

Finding Item-Level Causes of Differential Item Functioning: A Hierarchical IRT Model for
Explaining DIF

Jennifer A. Brussow

University of Kansas

Submitted to the graduate degree program in Educational Psychology and the
Graduate Faculty of the University of Kansas in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

Committee Members

William P. Skorupski, Ph.D., Chairperson

Bruce Frey, Ph.D.

Vicki Peyton, Ph.D.

Dave Hansen, Ph.D.

Paul Johnson, Ph.D.

Date Defended: March 1, 2018

The dissertation committee for Jennifer Brussow certifies that this is the approved version of the following dissertation:

Finding Item-Level Causes of Differential Item Functioning: A Hierarchical IRT Model for Explaining DIF

William P. Skorupski, Ph.D., Chairperson

Date Approved: _____

Abstract

This research explored the effectiveness of using a hierarchical 2-PL item response theory (IRT) model to explain differential item functioning (DIF) according to item-level features. Explaining DIF in terms of variance attributable to construct-irrelevant item-level features would allow testing programs to improve item writing and item review processes to account for the features shown to predict DIF. Whereas previous research in this area has used classical test theory for scaling and logistic regression for DIF detection, this study explained DIF in terms of a hierarchical IRT model. Latent trait models are more widely used in operational testing programs; additionally, simultaneous estimation allows uncertainty in parameter estimates to be considered during the estimation of item-level features' relationship with DIF and is more parsimonious than a two-stage model.

This simulation study assessed the parameter recovery and stability of the proposed model across 36 different conditions created by varying four parameters: the strength of the correlation between the amount of DIF and the item-level features, the proportion of examinees in the reference group, and the mean and mixture probability of the mixture distribution used to sample items' DIF. The model successfully recovered person and item parameters, differences in groups' mean ability, and the relationship between the amount of DIF observed in an item and the presence of DIF-related item-level features. Model performance varied according to the values of the four parameters used to create conditions, especially the proportion of examinees in the reference group, which exhibited meaningful effect sizes in ANOVAs used to assess the parameters' impact on MSE and affected the model's power to detect DIF. When there were equal numbers of examinees in the reference and focal groups, the power to detect DIF increased, but at the expense of higher false positive rates and poorer precision.

Keywords: Differential item functioning, item response theory, hierarchical models, explanatory item response theory models, simulation studies, construct-irrelevant variance

Acknowledgements

My interest in measurement and completion of this degree are largely due to my outstanding advisor, Billy Skorupski, whose fateful IRT workshop at Stats Camp 2013 started me along this path. Billy, you have taught me volumes about a wide range of measurement topics and supported me through many professional milestones. I am incredibly grateful to have benefited from your guidance and friendship.

My other committee members also provided invaluable support. Many thanks to Paul Johnson, who introduced me to cluster computing, Git, and many other bits of technical knowledge that continue to prove useful with each passing day. I will continue to think of you each time I exit vim. Dave Hansen, thank you for your support and unique perspective on this study – you helped me clearly articulate the practical applications of my work. Vicki Peyton, thank you for talking me into signing up for the program and for your encouragement along the way. Bruce Frey, thank you for stepping in to save my defense date and for providing excellent suggestions for connecting the model to practice.

All model estimation in this dissertation was conducted on the University of Kansas Community Cluster. Access was provided by the Center for Research Methods and Data Analysis and the College of Liberal Arts and Sciences, and special thanks go to Paul Johnson for his sponsorship of my account.

This degree taught me a lot about measurement, but it also connected me to many excellent friends. Jessica Loughran, thank you for your endless capacity for listening to my stories and complaints and for your consistently excellent advice. I highly doubt I would have survived without our many, many coffee dates. Jake Thompson, thank you for patiently guiding me through the depths of the R learning curve, for working through technical minutiae with me, and for generally being an excellent friend. Jenna Zucker, thanks for your unfailing optimism; your support helped me at least pretend to believe in myself at critical moments.

Finally, my warmest thanks go to my family. My parents, Chuck and Ann Brussow, taught me that I could accomplish anything I set my mind to, and they have provided unflagging love and support through an appalling number of years of post-secondary education. My husband, Aaron Miller, has been my biggest supporter throughout this degree. Aaron, it is a great understatement to say that I couldn't have done this without you. Two dozen years of education have not provided me with sufficient words to convey how lucky I am.

Table of Contents

Abstract iii

Acknowledgements v

List of Tables vii

List of Figures viii

Chapter I: Introduction 1

 Context of Study 1

 Significance of Study 2

 Research Questions 4

Chapter II: Literature Review 5

 Differential Item Functioning 5

 Hierarchical Linear Models 6

 Item Response Theory Models 8

 Explanatory IRT Models 10

 Applications of Explanatory IRT Models 12

 Pilot Study 16

 Methods. 16

 Results. 17

 Conclusions. 19

Chapter III: Methods 19

 Simulating Data 19

 Conditions 24

 Analysis 25

 Outcome Variables 26

 Technical Considerations for the Practitioner 27

Chapter IV: Results and Discussion 28

 Convergence 28

 Item and Person Parameter Recovery 29

 Bias. 30

 Mean Squared Error 35

 Correlation. 40

 Recovery of group means. 42

 Recovery of the Relationship between Item-Level Features and DIF 44

Credible Intervals..... 45

Decision Consistency..... 48

Chapter V: Conclusion..... 52

Conclusions..... 52

 Research Question 1..... 52

 Research Question 2..... 52

 Research Question 3..... 53

 Research Question 4..... 53

Uses of the Proposed Model 55

Limitations 56

Future Directions 56

References..... 58

Appendix A. Density plots of mixture distributions..... 62

Appendix B. Histograms of the mean \hat{R} value across parameters for the replications within each condition 69

Appendix C. Correlation scatterplots for a -parameter recovery..... 76

Appendix D. Correlation scatterplots for b -parameter recovery 83

Appendix E. Correlation scatterplots for D -parameter recovery 90

Appendix F. Correlation scatterplots for θ recovery 97

Appendix G. Mean recovery histograms for the reference group means 104

Appendix H. Mean recovery histograms for the focal group means 111

Appendix I. Decision consistency scatterplots for the 0.5 DIF flagging threshold 118

Appendix J. Decision consistency scatterplots for the 0.75 DIF flagging threshold..... 125

Appendix K. Decision consistency scatterplots for the 1.0 DIF flagging threshold..... 132

Appendix L. Decision consistency tables for each DIF flagging threshold 139

Appendix M. R and Stan code for data generation and model estimation 143

List of Tables

Table 1. Models as a function of the predictors..... 11

Table 2. Conditions examined in the previous study 17

Table 3. Parameter recovery statistics: means derived from parameter means over replications 18

Table 4. R^2 recovery: means derived from parameter means and medians over replications 18

Table 5. Conditions examined in the present study 24

Table 6. Priors used in estimation..... 26

Table 7. Proportion of R values less than 1.1 for each condition 28

Table 8. Mean bias by parameter type across replications 31

Table 9. ANOVA results for a -parameter bias 32

Table 10. ANOVA results for b -parameter bias 33

Table 11. ANOVA results for D -parameter bias 34

Table 12. ANOVA results for θ bias 34

Table 13. Mean MSE for each parameter type across replications..... 36

Table 14. ANOVA results for a -parameter MSEs..... 37

Table 15. ANOVA results for b -parameter MSEs..... 37

Table 16. ANOVA results for D -parameter MSEs..... 38

Table 17. ANOVA results for θ MSEs. 38

Table 18. Correlation between simulated and recovered item and person parameters..... 41

Table 19. Mean bias for group means for each condition across replications 42

Table 20. Mean proportion of credible intervals containing the true simulated value, level 1 terms
..... 46

Table 21. Mean proportion of credible intervals containing the true simulated value, level 2 terms
..... 47

Table 22. Mean decision consistency results by parameter type using a 0.5 DIF threshold 50

Table 23. Mean decision consistency results by parameter type using a 0.75 DIF threshold 50

Table 24. Mean decision consistency results by parameter type using a 1.0 DIF threshold 50

List of Figures

Figure 1. Density plots of R^2 recovery..... 45

Chapter I: Introduction

Context of Study

Differential item functioning (DIF) occurs when an item functions differently across groups of respondents after controlling for overall ability. An item exhibiting DIF has different statistical properties for different groups of examinees after the groups have been matched on proficiency (Angoff, 1993). One group – typically the group with more members – is denoted as the reference group, while the other group is considered the focal group. DIF detection procedures investigate whether the focal group is disadvantaged due to statistical differences in item performance. The presence of DIF in an item is due to construct-irrelevant variance introduced by item-level features that disadvantage respondents due to ethnic/racial, gender, and/or cultural differences (de Ayala, 2013).

Such construct-irrelevant variance poses a threat to the assessment's validity argument, as does the possibility of subgroups of examinees receiving artificially inflated or depressed scores. Therefore, DIF detection is an important process for operational testing programs. Most operational testing programs conduct DIF analyses to pre-screen new items and ensure the proper functioning of operational items. Items exhibiting DIF are typically eliminated or rewritten in order to ensure that examinees' scores are as accurate and valid as possible.

While many methods exist to detect DIF (e.g., Angoff, 1982; Dorans & Holland, 1992; Jodoin & Gierl, 2001; Swaminathan & Rogers, 1990), relatively few studies have attempted to provide quantitative methods for identifying the item-level features responsible for DIF. Existing studies on this topic (Cho, Lee, & Kingston, 2012; Haag, Heppt, Stanat, Kuhl, & Pant, 2013; Loughran, 2014; Loughran & Skorupski, 2014) have yielded inconsistent results. This inconsistency likely stems at least in part from the lack of a consistent modeling approach across

studies and a reliance on two-stage models. Two-stage modeling strategies fail to appropriately account for error since point estimates from the first stage are used to estimate parameters during the second stage. By failing to consider the uncertainty surrounding the first-stage estimates, the effects estimated during the second stage are more subject to bias.

A consistent, appropriate modeling strategy for identifying specific item-level features and quantifying their relationship to DIF could be used across various operational settings. Additionally, use of an item response theory (IRT) model rather than a regression model, as used in Loughran (2014) and Loughran & Skorupski (2014), would provide a method for DIF detection and explanation better suited to operational testing companies, since most testing programs use item response models to scale and score responses. Accurate identification of item-level features related to the amount of observed DIF would provide valuable information to inform the test development and item review processes in operational settings. Use of a consistent methodology would allow for item features to be examined across assessments and would support better inferences about which item features are related to DIF. Once identified, these item-level features could be controlled for during item development and screened for during item review processes, saving test developers time and money.

Significance of Study

As an example of item-level features that could be used to predict DIF, consider the issue of the linguistic complexity of items on math assessments. Unnecessary linguistic complexity could introduce construct-irrelevant variance, as reading comprehension skills would be required in addition to math skills in order to solve the problem. Such items may exhibit item bias against English Language Learners (ELLs). The *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014) address this issue as it relates to the comparability of inferences,

pointing out that when ELLs take an English-language assessment not designed to measure language ability, “one may not know whether the test score is, in whole or in part, a measure of the ability to read in the language of the test rather than a measure of the intended construct” (pp. 59-60). Standard 3.2 clearly addresses this concern by stating that

“Test developers are responsible for developing tests that measure the intended construct and for minimizing the potential for tests’ being affected by construct-irrelevant characteristics, such as *linguistic* [emphasis added], communicative, cognitive, cultural, and other characteristics” (AERA, APA, & NCME, 2014, p. 64).

Despite this guidance, test developers continue to include linguistically complex items on mathematics assessments. Consider the following item, which is provided on a state Department of Education website as an example of a selected response item on their math assessment.¹ Though the item ostensibly measures mathematical ability, this question stem consists of 46 words, 29 of which are in prepositional phrases. Prepositional phrases add to cognitive demand by complicating the item’s syntax (Saint-Dizier, 2006). The item has been formatted to show prepositions in bold text and words contained in prepositional phrases in italicized text:

Alexandra conducted a survey **of** *50 candles* made **by** *several different workers* **at** *a candle factory*. She determined that 4% **of** *the candles* had mistakes. Based **on** *the results* **of** *Alexandra’s survey*, what prediction can she make **about** *the next 500 candles* made **by** *these workers* **at** *the factory*?

A. 4 candles will have mistakes.

¹ http://mdk12.msde.maryland.gov/instruction/assessment/sample_grade8_math.html

- B. 20 candles will have mistakes.
- C. 40 candles will have mistakes.
- D. 200 candles will have mistakes.

With 63% of the words in its stem contained within prepositional phrases, this item illustrates how linguistically complex items may pose challenges for ELLs. Studies of similar item-level linguistic features' impact on ELLs (Loughran, 2014; Loughran & Skorupski, 2014) served as the initial motivation for this study. By identifying item-level features – including linguistic features such as prepositional phrases – that explain the amount of DIF observed, test developers can more accurately anticipate and avoid construct-irrelevant variance, thus strengthening the validity argument for scores' use and interpretation.

A hierarchical 2-PL IRT was chosen to accurately identify such item-level features. The following research questions will be investigated in order to assess the viability and stability of the proposed model over a range of contexts:

Research Questions

1. Can an explanatory hierarchical IRT model accurately recover person and item parameters?
2. Does the model accurately recover differences in mean group ability between the focal and reference groups?
3. Does the model accurately recover the relationship between the amount of DIF observed in an item and the presence of DIF-related item-level features?

4. Does the model perform similarly across different simulated conditions which vary the proportion of examinees in the reference group, the strength of the relationship of DIF to the item-level features, and the distribution of the amount of simulated DIF in each item?

Chapter II: Literature Review

Differential Item Functioning

Differential item functioning, or DIF, can be said to occur “when a test item does not have the same relationship to a latent variable across two or more examinee groups” (Embretson & Reise, 2013, p. 251). Detection of DIF is important because the presence of DIF may indicate item bias in the form of construct-irrelevant variance that differentially impacts examinees belonging to different groups (de Ayala, 2013). The presence of such construct-irrelevant variance may disadvantage examinees in one group, and it also weakens the assessment’s validity argument by calling the validity of inferences made from test scores into question.

Many methods exist for detecting DIF. The two most notable methods from classical test theory are the Mantel-Haenszel and logistic regression approaches. Within the framework of IRT models, Zumbo (1999) succinctly describes the assessment of DIF in terms of comparing the item characteristic curves (ICCs) of different groups on a given item. He furthermore clarifies that DIF is commonly conceptualized as a difference in placement of the two ICCs. In an item that does not exhibit DIF, the groups’ ICCs would be superimposed atop each other (de Ayala, 2013). If one or more item parameters differ significantly, then the item can be said to exhibit DIF (Embretson & Reise, 2013).

There are two main types of DIF: uniform and non-uniform. When uniform DIF is present, the item is more difficult for one group throughout the ability continuum, and the

resulting ICCs for each group are parallel when plotted. Non-uniform DIF favors one group for one portion of the ability continuum but another for another portion of the ability continuum. In a graphical representation of the ICCs for each group, uniform DIF would appear as one group's ICC being consistently higher than the other's across the ability continuum. For non-uniform DIF, the ICCs for each group cross where the direction of DIF changes. In sum, uniform DIF represents a difference in an item's difficulty, or b -parameter, across groups. Non-uniform group represents a difference in an item's a -parameter across groups, though differences in the b -parameter may also be present (de Ayala, 2013).

Hierarchical Linear Models

Multilevel modeling is commonly used to account for differences between individuals and/or items that may be due to shared characteristics, and differential item functioning can be modeled directly by using multilevel models (Gelman & Hill, 2007). While the term *multilevel modeling* refers to the larger discipline of models with several levels of data structures, hierarchical linear modeling is the most common type of multilevel analysis in the field of education. Within educational research, the classic example of data best suited for hierarchical modeling is students nested within classrooms; however, other examples of hierarchical data structures abound. For example, a study of worker productivity might examine workers nested within firms, while a study of international economic development might study households nested within countries (Raudenbush & Bryk, 2002, p. 3). The prevalence of hierarchical data structures in real world contexts has motivated the development of statistical modeling strategies to address these data.

Raudenbush and Bryk (2002) identify several benefits of using hierarchical models rather than standard regression methods, including the improved estimation of individual effects and

modeling of cross-level effects. Goldstein (2011) similarly identifies the benefits of hierarchical modeling: 1) statistically efficient estimates of regression coefficients are obtained; 2) the analysis yields correct estimates for standard errors, confidence intervals, and significance tests, and these estimates are generally more conservative due to the variance being partitioned appropriately; 3) covariates at different levels allow researchers to explore the explanatory value of units' characteristics; and 4) correct rankings of individual units can be obtained after adjusting for confounding factors (p. 14). Accurate standard errors are critically important when making claims about effects' significance and strength. If individuals within a sample are nested within groups but the modeling strategy does not account for these relationships, the local independence assumption is violated. Violation of this assumption results in underestimation of standard errors, which in turn leads to incorrect identification of effects that may not actually exist (Raudenbush & Bryk, 2002).

Snijders (2014) traces the development of hierarchical models back to Robinson (1950), where the idea of the ecological fallacy was introduced. This concept described how associations between variables at one level could be mistakenly interpreted as evidence for associations at a different level; to address this issue, statistical models were developed to include separate error terms for each level. Hierarchical modeling literature typically refers to units, or observations, which are grouped at levels (Goldstein, 2011; Raudenbush & Bryk, 2002).

In general, hierarchical linear models adapt the typical ordinary least squares regression model by allowing the slope and/or intercept terms to become random error terms by adding an additional level. If we consider the traditional example of students nested within schools, then a simple model for one school i could be written as

$$y_i = b_0 + b_1x_i + e_i. \tag{1}$$

To generalize this model and describe several schools simultaneously, the intercept b_{0j} and/or the coefficient b_{1j} can become random effects described by the second-level equations

$$b_{0j} = b_0 + u_{0j} \quad \text{and} \quad b_{1j} = b_1 + u_{1j}, \quad (2, 3)$$

where u_{0j} and u_{1j} are random variables possessing their own variance (adapted from Goldstein, 2011, and Raudenbush & Bryk, 2002). This system of equations treats the observed group values as a sample drawn from a larger population of possible group values.

Item Response Theory Models

Though the history of hierarchical modeling begins in the area of linear regression, the proposed model uses an item response theory (IRT) model in an effort to provide a more relevant modeling strategy to operational testing programs. Unlike linear modeling, IRT relies on characterizations of individuals' levels of a latent trait and items' latent properties in order to predict observed responses (de Ayala, 2013, p. 4). IRT models were primarily designed for use with dichotomous data such as that collected from a typical multiple-choice assessment. Perhaps as a result, IRT modeling is especially prevalent in ability testing, and it is now used widely in operational testing programs (Emberson & Reise, 2013).

The Rasch model is the simplest form of an IRT model. Within the Rasch model, the natural log of the odds ratio of the probability of success to the probability of failure is conceptualized as the difference between the person's level of the latent trait, or θ , and the item's difficulty, often referred to as the b -parameter. A person i 's response to item j can thus be expressed as

$$P(x_{ij} = 1 | \theta_i, b_j) = \frac{e^{(\theta_i - b_j)}}{1 + e^{(\theta_i - b_j)}} \quad (4)$$

(adapted from de Ayala, 2013, and Embretson & Reise, 2013). This equation represents a transformed version of the model that gives the dependent variable prediction in terms of a probability rather than as log odds.

The Rasch model is commonly extended to include parameters for item discrimination (the 2-PL model) and guessing behavior (the 3-PL model). In the 2-PL model, the parameter for item discrimination (a) is added to model the discriminating behavior of each item, which is related to the biserial correlation between responses on that item and total scores on the assessment (Embretson & Reise, 2013). Modeling item discrimination indicates a philosophical shift in the objective of the analysis. While Rasch modeling focuses on constructing an instrument consistent with the constraints of the Rasch model (Andrich, 1988, as cited in de Ayala, 2013), the 2-PL model is concerned with accurately modeling the data (de Ayala, 2013, p. 99).

The 3-PL model further extends the IRT framework to include a lower asymptote, or c -parameter, to reflect guessing behavior on items. This parameter can be said to represent “the probability of a response of 1 on an item due to chance alone” (de Ayala, 2013, p. 124). Though this parameter was originally referred to as the guessing parameter, it is now considered to be a pseudo-guessing parameter due to the fact that its value is typically lower than what would be predicted by random chance alone (de Ayala, 2013, p. 126). The 3-PL model can be expressed as

$$(x_{ij} = 1 | \theta_i, b_j, a_j, c_j) = c_j + (1 - c_j) \frac{e^{a_j(\theta_i - b_j)}}{1 + e^{a_j(\theta_i - b_j)}} \quad (5)$$

As de Ayala (2013) notes, the inclusion of this parameter assumes that a person’s likelihood of guessing is equal regardless of his or her quantity of the latent trait. Additionally, non-zero c -parameters lower individuals’ ability estimates (Wainer, 1983) and reduce the amount

of item information (de Ayala, 2013). These issues, combined with the more frequent use of 2-PL models in operational testing programs, informed the use of a 2-PL model in this study.

IRT research generally does not overlap with the research on linear modeling; indeed, two commonly used textbooks cite IRT's theoretical differences from linear modeling during their definitions (de Ayala, 2013; Embretson & Reise, 2013). However, item response models have been shown to be special cases of generalized linear or nonlinear mixed models (GLMMs or NLMMs) (Kamata, 2001; McCulloch & Searle, 2001). Within this framework, item responses can be understood as repeated observations from the same respondent; in keeping with this conceptualization, item response data must therefore be treated as nested within persons, since the assumption of local independence does not hold (De Boeck & Wilson, 2004a; Ravand, 2015).

Explanatory IRT Models

With the understanding that IRT models are a special case of GLMMs utilizing a logit link function (De Boeck & Wilson, 2004a), it follows that hierarchical IRT models are closely related to hierarchical linear models. Hierarchical IRT models can be used to simultaneously estimate individual differences in ability while also exploring systematic effects that could explain the observed data. De Boeck & Wilson (2004a) refer to this fusion of perspectives as explanatory measurement. Within the item response theory literature, efforts to explain the presence and amount of DIF have relied on such explanatory IRT models, in which properties of assessment items and/or respondents are used to explain responses (De Boeck & Wilson, 2015). Such explanatory models can be parameterized to include a variety of predictors.

Models are typically designed to answer the question of how well individual examinees perform on an exam – that is, how much of the latent construct they can be inferred to possess.

Explanatory models also address a second research question, which is how item responses may be explained by other variables relating to person- or item-level features (De Boeck & Wilson, 2004a). This shift in focus extends the level of interpretation for the data and model to broader inferences about external variables. This can be described as a shift from a measurement approach to an explanatory approach (De Boeck & Wilson, 2004a).

Depending on the construction of the model, both of these approaches may be answered with a single analysis. Explanatory models can be used to investigate person and/or item-level effects. De Boeck & Wilson (2004c) note that the predictors can be:

“(a) characteristics of items, of persons, and or combinations of persons and items; they can be (b) observed or latent (of either items or persons); and they can be (c) latent continuous or latent categorical” (p. vii).

They go on to identify four different types of modeling approaches, which are defined as a function of the predictors examined within the model. Table 1 shows how properties of item and person predictors can be included or excluded within different types of models.

Table 1. Models as a function of the predictors.

Item predictors	Person predictors	
	Absence of properties	Inclusion of (person) properties
Absence of properties	doubly descriptive	person explanatory
Inclusion of (item) properties	item explanatory	doubly explanatory

Note: Adapted from De Boeck & Wilson (2004b), p. 47.

De Boeck & Wilson (2004a) furthermore describe explanatory models as addressing two contexts: repeated observations (also called within-subjects design) and longitudinal design.

Observations of responses to items within an instrument or assessment, such as are collected during a traditional achievement test, can be considered under the within-subjects context; since each item response constitutes a single observation, an assessment can be conceptually considered a set of repeated observations.

Applications of Explanatory IRT Models

Explanatory IRT models have been applied to real-world data in several publications. Fox (2004) used multilevel IRT models to account for nested data and estimate school-level effects from three different educational assessments. Each study examined a measure of student achievement as the outcome variable and included student and/or school-level characteristics as explanatory variables. Multilevel IRT models were contrasted with traditional sum-score models; the author concluded that multilevel modeling provides more accurate and precise estimates that are more appropriate when rating school effectiveness.

Stevenson, Hickendorff, Resing, Heiser, and de Boeck (2013) used explanatory IRT models to investigate sources of individual differences in children's ability and improvement after training on a test of analogical reasoning. They used Rasch modeling to compare a null model, a model including a session parameter, and a model with additional person-level predictors. The authors found that the best fitting model showed significant fixed effects for session, working memory capacity, age-group, and prior math achievement as well as significant interactions between session and training-type and between session and math achievement; this result suggests that explanatory hierarchical IRT models can more accurately assess sources of individual changes in ability after accounting for confounding factors such as session effects, age group, and prior achievement.

Hartig, Frey, Nold, & Klieme (2012) used two-stage explanatory IRT models in order to predict item difficulty from task characteristics. The authors compared a two-stage 1-PL IRT model, the Linear Logistic Test Model (LLTM; Fischer, 1973, 1977), and a generalization of the LLTM that included a random item effect to allow for residual variation of item difficulties (LLTM + e). Data examined were collected from a German examination of German and English language proficiency; task characteristics included the need for global comprehension of a text and text complexity (where text complexity is defined in terms of vocabulary and grammatical structures). The two-stage 1-PL IRT model yielded the best fit, the LLTM + e fit similarly well, and the LLTM had the poorest fit. The authors found that the need for global comprehension was most predictive of item difficulty. They additionally used the estimated item difficulties to define cut points between proficiency levels, with the proficiency levels defined in terms of the examined task characteristics. This study illustrates how information gathered from an explanatory model could be used to improve item writing and inform the selection of cut points for proficiency levels.

De Boeck's review of random item Rasch models (2008) compares the fixed persons–fixed items (FPFI), random persons–fixed items (RPFI), fixed persons–random items (FPRI), and random persons-random items (RPRI) Rasch models. De Boeck recommends the use of random effects for persons for studies desiring to generalize items' measurements over persons (for example, when building an item bank for computer-adaptive testing) or to explain person-level variation by way of external person covariates. The random items approach is recommended for studies desiring to generalize person measurements over items or to explain item difficulties by way of external item covariates. Despite the increased difficulty in estimation due to the crossed random effects, the author concludes that random item profile models, which

are RPRI models for both groups with random difficulties estimated for each group, are preferable for DIF detection and identification. Such models allow DIF to be described with a bivariate or multivariate distribution rather than a discrete event; i.e., all items have some amount of DIF rather than items being identified simply as having or not having DIF. He further suggests that a robust regression can be used to flag items as having DIF. In this procedure, the focal group's item difficulties are regressed on the reference group's difficulties, and then a robust confidence interval is determined for the distance to the regression line. Items exceeding this confidence interval are considered DIF items.

Randall, Cheong, and Engelhard (2011) used a many-facet Rasch model (MFRM), which is an explanatory IRT model, as well as a more traditional hierarchical generalized linear model (HGLM) to model the effects of differing test conditions on assessment performance for students with and without disabilities. Test conditions included the presence or absence of a resource guide and/or calculator. In their modeling strategy, DIF could be interpreted as occurring where a significant cross-level interaction occurs between item difficulties, group membership, and/or test conditions. The two models performed similarly; the authors noted that the MFRM has the benefit of providing fit statistics for all of the facets, but it is also limited by the restrictions associated with being a fixed effect model.

Explanation of DIF Using Item-Level Features

Several studies have used multiple regression to explain DIF using item-level features. A study by Cho, Lee, and Kingston (2012) focused on predicting item difficulty and the presence of DIF from item-level features. Two datasets of math assessment results for third through eighth graders were examined. The first dataset was used to ascertain whether item features were associated with item difficulty, item discrimination, or DIF. The second dataset was used to

examine whether observed DIF was related to students' accommodation status, gender, race, disability, or overall ability. While item-level features were related to item difficulty, they did not predict the presence of DIF. The authors used a 2-PL model to obtain a - and b -parameter estimates, then conducted ANOVAs to determine whether groups according to item type differed in their estimates. They then analyzed items for the presence of DIF and re-fit a two-group IRT model with constraints for non-DIF items to be equal; group classifications were then examined with follow-up tests. Results found that item types and features were associated with item difficulty, but the effect differed by grade level. No consistent explanation for DIF was found.

Haag, Heppt, Stanat, Kuhl, & Pant (2013) examined results from a math assessment administered to third-grade German students. Their analysis sought to uncover which academic language features were associated with DIF for second language learners as well as the effects of the language features on DIF. The authors used a unidimensional IRT model to scale the data and calculate DIF, and they then applied multiple regressions with follow-up commonality analyses to determine the unique contributions of each predictor to explaining the dependent variables. The authors found that the item text length and the number of noun phrases were significant predictors of DIF, general academic vocabulary was marginally significant, and the number of mathematics-specific vocabulary words was not predictive of DIF.

In a similar study, Loughran (2014) examined data from math assessments for American fourth and eighth graders. In a two-stage process, the author used logistic regression to detect DIF; the resulting DIF statistics were regressed on the various linguistic complexity features using stepwise regression. While results indicated that some item-level features predicted DIF, results were not consistent across grade levels. At grade 4, adjective clauses and multi-meaning words were found to be associated with DIF; at grade 8, adjective clauses and whether an item

was primarily schematic were found to be associated with DIF. In addition to inconsistency across grade levels, results were not consistent with Haag, Heppt, Stanat, Kuhl, & Pant (2013), though both studies examined the effect of item-level features related to linguistic complexity. Different item features were identified as predicting DIF. In both studies, the true relationships between item features and the amount of DIF remained unknown, so parameter recovery could not be evaluated.

Results from these studies led to two follow-up studies evaluating the practicality and parameter recovery of two-stage and simultaneous logistic regression methods (Brussow, Skorupski, & Loughran, 2015; Skorupski, Brussow, & Loughran, 2016). Results from these studies indicated that the simultaneous model recovered items' DIF parameters and the R^2 value indicating the relationship between DIF and item-level features more accurately than the corresponding two-stage method. The present study attempts to improve on these modeling strategies by introducing a model for explaining DIF using a coefficient and grouping variable in a hierarchical IRT model. A hierarchical IRT model eliminates the need for logistic regression, resulting in a more parsimonious model, and a simultaneous rather than two-stage estimation process allows the parameters' error to be considered when evaluating the relationship between DIF and item-level features. Additionally, use of a latent trait model rather than a procedure with sum score as the conditioning variable is more directly applicable to operational programs using latent trait models for scaling.

Pilot Study

Methods.

In order to test the viability of this modeling strategy, a small-scale study was conducted across four conditions. The data simulation procedure was similar to the process outlined in the

Methods chapter of this manuscript. Differences included the number of test items simulated (40 items instead of 60) and the method for sampling the amount of DIF for a given item. Within each dataset, two of the 40 test items (5%) were simulated to have meaningful uniform DIF in favor of the reference group: $D \sim N(1, 0.1^2)$. The remaining 38 items were simulated to have random, negligible DIF, which could favor either the reference or the focal group: $D \sim N(0, 0.1^2)$.

The variables manipulated to form the four conditions included the proportion of examinees in the reference group and the correlation between the amount of DIF and item-level features, but there were only two levels of each of these variables. Additionally, due to the differences in the method used to sample items' amount of DIF, the variables manipulated to form conditions did not include the parameters related to the mixture distribution used to simulate D -parameters in the current study. The four conditions examined in the pilot study are listed in Table 2.

Table 2. Conditions examined in the previous study

Condition	Correlation between DIF and item-level features (ρ)	Proportion of examinees in reference group
1	0.4	0.50
2	0.8	0.50
3	0.4	0.90
4	0.8	0.90

The pilot study was also conducted to determine whether the posterior distributions would be more appropriately summarized using the mean or median to obtain a point estimate for each parameter.

Results.

The model accurately recovered item- and person-parameters; the simulated and estimated values were very highly correlated. The model recovered the focal and reference group means with relatively little bias, and bias was centered around zero. Additionally, results suggested that the mean is the appropriate statistic to use when summarizing parameters’ posterior distributions. Using mean values resulted in better parameter recovery, especially for the relationship between item-level features and the estimated amount of DIF. Mean values for parameter recovery statistics obtained from summarizing the posterior distributions according to their means are shown in Table 3.

Table 3. Parameter recovery statistics: means derived from parameter means over replications

True DIF corr.	Prop. in reference group	<i>a</i> -param corr.	<i>b</i> -param corr.	<i>D</i> -param corr.	θ corr.	True – est. focal mean	True – est. reference mean
0.4	0.5	0.973	0.995	0.876	0.959	0.018	0.000
0.4	0.9	0.965	0.996	0.763	0.966	-0.009	-0.001
0.8	0.5	0.967	0.995	0.917	0.959	0.012	-0.001
0.8	0.9	0.971	0.996	0.869	0.966	0.019	-0.007

R^2 values derived by summarizing posterior distributions according to their medians tended to exhibit negative bias, especially in conditions where the true correlation and therefore the expected R^2 value was larger. Due to this finding, the posterior means were recommended for obtaining estimates. Mean values for R^2 recovery obtained from summarizing the posterior distributions according to their means and medians are shown in Table 4.

Table 4. R^2 recovery: means derived from parameter means and medians over replications

True DIF corr.	Prop. in reference group	Parameter means used: True – est. R^2	Parameter medians used: True – est. R^2
0.4	0.5	0.073	-0.068
0.4	0.9	0.066	-0.050

0.8	0.5	-0.009	-0.244
0.8	0.9	-0.004	-0.248

Conclusions.

Results of the pilot study indicated that the hierarchical IRT model is a viable method for accurately recovering parameter values, estimating the amount of DIF exhibited by items, and quantifying the relationship between the estimated amount of DIF and identified item-level features. Based on the results of the pilot study, the current study uses the mean of the posterior distribution as the point estimate for parameters. After discussions regarding the methodology, the method for sampling the amount of DIF for items was changed to draw values from a mixture distribution rather than two discrete distributions. Mixture distributions more accurately simulate the distributions of DIF occurring in real-world contexts, which would not be organized into neat clusters around 0 and a given effect size. While evaluating the results of the pilot study, the idea of examining the proportion of estimated parameter values whose credible intervals include the originally simulated parameter value was raised. This method of evaluating the model's recovery of simulated values is included in the full-scale study.

Chapter III: Methods

Simulating Data

In order to explore the feasibility of the proposed model across a variety of conditions that may be experienced in real-world contexts, a simulation study was designed and carried out. Data were generated from a 2-PL IRT model; the 2-PL model was chosen to provide the greatest level of applicability for operational testing programs, many of which use 2-PL models to scale and score assessments. The typical 2-PL model was expanded to include an added grouping

parameter for persons and DIF coefficient for items (G_i and D_j , respectively). The grouping parameter G_i was dichotomously coded to be either 0, indicating membership in the reference group, or 1, indicating membership in the focal group. This dichotomously coded grouping parameter meant that the DIF coefficient D_j was only estimated when a respondent was in the focal group ($G_i = 1$). When estimated, the DIF coefficient represents a shift in the item's difficulty, or b -parameter, for respondents who are members of the focal group. Structuring the model in this way allows the item difficulties to vary across groups; the amount of difference can then be used to detect and quantify the amount of DIF to determine whether an item exhibits a meaningful quantity of DIF. Equation 6 shows how an examinee i 's probability of a correct response to an item j is calculated using a 2-PL IRT model with the added grouping variable G_i and DIF coefficient D_j .

$$P(x_{ij} = 1 | \theta_i, G_i, a_j, b_j, D_j) = \frac{e^{a_j(\theta_i - (b_j + D_j G_i))}}{1 + e^{a_j(\theta_i - (b_j + D_j G_i))}} \quad (6)$$

In this hierarchical model, the DIF coefficient D_j was estimated to have its own mean and variance:

$$D_j \sim N(\hat{D}_j, \sigma_D^2) \quad (7)$$

This formulation illustrates how the DIF coefficient can be explained according to item-level predictors, making this an explanatory model. The mean \hat{D}_j is then explained by a regression equation with the item-level features Z_j as a predictor, as shown in Equation 8:

$$\hat{D}_j = \beta_0 + \beta_1 Z_j \quad (8)$$

This linear regression predicts the estimated amount of DIF for each item from the identified item-level features. The intercept β_0 represent the mean value of DIF when the predictor is 0,

while the coefficient β_1 represents the units of DIF per unit of Z_j . During estimation, the same prior was used for both the intercept and coefficient because Z_j was standardized during data simulation.

The value of Z_j , or the item-level features for each item, was generated using the correlation between the amount of DIF and item-level features (ρ), the z-score of the amount of DIF (z_{D_j}), and a randomly generated error term $e_j \sim N(0, \sqrt{(1 - \rho^2)})$. Equation 9 shows this process.

$$Z_j = \rho z_{D_j} + e_j \quad (9)$$

While Z_j was a scalar for the purposes of this simulation study, it could also be a vector of item-level features in a real-world implementation of this model.

The focal group mean, μ_{focal} , was also estimated as part of the model. By allowing the focal group mean to differ from the reference group mean, the model can accurately capture differences in overall ability between groups. In real-world applications, group differences may be observed between the focal and reference groups; by allowing the group means to vary, the model more accurately explains the variance in group performance and ensures that mean ability differences do not interfere with the estimation of DIF.

For each generated dataset, item responses were simulated for 1,000 examinees taking a 60-item test. Ability (θ) was drawn from a standard normal distribution for examinees in the reference group, $\theta_{reference} \sim N(0, 1)$. Examinees in the focal group were simulated to be half a standard deviation lower, $\theta_{focal} \sim N(-0.5, 1)$. These values were chosen to facilitate comparison of the results with previous research (Skorupski, Brussow, & Loughran, 2016; Brussow, Skorupski, & Loughran, 2015); the initial selection of these values was driven by group differences

observed in Loughran and Skorupski (2014). This approach also draws from de Ayala, Kim, & Stapleton’s (2002) conceptualization of data as deriving from a mixture distribution from multiple latent populations, each with its own underlying scale. The proportion of examinees in the focal and reference groups was another variable explored in the creation of conditions (see Table 5). Previous research by Jodoin & Gierl (2001) reported that power to detect both uniform and non-uniform DIF decreased as the proportions of examinees in each group became more unequal.

The items’ b -parameters were drawn from a standard normal distribution, $b \sim N(0, 1)$; the a -parameters were drawn from a uniform distribution, $a \sim U(0.5, 3.0)$. These values were chosen to emulate commonly observed ranges of parameter values from IRT analyses. Within each dataset, items’ DIF parameters (D) were drawn from a bimodal mixture distribution. The mean and variance of a mixture distribution created from two normal distributions can be determined according to five parameters:

1. The proportion of values from the first distribution (α), also referred to as the “mixture probability,”
2. the mean of the first distribution (μ_1),
3. the variance of the first distribution (σ_1^2),
4. the mean of the second distribution (μ_2), and
5. the variance of the second distribution (σ_2^2).

By varying these parameters, the mean and variance of the mixture distribution can be manipulated. These expected values can be determined by the following equations:

$$\mu_{mixture} = \alpha\mu_1 + (1 - \alpha)\mu_2 \quad (10)$$

$$\sigma_{mixture}^2 = \alpha\sigma_1^2 + (1 - \alpha)\sigma_2^2 + \alpha(1 - \alpha)(\mu_1 - \mu_2)^2 \quad (11)$$

The mean of the first distribution (μ_1) was fixed at 0 in order to simulate an assessment where the majority of items exhibited negligible DIF. The standard deviations of each distribution (σ_1 and σ_2) were set at 0.1 (for a variance of 0.316) in order to ensure sufficient overlap between distributions yet allow for clear clustering of items with true DIF. The mean of the second distribution (μ_2) and the mixture probability (α) were systematically varied to simulate different assessment conditions with varying proportions of items exhibiting meaningful DIF and varying amounts of DIF being exhibited. Means of the second distribution were set to 0.5 and 1.0 in order to simulate medium and large effect sizes of DIF (Jodoin & Gierl, 2001). Density plots of the resulting distributions can be found in Appendix A.

This simulated true amount of DIF was then used to generate a covariate Z_j , which was correlated with the amount of DIF in the desired amount for the given condition (see Table 5). This level of correlation (ρ) was another variable that was varied across conditions. The covariate Z_j was in turn used as an explanatory variable for the amount of estimated DIF (see Equation 2). This approach is designed to simulate construct-irrelevant item-level features in order to explain the amount of DIF. An example of this phenomenon can be found in Loughran and Skorupski (2014), in which items' linguistic features are found to partially explain the amount of DIF detected in items on an operational math assessment. Other item-level features could also be selected for examination based on the characteristics of the assessment, the reference and focal groups, and the available information on item-level features.

With reference to De Boeck & Wilson's table of models as a function of their predictors (see Table 1), this model would be classified as doubly explanatory. Item properties – DIF-

related item-level features – are modeled with random effects. Person properties – membership in either the focal or reference group – are treated as fixed effects.

Conditions

By varying the level of correlation between the amount of DIF and item-level features (ρ), the proportion of examinees assigned to the reference group, the mean of the second distribution in the mixture distribution (μ_2) and the mixture probability (α), 36 conditions were established for this study. Table 5 lists these 36 conditions.

Table 5. Conditions examined in the present study

Condition	Correlation between DIF and item-level features (ρ)	Proportion of examinees in reference group	Mean of the second distribution (μ_2)	Mixture probability (α)
1	0.4	0.5	0.5	0.85
2	0.4	0.5	0.5	0.9
3	0.4	0.5	0.5	0.95
4	0.4	0.5	1	0.85
5	0.4	0.5	1	0.9
6	0.4	0.5	1	0.95
7	0.4	0.9	0.5	0.85
8	0.4	0.9	0.5	0.9
9	0.4	0.9	0.5	0.95
10	0.4	0.9	1	0.85
11	0.4	0.9	1	0.9
12	0.4	0.9	1	0.95
13	0.6	0.5	0.5	0.85
14	0.6	0.5	0.5	0.9
15	0.6	0.5	0.5	0.95
16	0.6	0.5	1	0.85
17	0.6	0.5	1	0.9
18	0.6	0.5	1	0.95
19	0.6	0.9	0.5	0.85
20	0.6	0.9	0.5	0.9
21	0.6	0.9	0.5	0.95
22	0.6	0.9	1	0.85
23	0.6	0.9	1	0.9
24	0.6	0.9	1	0.95
25	0.8	0.5	0.5	0.85
26	0.8	0.5	0.5	0.9

27	0.8	0.5	0.5	0.95
28	0.8	0.5	1	0.85
29	0.8	0.5	1	0.9
30	0.8	0.5	1	0.95
31	0.8	0.9	0.5	0.85
32	0.8	0.9	0.5	0.9
33	0.8	0.9	0.5	0.95
34	0.8	0.9	1	0.85
35	0.8	0.9	1	0.9
36	0.8	0.9	1	0.95

One hundred datasets were simulated for each of the conditions in Table 5, and each dataset was analyzed as follows. *R* code written to perform the data generation process can be found in Appendix M.

Analysis

Data were analyzed using the same 2-PL IRT model used to simulate datasets (see Equations 6 and 7). All analyses were conducted using fully Bayesian estimation in Stan in *R* via the *rstan* package (Stan Development Team, 2017). The *R* and Stan code written to conduct the analysis is available in Appendix M. Two chains with 12,000 total iterations including 5,000 warmup iterations were used for each analysis. Priors were specified as follows in Table 6. Priors were selected to match values in Brussow, Skorupski, and Loughran (2015) where possible in order to facilitate comparison of results. Priors were also selected to match the process of data simulation as closely as possible in order to speed up estimation. In a real-world scenario, more diffuse priors would be recommended in order to more comprehensively explore the available parameter space. Additionally, the priors for β_0 and β_1 would typically differ due to differing expected values. Since the simulation used a standardized scale, the priors for these parameters are identical here.

Table 6. Priors used in estimation

Parameter	Prior mean and variance
a -parameter	lognormal (0, 1)
b -parameter	normal (0, 1)
θ	normal (μ_θ , 1)
D -parameter	normal (μ_D , σ_D^2)
μ_{focal}	normal (0, 4)
β_0	normal (0, 1)
β_1	normal (0, 1)
σ_D^2	uniform (0, 10)

After fitting the model, posterior distributions were evaluated for convergence using Gelman and Shirley's \hat{R} statistic (2011). This statistic represents the square root of the mixture variance divided by the average within-chain variance, and it is referred to by the authors as the "potential scale reduction factor" (p. 170). At perfect convergence, the chains will have completely mixed, so the ratio of between to within chain variance should be identical, yielding an \hat{R} value of 1.0. In practice, the authors recommend aiming for \hat{R} values of 1.1 or less for all parameters of interest; this recommendation was used to investigate convergence for replications of this study.

Outcome Variables

Per the findings from the pilot study, posterior distributions were summarized according to their means. Results of analyses within each condition were summarized over replications in order to determine parameter recovery for the model parameters and the sampling distributions of the R^2 statistic from the regression relating the item features to the amount of DIF estimated. The estimated item parameters (a , b , and D), person parameters (θ), and group parameters (μ_{focal})

were compared to their simulated values to determine the correlation between estimated and simulated parameters as well as the distribution of bias in parameter recovery for each condition. To evaluate the model's ability to recover the simulated relationship between the item-level predictors and the amount of DIF, R^2 recovery was also evaluated using these metrics. The proportion of replications whose posterior means fell within a 95% credible interval was also examined to provide another metric for model performance.

Several thresholds for flagging items as possessing DIF were evaluated in terms of decision consistency. The thresholds examined for flagging items with DIF included shifts in item difficulty of 0.5, 0.75, and 1.0. These thresholds were chosen to mirror common DIF thresholds used by operational testing companies. DIF classification results for each of these thresholds were summarized for each condition in terms of the false positive rate, power, and precision.

Technical Considerations for the Practitioner

The hierarchical nature of the model combined with fully Bayesian estimation resulted in a relatively large amount of time required for each replication to run. The cluster computing facility at the University of Kansas was utilized in order to complete the study in a reasonable amount of time. The two sampling chains were run simultaneously using multicore capabilities; with this setup, each replication took 12 hours on average to complete. Since 100 replications were conducted for each condition, this resulted in approximately 1,200 hours of computing time for each of the 36 conditions, which works out to 43,200 hours or 1,800 days of total computing time. Cluster computing allowed multiple conditions to be run simultaneously, but several months were still required to complete the study. Practitioners using this model in real life should allow sufficient time for estimation to be completed.

After preliminary testing, 12,000 steps was chosen as the appropriate chain length needed to attain good convergence with the selected priors. As the Limitations section notes, less informative priors are recommended for use with a real dataset, since the true parameters are not known. However, practitioners should be aware that more diffuse priors will also require more steps to reach convergence, which will in turn increase the amount of time required to estimate this model. The desire to finish this project in a reasonable amount of time informed the selection of the relatively informative priors used for estimation in this study.

Chapter IV: Results and Discussion

Convergence

Convergence of posterior distributions was evaluated using Gelman and Shirley’s \hat{R} statistic (2011). The mean \hat{R} value for each replication was calculated as the mean of the \hat{R} values for each of the estimated parameters. The proportion of replications with mean \hat{R} values less than 1.1 is displayed in Table 7.

Table 7. Proportion of \hat{R} values less than 1.1 for each condition

ρ	Proportion reference group	μ_2	α	Proportion replications whose mean $\hat{R} \leq 1.1$
0.4	0.5	0.5	0.85	0.86
0.4	0.5	0.5	0.90	0.88
0.4	0.5	0.5	0.95	0.87
0.4	0.5	1.0	0.85	0.84
0.4	0.5	1.0	0.90	0.89
0.4	0.5	1.0	0.95	0.84
0.4	0.9	0.5	0.85	1
0.4	0.9	0.5	0.90	1
0.4	0.9	0.5	0.95	1
0.4	0.9	1.0	0.85	1
0.4	0.9	1.0	0.90	1
0.4	0.9	1.0	0.95	1
0.6	0.5	0.5	0.85	0.88

0.6	0.5	0.5	0.90	0.78
0.6	0.5	0.5	0.95	0.76
0.6	0.5	1.0	0.85	0.88
0.6	0.5	1.0	0.90	0.86
0.6	0.5	1.0	0.95	0.84
0.6	0.9	0.5	0.85	1
0.6	0.9	0.5	0.90	1
0.6	0.9	0.5	0.95	1
0.6	0.9	1.0	0.85	1
0.6	0.9	1.0	0.90	1
0.6	0.9	1.0	0.95	1
0.8	0.5	0.5	0.85	0.87
0.8	0.5	0.5	0.90	0.93
0.8	0.5	0.5	0.95	0.86
0.8	0.5	1.0	0.85	0.86
0.8	0.5	1.0	0.90	0.84
0.8	0.5	1.0	0.95	0.80
0.8	0.9	0.5	0.85	1
0.8	0.9	0.5	0.90	1
0.8	0.9	0.5	0.95	1
0.8	0.9	1.0	0.85	1
0.8	0.9	1.0	0.90	1
0.8	0.9	1.0	0.95	1

As Table 7 shows, convergence rates were impacted by the proportion of examinees in the reference group. When fewer examinees were in the reference group, it was more difficult to reach convergence. Mean \hat{R} values were less than 1.1 in 100% of replications for conditions whose proportion of examinees in the reference group was 0.9, but when the proportion of examinees in the reference group was 0.5, the mean \hat{R} was less than 1.1 for 76-93% of replications. Histograms of the mean \hat{R} values are shown in Appendix B.

