2023

Establishing Roots Before Branching Out: Parameter Recovery in Item Response Tree Models

Tyler Ryan
Wright State University

ESTABLISHING ROOTS BEFORE BRANCHING OUT: PARAMETER RECOVERY IN ITEM RESPONSE TREE MODELS

A thesis submitted in partial fulfillment of the
requirements for the degree of
Master of Science

By

TYLER RYAN
B.A. Wright State University, 2013
M.A. Wright State University, 2017

2023
Wright State University

WRIGHT STATE UNIVERSITY

COLLEGE OF GRADUATE PROGRAMS AND HONORS STUDIES

04/21/23

I HEREBY RECOMMEND THAT THE THESIS PREPARED UNDER MY SUPERVISION BY Tyler Ryan ENTITLED Establishing Roots Before Branching Out: Parameter Recovery in Item Response Tree Models BE ACCEPTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF Master of Science.

David LaHuis, Ph.D.
Thesis Director

Scott Watamaniuk, Ph.D.
Graduate Program Director

Debra Steele-Johnson, Ph.D.
Chair, Department of Psychology

Committee on Final Examination:

David LaHuis, Ph.D.

Debra Steele-Johnson, Ph.D.

Joseph Houpt, Ph.D.

Shu Schiller, Ph.D.
College of Graduate Programs & Honors Studies

ABSTRACT

Ryan, Tyler. M.S. Department of Psychology, Wright State University, 2023. Establishing Roots Before Branching Out: Parameter Recovery in Item Response Tree Models.

Item Response Trees are a type of item response model that incorporates information about conditional responding to items using a rooted tree graph structure. Researchers have used item response trees for common measurement tasks and for testing novel hypotheses. Previous simulation studies investigating item response trees either lack generalizability to the broad domain of their use or lack thorough investigation and reporting of the results. I conducted a simulation study to explore how sample size, test length, item characteristics, and tree structure affect both item and person parameter recovery for 1PL and 2PL models. The results suggested that, as with any item response model, item response tree models are unbiased. However, large samples and long test lengths are needed to minimize estimate uncertainty. Issues of sample size and test length are compounded by the conditional structure incorporated in item response tree models. In particular, the depth of the tree and low item endorsement can pose severe estimation issues when sample sizes are not large and test lengths are not long. I used posterior predictive simulations to provide the reader with a practical understanding of the limitations of item response trees in the context of item and personnel selection and prediction of external variables.

TABLE OF CONTENTS

Page

LIST OF FIGURES

LIST OF TABLES

Introduction

Item Response Theory (IRT) provides a useful framework for investigating the relationships between latent variables and manifest variables. IRT is commonly used in survey research because it allows researchers and practitioners to examine and calibrate their tests and surveys, potentially making them more valid and reliable. Simulation studies allow researchers to investigate the ability of a statistical model to estimate underlying hypothesized parameters accurately and efficiently. The IRT literature contains a plethora of simulation studies validating the most common models (Harwell et al., 1996). De Boeck and Partchev (2012) introduced a subset of IRT models called Item Response Trees (IRTrees) to model response processes with tree-like conditional probability structures. IRTree models incorporate multiple latent states or traits involved in sequential decision-making processes that result in observed behaviors or responses to items and stimuli. Simulation studies on IRTrees are sparse in the IRT literature. Existing IRTree simulation studies focus either on a specific type of IRTree model not in common use or do not thoroughly present the results of more generalizable IRTree simulations. Although there may be reason to believe that the findings of the past simulation studies investigating other more common IRT models apply to IRTree models, researchers have applied IRTrees in new unique ways that are not covered by past research. Most importantly, IRTrees often require coding of observed responses in a way that introduces missingness and may present unique problems for parameter estimation. Thus, the purpose of this study was to investigate parameter recovery with IRTree models by conducting a simulation study. First, I will review

IRT and IRTree models and the pertinent literature. Then, I will explain the simulation study method and analysis. Finally, I will discuss the results.

Item Response Theory

IRT is both a theory of model-based measurement and a family of statistical measurement models that relate latent unobserved variables to their observed or manifest variables (Embretson & Reise, 2000). In the domain of psychological assessment, test and survey items are the manifest variables, which are said to measure or indicate some unobserved latent psychological ability, trait, or state. As measurement models, IRT models separate parameters representing person-based characteristics (e.g., psychological constructs or abilities) from item-based characteristics (e.g., item difficulty). Characteristics of the person interact with the characteristics of the item to produce an observed response. IRT models may be distinguished from one another by the number of parameters involved and the functional form of the person-item interaction.

Recently, researchers have used IRTree models as a means of measuring different latent variables involved in the process of producing an observed response (De Boeck & Partchev, 2012). Most researchers who have used IRTree models have used either the one-parameter or two-parameter logistic function for dichotomous responses. Some authors have incorporated polytomous response functions in IRTree models as well (e.g., Meiser et al., 2019). Many IRT models are unidimensional and only measure a single latent trait. Most IRTree models are multidimensional models, such that they measure multiple latent person abilities. Hypotheses about the presence of multiple latent attributes can be tested with IRTree models by comparing typical multidimensional models to constrained unidimensional models. I review some basic dichotomous and polytomous models briefly here for two reasons. First, I use both one-parameter and two-parameter models in the present simulation study. Second, it is useful to understand the underlying theoretical model for typical dichotomous and polytomous models in order to compare them to the underlying theoretical model posed by IRTrees when analyzing survey response data.

One-Parameter Logistic Model

The simplest IRT model is the one-parameter logistic model (1PL), also known as the Rasch model. 1PL models relate dichotomous items to one or more latent traits. The odds of endorsing a given item are equivalent to the ratio of the probability of successfully endorsing an item to the probability of failing to endorse the item, $\frac{p_{ij}}{1-p_{ij}}$. Taking the log of the odds, we see that endorsement is a function of the discrepancy between the log of the probability of success and the log probability of failure. This discrepancy is equivalent to the difference between one's latent ability and the difficulty of the item, expressed in logits, z_{ij}. The univariate 1PL model has one item-parameter, β_j, which is an estimate of item j's difficulty, and one person-parameter, θ_i, which is the ability or latent trait level i, such that,

$$\log\frac{p_{ij}}{1-p_{ij}} = \log(p_{ij}) - \log\left(1-p_{ij}\right) = \theta_i - \beta_j = z_{ij}.$$

The resultant probability of endorsement for the logistic model is given by the item response function (IRF),

$$\Pr(Y_{ij} = 1|\theta_i, \beta_j) = \Psi(\theta_i - \beta_j) = \frac{1}{1+e^{-1(\theta_i-\beta_j)}} = \frac{1}{1+e^{-z_{ij}}}$$

When a respondent's ability is equivalent to the item difficulty, $\theta_i = \beta_j$, the discrepancy is $\theta_i - \beta_j = 0$, and the resultant probability is,

$$\Pr(Y_{ij} = 1 \,\middle|\, \theta_i, \beta_j) = \frac{1}{1+e^0} = \frac{1}{2} = 0.50.$$

When a respondent's ability is greater than the item difficulty, the discrepancy is positive, and the endorsement probability is greater than 0.50. Likewise, when the respondent's ability is less than the item difficulty, the discrepancy is negative, and the endorsement probability is less than 0.50. In the case of a single latent ability parameter, the unidimensional case, ability is commonly assumed to be normally distributed $\theta \sim N(0, 1)$.

The 1PL functional form assumes that a respondent chooses a response based on whether the item difficulty or location is greater than or less than their ability. This is most obvious in ability testing where items considered difficult by a respondent are not answered correctly. The function assumes that the probability of correct endorsement increases monotonically as the ability of the respondent increases or the difficulty of the item decreases, sometimes called a dominance response process (Stark et al., 2006). In contrast, ideal point response processes suppose that the absolute difference between a respondent's latent trait and the item characteristic is determinative of the endorsement probability. The ideal point process is often considered for non-cognitive test items such as personality items, where statements are endorsed based on their proximity to a respondent's belief. Researchers have used dominance process models for IRTrees, but there is nothing in principle that suggests ideal point models could not be used as well. I focus on dominance process models for the present study.

IRT models can be specified as either unidimensional or multidimensional. In the unidimensional case, the items are assumed to measure a single construct such that only one latent ability is involved in responding. A multidimensional model assumes that the items measure multiple constructs such that several latent abilities are involved in responding. In the multidimensional case, θ_i becomes a vector of length K equal to the number of latent ability parameters involved in responding to item j, where $\theta_i = \theta_{i1}, \dots, \theta_{ik}, \dots, \theta_{iK}$ and $\theta \sim$

$MVN(0, \Sigma_\theta)$. The $K \times K$ covariance matrix, Σ_θ, can be freely estimated or fixed to impose orthogonality/equality on the latent traits, depending on the goals of the researcher.

Multidimensional models may be specified as either compensatory or non-compensatory. Compensatory multidimensional models suggest that the probability of a given response to an item involves the sum of each ability, such that sufficiency in one dimension can compensate for insufficiency in another dimension. The logit discrepancy function becomes, $z_{ij} = \sum_{k=1}^{K} \alpha_{jk} (\theta_{ik} - \beta_j)$, and the logistic function becomes,

$$\Pr(Y_{ij} = 1 \,|\, \theta_i, \alpha_{jk}, \beta_j) = \frac{1}{1 + e^{-\sum_{k=1}^{K} \alpha_{jk}(\theta_{ik} - \beta_j)}} = \frac{1}{1 + e^{-\sum_{k=1}^{K} z_{ijk}}}.$$

The non-compensatory case requires some level of competency on each ability dimension in order to endorse a given response. The resultant logit function is, $z_{ijk} = \alpha_{jk}(\theta_{ik} - \beta_j)$, and the logistic function becomes,

$$\Pr(Y_{ij} = 1 \,|\, \theta_i, \alpha_{jk}, \beta_j) = \prod_{k=1}^{K} \frac{1}{1 + e^{-\alpha_{jk}(\theta_{ik} - \beta_j)}} = \prod_{k=1}^{K} \frac{1}{1 + e^{-z_{ijk}}}.$$

Although IRTrees may be specified as unidimensional models, IRTrees typically involve multiple latent abilities and are often formulated as non-compensatory models. IRTree models typically imply that the respondent is required to use multiple traits during the response process. Because of the flexibility of the IRTree modeling framework, researchers can specify models with a variety of latent factor structures, person parameters, and item parameters. Similarly, researchers may also specify different functional forms for the person-item interaction.

Two-Parameter Logistic Model

The two-parameter logistic (2PL) response model is an extension of the 1PL model where an additional item parameter, α_j, is used to estimate an item’s ability to discriminate between levels of respondents’ abilities. The discrimination parameter may be interpreted as the expected change in the log-odds of item endorsement given a one-unit change in respondent ability (or item difficulty). The discrepancy function becomes $z_{ij} = \alpha_j(\theta_i - \beta_j)$. The 1PL model can be viewed as a special case of the 2PL model in which the discrimination parameter is fixed to one (or some other constant), such that all items equally discriminate between different levels of ability and equally load onto the latent variable. The logistic function can then be written as,

$$\Pr(Y_{ij} = 1 \,\middle|\, \theta_i, \alpha_j, \beta_j) = \frac{1}{1 + e^{-\alpha_j(\theta_i - \beta_j)}}.$$

Polytomous Models

Polytomous response models, involving three or more response options, $m >= 3$, have largely been modeled with either difference measurement models, such as the graded response model (Samejima, 1969), or divide-by-total measurement models, such as the nominal response model (Bock, 1972). Tutz (1990) formulated the sequential response model, which is neither strictly a difference nor divide-by-total model for ordered polytomous outcomes. The sequential model is formulated as a series of binary response functions beginning at the lowest response option, progressively comparing each option with the next highest adjacent response option until failing to endorse. The response function can be written as,

$$\Pr(Y_{ij} = m \in \{1, \dots, M\} | \theta_i, \beta_{jr}) = \prod_{r=1}^{m-1} \frac{e^{(\theta_i - \beta_{jr})^{T[m,r]}}}{1 + e^{\theta_i - \beta_{jr}}}$$

where M is the number of response options on the ordered polytomous scale, $M - 1 = R$ is the total number of category thresholds, r indexes the $m - 1$ category thresholds crossed to progress from the first response to response m, T is an $M \times R$ binary matrix indexed by m and r such that $T_{[m,r]} = 1$ for successfully crossing threshold r and $T_{[m,r]} = 0$ for failing to cross threshold r. For example, say I have a respondent with average ability, $\theta = 0$, responding to an item with a 5-point scale, $M = 5$, $M - 1 = R = 4$ thresholds, and threshold difficulties equal to a r-length vector $\beta_j = [-2,-1,1,2]$. In order to respond with a 3, the respondent must first assess response options 1 versus 2,

$$\Pr(Y_{ij1}^{*} = 1|\theta_i = 0, \beta_{j1} = -2) = \frac{1}{1 + e^{-(0--2)}} = 0.881,$$

where $Y_{ij1}^{*} = 1$ indicates successful completion of response step 1. Given that they have successfully completed the first step, the respondent may then assess response options 2 versus 3,

$$\Pr(Y_{ij2}^{*} = 1|\theta_i = 0, \beta_{j2} = -1, Y_{ij1}^{*} = 1) = \frac{1}{1 + e^{-(0--1)}} = 0.731.$$

Finally, after successfully completing response step 2, the respondent must weigh options 3 versus 4,

$$\Pr(Y_{ij3}^{*} = 1|\theta_i = 0, \beta_{j3} = -1, Y_{ij2}^{*} = 1) = \frac{1}{1 + e^{-(0-1)}} = 0.269.$$

Assuming the respondent fails to successfully complete this third step, they finish their assessment and endorse response option 3. The resultant probability of endorsing a 3 is the product of the probabilities of each step attempted in the response process,

$$\Pr(Y_{ij} = 3|\theta_i, \beta_j) = \Pr(Y_{ij1}^{*} = 1)\Pr(Y_{ij2}^{*} = 1)[1 - \Pr(Y_{ij3}^{*} = 1)].$$

The probability of any response option is thus conditional on successful completion of previous steps. The response probabilities for a 5-points scale are,

$$\begin{aligned}
\Pr(Y_{ij} = 1|\theta_i, \beta_j) &= [1 - \Psi(\theta_i - \beta_{j1})] \\
\Pr(Y_{ij} = 2|\theta_i, \beta_j) &= \Psi(\theta_i - \beta_{j1})[1 - \Psi(\theta_i - \beta_{j2})] \\
\Pr(Y_{ij} = 3|\theta_i, \beta_j) &= \Psi(\theta_i - \beta_{j1})\Psi(\theta_i - \beta_{j2})[1 - \Psi(\theta_i - \beta_{j3})] \\
\Pr(Y_{ij} = 4|\theta_i, \beta_j) &= \Psi(\theta_i - \beta_{j1})\Psi(\theta_i - \beta_{j2})\Psi(\theta_i - \beta_{j3})[1 - \Psi(\theta_i - \beta_{j4})] \\
\Pr(Y_{ij} = 5|\theta_i, \beta_j) &= \Psi(\theta_i - \beta_{j1})\Psi(\theta_i - \beta_{j2})\Psi(\theta_i - \beta_{j3})\Psi(\theta_i - \beta_{j4}).
\end{aligned}$$

The sequential choice model requires that response option endorsement is conditional on endorsement and non-endorsement of other response options. The model does not require that all responses be considered because the response process terminates once the respondent fails to endorse the next highest option. They cannot proceed and assess higher response options unless they have successfully passed lower steps. These properties make sequential response models attractive when modeling a linear response process that has multiple steps.

Tutz and Draxler (2019) suggest that the sequential response model is advantageous because it models a single latent trait involved in a linear set of response steps with progressive achievement indicating greater ability. However, the assumption that the entire response process is linear assumes that respondents assess and respond to the problem at hand in a linear fashion. The sequential response model makes intuitive sense for problems that are structured to have some progressive or linear response process, such as solving a math problem. A simple algebra problem on a math competency test (e.g., $20 = .5x + 10$, solve for x) may require a respondent to show their work step-by-step to solve for an unknown variable. Each step is done in a specific order, and one cannot solve the entire problem without progressively completing each step. But much of the time, items in psychological testing and survey research do not have an obvious linear process for responding. On a personality survey measuring Conscientiousness, an item might read “I shirk my duties” and ask the respondent to rate their agreement to the statement on a 1 to 6 scale. Such an item does not readily suggest that a respondent must progressively assess the response options, beginning with the lowest option, as one would when solving a math problem. Likewise, Tutz’ (1990) formulation suggests that only one latent ability is involved when responding to an item, a restriction that may not hold in some contexts (e.g., word problems requiring both mathematical and verbal abilities). Item response trees offer a flexible extension to the sequential response model framework.

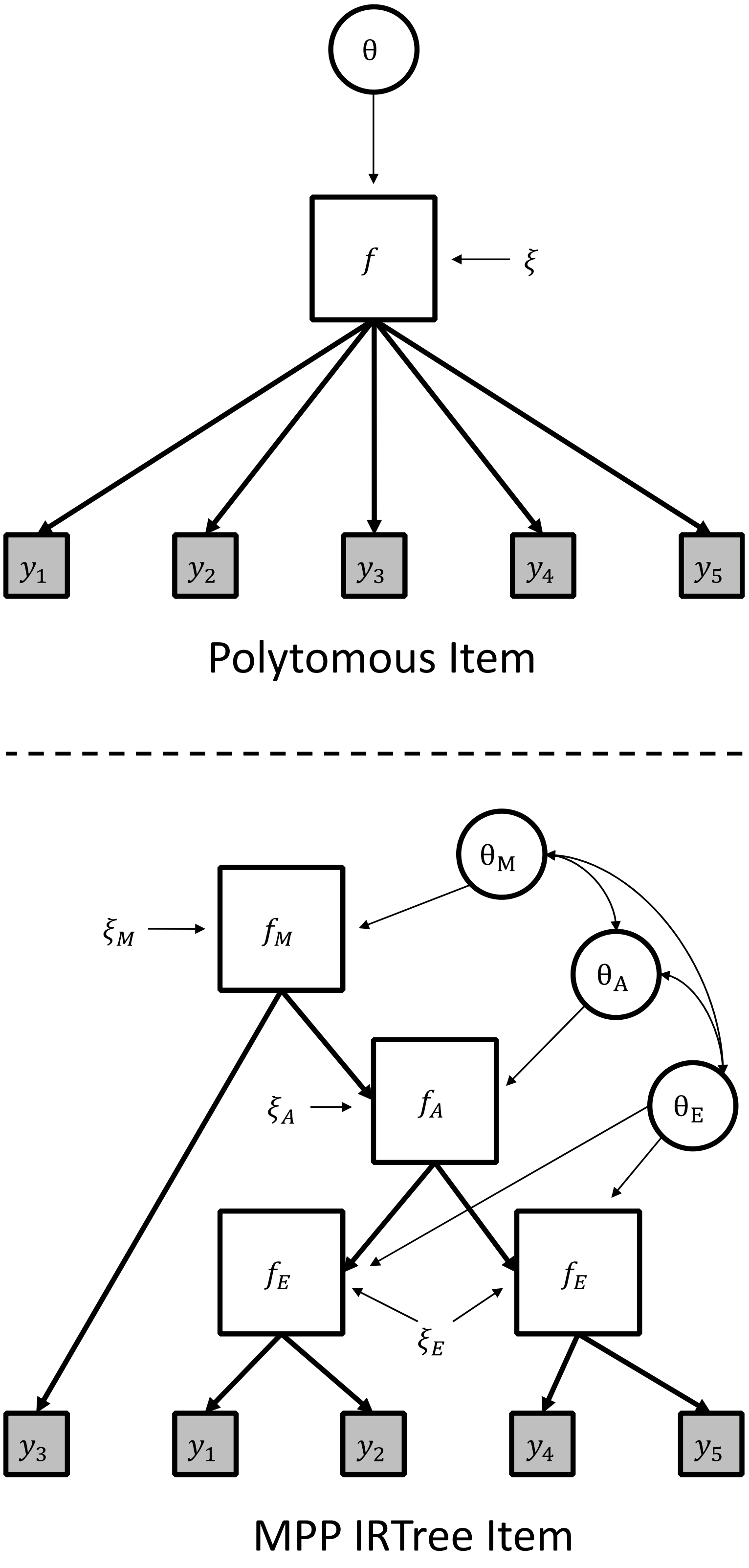
Item Response Trees

De Boeck and Partchev (2012) proposed a general class of IRT models that take on a tree-like structure, which they called Item Response Trees (IRTrees). De Boeck and Partchev classified these models as either response tree models or latent-variable tree models. Response tree models consist of a sequence of (possibly unobserved) responses leading to a terminal observed response. Latent-variable tree models consist of a sequence of latent-variables leading to an observed end response. Given that response tree models have received far more attention in previous literature and are relevant to the development and assessment of survey instruments, the present research and further mention of IRTree models will refer strictly to response tree models. I will first explain the terminology that I adopt to discuss IRTrees in this study. I will then explain the general structure of IRTrees and provide an example of coding observations for an IRTree model. Finally, I will discuss the general item response function of IRTrees.

IRTree terminology and implementation

I will borrow language used in graph theory to describe IRTree models and consider IRTrees as directed and rooted tree graphs. For purposes of consistency and clarity, I will refer to items in an IRTree model as either nodes when discussing the models generally or as auxiliary items when in the context of survey and test data that require recoding observed responses. Figure 1 displays a model of a polytomous item and that polytomous item coded into a Midpoint Primary Process (MPP) IRTree item. Both items utilize 5-point response outcomes, y, person ability parameter(s), θ, and item parameter(s), ξ, which are incorporated into response function(s), f. The MPP IRTree item contains several auxiliary items. The items are "auxiliary" in the sense that a single polytomous item is recoded into multiple, often dichotomous, items using external information such as a hypothesized structure or survey design. Auxiliary-items or nodes represent behaviors or decisions made by a respondent, often in response to some stimulus such as a test item. Responses to some nodes lead to subsequent nodes that entail further responding. Some nodes may evoke some final or terminal response without a subsequent node, referred to as a “leaf.” For example, in Figure 1, the node f_M has two possible responses, one of which leads to another node, the node f_A and the other to a terminal response leaf “3.” Nodes that lead directly to subsequent nodes are referred to as parent-nodes with regard to their direct descendants and are referred to as ancestor-nodes with regard to their direct and indirect descendants. In Figure 1, node f_M is a parent of node f_A, while node f_A is a parent of node f_E. f_M is also an ancestor to nodes f_A and f_E. Nodes that are resultant of decisions made at antecedent nodes are referred to as child-nodes with respect to their parent-node and descendant nodes with respect to their ancestors. A parent-node may have multiple child-nodes, a child-node may have multiple parent-nodes, and either may lead to multiple terminal response leaves.

