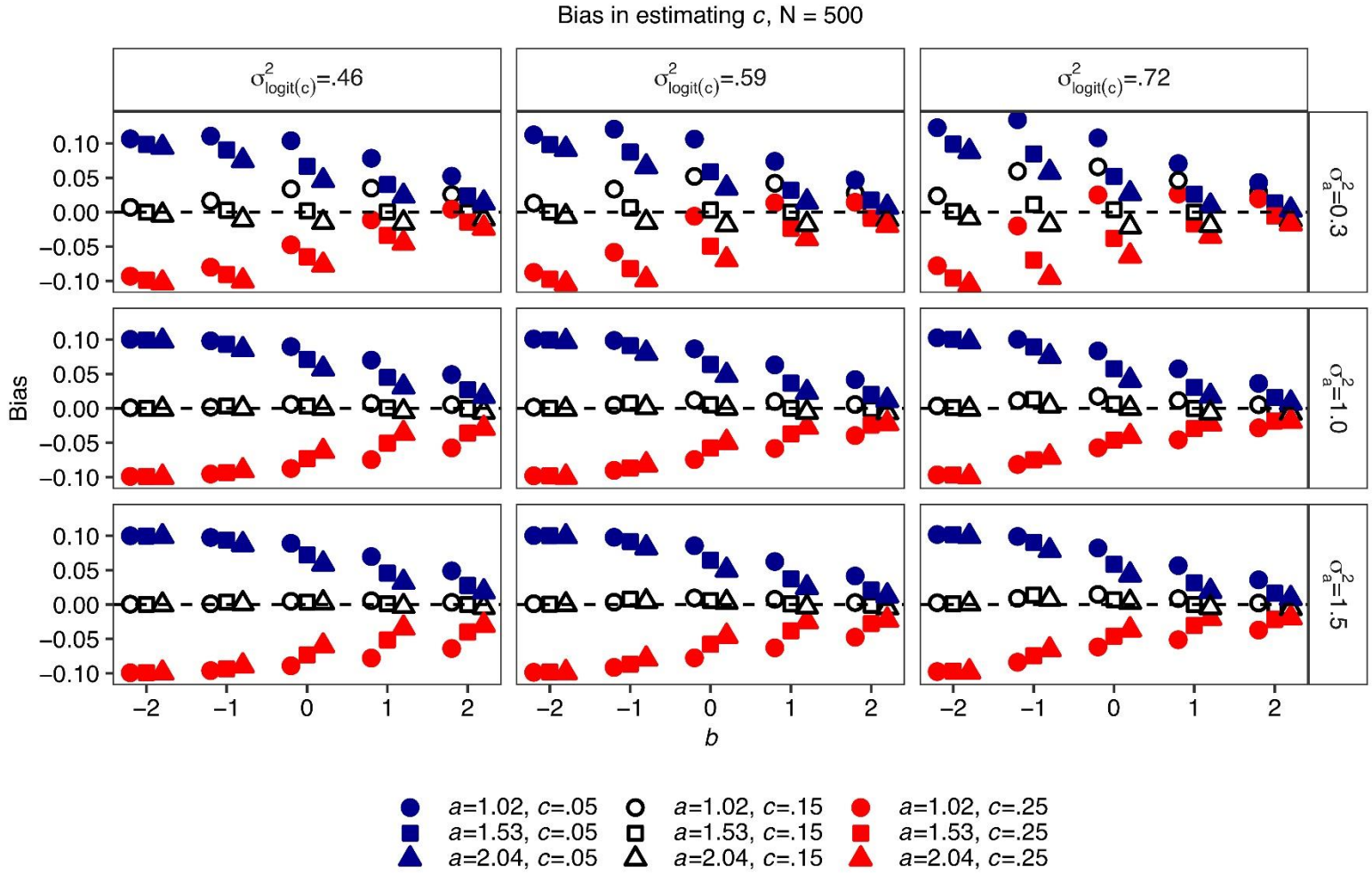
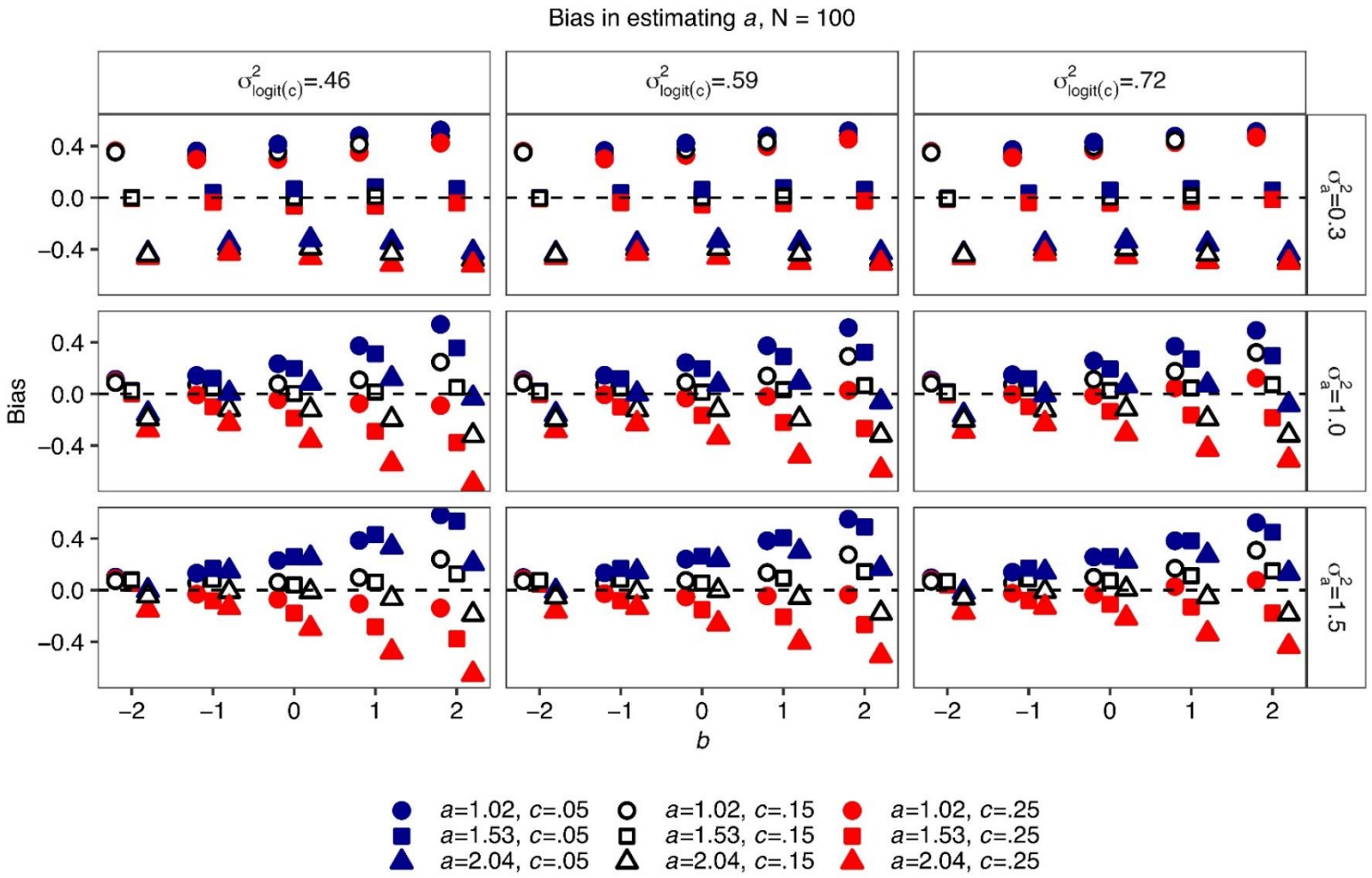