Item and Person Parameter Recovery

Parameter estimates were obtained from the means of the posterior distributions per the findings of the pilot study. The simulation recovered the items’ simulated a -parameters and b -parameters, the respondents’ simulated ability scores, and the means of the reference group (0.0)

and focal group (-0.5) with a high degree of accuracy. Recovery of the D -parameters was less accurate, likely due to the use of relatively informative priors during estimation. Informative priors are appropriate for studies focusing on DIF detection because we can reasonably assume that items were not constructed to exhibit DIF. Bias in the parameter estimates would likely be lessened with less informative priors, though at the cost of increased computational time. Given that a single replication of this simulation took 12-15 hours on average, informative priors were selected in order to ensure the study could be completed in a reasonable amount of time. While this study's purpose did not include evaluating the effects of prior selection, a follow-up study could investigate a variety of priors.

Bias.

Recovery of item and person parameters was evaluated for bias by subtracting the true simulated parameter value from the estimated parameter value. A value of 0 bias would represent perfect recovery of the simulated parameter value. Negative values show that the parameter estimates underestimated the true values; positive values show that the parameter estimates overestimated the true values. Bias amounts are on the same scale as their parameter types. Table 8 shows the mean bias for each type of parameter within each condition across replications; inspection of the bias estimates shows that the amount of bias in parameter recovery did not result in estimates that differ meaningfully from the true values. The possible exception are the bias estimates for the D -parameters; the mean bias exceeds -0.1 for all of the conditions where μ_2 is 1.0 and α is 0.85 except for the condition where ρ is 0.6 and the proportion of examinees in the reference group is 0.5, where the amount of bias is -0.099. This may indicate that this combination of parameters, in which the mean of the second distribution is further away from 0 and the proportion of draws coming from that second distribution is greatest, creates a

situation where the priors' effect of biasing D -parameter estimates toward zero is both more pronounced and applied to a larger proportion of D -parameter values.

Table 8. Mean bias by parameter type across replications

ρ	Proportion reference group	μ_2	α	Mean a -parameter bias	Mean b -parameter bias	Mean D -parameter bias	Mean θ bias
0.4	0.5	0.5	0.85	-0.018	0.007	-0.063	-0.022
0.4	0.5	0.5	0.90	-0.009	0.003	-0.033	-0.050
0.4	0.5	0.5	0.95	-0.019	0.002	0.006	0.029
0.4	0.5	1.0	0.85	-0.017	-0.006	-0.121	-0.026
0.4	0.5	1.0	0.90	-0.017	-0.002	-0.041	-0.119
0.4	0.5	1.0	0.95	-0.017	-0.001	-0.018	-0.046
0.4	0.9	0.5	0.85	-0.014	0.006	-0.046	-0.012
0.4	0.9	0.5	0.90	-0.015	-0.002	-0.025	-0.006
0.4	0.9	0.5	0.95	-0.019	-0.004	0.015	-0.009
0.4	0.9	1.0	0.85	-0.016	0.000	-0.107	-0.011
0.4	0.9	1.0	0.90	-0.012	0.001	-0.063	0.002
0.4	0.9	1.0	0.95	-0.017	-0.003	-0.017	-0.004
0.6	0.5	0.5	0.85	-0.022	0.005	-0.050	0.004
0.6	0.5	0.5	0.90	-0.030	-0.008	0.011	0.004
0.6	0.5	0.5	0.95	-0.016	-0.005	0.020	0.034
0.6	0.5	1.0	0.85	-0.011	-0.004	-0.099	-0.109
0.6	0.5	1.0	0.90	-0.012	-0.002	-0.034	-0.047
0.6	0.5	1.0	0.95	-0.023	-0.007	-0.012	0.036
0.6	0.9	0.5	0.85	-0.016	0.000	-0.036	-0.009
0.6	0.9	0.5	0.90	-0.018	0.002	-0.013	0.005
0.6	0.9	0.5	0.95	-0.019	-0.003	0.003	-0.015
0.6	0.9	1.0	0.85	-0.013	0.000	-0.104	0.003
0.6	0.9	1.0	0.90	-0.013	0.001	-0.067	-0.002
0.6	0.9	1.0	0.95	-0.019	-0.002	-0.018	-0.005
0.8	0.5	0.5	0.85	-0.016	-0.002	-0.013	-0.002
0.8	0.5	0.5	0.90	-0.008	0.007	-0.021	-0.042
0.8	0.5	0.5	0.95	-0.010	0.007	-0.001	-0.022
0.8	0.5	1.0	0.85	-0.015	-0.004	-0.141	-0.080
0.8	0.5	1.0	0.90	-0.024	-0.002	-0.045	0.045
0.8	0.5	1.0	0.95	-0.020	0.000	-0.030	-0.013
0.8	0.9	0.5	0.85	-0.015	-0.002	-0.029	0.004
0.8	0.9	0.5	0.90	-0.013	0.000	-0.014	0.023
0.8	0.9	0.5	0.95	-0.017	-0.006	0.014	-0.003
0.8	0.9	1.0	0.85	-0.023	-0.006	-0.129	-0.015
0.8	0.9	1.0	0.90	-0.013	-0.007	-0.065	-0.018
0.8	0.9	1.0	0.95	-0.013	-0.003	-0.020	-0.002

In order to investigate whether the parameters that were systematically varied across conditions explained a meaningful amount of the variance in bias in parameter recovery, a series of ANOVAs were conducted. The dataset analyzed for these ANOVAs consisted of the mean parameter bias for each replication. Since each of the 36 conditions was replicated 100 times, this yielded a 3,600-row dataset for analysis. Results from each ANOVA were evaluated for statistical significance as well as effect size.

Both η^2 and partial η^2 are reported for each predictor. These effect sizes differ in calculation and interpretation: while η^2 is calculated as the variance attributable to the predictor divided by the total variance, partial η^2 is calculated as the variance attributable to the predictor divided by that same variance plus the error variance. This difference in calculation allows partial η^2 to partial out the influence of other predictors. While the sum of the η^2 results in a given ANOVA can be said to equal the total amount of variance explained in the model, the sum of partial η^2 values can exceed 1, so a similar interpretation is not valid (Pierce, Block, & Aguinis, 2004). The results of the ANOVAs for each of the four parameter types are shown in Table 9, Table 10, Table 11, and Table 12. Bold font is used to denote significant p-values ($p < 0.05$). Effect sizes can be classified as small, medium, or large according to the value of η^2 or partial η^2 . Small effects (exceeding 0.0099) are denoted with italic font, medium effects (exceeding 0.0588) are denoted with bold font, and large effects (exceeding 0.1379) are denoted with bold italic font (Cohen, 1969, as cited in Richardson, 2011).

Table 9. ANOVA results for *a*-parameter bias

Predictor	Df	Sum Sq	Mean Sq	F value	Pr(>F)	eta.sq	eta.sq.part
rho	2	0.003	0.001	0.739	0.478	0.000	0.000
PREF	1	0.001	0.001	0.501	0.479	0.000	0.000
mu	1	0.000	0.000	0.005	0.942	0.000	0.000

alpha	2	0.003	0.001	0.673	0.510	0.000	0.000
rho:PREF	2	0.001	0.001	0.281	0.755	0.000	0.000
rho:mu	2	0.015	0.007	3.952	0.019	0.002	0.002
rho:alpha	4	0.006	0.002	0.861	0.487	0.001	0.001
PREF:mu	1	0.001	0.001	0.398	0.528	0.000	0.000
PREF:alpha	2	0.001	0.001	0.290	0.748	0.000	0.000
mu:alpha	2	0.001	0.000	0.199	0.819	0.000	0.000
rho:PREF:mu	2	0.005	0.003	1.350	0.259	0.001	0.001
rho:PREF:alpha	4	0.003	0.001	0.347	0.846	0.000	0.000
rho:mu:alpha	4	0.013	0.003	1.765	0.133	0.002	0.002
PREF:mu:alpha	2	0.008	0.004	2.246	0.106	0.001	0.001
rho:PREF:mu:alpha	4	0.009	0.002	1.267	0.280	0.001	0.001
Residuals	3564	6.646	0.002				

Note: Statistically significant p-values ($p < 0.05$) are denoted by bold font. Values of η^2 or partial η^2 that exceed the criteria to be considered a small effect (exceeding 0.0099) are denoted with italic font, medium effects (exceeding 0.0588) are denoted with bold font, and large effects (exceeding 0.1379) are denoted with bold italic font (Cohen, 1969, as cited in Richardson, 2011).

Table 10. ANOVA results for *b*-parameter bias

Predictor	Df	Sum Sq	Mean Sq	F value	Pr(>F)	eta.sq	eta.sq.part
rho	2	0.003	0.001	0.816	0.442	0.000	0.000
PREF	1	0.001	0.001	0.505	0.477	0.000	0.000
mu	1	0.007	0.007	4.381	0.036	0.001	0.001
alpha	2	0.001	0.001	0.404	0.668	0.000	0.000
rho:PREF	2	0.010	0.005	3.169	0.042	0.002	0.002
rho:mu	2	0.002	0.001	0.660	0.517	0.000	0.000
rho:alpha	4	0.006	0.002	0.952	0.433	0.001	0.001
PREF:mu	1	0.003	0.003	1.720	0.190	0.000	0.000
PREF:alpha	2	0.002	0.001	0.550	0.577	0.000	0.000
mu:alpha	2	0.003	0.002	1.044	0.352	0.001	0.001
rho:PREF:mu	2	0.001	0.000	0.239	0.787	0.000	0.000
rho:PREF:alpha	4	0.005	0.001	0.734	0.569	0.001	0.001
rho:mu:alpha	4	0.006	0.001	0.916	0.453	0.001	0.001
PREF:mu:alpha	2	0.001	0.000	0.218	0.804	0.000	0.000
rho:PREF:mu:alpha	4	0.005	0.001	0.763	0.549	0.001	0.001
Residuals	3564	5.830	0.002				

Note: Statistically significant p-values ($p < 0.05$) are denoted by bold font. Values of η^2 or partial η^2 that exceed the criteria to be considered a small effect (exceeding 0.0099) are denoted with italic font, medium effects (exceeding 0.0588) are denoted with bold font, and large effects (exceeding 0.1379) are denoted with bold italic font (Cohen, 1969, as cited in Richardson, 2011).

Table 11. ANOVA results for D -parameter bias

Predictor	Df	Sum Sq	Mean Sq	F value	Pr(>F)	eta.sq	eta.sq.part
rho	2	0.061	0.031	1.034	0.356	0.001	0.001
PREF	1	0.003	0.003	0.117	0.732	0.000	0.000
mu	1	2.022	2.022	68.213	0.000	<i>0.018</i>	<i>0.019</i>
alpha	2	3.281	1.641	55.353	0.000	<i>0.029</i>	<i>0.030</i>
rho:PREF	2	0.047	0.023	0.789	0.454	0.000	0.000
rho:mu	2	0.088	0.044	1.478	0.228	0.001	0.001
rho:alpha	4	0.024	0.006	0.203	0.937	0.000	0.000
PREF:mu	1	0.011	0.011	0.357	0.550	0.000	0.000
PREF:alpha	2	0.068	0.034	1.140	0.320	0.001	0.001
mu:alpha	2	0.410	0.205	6.917	0.001	0.004	0.004
rho:PREF:mu	2	0.006	0.003	0.101	0.904	0.000	0.000
rho:PREF:alpha	4	0.035	0.009	0.292	0.883	0.000	0.000
rho:mu:alpha	4	0.147	0.037	1.237	0.293	0.001	0.001
PREF:mu:alpha	2	0.026	0.013	0.440	0.644	0.000	0.000
rho:PREF:mu:alpha	4	0.036	0.009	0.301	0.878	0.000	0.000
Residuals	3564	105.628	0.030				

Note: Statistically significant p-values ($p < 0.05$) are denoted by bold font. Values of η^2 or partial η^2 that exceed the criteria to be considered a small effect (exceeding 0.0099) are denoted with italic font, medium effects (exceeding 0.0588) are denoted with bold font, and large effects (exceeding 0.1379) are denoted with bold italic font (Cohen, 1969, as cited in Richardson, 2011).

Table 12. ANOVA results for θ bias

Predictor	Df	Sum Sq	Mean Sq	F value	Pr(>F)	eta.sq	eta.sq.part
rho	2	0.147	0.074	5.373	0.005	0.003	0.003
PREF	1	0.344	0.344	25.076	0.000	0.006	0.007
mu	1	0.291	0.291	21.215	0.000	0.005	0.006
alpha	2	0.290	0.145	10.590	0.000	0.005	0.006
rho:PREF	2	0.081	0.041	2.967	0.052	0.002	0.002
rho:mu	2	0.055	0.028	2.019	0.133	0.001	0.001
rho:alpha	4	0.434	0.109	7.922	0.000	0.008	0.009
PREF:mu	1	0.193	0.193	14.103	0.000	0.004	0.004
PREF:alpha	2	0.382	0.191	13.931	0.000	0.007	0.008
mu:alpha	2	0.113	0.056	4.117	0.016	0.002	0.002
rho:PREF:mu	2	0.335	0.167	12.219	0.000	0.006	0.007
rho:PREF:alpha	4	0.535	0.134	9.768	0.000	<i>0.010</i>	<i>0.011</i>
rho:mu:alpha	4	0.391	0.098	7.143	0.000	0.007	0.008
PREF:mu:alpha	2	0.162	0.081	5.925	0.003	0.003	0.003
rho:PREF:mu:alpha	4	0.561	0.140	10.232	0.000	<i>0.011</i>	<i>0.011</i>

Residuals	3564	48.834	0.014
-----------	------	--------	-------

Note: Statistically significant p-values ($p < 0.05$) are denoted by bold font. Values of η^2 or partial η^2 that exceed the criteria to be considered a small effect (exceeding 0.0099) are denoted with italic font, medium effects (exceeding 0.0588) are denoted with bold font, and large effects (exceeding 0.1379) are denoted with bold italic font (Cohen, 1969, as cited in Richardson, 2011).

As the tables showing the results of the ANOVAs investigating bias show, the main effects for μ_2 and α constituted a small effect size for D -parameter bias. Two interactions in the bias ANOVA for θ recovery were also found to be small effects: the three-way interaction of ρ , the proportion of examinees in the reference group, and α ; and the four-way interaction between all four systematically varied parameters. The largest η^2 value observed during the ANOVAs investigating bias was for α when examining the bias in D -parameter recovery; this η^2 was 0.029, which can be interpreted to mean that this interaction term explained 2.9% of the observed variance. These results indicate that all of the systematically varied parameters affected the recovery of θ , while D -parameter recovery was only impacted by μ_2 and α , the two parameters governing the mixture distribution from which D -parameters were simulated. These findings, combined with an interpretation of the mean bias for each parameter type for each condition shown in Table 8, may indicate that the relatively informative priors' effect of biasing D -parameter estimates toward 0 was more pronounced when items were simulated to exhibit more DIF ($\mu_2 = 1$) and when fewer items' DIF parameters were drawn from the distribution centered around 0 ($\alpha = 0.85$).

Mean Squared Error.

As another means of evaluating parameter recovery, mean squared error (MSE) was evaluated for each parameter type. A value of 0 MSE would represent perfect recovery of the simulated parameter value. Table 13 shows the mean MSE for each type of parameter within each condition across replications.

Table 13. Mean MSE for each parameter type across replications

ρ	Proportion reference group	μ_2	α	Mean a -parameter MSE	Mean b -parameter MSE	Mean D -parameter MSE	Mean θ MSE
0.4	0.5	0.5	0.85	0.039	0.013	0.052	0.060
0.4	0.5	0.5	0.90	0.038	0.013	0.082	0.091
0.4	0.5	0.5	0.95	0.029	0.012	0.065	0.065
0.4	0.5	1.0	0.85	0.029	0.014	0.078	0.083
0.4	0.5	1.0	0.90	0.039	0.014	0.077	0.106
0.4	0.5	1.0	0.95	0.031	0.013	0.068	0.070
0.4	0.9	0.5	0.85	0.030	0.010	0.026	0.049
0.4	0.9	0.5	0.90	0.031	0.010	0.021	0.049
0.4	0.9	0.5	0.95	0.030	0.010	0.019	0.050
0.4	0.9	1.0	0.85	0.034	0.010	0.054	0.051
0.4	0.9	1.0	0.90	0.031	0.010	0.040	0.048
0.4	0.9	1.0	0.95	0.031	0.009	0.028	0.048
0.6	0.5	0.5	0.85	0.038	0.012	0.075	0.065
0.6	0.5	0.5	0.90	0.031	0.012	0.068	0.059
0.6	0.5	0.5	0.95	0.032	0.012	0.063	0.100
0.6	0.5	1.0	0.85	0.028	0.013	0.068	0.112
0.6	0.5	1.0	0.90	0.032	0.013	0.055	0.064
0.6	0.5	1.0	0.95	0.032	0.013	0.053	0.072
0.6	0.9	0.5	0.85	0.027	0.010	0.024	0.047
0.6	0.9	0.5	0.90	0.032	0.009	0.020	0.048
0.6	0.9	0.5	0.95	0.032	0.010	0.015	0.050
0.6	0.9	1.0	0.85	0.031	0.010	0.049	0.050
0.6	0.9	1.0	0.90	0.035	0.010	0.039	0.049
0.6	0.9	1.0	0.95	0.032	0.009	0.027	0.051
0.8	0.5	0.5	0.85	0.034	0.012	0.067	0.066
0.8	0.5	0.5	0.90	0.052	0.012	0.051	0.062
0.8	0.5	0.5	0.95	0.034	0.012	0.064	0.068
0.8	0.5	1.0	0.85	0.037	0.012	0.088	0.094
0.8	0.5	1.0	0.90	0.033	0.013	0.063	0.076
0.8	0.5	1.0	0.95	0.035	0.013	0.067	0.061
0.8	0.9	0.5	0.85	0.031	0.009	0.019	0.047
0.8	0.9	0.5	0.90	0.028	0.010	0.015	0.047
0.8	0.9	0.5	0.95	0.028	0.009	0.014	0.048
0.8	0.9	1.0	0.85	0.032	0.010	0.049	0.052
0.8	0.9	1.0	0.90	0.032	0.010	0.031	0.051
0.8	0.9	1.0	0.95	0.028	0.010	0.018	0.047

These MSEs were also examined via ANOVAs whose design was identical to the process described for the ANOVAs investigating bias. Results from these ANOVAs are given in Table

14, Table 15, Table 16, and Table 17. As in the tables containing the ANOVA results for bias, bold font is used to denote significant p-values ($p < 0.05$) and effect sizes exceeding the 0.1 criterion to be considered a “small” effect size (Cohen, 1969, as cited in Richardson, 2011).

Table 14. ANOVA results for a -parameter MSEs

Predictor	Df	Sum Sq	Mean	F value	Pr(>F)	eta.sq	eta.sq.part
			Sq				
rho	2	0.002	0.001	0.599	0.549	0.000	0.000
PREF	1	0.013	0.013	7.930	0.005	0.002	0.002
mu	1	0.001	0.001	0.318	0.573	0.000	0.000
alpha	2	0.006	0.003	1.921	0.147	0.001	0.001
rho:PREF	2	0.007	0.003	2.063	0.127	0.001	0.001
rho:mu	2	0.000	0.000	0.086	0.918	0.000	0.000
rho:alpha	4	0.003	0.001	0.510	0.729	0.001	0.001
PREF:mu	1	0.006	0.006	3.708	0.054	0.001	0.001
PREF:alpha	2	0.003	0.001	0.786	0.456	0.000	0.000
mu:alpha	2	0.001	0.000	0.272	0.762	0.000	0.000
rho:PREF:mu	2	0.000	0.000	0.127	0.881	0.000	0.000
rho:PREF:alpha	4	0.007	0.002	1.018	0.397	0.001	0.001
rho:mu:alpha	4	0.007	0.002	1.040	0.385	0.001	0.001
PREF:mu:alpha	2	0.004	0.002	1.176	0.308	0.001	0.001
rho:PREF:mu:alpha	4	0.013	0.003	1.976	0.095	0.002	0.002
Residuals	3564	5.749	0.002				

Note: Statistically significant p-values ($p < 0.05$) are denoted by bold font. Values of η^2 or partial η^2 that exceed the criteria to be considered a small effect (exceeding 0.0099) are denoted with italic font, medium effects (exceeding 0.0588) are denoted with bold font, and large effects (exceeding 0.1379) are denoted with bold italic font (Cohen, 1969, as cited in Richardson, 2011).

Table 15. ANOVA results for b -parameter MSEs

Predictor	Df	Sum Sq	Mean	F value	Pr(>F)	eta.sq	eta.sq.part
			Sq				
rho	2	0.000	0.000	4.680	0.009	0.002	0.003
PREF	1	0.007	0.007	415.896	0.000	0.103	0.104
mu	1	0.000	0.000	8.960	0.003	0.002	0.003
alpha	2	0.000	0.000	3.350	0.035	0.002	0.002
rho:PREF	2	0.000	0.000	1.079	0.340	0.001	0.001
rho:mu	2	0.000	0.000	0.539	0.583	0.000	0.000
rho:alpha	4	0.000	0.000	0.787	0.534	0.001	0.001
PREF:mu	1	0.000	0.000	12.236	0.000	0.003	0.003
PREF:alpha	2	0.000	0.000	0.357	0.700	0.000	0.000

mu:alpha	2	0.000	0.000	0.367	0.693	0.000	0.000
rho:PREF:mu	2	0.000	0.000	1.838	0.159	0.001	0.001
rho:PREF:alpha	4	0.000	0.000	0.647	0.629	0.001	0.001
rho:mu:alpha	4	0.000	0.000	0.294	0.882	0.000	0.000
PREF:mu:alpha	2	0.000	0.000	2.266	0.104	0.001	0.001
rho:PREF:mu:alpha	4	0.000	0.000	0.791	0.531	0.001	0.001
Residuals	3564	0.060	0.000				

Note: Statistically significant p-values ($p < 0.05$) are denoted by bold font. Values of η^2 or partial η^2 that exceed the criteria to be considered a small effect (exceeding 0.0099) are denoted with italic font, medium effects (exceeding 0.0588) are denoted with bold font, and large effects (exceeding 0.1379) are denoted with bold italic font (Cohen, 1969, as cited in Richardson, 2011).

Table 16. ANOVA results for D -parameter MSEs

Predictor	Df	Sum Sq	Mean Sq	F value	Pr(>F)	eta.sq	eta.sq.part
rho	2	0.020	0.010	1.988	0.137	0.001	0.001
PREF	1	1.347	1.347	273.142	0.000	0.070	0.071
mu	1	0.104	0.104	21.170	0.000	0.005	0.006
alpha	2	0.091	0.045	9.209	0.000	0.005	0.005
rho:PREF	2	0.010	0.005	0.965	0.381	0.000	0.001
rho:mu	2	0.019	0.009	1.882	0.152	0.001	0.001
rho:alpha	4	0.034	0.009	1.739	0.139	0.002	0.002
PREF:mu	1	0.046	0.046	9.416	0.002	0.002	0.003
PREF:alpha	2	0.012	0.006	1.251	0.286	0.001	0.001
mu:alpha	2	0.046	0.023	4.688	0.009	0.002	0.003
rho:PREF:mu	2	0.024	0.012	2.456	0.086	0.001	0.001
rho:PREF:alpha	4	0.033	0.008	1.663	0.156	0.002	0.002
rho:mu:alpha	4	0.010	0.002	0.482	0.749	0.000	0.001
PREF:mu:alpha	2	0.004	0.002	0.395	0.673	0.000	0.000
rho:PREF:mu:alpha	4	0.005	0.001	0.245	0.913	0.000	0.000
Residuals	3564	17.574	0.005				

Note: Statistically significant p-values ($p < 0.05$) are denoted by bold font. Values of η^2 or partial η^2 that exceed the criteria to be considered a small effect (exceeding 0.0099) are denoted with italic font, medium effects (exceeding 0.0588) are denoted with bold font, and large effects (exceeding 0.1379) are denoted with bold italic font (Cohen, 1969, as cited in Richardson, 2011).

Table 17. ANOVA results for θ MSEs.

Predictor	Df	Sum Sq	Mean Sq	F value	Pr(>F)	eta.sq	eta.sq.part
rho	2	0.014	0.007	3.564	0.028	0.002	0.002
PREF	1	0.671	0.671	351.188	0.000	0.084	0.090
mu	1	0.035	0.035	18.560	0.000	0.004	0.005

alpha	2	0.010	0.005	2.521	0.080	0.001	0.001
rho:PREF	2	0.011	0.005	2.801	0.061	0.001	0.002
rho:mu	2	0.001	0.000	0.252	0.777	0.000	0.000
rho:alpha	4	0.108	0.027	14.199	0.000	<i>0.014</i>	<i>0.016</i>
PREF:mu	1	0.021	0.021	11.162	0.001	0.003	0.003
PREF:alpha	2	0.007	0.004	1.950	0.142	0.001	0.001
mu:alpha	2	0.084	0.042	22.083	0.000	<i>0.011</i>	<i>0.012</i>
rho:PREF:mu	2	0.003	0.001	0.734	0.480	0.000	0.000
rho:PREF:alpha	4	0.107	0.027	14.056	0.000	<i>0.013</i>	<i>0.016</i>
rho:mu:alpha	4	0.020	0.005	2.605	0.034	0.002	0.003
PREF:mu:alpha	2	0.055	0.028	14.456	0.000	0.007	0.008
rho:PREF:mu:alpha	4	0.025	0.006	3.256	0.011	0.003	0.004
Residuals	3564	6.806	0.002				

Note: Statistically significant p-values ($p < 0.05$) are denoted by bold font. Values of η^2 or partial η^2 that exceed the criteria to be considered a small effect (exceeding 0.0099) are denoted with italic font, medium effects (exceeding 0.0588) are denoted with bold font, and large effects (exceeding 0.1379) are denoted with bold italic font (Cohen, 1969, as cited in Richardson, 2011).

The ANOVAs examining MSEs yielded several values of η^2 that met the criteria to be considered small or medium effect sizes. For the *b*-parameter ANOVA, the proportion of students in the reference group was found to be a meaningful predictor of the variance in the MSEs of the parameter estimates, with an η^2 value of 0.103, a medium effect. While the other three main effects and the interaction between the proportion of students in the reference group and the value of μ_2 were also statistically significant, their effect sizes were not meaningful. The sum of all η^2 values for this model was 0.118, which can be interpreted to mean that all predictors explained 11.8% of the total variance. A similar pattern can be observed in the *D*-parameter ANOVA results: the only meaningful predictor was the proportion of students in the reference group, which had a medium effect size with an η^2 value of 0.070.

For the ANOVA examining the MSEs of the θ estimates, ten of the predictors in the ANOVA were statistically significant. The statistically significant predictors included three of the four main effects (α was not statistically significant), three of the six two-way interactions,

three of the four three-way interactions, and the four-way interaction. The largest single η^2 value was for the proportion of students in the reference group, which had an η^2 of 0.084, a medium effect. Additionally, two of the two-way interactions constituted small effects: the interaction between ρ and α and the interaction between μ_2 and α . Finally, the three-way interaction between ρ , the proportion of examinees in the reference group, and α was also a small effect. The sum of all η^2 values for this model was 0.147. This summed effect size meets the criteria to be considered a large effect size, and it can be interpreted to mean that all predictors explained 14.7% of the total variance.

Taken together, the ANOVAs that sought to explain the variance in parameter estimates' MSEs in terms of the systematically varied parameters used to construct conditions indicate that those systematically varied parameters are meaningful in explaining the variance in the observed MSEs. The only meaningful main effect was the proportion of students in the reference group, which exhibited a medium effect size when conducting ANOVAs on the MSEs for the b -parameters, D -parameters, and θ . In the analysis of the MSEs for θ , the summed values of η^2 met the criteria to be considered a large effect size, indicating that the parameters that were systematically varied to create conditions had a large impact on the accuracy of θ recovery.

Correlation.

Simulated and recovered parameter values should be highly correlated when parameter recovery is accurate. Table 18 shows the average correlations for each type of parameter within each condition across replications. The relatively low correlations of D -parameters can be explained by two factors: negative bias in parameter recovery due to the priors selected, and the clustering pattern observed in these parameter values resulting from the way they were sampled from the mixture distribution. Both of these effects would likely be ameliorated by the use of less

informative priors during estimation. The correlations between simulated and estimated a , b , D , and θ parameters can be visualized with scatterplots, which are available in Appendices C, D, E, and F.

Table 18. Correlation between simulated and recovered item and person parameters

ρ	Proportion reference group	μ_2	α	a -parameter correlation	b -parameter correlation	D -parameter correlation	θ correlation
0.4	0.5	0.5	0.85	0.970	0.995	0.860	0.973
0.4	0.5	0.5	0.90	0.970	0.995	0.827	0.969
0.4	0.5	0.5	0.95	0.974	0.995	0.788	0.971
0.4	0.5	1.0	0.85	0.976	0.995	0.936	0.970
0.4	0.5	1.0	0.90	0.970	0.995	0.908	0.970
0.4	0.5	1.0	0.95	0.974	0.995	0.867	0.971
0.4	0.9	0.5	0.85	0.975	0.996	0.735	0.977
0.4	0.9	0.5	0.90	0.974	0.996	0.686	0.977
0.4	0.9	0.5	0.95	0.975	0.996	0.616	0.977
0.4	0.9	1.0	0.85	0.973	0.996	0.851	0.976
0.4	0.9	1.0	0.90	0.974	0.996	0.812	0.977
0.4	0.9	1.0	0.95	0.975	0.996	0.725	0.977
0.6	0.5	0.5	0.85	0.970	0.995	0.874	0.970
0.6	0.5	0.5	0.90	0.974	0.995	0.853	0.971
0.6	0.5	0.5	0.95	0.974	0.995	0.815	0.971
0.6	0.5	1.0	0.85	0.975	0.995	0.938	0.971
0.6	0.5	1.0	0.90	0.974	0.995	0.921	0.972
0.6	0.5	1.0	0.95	0.974	0.995	0.877	0.972
0.6	0.9	0.5	0.85	0.976	0.996	0.775	0.977
0.6	0.9	0.5	0.90	0.973	0.996	0.758	0.977
0.6	0.9	0.5	0.95	0.974	0.996	0.708	0.977
0.6	0.9	1.0	0.85	0.975	0.996	0.874	0.976
0.6	0.9	1.0	0.90	0.973	0.996	0.843	0.976
0.6	0.9	1.0	0.95	0.973	0.996	0.785	0.977
0.8	0.5	0.5	0.85	0.972	0.995	0.907	0.970
0.8	0.5	0.5	0.90	0.965	0.995	0.897	0.972
0.8	0.5	0.5	0.95	0.973	0.996	0.876	0.971
0.8	0.5	1.0	0.85	0.970	0.995	0.952	0.969
0.8	0.5	1.0	0.90	0.973	0.995	0.937	0.971
0.8	0.5	1.0	0.95	0.972	0.995	0.904	0.971
0.8	0.9	0.5	0.85	0.974	0.996	0.858	0.977
0.8	0.9	0.5	0.90	0.976	0.996	0.850	0.977
0.8	0.9	0.5	0.95	0.976	0.996	0.829	0.977
0.8	0.9	1.0	0.85	0.974	0.996	0.908	0.976
0.8	0.9	1.0	0.90	0.974	0.996	0.891	0.976

0.8	0.9	1.0	0.95	0.977	0.996	0.864	0.977
-----	-----	-----	------	-------	-------	-------	-------

Overall, these high levels of correlation indicate that the proposed model was able to accurately recover the simulated parameters.

Recovery of group means.

In addition to the item parameters and θ , the model estimated the focal and reference group means. During data simulation, individuals’ ability parameters were drawn from a distribution according to their group membership; these distributions were $\theta_{reference} \sim N(0, 1)$ and $\theta_{focal} \sim N(-0.5, 1)$. Reference group mean recovery was consistently accurate, with an overall mean bias across replications and conditions of 0.001. Focal group mean recovery was slightly less accurate, with an overall mean bias across replications and conditions of -0.049. Mean bias for each group mean for each condition are given in Table 19.

Table 19. Mean bias for group means for each condition across replications

ρ	Proportion reference group	μ_2	α	Mean reference group mean bias	Mean focal group mean bias
0.4	0.5	0.5	0.85	-0.065	-0.001
0.4	0.5	0.5	0.90	-0.031	0
0.4	0.5	0.5	0.95	-0.003	0
0.4	0.5	1.0	0.85	-0.130	-0.001
0.4	0.5	1.0	0.90	-0.058	0.002
0.4	0.5	1.0	0.95	-0.033	-0.001
0.4	0.9	0.5	0.85	-0.049	0.015
0.4	0.9	0.5	0.90	-0.017	0.014
0.4	0.9	0.5	0.95	0.002	0.003
0.4	0.9	1.0	0.85	-0.113	-0.001
0.4	0.9	1.0	0.90	-0.064	-0.010
0.4	0.9	1.0	0.95	-0.038	0.007
0.6	0.5	0.5	0.85	-0.066	-0.001
0.6	0.5	0.5	0.90	-0.008	0.002
0.6	0.5	0.5	0.95	0.013	-0.001
0.6	0.5	1.0	0.85	-0.111	0.002
0.6	0.5	1.0	0.90	-0.051	-0.002

0.6	0.5	1.0	0.95	-0.033	-0.001
0.6	0.9	0.5	0.85	-0.042	0.013
0.6	0.9	0.5	0.90	-0.031	-0.006
0.6	0.9	0.5	0.95	-0.012	-0.011
0.6	0.9	1.0	0.85	-0.125	0
0.6	0.9	1.0	0.90	-0.059	0.004
0.6	0.9	1.0	0.95	-0.019	0.002
0.8	0.5	0.5	0.85	-0.016	0
0.8	0.5	0.5	0.90	-0.021	0.001
0.8	0.5	0.5	0.95	0.004	0.001
0.8	0.5	1.0	0.85	-0.165	0.001
0.8	0.5	1.0	0.90	-0.058	0
0.8	0.5	1.0	0.95	-0.030	0
0.8	0.9	0.5	0.85	-0.070	-0.023
0.8	0.9	0.5	0.90	-0.013	0.001
0.8	0.9	0.5	0.95	0.001	0.025
0.8	0.9	1.0	0.85	-0.155	0.017
0.8	0.9	1.0	0.90	-0.082	0.006
0.8	0.9	1.0	0.95	-0.036	-0.014

In addition to overall mean bias for each condition, the distribution of bias for replications within each condition can also be considered. Histograms showing these distributions for the focal group means are given in Appendix G; similar histograms for the reference group means are given in Appendix H. The variance of the distribution of bias observed in recovery of group means primarily varied according to the proportion of examinees in the reference group. The reference group mean had less variance in the amount of bias per replication overall, and there was less variance in the amount of bias across replications when the proportion of examinees in the reference group was 0.5. The focal group mean exhibited more bias and more variance in the amount of bias across replications, and there was less variance in the amount of bias across replications when the proportion of examinees in the reference group was 0.9. The number of possible explanations for these patterns in the variance of the amount of bias makes it difficult to offer a compelling explanation. Future research is needed to better understand these relationships.

Recovery of the Relationship between Item-Level Features and DIF

Recovery of the R^2 values that summarized the relationship between the DIF parameters and the item-level predictors was less accurate than recovery of the level 1 parameters (a -parameters, b -parameters, D -parameters, θ values, and the reference and focal group means). However, recovered R^2 values were distributed around their true simulated values, as shown in Figure 1. The simulated R^2 values are simply squared values of ρ , the correlation between the item-level features and the amount of DIF. As Figure 1 shows, recovered R^2 values were relatively normally distributed around their true values. The distribution of recovered R^2 values for the condition where $\rho = 0.4$ (and therefore $R^2 = 0.16$) exhibits positive skewness, likely due to the distribution being truncated at a lower bound of 0.

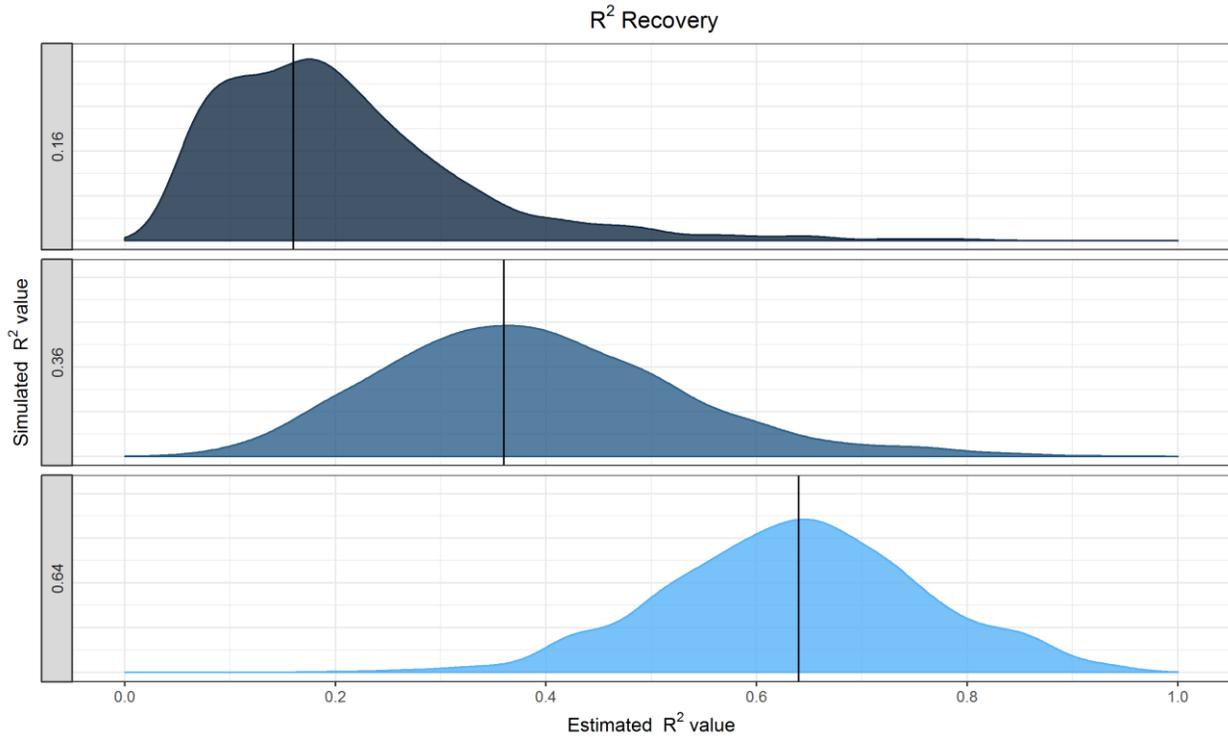


Figure 1. Density plots of R^2 recovery

Recovery of the amount of variance in the amount of DIF explained by the item-level features is a critical outcome for the proposed model. These results indicate adequate recovery of the simulated R^2 values across various values of R^2 . Less informative priors on the D -parameters may yield more accurate parameter recovery for the D -parameters, which would in turn improve recovery of R^2 .

Credible Intervals

The accuracy of parameter recovery was also assessed by examining the proportion of estimated credible intervals that contained the parameters' true values. The mean proportions across replications within each condition are given for the level 1 terms in Table 20 and for the level 2 terms in Table 21.

Table 20. Mean proportion of credible intervals containing the true simulated value, level 1 terms

ρ	Proportion reference group	μ_2	α	a - parameters	b - parameters	D - parameters	θ - parameters
0.4	0.5	0.5	0.85	0.95	0.95	1.00	0.98
0.4	0.5	0.5	0.90	0.95	0.95	1.00	0.98
0.4	0.5	0.5	0.95	0.94	0.96	1.00	0.98
0.4	0.5	1.0	0.85	0.95	0.94	1.00	0.97
0.4	0.5	1.0	0.90	0.94	0.95	1.00	0.98
0.4	0.5	1.0	0.95	0.94	0.95	1.00	0.98
0.4	0.9	0.5	0.85	0.94	0.94	1.00	0.96
0.4	0.9	0.5	0.90	0.94	0.95	1.00	0.96
0.4	0.9	0.5	0.95	0.95	0.95	1.00	0.96
0.4	0.9	1.0	0.85	0.94	0.95	1.00	0.95
0.4	0.9	1.0	0.90	0.94	0.95	1.00	0.96
0.4	0.9	1.0	0.95	0.94	0.95	1.00	0.96
0.6	0.5	0.5	0.85	0.95	0.96	1.00	0.98
0.6	0.5	0.5	0.90	0.94	0.95	1.00	0.98
0.6	0.5	0.5	0.95	0.93	0.95	1.00	0.98
0.6	0.5	1.0	0.85	0.95	0.95	1.00	0.98
0.6	0.5	1.0	0.90	0.94	0.95	1.00	0.98
0.6	0.5	1.0	0.95	0.93	0.94	1.00	0.97
0.6	0.9	0.5	0.85	0.94	0.95	1.00	0.96
0.6	0.9	0.5	0.90	0.94	0.95	1.00	0.96
0.6	0.9	0.5	0.95	0.94	0.94	1.00	0.96
0.6	0.9	1.0	0.85	0.95	0.94	1.00	0.95
0.6	0.9	1.0	0.90	0.94	0.94	1.00	0.95
0.6	0.9	1.0	0.95	0.94	0.95	1.00	0.96
0.8	0.5	0.5	0.85	0.94	0.94	1.00	0.98
0.8	0.5	0.5	0.90	0.94	0.95	1.00	0.98
0.8	0.5	0.5	0.95	0.95	0.94	1.00	0.98
0.8	0.5	1.0	0.85	0.95	0.96	1.00	0.98
0.8	0.5	1.0	0.90	0.94	0.95	1.00	0.98
0.8	0.5	1.0	0.95	0.94	0.94	1.00	0.98
0.8	0.9	0.5	0.85	0.95	0.95	1.00	0.96
0.8	0.9	0.5	0.90	0.95	0.95	1.00	0.96
0.8	0.9	0.5	0.95	0.94	0.95	1.00	0.96
0.8	0.9	1.0	0.85	0.94	0.94	1.00	0.96
0.8	0.9	1.0	0.90	0.95	0.94	1.00	0.95
0.8	0.9	1.0	0.95	0.95	0.95	1.00	0.96

Recovery of a , b , D , and θ was very accurate. For a -parameters, the mean proportion of credible intervals containing the true value ranged from 0.93 to 0.95 across conditions, with an

average of 0.94. Results for the b -parameters were similar, with a range of 0.94 to 0.96 and an average of 0.95. All estimated credible intervals for D -parameters contained the true simulated value across replications and conditions due to larger credible intervals being estimated for the D -parameters. Credible intervals for θ performed similarly to those for the a - and b -parameters, with a range of 0.95 to 0.98 and a mean of 0.97.

Table 21. Mean proportion of credible intervals containing the true simulated value, level 2 terms

ρ	Proportion reference group	μ_2	α	β_0 values	β_1 values	R^2 values
0.4	0.5	0.5	0.85	1.00	0.84	0.93
0.4	0.5	0.5	0.90	1.00	0.86	0.93
0.4	0.5	0.5	0.95	1.00	0.90	0.94
0.4	0.5	1.0	0.85	1.00	0.73	0.95
0.4	0.5	1.0	0.90	1.00	0.81	0.97
0.4	0.5	1.0	0.95	1.00	0.86	0.97
0.4	0.9	0.5	0.85	1.00	0.85	0.97
0.4	0.9	0.5	0.90	1.00	0.90	0.92
0.4	0.9	0.5	0.95	1.00	0.93	0.97
0.4	0.9	1.0	0.85	1.00	0.89	0.98
0.4	0.9	1.0	0.90	1.00	0.81	0.95
0.4	0.9	1.0	0.95	1.00	0.88	0.99
0.6	0.5	0.5	0.85	1.00	0.58	0.90
0.6	0.5	0.5	0.90	1.00	0.70	0.93
0.6	0.5	0.5	0.95	1.00	0.78	0.96
0.6	0.5	1.0	0.85	1.00	0.49	0.89
0.6	0.5	1.0	0.90	1.00	0.52	0.94
0.6	0.5	1.0	0.95	1.00	0.67	0.91
0.6	0.9	0.5	0.85	1.00	0.68	0.92
0.6	0.9	0.5	0.90	1.00	0.78	0.93
0.6	0.9	0.5	0.95	1.00	0.87	0.95
0.6	0.9	1.0	0.85	1.00	0.57	0.95
0.6	0.9	1.0	0.90	1.00	0.65	0.94
0.6	0.9	1.0	0.95	1.00	0.72	0.97
0.8	0.5	0.5	0.85	1.00	0.21	0.94
0.8	0.5	0.5	0.90	1.00	0.41	0.94
0.8	0.5	0.5	0.95	1.00	0.67	0.94
0.8	0.5	1.0	0.85	1.00	0.08	0.92
0.8	0.5	1.0	0.90	1.00	0.16	0.92
0.8	0.5	1.0	0.95	1.00	0.41	0.92
0.8	0.9	0.5	0.85	1.00	0.41	0.98

0.8	0.9	0.5	0.90	1.00	0.56	0.98
0.8	0.9	0.5	0.95	1.00	0.79	0.97
0.8	0.9	1.0	0.85	1.00	0.22	0.95
0.8	0.9	1.0	0.90	1.00	0.40	0.95
0.8	0.9	1.0	0.95	1.00	0.52	0.93

The proportion of credible intervals containing the true simulated parameters varied for the level 2 parameters. All estimated credible intervals for β_0 values contained the true simulated value across replications and conditions due to larger credible intervals being estimated for the β_0 values. However, β_1 values’ recovery looked very different; the mean proportion of credible intervals containing the true simulated parameters ranged from 0.08 (for the condition where $\rho = 0.8$, the proportion of examinees in the reference group was 0.5, $\mu_2 = 1.0$, and $\alpha = 0.85$) to 0.93 (for the condition where $\rho = 0.4$, the proportion of examinees in the reference group was 0.9, $\mu_2 = 0.5$, and $\alpha = 0.95$), with an overall mean of 0.64. Though β_0 and β_1 were not parameters of interest for this study, their sub-optimal recovery suggests that future research is needed to determine the factors affecting their estimation. The proportion of R^2 values whose estimated credible intervals contained the true value ranged from 0.89 to 0.99, with an overall mean value of 0.94.

Decision Consistency

Results can also be evaluated in terms of their decision consistency; that is, the proportion of time that items’ simulated and estimated values of DIF fall on the same side of a given threshold for determining whether the item exhibits meaningful DIF. Such thresholds are commonly used in operational testing programs to decide which items merit further review by item writers and/or a bias committee. Several thresholds for the value of the D -parameter that would cause an item to be flagged as exhibiting a meaningful amount of DIF were investigated: 0.5, 0.75, and 1.0. These values represent plausible thresholds that could be selected by testing

programs with varying tolerances for DIF and/or varying consequences for bias against a disadvantaged group. Scatterplots showing the decision consistency for each of the conditions for each of these thresholds are shown in Appendices I, J, and K. Each combination of simulation condition and decision consistency threshold was evaluated for its false positive rate, power, and precision. Tables showing results for each condition on these metrics for the three DIF thresholds can be found in Appendix L.

At higher thresholds, the proposed model's power to detect items simulated to exhibit DIF above the threshold was relatively low. Power exceeded 50% only when using the 0.5 DIF threshold. Power was greatest for the conditions where μ_2 was 1.0 – more distant from zero – and α was 0.85 – the lowest proportion of draws coming from the distribution centered around 0. The combination of these parameter values shows that the method is best able to detect DIF when the magnitude of observed DIF is relatively high and more items exhibit DIF.

When examining model performance across the three DIF thresholds, it is evident that the farther the threshold is placed from 0, the lower the values for power and precision, which are not preferred. The false positive rate exhibited an uneven pattern across thresholds, and it was clearly impacted by the value of μ_2 . When μ_2 was 1.0, false positive rates were much lower. Just as the accuracy of parameter recovery varied across conditions, model performance varied according to each of the systematically varied parameters used to create conditions. In order to facilitate interpretation, the false positive rate, power, and precision were also averaged for each possible value of each of systematically varied parameters, disregarding variation in other systematically varied parameters. This summarization allows for interpretation of patterns in results for each of the four parameters used to create conditions. Results for the 0.5 DIF

threshold are given in Table 22; results for the 0.75 DIF threshold are given in Table 23, and results for the 1.0 DIF threshold are given in Table 24.