Figure 1. *Polytomous Item and Midpoint Primary Process IRTree Item*



I will refer to the initial node in a tree as the “root” from which all other nodes and leaves stem from. The paths between nodes in the tree are directed and are referred to as branches. The hypothesized response process begins at the root-node and proceeds to each descendant node along a given branch to a terminal response leaf. Each node has a probability of endorsement, determined by a set of item and person parameters. The probability of observing some terminal response leaf is conditional on the endorsement or non-endorsement of all of its ancestor nodes. The theoretical or graphical structure of a given tree implies a directed path. However, it does not necessarily imply some temporal or causal ordering of the ancestor nodes. For illustrative purposes, I will discuss IRTrees and their application in psychological research with language that implies temporal or causal ordering. Finally, although Figure 1 depicts a model with nodes that have only dichotomous response options, IRTrees may also generalize to polytomous responses. For the present study, I will focus solely on IRTrees with dichotomous nodes.

IRTree models involve sequences of responses, either observed or hypothesized. Researchers have used IRTrees to model response processes of respondents to survey instruments with polytomous rating scales. Instead of assuming that respondents weigh each response option against all other response options in a single decision (e.g., partial credit model, nominal response model), or assuming that respondents progressively weigh one option against the next highest option (e.g., sequential response model), IRTree models suggest that multiple (possibly unobserved) decisions are made which eventually lead to the observed response. In the context of a polytomous test item, an IRTree model could be constructed by observing responses to the item with a polytomous rating scale and hypothesizing a series of sub-questions the respondent poses to themselves to arrive at the observed response. A researcher could construct many different IRTree models depending on how the polytomous items are recoded and the hypothesized ordering of the response process.

For example, to model an item with a five-point scale, a midpoint primary process IRTree model (MPP, LaHuis et al., 2019) supposes that a respondent first decides whether to provide a neutral or directed response for the question posed. If a respondent chooses to endorse a neutral position, the midpoint (i.e., 3) of the scale is selected and the response process ends. If a respondent chooses not to endorse the midpoint, they proceed to the next decision process which involves deciding whether to agree or disagree to the question posed. Finally, the respondent proceeds to decide whether to provide an extreme response or not. Figure 1 depicts a polytomous rating scale item and an MPP IRTree item. White boxes indicate (possibly unobserved) response steps or nodes, grey boxes indicate observed response outcomes, and circles indicate latent person abilities used at each node. For the first decision, the MPP model hypothesizes that a respondent employs their propensity to endorse the midpoint of a scale, a construct from the response style literature (Van Vaerenbergh & Thomas, 2013). The first decision may also reflect the relevancy of the item to the respondent or the neutrality of the respondent towards the item. The last decision in the MPP model, the decision to provide an extreme response or not, is typically assumed to be the result of one’s extreme response style. The second decision, the agreement vs disagreement decision, is a measure of the substantive construct of interest measured without influence from the respondent’s midpoint and extreme response styles. The MPP model hypothesizes a specific ordering of the decisions in the response process and can be contrasted from alternative orderings such as the agreement primary process (LaHuis et al., 2019). The agreement primary process IRTree model hypothesizes that the respondent first decides to agree or disagree with the item content, then decides whether to provide an extreme response or not, and then finally decides whether to endorse a neutral position if an extreme response is not desired.

Many researchers using IRTrees recode observed responses to polytomous test items into auxiliary-items, one for each hypothesized decision process. In the MPP model, each node is coded as either a zero or one, indicating non-endorsement and endorsement, respectively. Some IRTree models, such as the MPP model, suggest that the response process can terminate without activating child-nodes further down the tree. If I observed that the respondent endorsed a 3 on the five-point polytomous rating scale, I would code the midpoint auxiliary-item as 1 and the agreement and extreme auxiliary-items as missing data. In this case, the model hypothesizes that the respondent terminated their response process after the first decision without engaging subsequent decision processes. Table 1 displays the coding matrix that provides the recoding scheme for a midpoint primary process IRTree model with five response options and three decision processes.

Table 1. *Midpoint Primary Process Coding Matrix.*

Rating Scale Response	M	A	E
1	0	0	1
2	0	0	0
3	1	-	-
4	0	1	0
5	0	1	1

Table 2. *Midpoint Primary Process D-Matrix.*

$$D = \begin{array}{c} \begin{array}{ccc} M & A & E \end{array} \\ \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix} \end{array}$$

Table 3. *Midpoint Primary Process T-Matrix.*

$$T = \begin{array}{c} \begin{array}{ccc} M & A & E \end{array} \\ \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \end{array}$$

IRTree models involve both person-specific and item-specific parameters. IRTree models are measurement models of hypothesized latent factor structures that characterize the respondents and their standing on some latent construct. A model may be specified as measuring a single factor by all nodes, measuring multiple factors with each node measuring its own, or a combination of the two (e.g., bifactor structure). In Figure 1, three latent factors are hypothesized for the MPP model, one for each node. Each node has some set of fixed parameters such as the difficulty of endorsing a node or the strength of association between the latent construct and node endorsement. Node parameters may be freely estimated or constrained to test hypotheses about specific decision processes. For example, in the midpoint primary process model, a researcher could specify the auxiliary-item parameters of the two extreme nodes as equivalent, such that the decision to endorse an extreme response is identical regardless of whether one chooses to agree or disagree with the agreement node. Conversely, a researcher could specify an extreme-disagreement and extreme-agreement node with a simple modification to the coding matrix and estimate unique auxiliary-item parameters for each extreme node. The researcher may then test hypotheses about the equivalence between the nodes by comparing the fit of this model to one that fixes the parameters. The response function of the model specifies how the node or item parameters and the underlying factors being measured interact to produce the observed responses.

IRTree Response Function.

The basic 1PL IRTree model is an extension of the non-compensatory 1PL model. Equation 1 gives the probability of a given observed response m to item j for respondent i with a set of abilities θ_i.

$$\Pr(Y_{ij} = m|\theta_i, \beta_j) = \prod_{r=1}^{R} \left[\frac{e^{(\theta_{ir} - \beta_{jr})^{d_{mr}}}}{1 + e^{\theta_{ir} - \beta_{jr}}} \right]^{t_{mr}} \tag{1}$$

The difficulty parameter, β_{j_r}, becomes specific to a given node, r. The coding matrix can be represented with two separate matrices that represent endorsement/non-endorsement probability for a given node and contribution of a node to a given response on the original observed rating scale. The exponent inside the brackets, d_{mr}, indexes the matrix *D*, depicted in Table 2, where m indexes the rows of terminal response outcomes and r indexes the columns of nodes. Indexing the *D*-matrix across nodes, when $d_{mr} = 1$, the equation inside the brackets equals the probability of endorsing node r. When $d_{mr} = 0$, the numerator equals one and the equation inside the brackets equals the inverse probability. The exponent outside the brackets, t_{mr}, indexes a matrix T, depicted in Table 3, that indicates whether a given node contributes to the (non-)endorsement of a given terminal response. When $t_{mr} = 1$, the inside equation contributes to the terminal response option. When the terminal response occurs before a child-node is reached, $t_{mr} = 0$. Then, the value inside the brackets equals 1 and does not contribute to the terminal response option probability. Note that the choice of defining missing values in the *D*-matrix as either 0 or 1 is arbitrary because the T-matrix ensures that they will not contribute to the terminal response probability. Response probabilities to each node are multiplied to produce the probability of endorsing a given terminal response option on the original polytomous rating scale.

To demonstrate the midpoint primary process model, suppose a respondent with a midpoint, agreement, and extreme latent ability of 0.5, 1.0, and -1.0, respectively, is given an item with node-difficulty thresholds 0.0, 0.5, and 1.0, respectively. Given the coding scheme for an MPP IRTree model, we could expect response option probabilities to be,

$$\begin{aligned}
\Pr(Y_j = 1|\Theta) &= \Psi(0.50 - 0.00)^{-1} \times \Psi(1.00 - 0.50)^{-1} \times \Psi(-1.00 - 1.00) = 0.017 \\
\Pr(Y_j = 2|\Theta) &= \Psi(0.50 - 0.00)^{-1} \times \Psi(1.00 - 0.50)^{-1} \times \Psi(-1.00 - 1.00)^{-1} = 0.126 \\
\Pr(Y_j = 3|\Theta) &= \Psi(0.50 - 0.00) = 0.622 \\
\Pr(Y_j = 4|\Theta) &= \Psi(0.50 - 0.00)^{-1} \times \Psi(1.00 - 0.50) \times \Psi(-1.00 - 1.00)^{-1} = 0.207 \\
\Pr(Y_j = 5|\Theta) &= \Psi(0.50 - 0.00)^{-1} \times \Psi(1.00 - 0.50) \times \Psi(-1.00 - 1.00) = 0.028
\end{aligned}$$

where Ψ indicates the logistic function. Note that the terminal response probabilities sum to one.

There is some caution required when interpreting IRTree models. The language used to describe IRTrees as sequential or process models may imply some casual ordering. Despite the conditional nature of response probabilities for child-nodes, the mathematical form of the IRTree response function does not itself imply temporal or causal ordering. Such an interpretation would require a study design that presents items or stimuli in a determined order. This is not to say that a theory of causal order cannot inform the construction of an IRTree model.

Many applications of IRTree models involve recoding a single observed response per item into multiple hypothesized unobserved auxiliary-item responses. A further issue with recoding items into multiple hypothesized auxiliary-items may be the indeterminacy of their conditional ordering. Leventhal (2020) suggested that the actual ordering of the decision nodes is indeterminate because the probabilities are commutative, such that the same terminal response probabilities are achieved regardless of the order in which the decision probabilities are multiplied. This is the case for models that either have no responses coded as missing or have the same nodes coded as missing under the same conditions. Take the agreement primary process (APP) model for example, another possible formulation of an IRTree model like the MPP. In the APP model, the agreement decision is made first, followed by the extreme, and finally the midpoint. This would result in the same probabilities as an extreme primary process (EPP) model, in which the extreme decision is made first, followed by the agreement decision, and finally the midpoint decision. In both APP and EPP models, the missing values encoded in the auxiliary-items are in the same places. In contrast, the same cannot be said for the MPP model as the missing values are not encoded the same way and given identical person and item parameters, result in different terminal response probabilities. The determining factor in distinguishing models that hypothesize the same nodes but in different orders is whether the coding schemes differ for the auxiliary-items and result in different instances of missingness.

Multinomial Processing Trees

A related class of models worth mentioning found in literature on cognitive models of decision making are Multinomial Processing Tree models (MPT, Batchelder, 1998). The main differences between IRTrees and MPT models are the research questions involved, the levels of analysis, and the functional forms of the models themselves. MPT models are often used to study cognitive processes, such as information encoding, recall, and response selection whereas IRTree models typically focus on higher order features and constructs such as personality or preference. Finally, although not in absolute terms, MPT models tend to model aggregate response category frequencies for the whole sample as a function of a small number of parameters involved in one or several nodes in a tree. This contrasts with IRTrees, which separate characteristics of the individual items from characteristics of the individual respondent, typically for each node. MPT models often estimate separate node parameters based on experimental manipulations that characterize groups of items and some extensions allow for person specific parameters (Matzke et al., 2015). Although nothing restricts MPT models to such configurations in principle, MPT models used in the literature do not appear to fully separate person and item characteristics like IRTree models do. As a result, MPT models typically have few parameters in comparison. The results of the present study may be informative for the MPT modeling literature to the extent that MPT models functionally resemble IRTree models.

Previous Research and Use.

Often, researchers have used IRT models to construct and investigate items in psychological tests. Researchers have applied IRTree models to a wide variety of observed phenomena. Several have used respondents' response times to survey items to differentiate between types of cognitive processes used when responding. Böckenholt (2012) used IRTree models to distinguish between System 1 and System 2 thinking when responding to items on a cognitive reflection test, demonstrating their ability to differentiate respondents' tendency to use each system and identify items that evoke the use of one system more than the other. Blacksmith et al. (2019) conducted a similar study but found no evidence that System 1 and System 2 usage are distinguishable traits when responding to the cognitive reflection test.

De Boeck and Partchev (2012) used 1PL IRTrees to model fast- and slow-intelligence measured by a cognitive abilities test. They found evidence that fast- and slow-intelligence are distinct cognitive abilities and that separate response processes are used when responding quickly versus responding slowly to the test items. However, the fast and slow abilities were highly correlated. DiTrapani et al. (2016) found no evidence of different intelligence resources when fitting 2PL IRTrees to the same test but did find higher discrimination parameters for the fast-intelligence response process. This suggests that the fast-intelligence process may contain more information about a respondent's intelligence than the slow-intelligence response process.

Many researchers have applied IRTrees to the study of processes and abilities involved in responding to rating scales on psychological tests. Some researchers have used IRTrees for separating response styles from the primary latent trait being measured. Plieninger and Meiser

(2014) found that an IRTree model fit response data to assessments of self-confidence better than a unidimensional model that did not account for response style. Zettler et al. (2016) used IRTrees to investigate the HEXACO model of personality assessed by self- and observer-reports. They found that the directional estimates of each facet correlated stronger between self- and observer-reports after accounting for response style and that these correlations differed substantially between facets.

A common assumption made by researchers who utilize IRTree models to evaluate extreme response styles is that extreme response styles operate as a single latent trait in both directions of a rating scale. Put another way, researchers have commonly assumed that a respondent's extreme response style influences responding in a negative direction to the same extent as it influences responding in a positive direction. Jeon and De Boeck (2019) tested this assumption finding evidence that directional extreme response style invariance does hold, such that extreme response styles do not have different effects on responding depending on which direction a respondent chooses. This could be interpreted as evidence that extreme response style exists as a single latent trait rather than two separate (negative/positive) traits. Böckenholt and Meiser (2017) compared both mixed Rasch models and IRTree models, finding that both adequately separate substantive trait estimates from response styles. They do suggest that IRTrees offer more theoretical purchase on the processes behind response styles compared to mixed Rasch models. However, mixture models are able to separate individuals that use response styles from those that do not when responding, something IRTree models cannot do.

A related issue with self-report survey data is the presence of incomplete and missing responses. Our inferences often rely on the assumption that data are at least missing at random, such that substantive criteria under study are not related to any patterns of missingness (Debeer et al., 2017). Jeon and De Boeck (2016) investigated missing data in a survey of perceptions of trust in charity organizations. Using an IRTree model, they incorporated response omission as a primary latent trait that determined responses to the items. They found that omitting responses was moderately and positively correlated with strong negative attitudes towards charity organizations and modestly and negatively correlated with strong positive attitudes. Their model suggested that response omission can be considered a response style itself that is related to extreme response styles. This suggests that the assumption that the data were missing at random was not supported, thus requiring one to account for missing data patterns in their model. Conveniently, IRTrees already do this. Debeer et al. (2017) built on this, demonstrating that ignoring missing not at random data can cause biased estimates with moderate amounts of missing data. They demonstrated the adequacy of using IRTrees to incorporate response omission into the modeling framework and recover the underlying parameters.

Past Simulation Research

When a new measurement model in IRT is introduced, it can be studied using Monte Carlo simulation to investigate parameter recovery properties. Simulation studies involve generating some data using a statistical model with a set of parameters, adding some random variation, and running a proposed model on the generated data to inspect how well the procedure can recover the parameters that generated the data and the validity of inferences often made with such models (Harwell et al., 1996). Some researchers have investigated IRTree models with simulation studies. Jeon and De Boeck (2016) provide supplementary material for a simulation demonstrating parameter recovery for a two-node 2PL IRTree model with 24 items, 317 respondents, and three response options across 100 replications. They provide a graph of estimated item parameter bias and standard errors across the 97 estimated parameters and note that bias ($M = 0.03$) and mean squared error ($MSE = 0.04, \sqrt{MSE} = 0.20$) were relatively small. Debeer et al. (2017) examined the utility of IRTree models for accounting for missing responses and testing for MNAR and MAR missingness. Parameter estimate bias was diminished when processes for not reaching and skipping items were included in the response process compared to simply omitting observations.

Plieninger (2017) studied the impact of extreme and agreement response styles on validity and reliability of test data. Plieninger found that bias was substantial when the underlying latent trait was correlated with the response style. Tijmstra et al. (2018) used a Bayesian approach to simulate mixture models to distinguish respondents that respond with a graded partial credit response model from those that respond with an IRTree response model. They found that item difficulty/threshold parameters for the IRTree class were adequately recoverable except when sample sizes were small or IRTree class membership was only 25% of the sample (i.e., $n = 250$). Person ability parameters were also severely biased when IRTree class membership was only 25%. Jin et al. (2019) performed a simulation study on detection of differential item functioning when data were generated from IRTree models. They simulated several conditions in which DIF occurred in one or multiple decision processes, the proportion of items exhibiting DIF, and the method of detection. Of note, true positive detections of DIF were severely diminished for processes further along the response tree because they contained missing data—and therefore less information—due to the recoding scheme.

Jeon and De Boeck (2019) provide results to a limited simulation study as supplemental material to investigate whether modeling positive and negative extreme responding with separate latent traits induces confounding when estimating item and person parameters. Mean item-parameter normalized bias for each node ranged between -0.04 and 0.17, while latent correlation normalized bias between each pair of latent factors ranged between -1.92 and 0.88. They did not report parameter estimate variability such as RMSE. DiTrapani (2019) conducted a series of simulations demonstrating the validity of a new fit index (RORME) based on out-of-sample RMSE for comparing different IRTree models and non-IRTree models. DiTrapani found that the index performed adequately, but he did not thoroughly investigate parameter recovery. DiTrapani (2019) performed a second simulation study comparing IRTree models to graded response models, finding that IRTree models recover latent trait estimates adequately even when the data were generated with a graded response model.

Leventhal (2019) found that IRTree models adequately recover parameters under different conditions, manipulating test length, number of respondents, and response options in a Bayesian IRT framework. Leventhal's focus was comparing IRTree models to multidimensional nominal response models and modified partial credit models. The criterion was mean item mean squared error (IMSE), which is a measure of average discrepancy between observed and expected test scores. Parameter recovery was not explicitly investigated. Leventhal also used a two-decision IRTree model for the 6-point scale data, similar to Thissen-Roe and Thissen (2013), which assumes that the initial decision is to agree versus disagree, and the second decision is modeled as a graded response decision between three levels of a directed response. Huang (2020) conducted a simulation study demonstrating the validity of mixture IRTree models to differentiate between normal and aberrant respondents. They found that test length decreased person parameter bias and RMSE for the latent variance-covariance matrix. RMSE for all parameters decreased as sample size increased. Cho et al. (2020) demonstrated the validity of using dynamic IRTree models to investigate eye-tracking time-series data. Using a two-node model, they found that a correctly specified model displayed very little parameter bias and

RMSE for both fixed and random effects. They also noted that the model standard errors provided adequate coverage for the fixed effects. Of note, the RMSE values for both fixed and random effects were substantially (2 to 5 times) larger for the second node than the first node.

In sum, the results from previous studies provide some evidence that IRTree models perform adequately when correctly specified. However, many of these studies investigated particular applications of IRTree models (e.g., response missingness, eye-tracking, mixture-modeling) rather than simple parameter recovery for their typical usage. Secondly, those that did perform a simulation study with typical IRTree models did not thoroughly report the results for parameter bias, parameter estimate variability, or item characteristic curve recovery. Thus, one purpose of my study will be to conduct a more thorough investigation of parameter recovery for IRTree models. I am interested in the effects of sample size, test length, and number of item parameters on parameter recovery.

RQ1: How do sample size, test length, and number of item parameters, and their interactions, affect parameter estimation in IRTree models?

Another issue that has yet to be addressed in previous literature is the potential limitation of IRTrees due to the conditional nature of the responses that can result in missingness. IRTree models often require coding or recoding observed responses to items. This method is not found in other IRT models and should be given special attention. The coding procedure involved in creating auxiliary-items for IRTree models can introduce missing data due to terminal responses occurring prior to other nodes or divergent branches. For example, in the MPP model, the first decision is whether to provide a directed response or a neutral response. If respondent i provides a neutral response (i.e., endorses a three) to item j, the decision process ends, and the auxiliary-items for the agreement and extreme response processes are coded as missing data. Thus, the proportion of missing data in the agreement and extreme auxiliary-items is determined by the proportion of midpoint endorsements. If an item with a rating scale discourages providing a directed response, possibly due to sensitive or socially obligating item content, we would expect to see an overabundance of neutral or indifferent responses and endorsements of the midpoint. This means there is little information about the construct of interest and about respondents' tendencies to endorse extreme responses. This generalizes beyond recoding rating scales to survey designs that elicit sequential decisions. Nodes that are conditional on the responses to ancestor nodes will potentially exhibit greater amounts of missingness. The greater the number of ancestors that a node is conditioned on in a path, the greater the amount of missingness that node could result in. In other words, the greater the depth of a node in a branch of conditional responses, the greater the potential missingness. This would likely compound the effects of low sample size or item parameters that reduce the number of observations available to child nodes. Thus, the second purpose of my study will be to investigate the effect of this conditional dependency on item parameter recovery.

RQ2: How does node depth affect item parameter recovery?