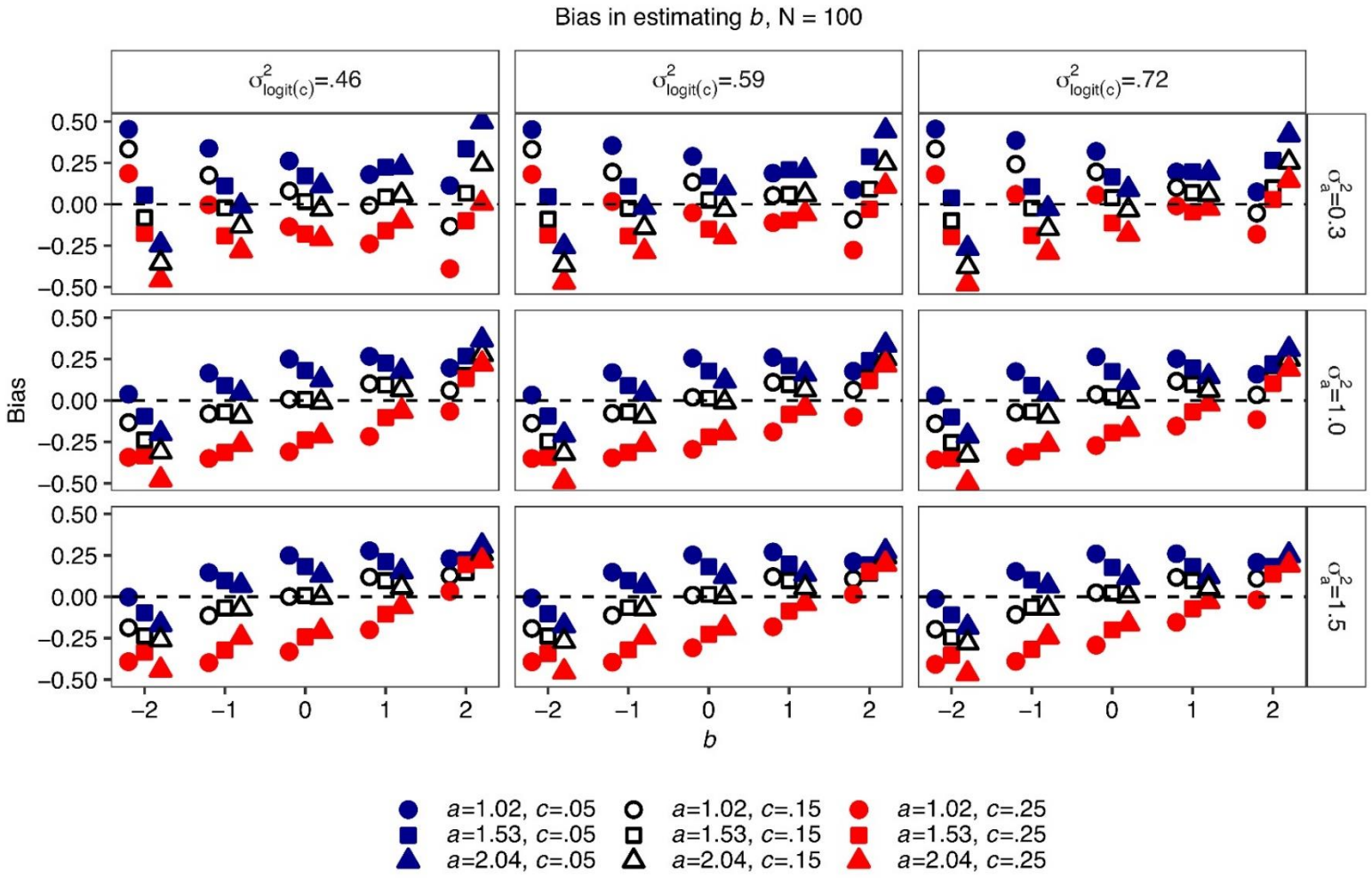