Table 22. Mean decision consistency results by parameter type using a 0.5 DIF threshold

Parameter type	Value	False positive rate	Power	Precision
ρ	0.4	0.361	0.490	0.639
ρ	0.6	0.365	0.541	0.635
ρ	0.8	0.360	0.621	0.640
P_REF	0.5	0.580	0.617	0.420
P_REF	0.9	0.144	0.484	0.856
μ_2	0.5	0.491	0.282	0.509
μ_2	1.0	0.234	0.819	0.766
α	0.85	0.331	0.558	0.669
α	0.90	0.345	0.558	0.655
α	0.95	0.410	0.535	0.590

Table 23. Mean decision consistency results by parameter type using a 0.75 DIF threshold

Parameter type	Value	False positive rate	Power	Precision
ρ	0.4	0.432	0.232	0.485
ρ	0.6	0.491	0.278	0.509
ρ	0.8	0.524	0.383	0.476
P_REF	0.5	0.564	0.378	0.436
P_REF	0.9	0.401	0.217	0.544
μ_2	0.5	0.894	0.127	0.051
μ_2	1.0	0.071	0.468	0.929
α	0.85	0.492	0.323	0.508
α	0.90	0.483	0.351	0.517
α	0.95	0.472	0.218	0.445

Table 24. Mean decision consistency results by parameter type using a 1.0 DIF threshold

Parameter type	Value	False positive rate	Power	Precision
ρ	0.4	0.410	0.087	0.340
ρ	0.6	0.388	0.092	0.363

ρ	0.8	0.384	0.121	0.366
P_REF	0.5	0.683	0.154	0.317
P_REF	0.9	0.104	0.046	0.396
μ_2	0.5	0.500	0.000	0.000
μ_2	1.0	0.288	0.200	0.712
α	0.85	0.389	0.097	0.361
α	0.90	0.403	0.103	0.347
α	0.95	0.389	0.100	0.361

Changes in the values of the systematically varied parameters affected the decision consistency metrics in different ways. Inspection of the full tables in Appendix L and the scatterplots in Appendices I, J, and K shows that overall, the proportion of examinees in the reference group was the most impactful parameter for these metrics. Across all thresholds and regardless of the other systematically varied parameters' values, the mean false positive rate is minimized and the precision is increased when the proportion of examinees in the reference group is 0.9 rather than 0.5. However, these gains come at the expense of power, especially at the higher thresholds.

For the 0.5 DIF threshold, setting the proportion of examinees in the reference group to 0.9 instead of 0.5 decreased the mean false positive rate from 0.580 to 0.144 and increased the precision from 0.420 to 0.856; however, power decreased from 0.617 to 0.484. This pattern was present across DIF thresholds, and the magnitude of the difference between conditions with differing proportions of examinees in the reference group increased as the DIF threshold increased. When the DIF threshold was 1.0, setting the proportion of examinees in the reference group to 0.9 instead of 0.5 decreased the mean false positive rate from 0.683 to 0.104 and increased the precision from 0.317 to 0.396; however, power decreased from 0.154 to 0.046. These findings indicate that the proposed model provides more accurate classifications when a

greater proportion of examinees are in the reference group, though this accuracy comes at the price of the power to detect DIF when it is truly present.

Chapter V: Conclusion

Conclusions

The results given above can be interpreted in light of the research questions posed to evaluate this proposed model.

Research Question 1.

The first research question posed was, “Can an explanatory hierarchical IRT model accurately recover person and item parameters?” Recovery of these parameters was assessed in terms of bias, mean squared error, correlation and the proportion of parameters whose credible interval included the true simulated value. Results of this study demonstrate that the proposed model exhibited adequate recovery of person and item parameters, though parameter recovery varied according to the simulation conditions. The proposed model provides accurate estimates of person and item parameters while simultaneously estimating the amount of DIF exhibited by items and the relationship between the estimated amount of DIF and item-level explanatory features.

Research Question 2.

In addition to person and item parameters, the proposed model should also be able to accurately recover group means. This led to the second research question, “Does the model accurately recover differences in mean group ability between the focal and reference groups?” Group mean recovery was addressed by examining the amount of bias in group mean ability estimates. Results showed that recovery of the reference and focal group means group mean was

highly accurate overall. Reference group means had an overall mean bias across replications and conditions of 0.001. Focal group mean recovery exhibited a slight amount of negative bias, with an overall mean bias across replications and conditions of -0.049.

Additionally, the variance of the distribution of bias observed in recovery of group means was examined. The primary condition variable affecting this variance was the proportion of examinees in the reference group. When the proportion of examinees in the reference group was 0.5, there was less variance in the bias of the reference group mean estimates. In contrast, when the proportion of examinees in the reference group was 0.9, there was less variance in the bias of the focal group mean estimates. Overall, the estimates of the reference group mean exhibited less variance in the amount of bias across replications.

Research Question 3.

After examining recovery of the level 1 parameters, the next research question was, “Does the model accurately recover the relationship between the amount of DIF observed in an item and the presence of DIF-related item-level features?” While recovery of the R^2 value was less accurate than recovery of the person and item parameters, results indicated adequate recovery of the simulated R^2 values across the three simulated values of R^2 . Estimated R^2 values were distributed normally around their simulated values, and the credible intervals surrounding the estimated R^2 values contained the true values 94% of the time on average. Recovery of this relationship may be improved by using less informative priors when estimating the D -parameters, which may in turn result in more accurate parameter recovery. More accurate estimates of the D -parameters would improve recovery of R^2 .

Research Question 4.

The study also examined the model's performance across conditions in order to answer the final research question: "Does the model perform similarly across different simulated conditions which vary the proportion of examinees in the reference group, the strength of the relationship of DIF to the item-level features, and the distribution of the amount of simulated DIF in each item?" ANOVAs were conducted for bias and MSE to examine the systematically varied parameters used to create conditions' effect on parameter recovery. Many of the predictors and interactions were statistically significant at the $p < 0.05$ level, but when effect sizes were considered, many of the results were not meaningful.

The ANOVA examining bias in D -parameter recovery yielded two η^2 values that met the criteria to be considered small effect sizes; these effects were for μ_2 and α , the parameters that determine the shape of the mixture distribution. These results suggest that the negative bias in D -parameter estimates, likely due to the use of relatively informative, was more pronounced the mixture distribution used to simulate D -parameters resulted in items with greater amounts of DIF ($\mu_2 = 1$) and fewer items with DIF parameters close to 0 ($\alpha = 0.85$). Additionally, two interactions in the ANOVA examining bias in θ recovery were also found to be small effects: the three-way interaction of ρ , the proportion of examinees in the reference group, and α ; and the four-way interaction between all four systematically varied parameters. These results suggest that the systematically varied parameters' combined effect was meaningful for the accuracy of θ recovery.

Results from the ANOVAs examining MSEs indicate that the systematically varied parameters used to create conditions had a meaningful impact on parameter recovery. Of those parameters, the most meaningful was the proportion of students in the reference group, which exhibited a medium effect size when conducting ANOVAs on the MSEs for the b -parameters, D -

parameters, and θ . However, evidence for all parameters' impact on parameter recovery can be seen in the results for θ . For this ANOVA, the summed values of η^2 met the criteria to be considered a large effect size; the sum of 0.147 can be interpreted to mean that all predictors explained 14.7% of the total variance in the amount of error observed in parameter recovery.

Power was greater for conditions with equal proportions of examinees in the reference and focal groups, which mirrors the findings of Jodoin and Gierl (2001). However, the improved power came at the expense of higher false positive rates and poorer precision. Additionally, false positive rates were quite high when the amount of simulated DIF was 0.5 rather than 1.0. Testing programs conducting DIF detection studies on real world datasets should carefully consider their priorities when selecting a threshold for flagging items as exhibiting meaningful DIF. While lower thresholds identify more of the items which truly exhibit DIF (power), they also result in more false positives, which could result in unnecessary, costly item review and rewriting.

Uses of the Proposed Model

The proposed model could prove useful to operational testing programs seeking to improve their item writing and review processes. The model is able to accurately estimate item and person parameters while simultaneously providing estimates of the amount of DIF and its relationship to explanatory item-level features. By explaining the amount of DIF observed in terms of item-level features, this method provides actionable results that could inform item writing and review processes.

An operational testing program could use this modeling strategy with pilot data in order to identify item-level features' relationship to the amount of DIF. Once meaningful item features were identified, the program could design process improvements. Item writers could be trained to avoid item-level features found to be associated with the amount of DIF observed, and item

reviewers could be instructed to look for DIF-causing features during the item review process. If test developers could produce fewer items exhibiting DIF, they could save time and money on the item review process. Additionally, attention to construct-irrelevant item level features during the development process would strengthen the validity evidence for scores from that assessment.

Limitations

While this study attempted to select conditions that would be seen in real-world assessments, this model has not yet been applied to a real dataset. Additionally, the model does not account for non-uniform DIF, which can also bias item performance.

Results from this simulation study were also impacted by the selection of priors. Relatively informative priors were used in order to ensure each of the 6,000 replications of this study could be completed in a reasonable amount of time. More diffuse priors are recommended when using this method with a real dataset whose true values are unknown. While less informative priors will likely increase the time required for estimation, time constraints would be less of an issue for a single analysis. Please see the Technical Considerations for the Practitioner section for more information on the computing power required to run this model.

Future Directions

Future simulation studies could examine the model's performance across additional conditions such as test length, proportion of items exhibiting DIF, and number of examinees. While β_0 and β_1 were not parameters of interest for this study, additional simulations may wish to vary the true values of β_0 and β_1 to determine the factors affecting recovery of those parameters. The negative bias in estimates of the D -parameters may be responsible for the poor recovery of these parameters, but future study is needed to confirm or deny this hypothesis.

Since real world operational tests may exhibit non-uniform DIF in addition to uniform DIF, future studies may wish to add an interaction term to the model and assess its effectiveness in recovering both types of DIF. Future research could also investigate the relationships between the proportion of examinees in the reference group and the amount of bias observed in the recovery of groups' mean ability. Finally, the method for assessing decision consistency could be improved by retaining the credible intervals for D -parameter estimates and flagging items whose credible intervals do not cross zero as possessing DIF. Recovery could be evaluated by tracking which component distribution of the mixture distribution was used to simulate the item's true D -parameter.

An application of this model with a real-world dataset that has been coded for item-level features would provide a demonstration of its performance with actual data and illustrate its utility to operational testing programs. Applications of the model across several real world datasets could provide valuable information on which item-level features are consistently related to DIF for various commonly assessed constructs.

References

- Andrich, D. (1988). *Rasch models for measurement*. (Sage University Paper series on Quantitative Applications in the Social Sciences 07-068). Beverly Hills, CA: Sage.
- Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3-23). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Angoff, W. H. (1982). Use of difficulty and discrimination indices for detecting item bias. In R.A. Berk (Ed.), *Handbook of Methods for Detecting Test Bias*, 96-116. Baltimore, MD: Johns Hopkins University Press.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing. (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.
- Brussow, J.A., Skorupski, W.P., & Loughran, J. (2015, April). *Explanatory models for understanding differential item functioning*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL.
- Cho, H. J., Lee, J., & Kingston, N. (2012). Examining the effectiveness of test accommodation using DIF and a mixture IRT model. *Applied Measurement in Education*, 25(4), 281-304.
- De Ayala, R. J. (2013). *The theory and practice of item response theory*. Guilford Publications.
- De Ayala, R. J., Kim, S. H., Stapleton, L. M., & Dayton, C. M. (2002). Differential item functioning: A mixture distribution conceptualization. *International Journal of Testing*, 2(3-4), 243-276.
- De Boeck, P. (2008). Random item IRT models. *Psychometrika*, 73(4), 533-559.
- De Boeck, P. & Wilson, M. (2004a). A framework for item response models. In P. DeBoeck & M. Wilson (Eds.), *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach*, 3-41. New York: Springer.
- De Boeck, P. & Wilson, M. (2004b). Descriptive and explanatory item response models. In P. DeBoeck & M. Wilson (Eds.), *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach*, 43-74. New York: Springer.

- De Boeck, P. & Wilson, M. (2004c). Preface. In P. DeBoeck & M. Wilson (Eds.), *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach*, vii-ix. New York: Springer.
- De Boeck, P. & Wilson, M. (2015). Multidimensional explanatory item response modeling. In S.P. Reise & D.A. Revicki (Eds.), *Handbook of Item Response Theory Modeling*, 252-271. New York: Routledge.
- Dorans, N. J., & Holland, P. W. (1992). DIF detection and description: Mantel-Haenszel and standardization. *ETS Research Report Series*, 1992(1).
- Embretson, S. E., & Reise, S. P. (2013). *Item response theory for psychologists*. Psychology Press.
- Fox, J. (2004). Applications of multilevel IRT modeling. *School Effectiveness and School Improvement*, 15(3-4), 261-280.
- Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. New York, NY: Cambridge University Press.
- Gelman, A., & Shirley, K. (2011). Inference from simulations and monitoring convergence. In S.B. Brooks, A. Gelman, & G. Jones (Eds.), *Handbook of Markov Chain Monte Carlo*, 163-174. Boca Raton, FL: CRC Press.
- Goldstein, H. (2011). *Multilevel statistical models* (4th ed.). John Wiley & Sons.
- Haag, N., Heppt, B., Stanat, P., Kuhl, P., & Pant, H. A. (2013). Second language learners' performance in mathematics: Disentangling the effects of academic language features. *Learning and Instruction*, 28, 24-34.
- Hartig, J., Frey, A., Nold, G., & Klieme, E. (2012). An application of explanatory item response modeling for model-based proficiency scaling. *Educational and Psychological Measurement*, 72(4), 665-686.
- Jodoin, M. G., & Gierl, M. J. (2001). Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education*, 14(4), 329-349.

- Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement*, 38(1), 79-93.
- Loughran, J.T. (2014). *Understanding differential item functioning for English language learners: The influence of linguistic complexity features* (Doctoral dissertation). Retrieved from https://kuscholarworks.ku.edu/bitstream/handle/1808/18649/Loughran_ku_0099D_13735_DATA_1.pdf
- Loughran, J.T., & Skorupski, W.P. (2014). *Understanding differential item functioning with the general linear model*. Paper presented at the 2014 annual meeting of the National Council on Measurement in Education (NCME), Philadelphia, PA.
- McCulloch, C. E., & Searle, S. R. (2001). *Generalized, Linear, and Mixed Models*. New York: Wiley.
- Pierce, C. A., Block, R. A., & Aguinis, H. (2004). Cautionary note on reporting eta-squared values from multifactor ANOVA designs. *Educational and psychological measurement*, 64(6), 916-924.
- Randall, J., Cheong, Y.F., & Engelhard, G., Jr. (2011). "Using explanatory Item Response Theory modeling to investigate context effects of differential item functioning for students with disabilities." *Educational and Psychological Measurement*, 71, 129-147.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd Ed.). Sage.
- Ravand, H. (2015). Item Response Theory Using Hierarchical Generalized Linear Models. *Practical Assessment, Research & Evaluation*, 20(7).
- Richardson, J. T. (2011). Eta squared and partial eta squared as measures of effect size in educational research. *Educational Research Review*, 6(2), 135-147.
- Robinson, W. S. (1950). Ecological Correlations and the Behavior of Individuals. *American Sociological Review*, 15(3), 351-357.
- Saint-Dizier, P. (2006). Introduction to the syntax and semantics of prepositions. In P. Saint-Dizier (Ed.), *Syntax and semantics of prepositions*, 1-26. Dordrecht, The Netherlands: Springer.

Skorupski, W.P., Brussow, J.A., and Loughran, J.T. (2016, April). *Simultaneous evaluation of DIF and its sources using hierarchical explanatory models*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Washington, D.C.

Snijders, T. A. (2014). Multilevel analysis. In *International Encyclopedia of Statistical Science* (pp. 879-882). Springer Berlin Heidelberg.

Stan Development Team (2016). RStan: the R interface to Stan. R package version 2.14.1. <http://mc-stan.org/>.

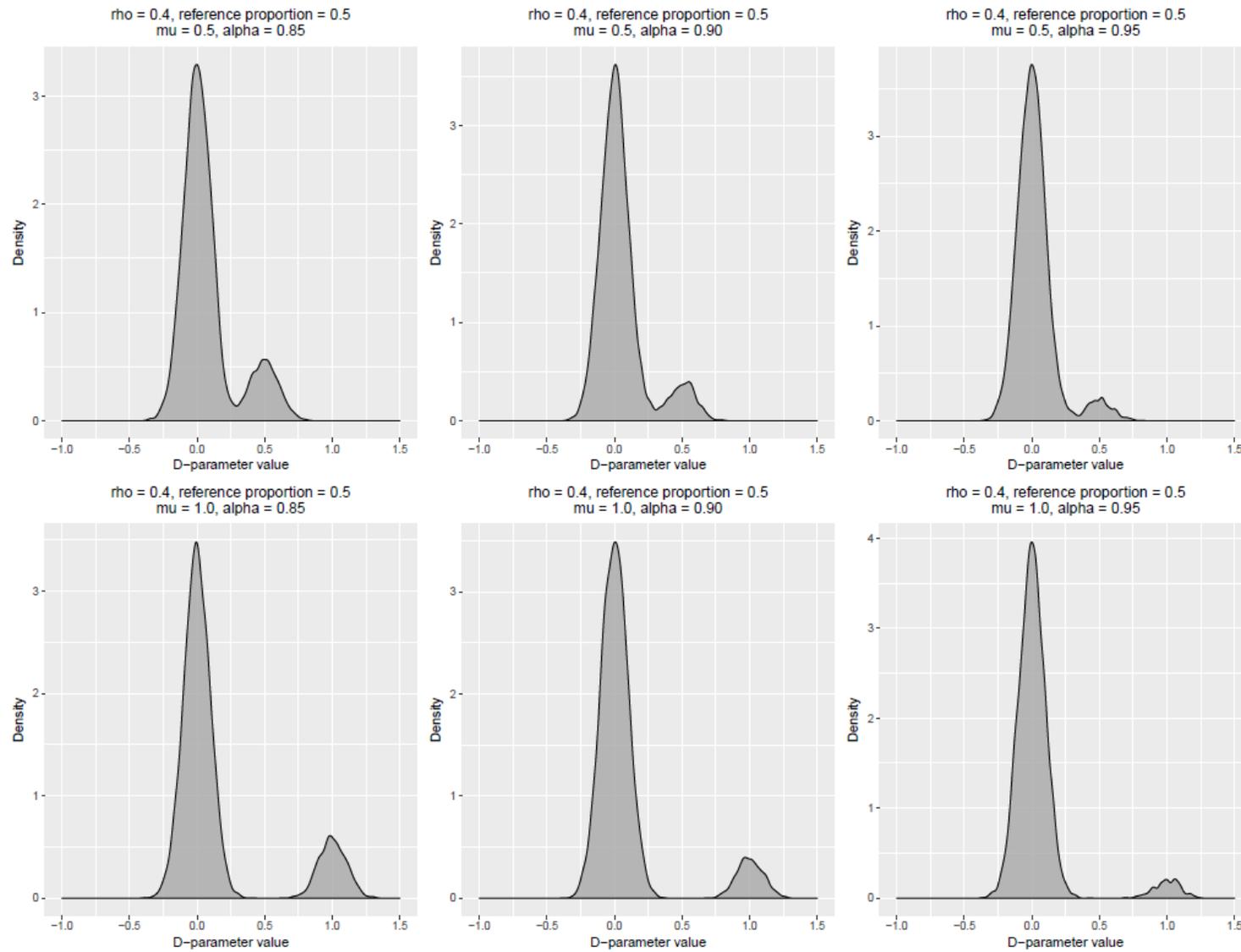
Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational measurement*, 27(4), 361-370.

Wainer, H. (1983). Are we correcting for guessing in the wrong direction? In D.J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 63-80). Hillsdale, NJ: Erlbaum.

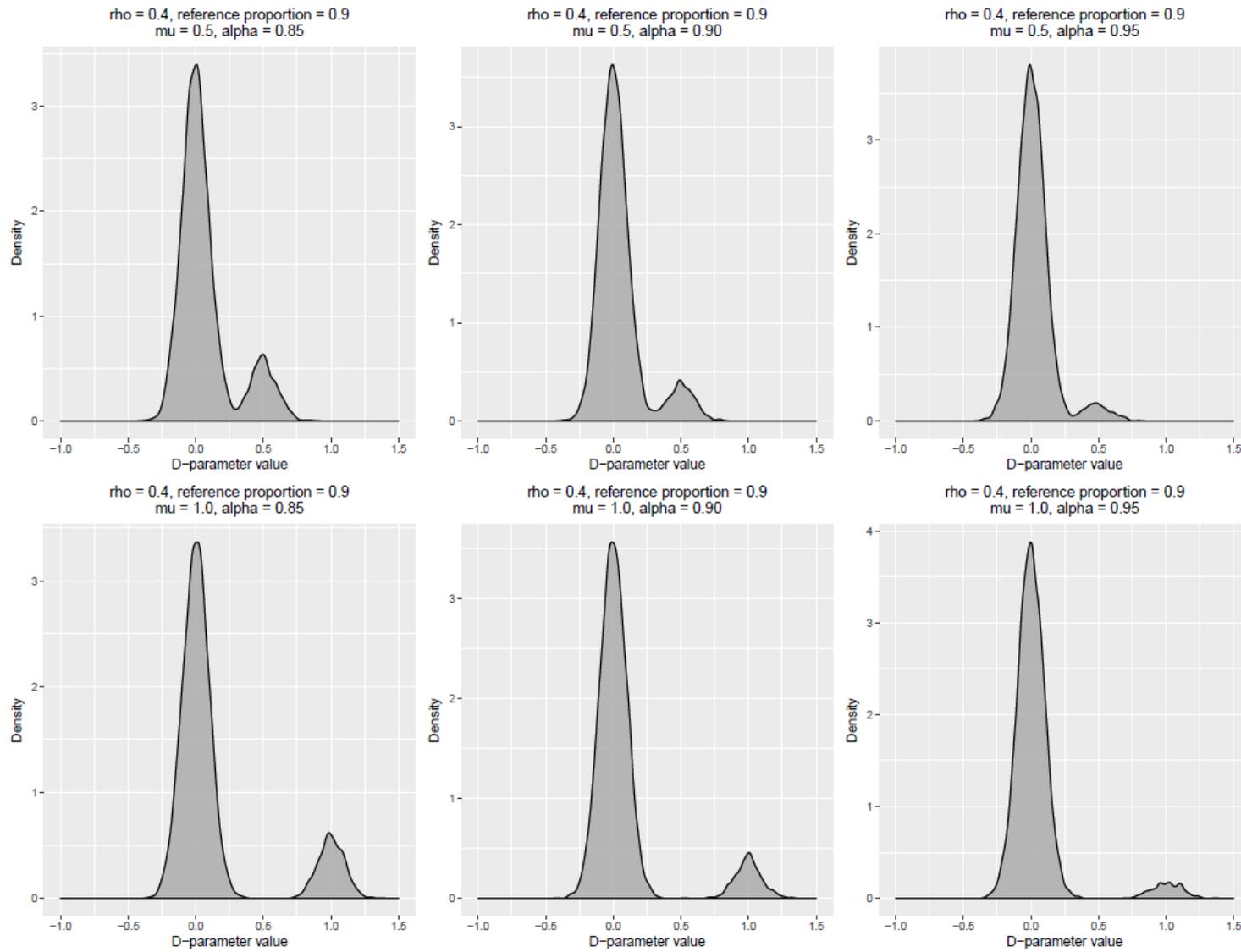
Zumbo, B.D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores*. Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Appendix A. Density plots of mixture distributions

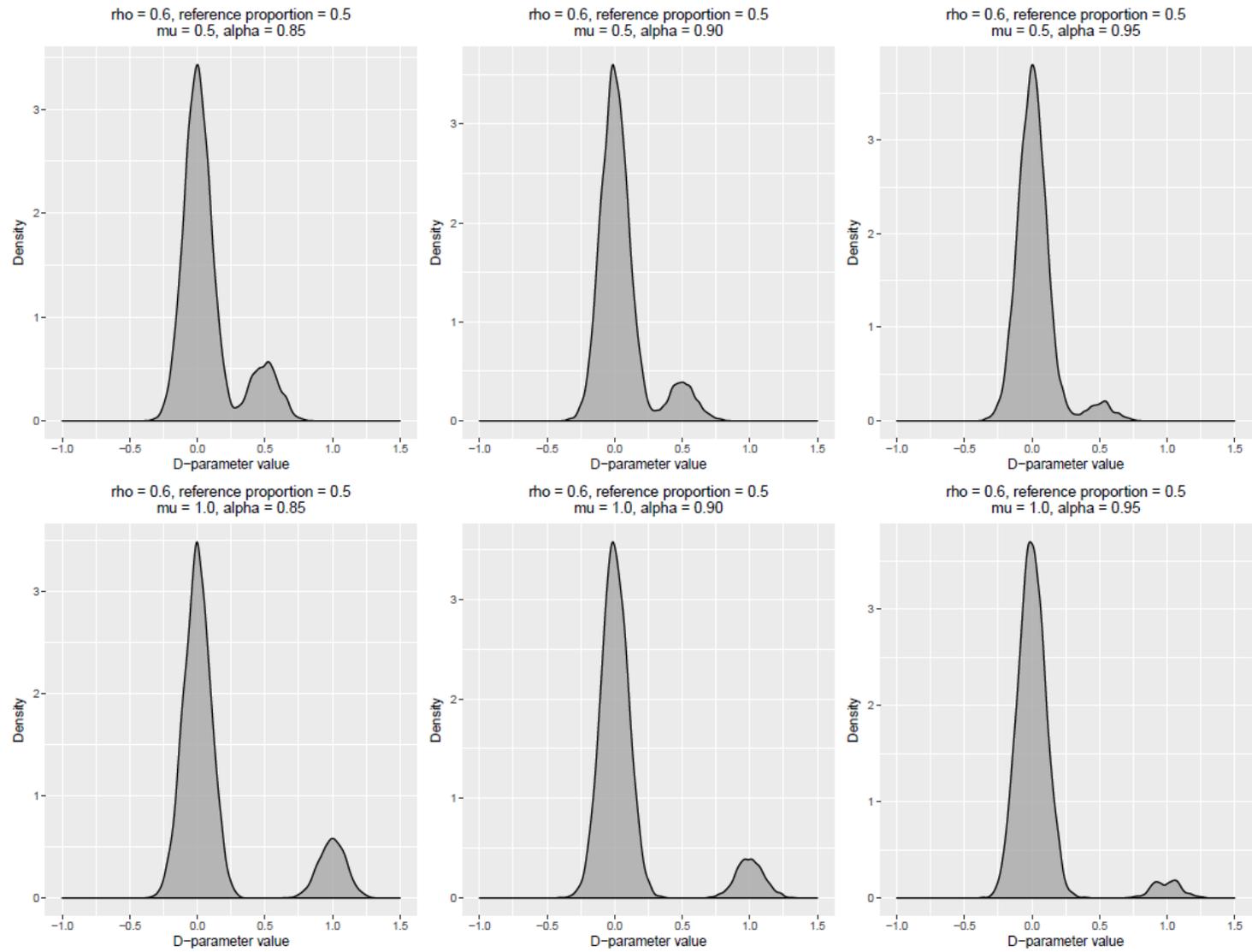
D-parameter mixture distributions



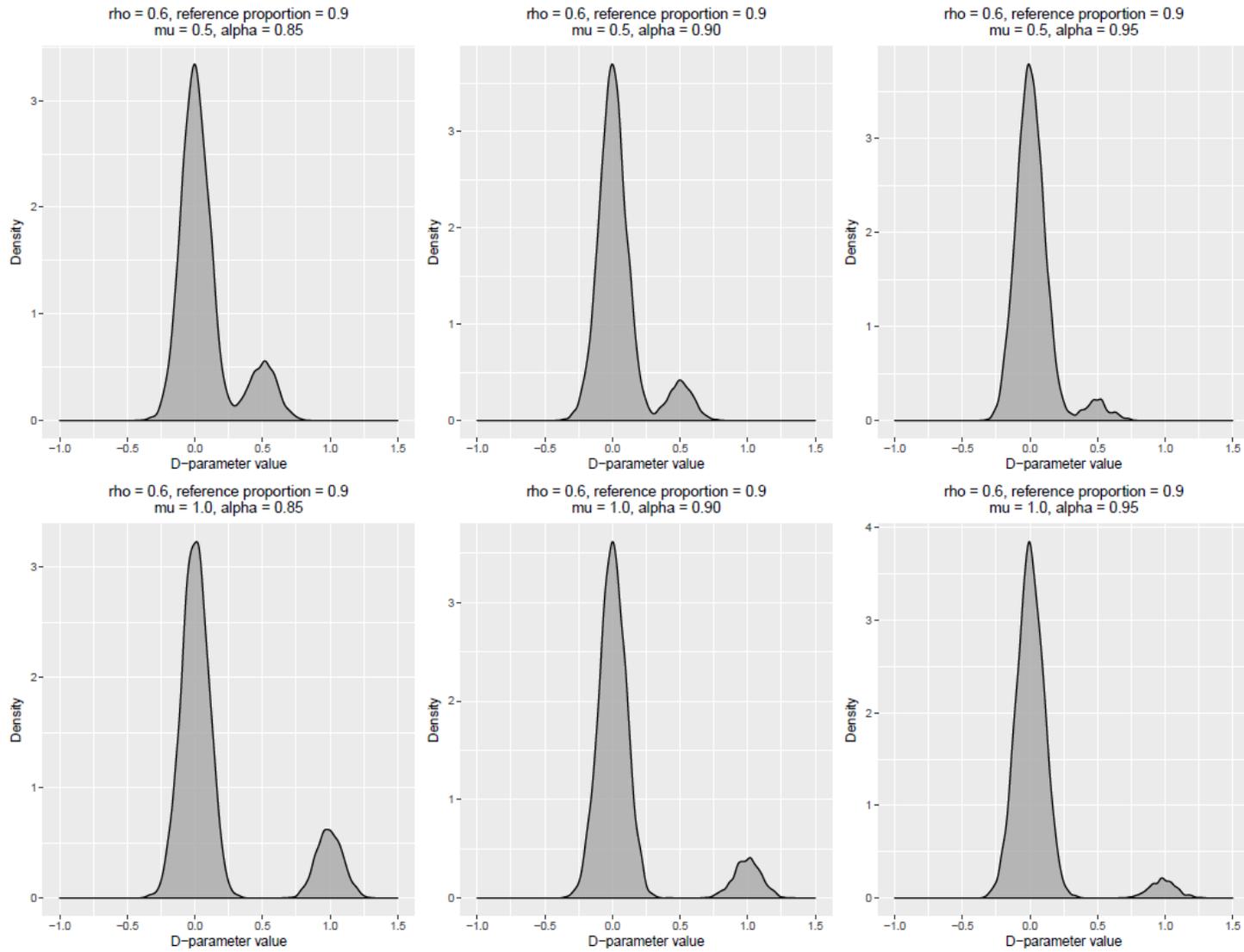
D-parameter mixture distributions



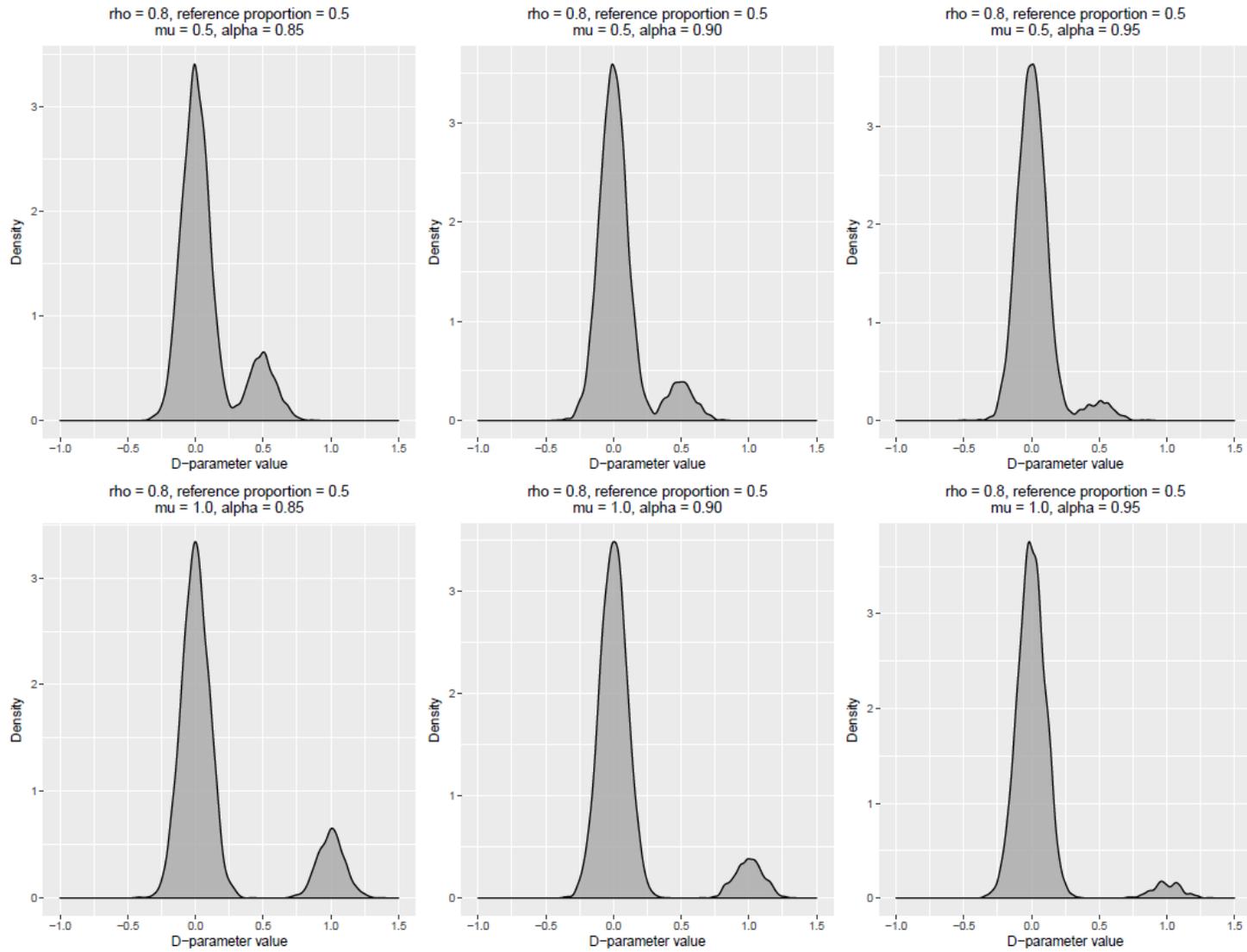
D-parameter mixture distributions



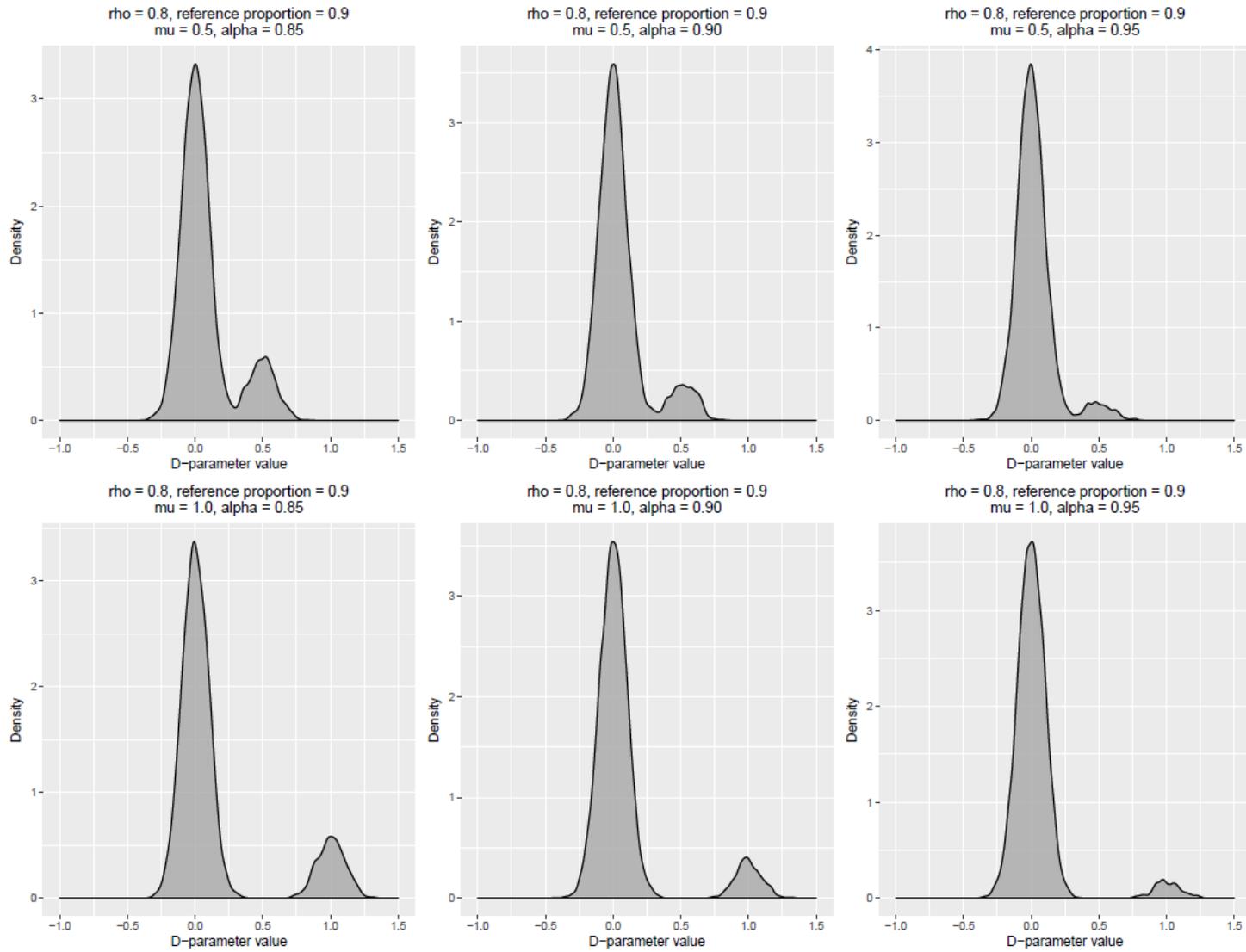
D-parameter mixture distributions



D-parameter mixture distributions



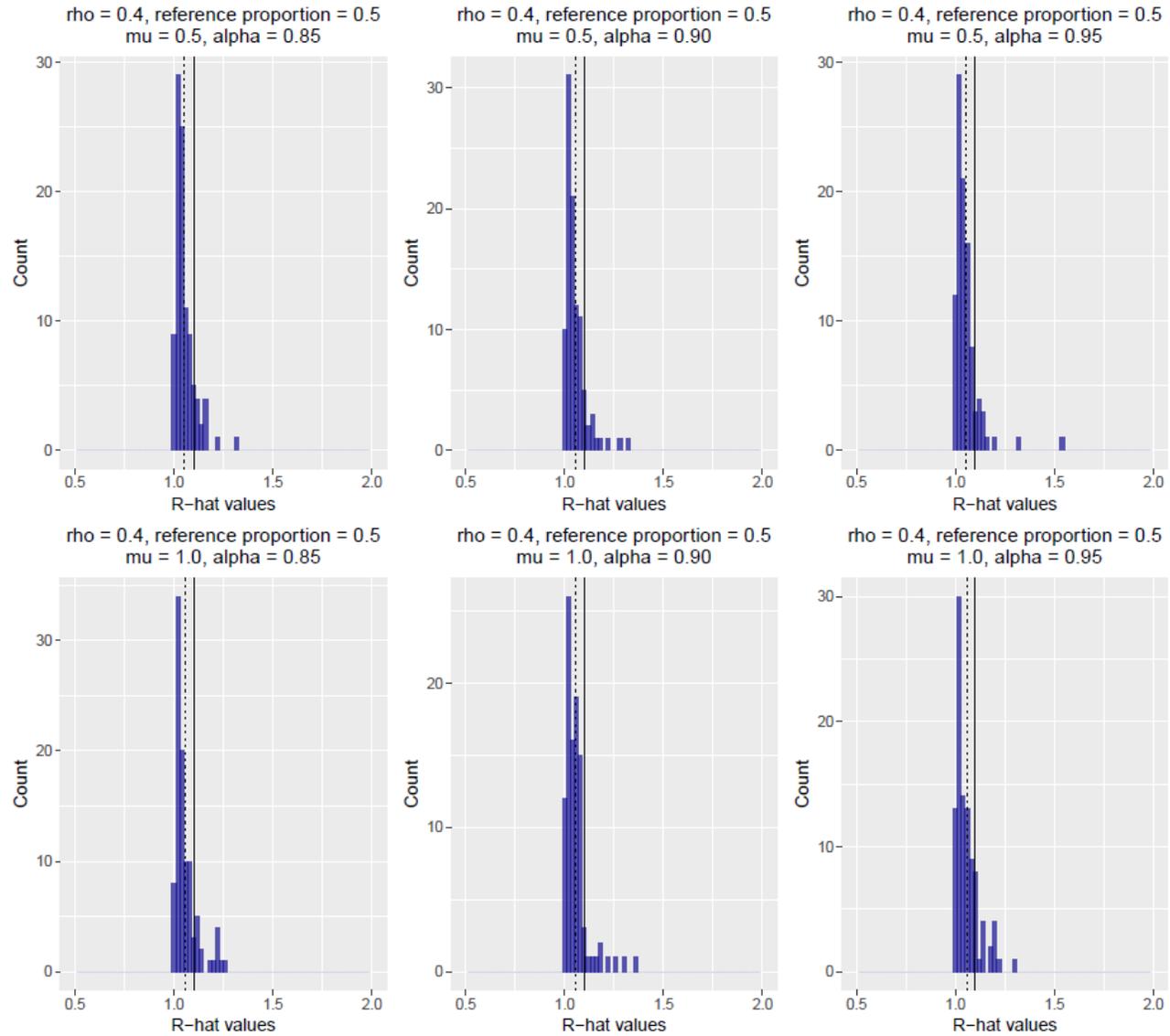
D-parameter mixture distributions



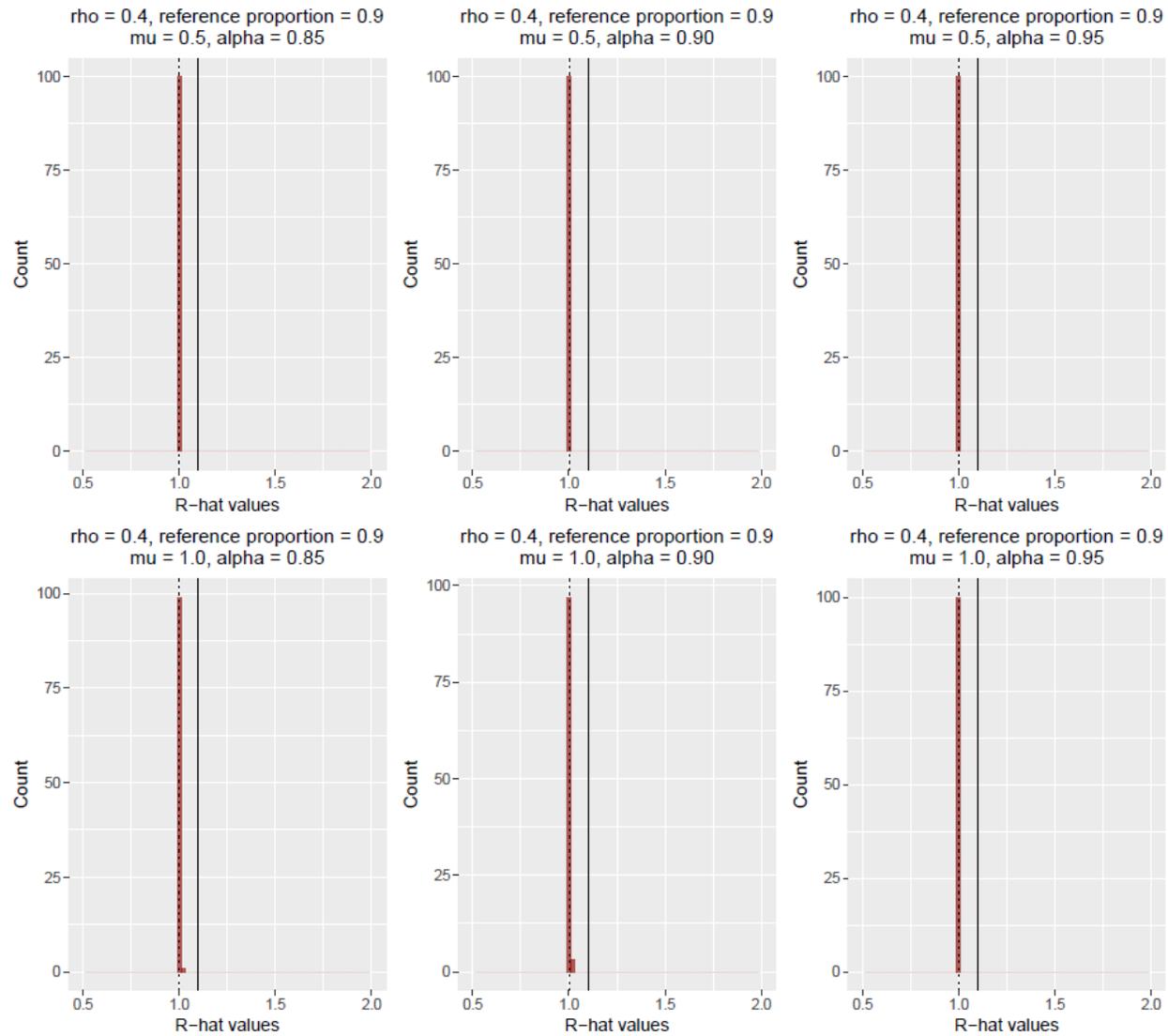
Appendix B. Histograms of the mean \hat{R} value across parameters for the replications within each condition

Note: Dotted vertical lines represent the mean of the distribution. Solid vertical lines represent the 1.1 threshold recommended by Gelman and Shirley (2011).