Present Study

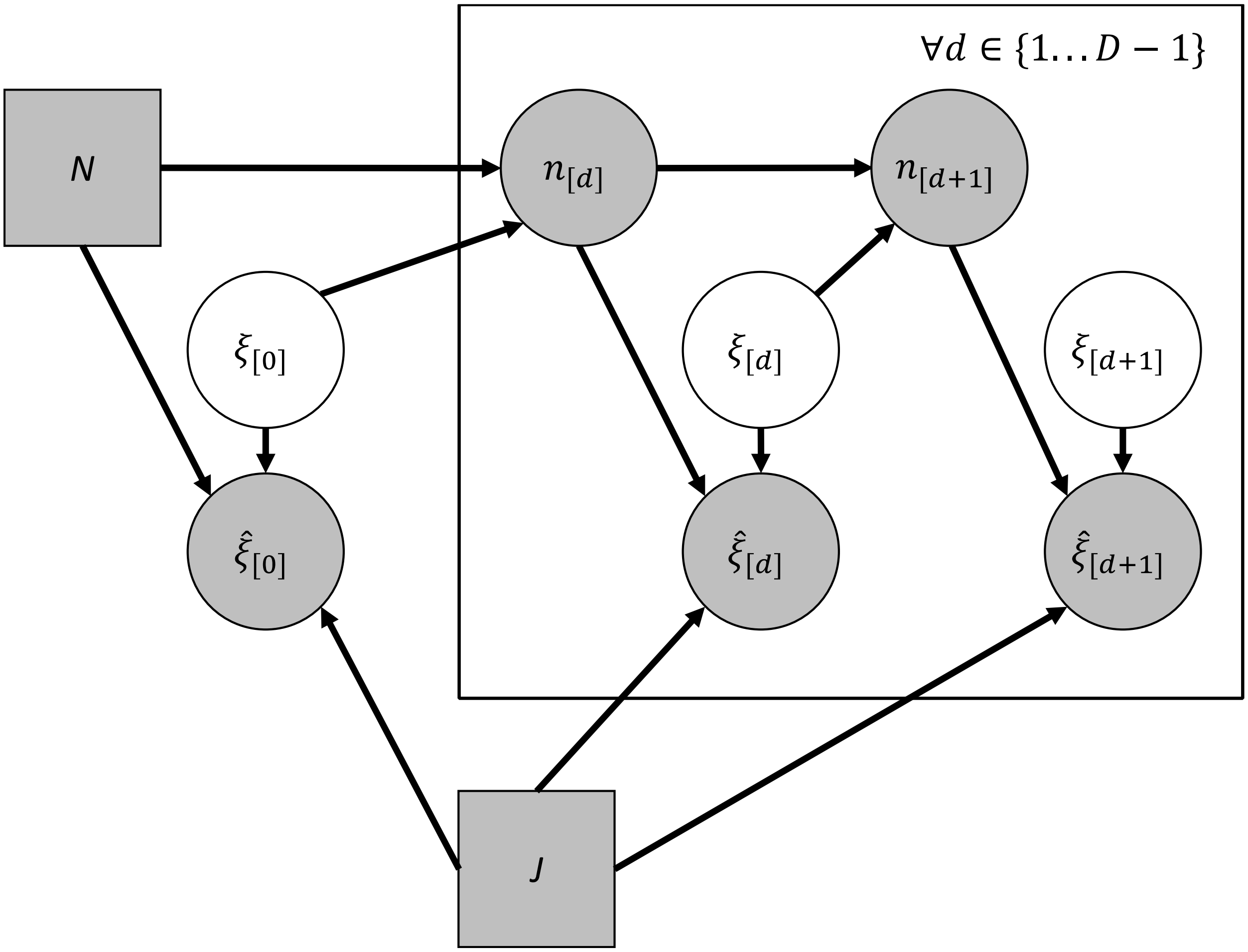
Given the limitations of past simulation research involving IRTree models, there is a need for a thorough investigation of parameter recovery in IRTree models. The present study will help to quantify the validity and reliability of parameter estimates from IRTree models. Validity refers to bias or the discrepancy between the estimated parameter and the true parameter used to generate the simulated data. Invalid parameter estimates can affect recovery of the item characteristic curve which is used for multiple practical inferences such as ability estimation, investigation of item bias, test equating, and computer adaptive testing (Thissen & Wainer, 1982). By parameter estimate reliability, I refer to the variability of the parameter estimate, which provides a measure of uncertainty about the point estimate of a given parameter. High variability in the estimate induces greater uncertainty in the validity of the estimate for a given sample or measurement. Plainly, even if an estimation method is valid with regard to its expected value, low reliability can render the practical use of the method doubtful.

Parameter bias and variability have been thoroughly investigated in past simulation studies with other IRT models (Harwell et al., 1996). The most common manipulations in these studies are test length, sample size, and number of item parameters. The common finding is that both sample size and test length have small positive effects on estimated validity and reliability, and their interaction can have moderate to strong effects on validity and reliability (e.g., Drasgow, 1989; Hulin et al., 1982). Models using small sample sizes with long tests are likely to lead to poor item parameter estimates. The number of item parameters to estimate for each item has a negative effect on estimate validity and reliability and interacts in a similar way with sample size and test length such that longer tests with more parameters require larger samples for valid inference (Hulin et al., 1982). For clarity, I do not have sufficient reason to predict whether the IRTree model will be upwardly or downwardly biased, so I refer to bias in an absolute sense. There is little reason to suspect that IRTrees will behave differently than other IRT models, as they are based on the same functional form as those models investigated in previous studies. However, it is likely that the conditional nature of nodes at greater depths in a tree will make them more vulnerable to estimate inaccuracy and unreliability with smaller sample sizes, shorter test lengths, and more item parameters to estimate. For the present study, this implies interaction effects between these three factors and the depth of a given node. The fact that nodes are conditioned on one another implies that the characteristics of the conditioning ancestor nodes have some causal effect on their descendants. I will explain this effect further and how I plan to measure it.

Hypothesized Causal Model, Node Depth, and Propagation

Figure 2 displays the hypothesized causal model for item parameter estimation in plate notation, where *N* represents total sample size, *J* represents test length, *D* represents maximum node depth, *n* represents descendant node-specific sample size, ξ represents true item parameters, and $\hat{\xi}$ represents estimated item parameters. Squares indicate deterministic variables, circles represent stochastic variables, grey shapes represent observed variables, and white shapes represent unobserved or latent variables in a typical measurement setting. Sample size, test length, and the true data generating parameter values should have some direct effect on estimation of the root-node item parameters, $\hat{\xi}_{[0]}$. After the root-node, node depth plays an integral role in the causal model. I will use node depth as a predictor in the analyses below, but node depth is not represented as its own variable in the causal model. Time is often used as a predictor in many regression models, but time itself does not exert causal effects and simply serves as a proxy for unmeasured causal interactions and change. Much like how time in other models is not a true causal variable, node depth is not a properly causal variable in the sense that it merely indexes position and sets a unit for distance and conditionality in the graph. Node depth is represented by the box (plate) outlining the descendant node sample size and item parameters and expressing the causal paths between these variables across node depths 1 ... d ... D. The causal effect across nodes is due to the missingness incorporated into the data resulting in lower sample sizes at greater depths. Node sample size *n* is a fully mediating variable across nodes, assuming uncorrelated latent variables for simplicity. Incorporating sample size at each node as a predictor in the analyses below may be interesting and useful in some regard for researchers prior to analyzing their data. If the study results can provide some suggestion about minimum sample sizes required for each node, a researcher could inspect their data prior to analysis and determine whether they should proceed with a particular model. This is less helpful in the planning stages of a study because researchers do not yet have data to inspect. Node depth is a characteristic of the specific IRTree model a researcher chooses. A description of how node depth affects parameter estimation is useful to researchers for making decisions about what IRTree model to use prior to data collection and analysis.

Figure 2. *Hypothesized Causal Model for Item Parameter Bias*.



Note. N = total sample size, J = test length, D = node depth, n = sample size of descendant node, ξ = true item parameters, $\hat{\xi}$ = estimated item parameters. Squares indicate deterministic variables, circles indicate stochastic variables, white shapes indicate unobserved variables, and grey shapes indicate observed variables. The box (plate) is a condensed representation of node depth.

The conditional nature of IRTree items means that fewer observations are available for estimating parameters for items occurring deeper in a branch. I will refer to this process of ancestor nodes “passing on” observations to descendant nodes as *propagation*. Propagation describes not only how a parent-node affects or conditions a child-node, but also how properties of ancestor nodes can indirectly affect or condition descendant nodes at greater depths. Put another way, propagation entails cumulative effects across depths. This propagation mechanism is what makes IRTrees unique from other IRT models and requires special consideration regarding its effects on parameter estimation.

Propagation is a function of the total sample size, node depth, and item parameters determining endorsement probability. The number of observations propagated $n_{[d]}$ to a descendant node at depth d is conditional on the total or root-node sample size N and on endorsement or non-endorsement of $d-1$ ancestor items each with endorsement probability $p_{[d]}$. Endorsement probability of ancestor items is conditional on the parameters of the items and respondents. For the 1PL model, the item $\beta_{[d]}$ and person parameters $\theta_{[d]}$ of an ancestor-node at depth d affects its endorsement probability, $p_{[d]} = \Psi(\theta_{[d]} - \beta_{[d]})$. Marginalizing over the person ability parameter, this in turn affects the expected proportion of respondents, $E\left[p_{[d]}\right]$, that will endorse the item. Again, the discrepancy between the person ability and item difficulty parameters determines the endorsement probability, so the location of the item difficulty parameter determines the expected probability of endorsing the item. The inclusion of an item discrimination parameter for the 2PL model increases or decreases the influence the item difficulty has on the expected endorsement probability. The expected number of respondents that reach the second node in a branch is determined by the expected probability of endorsement for the first node times the number of respondents at the first node (i.e. the total sample size), or

$E\big[n_{[d=2]}\big] = \ n_{[d=1]}E\big[p_{[d=1]}\big] = N \cdot E\big[p_{[d=1]}\big]$. The expected number of respondents a child-node d is, $E\big[n_{[d]}\big] = E\big[n_{[d-1]}\big]E\big[p_{[d-1]}\big]$. The expected number of propagated observations from the total sample size to a given descendant node D can be written as the root-node sample size N times the expected propagation rate, which is the product of expected probabilities for all ancestor nodes,

$$E\big[n_{[D]}\big] = N \ \prod_{d=1}^{D-1} E[p_{[d]}].$$

In most studies, researchers do not have strong prior knowledge on either item parameters or item marginal probabilities. This makes choosing sample sizes for data collection more difficult. It would be useful for a researcher to know the average amount of estimate bias they can expect regardless of the item parameters. We can break down the expected propagation proportion into average propagation for a given node, p_μ, times the deviation from this average, $p^*_{[D]}$ which is the expected propagation proportion divided by the average propagation. I will assume that the average proportion of observations propagated to node D is the probability of an average person endorsing an average item raised to the power of $D-1$, or $p_\mu^{D-1} = \Psi(0-0)^{D-1} = 0.5^{D-1}$. For example, the average proportion of observations propagated to the first node is obviously $0.5^{1-1} = 1$, the second node is $0.5^{2-1} = 0.5$, the third is $0.5^{3-1} = 0.25$, and so on. The expected deviation from the average propagation rate is the ratio of expected propagation to average propagation, $p^*_{[D]} = \frac{\prod_{d=1}^{D-1} E[p_{[d]}]}{p_\mu^{D-1}}$, which I will call the *relative propagation rate*. Relative propagation denotes whether and to what degree a given item is propagated greater or fewer observations from their ancestors compared to a set of average ancestor items. When an ancestor item is less likely to be endorsed, perhaps because of difficult item content, the relative propagation rates of their children decrease. When an ancestor item is more likely to be endorsed, this rate increases. Rearranging and substituting the equations above, $\prod_{d=1}^{D-1} E[p_{[d]}] = p_{\mu}^{D-1} \frac{\prod_{d=1}^{D-1} E[p_{[d]}]}{p_{\mu}^{D-1}} = p_{\mu}^{D-1} p_{[D]}^{*}$, we get, $E\left[n_{[D]}\right] = N\, p_{\mu}^{D-1} p_{[D]}^{*}$, or the expected number of propagated observations to node D is the product of the average and relative propagation rates of the immediate parent node and the total sample size.

This is analogous to using a non-centered parameterization in hierarchical modeling where the latent means and variances are modeled as separate parameters. Separating total node propagation into average propagation and relative propagation allows me to quantify the effects of node depth and item endorsement probability separately. Node depth can now be used as a meaningful predictor whereby it serves as a proximal measure for the cumulative average propagation rate at a given node depth. Relative propagation then captures the cumulative effects imposed on descendant nodes by their ancestors due to their item parameters beyond the cumulative effects of node depth.

In sum, node depth measures the average rate of propagation of observations across nodes. The deeper the node, the fewer the propagated observations which should lead to greater estimate variability. Relative propagation entails deviations from the average propagation rate due to item-specific characteristics. Items that have fewer endorsements will propagate fewer observations than average and result in greater estimate variability.

Sample Size

Adequate sample sizes are required in order to sufficiently estimate the parameters of interest in a model. Smaller sample sizes often entail less information available to estimate model parameters accurately and reliably. I manipulated sample size in the present study. As with previous research investigating the effects of sample size on item parameter recovery, I expected no significant effect of sample size on estimate bias and a negative effect on parameter variability. Sample size should also negatively moderate the effects of node depth, relative propagation, and the interaction between them, such that their effects weaken with larger sample sizes.

Test length

Previous research suggests that shorter tests increase test bias and sampling variability (Lord, 1968; McCauley & Mendoza, 1985; Stone, 1992). This is likely because shorter tests contain less information about the latent trait of interest. Shorter tests also result in fewer possible response patterns and therefore fewer latent ability scores to differentiate respondents. Recommendations for adequate test lengths vary and often depend on the sample size and model complexity. For example, Harwell and Janosky (1991) suggest that 15 items and a sample of 250 respondents for a unidimensional 2PL model provides satisfactory parameter estimates and standard errors. On the other hand, Drasgow (1989) suggests that 5 items with 200 respondents provides reasonable parameter estimates and standard errors for a 2PL model. The effects of test length on estimation are better understood relative to the size of the sample. Test length should have a small negative main effect on parameter bias and variability. At large sample sizes, differences in parameter estimate validity and reliability due to different test lengths will likely be negligible. Therefore, the effect of test length on estimate bias and variability should be negatively moderated by sample size.

Model Type

A one-parameter model implies that items are characterized by one item parameter, namely the item difficulty. Two-parameter models imply that the item is characterized by two parameters, a difficulty parameter and a discrimination parameter. The addition of another parameter to estimate often requires greater sample sizes (Hulin et al., 1982). The above discussion should apply to both one- and two-parameter models. The main difference between the two are that parameter estimates in the two-parameter model should be more severely impacted by sample size and node-depth than the one-parameter model. In order to avoid making the analysis and interpretation of the results overly cumbersome, I will conduct separate simulations and analyses for the two types of models. Statistical tests of differences between the types of models, and differences in their interaction with the other factors, may be helpful for some looking for advice on which type of model to use. I argue that such information can be had without directly comparing the two with statistical tests. I also argue that decisions about which model to choose should involve theoretical considerations rather than simply sample size and model fit (Andrich, 2004). I will instead discuss descriptive comparisons between the two sets of analyses.

Method

In order to investigate the validity of IRTree models, I conducted a Monte Carlo simulation study. Harwell et al. (1996) provide an overview of IRT simulation studies and provides recommendations for conducting them. I generated data with “true” underlying parameters and used IRTree models to estimate the underlying parameters. I used the discrepancies between the true and estimated underlying parameters as criteria for assessing the quality of the model. To understand the limitations and factors that can affect IRTree validity and reliability, I included multiple manipulations that affect the generated data.

Design

I manipulated sample size (500, 2,000), test length (10 items, 30 items), and the type of model (1PL versus 2PL), resulting in $2 \times 2 \times 2 = 8$ conditions. Each condition was replicated a total of 100 times. Within each replication, I simulated multiple items and multiple respondent abilities. The criteria within a set of items were potentially correlated because they were simulated within the same replication. The analysis thus modeled the intercepts as randomly varying between replications nested within a given condition.

Sample Size

The simulation studies reviewed in Harwell et al. (1996) simulated samples of respondents that ranged from 100 or fewer to 1000 or greater. The median sample size simulated was between 300 and 500 respondents. For the present study, the two levels of the sample size factor were 500 and 2,000 representing small and large sample sizes, respectively.

Test Length

Again, referring to the Harwell et al. (1996) review, the test lengths researchers used in previous studies ranged from 5 items to 60 items, the mode of which was 25 items. The test length factor consisted of two levels representing short (10 items) and long test lengths (30 items). A thirty-item test was chosen for the long test length because estimation time becomes prohibitive for the present simulation study with an increasing number of items.

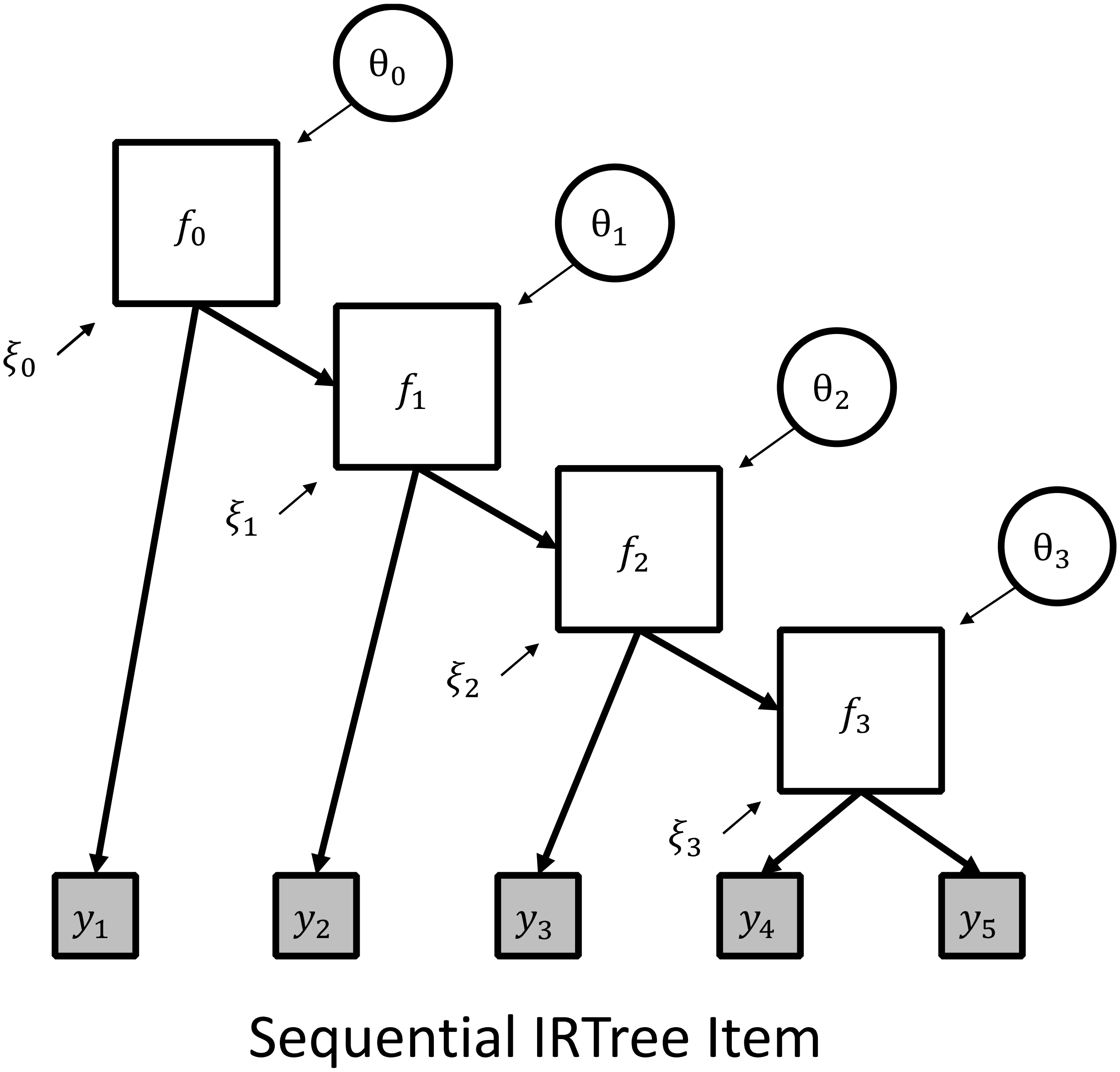
Data Generation Procedure

Previous simulation studies often relied on simulating a single large set of data and using subsets of this data set to study parameter recovery. A consequence of this method is that the different experimental conditions are not independent of each other, confounding measurement of the manipulations of interest and making analysis and inference more difficult. Another issue is that these researchers often used a fixed set of item parameters and only generated random person parameters and item responses. Validity can be established with this method, but the generalizability of their results to the often-random distributions of items parameters found in practice is questionable (Harwell, 1997). A design that allows the random generation of item parameters should increase the generalizability of the simulation study results. For the present study, I generated item and person parameters from specified distributions. I generated data in the R programming environment (R Core Team, 2019). Pseudo-code for the simulation procedure is provided below.

Sequential IRTree Model

The most common models in the IRTree literature are the two-node binary models with four response options and the three-node binary models with five response options. The present study, however, focused on the question of node depth and observation propagation as possible concerns for parameter estimation. To systematically quantify these effects, a model that allows greater control over node depth, the number of nodes in total, the item parameters, and the latent factor variances and covariances was required. Figure 3 depicts the type of IRTree model that I used. The coding matrix specified was a four-decision response process. Failure on the first decision would result in an observed response of 1. Success on the first decision would lead to a second decision, wherein failure would result in an observed response of 2. The response process continues for the third and fourth nodes, resulting in a total of 5 possible observed outcomes. The general IRTree model adopted was similar in form to the sequential model (Tutz, 1990) and resilience model (DiTrapani, 2019). The sequential model involves a single latent factor used at each node in the response process, whereas the resilience model involved two latent factors measured by different nodes and allowed them to covary. The relationship between nodes at different stages in the response process through the latent factor parameters allows them to covary beyond their conditional dependency. Although I do not believe the threat to be severe, this is likely to confound estimates of parameter bias and variation due to the conditional dependency. To avoid this in the simulation, each node measured its own latent factor, and the latent factors were orthogonal to one another. Another question for the present study was whether there are substantial differences in parameter estimation between 1PL and 2PL models. Each condition was simulated with both 1PL and 2PL IRTree models.