SE of b, N = 500
σ²logit(c)=.46
σ²logit(c)=.59
σ²logit(c)=.72
σ²a=0.3
σ²a=1.0
σ²a=1.5
SE
b
a=1.02, c=.05
a=1.53, c=.05
a=2.04, c=.05
a=1.02, c=.15
a=1.53, c=.15
a=2.04, c=.15
a=1.02, c=.25
a=1.53, c=.25
a=2.04, c=.25

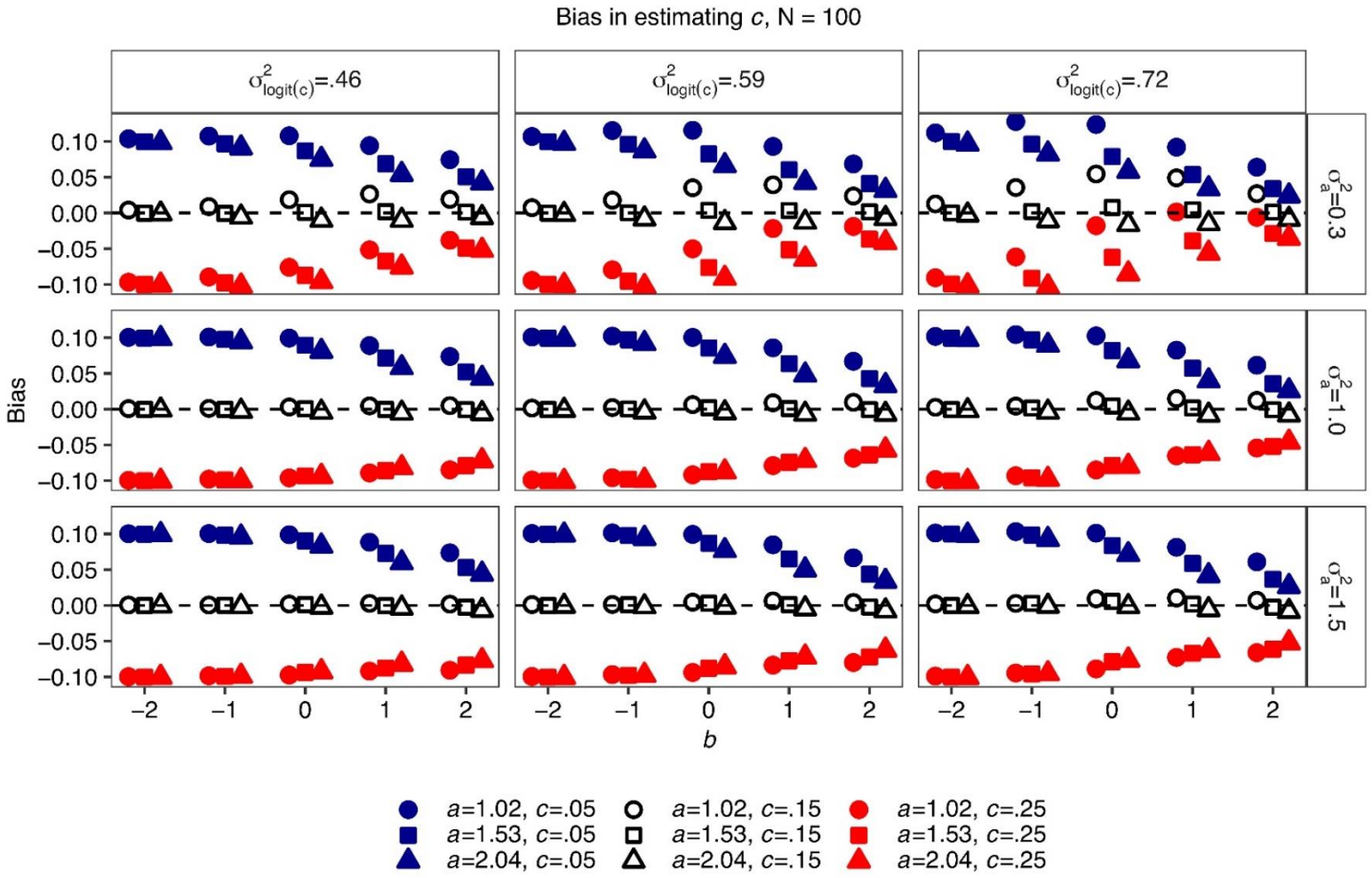
Figure 8

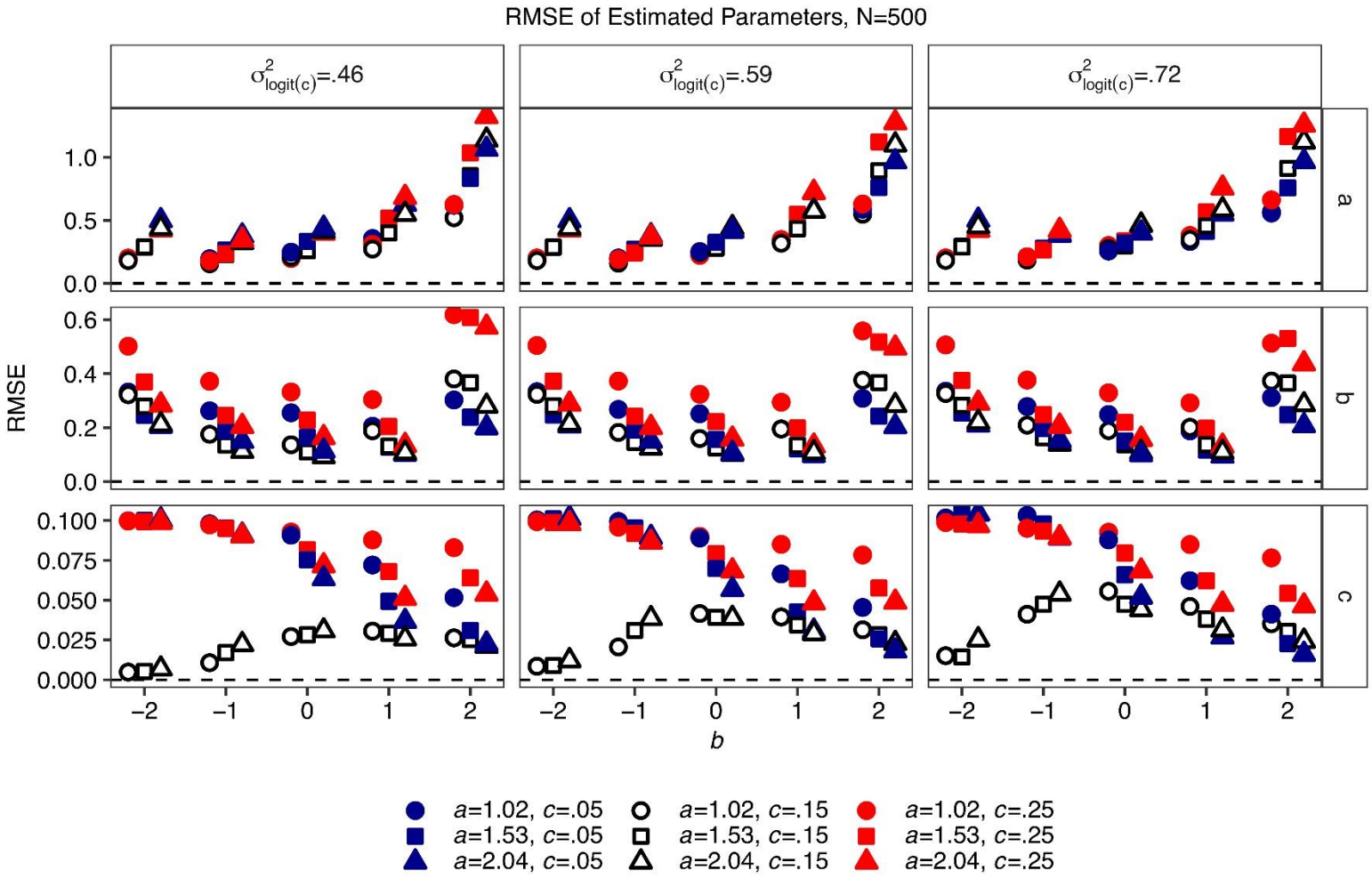

Bias in estimating c, N = 500
σ²logit(c)=.46
σ²logit(c)=.59
σ²logit(c)=.72
σ²a=0.3
σ²a=1.0
σ²a=1.5
Bias
b
0.10
0.05
0.00
−0.05
−0.10
−2
−1
0
1
2
a=1.02, c=.05
a=1.53, c=.05
a=2.04, c=.05
a=1.02, c=.15
a=1.53, c=.15
a=2.04, c=.15
a=1.02, c=.25
a=1.53, c=.25
a=2.04, c=.25