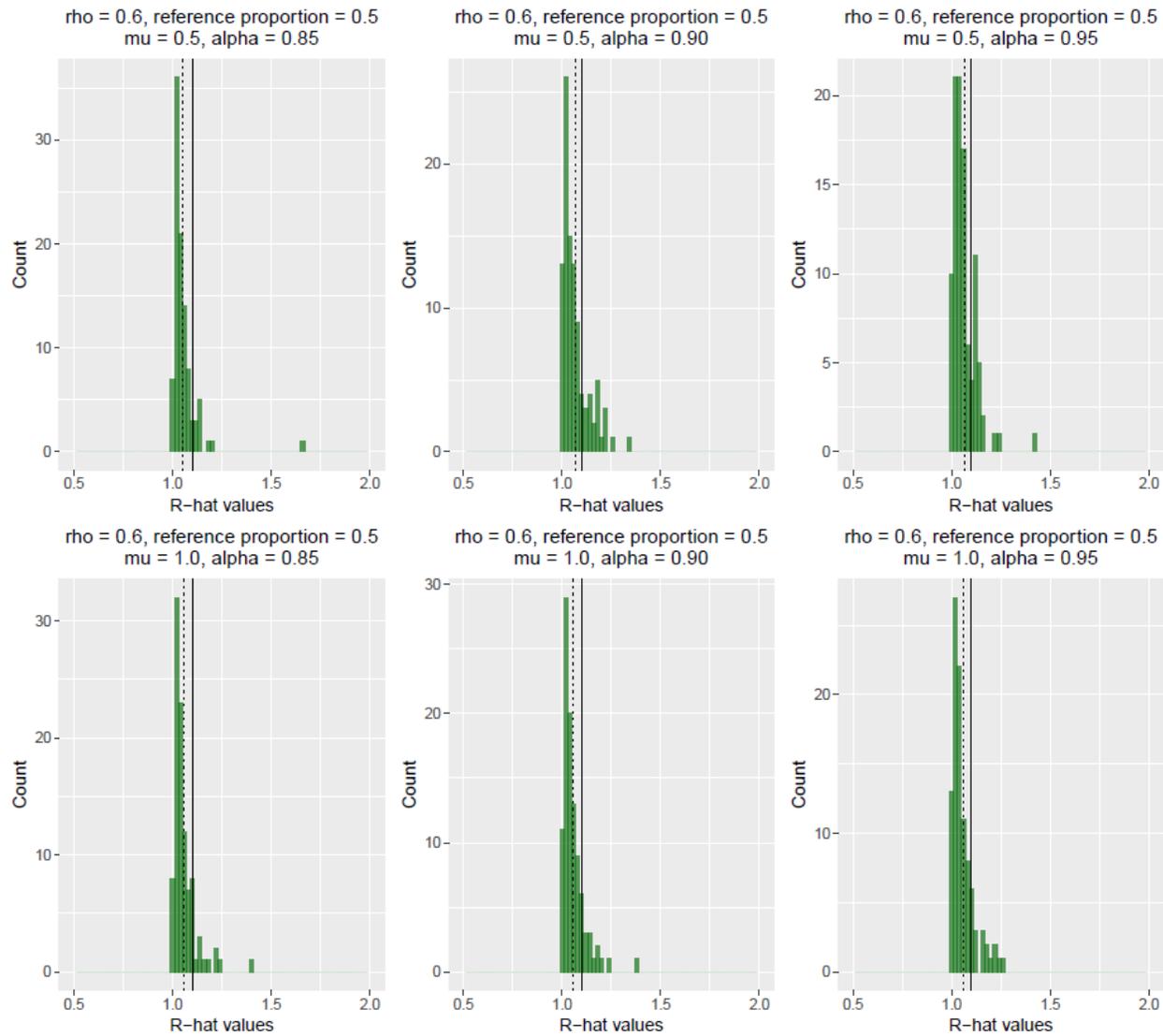
Histogram of mean R-hat values



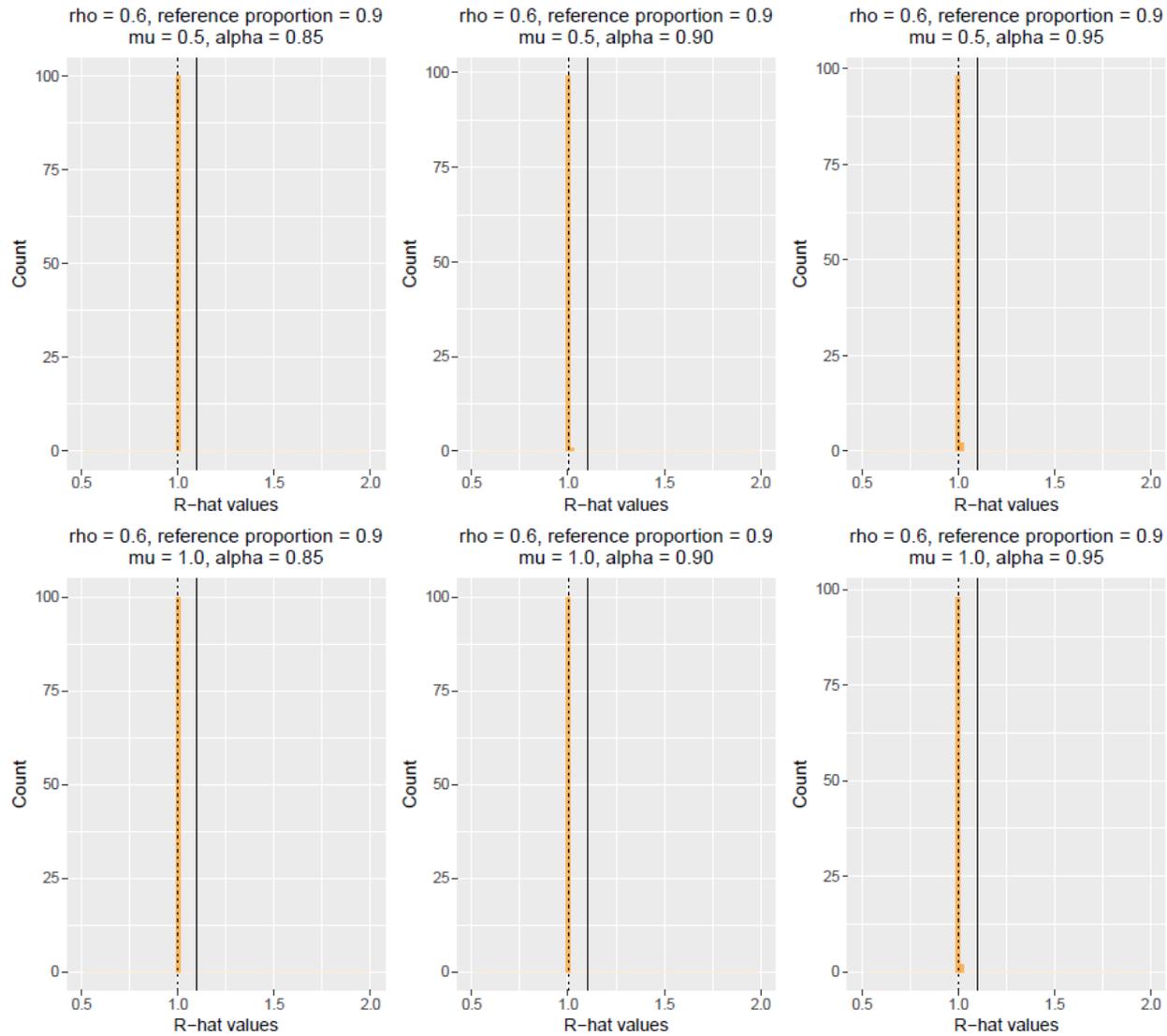
Histogram of mean R-hat values



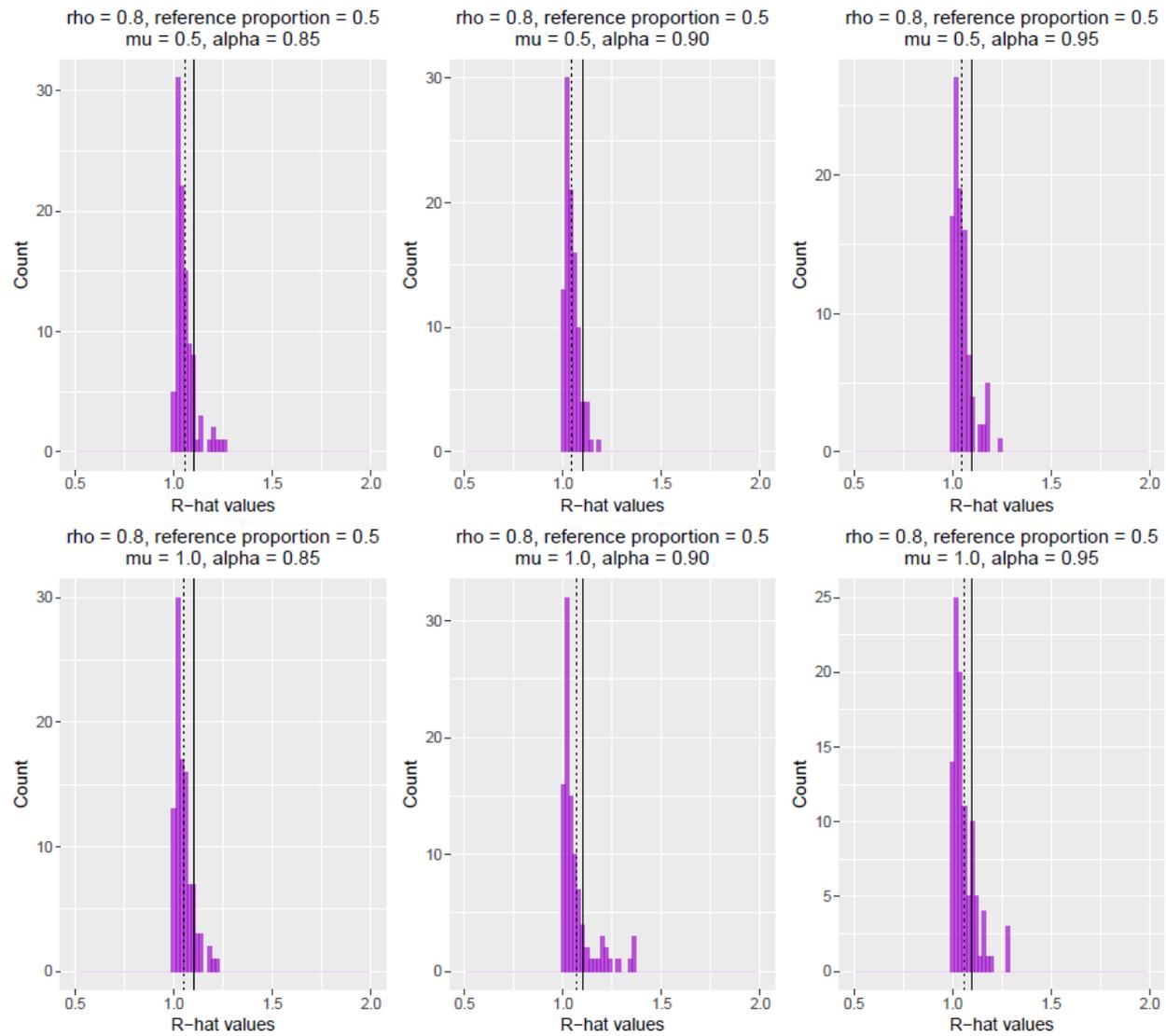
Histogram of mean R-hat values



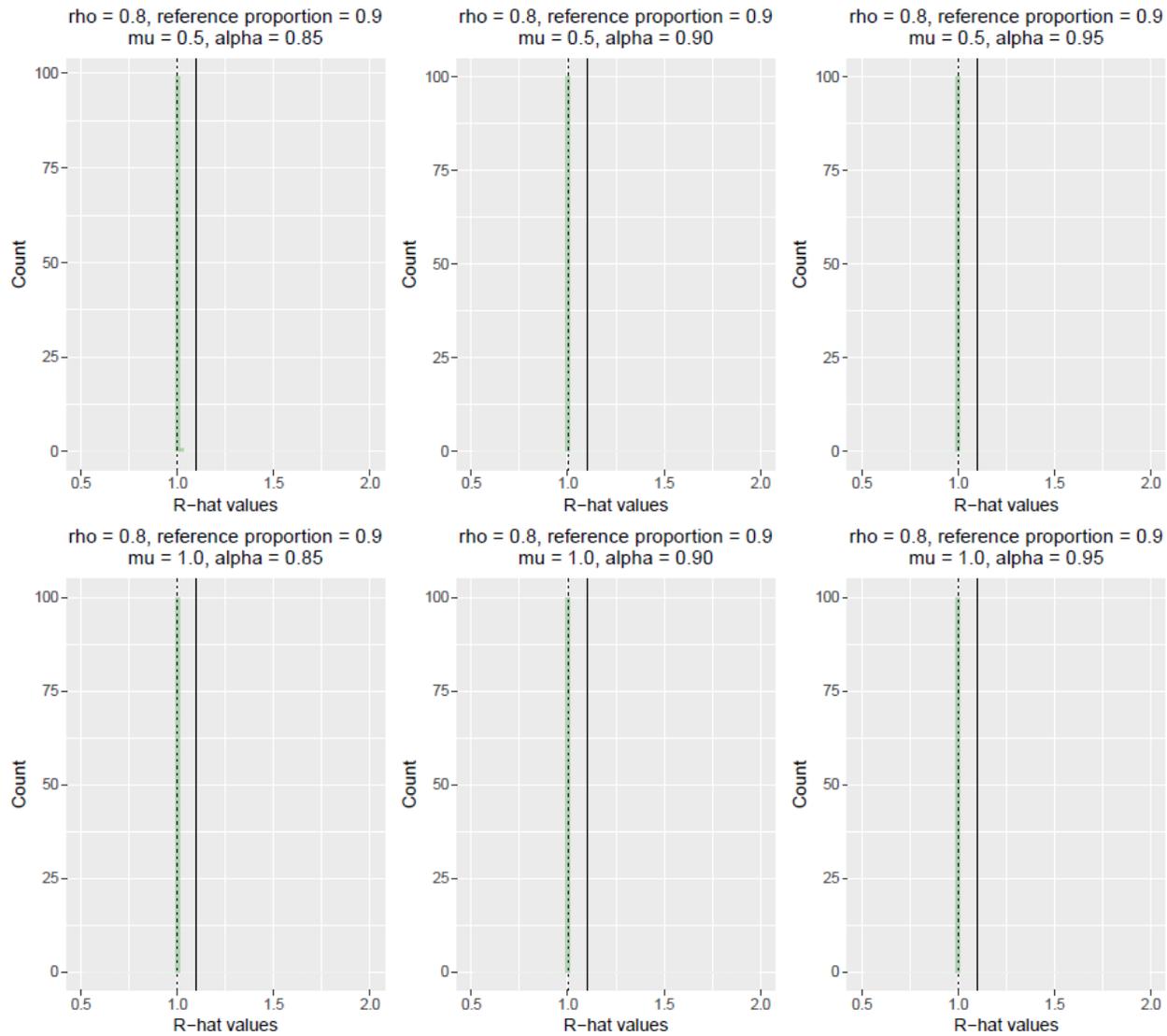
Histogram of mean R-hat values



Histogram of mean R-hat values



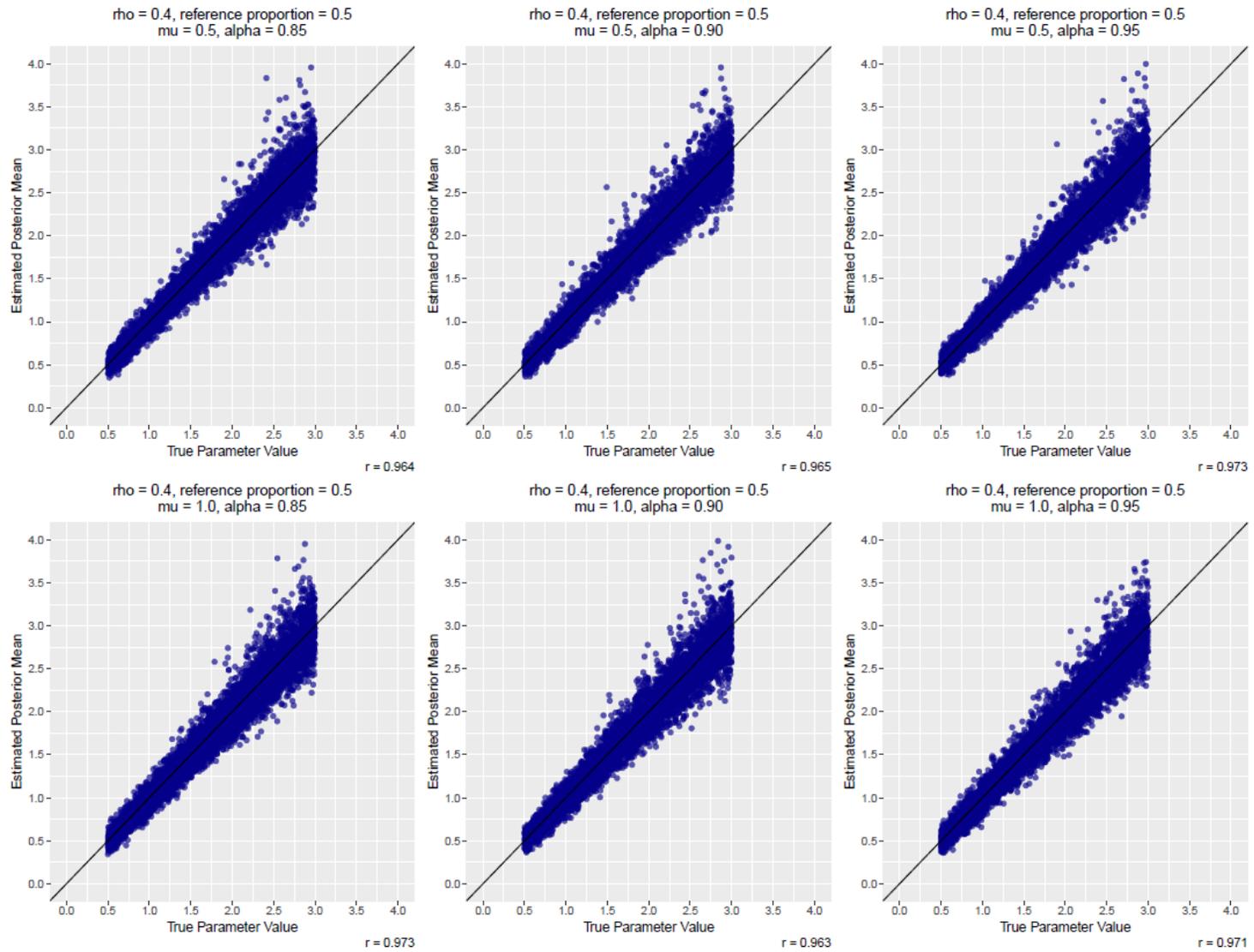
Histogram of mean R-hat values



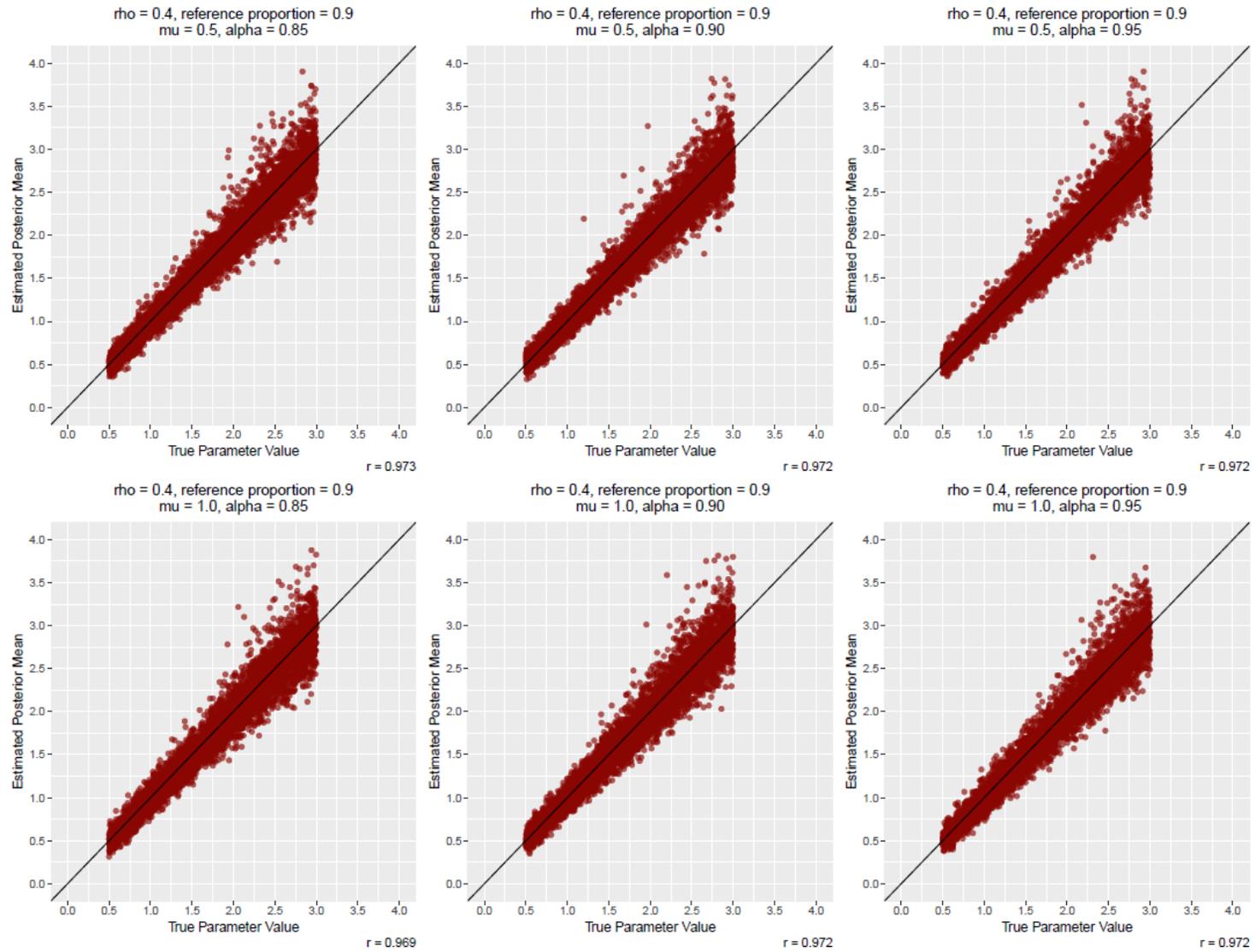
Appendix C. Correlation scatterplots for a -parameter recovery

Note: The solid diagonal line represents a perfect correlation of 1.

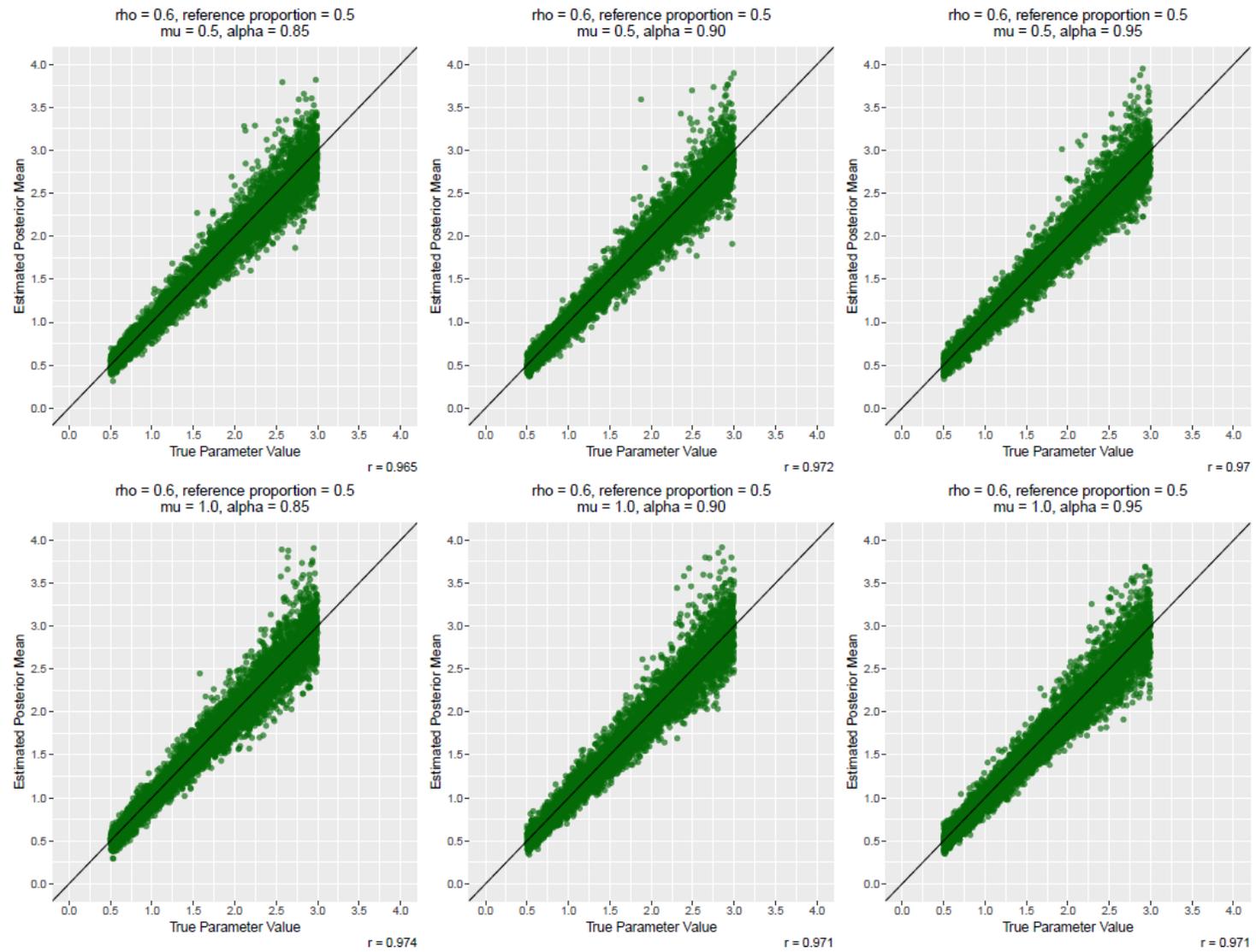
a-parameter scatterplots



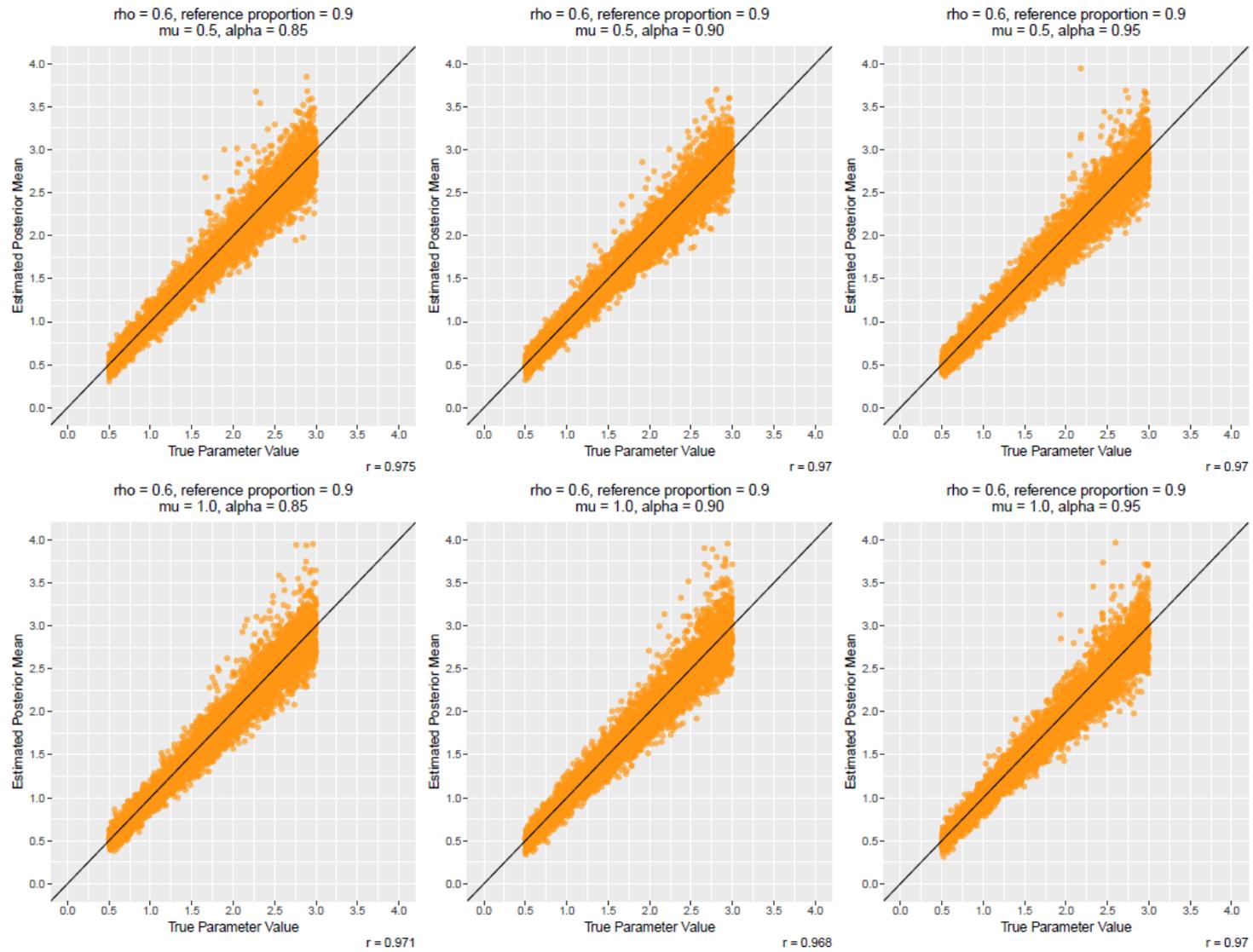
a-parameter scatterplots



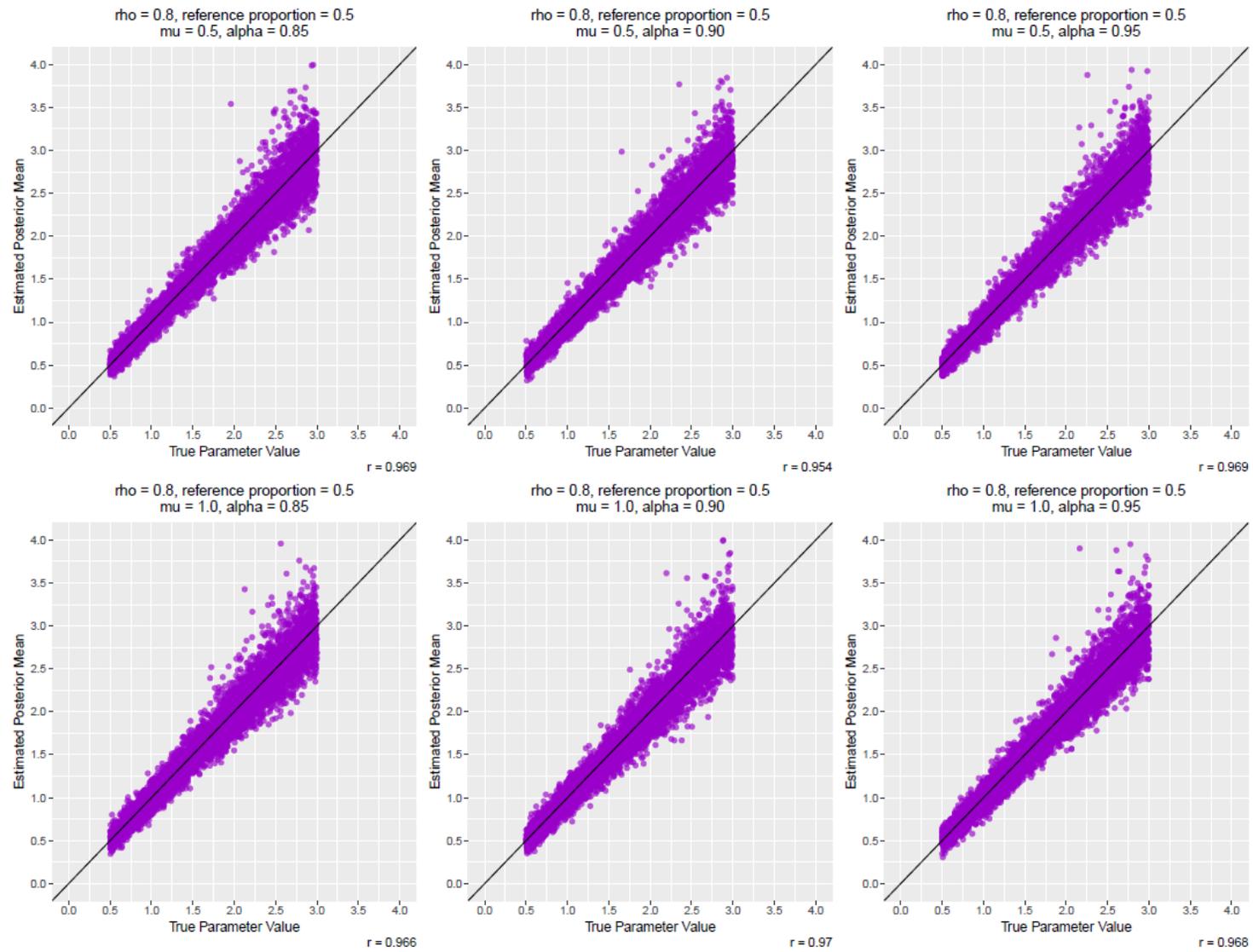
a-parameter scatterplots



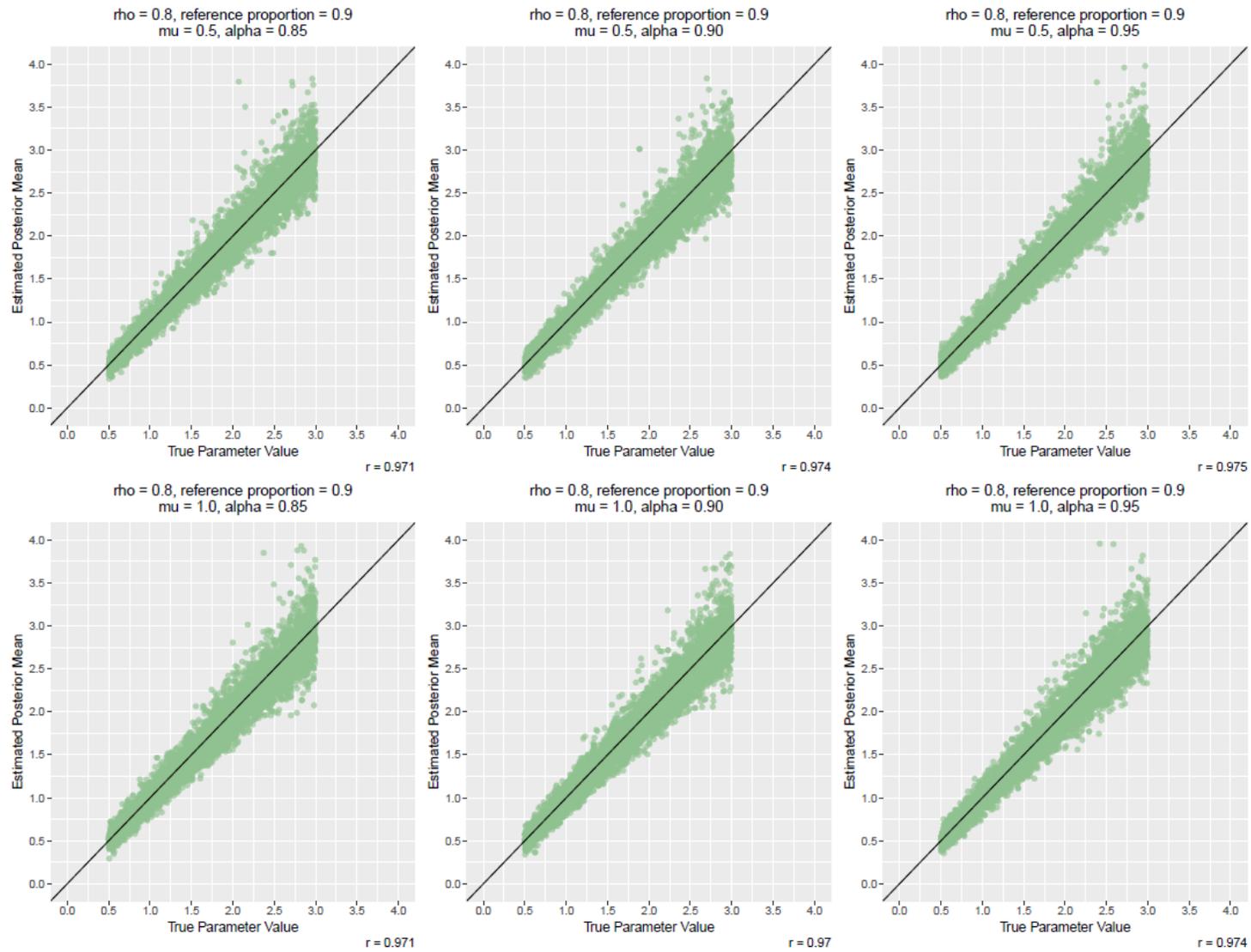
a-parameter scatterplots



a-parameter scatterplots



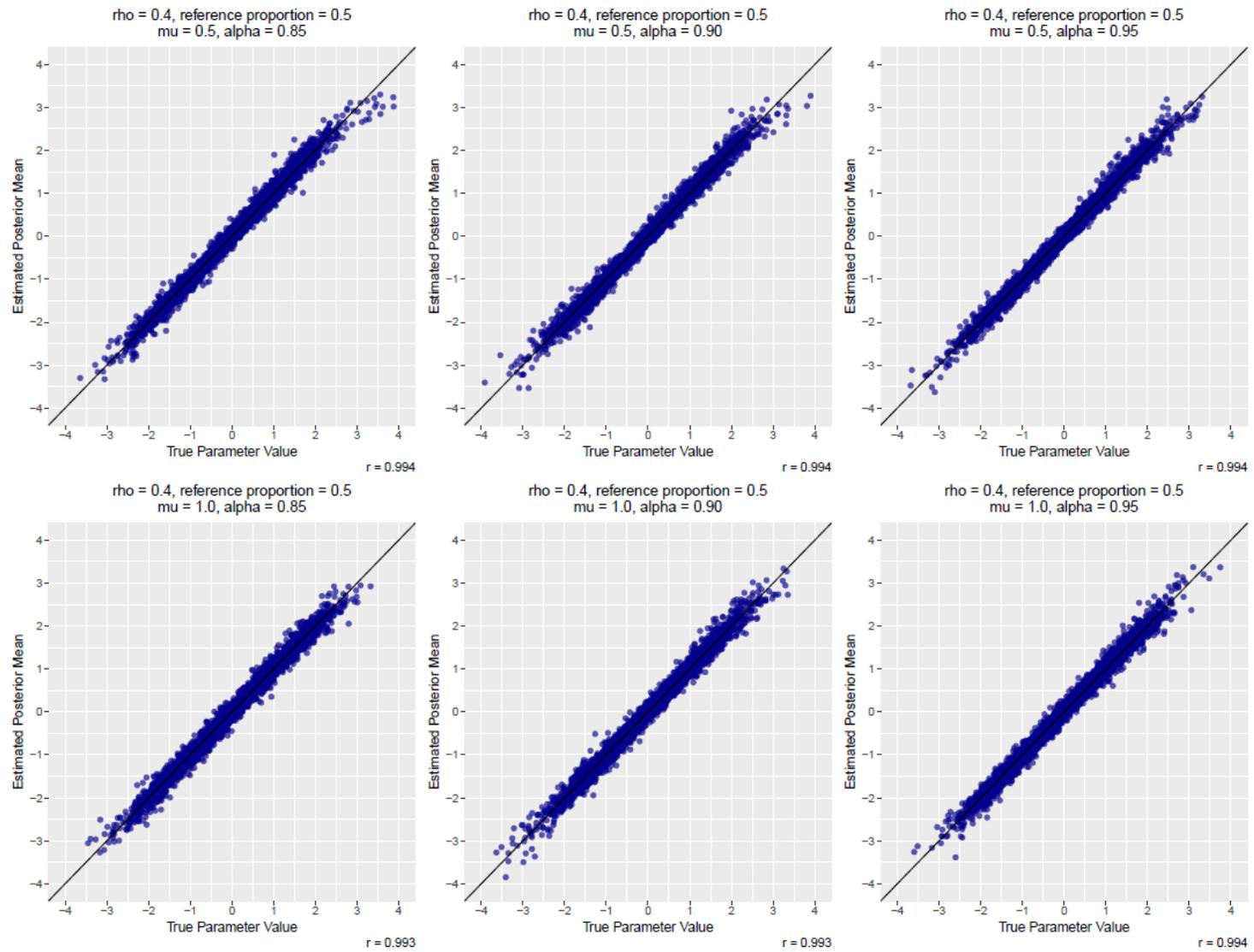
a-parameter scatterplots



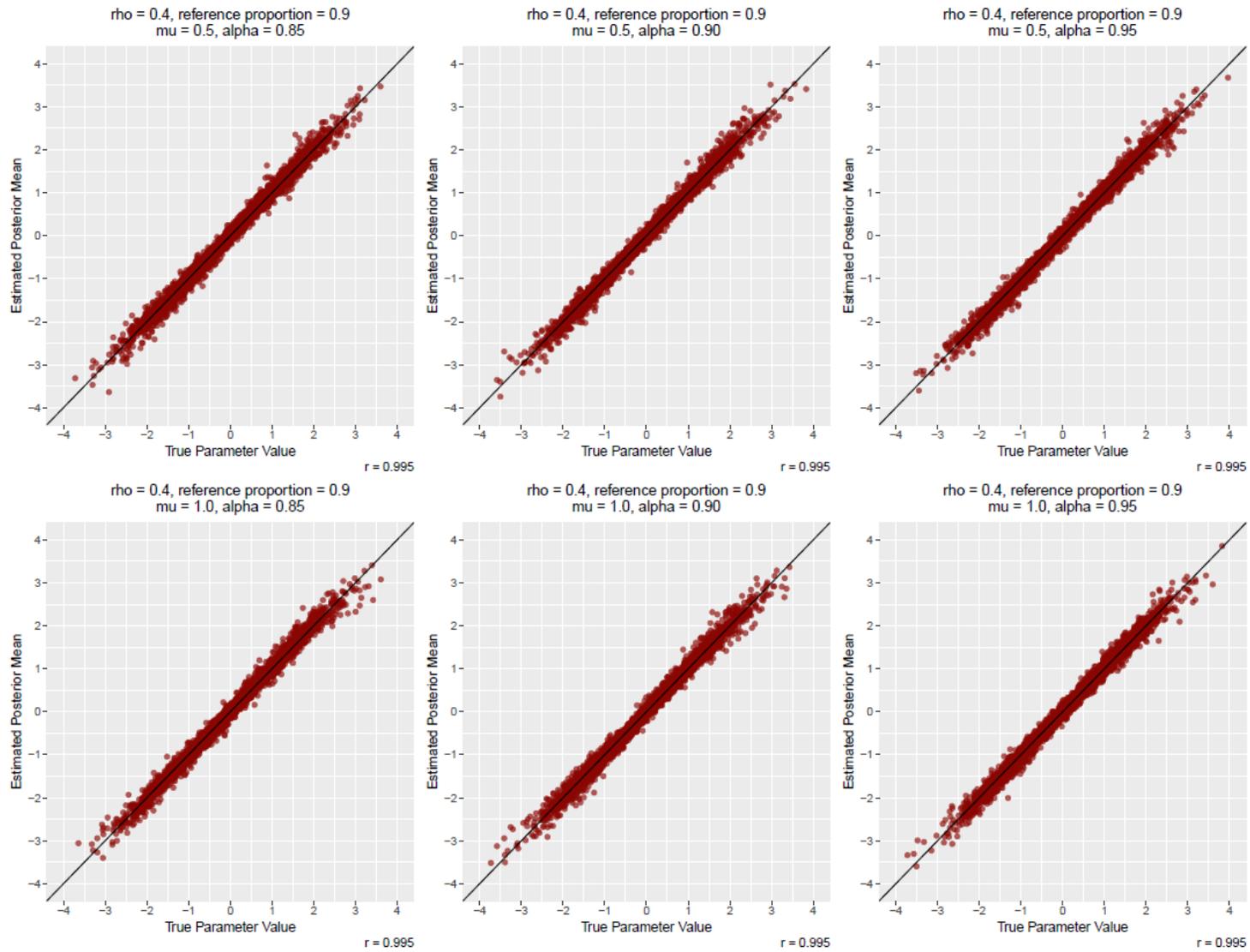
Appendix D. Correlation scatterplots for b -parameter recovery

Note: The solid diagonal line represents a perfect correlation of 1.