Figure 3. *Sequential IRTree Model with 4 Nodes*



Item Discrimination Parameters

Researchers have previously used log-normal distributions to simulate item discrimination parameters for IRTree models. Tijmstra et al. (2018) drew discrimination parameters from a log-normal distribution with log-means of -0.50 and log-standard deviations of 0.25. DiTrapani (2019) generated correlated discrimination parameters between positive and negative extreme response processes. First, DiTrapani generated parameters for the negative extreme response node from a uniform distribution with a range of 0.4 and 1. Then, parameters for the positive extreme response node were generated from a normal distribution with means equal to the generated negative extreme response node parameters and a standard deviation of 0.20. In this study, I generated the item discrimination parameters, a_j, for each node from a log-normal distribution with log-mean and log-standard deviation of 0 and 1, respectively. The resultant sampling distribution should have 95% density interval roughly bounded between 0.03 and 5.18 with a median at 1, ensuring that extremely high or extremely low discrimination parameters will not likely be generated. I fixed the discrimination parameters for the 1PL models to one.

Item Difficulty Parameters

Researchers have used several methods for generating difficulty parameters for IRTree models. Debeer et al. (2017) simulated item difficulty parameters for 20 items by selecting 20 values at evenly spaced intervals between -1.50 and 1.00. Jin et al. (2019) generated item difficulty parameters from a uniform distribution ranging between -1.50 and 1.50. Tijmstra et al. (2018) simulated parameters from a normal distribution with a mean of 2.00 and a standard deviation of 0.25. DiTrapani (2019) generated correlated difficulty parameters between the substantive trait and the positive and negative extreme response processes. DiTrapani first generated parameters for the substantive trait in the first node from a normal distribution with a mean and standard deviation equal to 0.00 and 1.00. Then, parameters for the extreme response nodes were generated with means equal to the parameters generated in the first node and standard deviations of 1.00. For the present study, I generated difficulty parameters from a normal distribution with a mean of zero, and standard deviation of one. Ninety-five percent of the resulting distribution should range between roughly -1.96 and 1.96.

Latent Factor Correlations, Variances, Means, and Scores

The latent factors that are measured with multidimensional item response models are often correlated. Tijmstra et al. (2018) drew latent factor scores from independent univariate normal distributions. Debeer et al. (2017), Jin et al. (2019), DiTrapani (2019), and Huang (2020) simulated latent factor scores from multivariate normal distributions with zero means and pre-specified covariance matrices. Setting the latent factors to be uncorrelated is an unreasonable assumption with regard to data one is likely to encounter outside of a simulation. However, the primary purpose of the study is to explore and quantify the effect of the conditional coding involved in IRTrees on item parameter estimate bias. Correlated latent factors will introduce "backdoor" relationships between nodes beyond the relationship implied by the conditional response process. This may confound the effect that one parent node has on a child node through their correlation between their respective latent factors. Therefore, I used an orthogonal set of latent factors to generate the ability parameters and ensure that the only relationship between items of different nodes was through their conditional response dependencies. I generated the latent factor scores themselves from a univariate normal distribution with means equal to zero and variances equal to one.

Response Generation

Once the parameters were generated for a given replication in a particular cell, I used the formulas given by Equation (1) to generate response probabilities for each rating scale option utilizing the coding matrix in Table 1. I then randomly sampled response options from a 5-point rating scale using the response probabilities. Finally, I coded these rating scale responses according to the coding matrix for analysis. I then saved the data generating parameters for comparison to model estimates. In IRT, if an item does not contain more than one response, item parameters cannot be estimated for that item and the item is typically dropped from analysis. Prior to parameter estimation, I assessed each simulated data set to ensure each auxiliary-item had more than one response option endorsed. If a data set had one or more auxiliary items that had only one response option endorsed, another data set was generated in its place.

Estimation of simulation models

I used the mirt package (Chalmers, 2012) in R to estimate each IRTree model. For each simulated data set, I estimated an IRTree model according to the coding matrix and experimental condition (1PL or 2PL). I estimated models within a Bayesian framework. I specified priors for the difficulty parameters with a normal distribution with mean and standard deviation equal to 0 and 1, respectively. I specified priors for the discrimination parameters with a log-normal distribution with a mean and standard deviation equal to 0 and 1, respectively.

Preliminary analyses suggested that estimating the models with the Bock and Aitkin (1981) EM procedure led to consistent non-convergence with several models, an issue DiTrapani (2019) encountered when simulating IRTree models as well. For marginal maximum likelihood, in order to integrate the *k*-dimensional latent factors out of the likelihood equation, the EM algorithm uses quadrature that increases exponentially with more latent factors (Chalmers, 2012). With regard to Chalmer’s (2012) implementation in mirt, the number of quadrature points per factor is reduced to make estimation more efficient. This can make estimates of the latent factor variance-covariance matrix less accurate with high dimensional models. In the case of IRTree models with four latent factors, the latent factor variance-covariance matrix frequently becomes non-positive definite, particularly when sample sizes and test lengths are low. The degenerate latent factors then cannot be integrated out of the likelihood equation and the maximum likelihood solution becomes unreliable. It is possible to increase the number of quadrature points per latent factor; however, computation time can become unwieldy.

Alternatively, the Metropolis-Hastings Robbins-Monro algorithm (MH-RM; Cai, 2010) uses stochastic imputation to estimate the latent factor variance-covariance matrix, often resulting in faster and more accurate estimation for high-dimensional models. Preliminary analyses using this algorithm had fewer issues with convergence and was adopted instead. Models with many items tend to take a large number of MH-RM cycles to converge. In order to reduce simulation time, I specified a maximum of 10,000 MH-RM cycles for estimation. Models that reached this cycle maximum were discarded and another dataset was generated in its place. The expected a posteriori estimates of the latent factor scores were then calculated and the model parameters were extracted when each model completed.

Replications

Multiple replications are needed to adequately account for sampling variability from the data generation process for the overall simulation. The number of replications used in previous studies of Monte Carlo simulations of IRT models ranges between 5 and 1000 (Harwell et al., 1996). For the present study, I replicated each condition 100 times. Although a greater number per condition would provide larger sample sizes for the proposed analyses, the time required for simulation would be too great relative to the reduction in the credibility intervals of the regression model parameters. Preliminary analyses suggested that the most computationally intensive models required between 4 to 5 minutes of estimation time. This would roughly take 6.5 to 8.5 hours to complete 100 replications for a single condition. The precision gained from more replications would likely be negligible relative to the increased simulation time.

Simulation Data Generating Pseudo-code.

```
for $r$ in $1 \dots R$ do:
    for $j$ in $1 \dots J$ do:
        for $k$ in $1 \dots K$ do:
            $\beta_{[j,k]} \sim Normal(0,1)$
            if model = 1PL:
                $\alpha_{[j,k]} = 1$
            else:
                $\alpha_{[j,k]} \sim LogNormal(0,1)$
        end
    end
    for $i$ in $1 \dots N$ do:
        for $k$ in $1 \dots K$ do:
            $\theta_{[i,k]} \sim Normal(0,1)$
            for $j$ in $1 \dots J$ do:
                $z_{[i,j,k]} = \alpha_{[j,k]}\left(\theta_{[i,k]} - \beta_{[j,k]}\right)$
            end
        end
    end
    for $i$ in $1 \dots N$ do:
        for $j$ in $1 \dots J$ do:
            for $m$ in $1 \dots M$ do:
                $p_{[i,j,m]} = \prod_{k=1}^{k=K}\left(\frac{e^{z_{[i,j,k]} \times D_{[m,k]}}}{1+e^{z_{[i,j,k]}}}\right)^{T_{[m,k]}}$
            end
            $y_{[i,j]} \sim Multinomial\left(p_{[i,j,\dots]}\right)$
        end
    end
    $\left\{\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\theta}}\right\}_{[r]} = f_{MH-RM}(\boldsymbol{y})$
end
```

Parameter Estimate Standard Errors

I collected item parameter standard errors for the difficulty parameters. The standard errors represent uncertainty in the model parameter estimates. The standard errors can be used to construct confidence intervals around the parameter estimates. These are informative particularly for item selection and item bias investigations. The 2PL models did not produce standard errors due to estimation issues. I will discuss this further in the results and discussion sections.

Two criteria were explored regarding the standard errors. The first was whether the calculated standard errors provided adequate coverage for the true underlying parameter values such that traditional 95% confidence intervals encompassed the true parameter value. This was done by constructing 95% confidence intervals for each of the item parameter estimates and their associated standard errors. I created a dummy variable that indicated whether the true underlying parameter value was encompassed by the estimated 95% confidence intervals constructed around the parameter estimate. Low rates of coverage would suggest that the standard errors underestimate the true sampling variability of the item parameters or that the estimate procedure is highly biased. Differences in coverage rates between the manipulated conditions may indicate that some situations threaten the validity of the standard errors. The second criterion was the standard errors themselves and were be analyzed similarly to the item parameter estimate variability analyses. This may reveal conditions that lead to overly conservative or anti-conservative estimates of the parameter estimate sampling variability.

Analyses of Simulation Results

Given the large sample size and the low complexity of the proposed regression models, descriptive statistics would likely provide adequate evidence to draw broad conclusions such as the direction of the manipulated factor effects. My goal with this study was to provide more precise guidance for sample size, test length, and IRTree design by means of posterior predictive simulations. Posterior predictive simulations involve drawing samples from the posterior of a set of estimated model parameters conditioned on observed data to make predictions about unobserved data.

Many simulation studies in the past aggregated across items within each iteration for each condition. There is little reason to expect a large amount of sampling variability between iterations. However, a random intercepts model would allow me to quantify how much variability in the criteria is attributable to randomness introduced by the data simulation process. A substantial amount of variation between simulations that is not attributable to the manipulated factors would likely indicate some issue with the data generating process and would require a more remedial inspection. Secondly, the ancestor-propagation and node-depth factors are both within-iteration factors. Aggregating the data to the iteration level, as is commonly done with simulation studies, would render those as useless predictors. Instead, I estimated both bias and variability directly with a distributional model that uses predictors for both the location (e.g., mean) and scale (e.g., variance) of the observed bias distributions. I used Stan (Stan Development Team, 2019) for estimation of the explanatory models of bias.

Dependent Variables

Regarding the outcomes being explained, estimate bias for the item difficulty and person ability parameters is straightforward to calculate as the difference between the estimated and true parameters. I simulated true difficulty and ability parameters from a normal distribution. I did not expect any systematic bias to affect the distributional assumptions after accounting for the effects of the predictors, so the estimated parameters should follow a normal distribution. The resultant bias, calculated as the difference between two normally distributed dependent random variables, should also follow a normal distribution, with a mean $\mu^* = \hat{\mu} - \mu$ and standard deviation $\sigma^* = \sqrt{\hat{\sigma}^2 + \sigma^2 - 2\rho\hat{\sigma}\sigma}$. The task is then to estimate two sets of regression parameters from the set of predictors to explain both average bias and variability.

The rank order of person ability parameter estimates is important if a researcher wants to use the estimates as a linear predictor of some criterion. Systematic bias of the estimates will not affect the rank order. Increased sampling variability or estimate uncertainty will affect the rank ordering, but this affect may not be apparent by simply observing changes in the variability of the estimate bias distribution. A more direct approach would be to analyze the correlation between the true and estimated person ability parameters. I included an analysis of the correlation between the true and estimated person ability parameters, which I detail below.

I took a similar approach with the discrimination parameter estimate bias. I drew the true discrimination parameters from a log-normal distribution. Preliminary simulations suggested that the discrimination parameter estimates are also approximately log normal. I took the ratio of the estimated and true discrimination parameters. The ratio of two log-normal distributions is also log-normally distributed with a mean $\mu_\alpha^* = \hat{\mu}_\alpha - \mu_\alpha$ and standard deviation $\sigma_\alpha^* = \sqrt{\hat{\sigma}_\alpha^2 + \sigma_\alpha^2 - 2\rho_\alpha\hat{\sigma}_\alpha\sigma_\alpha}$.

Independent variables

I used four predictors and their two-, three-, and four-way interactions. For each analysis, I standardized the sample size and test length predictors to facilitate estimation. I coded node depth so that nodes one through four were coded as zero through three. To help with interpretation, I calculated the relative propagation predictor as the log of the odds-ratio of expected propagation proportion to average propagation proportion for each node D, $\log p_{[D]}^* = \log \frac{\prod_{d=1}^{D-1} E[p_{[d]}]}{p_\mu^{D-1}}$, where $E\left[p_{[d]}\right] = \sum \Psi\left(\theta_{[q]} - b_{[d]}\right) w_{[q]}$, evaluated using Gauss-Hermite quadrature, and $b_{[d]}$ is the true item difficulty for that simulated item. By taking the log of the relative propagation ratio, items with an expected propagation rate equal to the average propagation rate receive a score of one on the odds scale and a score of zero on the log-odds scale. Conveniently, the node-depth factor then becomes the contribution to the node-depth effect for an average auxiliary-item and the propagation predictor indicates greater or lesser observation propagation compared to an average auxiliary-item.

Interpretation of Results

A common difficulty with simulation studies is how to define adequate parameter recovery, especially without some criterion with an obvious standard driving the research question at hand. With regard to estimate variability for the difficulty parameters, I will consider a limit of $\sigma_\beta = 0.25$. Assuming that difficulty parameters are unbiased and are generated from a normal distribution, a standard deviation of 0.25 suggests that with 95% certainty the true item difficulty, β_j, resides on the logit scale interval $\hat{\beta}_j - 1.96 \times 0.25 < \beta_j < \hat{\beta}_j + 1.96 \times 0.25$. With this level of measurement certainty, a researcher could distinguish two items with a one-logit difference between their true data generating difficulty parameters, assuming the standard errors are close approximations of the true estimate uncertainty. For example, given a set of easy, average, and difficult items with difficulties of $\beta_{easy} = -1,\ \beta_{average} = 0,\ \beta_{hard} = 1$, I would expect respondents from a population with a standard normal ability distribution to produce endorsements at rates of 16%, 50%, and 84%, respectively. For an item selection task, adequate item parameter estimation should be able to distinguish an average item from one with a low or high rate of endorsement. With 95% confidence, $\sigma_\beta = 0.25$ is the maximum standard deviation that would allow the location and rank ordering of these items to be adequately estimated. In practice, the σ_β varies between items and many tasks such as item selection aim to distinguish items with much less than a one-logit difference in difficulty and thus require greater measurement certainty. I used this same criterion for interpreting person ability estimate bias. I also interpreted person ability parameter recovery in terms of a predictive validity task, where the estimated latent ability is used to predict some external criterion. I primarily focus on whether the estimates allow a researcher to make valid directional hypothesis tests regarding the relationship between the latent ability and the external criterion. Regarding the item discrimination parameters, defining a criterion is more difficult. However, due to issues with the simulation, this issue was not addressed.

These criteria provide some principled starting point. I will first report the model results and then provide an interpretation with these criteria in mind. Although I discuss the regression model parameters in the results section in terms of their size and direction, I will not interpret the regression model parameters using significance tests or in reference to null hypotheses. The emphasis of my analysis is to highlight conditions that may produce poor parameter estimates in concrete terms rather than vague notions of significant differences or standardized effect sizes which obfuscate the practical reasons for conducting a simulation study.

I adopted a Bayesian framework for estimation and inference. Parameter bias is less relevant for Bayesian inference. From a Bayesian perspective, posterior point estimates are not likely equal their data generating values except with very large sample sizes. Furthermore, inferences based on point estimates are often less informative than the distribution of plausible values the data generating parameter could have taken on or may realize in the future. I place a greater emphasis on estimate variability or uncertainty, which quantifies the range or distribution of plausible data generating values. I investigate parameter bias for the item and person parameters for the interested reader, but my primary focus is on estimate variability.

.

Results

Regression Model Estimation

The results of the simulations produced very large datasets. I required greater computing resources to create regression models for quantifying the effects of the manipulated factors. I used the supercomputing resources available through the Ohio Supercomputer Center (Ohio Supercomputer Center, 1987) to shorten computation time for the regression analyses. Researchers in the past have calculated summary statistics such as mean bias and root mean square error for each simulation iteration and then conducted separate regression analyses. This requires aggregation of data and separate sets of analyses. I used distributional regression models (Bürkner, 2018; Rigby & Stasinopoulos, 2005) which simultaneously estimate predictors for all dependent variable distribution parameters and allowed me to directly quantify IRTree estimate bias and variability together.

For the item parameter regression models, I used all of the previously mentioned predictors and their interactions. For the node depth factor, I coded nodes 1, 2, 3, and 4 as 0, 1, 2, and 3, respectively. I standardized the sample size and test length predictors to facilitate estimation. I took the log of the relative propagation predictor to center the average expected propagation at zero. By coding the predictors this way, the intercept represents the expected bias or variance for the root-node. The main effects for node depth predictor represent increases or decreases in the outcomes compared to the root-node. The main effects for sample size, test length, and log-relative observation propagation are effects on bias and variance for the root-node. Interactions between these predictors and node depth represent effects on the outcome for deeper nodes. The sample size and test length factors were centered and standardized to facilitate estimation. The predictors of the variance for the recovered parameter estimates are on the log scale. I used the same predictors for the person ability parameters except for the relative propagation factor. I present the regression model results for each analysis, but I will rely on posterior predictions of bias and variance for more intuitive interpretations. Additional tables and figures of the posterior distributions are in the appendix, including parameter density plots, MCMC chain trace plots, and parameter correlation plots.

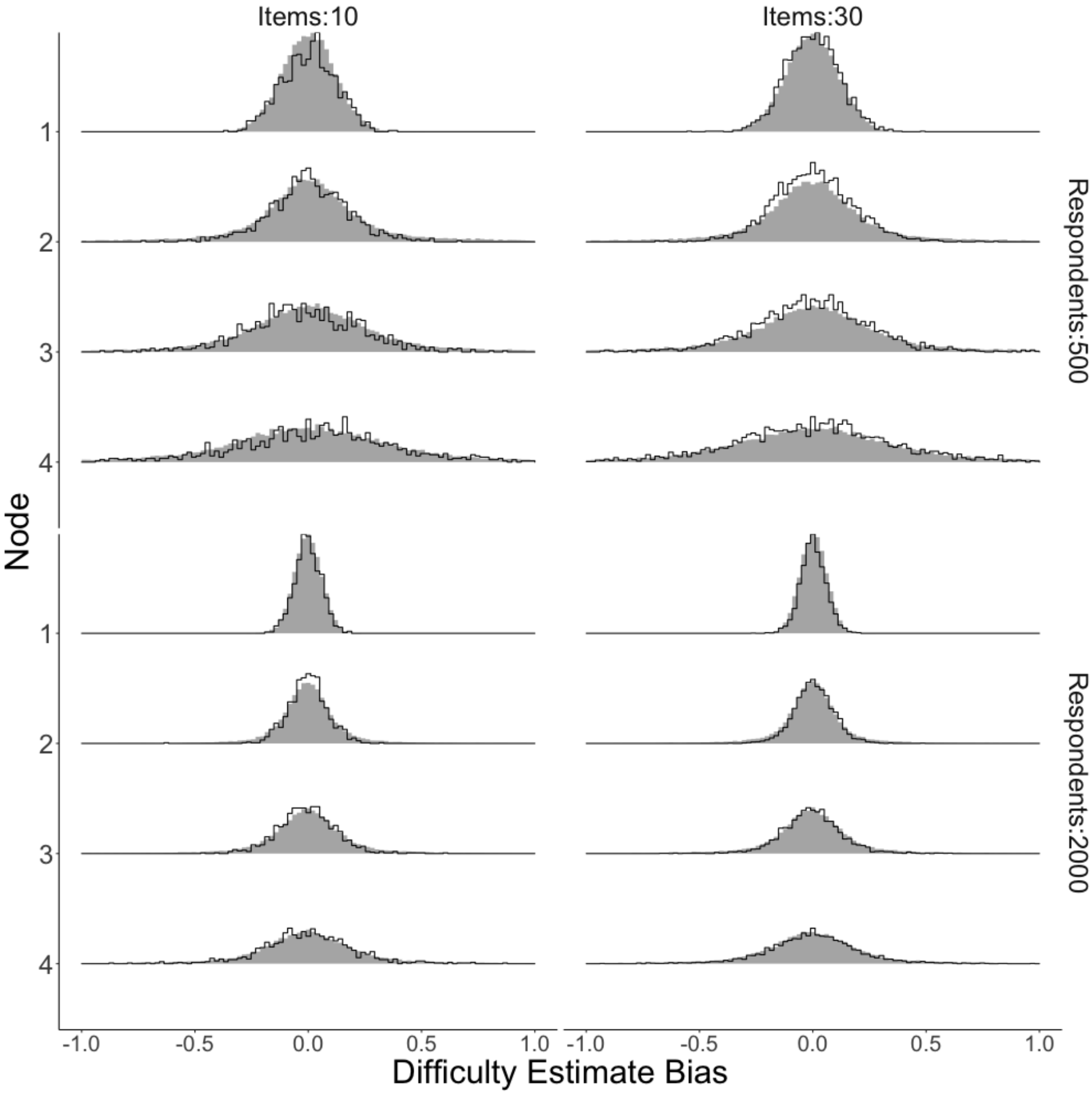
I used the same analytic strategy for 1PL and 2PL simulation item difficulty and person ability parameters. I encountered estimation issues with the 2PL models, which prevented estimation of the standard errors. The discrimination parameters were also poorly estimated, and I will detail this further in the section on discrimination parameter bias. Despite these issues, the results for two types of models were very similar for the item difficulty and person ability estimates. Below, I will present the results for both models grouped by parameter type. I will discuss the results in general terms that apply to both models for the sake of clarity and brevity instead of restating the same conclusions for each model separately. However, I will note when the results differ between the two models.