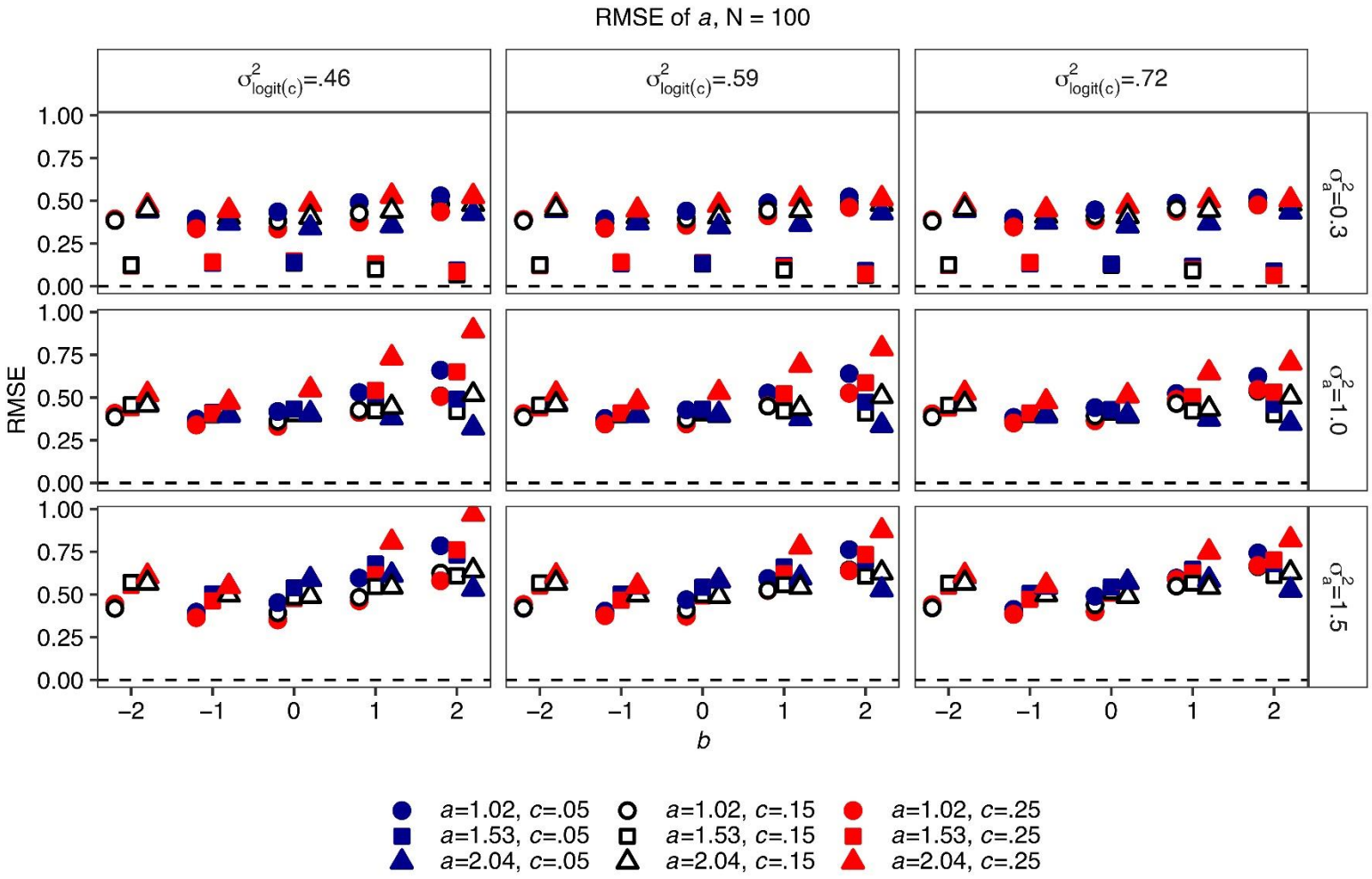
Appendix

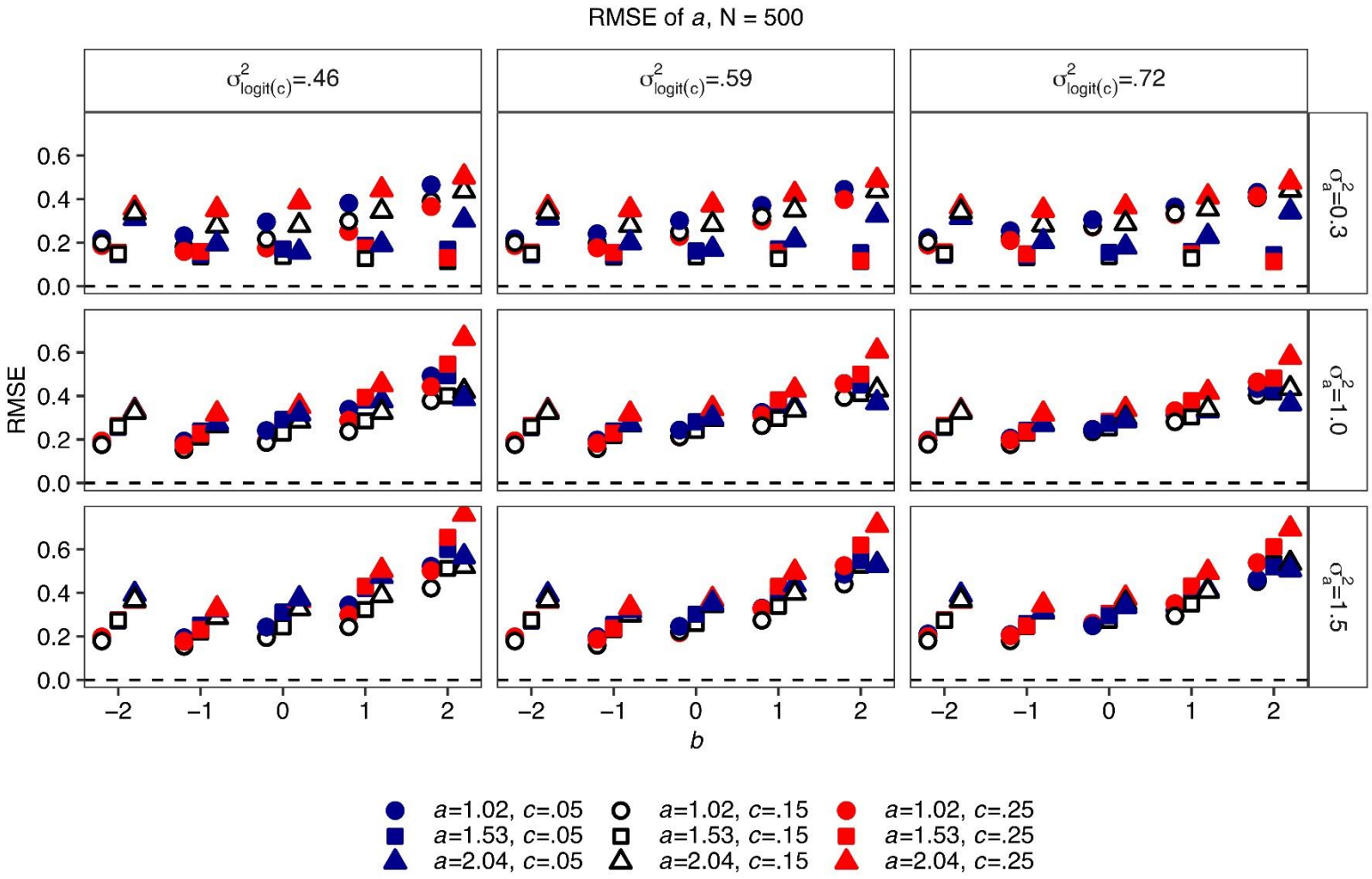

Bias in estimating a, N = 100
σ²logit(c)=.46
σ²logit(c)=.59
σ²logit(c)=.72
σ²a=0.3
σ²a=1.0
σ²a=1.5
Bias
b
0.4
0.0
−0.4
−2
−1
0
1
2
a=1.02, c=.05
a=1.53, c=.05
a=2.04, c=.05
a=1.02, c=.15
a=1.53, c=.15
a=2.04, c=.15
a=1.02, c=.25
a=1.53, c=.25
a=2.04, c=.25