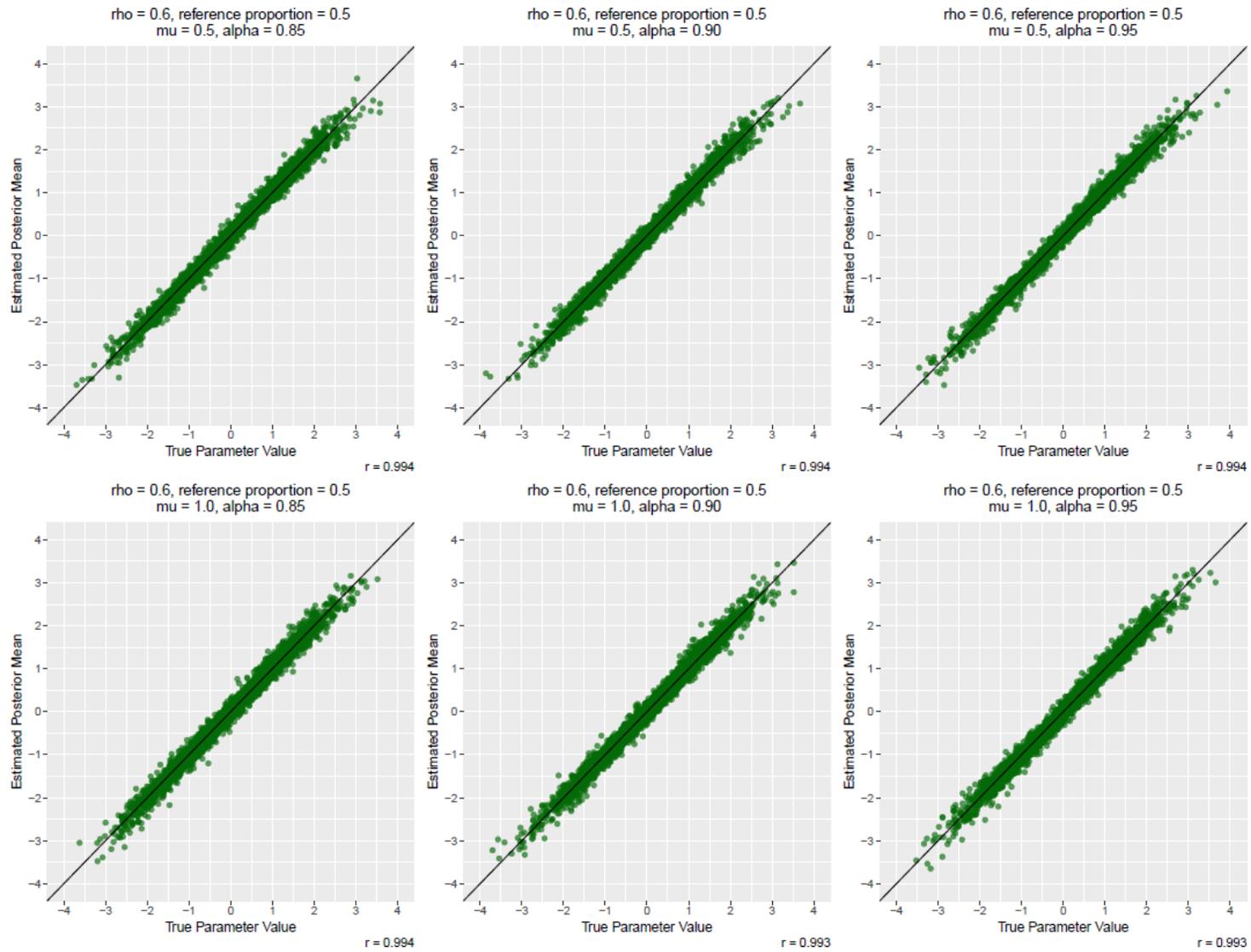
b-parameter scatterplots



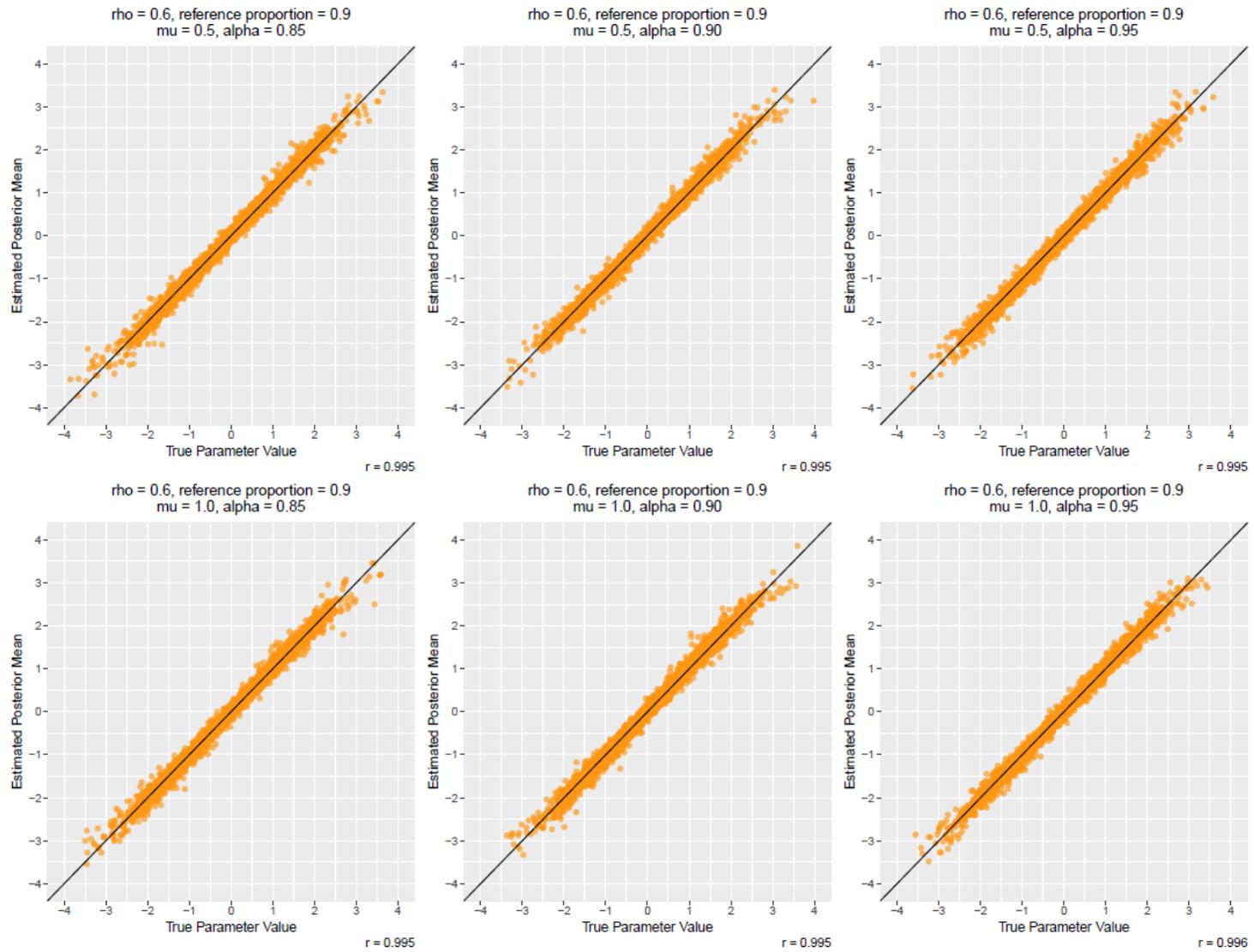
b-parameter scatterplots



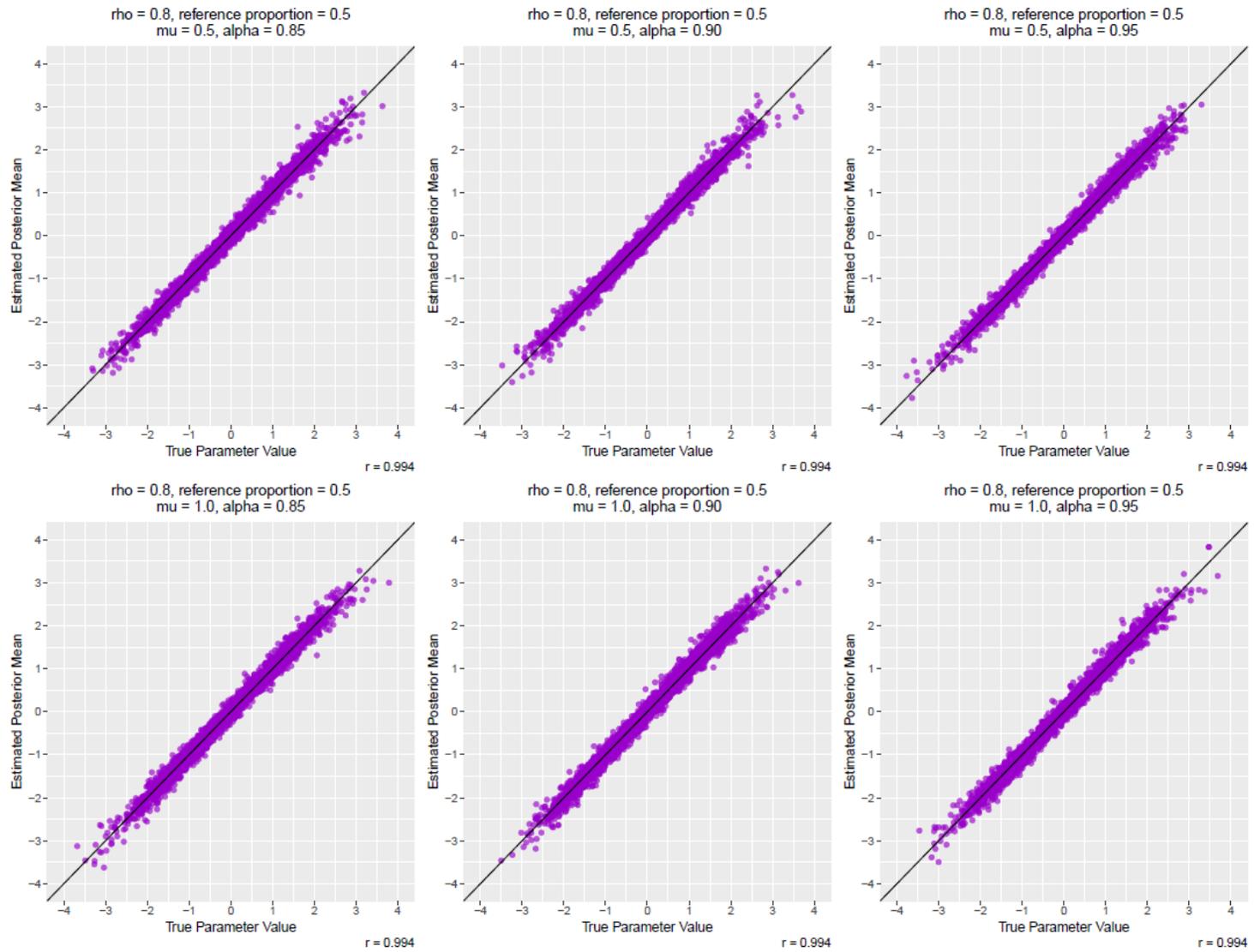
b-parameter scatterplots



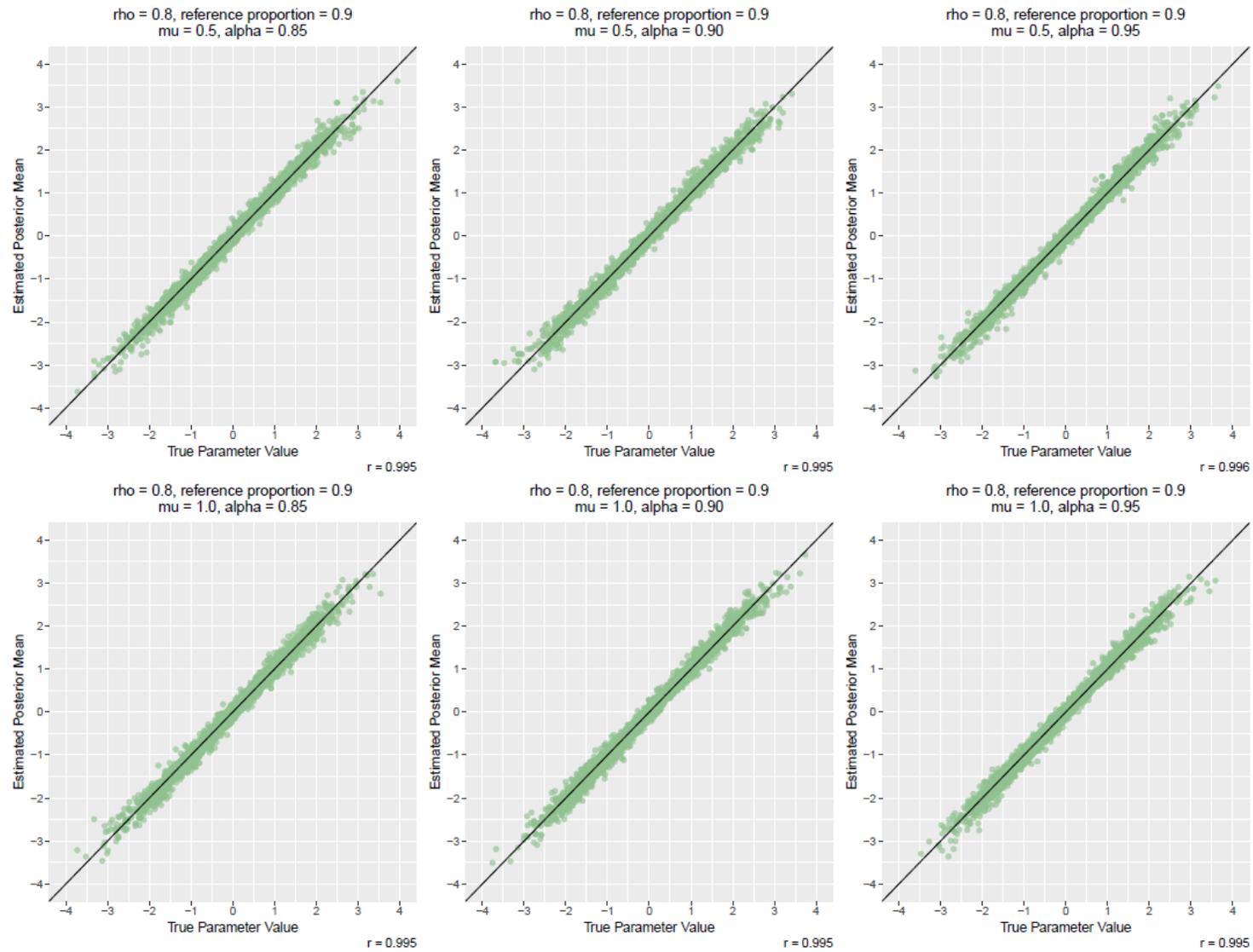
b-parameter scatterplots



b-parameter scatterplots



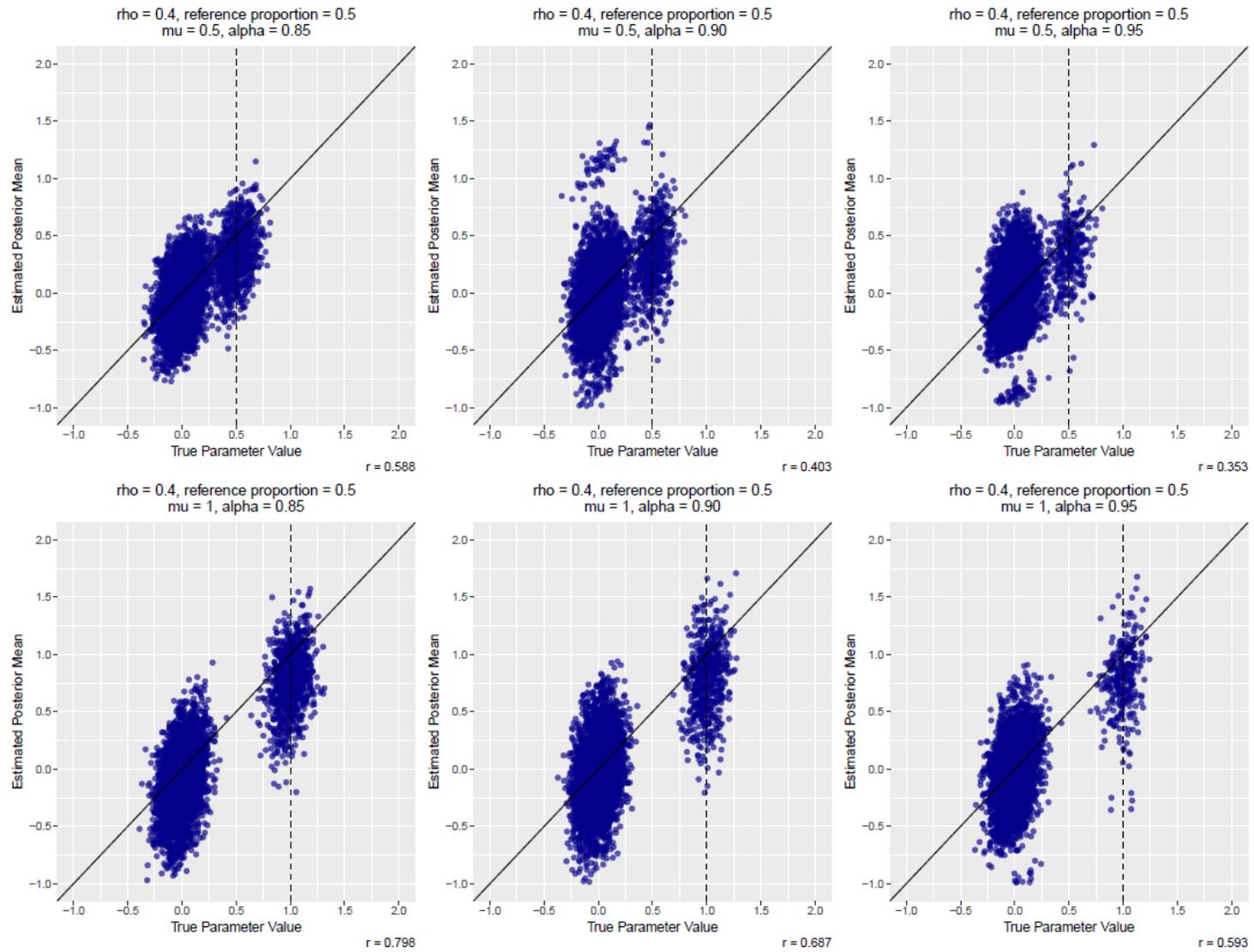
b-parameter scatterplots



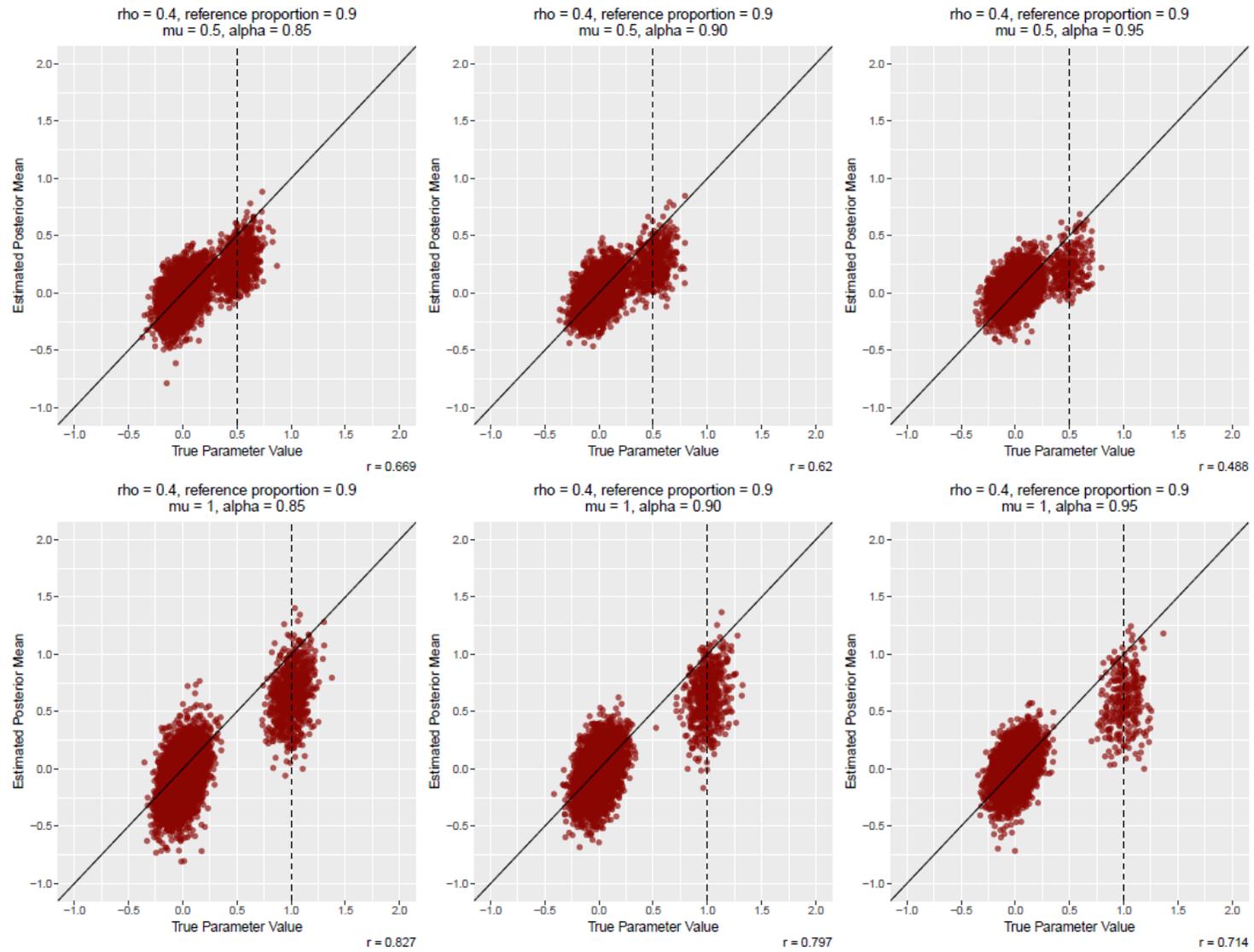
Appendix E. Correlation scatterplots for D -parameter recovery

Note: The solid diagonal line represents a perfect correlation of 1. The dotted vertical line shows μ_2 , the mean of the second distribution used to create the mixture distribution of D -parameters.

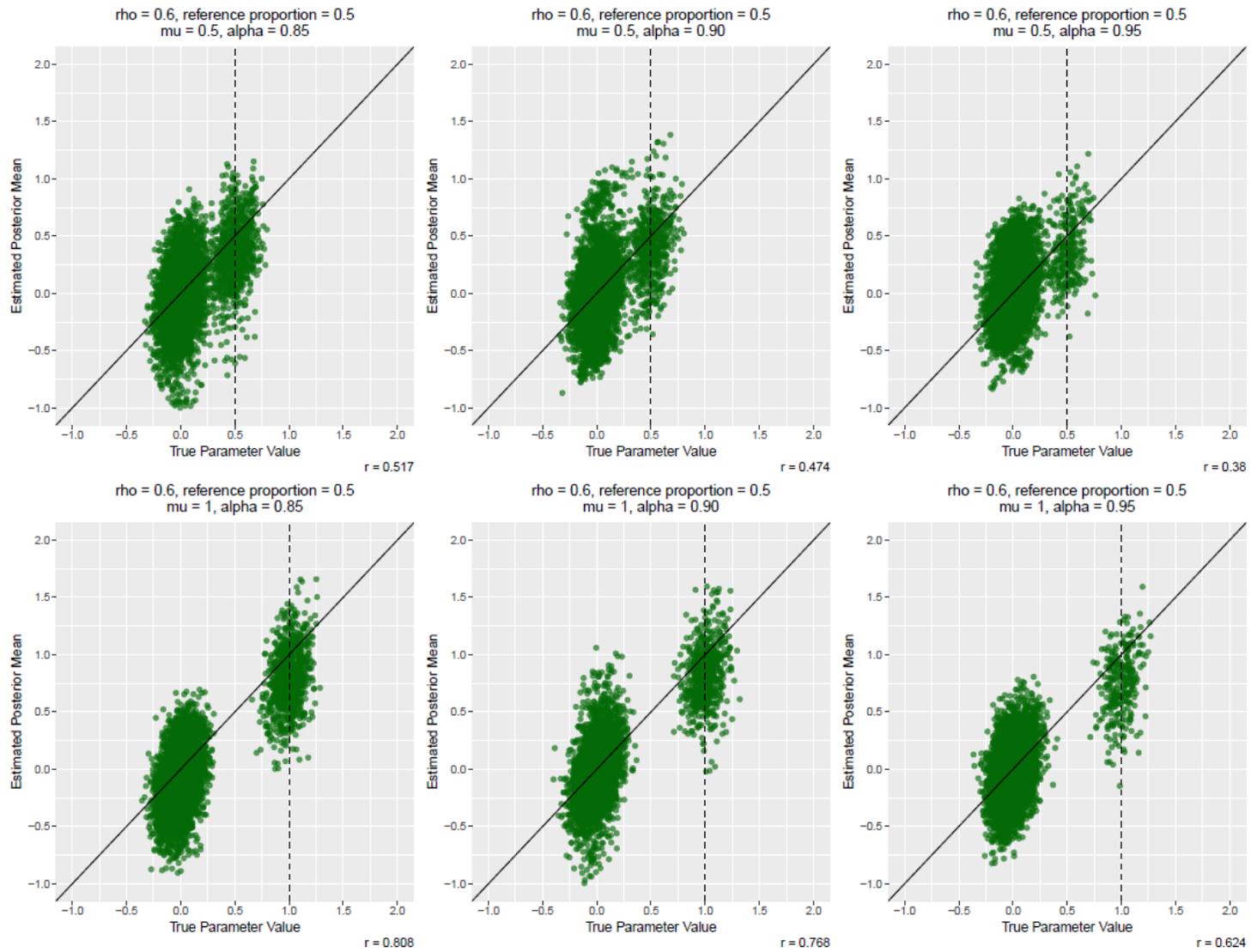
D-parameter scatterplots



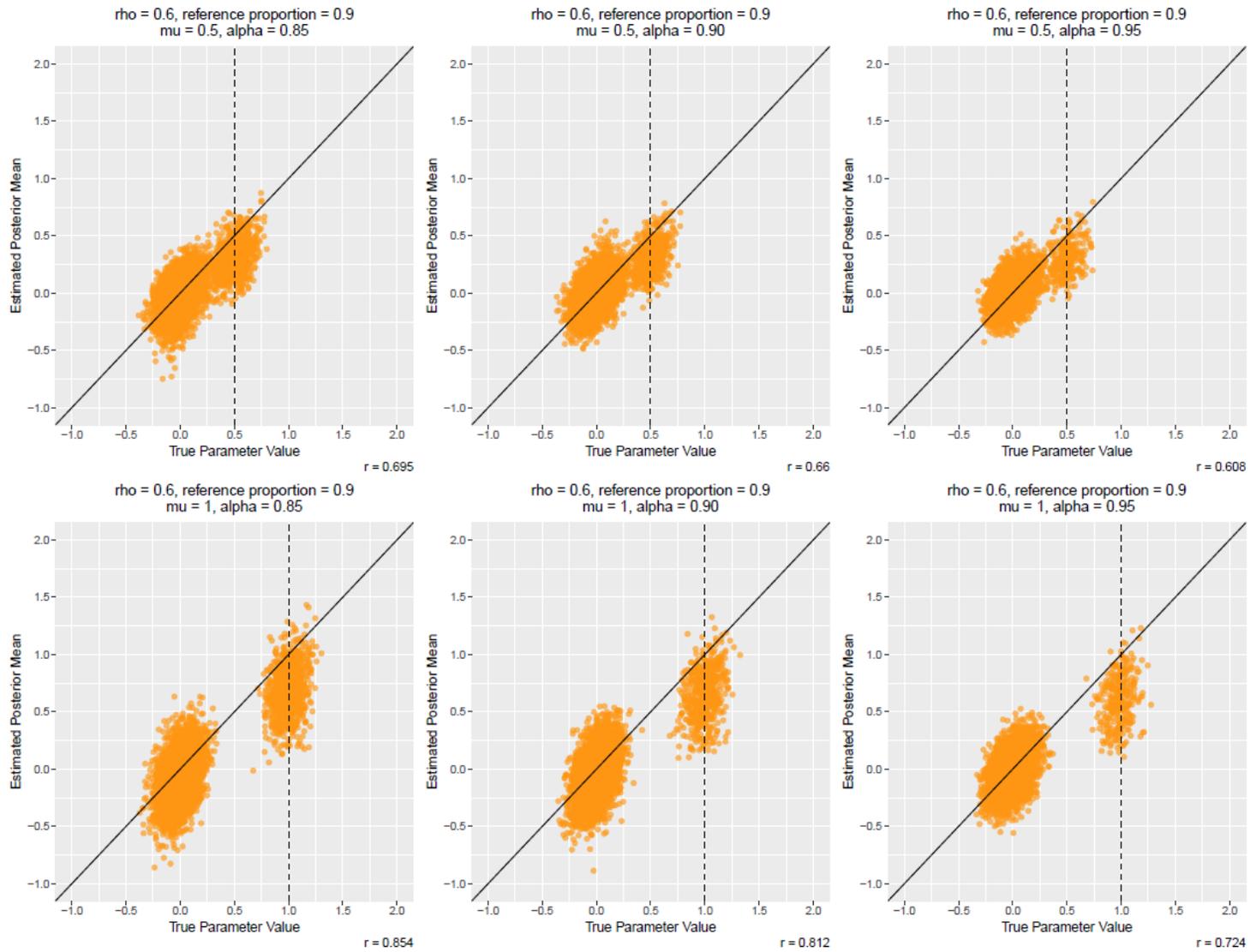
D-parameter scatterplots



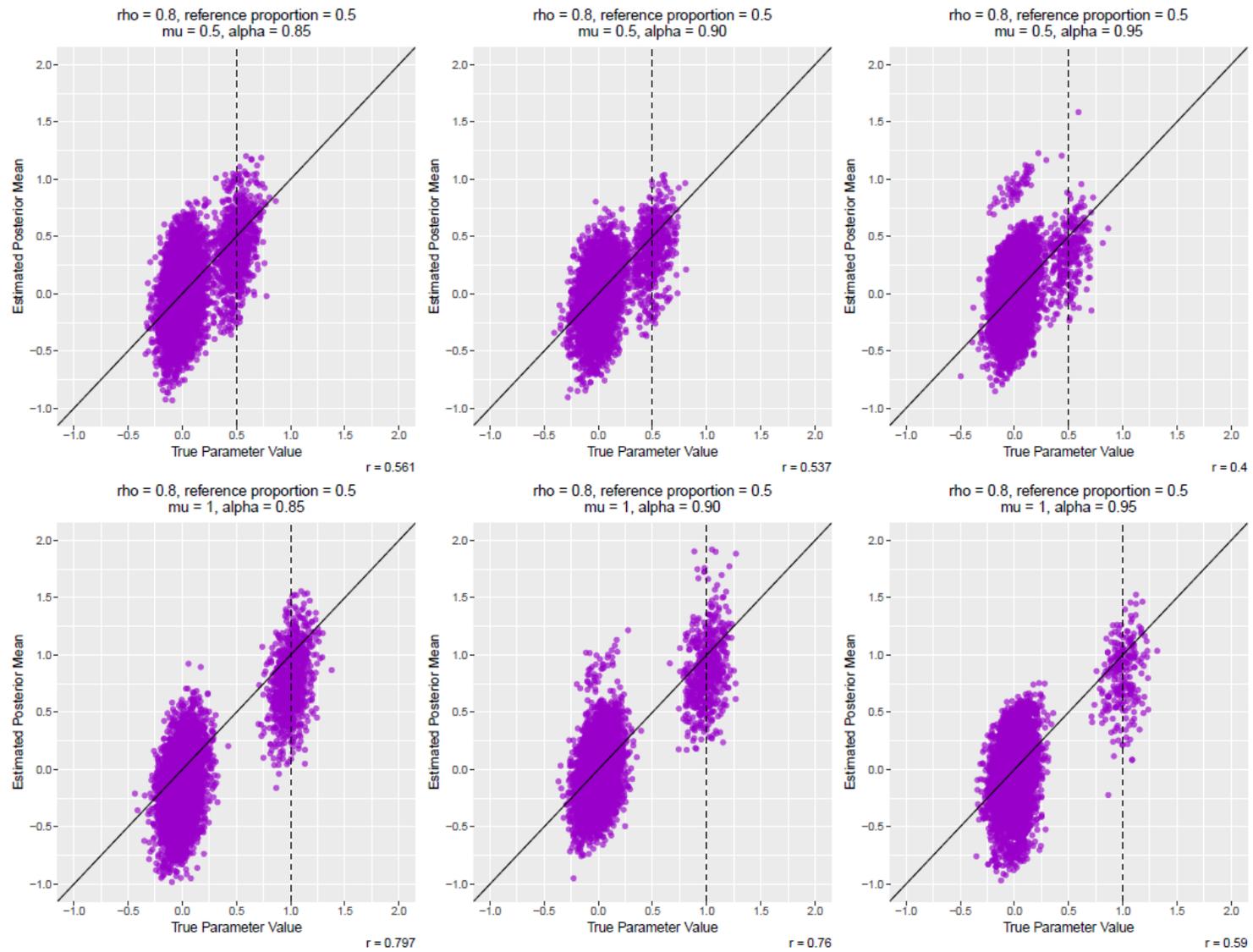
D-parameter scatterplots



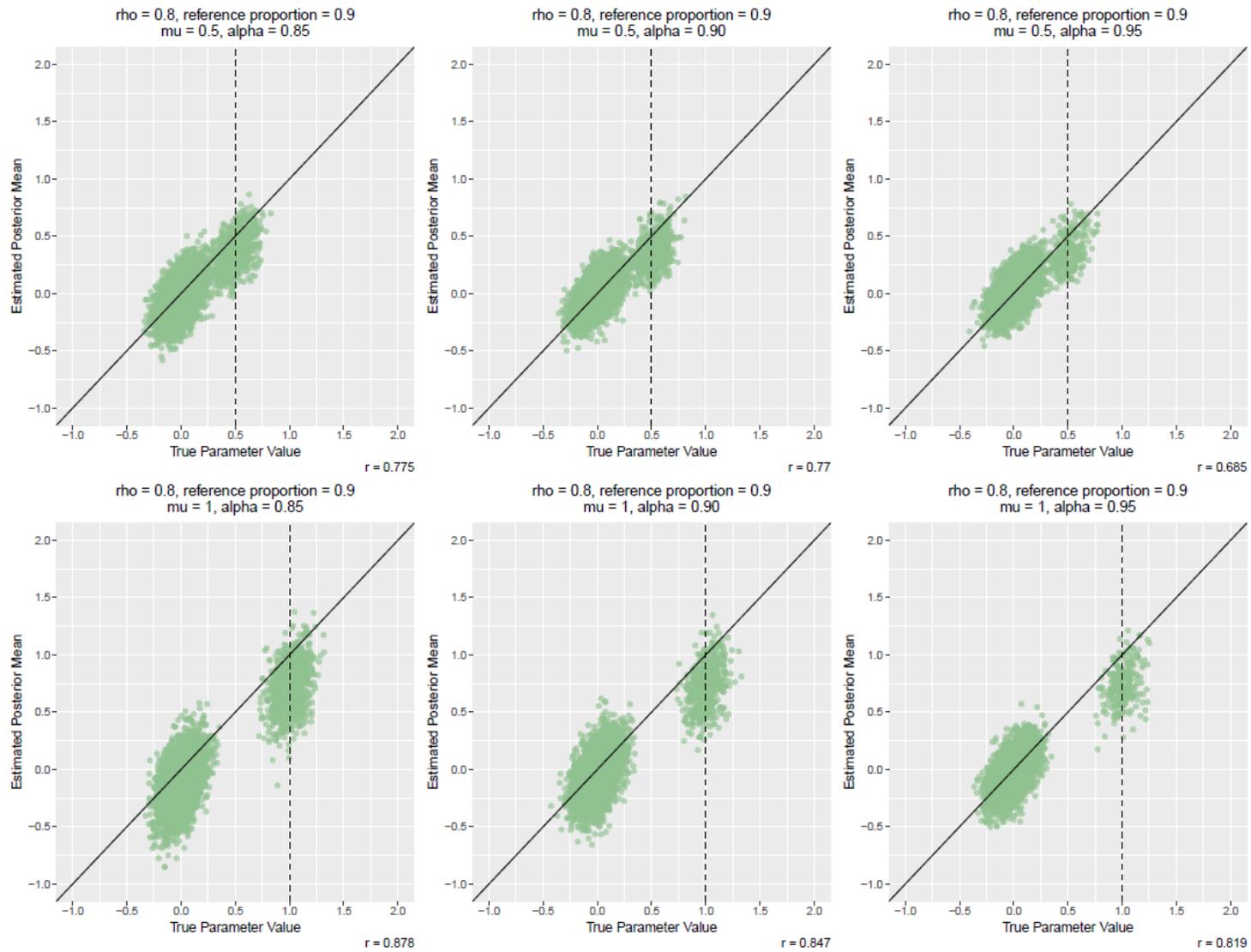
D-parameter scatterplots



D-parameter scatterplots



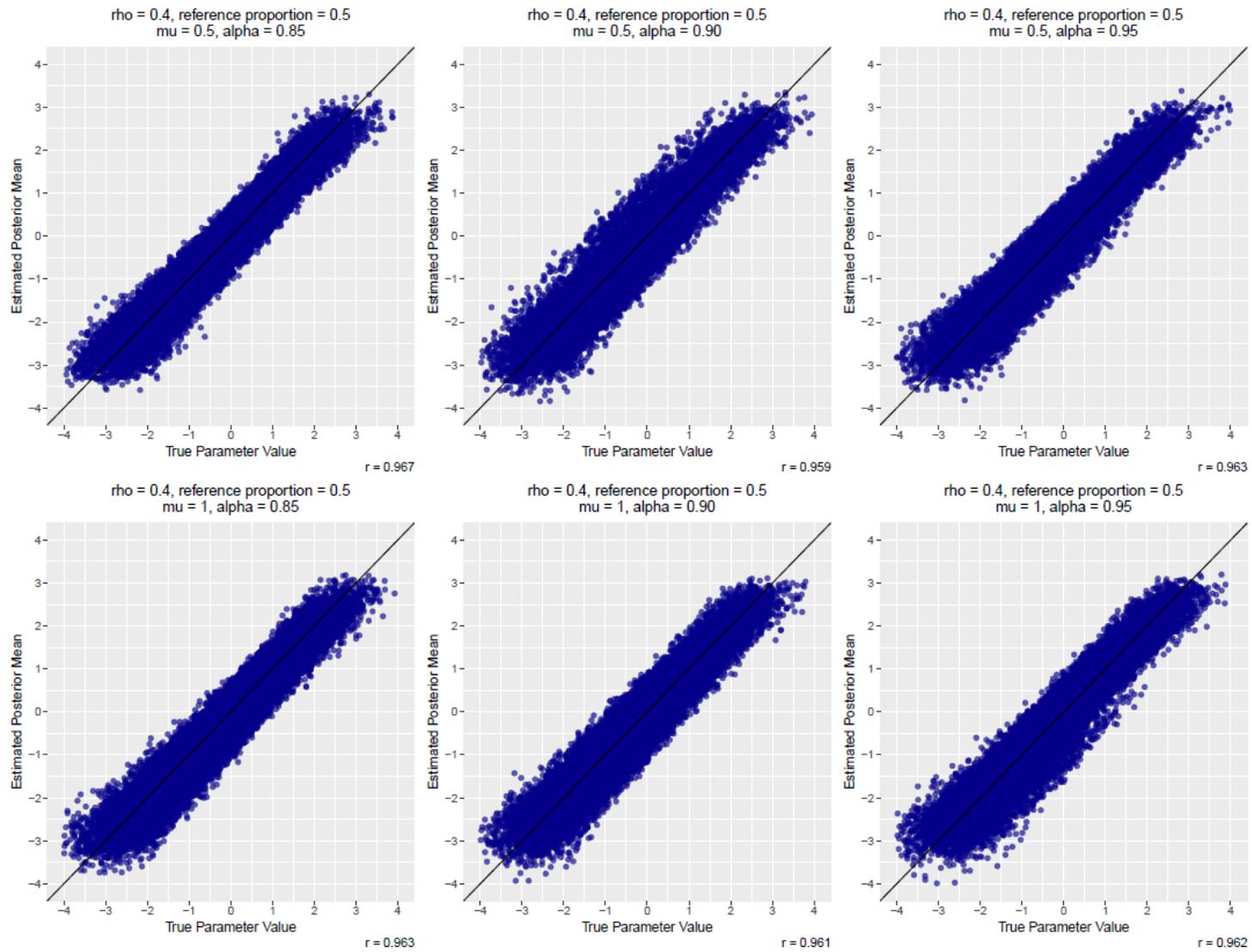
D-parameter scatterplots



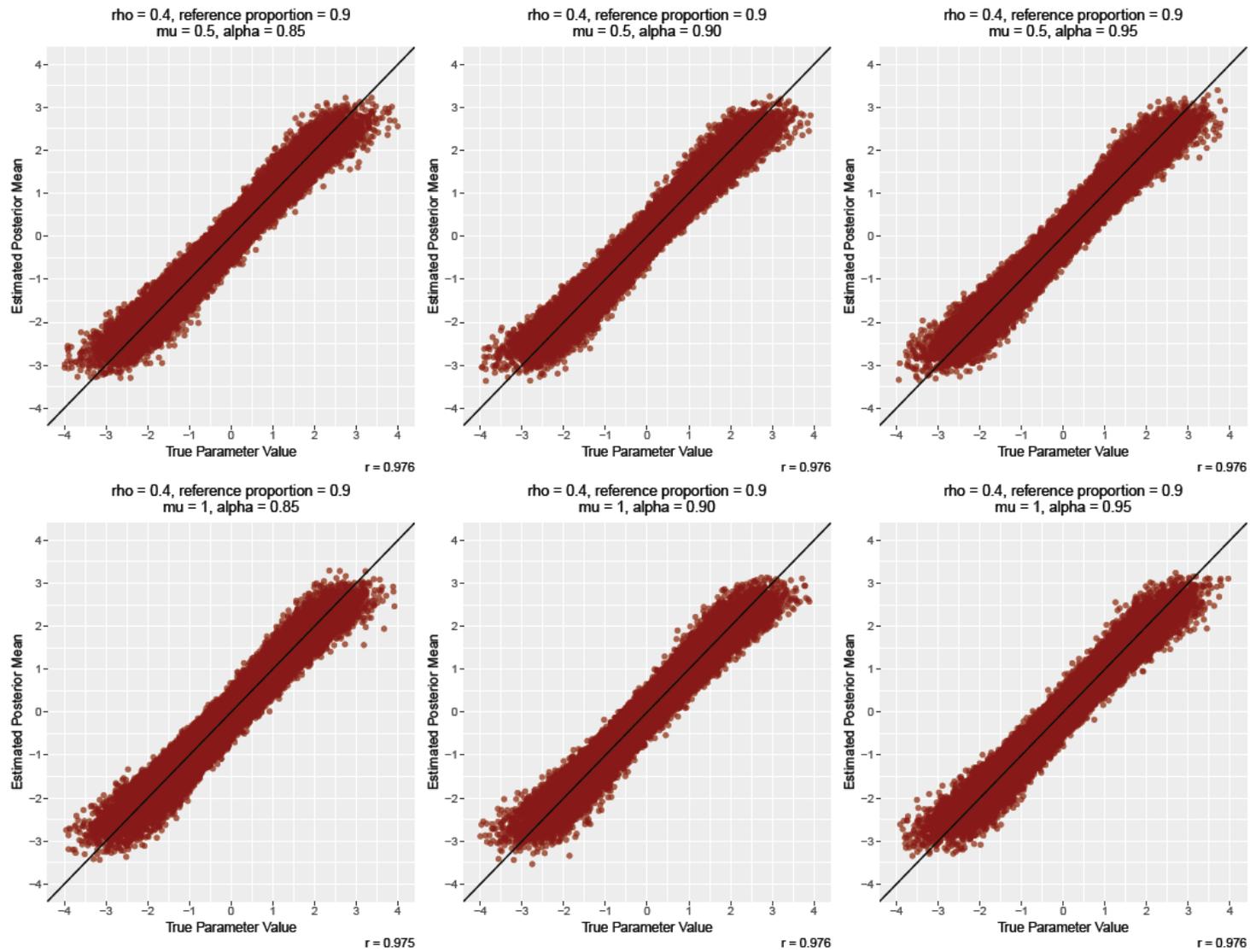
Appendix F. Correlation scatterplots for θ recovery

Note: The solid diagonal line represents a perfect correlation of 1.