Item Difficulty Parameter Bias

Descriptive Analysis

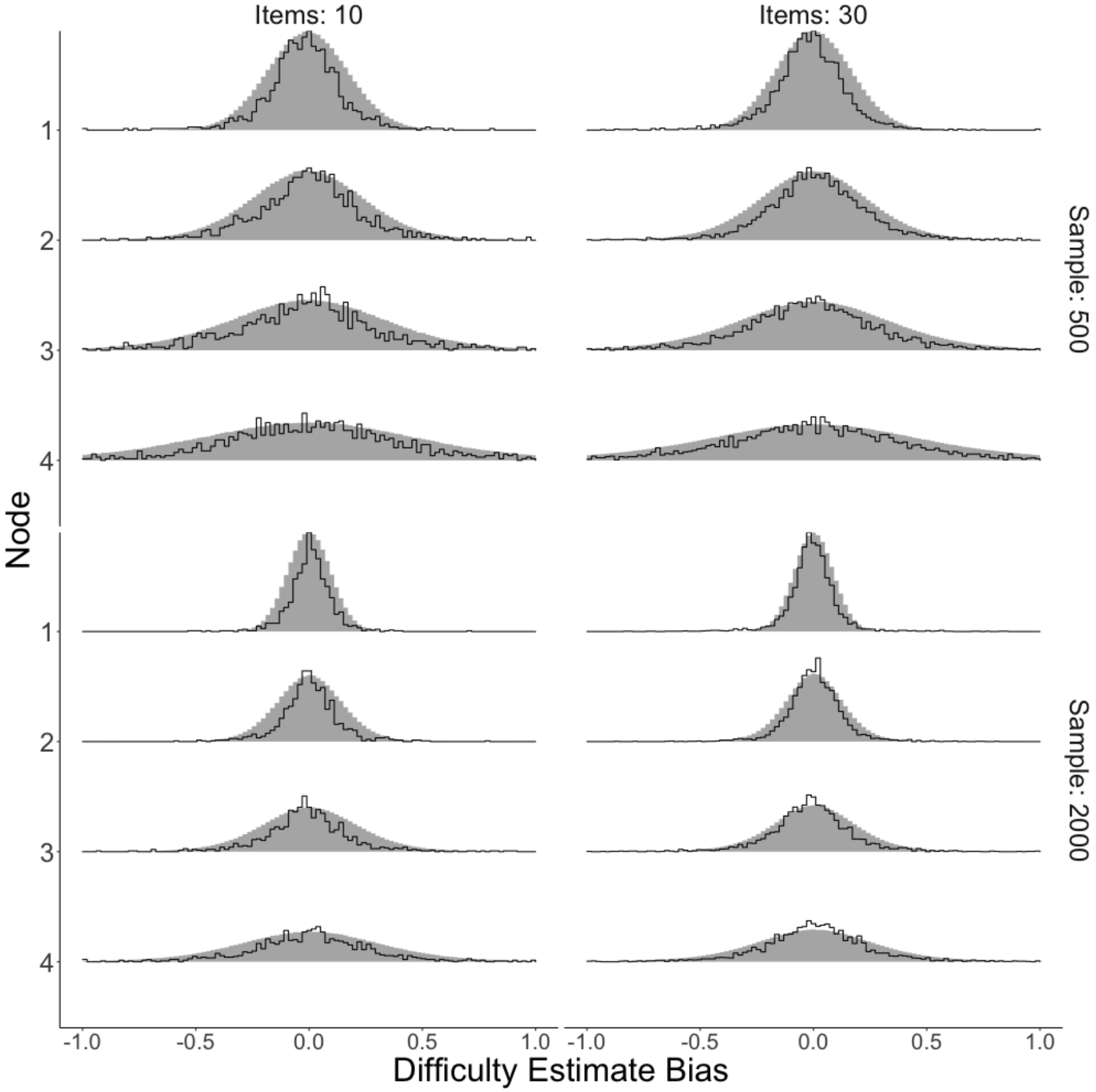
Adequate estimation of the item parameters requires a high degree of correlation between the true and estimated parameters. Strong positive correlations are indicative of estimated parameters with similar rank ordering to the true parameters. For both 1PL and 2PL simulation results, I inspected histograms (Figures 4 and 5) and scatter plots (Figures 6 and 7) of the item difficulty parameters. The true and estimated parameters were strongly and positively correlated for both the 1PL and 2PL models. Nodes 3 and 4 for the 10-item 500-respondent condition appear to have slightly weaker average correlations than the other conditions.

Figure 4. *Histograms of 1PL Item Difficulty Observed and Model Predicted Bias*



Note. Observed estimate of bias is displayed as a solid black line. Model predicted estimate bias is displayed as the grey shaded region

Figure 5. *Histograms of 2PL Item Difficulty Observed and Model Predicted Bias*.



Note. Observed estimate of bias is displayed as a solid black line. Model predicted estimate bias is displayed as the grey shaded region

Figure 6. *Scatterplot of True and Estimated 1PL Item Difficulty Parameters.*

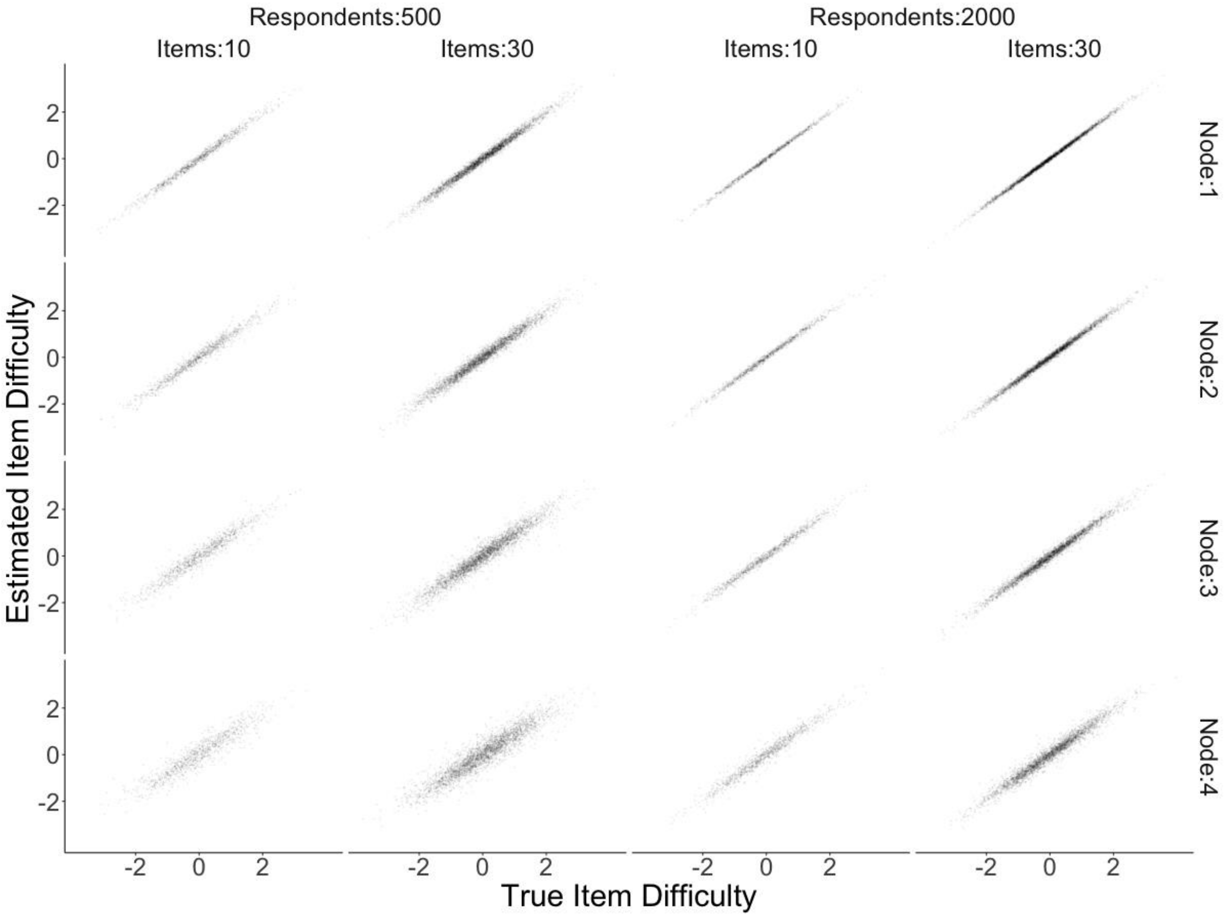
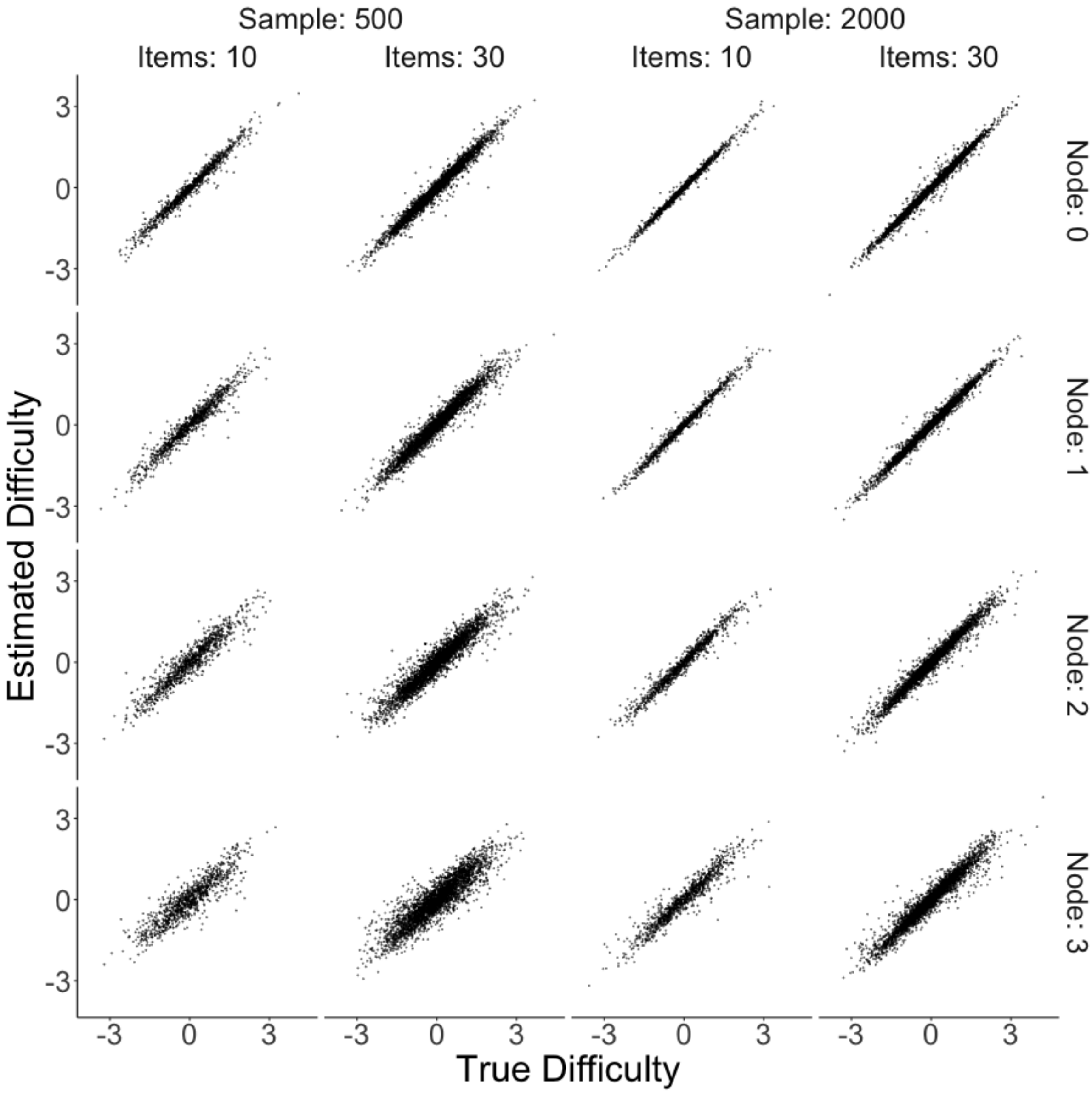


Figure 7. *Scatterplot of True and Estimated 2PL Item Difficulty Parameters*.



Regression Model Results

I conducted separate regression analyses for the 1PL and 2PL simulation data. I created each model with 3 MCMC chains, 1,000 warm-up iterations per chain, and 2,000 sampling iterations per chain. Inspection of the chains with trace plots suggested adequate mixing. All $\hat{R}$ values were equal to 1.00 after rounding which suggests adequate sampling from the posterior. The shaded regions in Figures 4 and 5 displays posterior predictions of bias across each condition. Table 4 displays predicted and observed bias means and standard deviations across each condition for the 1PL and 2PL models. The model predictions for the 1PL simulation appear to adequately approximate the observed bias distributions for all conditions. The 2PL predictions appear to overestimate the variance, particularly for nodes 3 and 4 with a sample size of 500. The overestimation does not appear to be severe, but some caution is warranted. Table 5 provides the model parameter means and 95% highest posterior density intervals (HPDI).

Table 4. *1PL Item Difficulty Model Predicted and Observed Means and Standard Deviations.*

			1PL Model				2PL Model			
	Node	Test Length	Sample Size = 500		Sample Size = 2000		Sample Size = 500		Sample Size = 2000	
			Predicted	Observed	Predicted	Observed	Predicted	Observed	Predicted	Observed
Mean of Bias	1	10	0.00	0.00	0.00	0.00	-0.01	-0.02	0.00	0.01
	2		0.00	0.00	0.00	0.00	-0.01	-0.01	0.00	0.00
	3		0.00	-0.02	0.00	-0.01	-0.01	-0.01	0.00	-0.01
	4		0.01	0.03	0.00	0.00	-0.01	0.01	0.01	0.01
	1	30	0.00	0.00	0.00	0.00	0.00	-0.02	0.00	-0.01
	2		0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.01
	3		0.00	0.00	0.00	-0.01	0.00	0.00	0.00	-0.01
	4		0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.01
Standard Deviation of Bias	1	10	0.11	0.12	0.06	0.06	0.18	0.18	0.10	0.10
	2		0.18	0.18	0.09	0.09	0.27	0.26	0.15	0.14
	3		0.26	0.26	0.13	0.13	0.38	0.34	0.23	0.20
	4		0.37	0.36	0.20	0.20	0.53	0.43	0.35	0.32
	1	30	0.11	0.12	0.06	0.06	0.17	0.18	0.09	0.12
	2		0.17	0.18	0.09	0.09	0.25	0.23	0.14	0.13
	3		0.26	0.26	0.14	0.13	0.38	0.33	0.21	0.20
	4		0.38	0.37	0.21	0.20	0.56	0.45	0.31	0.28

Mean parameter bias was very low. Posterior means of the model linear predictor parameters were zero with very little uncertainty. None of the predictors provided substantial explanation of bias because mean bias was essentially zero under all conditions. The variance in the mean bias intercepts between simulated datasets was also small for both the 1PL simulation, $\bar{\tau}_{\sigma_{[1PL]}} = 0.02,\ 95\%_{\text{HPDI}} = [0.02, 0.02]$, and 2PL simulation, $\bar{\tau}_{\sigma_{[2PL]}} = 0.00,\ 95\%_{\text{HPDI}} = [0.00, 0.00]$. This suggests that there is very little variability in mean bias between simulated datasets.

The between-simulation differences in estimate variability were small for the 1PL, $\bar{\tau}_{\sigma_{[1PL]}} = 0.02, 95\%_{\text{HDPI}}\ [0.00, 0.04]$, and 2PL simulations, $\bar{\tau}_{\sigma_{[2PL]}} = 0.06, 95\%_{\text{HDPI}}\ [0.05, 0.08]$. The variability in estimate bias is not meaningfully influenced by simulation iteration specific factors that are not already accounted for by the model predictors. The variability for the first node (i.e. the model intercept) was relatively small for the 1PL, $\bar{b}_{\sigma_{[1PL]}} = -2.54$, $95\%_{\text{HDPI}}[-2.56, -2.53]$, and 2PL models, $\bar{b}_{\sigma_{[2PL]}} = -2.10, 95\%_{\text{HDPI}}[-2.11, -2.09]$. The 2PL model had notably larger estimate variance than the 1PL model. Node depth had a reliable, small, and positive effect on estimate variability, $\bar{b}_{\sigma_{[1PL]}} = 0.34, 95\%_{\text{HPDI}}[0.34, 0.35]$, $\bar{b}_{\sigma_{[2PL]}} = 0.35, 95\%_{\text{HDPI}}[0.34, 0.36]$, such that deeper nodes exhibited greater estimate uncertainty. Sample size had a small negative reliable effect on estimate variability, $\bar{b}_{\sigma_{[1PL]}} = -0.36$, $95\%_{\text{HPDI}}[-0.37,\ -0.35]$, $\bar{b}_{\sigma_{[2PL]}} = -0.33, 95\%_{\text{HDPI}}[-0.34,\ -0.31]$, such that smaller sample sizes increase estimate uncertainty for the root-node. Relative propagation had a small negative reliable effect on estimate uncertainty, $\bar{b}_{\sigma_{[1PL]}} = -0.42, 95\%_{\text{HDPI}}\ [-0.46, -0.38]$, $\bar{b}_{\sigma_{[2PL]}} = -0.42, 95\%_{\text{HPDI}}\ [-0.44, -0.39]$, such that lower than average propagation increases estimate variability. The two-way and three-way interactions between sample size, node depth, and propagation were all reliably different from zero but were very small and did not provide much practical explanation of estimate variability. Test length did not have a reliable effect on difficulty parameter estimate uncertainty, nor did any of its interactions with the other predictors. Tables 6, 7, 8, and 9 display posterior predictions of difficulty parameter bias across different levels of sample size, propagation rates, and node depth for both models with 10- and 30-item tests. In sum, bias is negligible, the 2PL model exhibits greater estimate variability, and smaller sample sizes, lower relative propagation rates, and deeper nodes exhibit greater estimate variability.

Table 5. *Item Difficulty Models Posterior Means and Credibility Intervals of Predictors of Estimate Bias and Variability.*

	1PL Model				2PL Model			
	Estimate Bias		Estimate Variability†		Estimate Bias		Estimate Variability†	
	$\overline{b_\mu}$	$95\% HPDI$	$\overline{b_\sigma}$	$95\% HPDI$	$\overline{b_\mu}$	$95\% HPDI$	$\overline{b_\sigma}$	$95\% HPDI$
Intercept	0.00	[0.00, 0.00]	-2.54	[-2.56,-2.53]	0.00	[0.00, 0.00]	-2.10	[-2.11,-2.09]
D	0.00	[0.00, 0.00]	0.34	[0.34, 0.35]	0.00	[0.00, 0.00]	0.35	[0.34, 0.36]
S	0.00	[0.00, 0.00]	-0.36	[-0.37,-0.34]	0.00	[0.00, 0.00]	-0.33	[-0.34,-0.31]
T	0.00	[0.00, 0.00]	0.00	[-0.02, 0.01]	0.00	[0.00, 0.00]	-0.03	[-0.04,-0.02]
P	0.00	[-0.01, 0.01]	-0.42	[-0.46,-0.38]	0.01	[0.00, 0.01]	-0.42	[-0.44,-0.39]
D x S	0.00	[0.00, 0.00]	0.01	[0.00, 0.02]	0.00	[0.00, 0.00]	0.01	[0.01, 0.02]
D x T	0.00	[0.00, 0.00]	0.00	[0.00, 0.01]	0.00	[0.00, 0.00]	0.00	[-0.01, 0.00]
S x T	0.00	[0.00, 0.00]	0.00	[-0.02, 0.01]	0.00	[0.00, 0.00]	-0.01	[-0.02, 0.00]
D x P	0.00	[0.00, 0.01]	0.02	[0.00, 0.03]	0.00	[-0.01, 0.00]	0.04	[0.02, 0.06]
S x P	0.00	[-0.01, 0.01]	0.04	[0.00, 0.08]	-0.01	[-0.01, 0.00]	-0.03	[-0.06, 0.00]
T x P	0.00	[-0.01, 0.01]	-0.01	[-0.05, 0.04]	0.00	[-0.01, 0.01]	0.00	[-0.03, 0.02]
D x S x T	0.00	[0.00, 0.00]	0.00	[-0.01, 0.01]	0.00	[0.00, 0.00]	-0.01	[-0.02, 0.00]
D x S x P	0.00	[-0.01, 0.00]	-0.02	[-0.04, 0.00]	0.00	[0.00, 0.01]	0.01	[-0.01, 0.02]
D x T x P	0.00	[0.00, 0.00]	0.00	[-0.02, 0.02]	0.00	[-0.01, 0.01]	-0.01	[-0.02, 0.01]
S x T x P	0.00	[-0.01, 0.01]	-0.02	[-0.06, 0.02]	0.00	[-0.01, 0.00]	-0.02	[-0.05, 0.00]
D x S x T x P	0.00	[0.00, 0.00]	0.01	[-0.01, 0.03]	0.00	[0.00, 0.01]	0.01	[-0.01, 0.03]
τ	0.02	[0.02, 0.02]	0.02	[0.00, 0.04]	0.00	[0.00, 0.00]	0.06	[0.05, 0.08]

Note. *N* = 32,000. D = node depth, S = sample size (standardized, *M* = 1,250, *SD* = 750.01), T = test length (standardized, *M* = 25.00, *SD* = 8.66), P = log-relative propagation, τ = between-simulation iteration intercept variance, HPDI = highest posterior density interval. † Model parameters are on the log scale.