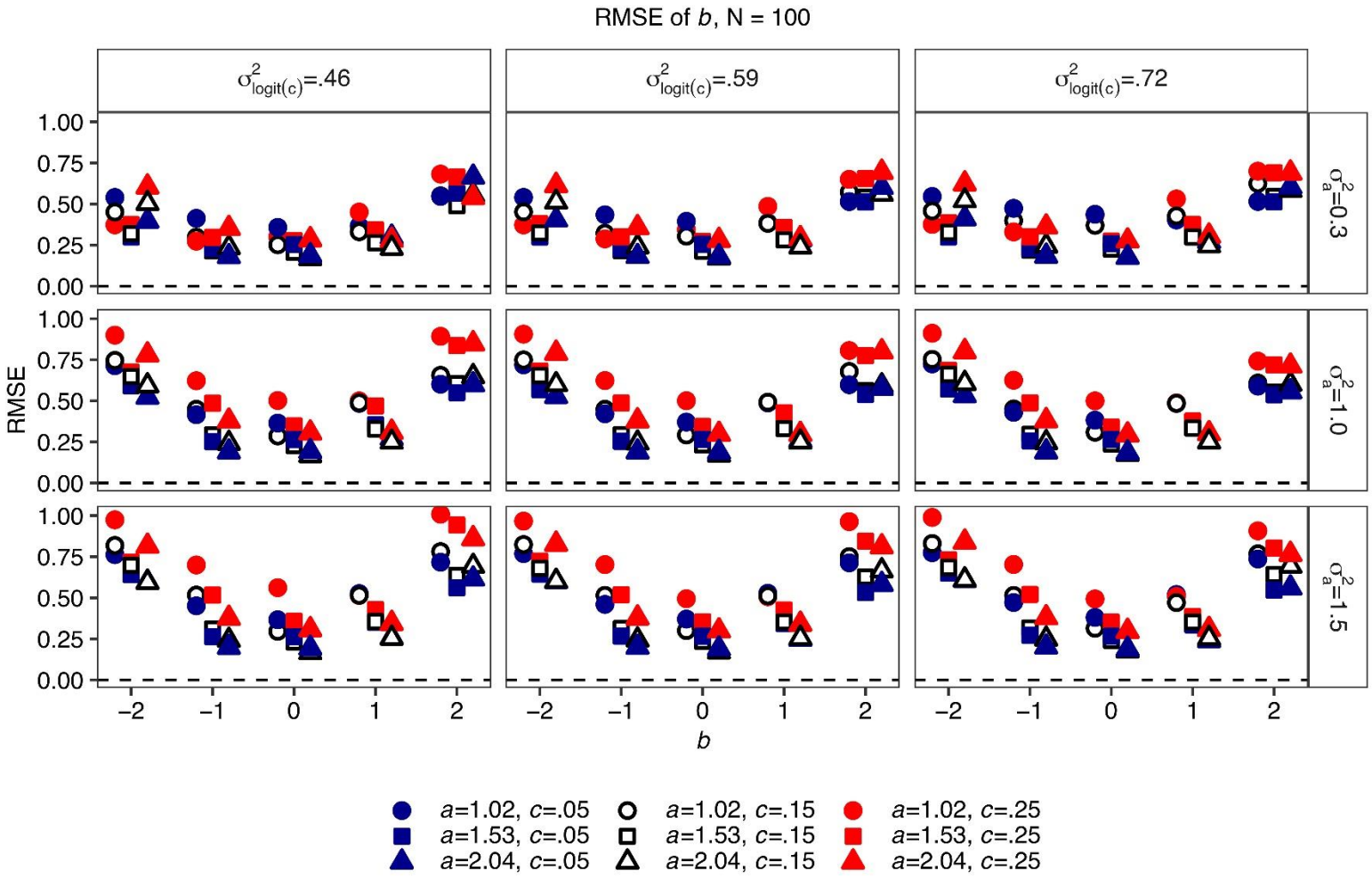

Bias in estimating b, N = 100
σ²logit(c)=.46
σ²logit(c)=.59
σ²logit(c)=.72
σ²a=0.3
σ²a=1.0
σ²a=1.5
Bias
0.50
0.25
0.00
−0.25
−0.50
−2
−1
0
1
2
b
a=1.02, c=.05
a=1.53, c=.05
a=2.04, c=.05
a=1.02, c=.15
a=1.53, c=.15
a=2.04, c=.15
a=1.02, c=.25
a=1.53, c=.25
a=2.04, c=.25

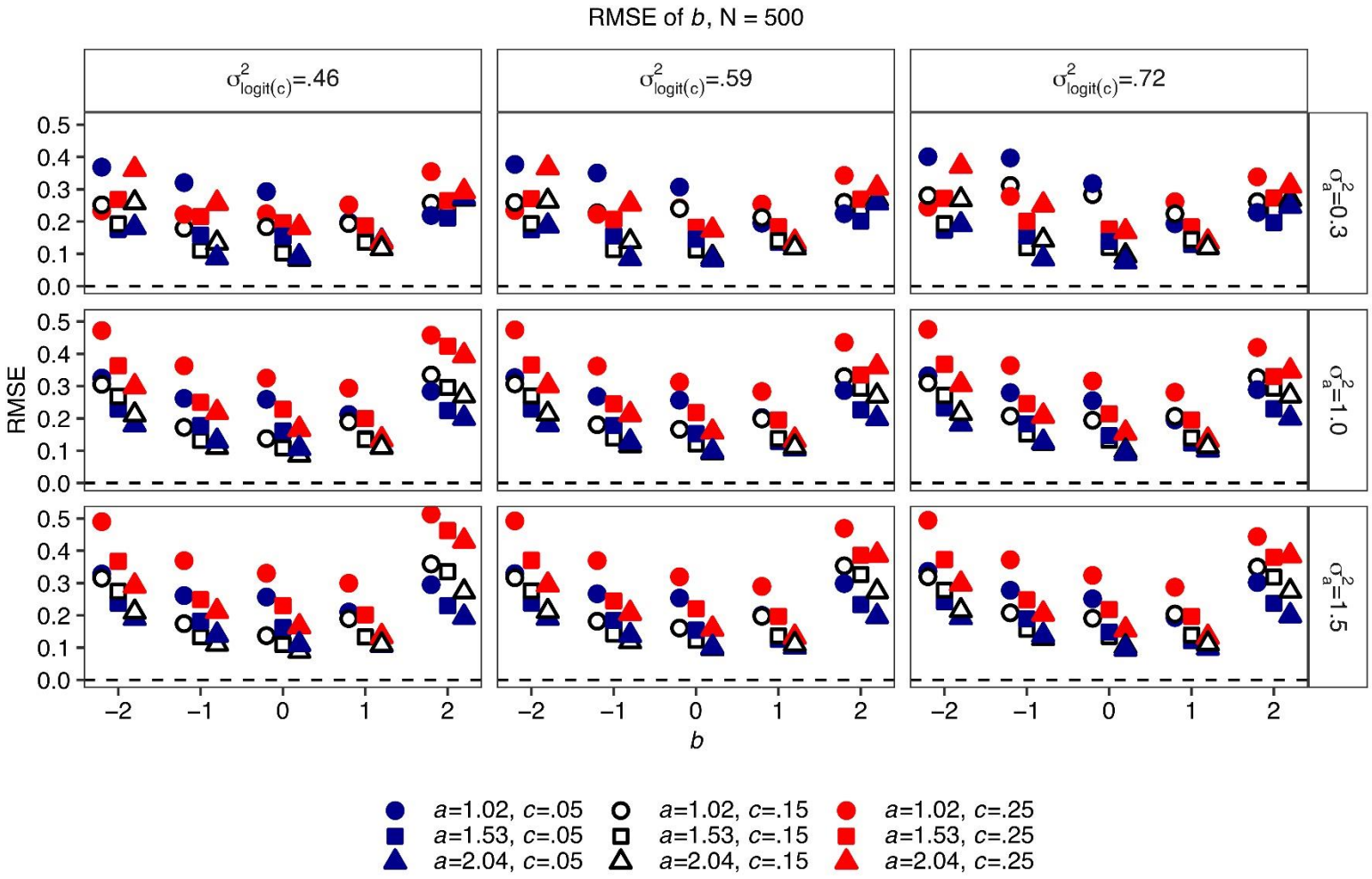

Bias in estimating c, N = 100
σ²logit(c)=.46
σ²logit(c)=.59
σ²logit(c)=.72
σ²a=0.3
σ²a=1.0
σ²a=1.5
Bias
0.10
0.05
0.00
−0.05
−0.10
−2
−1
0
1
2
b
a=1.02, c=.05
a=1.53, c=.05
a=2.04, c=.05
a=1.02, c=.15
a=1.53, c=.15
a=2.04, c=.15
a=1.02, c=.25
a=1.53, c=.25
a=2.04, c=.25