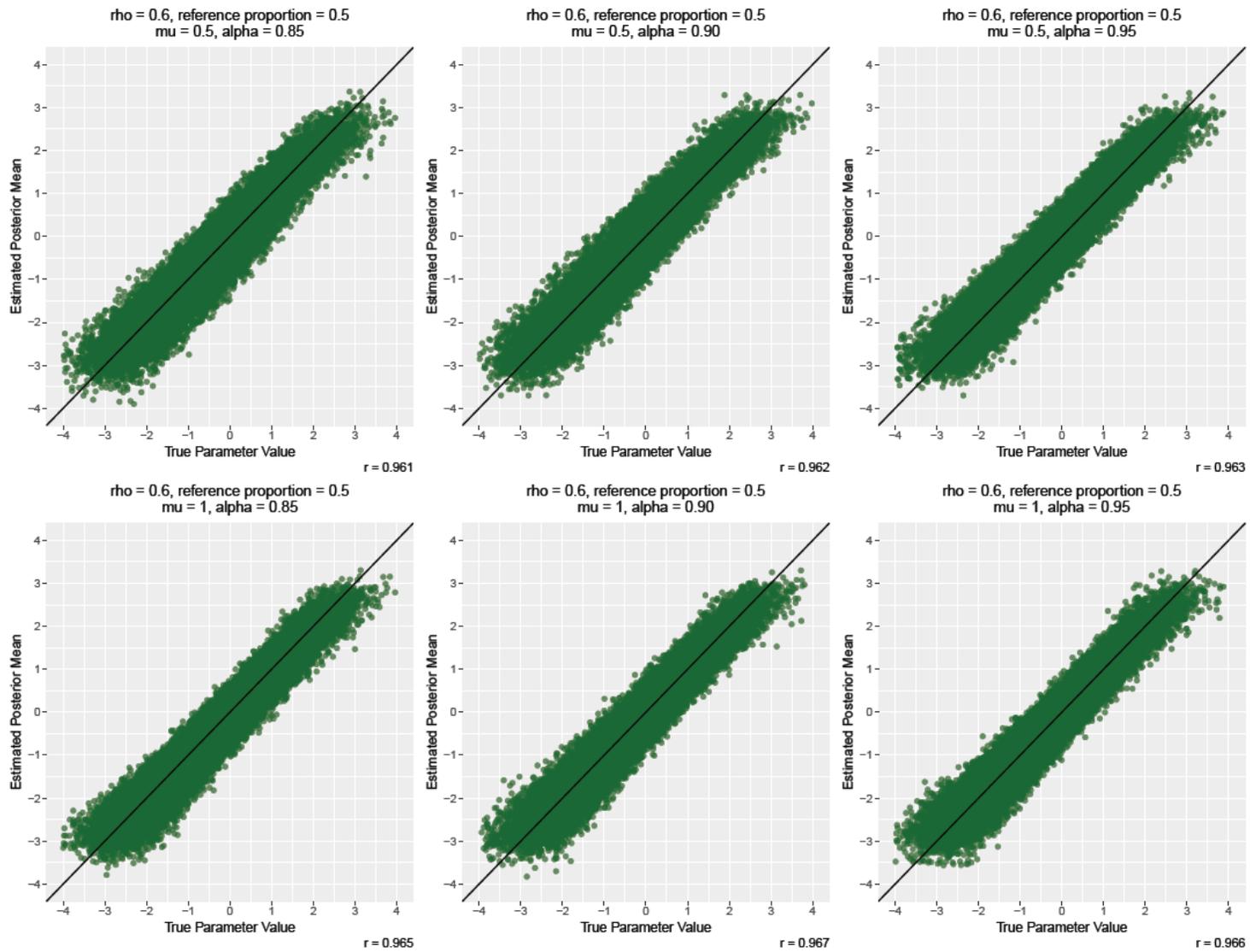
theta-parameter scatterplots



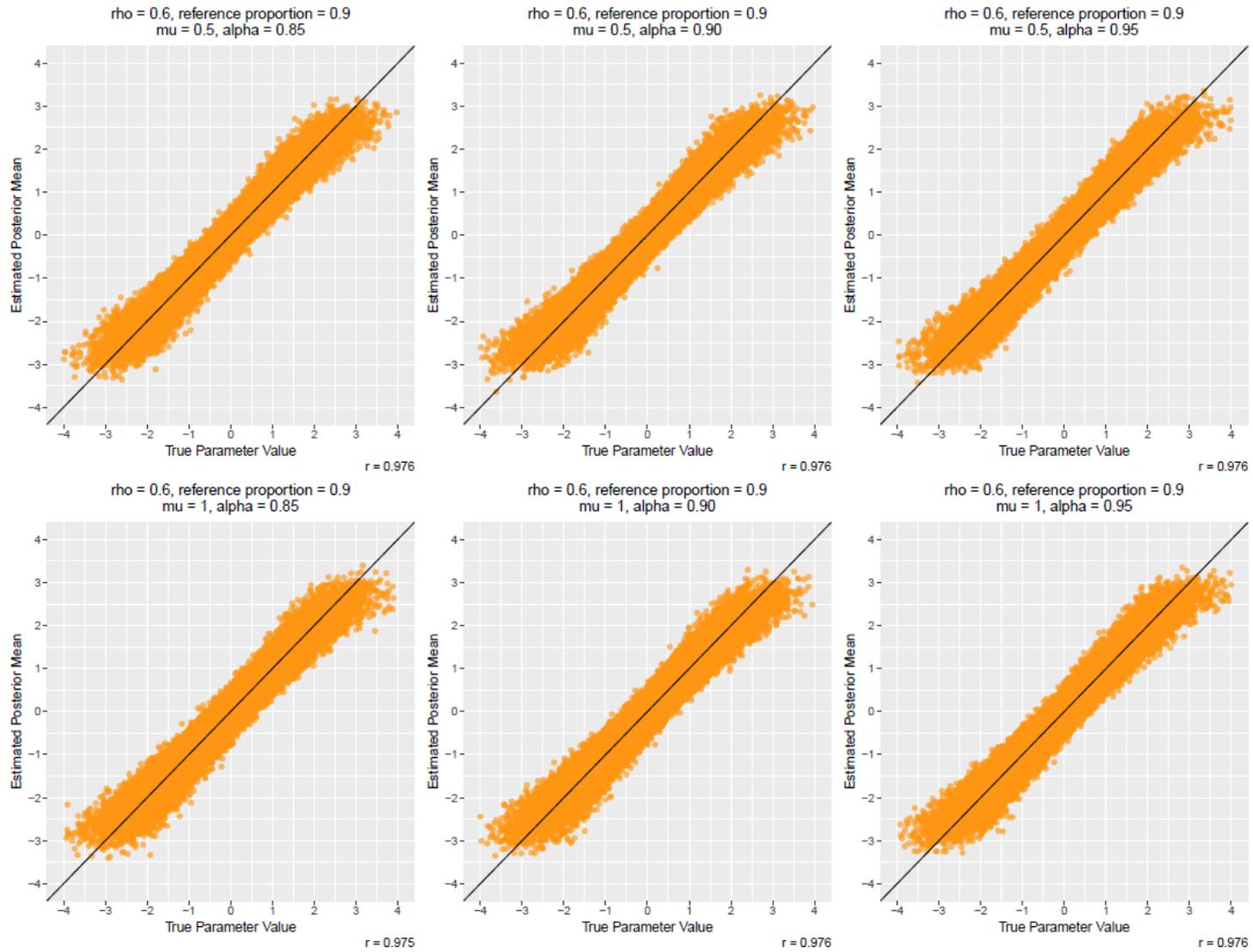
theta-parameter scatterplots



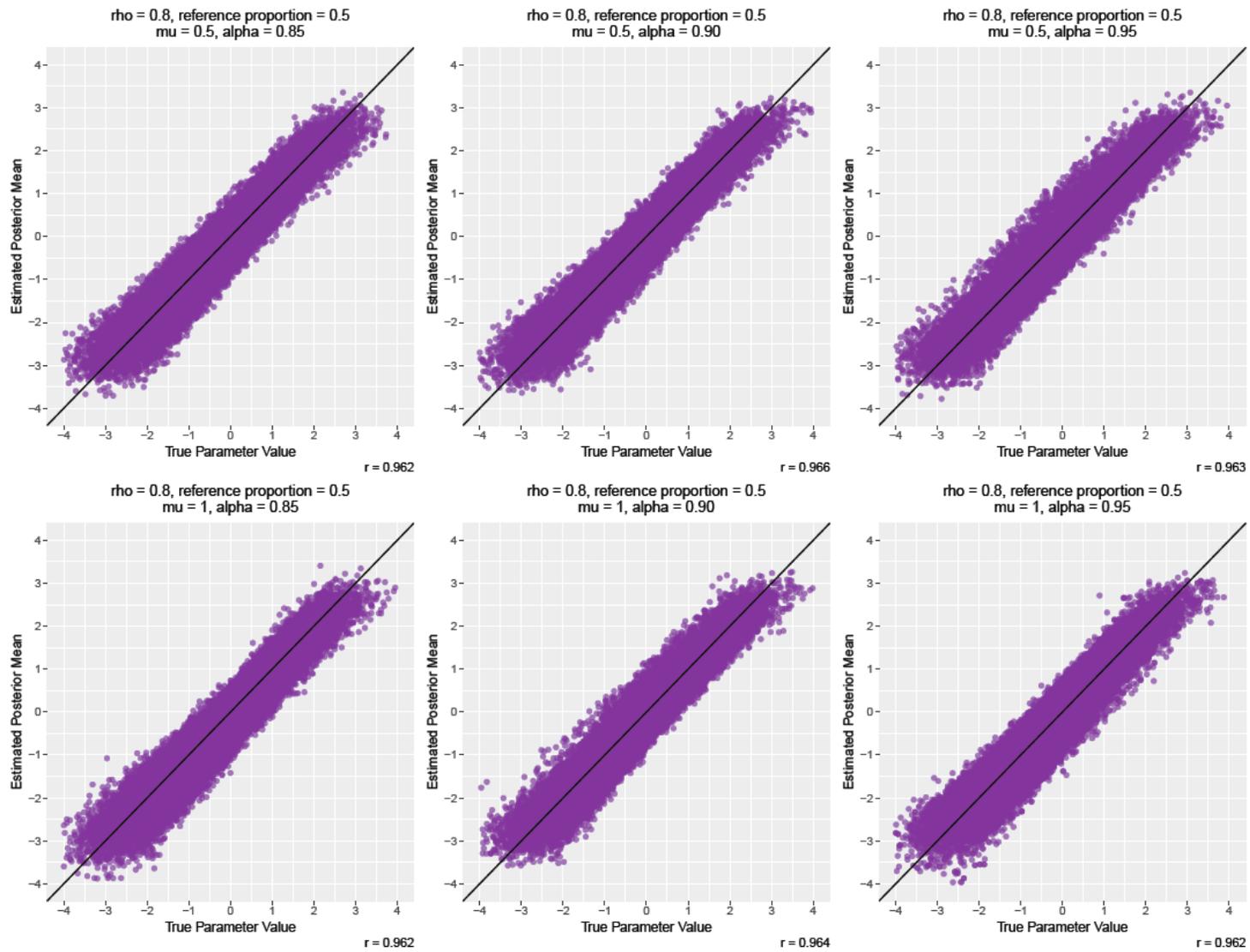
theta-parameter scatterplots



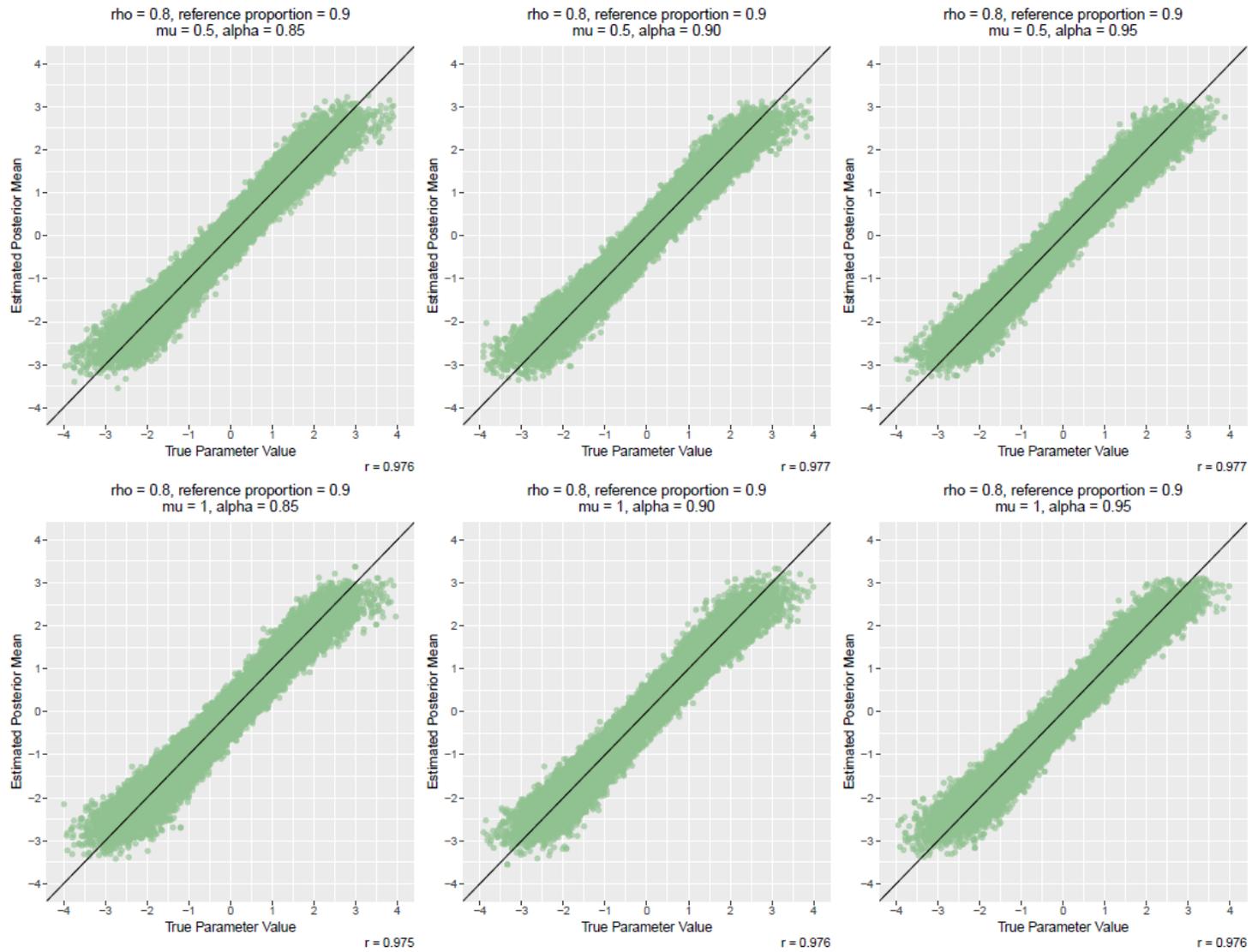
theta-parameter scatterplots



theta-parameter scatterplots



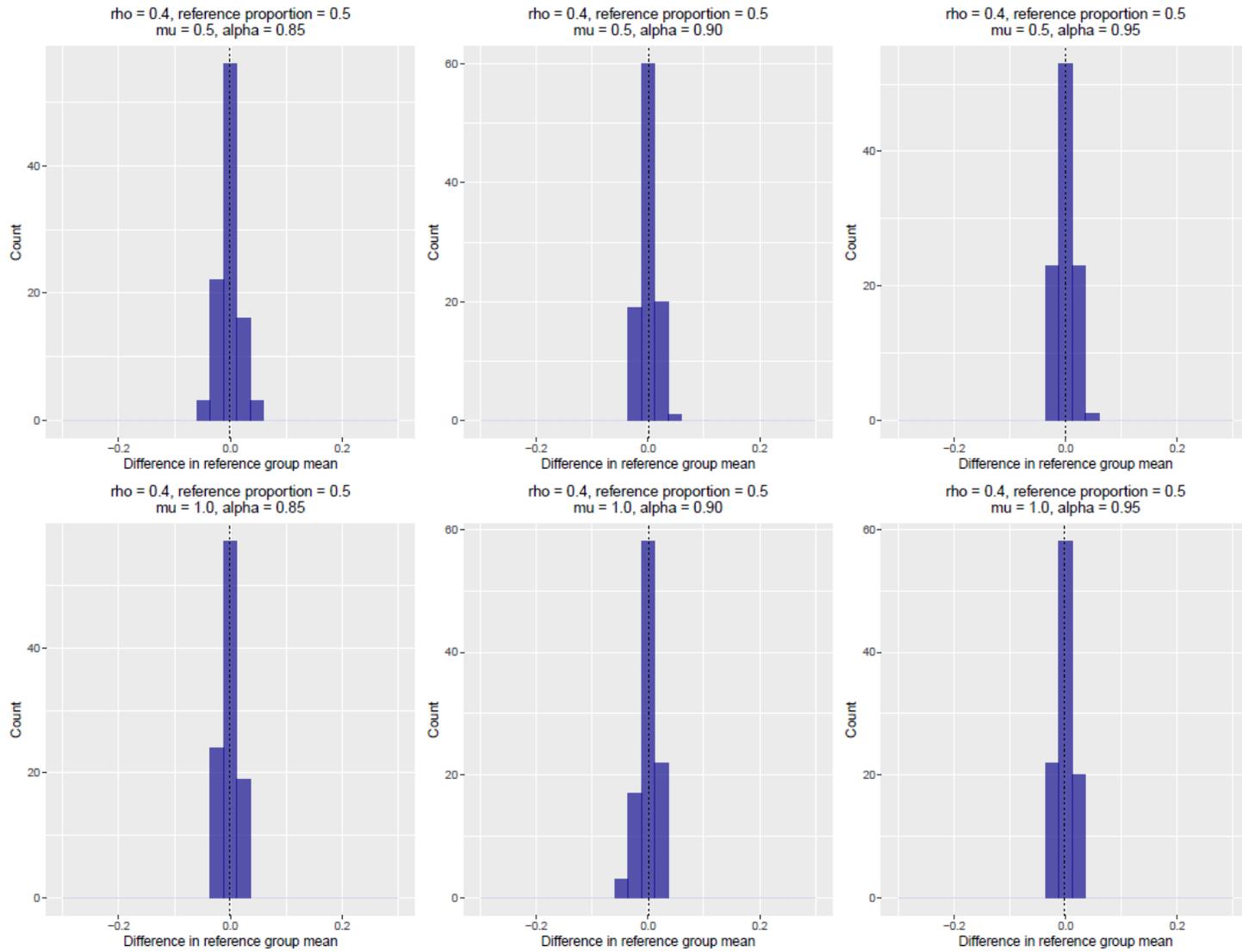
theta-parameter scatterplots



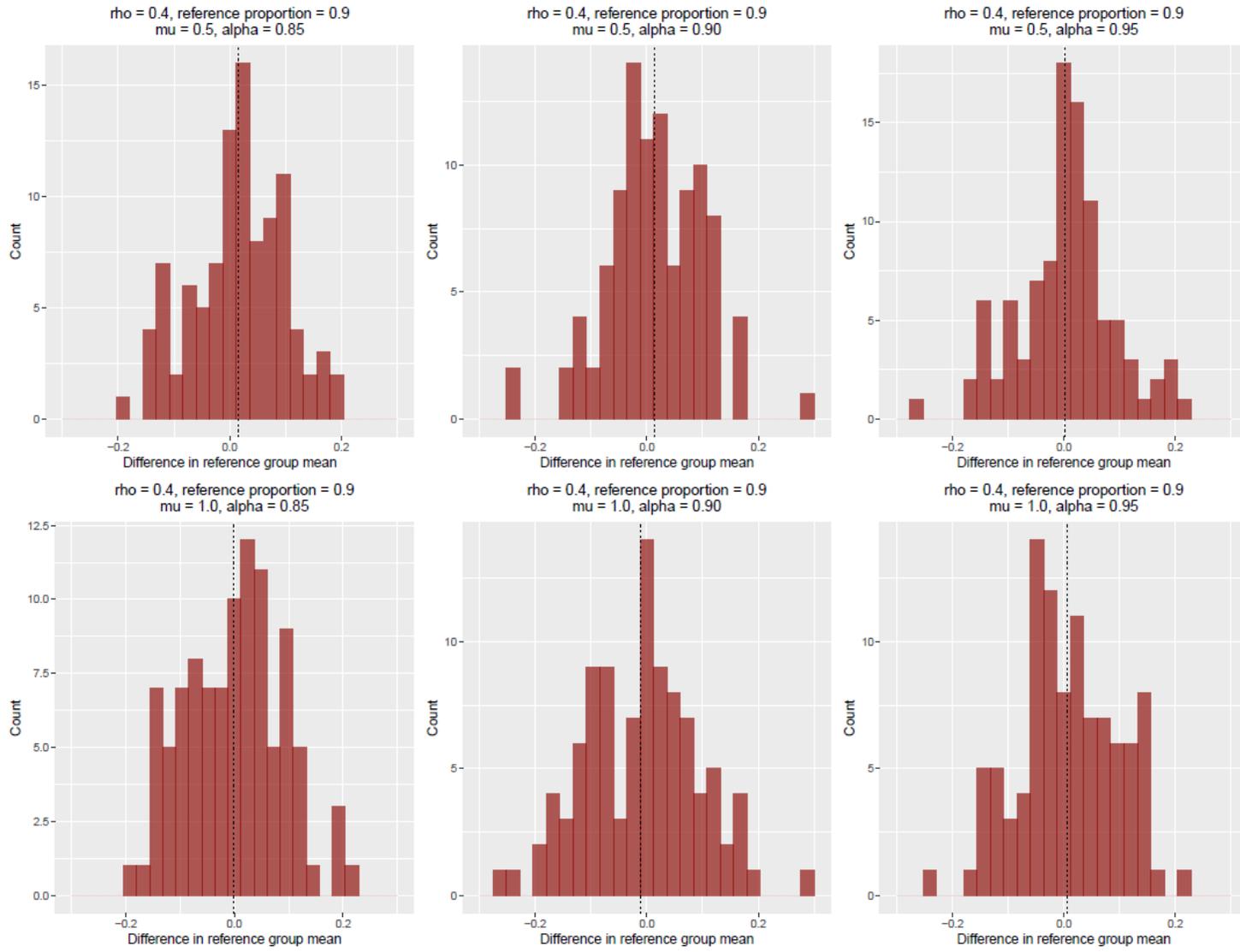
Appendix G. Mean recovery histograms for the reference group means

Note: The dotted vertical line represents the mean of the distribution.

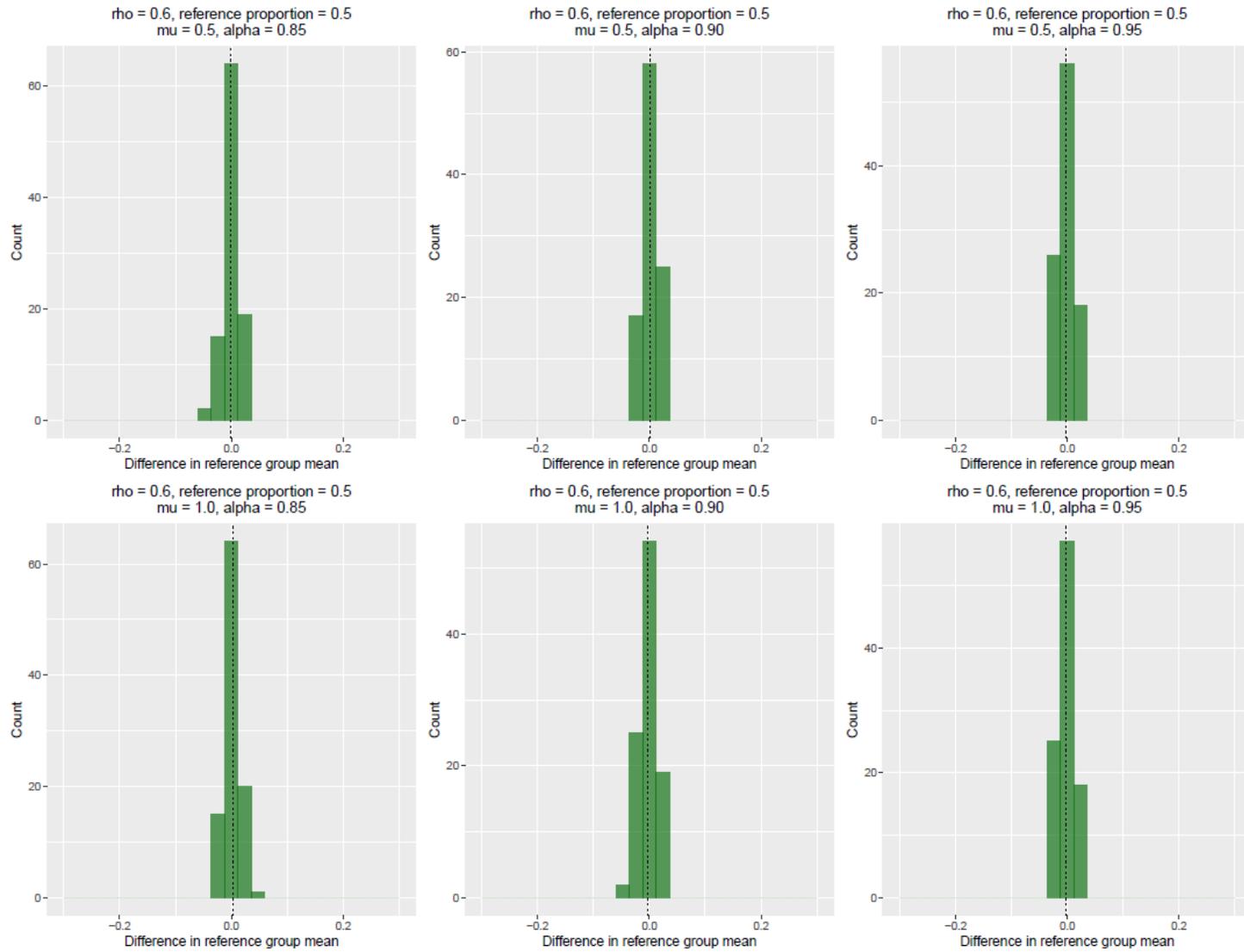
Reference mean recovery bias



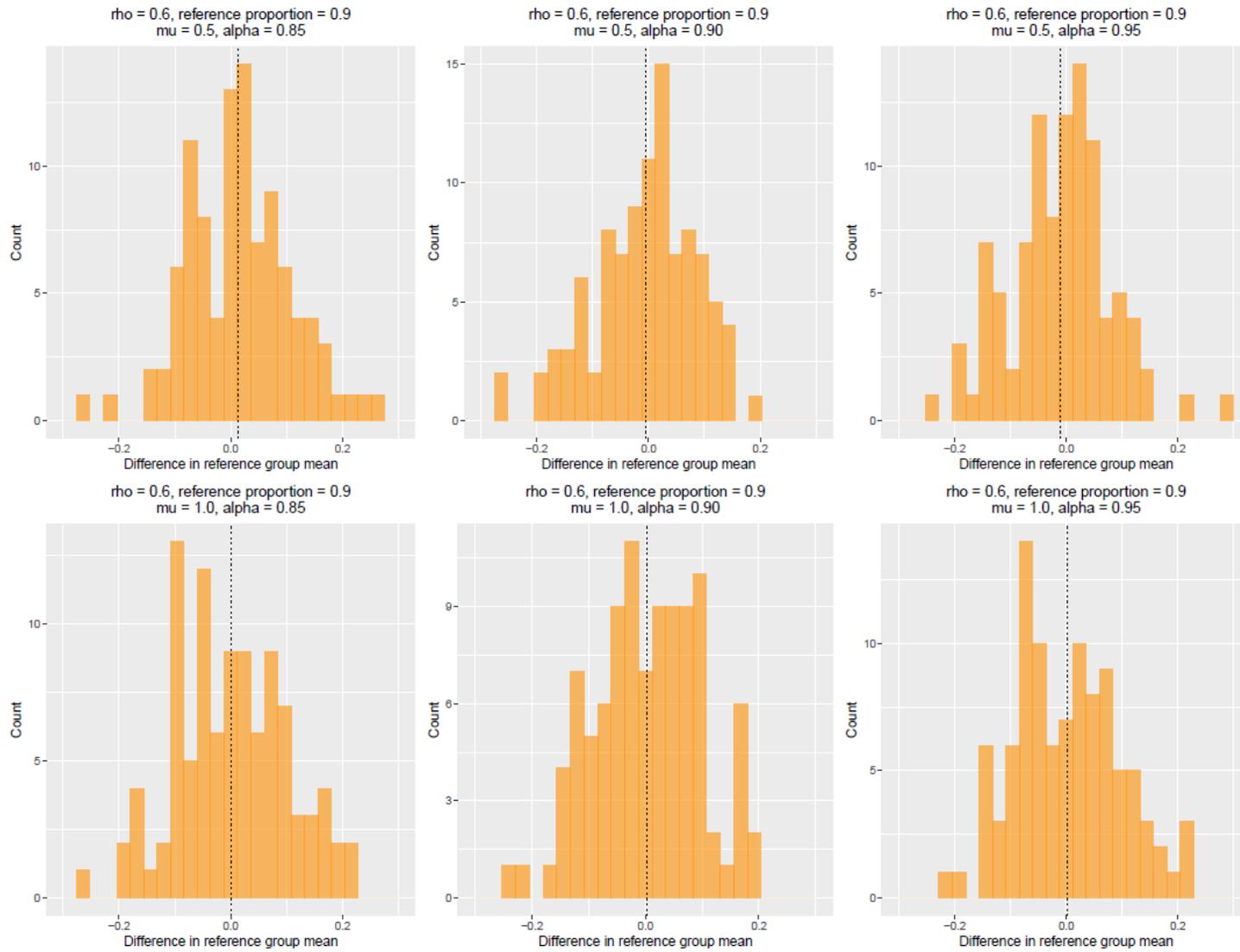
Reference mean recovery bias



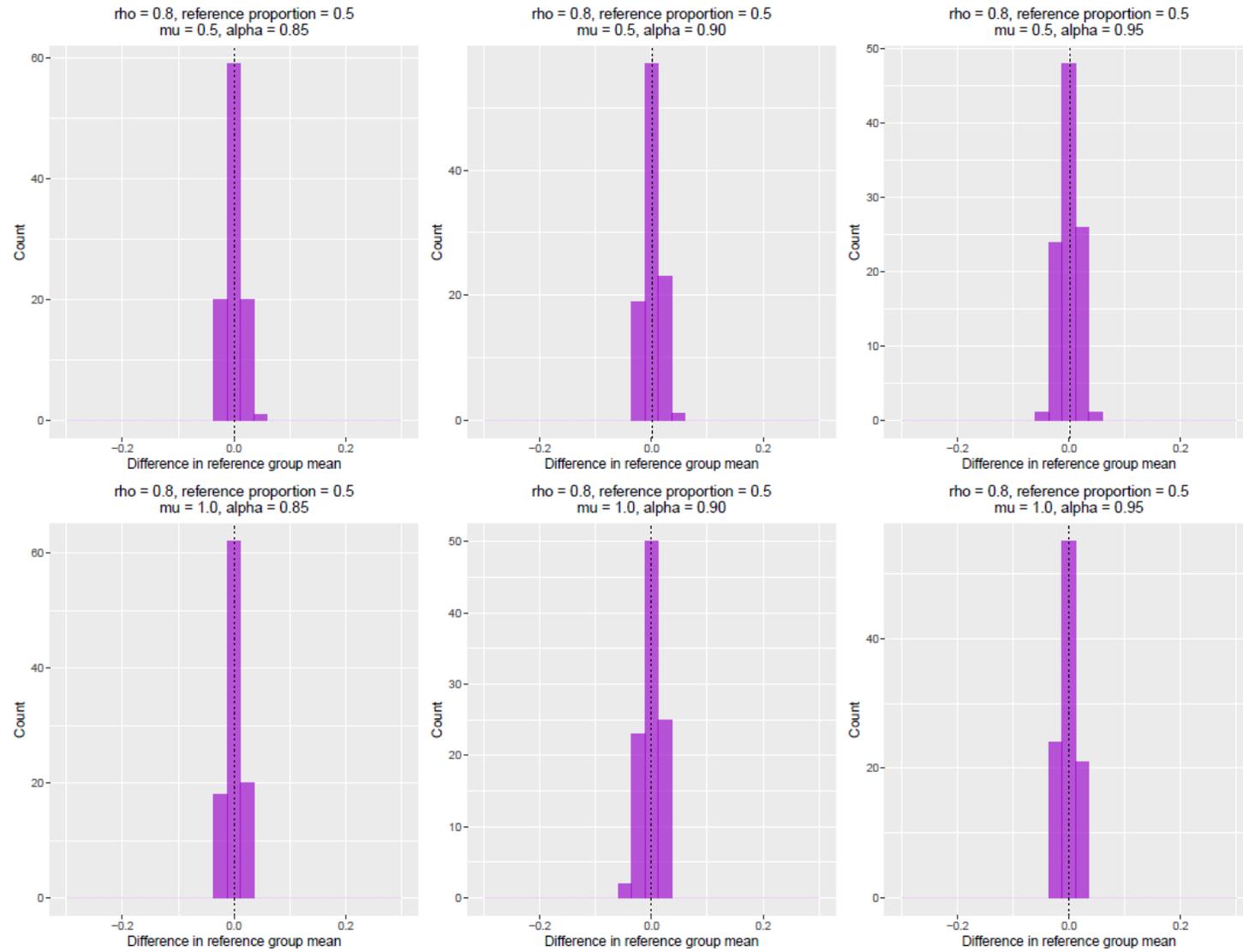
Reference mean recovery bias



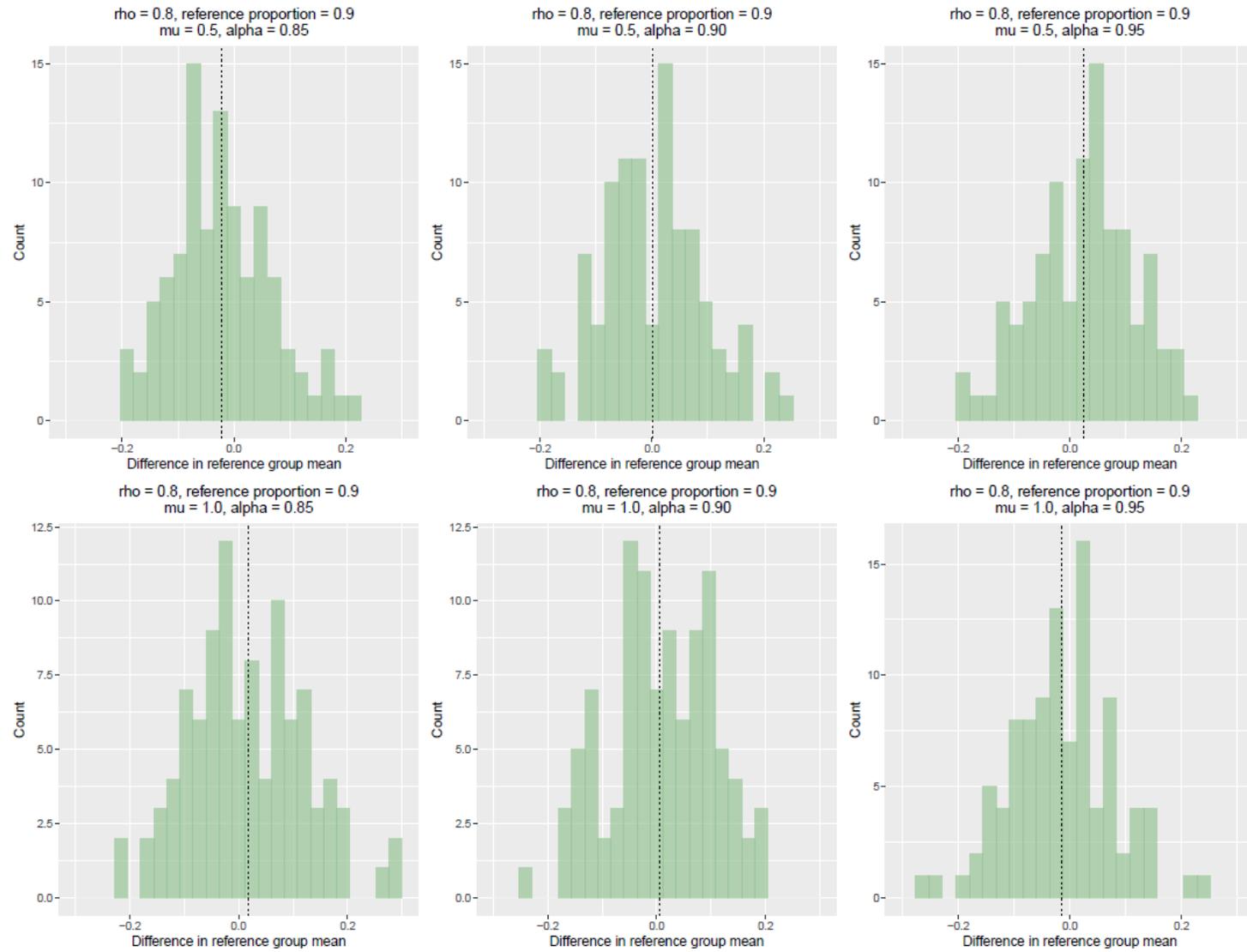
Reference mean recovery bias



Reference mean recovery bias



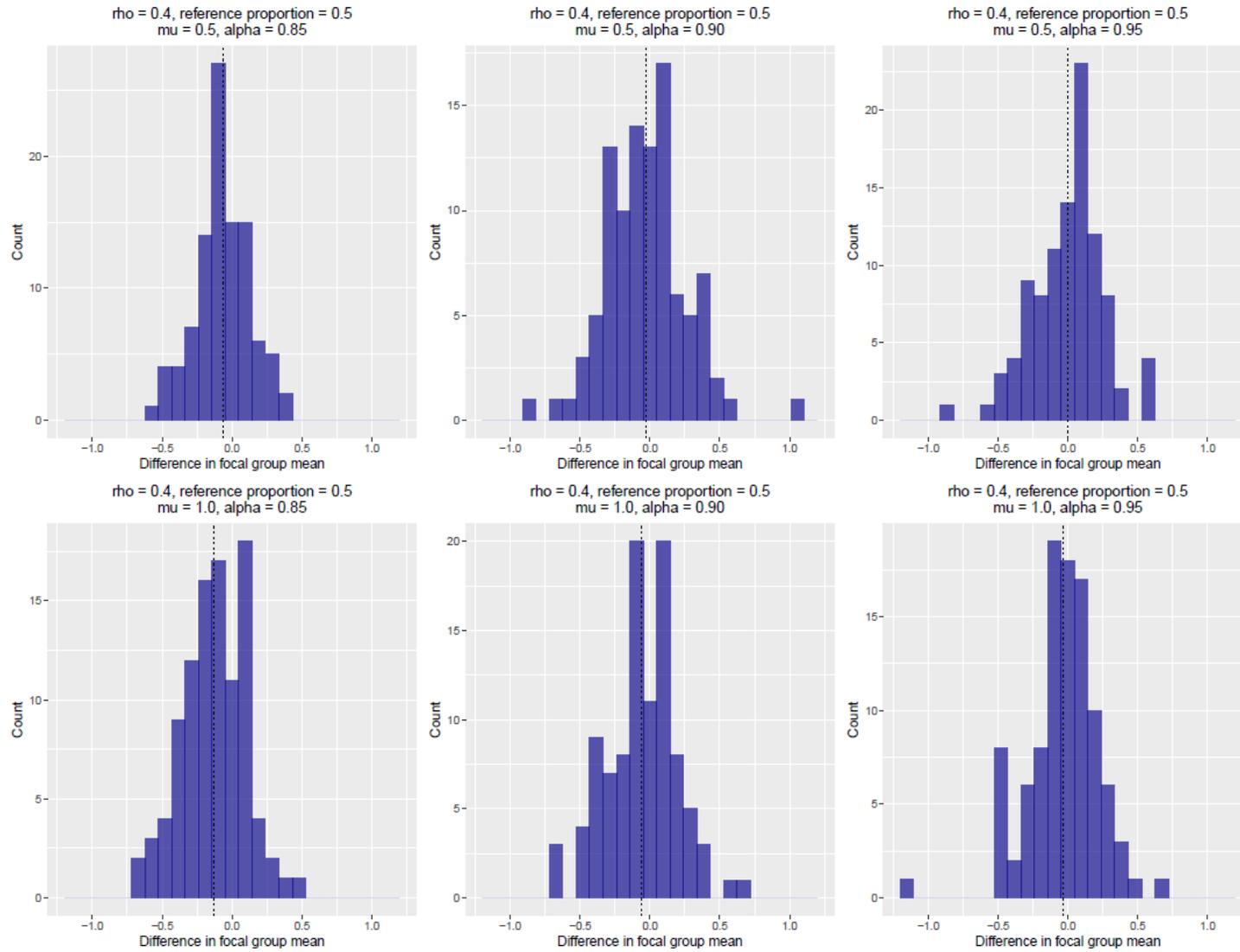
Reference mean recovery bias



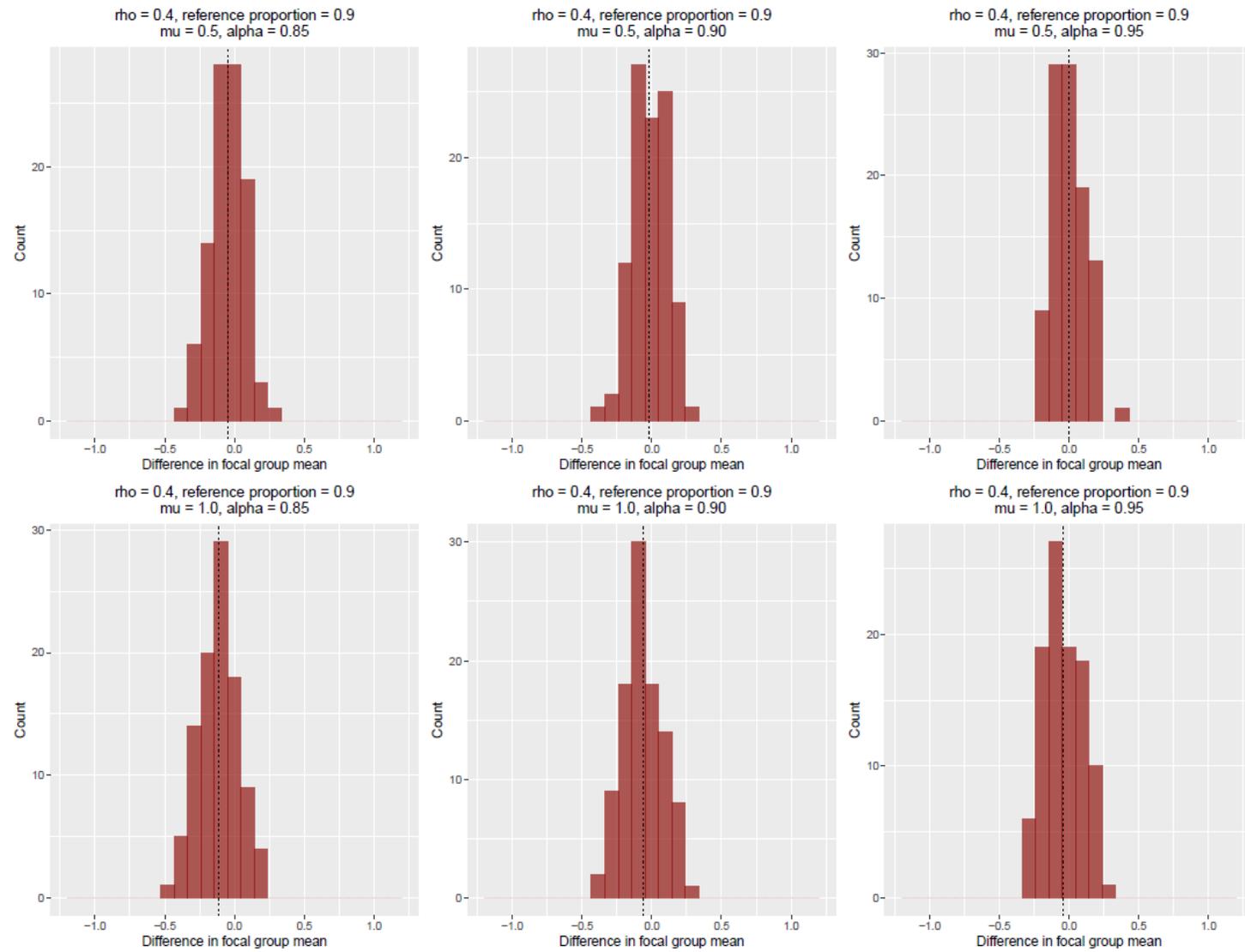
Appendix H. Mean recovery histograms for the focal group means

Note: The dotted vertical line represents the mean of the distribution.

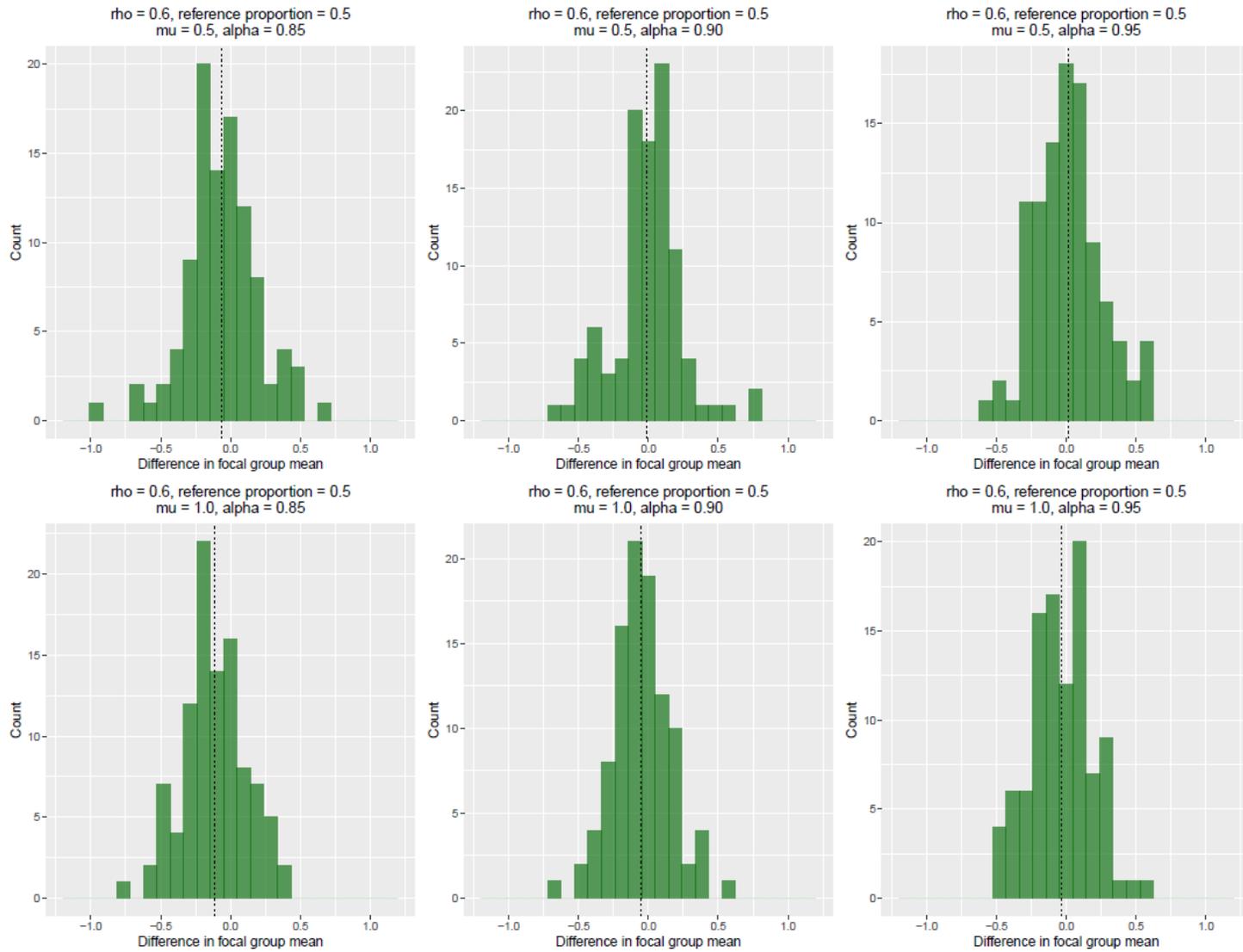
Focal mean recovery bias



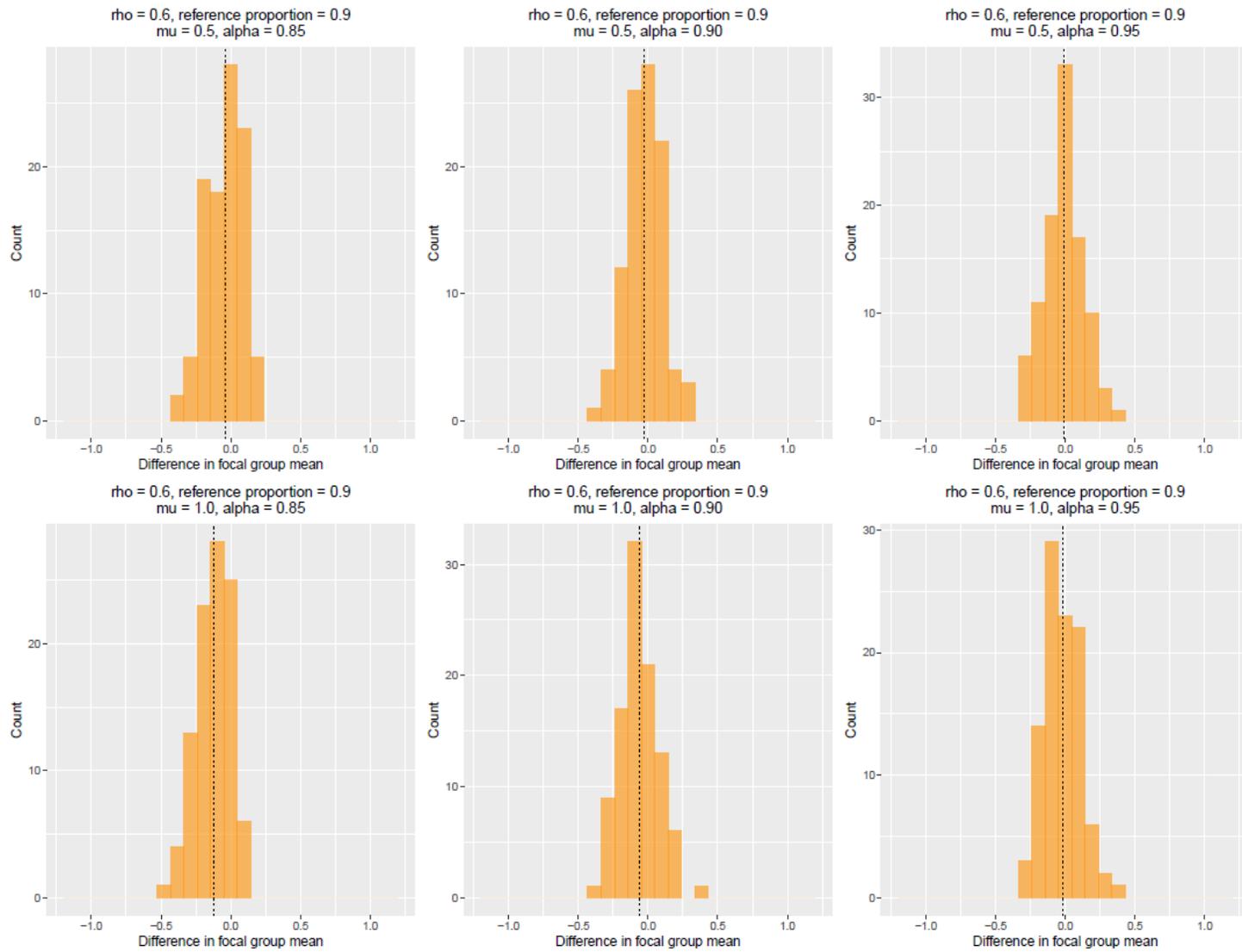
Focal mean recovery bias



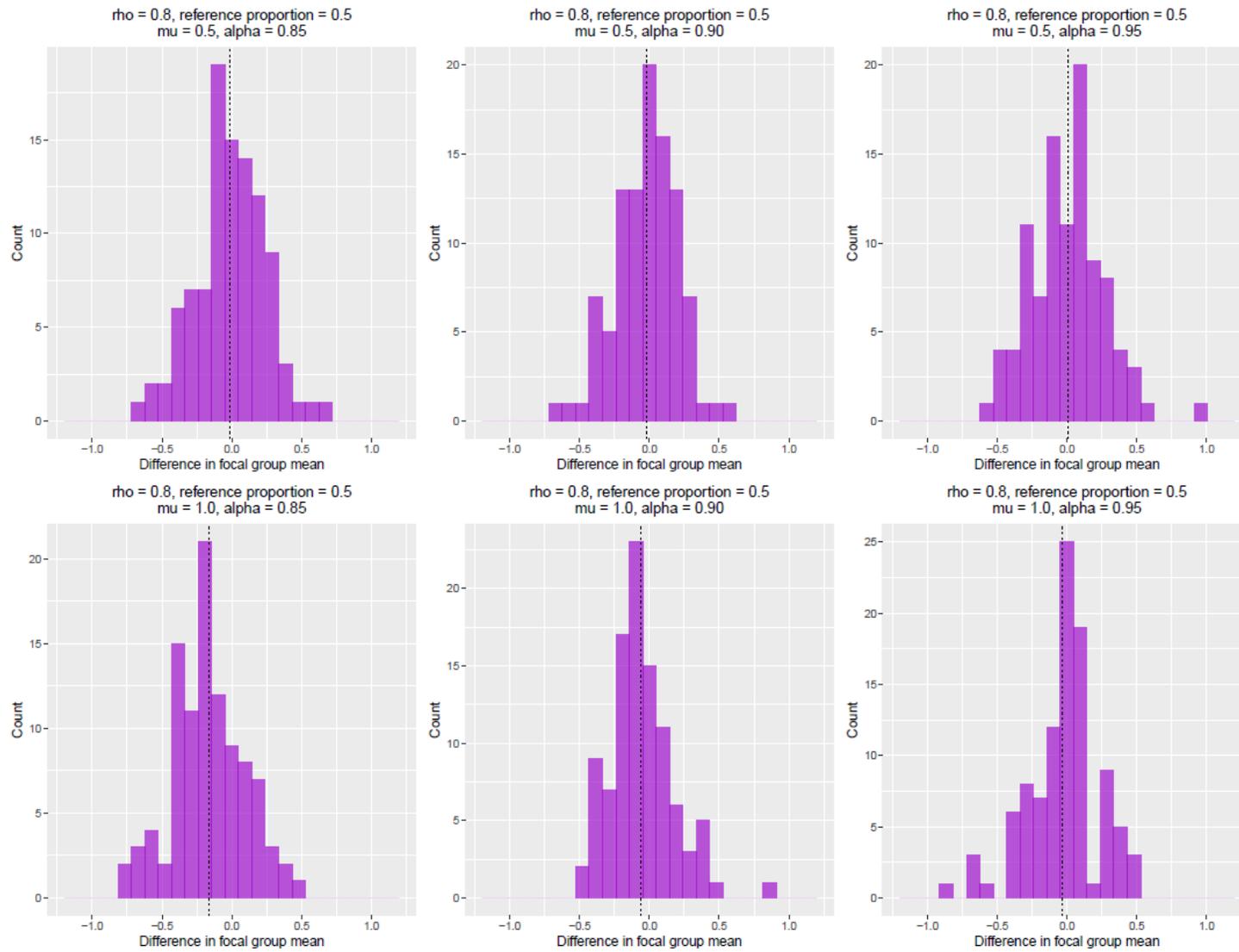
Focal mean recovery bias



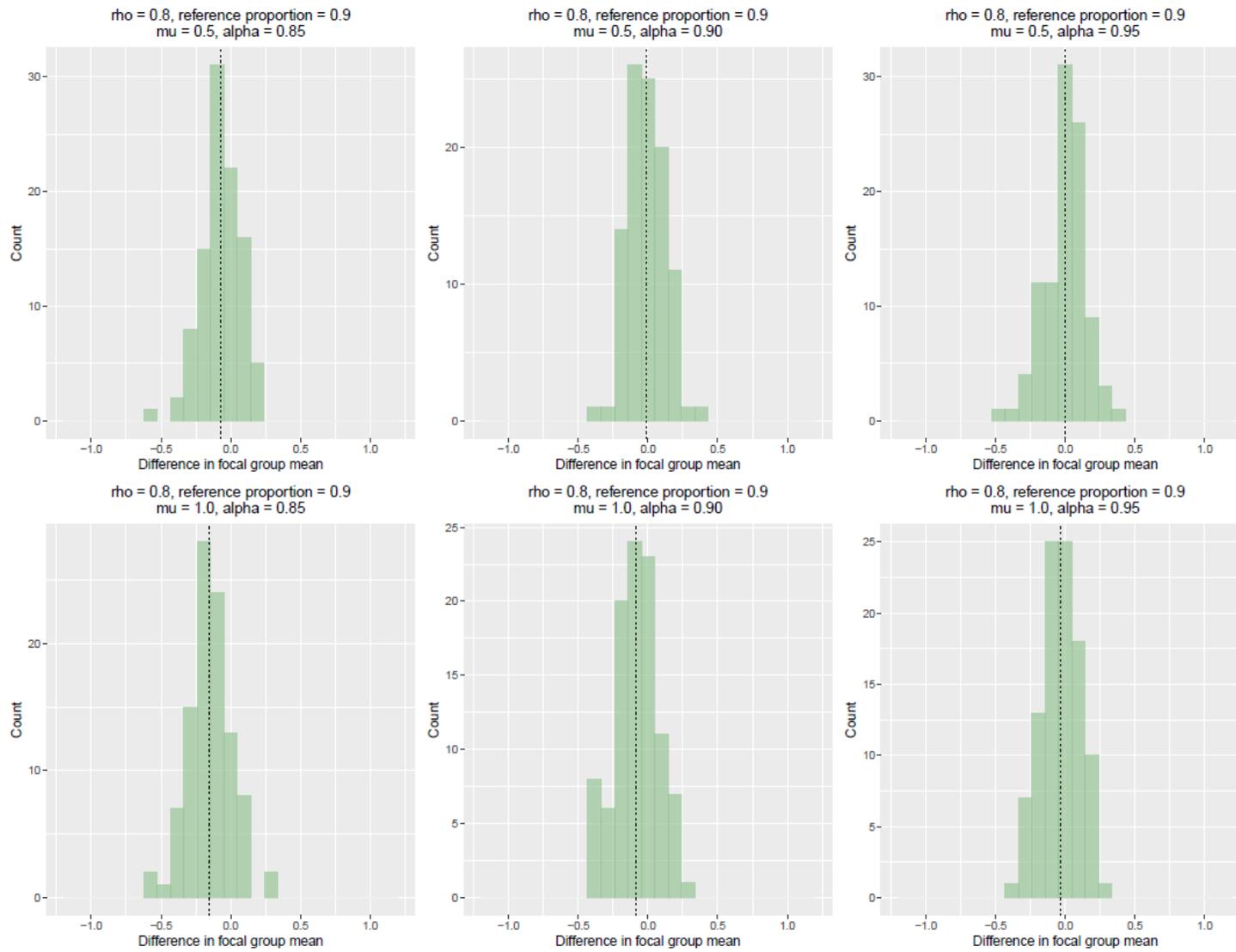
Focal mean recovery bias



Focal mean recovery bias



Focal mean recovery bias



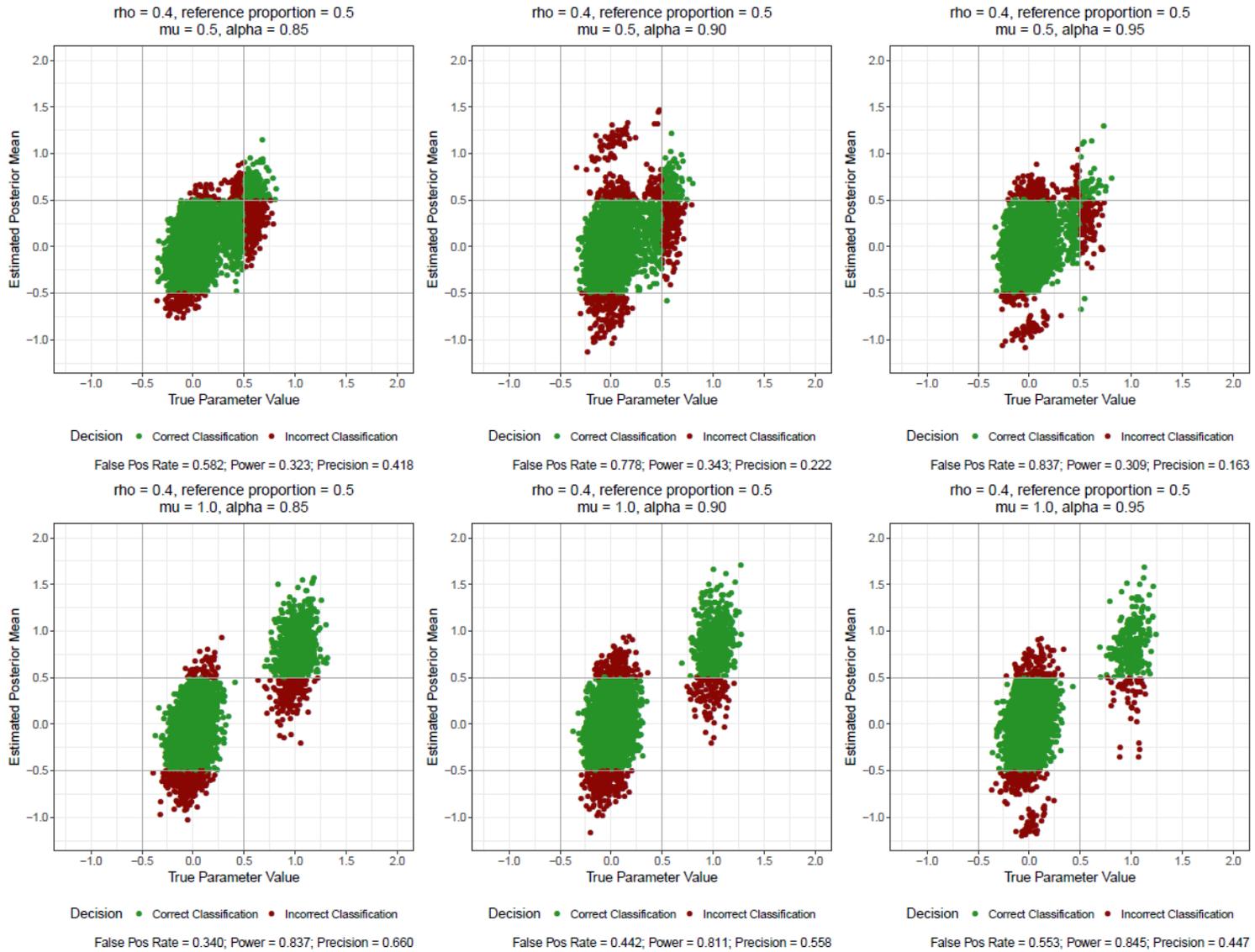
Appendix I. Decision consistency scatterplots for the 0.5 DIF flagging threshold

Note: The vertical lines show the flag threshold values for the simulated D -parameters, while the horizontal lines show the flag threshold values for the estimated D -parameters.