Table 6. *Posterior Prediction Means and Credibility Intervals of 1PL Item Difficulty Parameter Estimate Bias for 10-item test.*

Node	Propagation Ratio[†]	Sample Size = 250		Sample Size = 500		Sample Size = 1,000		Sample Size = 2,000	
		$\overline{\beta^*}$	95% HPDI	$\overline{\beta^*}$	95% HPDI	$\overline{\beta^*}$	95% HPDI	$\overline{\beta^*}$	95% HPDI
1	—	0.00	[-0.24,0.26]	0.00	[-0.22,0.23]	0.00	[-0.18,0.17]	0.00	[-0.12,0.12]
2	0.05	0.03	[-1.29,1.36]	0.02	[-1.20,1.11]	0.02	[-0.82,0.88]	0.00	[-0.45,0.45]
	0.5	0.01	[-0.47,0.47]	0.01	[-0.43,0.42]	0.00	[-0.32,0.32]	0.00	[-0.20,0.20]
	1	0.01	[-0.35,0.35]	0.00	[-0.31,0.31]	0.00	[-0.26,0.23]	0.00	[-0.17,0.15]
	2	0.00	[-0.28,0.25]	-0.01	[-0.24,0.23]	0.00	[-0.19,0.19]	0.00	[-0.13,0.12]
3	0.05	0.03	[-1.42,1.58]	0.01	[-1.25,1.38]	0.01	[-1.01,1.09]	0.01	[-0.68,0.67]
	0.5	0.01	[-0.60,0.63]	0.00	[-0.54,0.56]	0.01	[-0.44,0.44]	0.00	[-0.29,0.28]
	1	0.00	[-0.47,0.48]	0.00	[-0.42,0.44]	0.00	[-0.35,0.33]	-0.01	[-0.23,0.21]
	2	0.00	[-0.36,0.38]	0.00	[-0.33,0.33]	0.00	[-0.26,0.28]	0.00	[-0.19,0.16]
4	0.05	0.02	[-1.79,1.63]	0.00	[-1.62,1.60]	0.01	[-1.41,1.36]	0.00	[-1.03,1.00]
	0.5	0.01	[-0.79,0.85]	0.00	[-0.74,0.76]	0.00	[-0.59,0.65]	0.00	[-0.41,0.40]
	1	0.00	[-0.64,0.70]	0.00	[-0.62,0.59]	0.00	[-0.45,0.50]	0.00	[-0.32,0.30]
	2	0.00	[-0.57,0.51]	0.01	[-0.44,0.53]	0.00	[-0.37,0.40]	-0.01	[-0.24,0.24]

Note. $\overline{\beta^*}$ = mean posterior predicted difficulty estimate bias, HPDI = highest posterior density interval. [†]Formula for approximate number of observed responses: $n = Sample\ Size \times Propagation\ Ratio \times 0.5^{Node-1}$.

Table 7. *Posterior Prediction Means and Credibility Intervals of 1PL Item Difficulty Parameter Estimate Bias for 30-item test.*

Node	Propagation Ratio†	Sample Size = 250		Sample Size = 500		Sample Size = 1,000		Sample Size = 2,000	
		$\overline{\beta^*}$	95% HPDI	$\overline{\beta^*}$	95% HPDI	$\overline{\beta^*}$	95% HPDI	$\overline{\beta^*}$	95% HPDI
1	—	0.00	[-0.26,0.25]	0.00	[-0.23,0.22]	0.00	[-0.18,0.17]	0.00	[-0.11,0.12]
2	0.05	0.01	[-1.25,1.24]	0.01	[-1.02,1.14]	0.01	[-0.83,0.83]	0.00	[-0.50,0.48]
	0.5	0.00	[-0.48,0.46]	0.00	[-0.42,0.40]	0.00	[-0.33,0.32]	0.00	[-0.20,0.20]
	1	-0.01	[-0.35,0.34]	0.00	[-0.29,0.32]	0.00	[-0.23,0.25]	0.00	[-0.15,0.16]
	2	0.00	[-0.26,0.26]	0.00	[-0.22,0.24]	0.00	[-0.19,0.18]	0.00	[-0.13,0.11]
3	0.05	0.00	[-1.56,1.56]	-0.01	[-1.40,1.37]	0.00	[-1.12,1.07]	0.01	[-0.68,0.71]
	0.5	0.00	[-0.61,0.63]	0.00	[-0.56,0.56]	0.00	[-0.44,0.45]	0.00	[-0.30,0.27]
	1	0.00	[-0.50,0.47]	0.00	[-0.45,0.42]	0.00	[-0.34,0.36]	0.00	[-0.22,0.22]
	2	0.00	[-0.37,0.37]	0.00	[-0.32,0.37]	0.00	[-0.27,0.26]	0.00	[-0.17,0.17]
4	0.05	0.01	[-1.90,1.89]	0.00	[-1.70,1.76]	0.00	[-1.38,1.44]	-0.01	[-0.95,0.99]
	0.5	0.00	[-0.90,0.80]	-0.01	[-0.76,0.75]	0.00	[-0.63,0.61]	-0.01	[-0.45,0.37]
	1	-0.01	[-0.67,0.69]	0.00	[-0.60,0.61]	0.00	[-0.49,0.47]	0.00	[-0.32,0.31]
	2	0.01	[-0.52,0.53]	0.00	[-0.46,0.47]	0.00	[-0.37,0.38]	0.00	[-0.24,0.24]

Note. $\overline{\beta^*}$ = mean posterior predicted difficulty estimate bias, HPDI = highest posterior density interval. †Formula for approximate number of observed responses: $n = Sample\ Size \times Propagation\ Ratio \times 0.5^{Node-1}$.

Table 8. *Posterior Prediction Means and Credibility Intervals of 2PL Item Difficulty Parameter Estimate Bias for 10-item Tests.*

Node	Propagation Ratio†	Sample Size = 250		Sample Size = 500		Sample Size = 1,000		Sample Size = 2,000	
		$\overline{\beta^*}$	95% HPDI	$\overline{\beta^*}$	95% HPDI	$\overline{\beta^*}$	95% HPDI	$\overline{\beta^*}$	95% HPDI
1	—	-0.01	[-0.39,0.38]	-0.01	[-0.35,0.34]	-0.01	[-0.30,0.27]	0.00	[-0.19,0.19]
2	0.05	-0.05	[-1.65,1.46]	-0.05	[-1.46,1.43]	-0.03	[-1.18,1.22]	-0.01	[-0.81,0.80]
	0.5	-0.02	[-0.71,0.66]	-0.02	[-0.66,0.61]	-0.01	[-0.51,0.50]	0.00	[-0.35,0.35]
	1	0.00	[-0.55,0.51]	-0.01	[-0.50,0.46]	0.00	[-0.40,0.39]	0.00	[-0.26,0.28]
	2	0.00	[-0.41,0.43]	0.00	[-0.37,0.38]	0.00	[-0.31,0.32]	0.00	[-0.22,0.21]
3	0.05	-0.06	[-1.71,1.80]	-0.06	[-1.62,1.59]	-0.04	[-1.40,1.36]	0.03	[-1.04,1.04]
	0.5	-0.01	[-0.86,0.89]	-0.02	[-0.81,0.81]	0.00	[-0.71,0.67]	0.01	[-0.52,0.50]
	1	-0.01	[-0.73,0.71]	0.00	[-0.69,0.65]	0.00	[-0.55,0.56]	0.00	[-0.39,0.40]
	2	0.00	[-0.59,0.61]	0.00	[-0.55,0.54]	0.00	[-0.45,0.46]	0.00	[-0.33,0.32]
4	0.05	-0.05	[-2.03,2.04]	-0.07	[-2.09,1.73]	-0.02	[-1.65,1.69]	0.04	[-1.28,1.32]
	0.5	-0.02	[-1.22,1.12]	-0.03	[-1.12,1.05]	-0.01	[-0.92,0.98]	0.01	[-0.68,0.73]
	1	-0.01	[-1.04,0.95]	-0.01	[-0.91,0.96]	-0.01	[-0.73,0.85]	0.01	[-0.60,0.58]
	2	0.01	[-0.85,0.84]	0.02	[-0.77,0.81]	0.00	[-0.71,0.66]	-0.01	[-0.51,0.49]

Note. $\overline{\beta^*}$ = mean posterior predicted difficulty estimate bias, HPDI = highest posterior density interval. †Formula for approximate number of observed responses: $n = Sample\ Size \times Propagation\ Ratio \times 0.5^{Node-1}$.

Table 9. *Posterior Prediction Means and Credibility Intervals of 2PL Item Difficulty Parameter Estimate Bias for 30-item Tests.*

Node	Propagation Ratio†	Sample Size = 250		Sample Size = 500		Sample Size = 1,000		Sample Size = 2,000	
		$\overline{\beta^*}$	95% HPDI	$\overline{\beta^*}$	95% HPDI	$\overline{\beta^*}$	95% HPDI	$\overline{\beta^*}$	95% HPDI
1	—	0.00	[-0.37, 0.37]	0.00	[-0.34, 0.33]	0.00	[-0.26, 0.27]	0.00	[-0.16, 0.17]
2	0.05	-0.04	[-1.43, 1.45]	-0.03	[-1.36, 1.28]	-0.02	[-1.20, 1.10]	0.00	[-0.89, 0.80]
	0.5	-0.01	[-0.67, 0.66]	-0.01	[-0.62, 0.56]	-0.01	[-0.52, 0.44]	0.00	[-0.33, 0.31]
	1	-0.01	[-0.50, 0.54]	0.00	[-0.47, 0.46]	0.01	[-0.36, 0.38]	0.01	[-0.22, 0.26]
	2	0.01	[-0.39, 0.41]	0.01	[-0.35, 0.36]	0.01	[-0.27, 0.30]	0.00	[-0.19, 0.17]
3	0.05	-0.01	[-1.93, 1.89]	-0.03	[-1.84, 1.67]	0.00	[-1.55, 1.40]	0.01	[-1.02, 1.08]
	0.5	0.00	[-0.89, 0.89]	-0.01	[-0.79, 0.83]	0.00	[-0.66, 0.66]	0.00	[-0.43, 0.45]
	1	0.00	[-0.72, 0.72]	0.01	[-0.65, 0.65]	0.01	[-0.53, 0.53]	0.00	[-0.34, 0.35]
	2	0.00	[-0.60, 0.55]	0.01	[-0.51, 0.55]	0.01	[-0.40, 0.44]	0.00	[-0.27, 0.26]
4	0.05	0.02	[-2.47, 2.43]	0.01	[-2.27, 2.21]	0.00	[-1.82, 1.86]	-0.01	[-1.28, 1.33]
	0.5	0.00	[-1.27, 1.21]	-0.01	[-1.05, 1.16]	0.01	[-0.90, 0.94]	0.00	[-0.65, 0.60]
	1	-0.01	[-1.05, 1.00]	0.00	[-0.94, 0.91]	0.00	[-0.71, 0.77]	0.01	[-0.47, 0.53]
	2	0.01	[-0.80, 0.86]	0.00	[-0.72, 0.74]	0.01	[-0.57, 0.61]	0.01	[-0.38, 0.40]

Note. $\overline{\beta^*}$ = mean posterior predicted difficulty estimate bias, HPDI = highest posterior density interval. †Formula for approximate number of observed responses: $n = Sample\ Size \times Propagation\ Ratio \times 0.5^{Node-1}$.

Adequate Recovery Criterion

I conducted posterior predictive simulations to get a better practical understanding for when variability exceeds the $\sigma_{\hat{\beta}} = 0.25$ limit mentioned above. Propagation proportion for the 1PL is determined by the combination of auxiliary-item difficulty parameters in a tree, which in turn determines estimate uncertainty. These limits suggest some constraint on the possible combination of auxiliary-item difficulties in an item set. Using all 6,000 posterior draws from each regression model, I simulated data for a 4-node tree with sample sizes of 500 and 2,000, test lengths of 10 items and 30 items, and relative propagation rates ranging from 0.10 to 3 in increments of 0.10. For each combination of sample size, test length, and node depth, I then identified the minimum propagation rate required to maintain the $\sigma_{\hat{\beta}} = 0.25$ limit. I then simulated 1,000,000 draws from a standard normal distribution for 3 nodes (depths 1 through 3), converted them to probabilities with the logistic function, and calculated their cumulative product. For each condition and each node, I then calculated the proportion of these simulated products that propagated at least the minimum proportion of observations identified above that propagate enough observations. Table 10 presents these propagation minimums, approximate minimum node sample sizes, and the estimated proportion of all possible combinations of normally distributed difficulty parameters that would produce at least the minimum propagation rate. These proportions represent how likely a randomly drawn set of auxiliary-items would propagate enough observations to estimate the underlying item difficulties within the $\sigma_{\hat{\beta}} = 0.25$ uncertainty interval.

For the 1PL model, across all conditions the minimum sample size for a given node was roughly 100 observations. With large sample sizes and short maximum tree lengths (e.g., depth of 1 or 2), it is likely for items to exhibit difficulty that allows for sufficiently narrow uncertainty intervals. Approximately 98% of all possible node 1 item difficulties result in adequate propagation to node 2 (depth of 1). This drops to 86% for node 3 (depth of 2). For nodes at a depth of 3 (node 4) and large sample sizes, the percentage of difficulty parameter combinations drops to around 60-70% depending on the test length. In more concrete terms, a researcher could expect 3-4 items on a 10-item test to have large uncertainty intervals at a depth of 3. The results are more concerning for small sample sizes. 84% of all possible root-node items will propagate enough observations to nodes at depths of 1. The possible combinations drop to random chance levels for depths of 2. Finally, it is unlikely that ancestor items will propagate enough observations to nodes at depths of 3 or greater. In concrete terms, a 10-item test would likely have 8-9 items with impractically large uncertainty intervals for nodes at depths of 3. These results suggest that small sample sizes and long tree designs cannot produce data which provide sufficient certainty for the deepest node item parameters estimates.

I used the same procedure for the 2PL regression model to estimate the minimum propagation rates. When calculating the proportion of parameter combinations, I included a discrimination parameter using 1,000,000 draws from a standard log-normal distribution. The results for the 2PL models are more restrictive. Unlike the 1PL model, the minimum node-specific sample size required increases with depth and decreases slightly with longer tests and larger total samples sizes. Large sample sizes make adequate uncertainty estimation likely for nodes at a depth of 1. This drops to around 60-66% for depth 2 with large sample sizes. Adequate estimation becomes unlikely for both 10- and 30-item tests for nodes at a depth of 3. The proportion of root-node item difficulty parameters that would propagate enough observations to the second node (depth of 1) with small sample sizes is near chance levels. This proportion becomes highly unlikely for small sample sizes with depths of 2. Finally, the simulation did not produce any parameter combinations that would propagate enough observations to nodes at a depth of 3 when using small sample sizes.

Both the 1PL and 2PL minimum propagation predictions are anti-conservative. The regression models do not include predictors for differences in item difficulty or discrimination. The regression model predictors make predictions for the bias distribution of an average item. Items that have difficulty parameters further from the average, or discrimination parameters closer to zero, exhibit greater estimate variability (Thissen & Wainer, 1982).

Table 10. *Item Difficulty Minimum Propagation.*

			1PL Model			2PL Model		
Test Length	Total Sample Size	Node[†]	Minimum Propagation Rate	Node Sample Size	Parameter Combinations[††]	Minimum Propagation Rate	Node Sample Size	Parameter Combinations[††]
10 Items	500	2	0.20	100	0.84	0.50	250	0.50
		3	0.18	87.50	0.55	0.63	312.50	0.06
		4	0.24	118.75	0.16	–	–	–
	2000	2	0.05	100	0.98	0.10	200	0.94
		3	0.05	100	0.86	0.15	300	0.60
		4	0.05	100	0.63	0.24	475	0.16
30 Items	500	2	0.20	100	0.84	0.45	225	0.56
		3	0.20	100	0.50	0.60	300	0.07
		4	0.21	106.25	0.20	–	–	–
	2000	2	0.05	100	0.98	0.10	200	0.94
		3	0.05	100	0.86	0.13	250	0.66
		4	0.04	75	0.71	0.14	275	0.33

Note. [†] Node 1 (i.e., depth of 0 or root-node) is excluded as it is not affected by propagation. [††] Proportion of possible item difficulty parameter combinations for ancestor nodes that would propagate enough observations on average to achieve a two-standard deviation uncertainty interval ranging between -0.5 and 0.5 for item difficulty estimates. Minimum Propagation is the proportion of observations from the total sample size that must be propagated. Node Sample Size is the Total Sample Size multiplied by the Minimum Propagation. There are missing values for the 2PL model with for node 4 and a sample size of 500 because these conditions did not produce sufficiently narrow uncertainty intervals across the range of simulated relative propagation values.

Difficulty Estimate Standard Errors

The mirt estimation procedure did not produce standard errors for the 2PL models, so I only analyzed the standard errors for the 1PL models. I calculated coverage rates of the standard errors as the average number of true item difficulty estimates that resided within 95% normal theory confidence intervals using the standard errors produced by the mirt estimation procedure. The rates for all conditions were high (*M* = 94%, *SD* = 1%), with the lowest (88%) occurring for 500 respondent sample sizes with a 30-item test for the root-node. Regarding coverage, the standard errors are large enough to produce confidence intervals that encompass the true difficulty parameter estimates.

Using posterior predictive simulations, I estimated the regression model predicted estimate standard deviation and compared this to the standard errors produced by the mirt package estimation procedure for each item in the simulated data sets. I consider a standard error of an item that is smaller than its model predicted estimate standard deviation to be anti-conservative relative to the regression model predictions. This assumes that the regression model predicted standard deviations are valid. I calculated rates of anti-conservative standard errors in each condition. Table 11 display these rates. The lowest rate, 0.55, occurred in 30-item tests with 500 participants at the root-node, and the highest rate, 0.80, occurred for 30-item tests with 2,000 participants with a higher-than-average propagation rate at node 4. This suggests that the standard errors are not large enough to provide an adequate estimate of the uncertainty around the parameter estimates, a conclusion inconsistent with the coverage estimates. This could be an indication that the regression model overestimates the variance in estimate bias. These results suggest that the standard errors produced estimation procedure are adequate for producing confidence intervals that encompass the true parameter. However, assuming the regression model predictions are valid, the standard errors of the difficulty estimates may not be conservative enough for tasks such as item selection.

Table 11. *Rates of Anti-Conservative Standard Error Estimates and Coverage for 1PL Item Difficulty Parameter Estimates.*

			Anti-Conservative SE		SE Coverage	
			Relative Propagation		Relative Propagation	
Node	Sample Size	Test Length	High	Low	High	Low
1	500	10	0.56	—	0.95	—
		30	0.55	—	0.93	—
	2,000	10	0.75	—	0.94	—
		30	0.66	—	0.95	—
2	500	10	0.57	0.72	0.95	0.94
		30	0.61	0.71	0.94	0.95
	2,000	10	0.76	0.57	0.96	0.95
		30	0.64	0.64	0.95	0.94
3	500	10	0.69	0.65	0.94	0.94
		30	0.70	0.68	0.95	0.94
	2,000	10	0.75	0.63	0.96	0.96
		30	0.77	0.68	0.94	0.94
4	500	10	0.79	0.62	0.94	0.96
		30	0.79	0.71	0.94	0.93
	2,000	10	0.77	0.66	0.94	0.94
		30	0.80	0.69	0.95	0.94

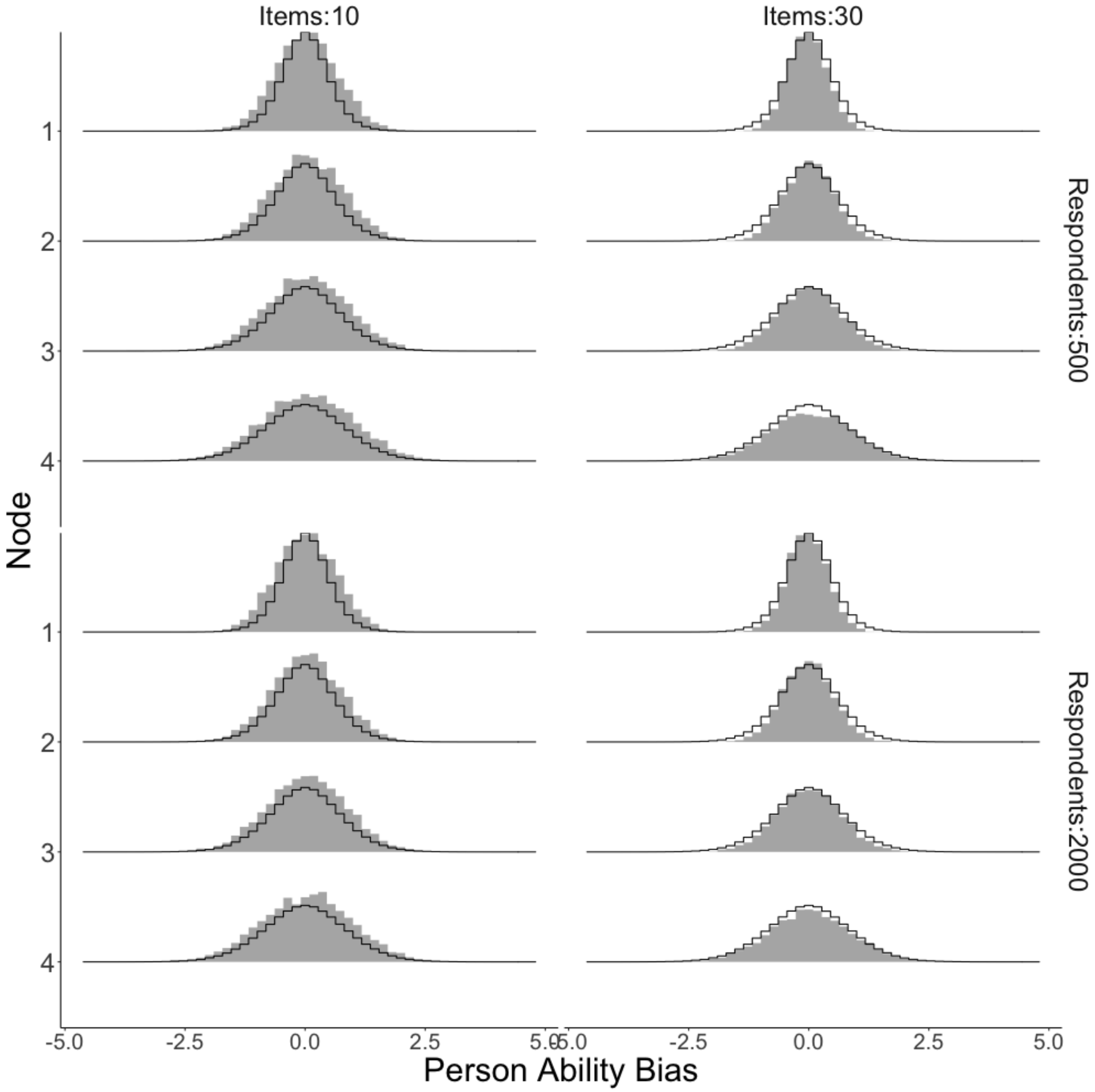
Note. SE = IRTree model standard error. Relative Propagation indicates whether the relative observation propagation rate was higher/equal to or lower than the average (i.e., 1).