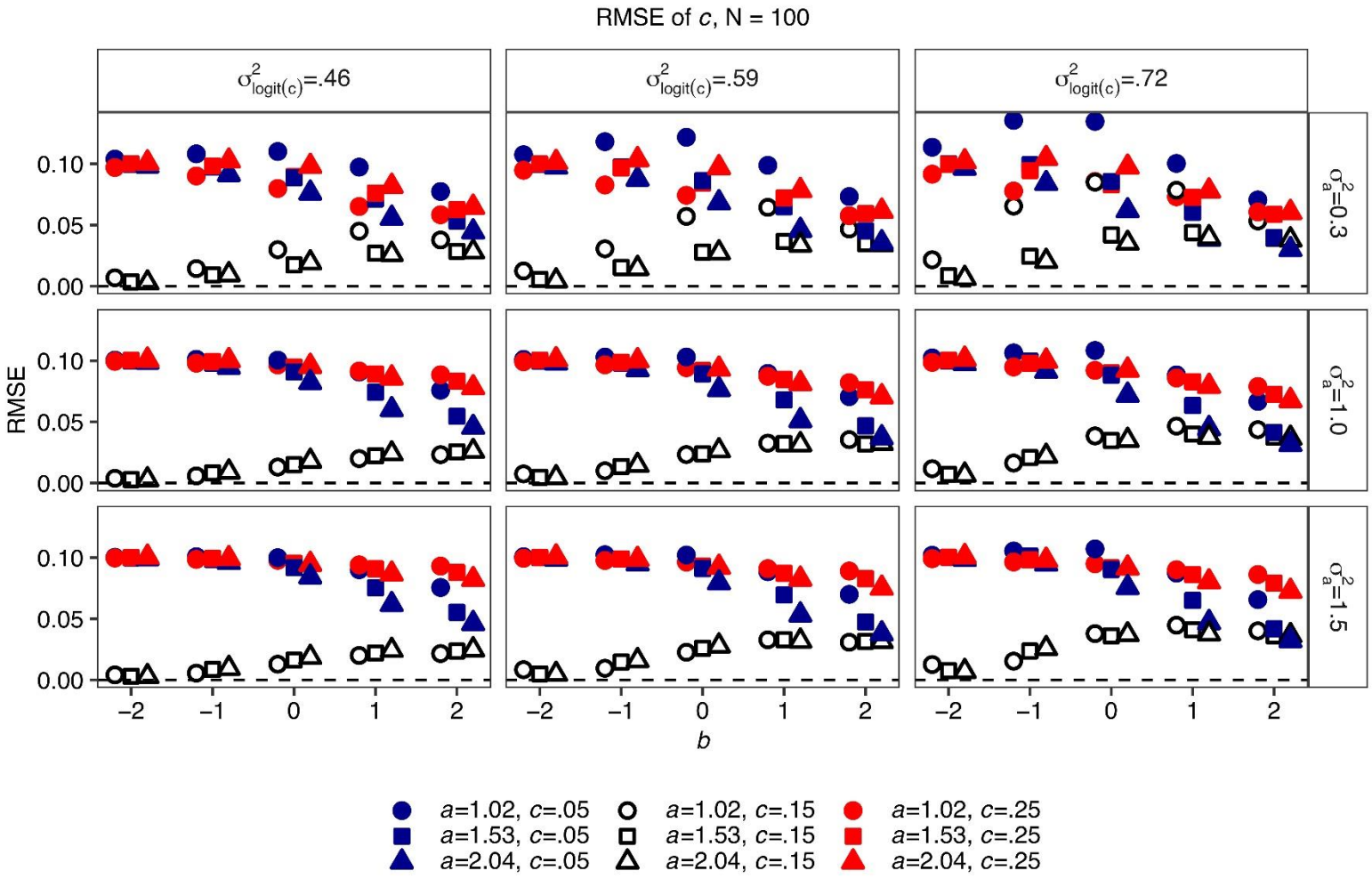

RMSE of Estimated Parameters, N=500
$\sigma^2_{logit(c)}$=.46
$\sigma^2_{logit(c)}$=.59
$\sigma^2_{logit(c)}$=.72
a
b
c
RMSE
1.0
0.5
0.0
0.6
0.4
0.2
0.0
0.100
0.075
0.050
0.025
0.000
−2
−1
0
1
2
b
a=1.02, c=.05
a=1.53, c=.05
a=2.04, c=.05
a=1.02, c=.15
a=1.53, c=.15
a=2.04, c=.15
a=1.02, c=.25
a=1.53, c=.25
a=2.04, c=.25

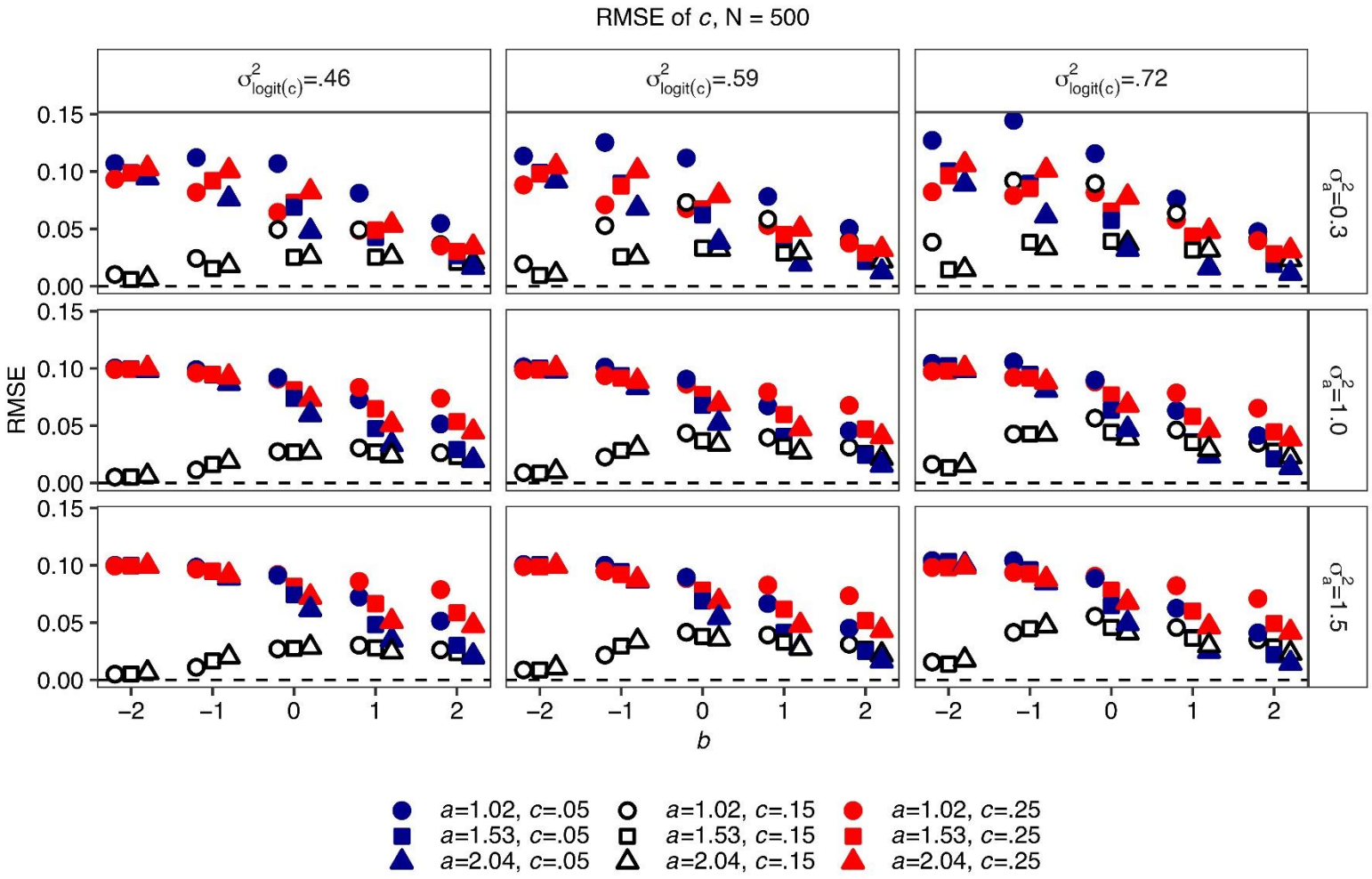

RMSE of a, N = 100
σ²logit(c)=.46
σ²logit(c)=.59
σ²logit(c)=.72
σ²a=0.3
σ²a=1.0
σ²a=1.5
RMSE
1.00
0.75
0.50
0.25
0.00
−2
−1
0
1
2
b
a=1.02, c=.05
a=1.53, c=.05
a=2.04, c=.05
a=1.02, c=.15
a=1.53, c=.15
a=2.04, c=.15
a=1.02, c=.25
a=1.53, c=.25
a=2.04, c=.25