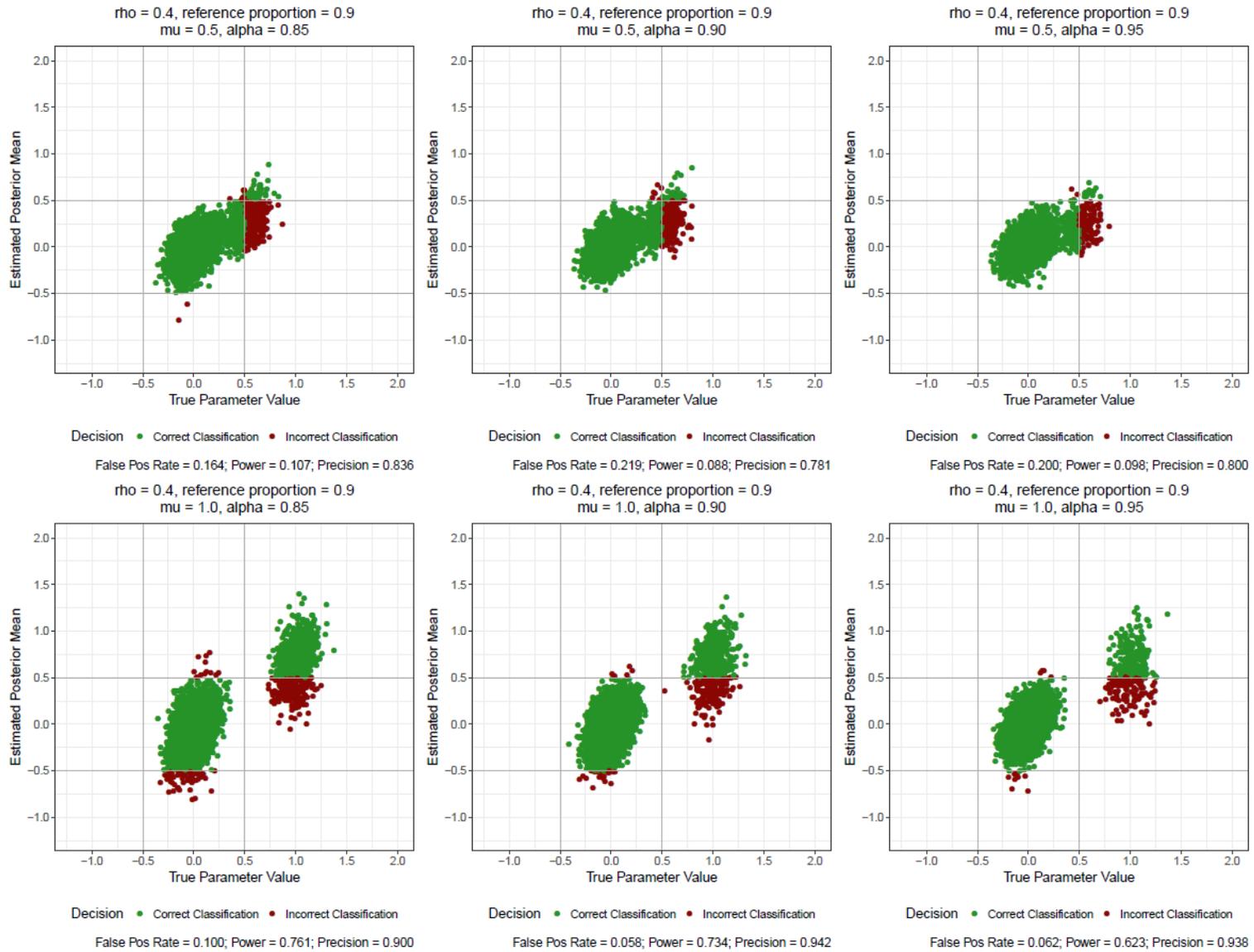
D -parameter values that were correctly classified, i.e., the simulated and estimated values were both above the flag threshold (true positive) or below the flag threshold (true negative), are colored green.

D -parameter values that were incorrectly classified, i.e., the simulated value was above the flag threshold but the estimated value was below the flag threshold (false negative) or the simulated value was below the flag threshold but the estimated value was above the flag threshold (false positive), are colored red.

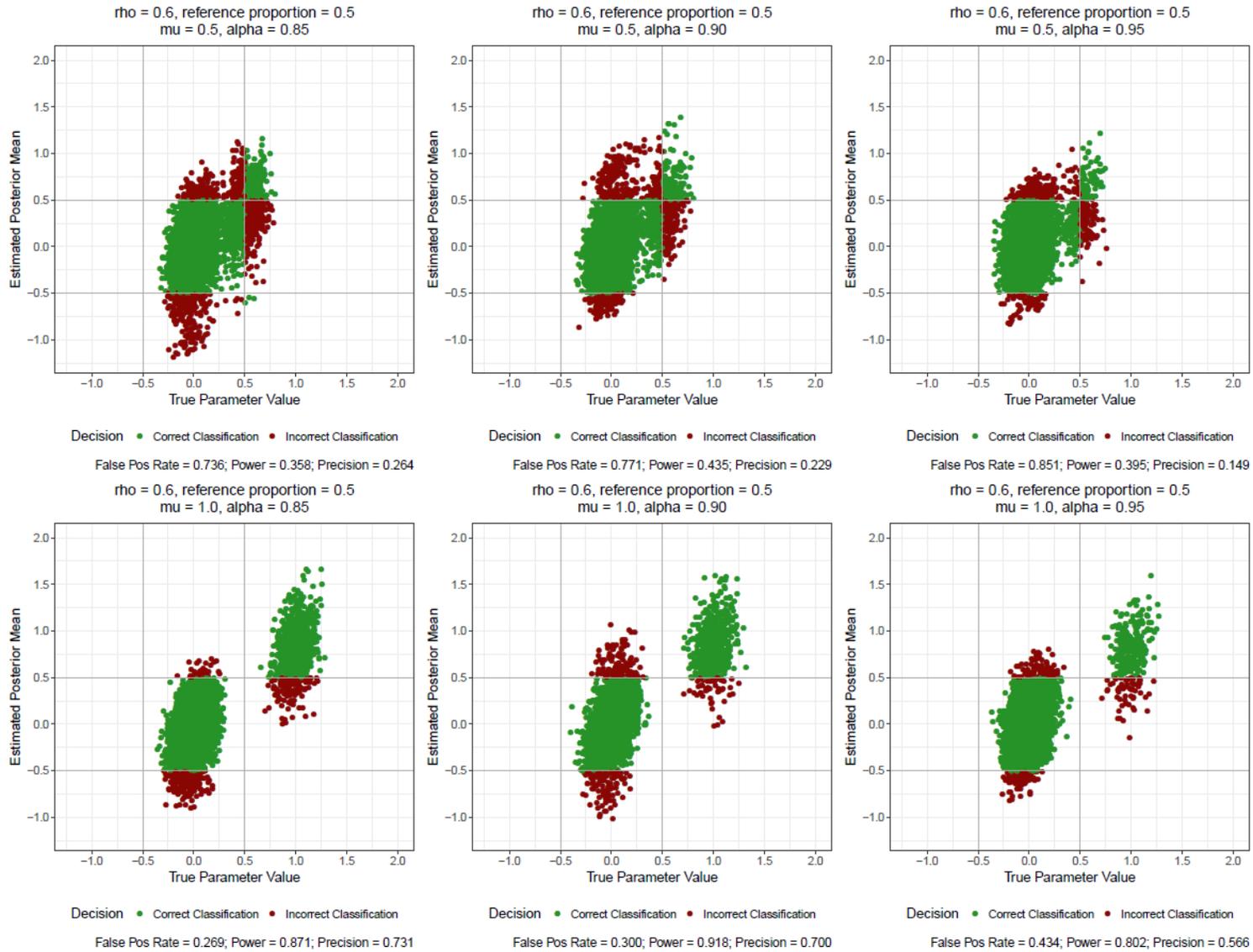
Decision Consistency for flag threshold 0.5



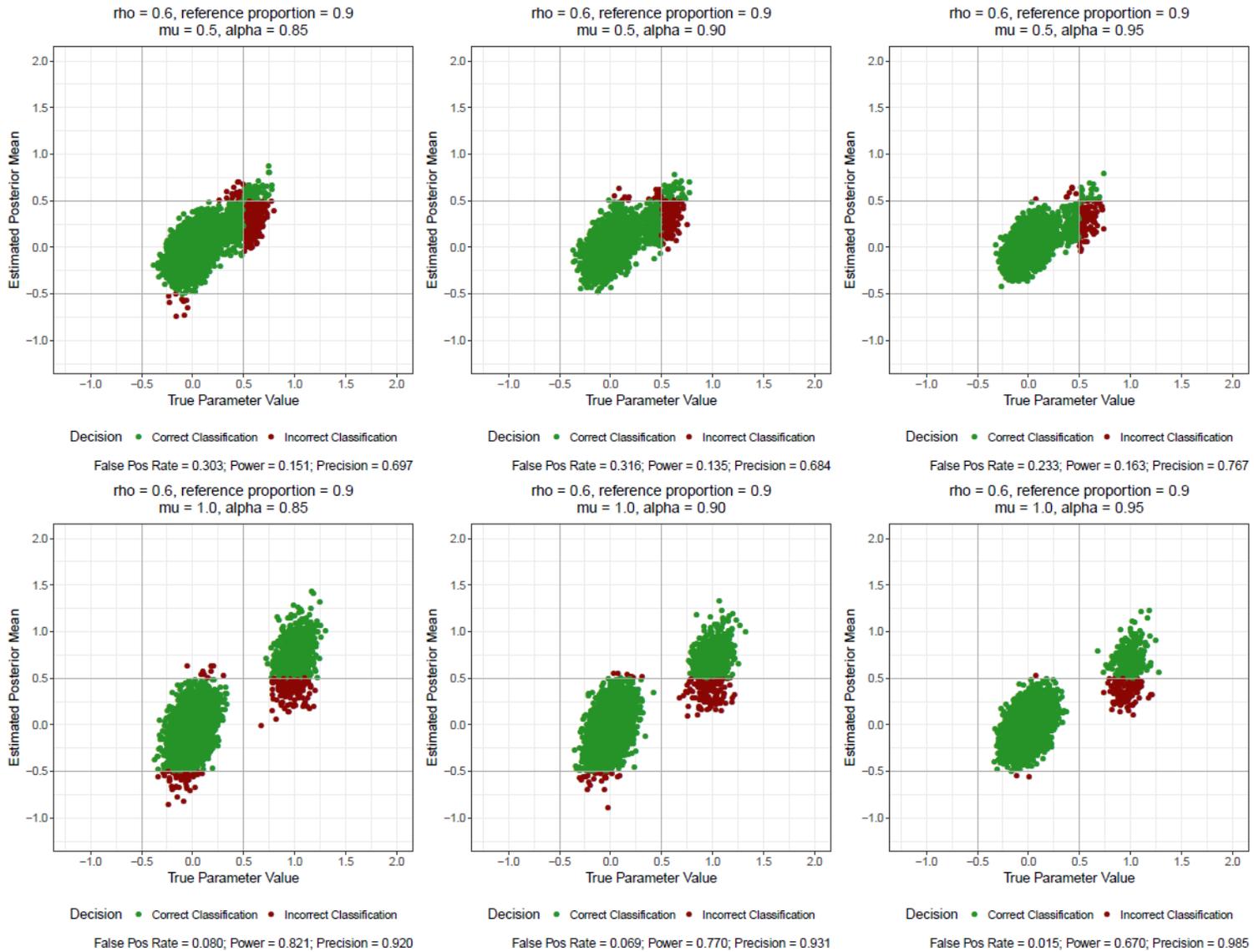
Decision Consistency for flag threshold 0.5



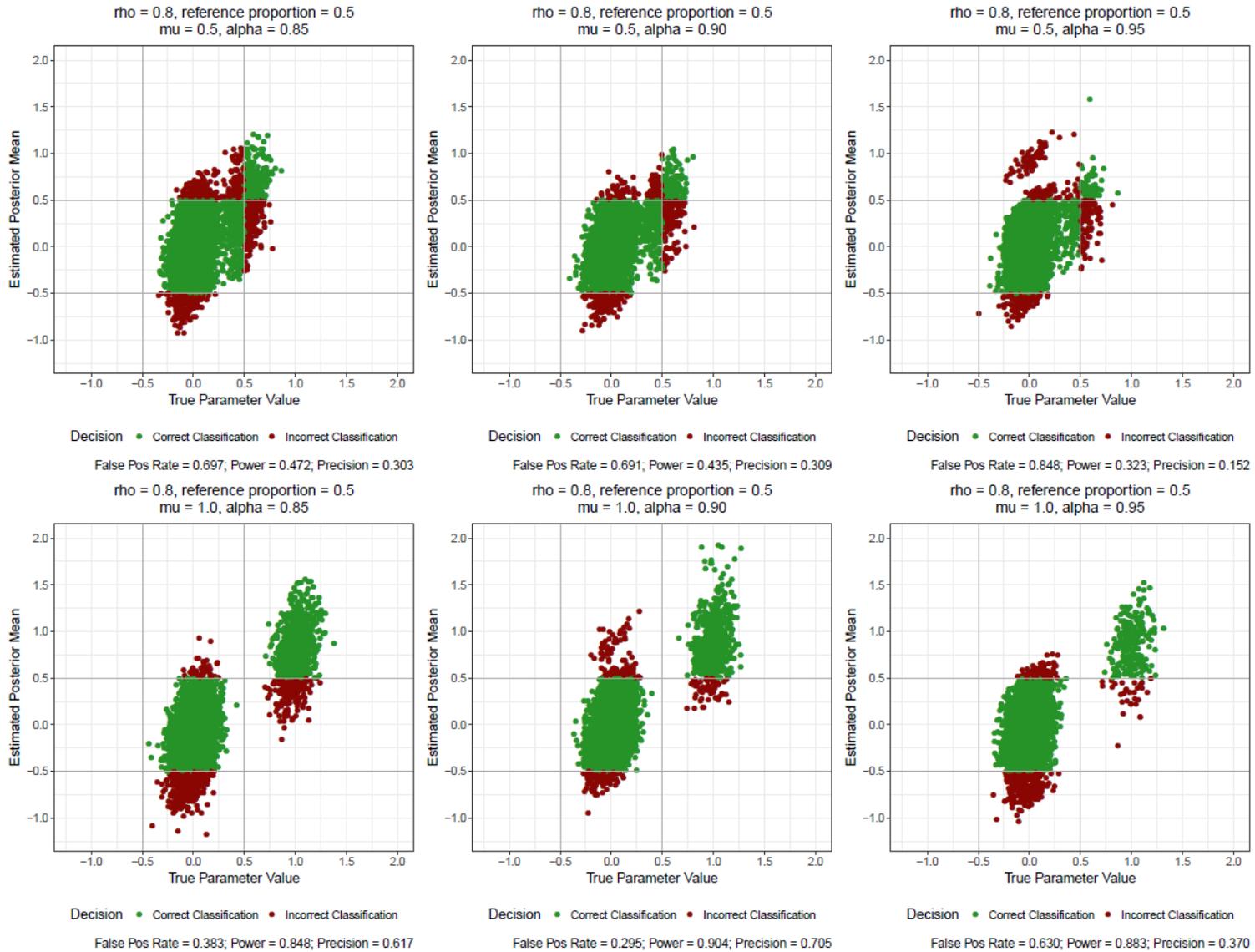
Decision Consistency for flag threshold 0.5



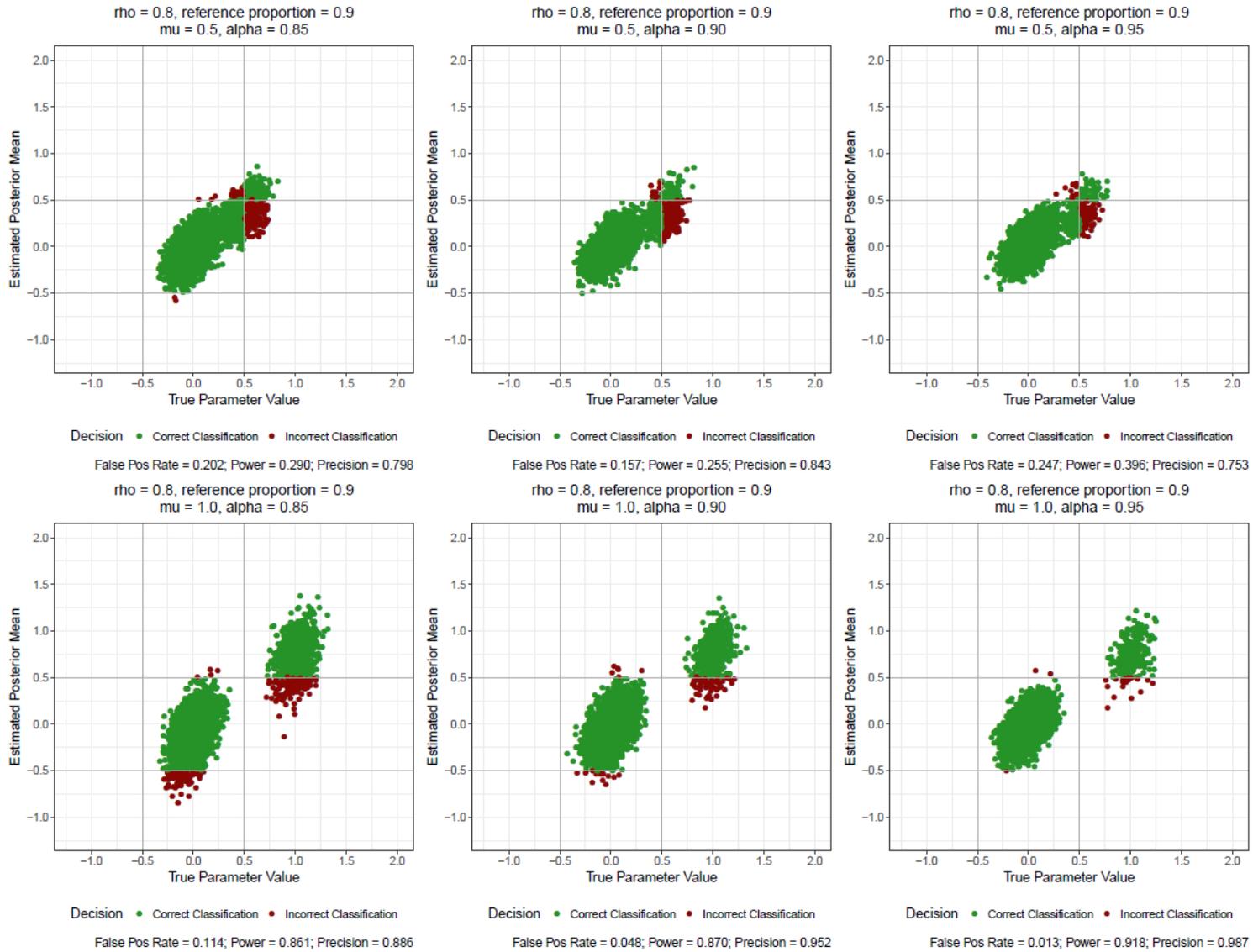
Decision Consistency for flag threshold 0.5



Decision Consistency for flag threshold 0.5



Decision Consistency for flag threshold 0.5



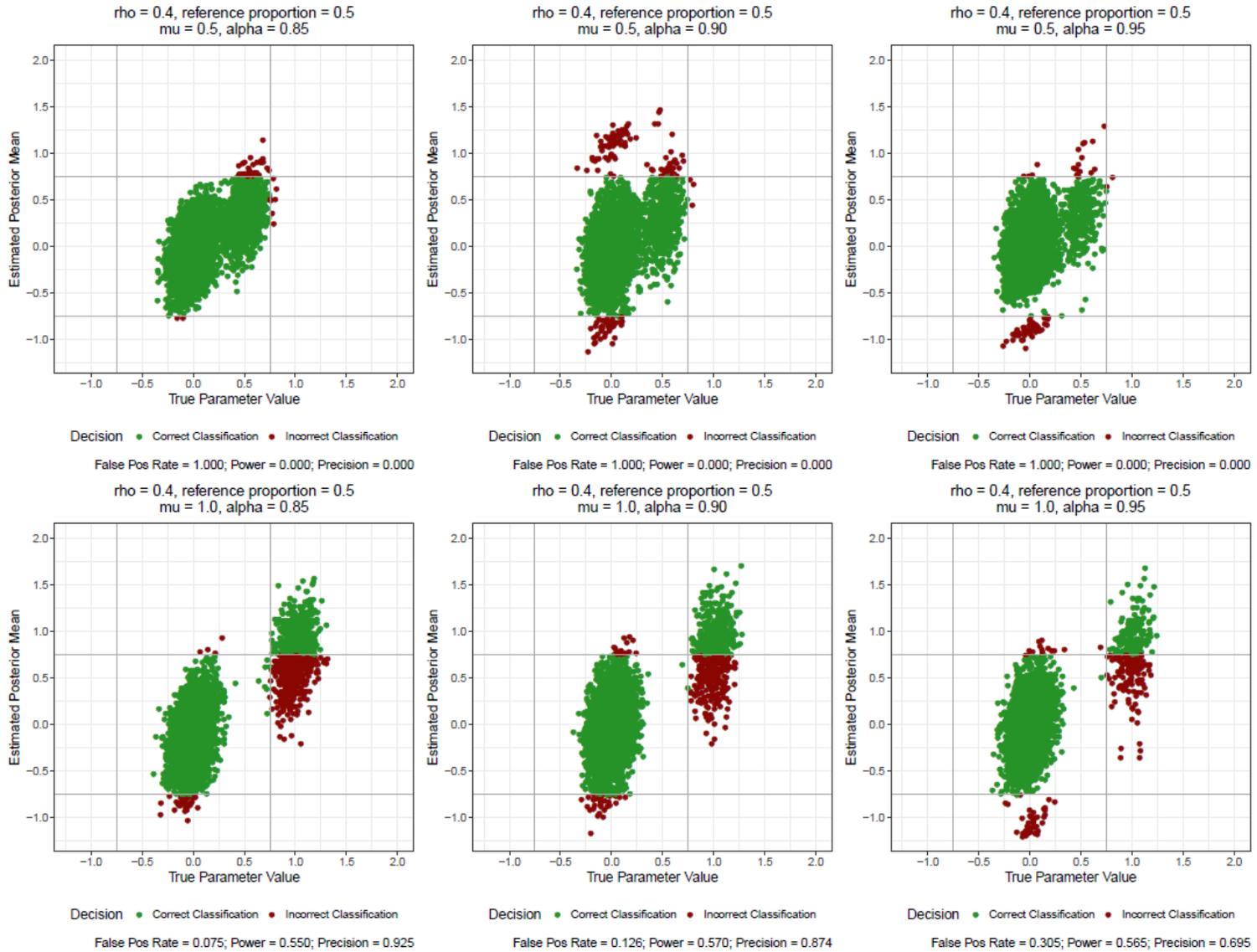
Appendix J. Decision consistency scatterplots for the 0.75 DIF flagging threshold

Note: The vertical lines show the flag threshold values for the simulated D -parameters, while the horizontal lines show the flag threshold values for the estimated D -parameters.

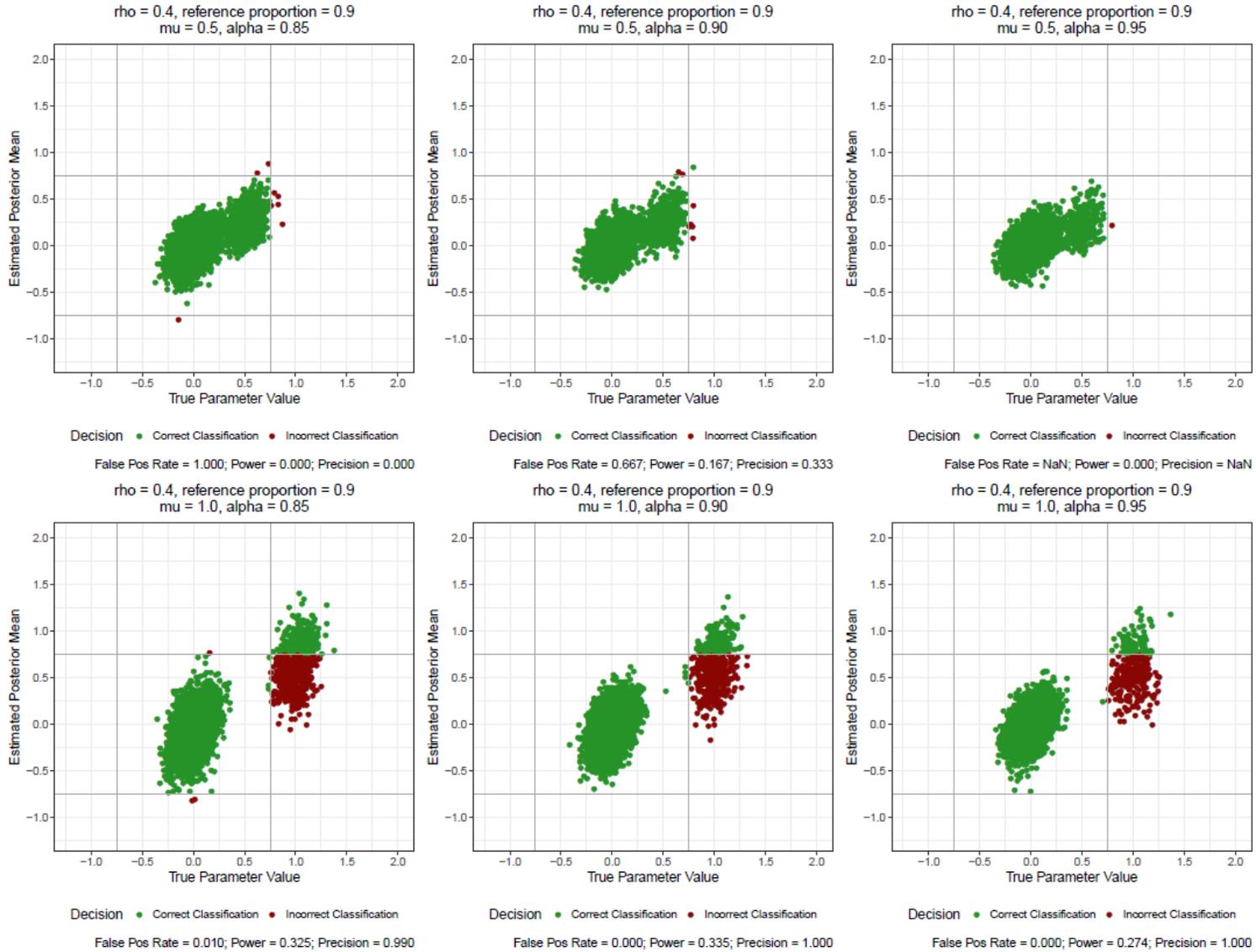
D -parameter values that were correctly classified, i.e., the simulated and estimated values were both above the flag threshold (true positive) or below the flag threshold (true negative), are colored green.

D -parameter values that were incorrectly classified, i.e., the simulated value was above the flag threshold but the estimated value was below the flag threshold (false negative) or the simulated value was below the flag threshold but the estimated value was above the flag threshold (false positive), are colored red.

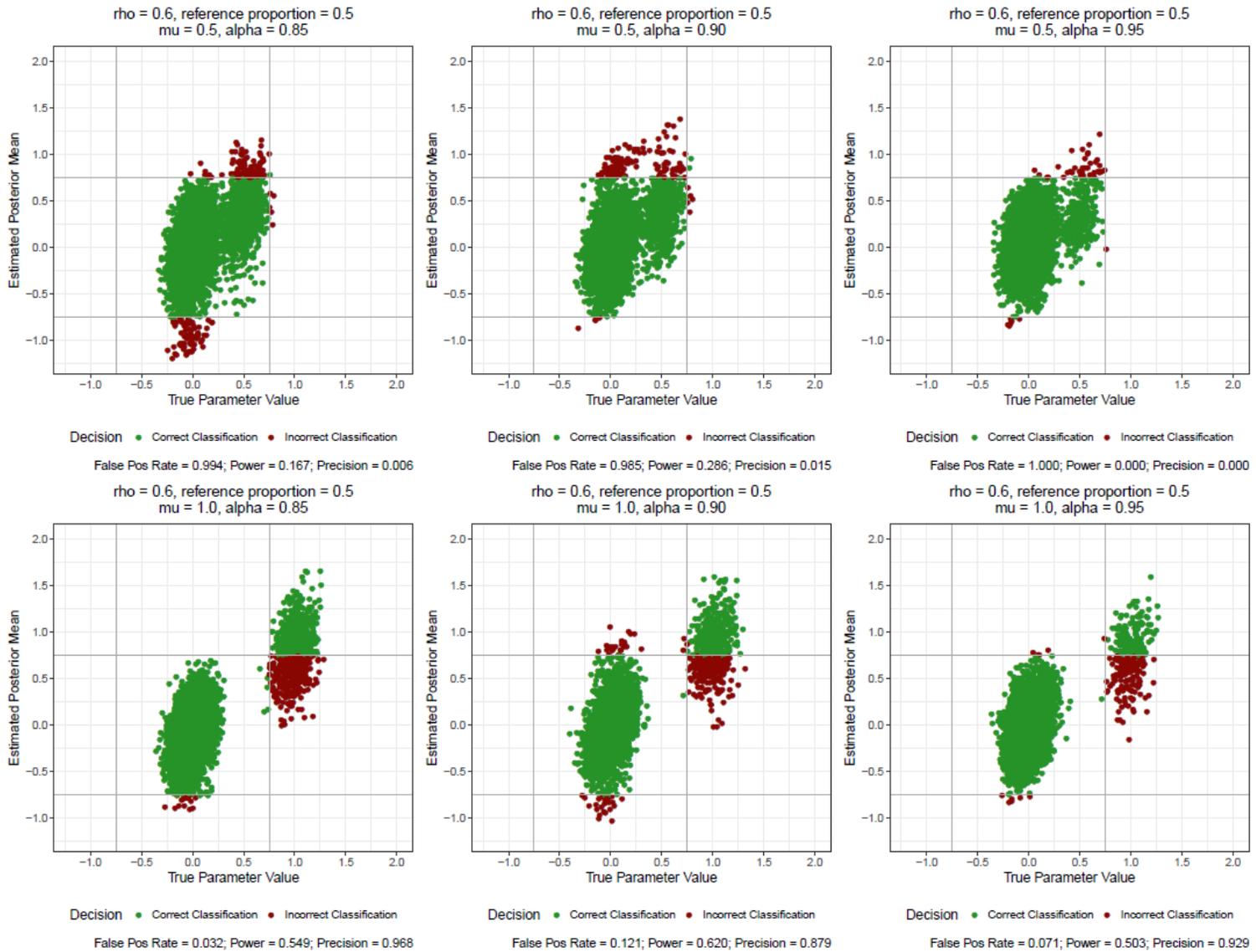
Decision Consistency for flag threshold 0.75



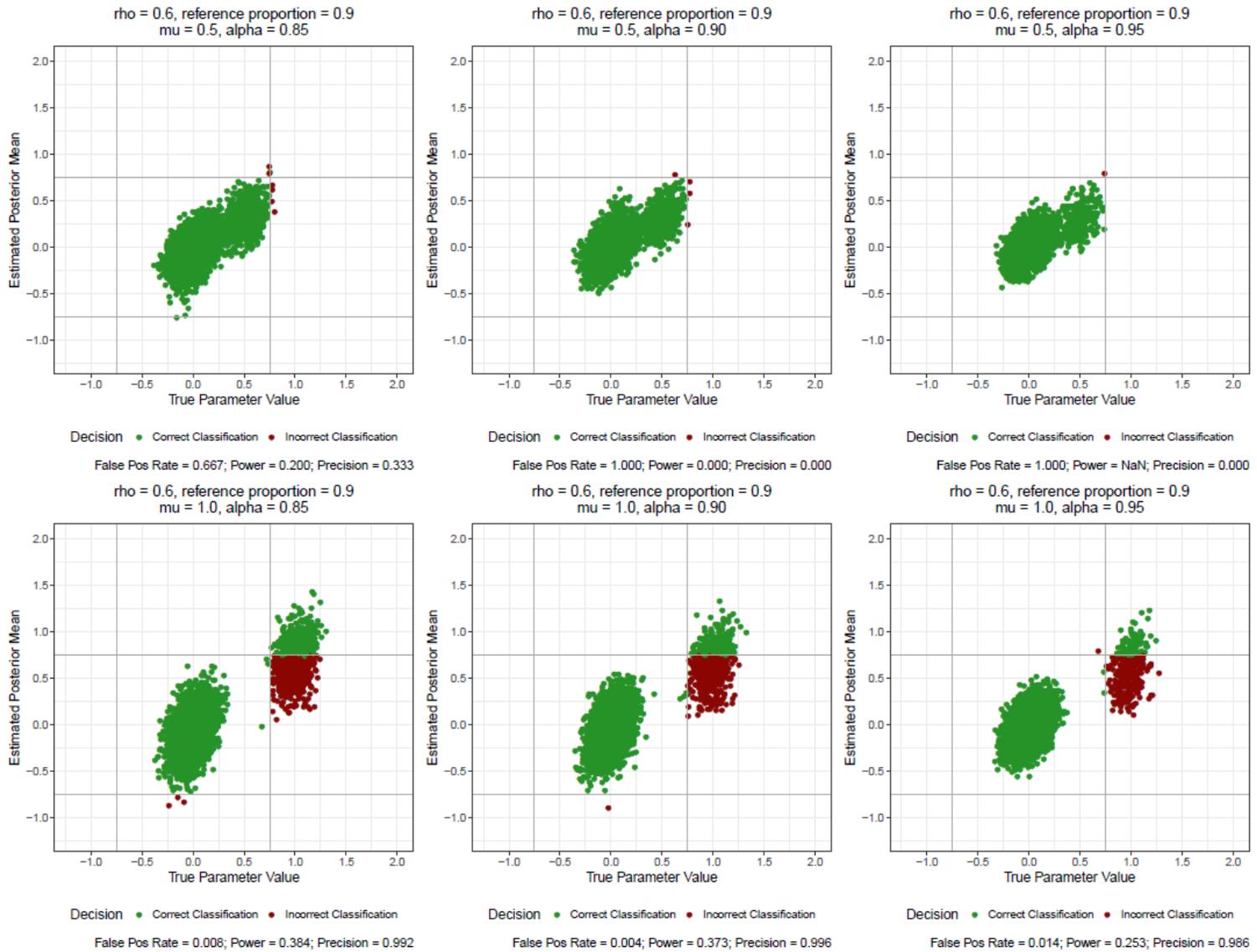
Decision Consistency for flag threshold 0.75



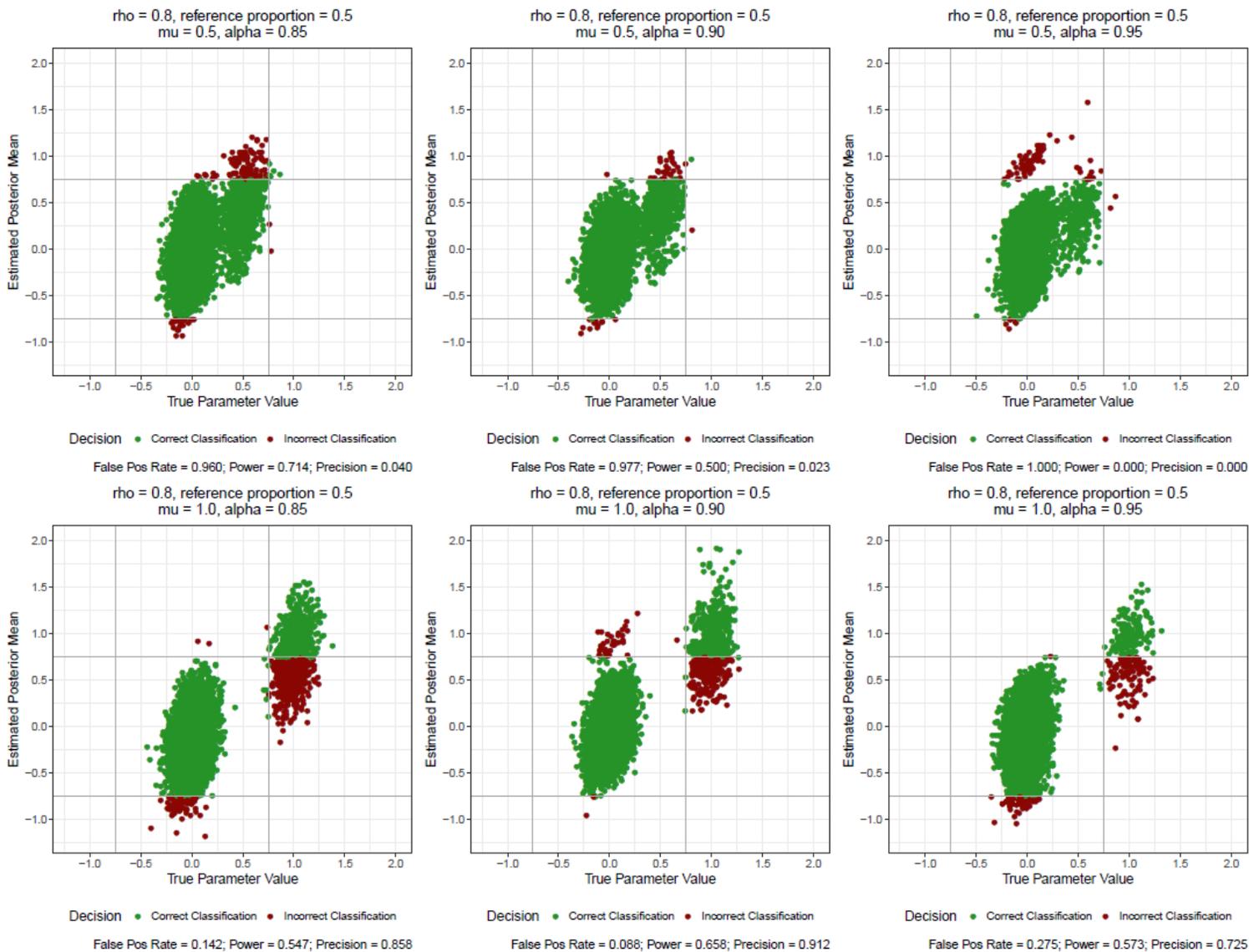
Decision Consistency for flag threshold 0.75



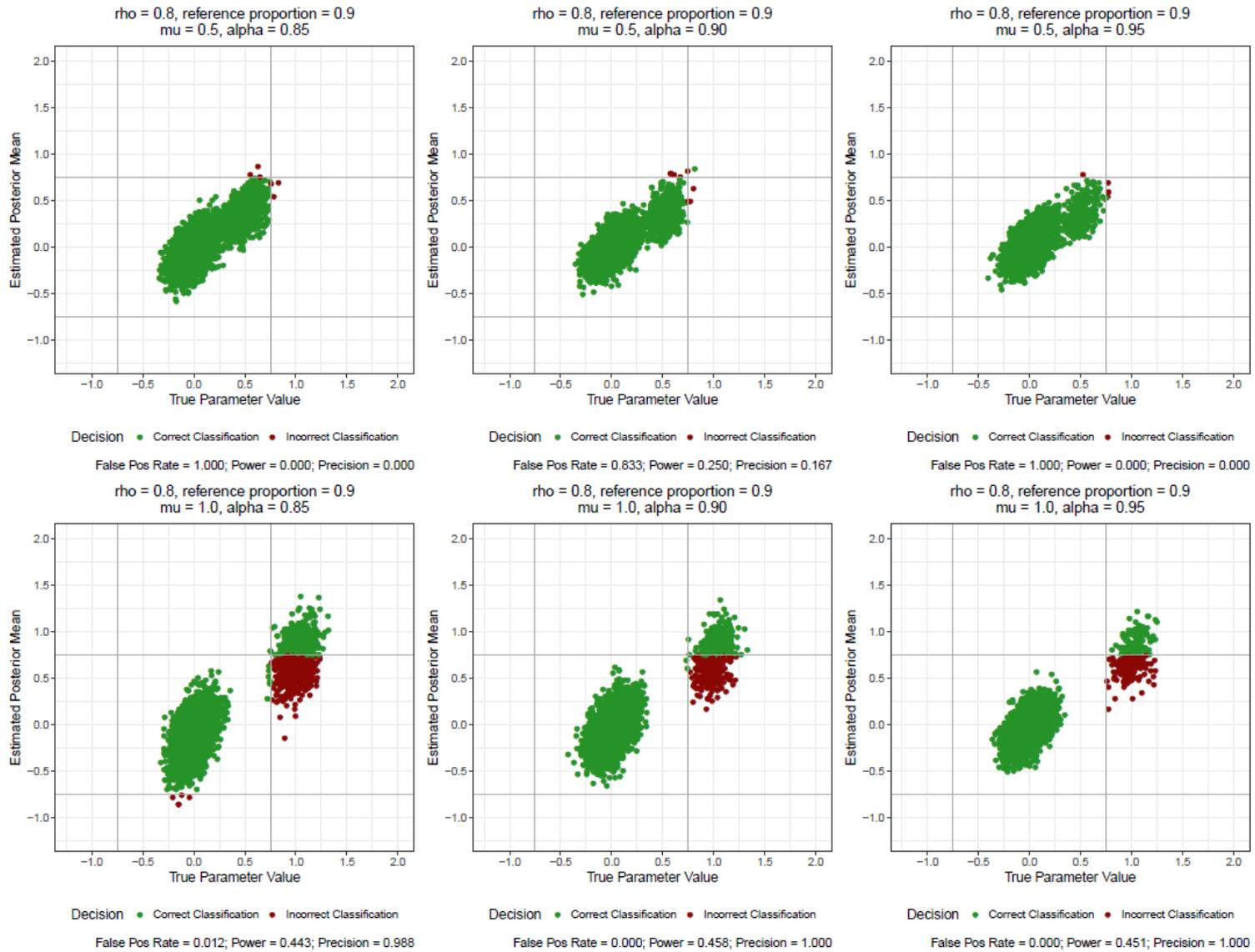
Decision Consistency for flag threshold 0.75



Decision Consistency for flag threshold 0.75



Decision Consistency for flag threshold 0.75



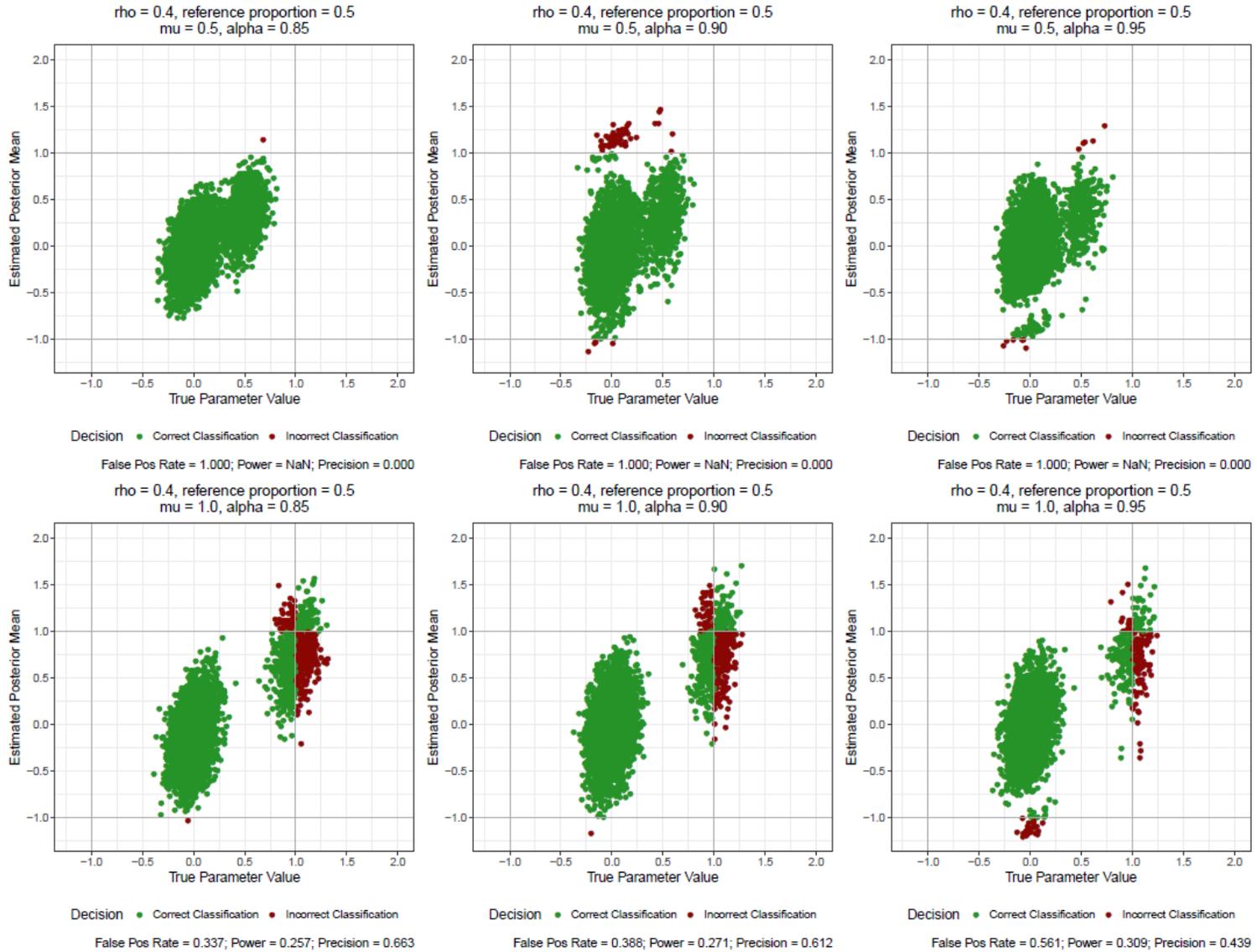
Appendix K. Decision consistency scatterplots for the 1.0 DIF flagging threshold

Note: The vertical lines show the flag threshold values for the simulated D -parameters, while the horizontal lines show the flag threshold values for the estimated D -parameters.