Person Ability Parameter Bias

Descriptive Analysis

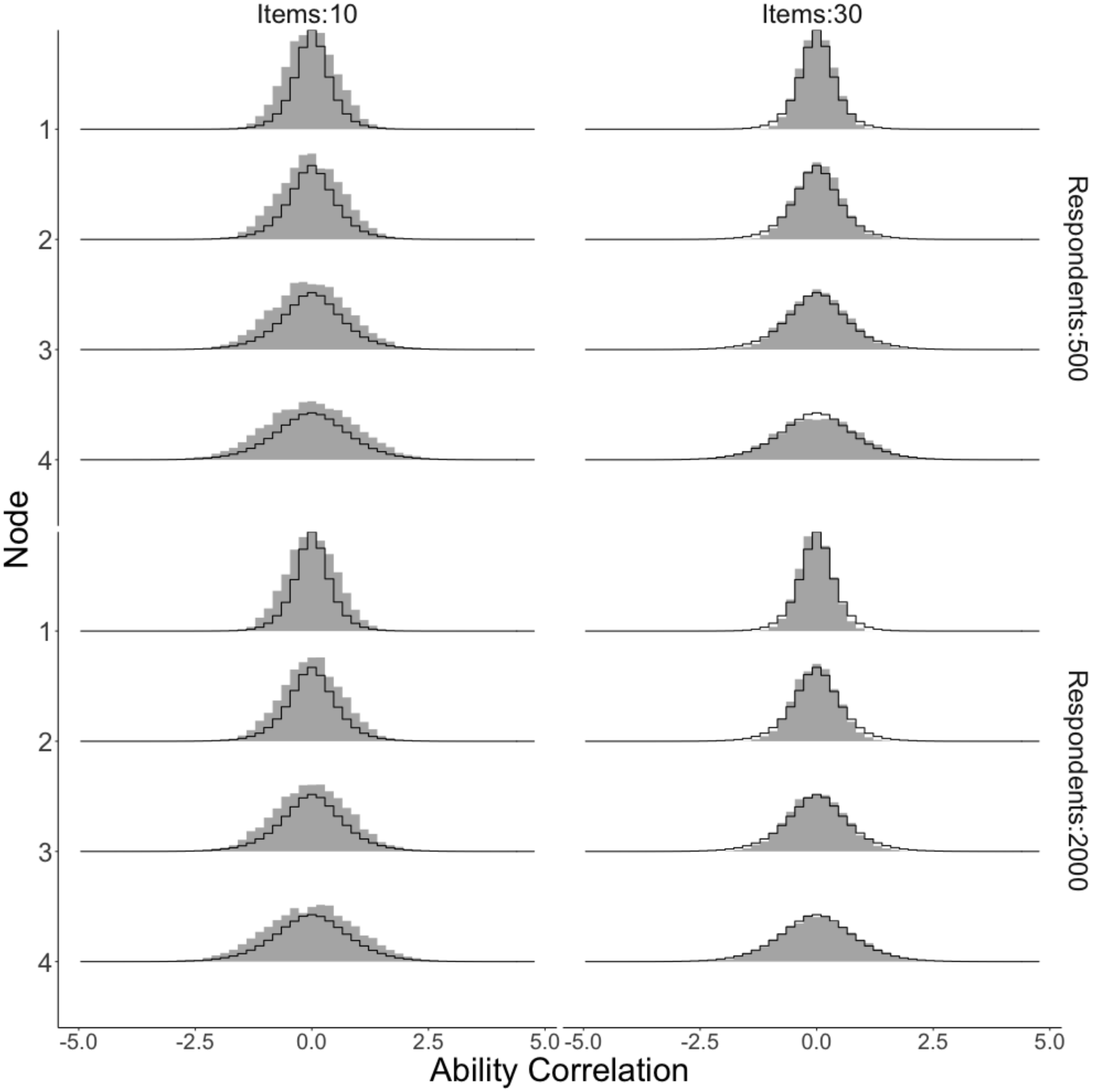
Figures 8 and 9 displays histograms of the person ability bias distributions, which appear normally distributed and increase in variance with deeper nodes. I inspected a scatter plot of the true and estimated person ability parameters (Figures 10 and 11). Two observations were immediately obvious. The strength and reliability of the positive linear relationships between the true and estimated parameters diminishes as node depth increases. There is also a flat horizontal line of estimated parameters at $\hat{\theta} = 0$ for nodes 3 and 4 for all conditions. The line is much more prominent for node 4 and for the 10-item test lengths. This line is also barely visible for the 10-item tests at node 2. Node 1 does not appear to exhibit this trend. The line suggests that a large portion of the estimated parameters are shrunk to zero regardless of true ability. I then inspected density plots of the estimated ability parameters for each simulation iteration, separated by condition and node depth (Figure 12 and 13). These plots further indicate that large portions of the estimates in each iteration are shrunk to near zero.

Figure 8. *Observed and Model Predicted 1PL Person Ability Parameter Bias.*



Note. True ability parameter distribution is displayed with a black line. Estimated ability parameter distribution is displayed with the grey shaded region.

Figure 9. *Observed and Model Predicted 2PL Person Ability Estimate Bias.*



Note. True ability parameter distribution is displayed with a black line. Estimated ability parameter distribution is displayed with the grey shaded region.

Figure 10. *Scatterplot of 1PL True and Estimated Person Ability Parameters*

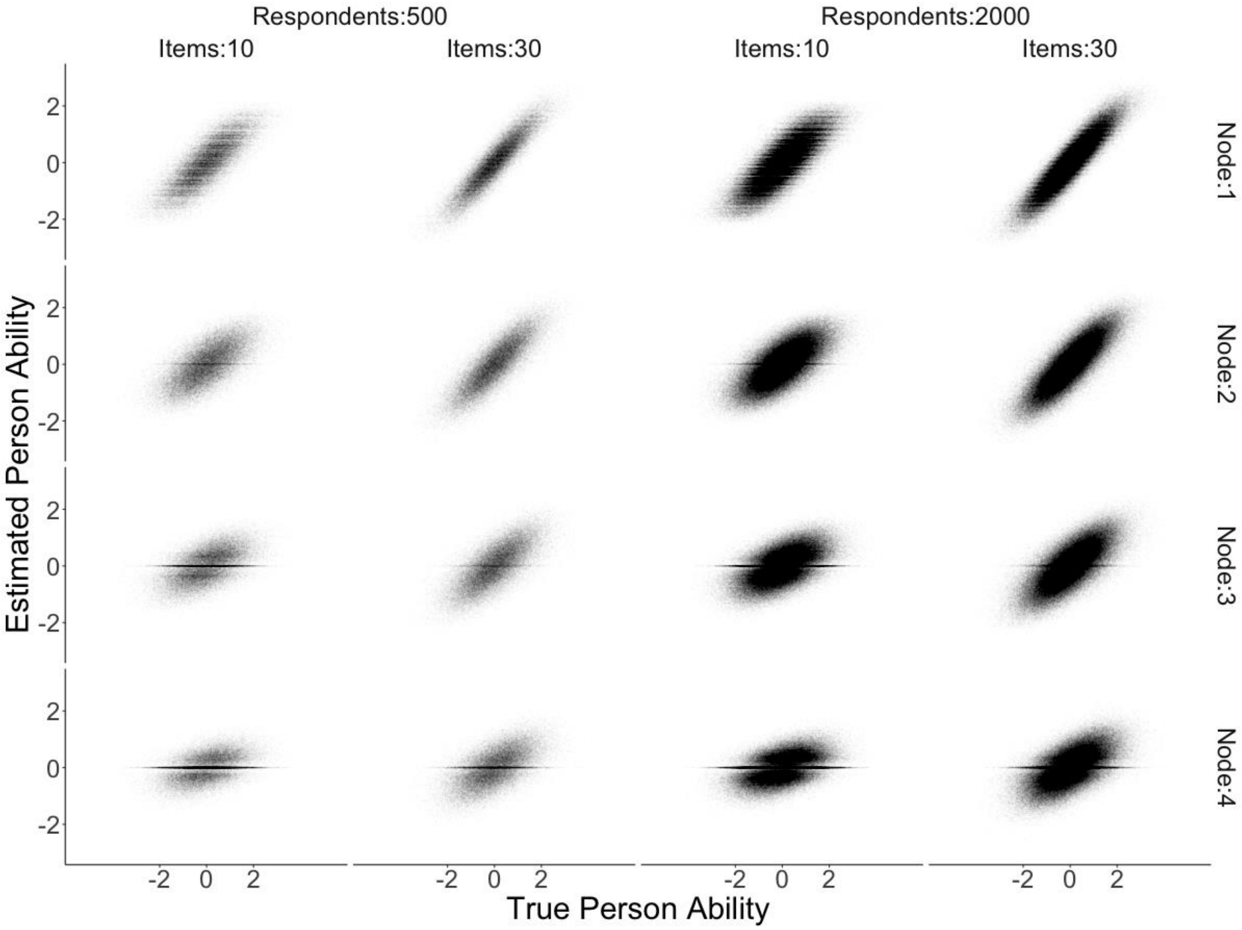


Figure 11. *Scatterplot of 2PL True and Estimated Person Ability Parameters*.

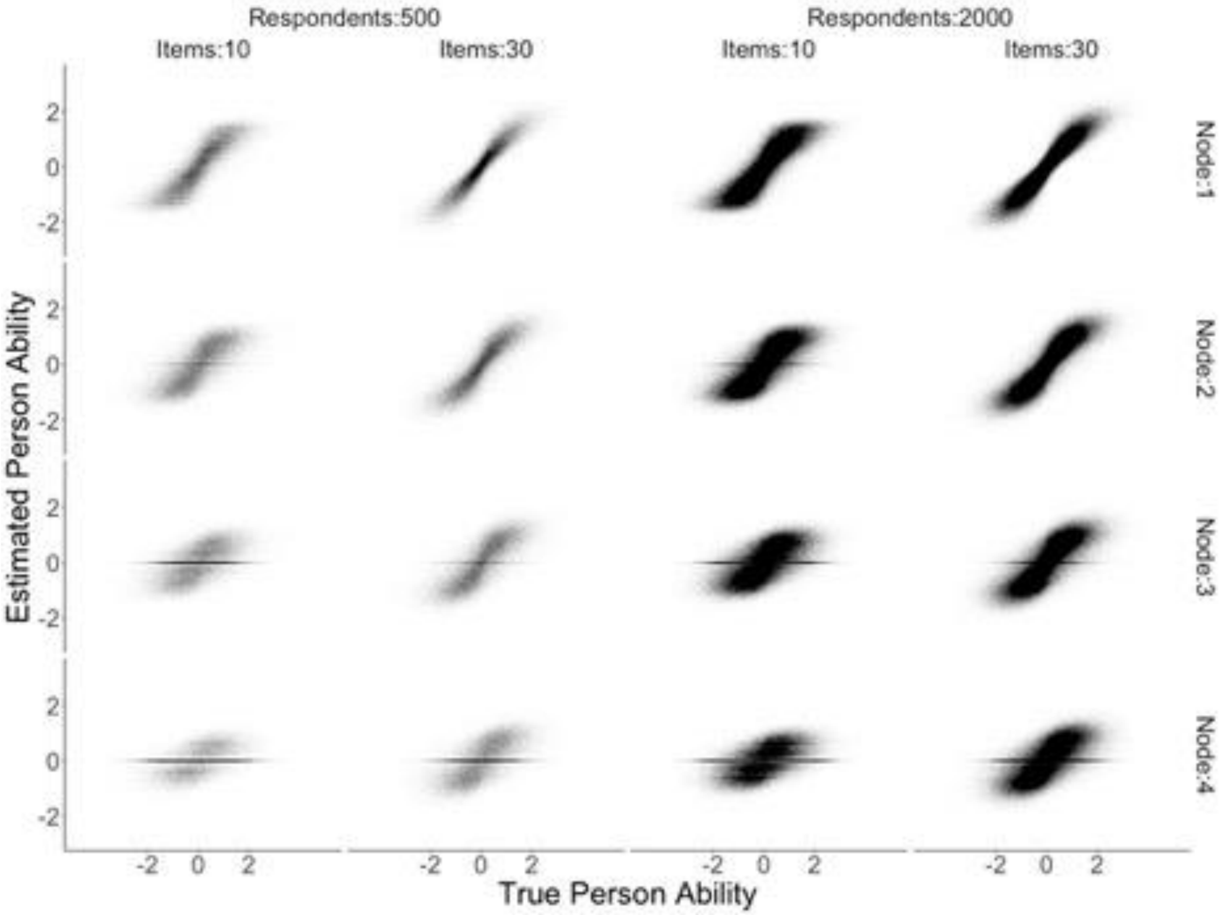


Figure 12. *Histograms of 1PL Person Ability Parameter Estimates*.

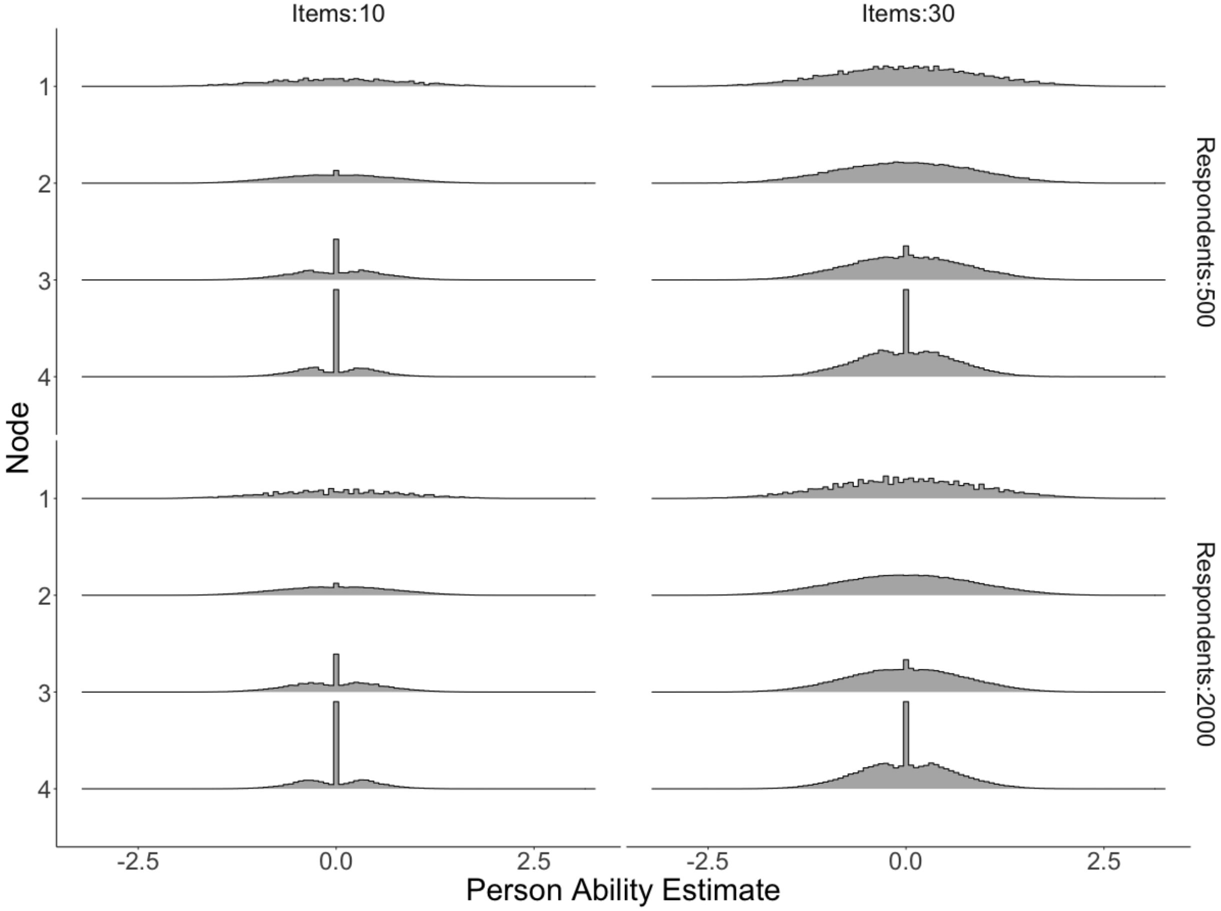
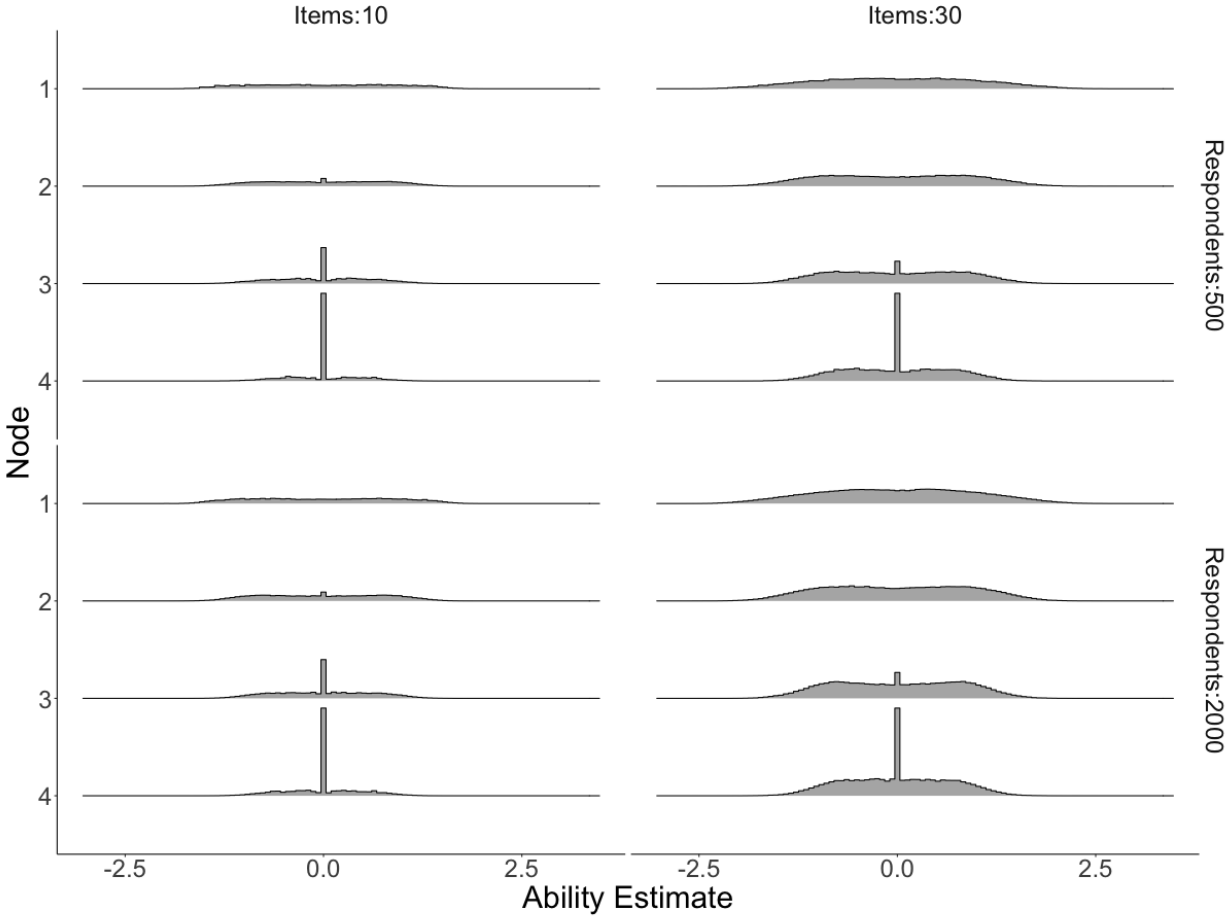


Figure 13. *Histograms of 2PL Person Ability Estimates.*

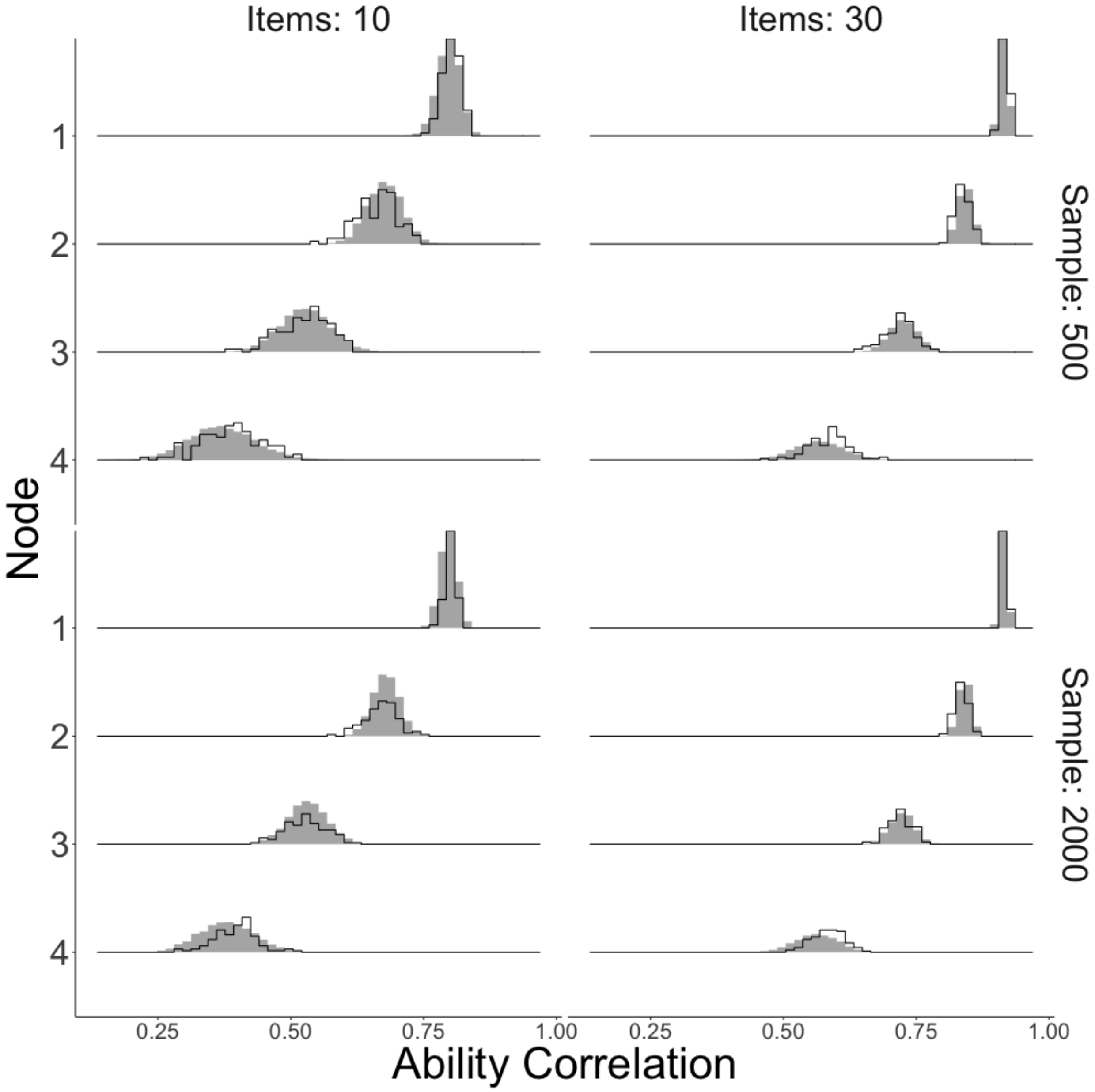


The ability parameters were estimated with the expected a posteriori (EAP) procedure in the mirt package, which estimates the mean of the posterior of θ. This requires calculating the likelihood of the data given the item parameters over some distribution of θ (typically a standard normal distribution). If there is no data to calculate the likelihood and assuming θ is normally distributed with a mean of zero, the EAP estimate should be zero. Participants that never reach some nodes in an IRTree for any item would have completely missing data for those latent traits. There is no information available to estimate these participants' latent ability parameters. This results in zero or near zero estimates and is likely what produced these horizontal lines in the scatter plots.