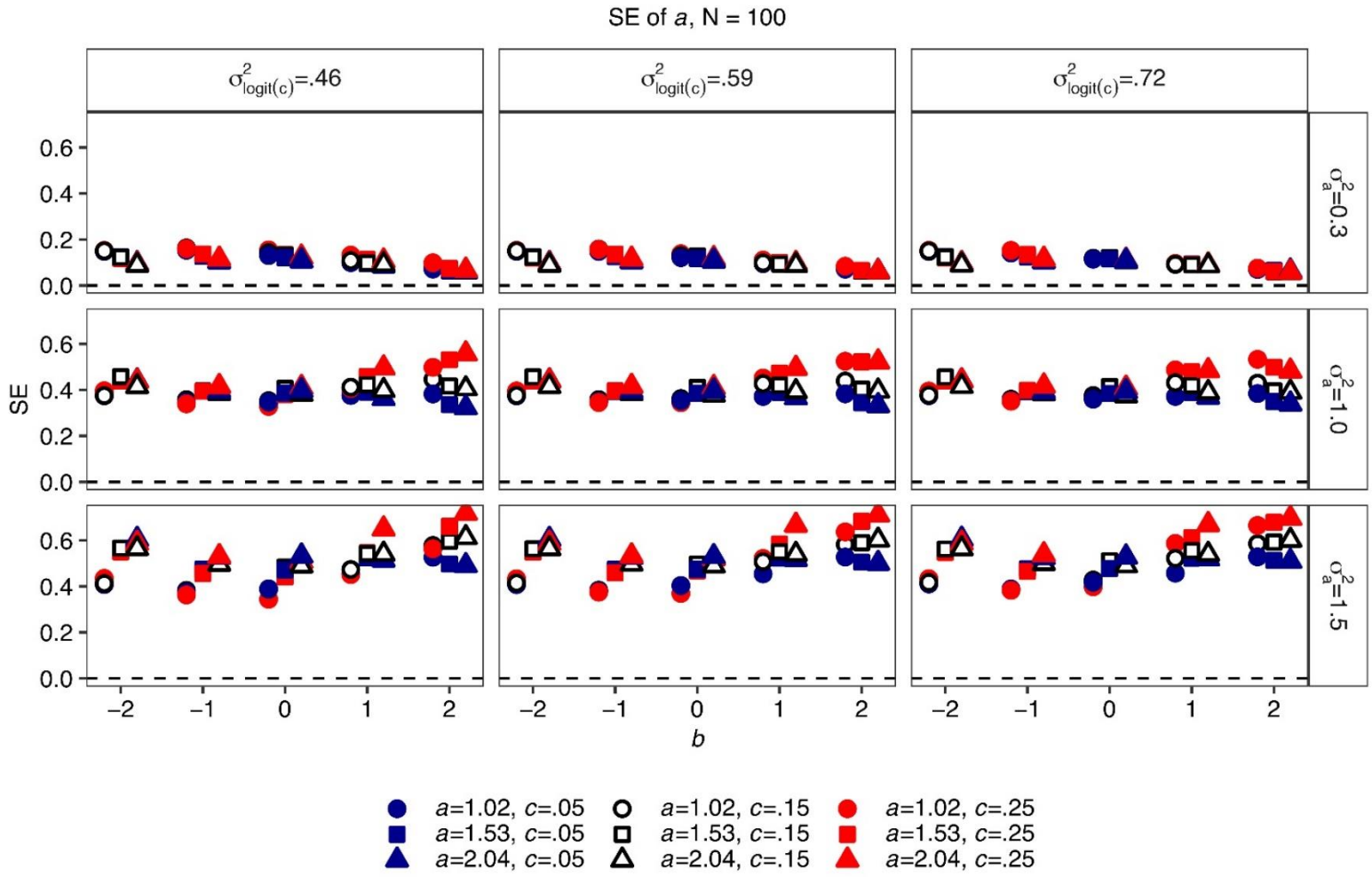

RMSE of a, N = 500
σ²logit(c)=.46
σ²logit(c)=.59
σ²logit(c)=.72
σ²a=0.3
σ²a=1.0
σ²a=1.5
RMSE
0.0
0.2
0.4
0.6
−2
−1
0
1
2
b
a=1.02, c=.05
a=1.53, c=.05
a=2.04, c=.05
a=1.02, c=.15
a=1.53, c=.15
a=2.04, c=.15
a=1.02, c=.25
a=1.53, c=.25
a=2.04, c=.25

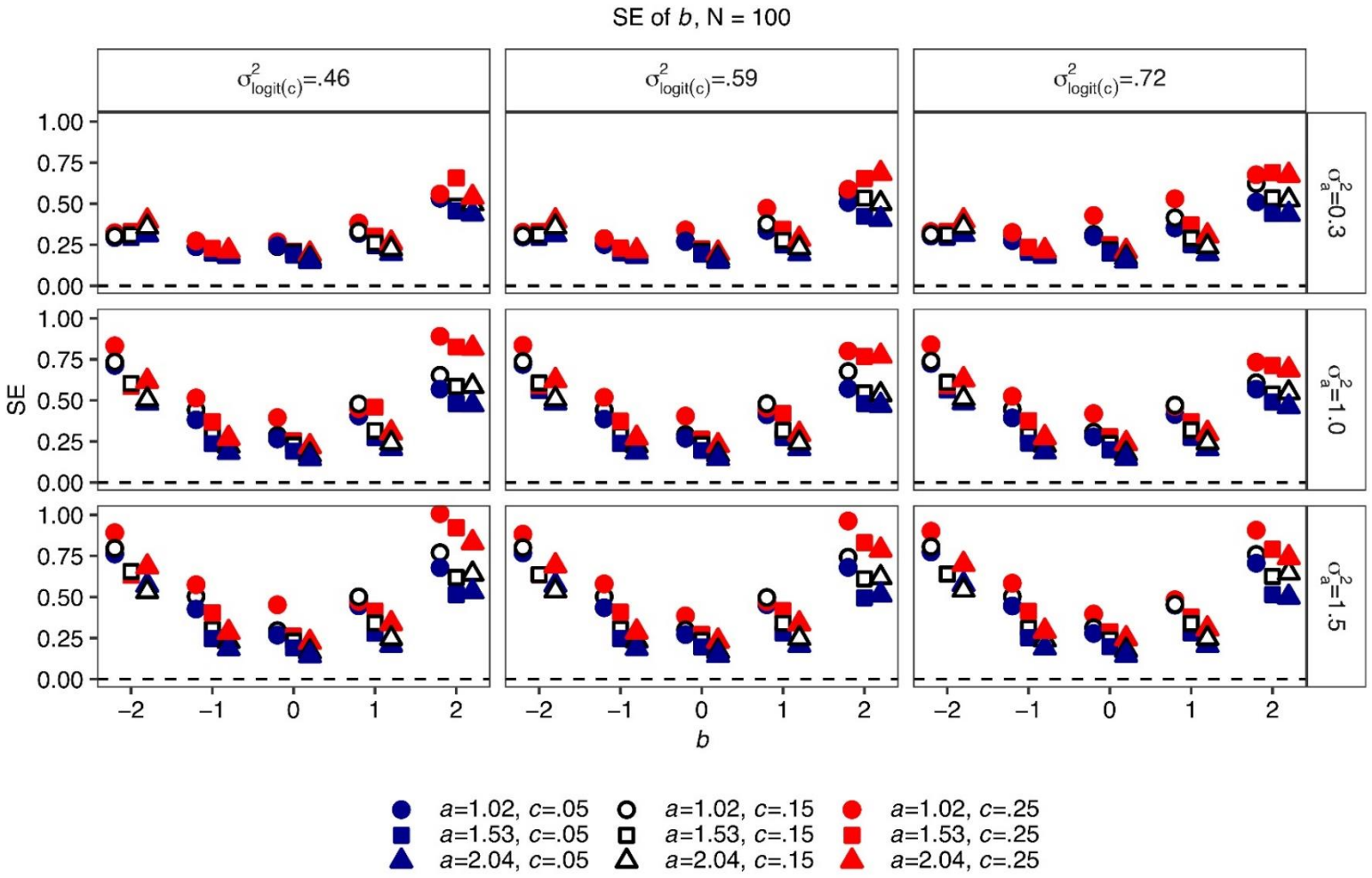

RMSE of b, N = 100
σ²logit(c)=.46
σ²logit(c)=.59
σ²logit(c)=.72
σ²a=0.3
σ²a=1.0
σ²a=1.5
RMSE
1.00
0.75
0.50
0.25
0.00
−2
−1
0
1
2
b
a=1.02, c=.05
a=1.53, c=.05
a=2.04, c=.05
a=1.02, c=.15
a=1.53, c=.15
a=2.04, c=.15
a=1.02, c=.25
a=1.53, c=.25
a=2.04, c=.25