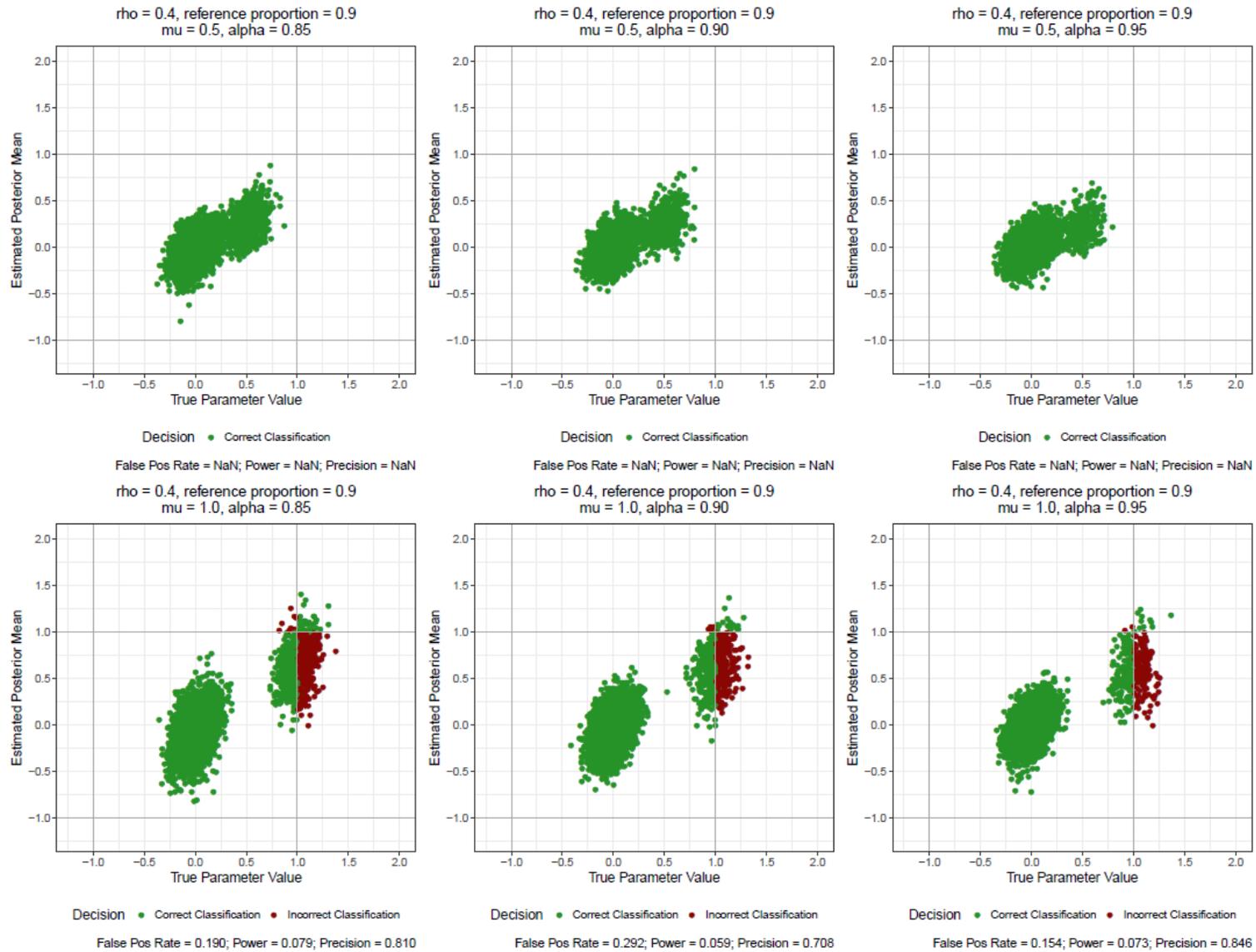
D -parameter values that were correctly classified, i.e., the simulated and estimated values were both above the flag threshold (true positive) or below the flag threshold (true negative), are colored green.

D -parameter values that were incorrectly classified, i.e., the simulated value was above the flag threshold but the estimated value was below the flag threshold (false negative) or the simulated value was below the flag threshold but the estimated value was above the flag threshold (false positive), are colored red.

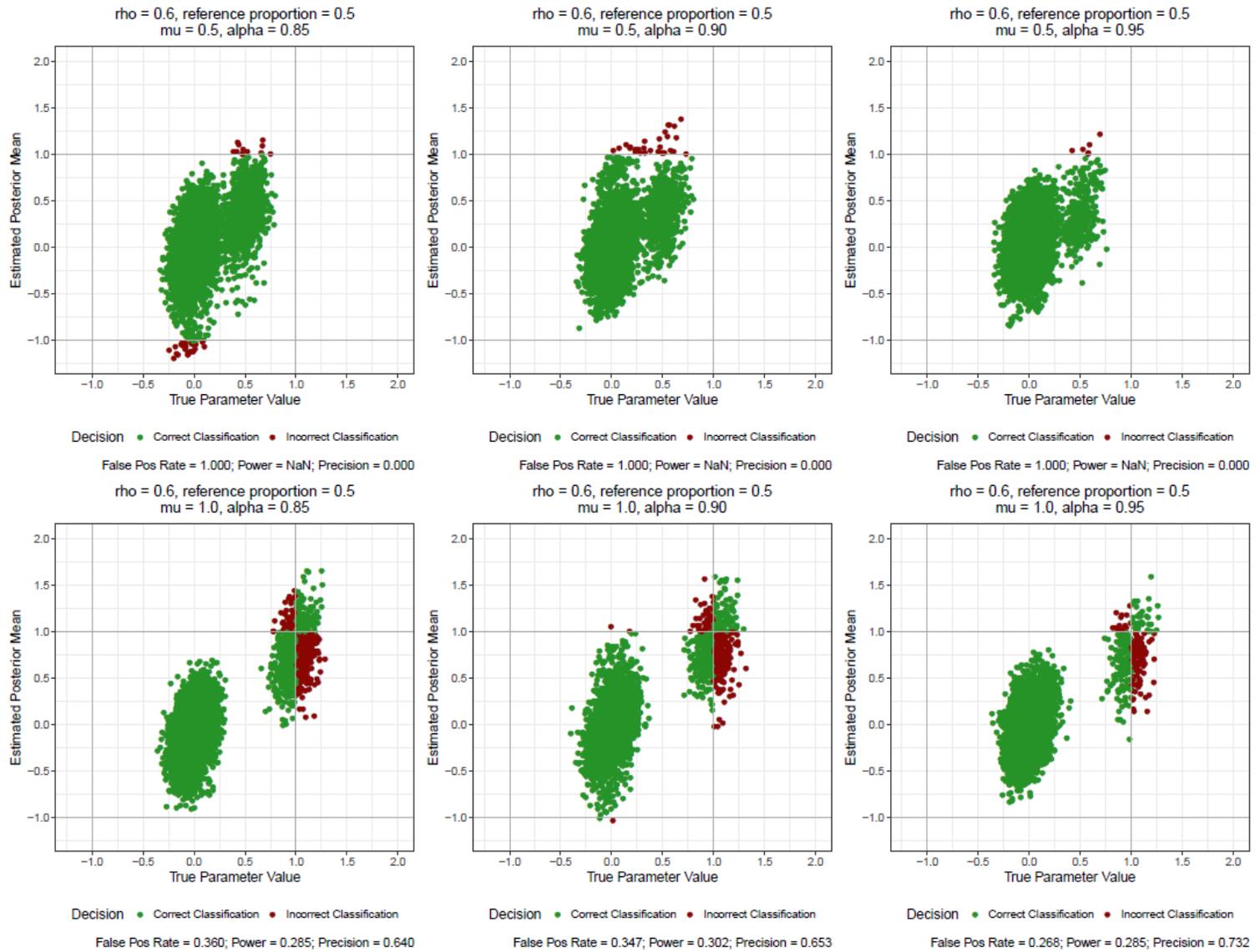
Decision Consistency for flag threshold 1



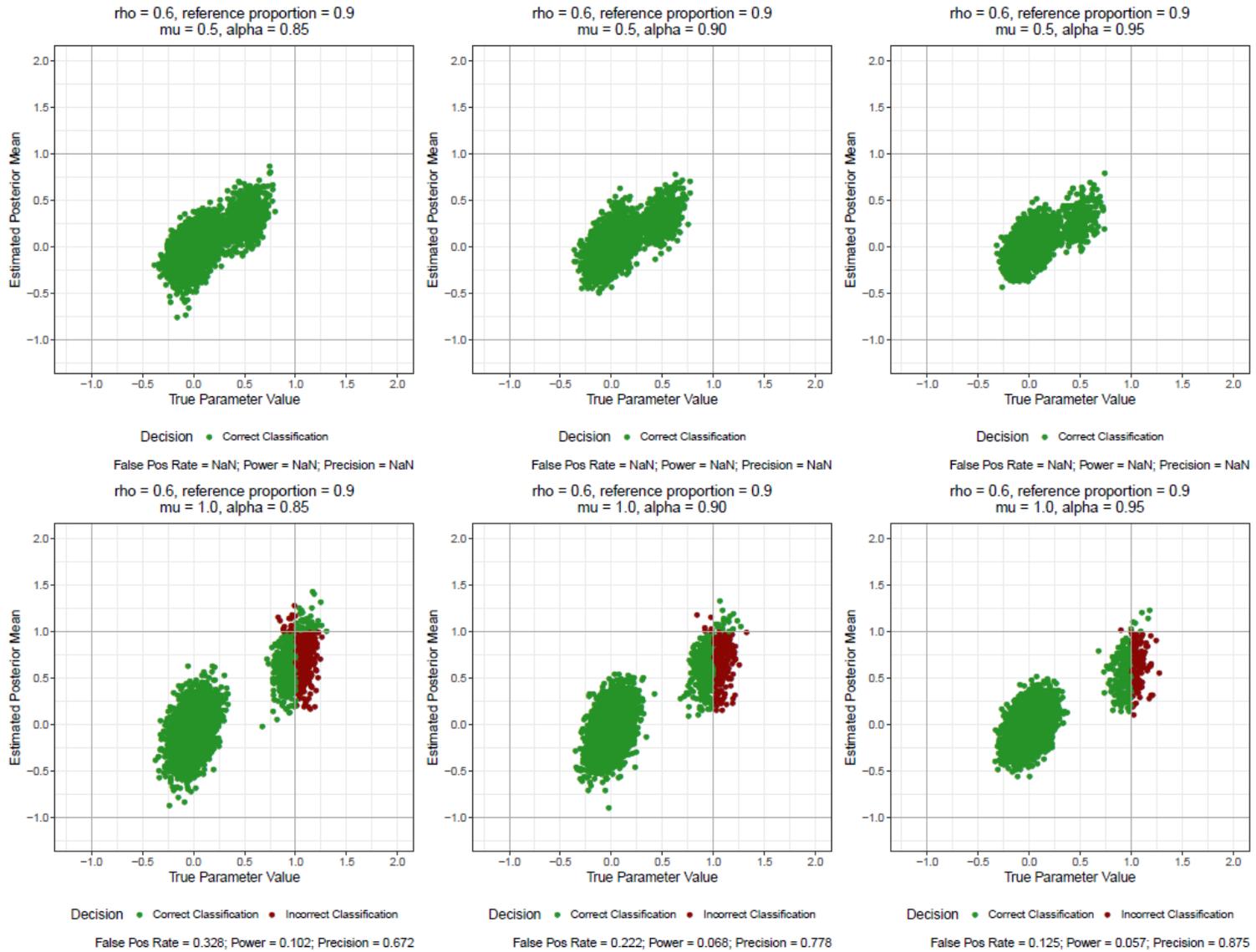
Decision Consistency for flag threshold 1



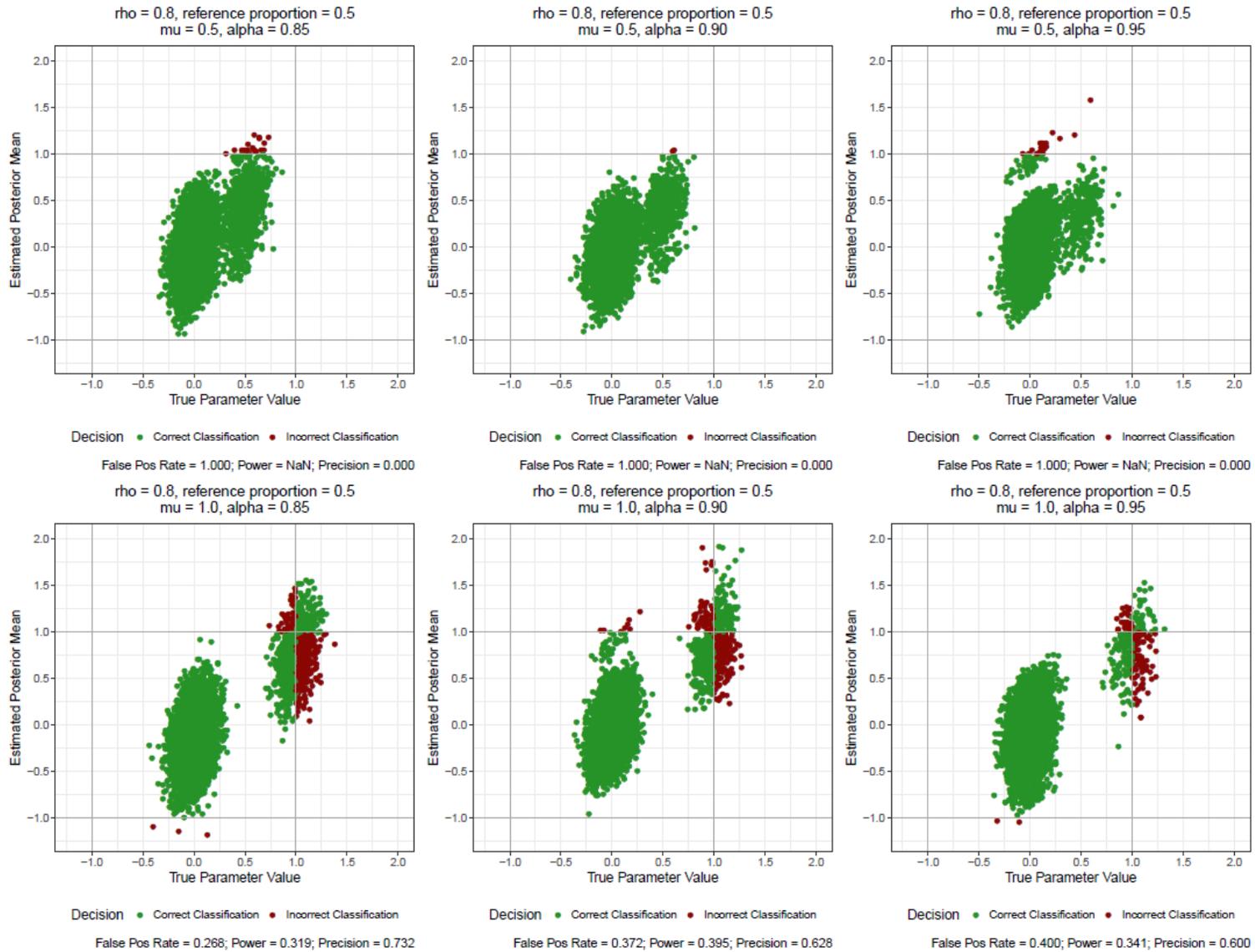
Decision Consistency for flag threshold 1



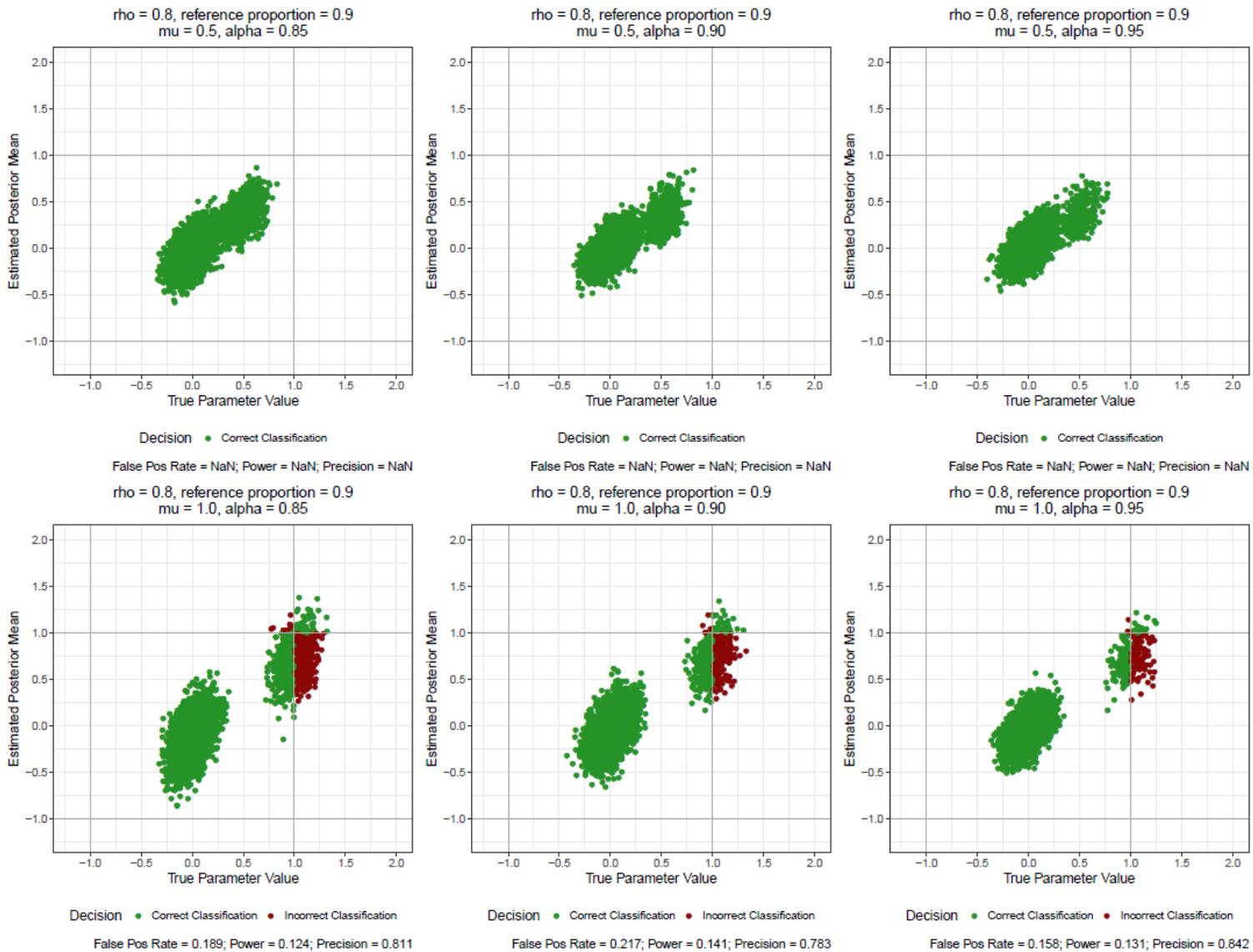
Decision Consistency for flag threshold 1



Decision Consistency for flag threshold 1



Decision Consistency for flag threshold 1



Appendix L. Decision consistency tables for each DIF flagging threshold

Table L1. False positive rate, power, and precision for the 0.5 DIF flagging threshold.

ρ	Proportion reference group	μ_2	α	False positive rate	Power	Precision
0.4	0.5	0.5	0.85	0.582	0.323	0.418
0.4	0.5	0.5	0.90	0.778	0.343	0.222
0.4	0.5	0.5	0.95	0.837	0.309	0.163
0.4	0.5	1.0	0.85	0.340	0.837	0.660
0.4	0.5	1.0	0.90	0.442	0.811	0.558
0.4	0.5	1.0	0.95	0.553	0.845	0.447
0.4	0.9	0.5	0.85	0.164	0.107	0.836
0.4	0.9	0.5	0.90	0.219	0.088	0.781
0.4	0.9	0.5	0.95	0.200	0.098	0.800
0.4	0.9	1.0	0.85	0.100	0.761	0.900
0.4	0.9	1.0	0.90	0.058	0.734	0.942
0.4	0.9	1.0	0.95	0.062	0.623	0.938
0.6	0.5	0.5	0.85	0.736	0.358	0.264
0.6	0.5	0.5	0.90	0.771	0.435	0.229
0.6	0.5	0.5	0.95	0.851	0.395	0.149
0.6	0.5	1.0	0.85	0.269	0.871	0.731
0.6	0.5	1.0	0.90	0.300	0.918	0.700
0.6	0.5	1.0	0.95	0.434	0.802	0.566
0.6	0.9	0.5	0.85	0.303	0.151	0.697
0.6	0.9	0.5	0.90	0.316	0.135	0.684
0.6	0.9	0.5	0.95	0.233	0.163	0.767
0.6	0.9	1.0	0.85	0.080	0.821	0.920
0.6	0.9	1.0	0.90	0.069	0.770	0.931
0.6	0.9	1.0	0.95	0.015	0.670	0.985
0.8	0.5	0.5	0.85	0.697	0.472	0.303
0.8	0.5	0.5	0.90	0.691	0.435	0.309

0.8	0.5	0.5	0.95	0.848	0.323	0.152
0.8	0.5	1.0	0.85	0.383	0.848	0.617
0.8	0.5	1.0	0.90	0.295	0.904	0.705
0.8	0.5	1.0	0.95	0.630	0.883	0.370
0.8	0.9	0.5	0.85	0.202	0.290	0.798
0.8	0.9	0.5	0.90	0.157	0.255	0.843
0.8	0.9	0.5	0.95	0.247	0.396	0.753
0.8	0.9	1.0	0.85	0.114	0.861	0.886
0.8	0.9	1.0	0.90	0.048	0.870	0.952
0.8	0.9	1.0	0.95	0.013	0.918	0.987

Table L2. False positive rate, power, and precision for the 0.75 DIF flagging threshold.

ρ	Proportion reference group	μ_2	α	False positive rate	Power	Precision
0.4	0.5	0.5	0.85	1.000	0.000	0.000
0.4	0.5	0.5	0.90	1.000	0.000	0.000
0.4	0.5	0.5	0.95	1.000	0.000	0.000
0.4	0.5	1.0	0.85	0.075	0.550	0.925
0.4	0.5	1.0	0.90	0.126	0.570	0.874
0.4	0.5	1.0	0.95	0.305	0.565	0.695
0.4	0.9	0.5	0.85	1.000	0.000	0.000
0.4	0.9	0.5	0.90	0.667	0.167	0.333
0.4	0.9	0.5	0.95	NaN	0.000	NaN
0.4	0.9	1.0	0.85	0.010	0.325	0.990
0.4	0.9	1.0	0.90	0.000	0.335	1.000
0.4	0.9	1.0	0.95	0.000	0.274	1.000
0.6	0.5	0.5	0.85	0.994	0.167	0.006
0.6	0.5	0.5	0.90	0.985	0.286	0.015
0.6	0.5	0.5	0.95	1.000	0.000	0.000

0.6	0.5	1.0	0.85	0.032	0.549	0.968
0.6	0.5	1.0	0.90	0.121	0.620	0.879
0.6	0.5	1.0	0.95	0.071	0.503	0.929
0.6	0.9	0.5	0.85	0.667	0.200	0.333
0.6	0.9	0.5	0.90	1.000	0.000	0.000
0.6	0.9	0.5	0.95	1.000	NaN	0.000
0.6	0.9	1.0	0.85	0.008	0.384	0.992
0.6	0.9	1.0	0.90	0.004	0.373	0.996
0.6	0.9	1.0	0.95	0.014	0.253	0.986
0.8	0.5	0.5	0.85	0.960	0.714	0.040
0.8	0.5	0.5	0.90	0.977	0.500	0.023
0.8	0.5	0.5	0.95	1.000	0.000	0.000
0.8	0.5	1.0	0.85	0.142	0.547	0.858
0.8	0.5	1.0	0.90	0.088	0.658	0.912
0.8	0.5	1.0	0.95	0.275	0.573	0.725
0.8	0.9	0.5	0.85	1.000	0.000	0.000
0.8	0.9	0.5	0.90	0.833	0.250	0.167
0.8	0.9	0.5	0.95	1.000	0.000	0.000
0.8	0.9	1.0	0.85	0.012	0.443	0.988
0.8	0.9	1.0	0.90	0.000	0.458	1.000
0.8	0.9	1.0	0.95	0.000	0.451	1.000

Table L3. False positive rate, power, and precision for the 1.0 DIF flagging threshold.

ρ	Proportion reference group	μ_2	α	False positive rate	Power	Precision
0.4	0.5	0.5	0.85	1.000	NaN	0.000
0.4	0.5	0.5	0.90	1.000	NaN	0.000
0.4	0.5	0.5	0.95	1.000	NaN	0.000
0.4	0.5	1.0	0.85	0.337	0.257	0.663

0.4	0.5	1.0	0.90	0.388	0.271	0.612
0.4	0.5	1.0	0.95	0.561	0.309	0.439
0.4	0.9	0.5	0.85	NaN	NaN	NaN
0.4	0.9	0.5	0.90	NaN	NaN	NaN
0.4	0.9	0.5	0.95	NaN	NaN	NaN
0.4	0.9	1.0	0.85	0.190	0.079	0.810
0.4	0.9	1.0	0.90	0.292	0.059	0.708
0.4	0.9	1.0	0.95	0.154	0.073	0.846
0.6	0.5	0.5	0.85	1.000	NaN	0.000
0.6	0.5	0.5	0.90	1.000	NaN	0.000
0.6	0.5	0.5	0.95	1.000	NaN	0.000
0.6	0.5	1.0	0.85	0.360	0.285	0.640
0.6	0.5	1.0	0.90	0.347	0.302	0.653
0.6	0.5	1.0	0.95	0.268	0.285	0.732
0.6	0.9	0.5	0.85	NaN	NaN	NaN
0.6	0.9	0.5	0.90	NaN	NaN	NaN
0.6	0.9	0.5	0.95	NaN	NaN	NaN
0.6	0.9	1.0	0.85	0.328	0.102	0.672
0.6	0.9	1.0	0.90	0.222	0.068	0.778
0.6	0.9	1.0	0.95	0.125	0.057	0.875
0.8	0.5	0.5	0.85	1.000	NaN	0.000
0.8	0.5	0.5	0.90	1.000	NaN	0.000
0.8	0.5	0.5	0.95	1.000	NaN	0.000
0.8	0.5	1.0	0.85	0.268	0.319	0.732
0.8	0.5	1.0	0.90	0.372	0.395	0.628
0.8	0.5	1.0	0.95	0.400	0.341	0.600
0.8	0.9	0.5	0.85	NaN	NaN	NaN
0.8	0.9	0.5	0.90	NaN	NaN	NaN
0.8	0.9	0.5	0.95	NaN	NaN	NaN
0.8	0.9	1.0	0.85	0.189	0.124	0.811
0.8	0.9	1.0	0.90	0.217	0.141	0.783
0.8	0.9	1.0	0.95	0.158	0.131	0.842

Appendix M. R and Stan code for data generation and model estimation

Code from three files is given in this appendix:

- *cluster_base_script.R* sources the other two files, then generates data, estimates the model, and writes relevant output to disk.
- *stan_scripts.R* contains the Stan syntax used to run the model.
- *mixture_functions.R* contains the functions used to simulate data from a mixture distribution as well as various other helper functions used during data simulation and processing.

The simulations were run in a cluster computing environment. The arguments specifying the systematically varied parameters used to create conditions were specified in the command line when starting each job in the cluster computing environment. A section of commented code in *cluster_base_script.R* shows how these values could also be set directly within R.

Stan result objects were not saved to disk for these simulations due to space constraints in the cluster environment. Commented lines of code in *cluster_base_script.R* show how result objects could be stored and saved; however, these objects can become very large, so exercise caution.

cluster_base_script.R

```
##### SETUP #####
```

```
work_dir <- getwd()
```

```
source("mixture_functions.R")
date <- format.Date(Sys.Date(), "%Y%m%d")
options(scipen = 999)
```

```
needed_packages <- c("tidyr", "dplyr", "rstan", "rstudioapi", "robustbase", "portableParallelSeeds")
for(i in 1:length(needed_packages)){
  library(needed_packages[i], character.only = TRUE)
}
```

```
##### COMMAND LINE ARGUMENT SETUP #####
```

```
#comment out when not testing
# comm_args <- c("rho=0.8", "P_REF=0.5", "mu2=0.5", "alpha=0.95")
```

```

#uncomment for real run
comm_args <- commandArgs(trailingOnly = TRUE)

args <- strsplit(comm_args,"=",fixed=TRUE)

for (arg in 1:length(args)){
  argname <- args[[arg]][1]
  argval <- as.numeric(args[[arg]][2])
  assign(argname,argval)
}

#### SPECIFICATIONS ####
#number of people
n_people <- 1000
#number of items
n_items <- 60
#number of reps
nreps <- 10
#rho is the amount of DIF explained by the second-order factors - to be specified in job script
#P_REF is the proportion of people in the reference group - to be specified in job script
#alpha is mixture parameter - to be specified in job script
#mu1 is mean of distribution 1 - items with negligible DIF
mu1 <- 0
#mu2 is the mean of distribution 2 - items with "true" DIF - to be specified in job script
#sdev is the standard deviation of each distribution. sdev is equal for both distributions
sdev <- .1
#sdev_D is used to calculate beta1 and beta0
sdev_D <- sqrt(alpha*(sdev^2)) + ((1-alpha)*(sdev^2)) + (alpha*(1-alpha)*((mu1-mu2)^2))
#R2 is the amount of variance attributed to item-level features
R2_true <- (rho^2)

#### SEED SETUP ####
filename <- paste0("seeds_", gsub(".", "-", as.character(rho), fixed = TRUE), "rho_",
                  gsub(".", "-", as.character(P_REF), fixed = TRUE), "PREF_",
                  gsub(".", "-", as.character(mu2), fixed = TRUE), "mu_",
                  gsub(".", "-", as.character(alpha), fixed = TRUE), "alpha.rds")

seeds <- readRDS(filename)

#### STAN SETUP ####
#load stan model scripts
source("stan_scripts.R")

b.dat_long <- list("n_people", "n_items", "n_observations", "respondentid",

```

```

        "itemid", "response", "group", "group_long",
        "DIFpredict")

#analysis setup
precomp <- stanc(model_code = stancode_long)
precomp_model <- stan_model(stanc_ret = precomp)

#### DATA SAVE SETUP ####
true_params <- vector("list", nreps)
# result_objs <- vector("list", nreps)
est_param_summary <- vector("list", nreps)
est_param_means <- vector("list", nreps)
correlations <- vector("list", nreps)
params_extraction <- vector("list", nreps)
CIs_analysis <- vector("list", nreps)
CIs_proportion <- vector("list", nreps)

#setup output folder for use later
folder_name <- paste0(date, "_simulation-results")
file_tag <- paste0(nreps, "reps_",
                  gsub(".", "-", as.character(rho), fixed = TRUE), "rho_",
                  gsub(".", "-", as.character(P_REF), fixed = TRUE), "PREF_",
                  gsub(".", "-", as.character(mu2), fixed = TRUE), "mu_",
                  gsub(".", "-", as.character(alpha), fixed = TRUE), "alpha_",
                  date)

if(!dir.exists(paste0(work_dir, "/", folder_name))){
  dir.create(paste0(work_dir, "/", folder_name))
}
setwd(paste0(work_dir, "/", folder_name))

for(i in 1:nreps){
  setSeeds(seeds, run = i)

  #### SIMULATION ####
  #simulate a set of items
  true_item_params <- item_sim(n_items, b_mean = 0, b_sd = 1, a_min = 0.5, a_max = 3,
                             mix_alpha = alpha, mix_mu1 = mu1, mix_mu2 = mu2, mix_sdev = sdev)

  #simulate a set of people's ability scores
  true_ability <- ability_sim(n_people, P_REF = P_REF, ref_theta_mean = 0, ref_theta_sd = 1,
                             focal_theta_mean = -0.5, focal_theta_sd = 1)

  #get responses for a set of people to a set of items
  dataset <- one_dataset(true_ability, true_item_params)

```

```

#get values for the DIF predictor
DIFpredict <- DIF_predictor(true_item_params, rho = rho)

#calculate beta1 and beta0
beta1_true <- rho*sdev_D
beta0_true <- mean(DIFpredict) - (beta1_true*mean(true_item_params[, "dif_param"]))

#save the true parameters
true_params[[i]] <- list(true_item_params, true_ability, dataset, DIFpredict,
                        beta1_true, beta0_true)
names(true_params[[i]]) <- c("true_item_params", "true_ability", "dataset",
                             "DIFpredict", "beta1_true", "beta0_true")

#set up grouping variable
group <- true_ability[,2]
n_ref <- sum(group)

#restructuring the data to long format
dataset <- long_format(dataset, group)

#pulling the individual parts back out
respondentid <- dataset$respondentid
itemid <- as.numeric(dataset$itemid)
response <- dataset$response
group_long <- dataset$group

n_observations <- nrow(dataset)

#### ANALYSIS ####
#conducting the analysis
analysis <- sampling(precomp_model, data = b.dat_long,
                    iter = 12000, warmup = 5000, chains = 2, verbose = FALSE, cores = 2)

#save the analysis object
# result_objs[[i]] <- analysis

#### OUTPUT ####
#pull out the summary of the estimated parameters
params_summary <- summary(analysis, pars = c("a", "b", "D", "beta0", "beta1", "mu",
                                             "sigma2", "R2", "theta",
                                             "foc_mean"),
                          probs = c(0.025, 0.25, 0.5, 0.75, 0.975))$summary

#save the summary of the estimated parameters
est_param_summary[[i]] <- params_summary

```

```

#calculate the means of the estimated parameters
params <- extract(analysis, pars = c("a", "b", "D", "beta0", "beta1", "mu", "sigma2",
                                   "R2", "theta", "foc_mean"))

params_extraction[[i]] <- params

a_params <- as.matrix(colMeans(params$a))
b_params <- as.matrix(colMeans(params$b))
D_params <- as.matrix(colMeans(params$D))
beta0 <- mean(params$beta0)
beta1 <- mean(params$beta1)
mu <- as.matrix(colMeans(params$mu))
sigma2 <- mean(params$sigma2)
R2 <- mean(params$R2)
theta <- as.matrix(colMeans(params$theta))
foc_mean <- mean(params$foc_mean)

#save the means of estimated parameters
est_param_means[[i]] <- list(a_params, b_params, D_params, beta1, beta0,
                             mu, sigma2, R2, theta, foc_mean)
names(est_param_means[[i]]) <- c("a_params", "b_params", "D_params",
                                "beta1", "beta0", "mu", "sigma2",
                                "R2", "theta", "foc_mean")

#save the mean correlations & differences from the expected values
a_corr <- cor(a_params, true_item_params[, "a_param"])
b_corr <- cor(b_params, true_item_params[, "b_param"])
D_corr <- cor(D_params, true_item_params[, "dif_param"])
theta_corr <- cor(theta, true_ability[, 1])
foc_mean_diff <- foc_mean - (-.5)
ref_mean_diff <- mean(theta[1:n_ref]) - 0
R2_diff <- R2 - (rho^2)
beta1_diff <- beta1_true - beta1
beta0_diff <- beta0_true - beta0

correlations[[i]] <- list(a_corr, b_corr, D_corr, theta_corr,
                        foc_mean_diff, ref_mean_diff, R2_diff,
                        beta1_diff, beta0_diff)
names(correlations[[i]]) <- c("a_corr", "b_corr", "D_corr", "theta_corr",
                            "foc_mean_diff", "ref_mean_diff", "R2_diff",
                            "beta1_diff", "beta0_diff")

#save the number of true values falling within the confidence interval
param_types <- as.data.frame(unique(gsub("\\[.*", "", rownames(params_summary))))
colnames(param_types) <- "param"

```

```

param_types$dim <- c(rep("vec", 3), rep("scalar", 2),
                    rep("vec", 1), rep("scalar", 2),
                    rep("vec", 1), rep("scalar", 1))

for(j in 1:nrow(param_types)){
  if(param_types[j, "dim"] == "vec"){
    assign(paste0(param_types[j, "param"], "_params_summary"), params_summary[
      grep(paste0("^", param_types[j, "param"], "\\[", rownames(params_summary)),)]])
  } else if(param_types[j, "dim"] == "scalar"){
    assign(paste0(param_types[j, "param"], "_params_summary"), params_summary[
      grep(paste0("^", param_types[j, "param"]), rownames(params_summary)),])
  }
}

#item_params: a, b, D
a_param_CIs <- CI_retrieval(true_item_params[, "a_param"], a_params_summary)
a_param_CI_prop <- sum(a_param_CIs)/n_items

b_param_CIs <- CI_retrieval(true_item_params[, "b_param"], b_params_summary)
b_param_CI_prop <- sum(b_param_CIs)/n_items

D_param_CIs <- CI_retrieval(true_item_params[, "dif_param"], D_params_summary)
D_param_CI_prop <- sum(D_param_CIs)/n_items

#ability_params: theta
theta_param_CIs <- CI_retrieval(true_ability[, "theta"], theta_params_summary)
theta_param_CI_prop <- sum(theta_param_CIs)/n_people

#scalar params: beta0, beta1, R2
beta0_CIs <- CI_retrieval(beta0_true, t(beta0_params_summary))

beta1_CIs <- CI_retrieval(beta1_true, t(beta1_params_summary))

R2_CIs <- CI_retrieval(R2_true, t(R2_params_summary))

CIs_analysis[[i]] <- list(a_param_CIs, b_param_CIs, D_param_CIs,
                        theta_param_CIs, beta0_CIs, beta1_CIs,
                        R2_CIs)
names(CIs_analysis[[i]]) <- c("a_param_CIs", "b_param_CIs", "D_param_CIs",
                             "theta_param_CIs", "beta0_CIs", "beta1_CIs",
                             "R2_CIs")

CIs_proportion[[i]] <- list(a_param_CI_prop, b_param_CI_prop,
                            D_param_CI_prop, theta_param_CI_prop)
names(CIs_proportion[[i]]) <- list("a_param_CI_prop", "b_param_CI_prop",
                                   "D_param_CI_prop", "theta_param_CI_prop")

```

```
#### SAVE TO DISK ####
#write all the good stuff out to disk
saveRDS(true_params, paste0("true_params_", file_tag, ".rds"))
# saveRDS(result_objs, paste0("result_objs_", file_tag, ".rds"))
saveRDS(est_param_summary, paste0("est_param_summary_", file_tag, ".rds"))
saveRDS(params_extraction, paste0("params_extraction_", file_tag, ".rds"))
saveRDS(est_param_means, paste0("est_param_means_", file_tag, ".rds"))
saveRDS(correlations, paste0("correlations_", file_tag, ".rds"))
saveRDS(CIS_analysis, paste0("CIS_analysis_", file_tag, ".rds"))
saveRDS(CIS_proportion, paste0("CIS_proportion_", file_tag, ".rds"))
}
```

stan_scripts.R

```
stancode_long <- "
data {
  int<lower=0> n_people;
  int<lower=0> n_items;
  int<lower=0> n_observations;
  int<lower=0, upper=n_people> respondentid[n_observations];
  int<lower=0, upper=n_items> itemid[n_observations];
  int<lower=0, upper=1> response[n_observations];
  int<lower=0, upper=1> group_long[n_observations];
  vector[n_people] group;
  vector[n_items] DIFpredict;
}

parameters {
  vector<lower=0>[n_items] a;
  vector[n_items] b;
  vector[n_people] theta;
  vector[n_items] D_raw;
  real beta0;
  real beta1;
  real<lower=0> sigma2;
  real foc_mean;
  // real D[n_items];
}

transformed parameters {
  vector[n_items] mu;
  vector[n_items] ss_err;
  vector[n_items] ss_reg;
}
```

```

vector[n_people] mu_theta;
real R2;
vector[n_items] D;

mu_theta = foc_mean*group;

for (j in 1:n_items) {
  mu[j] = beta0 + beta1*DIFpredict[j];
}

D = mu + sigma2*D_raw;

for (j in 1:n_items) {
  ss_err[j] = pow((D[j]-mu[j]),2);
  ss_reg[j] = pow((mu[j]-mean(D[])),2);
}

R2 = sum(ss_reg[])/(sum(ss_reg[])+sum(ss_err[]));
}

model {
  vector[n_observations] eta;

  a ~ lognormal(0, 1);
  b ~ normal(0, 1);
  theta ~ normal(mu_theta, 1);
  D_raw ~ normal(0, 1);
  // D ~ normal(mu, sigma2);
  foc_mean ~ normal(0, 4);
  beta0 ~ normal(0, 1);
  beta1 ~ normal(0, 1);
  sigma2 ~ normal(0, 10);

  for(i in 1:n_observations){
    eta[i] = a[itemid[i]]*(theta[respondentid[i]] - (b[itemid[i]] + D[itemid[i]] * group_long[i]));
  }

  response ~ bernoulli_logit(eta);
}

```

mixture_functions.R

DATA GENERATION

```

#simulate a set of items
item_sim <- function(n_items, b_mean, b_sd, a_min, a_max,
                    mix_alpha, mix_mu1, mix_mu2, mix_sdev){
  item_param <- matrix(NA, nrow = n_items, ncol = 3)
  colnames(item_param) <- c("b_param", "a_param", "dif_param")

  item_param[, "b_param"] <- rnorm(nrow(item_param), b_mean, b_sd)
  item_param[, "a_param"] <- runif(nrow(item_param), a_min, a_max)
  k <- rbinom(nrow(item_param), 1, alpha)
  item_param[, "dif_param"] <- (rnorm(nrow(item_param), mix_mu1, mix_sdev)^k) *
    (rnorm(nrow(item_param), mix_mu2, mix_sdev)^(1-k))

  return(item_param)
}

#simulate a set of people's ability scores
ability_sim <- function(N_people, P_REF, ref_theta_mean, ref_theta_sd,
                       focal_theta_mean, focal_theta_sd){
  ability_scores <- matrix(NA, nrow = N_people, ncol = 2)
  colnames(ability_scores) <- c("theta", "group")
  ref_cutoff <- nrow(ability_scores)*P_REF
  ref_rows <- c(1:ref_cutoff)
  focal_rows <- c((ref_cutoff+1):nrow(ability_scores))

  ability_scores[ref_rows, "theta"] <- rnorm(length(ref_rows),
                                             ref_theta_mean, ref_theta_sd)
  ability_scores[ref_rows, "group"] <- 0

  ability_scores[focal_rows, "theta"] <- rnorm(length(focal_rows),
                                               focal_theta_mean, focal_theta_sd)
  ability_scores[focal_rows, "group"] <- 1

  return(ability_scores)
}

#get the responses for a single item
response_sim <- function(person_vec, item_vec){
  guts <- item_vec["a_param"]*(person_vec["theta"]-
                              (item_vec["b_param"]+item_vec["dif_param"]*person_vec["group"]))
  prob <- exp(guts)/(1+exp(guts))
  ifelse(runif(1, 0, 1) <= prob, return(1), return(0))
}

#get responses for a single person to a set of items
person_sim <- function(person_vec, item_param = item_param){
  responses_vec <- matrix(NA, nrow=nrow(item_param))

```

```

  for(i in 1:nrow(item_param)){
    responses_vec[i] <- response_sim(person_vec, item_param[i,])
  }
  return(responses_vec)
}

#get responses for a set of people to a set of items
one_dataset <- function(person_param, item_param){
  responses <- matrix(NA, nrow = nrow(person_param), ncol = nrow(item_param))
  for(i in 1:nrow(person_param)){
    responses[i,] <- person_sim(person_param[i,], item_param)
  }
  #colnames(responses) <- paste0("V", 1:nrow(item_param))
  return(responses)
}

#### PREPARATION ####
#get DIF predictor
DIF_predictor <- function(item_param, rho){
  mean_DIF <- mean(item_param[, "dif_param"])
  sd_DIF <- sd(item_param[, "dif_param"])
  zscores <- (item_param[, "dif_param"] - mean_DIF)/sd_DIF

  e1 <- rnorm(nrow(item_param), 0, sqrt(1-rho^2))

  DIF_predict <- rho*zscores + e1

  # DIF_predict <- sqrt(rho^2)*zscores + e1
  return(DIF_predict)
}

#### LONG FORMAT RESTRUCTURING ####
long_format <- function(data = dataset, group_data = group){
  #prep for reformatting
  data <- as.data.frame(data)
  names(data) <- paste0("Item", 1:ncol(data))
  data$respondentid <- c(1:nrow(data))

  #move to long format
  dataset_long <- gather(data, key = respondentid, value = response)
  names(dataset_long)[2] <- "itemid"

  #joining group
  group_data <- as.data.frame(group_data)
  group_data$respondentid <- c(1:nrow(group_data))
  dataset_long <- left_join(dataset_long, group_data, by = "respondentid")
}

```

```

dataset_long$itemid <- gsub("Item", "", dataset_long$itemid)
names(dataset_long) <- c("respondentid", "itemid", "response", "group")
}
return(dataset_long)
}

#### ANALYSIS ####

#do the analysis for one set of responses
one_analysis <- function(x, n_iter = 2000, n_burn = 1000, n_chains = 2,
                        modelname = "stan_model", b_dat = b.dat,
                        n_cores = 2, debug = FALSE){
  if(class(x) == "stanmodel"){
    OUT <- sampling(x, data = b.dat,
                  iter = n_iter, warmup = n_burn, chains = n_chains,
                  verbose = debug, cores = n_cores)
  } else if(class(x) == "character"){
    OUT <- stan(x, model_name = modelname, data = b.dat,
              iter = n_iter, warmup = n_burn, chains = n_chains,
              verbose = debug, cores = n_cores)
  } else {
    stop("Please specify a pre-compiled Stan model or a character variable
         containing a model specification in the Stan modeling language")
  }
  return(OUT)
}

one_analysis_BUGS <- function(x, n_iter = 1000, n_burn = 300, b_dat = b.dat,
                             b_par = b.par, model_file = "BUGScode.txt"){
  vars <- c(unlist(b_dat))
  mget(vars, envir = globalenv())
  OUT <- R2OpenBUGS::bugs(data = b_dat, inits = NULL, parameters.to.save = b_par,
                        model.file = model_file, n.chains = 2,
                        n.iter = n_iter, n.burn = n_burn, n.thin = 1, debug = TRUE)
  return(OUT)
}

#### PROCESSING - DATA RETRIEVAL ####
correlation_get <- function(condition, file_list){
  output <- readRDS(paste0(file_list[grepl(condition, file_list)]))
  output <- lapply(output, unlist, recursive = FALSE)
  output <- do.call(rbind, output)
  return(output)
}

```

```

}
true_param_get <- function(condition, file_list, param_type, param_name){
  output <- readRDS(paste0(file_list[grep1(condition, file_list)]))

  param <- vector("list", length(output))

  for(i in 1:length(output)){
    param[[i]] <- as.data.frame(output[[i]][param_type])
  }

  for(i in 1:length(param)){
    param[[i]] <- as.data.frame(param[[i]][grep(param_name, names(param[[i]])]))
  }

  param <- bind_rows(param, .id = names(output))
  return(param)
}

est_param_get <- function(condition, file_list, param_name){
  output <- readRDS(paste0(file_list[grep1(condition, file_list)]))

  param <- lapply(output, as.data.frame)
  param <- bind_rows(param, .id = names(output))
  return(param)
}

est_param_means_get <- function(condition, file_list, param_name){
  output <- readRDS(paste0(file_list[grep1(condition, file_list)]))

  param <- lapply(output, function(x) as.data.frame(x[param_name]))
  param <- bind_rows(param, .id = names(output))
  return(param)
}

#### CI ANALYSIS ####
CI_retrieval <- function(true_param_vec, est_param_mat){
  if(NROW(true_param_vec) != NROW(est_param_mat)){
    stop("Unequal param count!")
  }

  CI_out <- c(rep(NA, NROW(true_param_vec)))

  for(i in 1:NROW(true_param_vec)){
    CI_out[i] <- ifelse(true_param_vec[i] > est_param_mat[i, "2.5%"] &
      true_param_vec[i] < est_param_mat[i, "97.5%"], TRUE, FALSE)
  }
}

```

```
    }
  return(CI_out)
}

#### GRAPHING ####
scale_def <- function(list, column){
  scale <- NA
  for(i in 1:length(list)){
    rounded <- abs(c(round(max(list[[i]][, column]), digits = 1),
                    round(min(list[[i]][, column]), digits = 1)))
    scale[i] <- rounded[which.max(rounded)]
  }
  scale <- scale[which.max(scale)]
  return(scale)
}

scale_def_corr <- function(list, column){
  scale <- NA
  for(i in 1:length(list)){
    scale[i] <- (floor(((min(list[[i]][, column])) * 10)) / 10)
  }
  scale <- c(scale[which.min(scale)], 1)
  return(scale)
}
```