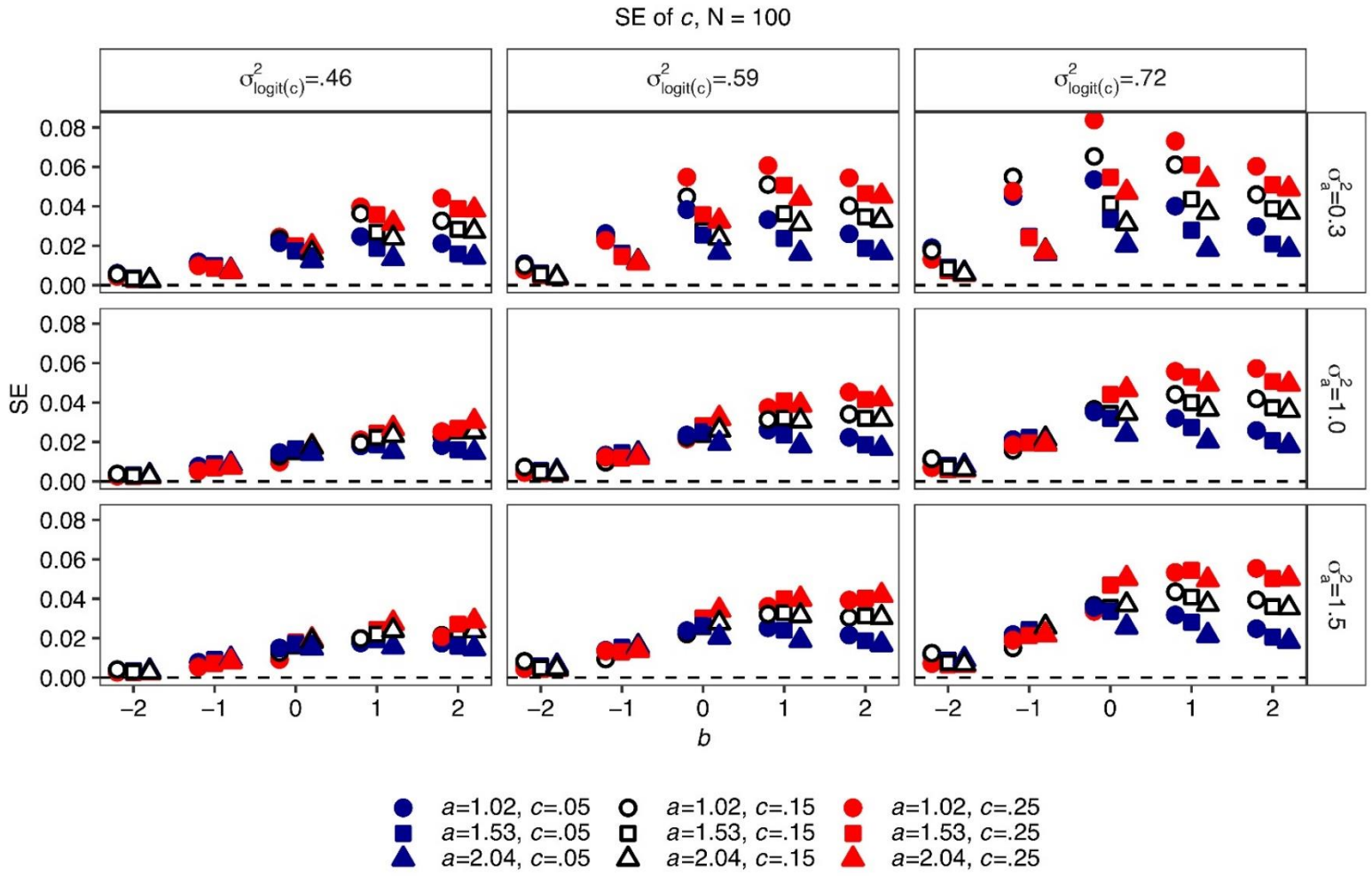

RMSE of b, N = 500
σ²logit(c)=.46
σ²logit(c)=.59
σ²logit(c)=.72
σ²a=0.3
σ²a=1.0
σ²a=1.5
RMSE
b
a=1.02, c=.05
a=1.53, c=.05
a=2.04, c=.05
a=1.02, c=.15
a=1.53, c=.15
a=2.04, c=.15
a=1.02, c=.25
a=1.53, c=.25
a=2.04, c=.25


RMSE of c, N = 100
σ²logit(c)=.46
σ²logit(c)=.59
σ²logit(c)=.72
σ²a=0.3
σ²a=1.0
σ²a=1.5
RMSE
0.00
0.05
0.10
−2
−1
0
1
2
b
a=1.02, c=.05
a=1.53, c=.05
a=2.04, c=.05
a=1.02, c=.15
a=1.53, c=.15
a=2.04, c=.15
a=1.02, c=.25
a=1.53, c=.25
a=2.04, c=.25


RMSE of c, N = 500
σ²logit(c)=.46
σ²logit(c)=.59
σ²logit(c)=.72
σ²a=0.3
σ²a=1.0
σ²a=1.5
RMSE
0.15
0.10
0.05
0.00
−2
−1
0
1
2
b
a=1.02, c=.05
a=1.53, c=.05
a=2.04, c=.05
a=1.02, c=.15
a=1.53, c=.15
a=2.04, c=.15
a=1.02, c=.25
a=1.53, c=.25
a=2.04, c=.25


SE of a, N = 100
σ²logit(c)=.46
σ²logit(c)=.59
σ²logit(c)=.72
σ²a=0.3
σ²a=1.0
σ²a=1.5
SE
b
a=1.02, c=.05
a=1.53, c=.05
a=2.04, c=.05
a=1.02, c=.15
a=1.53, c=.15
a=2.04, c=.15
a=1.02, c=.25
a=1.53, c=.25
a=2.04, c=.25


SE of b, N = 100
σ²logit(c)=.46
σ²logit(c)=.59
σ²logit(c)=.72
σ²a=0.3
σ²a=1.0
σ²a=1.5
SE
b
1.00
0.75
0.50
0.25
0.00
−2
−1
0
1
2
a=1.02, c=.05
a=1.53, c=.05
a=2.04, c=.05
a=1.02, c=.15
a=1.53, c=.15
a=2.04, c=.15
a=1.02, c=.25
a=1.53, c=.25
a=2.04, c=.25


SE of c, N = 100
σ²logit(c)=.46
σ²logit(c)=.59
σ²logit(c)=.72
σ²a=0.3
σ²a=1.0
σ²a=1.5
SE
0.00
0.02
0.04
0.06
0.08
−2
−1
0
1
2
b
a=1.02, c=.05
a=1.53, c=.05
a=2.04, c=.05
a=1.02, c=.15
a=1.53, c=.15
a=2.04, c=.15
a=1.02, c=.25
a=1.53, c=.25
a=2.04, c=.25