I simulated 100 datasets with 2,000 respondents for 10-item and 30-items tests and calculated the proportion of respondents with zero responses for each node. For a 10-item test, the resultant proportions of not answering any items at depths 1, 2, and 3, are $0.02\ 95\%_{\mathrm{HDPI}}[0.00, 0.05], 0.13\ 95\%_{\mathrm{HDPI}}[0.04, 0.24]$, and $0.36\ 95\%_{\mathrm{HDPI}}[0.21, 0.49]$. For a 30-item test, the resultant proportions of not answering any items at depths 1, 2, and 3, are $0.00\ 95\%_{\mathrm{HDPI}}[0.00, 0.00], 0.02\ 95\%_{\mathrm{HDPI}}[0.00, 0.05]$, and $0.10\ 95\%_{\mathrm{HDPI}}[0.05, 0.19]$. These are estimates of the expected proportion of the total sample size for which the latent ability parameters cannot be estimated from the data for a given node.

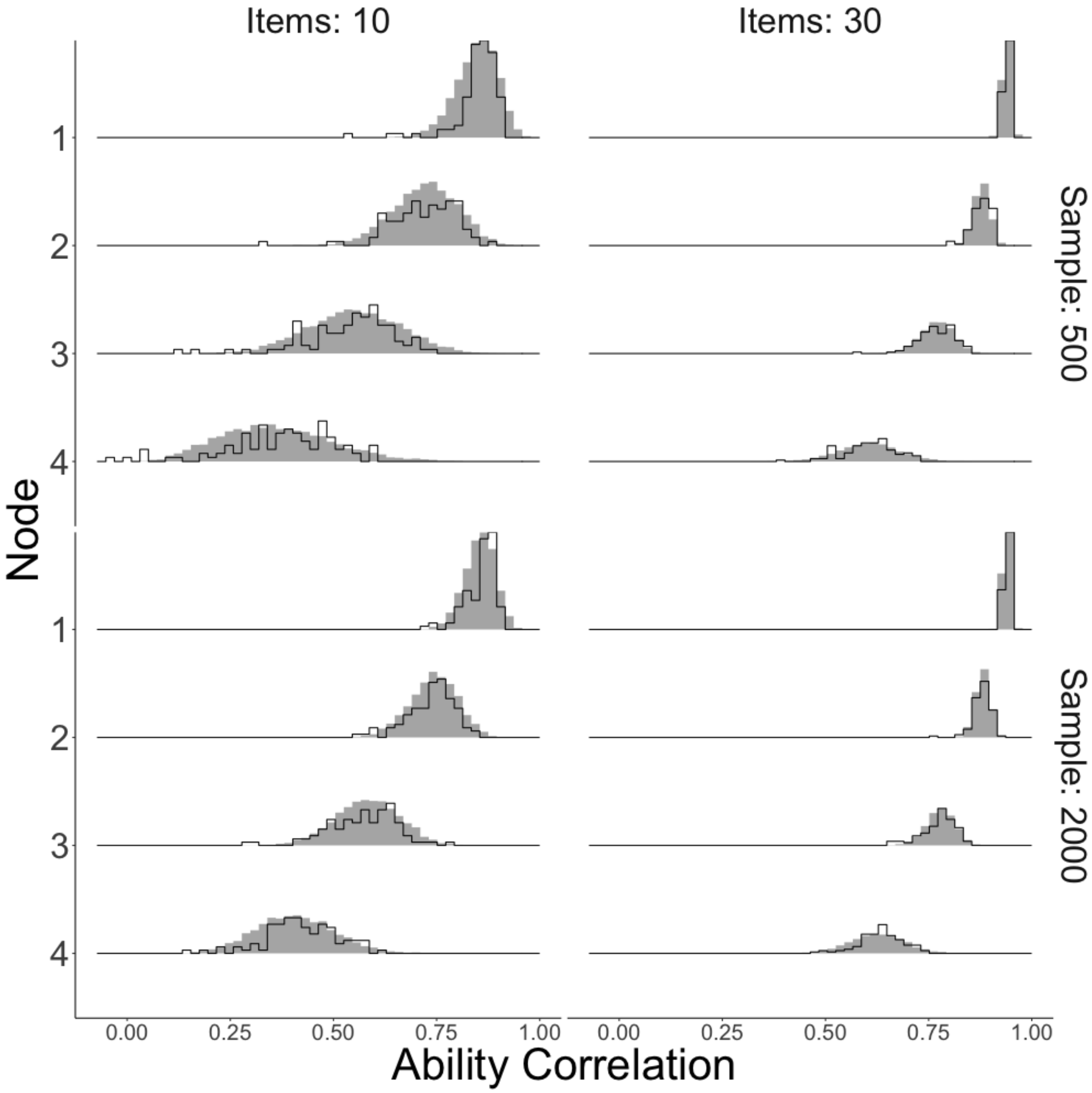
This mixture of estimable and non-estimable parameters is not revealed in density plots of estimate bias, which all look uni-modal and normally distributed. EAP estimates shrunk to zero are canceled out when estimate bias is calculated, so this spike at zero is not apparent unless the actual estimates are inspected. Although the shrunken estimates are not affecting estimate bias or variability, they likely affect the rank order of the estimates. I inspected the distributions of within-simulation iteration correlations between the estimated and true ability parameters for each condition and each node (Figure 14). The distributions suggest that the correlations are all positive but decrease sharply at greater depths in all conditions. The shrunken estimates are likely attenuating the correlations between the true and estimated parameters. I suspect that removal of respondents that have completely missing data on a given latent trait from the EAP estimate distribution will improve these correlations to some degree. I did not save information about each respondent’s response pattern, so I do not have information about which respondents have completely missing data. Another study is required to investigate this further.

Figure 14. *Histograms of Correlations Between True and Estimated 1PL Person Ability Parameters.*



Note. Observed correlation distributions are shown by the solid black line. Posterior predictions of the correlations are shown by the grey shaded regions.

Figure 15. *Histograms of Correlations Between True and Estimated 2PL Person Ability Parameters*.



Note. Observed correlation distributions are shown by the solid black line. Posterior predictions of the correlations are shown by the grey shaded regions.

Regression Model Results

Assuming that the shrunken EAP estimates are approximately zero or approximate a normal distribution with a mean of zero and very small standard deviation, the mean and variance of the bias distribution should not be affected. It may still be informative to quantify the relationships between the manipulated factors and ability estimate bias and variability.

Due to the number of observations in the sample ($N = 2{,}000{,}000$), using the entire dataset for the regression model was neither practical nor necessary. Instead, I took a random sample of $10{,}000$ observations to use for the regression analysis. I took the same analytic approach as I did for the item parameters except that I did not include the item-level relative response propagation factor. I used a standard normal prior distribution for the mean and standard deviation regression parameters. I estimated the model parameters with 3 chains, $1{,}000$ warm-up iterations, and $2{,}000$ sampling iterations. Again, $\hat{R}$ values were all approximately equal to 1.00 and the trace plots of the posterior distribution of the parameters suggested adequate chain mixing. Table 12 displays model predicted and observed means and standard deviations of the person ability estimate bias distributions. Figures 8 and 9 depict observed and model predicted distributions of the person ability estimate bias. The model slightly overestimates the variance of the bias distributions for nodes 1 and 4 and underestimates the variance for nodes 2 and 3 for most conditions. I suspect this is due to the overrepresentation of near zero values creating a leptokurtic observed bias distribution for which the model, attempting to fit a normal distribution, compensates for with slight biases in the variance. The over- and under-estimation is not severe, and the model appears to provide an adequate description of the observed bias distributions.

Table 12. *True and Estimated Person Ability Model Predicted and Observed Bias Means and Standard Deviations.*

	Test Length	Node	1PL Model				2PL Model			
			Sample Size = 500		Sample Size = 2000		Sample Size = 500		Sample Size = 2000	
			Predicted	Observed	Predicted	Observed	Predicted	Observed	Predicted	Observed
Mean Bias	10	1	0.04	0.00	0.02	0.00	-0.02	0.04	-0.01	0.01
		2	0.05	0.01	0.03	-0.05	-0.02	0.08	0.00	-0.01
		3	0.04	0.04	0.03	-0.01	-0.04	-0.10	-0.01	0.01
		4	0.06	0.10	0.03	0.05	-0.03	0.02	-0.02	0.04
	30	1	-0.02	-0.01	-0.03	-0.01	0.01	-0.04	-0.02	-0.01
		2	0.01	-0.03	-0.02	0.02	0.03	0.03	-0.02	0.00
		3	0.02	-0.09	0.00	-0.04	0.02	-0.05	-0.01	0.00
		4	0.05	0.02	0.03	0.02	0.03	0.04	0.02	-0.01
Standard Deviation of Bias	10	1	0.64	0.59	0.63	0.58	0.54	0.47	0.53	0.52
		2	0.74	0.74	0.72	0.75	0.66	0.72	0.65	0.66
		3	0.86	0.88	0.82	0.87	0.80	0.84	0.79	0.85
		4	0.96	0.94	0.94	0.93	0.95	0.93	0.96	0.92
	30	1	0.41	0.38	0.41	0.41	0.37	0.35	0.35	0.35
		2	0.52	0.54	0.52	0.57	0.48	0.49	0.46	0.47
		3	0.65	0.72	0.66	0.69	0.64	0.67	0.61	0.63
		4	0.84	0.82	0.84	0.81	0.86	0.81	0.81	0.76

Similar to the item difficulty parameters, none of the mean bias predictors, including the intercept, were different from zero in any practical sense. However, posterior predictions suggest that node 1 ability estimates from small to medium samples sizes ($N \leq 1{,}000$) with long tests ($J \geq 100$) are downwardly biased by roughly one-fifth of a standard deviation on average. The regression model did not "learn" from observations that were generated from very long tests. These conditions may be out of the range of valid predictions this model can make. It may also be indicative of the need for an adequate number of respondents in the sample to estimate the large number of item parameters for longer tests.

With regard to estimate variability, the intercept suggests that the average estimate variability for the first node, $\bar{b}_{\sigma[1PL]} = -0.67, 95\%_{\text{HDPI}}[-0.70, -0.65]$, $\bar{b}_{\sigma[2PL]} = -.84, 95\%_{\text{HDPI}}[-0.86, -0.81]$, is approximately 0.51 standard deviations for the 1PL model and 0.43 for the 2PL model. Node depth had a small positive effect on estimate variability, $\bar{b}_{\sigma[1PL]} = 0.19, 95\%_{\text{HDPI}}[0.17, 0.20]$, $\bar{b}_{\sigma[2PL]} = 0.24, 95\%_{\text{HDPI}}[0.22, 0.25]$, such that deeper nodes exhibit greater estimate variability. Test length had a small negative effect on estimate variability, $\bar{b}_{\sigma[1PL]} = -0.21, 95\%_{\text{HDPI}}[-0.23, -0.19]$, $\bar{b}_{\sigma[2PL]} = -0.20, 95\%_{\text{HDPI}}[-0.23, -0.18]$, such that longer tests provide more reliable estimates of the person ability parameters for the first node. The interaction between node depth and test length had a reliably positive but negligible effect on estimate variability, $\bar{b}_{\sigma[1PL]} = 0.05, 95\%_{\text{HDPI}}[0.04, 0.06]$, $\bar{b}_{\sigma[2PL]} = 0.04, 95\%_{\text{HDPI}}[0.03, 0.05]$, such that deeper nodes weaken the uncertainty reducing benefits of longer test lengths. Sample size did not have an effect reliably different from zero. The two-way interactions between node depth and sample size, sample size and test length, and the three-way interaction between node depth, sample size, and test length, were not reliably different from zero. Table 13 provides the results for the regression model. Table 14 provides posterior predictions for the average estimate bias. Figure 9 displays posterior predictions of the estimate bias for each node and condition.

Table 13. *Person Ability Parameter Bias Model Posterior Means and Credibility Intervals of Predictors of Estimate Mean and Variance.*

	1PL Model				2PL Model			
	Estimate Bias		Estimate Variability†		Estimate Bias		Estimate Variability†	
	$\overline{b_{\mu}}$	$95\%HPDI$	$\overline{b_{\sigma}}$	$95\%HPDI$	$\overline{b_{\mu}}$	$95\%HPDI$	$\overline{b_{\sigma}}$	$95\%HPDI$
Intercept	0.00	[-0.02, 0.02]	-0.67	[-0.70, -0.65]	-0.01	[-0.02, 0.01]	-0.84	[-0.86, -0.81]
D	0.01	[0.00, 0.02]	0.19	[0.17, 0.20]	0.00	[-0.01, 0.01]	0.24	[0.22, 0.25]
S	-0.01	[-0.02, 0.01]	-0.01	[-0.03, 0.02]	0.00	[-0.02, 0.01]	-0.01	[-0.04, 0.01]
T	-0.02	[-0.04, 0.00]	-0.21	[-0.23, -0.19]	0.00	[-0.02, 0.02]	-0.20	[-0.23, -0.18]
D x S	0.00	[-0.01, 0.01]	0.00	[-0.01, 0.01]	0.00	[-0.01, 0.01]	0.00	[-0.01, 0.01]
D x T	0.00	[-0.01, 0.02]	0.05	[0.04, 0.06]	0.00	[-0.01, 0.02]	0.04	[0.03, 0.05]
S x T	0.00	[-0.01, 0.02]	0.01	[-0.01, 0.03]	-0.01	[-0.03, 0.01]	0.00	[-0.03, 0.02]
D x S x T	0.00	[-0.02, 0.01]	0.00	[-0.01, 0.01]	0.00	[-0.01, 0.01]	0.00	[-0.02, 0.01]

Note. N = 10,000 randomly drawn from total simulated cases of 4×10^6 for each model. D = node depth, S = sample size (standardized, M_{1PL} = 1,687.25, SD_{1PL} = 609.38; M_{2PL} = 1,706.75, SD_{2PL} = 594.91), T = test length (standardized, M_{1PL} = 19.86, SD_{1PL} = 9.99; M_{2PL} = 20.07, SD_{2PL} = 10.00), HPDI = highest posterior density interval. † Model parameters are on the log scale.

Table 14. *Posterior Prediction Means and Credibility Intervals of 1PL Person Ability Parameter Estimate Bias.*

Node	Test Length	Sample Size = 250		Sample Size = 500		Sample Size = 1,000		Sample Size = 2,000	
		$\overline{\theta^*}$	95% HPDI	$\overline{\theta^*}$	95% HPDI	$\overline{\theta^*}$	95% HPDI	$\overline{\theta^*}$	95% HPDI
1	5	0.06	[-1.35, 1.50]	0.04	[-1.32, 1.53]	0.06	[-1.29, 1.54]	0.02	[-1.26, 1.38]
	10	0.05	[-1.20, 1.36]	0.04	[-1.24, 1.25]	0.04	[-1.27, 1.20]	0.02	[-1.17, 1.27]
	30	0.00	[-0.81, 0.82]	-0.02	[-0.79, 0.81]	-0.02	[-0.85, 0.76]	-0.03	[-0.87, 0.75]
	50	-0.07	[-0.58, 0.45]	-0.07	[-0.57, 0.46]	-0.07	[-0.59, 0.46]	-0.06	[-0.63, 0.45]
	100	-0.23	[-0.63, 0.18]	-0.22	[-0.59, 0.13]	-0.20	[-0.47, 0.08]	-0.14	[-0.40, 0.10]
2	5	0.08	[-1.50, 1.66]	0.06	[-1.55, 1.70]	0.05	[-1.54, 1.61]	0.03	[-1.55, 1.51]
	10	0.05	[-1.45, 1.57]	0.05	[-1.40, 1.49]	0.03	[-1.41, 1.47]	0.03	[-1.27, 1.57]
	30	0.01	[-1.07, 0.99]	0.01	[-0.97, 1.03]	0.01	[-1.01, 1.01]	-0.02	[-1.02, 1.02]
	50	-0.04	[-0.77, 0.65]	-0.03	[-0.73, 0.73]	-0.04	[-0.76, 0.69]	-0.03	[-0.81, 0.73]
	100	-0.14	[-0.53, 0.26]	-0.14	[-0.53, 0.22]	-0.13	[-0.49, 0.21]	-0.11	[-0.48, 0.25]
3	5	0.09	[-1.61, 1.93]	0.04	[-1.70, 1.83]	0.05	[-1.71, 1.79]	0.03	[-1.56, 1.77]
	10	0.05	[-1.49, 1.80]	0.04	[-1.63, 1.71]	0.05	[-1.66, 1.61]	0.03	[-1.58, 1.61]
	30	0.04	[-1.26, 1.33]	0.02	[-1.23, 1.29]	0.03	[-1.22, 1.34]	0.00	[-1.31, 1.28]
	50	0.01	[-1.00, 1.01]	0.00	[-1.05, 1.01]	-0.02	[-1.01, 1.06]	-0.02	[-1.08, 1.04]
	100	-0.05	[-0.69, 0.62]	-0.06	[-0.66, 0.59]	-0.07	[-0.65, 0.57]	-0.09	[-0.75, 0.55]
4	5	0.07	[-1.86, 2.12]	0.06	[-1.82, 2.14]	0.05	[-1.92, 1.96]	0.03	[-1.83, 1.88]
	10	0.07	[-1.81, 2.04]	0.06	[-1.76, 1.95]	0.05	[-1.87, 1.88]	0.03	[-1.91, 1.77]
	30	0.05	[-1.58, 1.65]	0.05	[-1.58, 1.66]	0.04	[-1.63, 1.67]	0.03	[-1.72, 1.57]
	50	0.05	[-1.32, 1.49]	0.03	[-1.46, 1.47]	0.01	[-1.44, 1.46]	-0.01	[-1.40, 1.51]
	100	0.05	[-1.17, 1.21]	0.03	[-1.07, 1.21]	0.01	[-1.11, 1.08]	-0.05	[-1.19, 1.09]

Note. $\overline{\theta^*}$ = mean posterior predicted difficulty estimate bias, HPDI = highest posterior density interval.

Table 15. *Posterior Prediction Means and Credibility Intervals of 2PL Person Ability Parameter Estimate Bias.*

Node	Test Length	Sample Size = 250		Sample Size = 500		Sample Size = 1,000		Sample Size = 2,000	
		$\overline{\theta^*}$	95% HPDI	$\overline{\theta^*}$	95% HPDI	$\overline{\theta^*}$	95% HPDI	$\overline{\theta^*}$	95% HPDI
1	5	-0.04	[-1.23, 1.13]	-0.04	[-1.22, 1.19]	-0.01	[-1.22, 1.16]	0.00	[-1.12, 1.13]
	10	-0.01	[-1.13, 1.02]	-0.02	[-1.11, 0.98]	-0.01	[-1.12, 0.98]	-0.01	[-1.07, 1.01]
	30	0.03	[-0.66, 0.81]	0.01	[-0.69, 0.74]	0.00	[-0.70, 0.72]	-0.02	[-0.72, 0.66]
	50	0.07	[-0.43, 0.59]	0.06	[-0.47, 0.52]	0.02	[-0.44, 0.51]	-0.03	[-0.50, 0.41]
	100	0.18	[-0.19, 0.58]	0.15	[-0.19, 0.50]	0.08	[-0.19, 0.33]	-0.05	[-0.26, 0.17]
2	5	-0.03	[-1.51, 1.27]	-0.04	[-1.47, 1.39]	-0.03	[-1.37, 1.43]	0.00	[-1.41, 1.35]
	10	-0.03	[-1.36, 1.31]	-0.02	[-1.29, 1.26]	-0.02	[-1.29, 1.26]	0.00	[-1.29, 1.25]
	30	0.03	[-0.93, 1.02]	0.03	[-0.90, 0.97]	0.02	[-0.92, 0.94]	-0.02	[-0.87, 0.93]
	50	0.08	[-0.61, 0.81]	0.07	[-0.63, 0.79]	0.04	[-0.63, 0.73]	-0.01	[-0.68, 0.64]
	100	0.23	[-0.23, 0.64]	0.19	[-0.21, 0.60]	0.12	[-0.24, 0.45]	-0.02	[-0.32, 0.29]
3	5	-0.02	[-1.61, 1.68]	-0.06	[-1.72, 1.57]	-0.03	[-1.63, 1.65]	-0.01	[-1.65, 1.58]
	10	-0.04	[-1.50, 1.57]	-0.04	[-1.64, 1.48]	-0.02	[-1.63, 1.45]	-0.01	[-1.55, 1.51]
	30	0.04	[-1.22, 1.32]	0.02	[-1.22, 1.24]	0.02	[-1.17, 1.27]	-0.01	[-1.21, 1.18]
	50	0.10	[-0.98, 1.13]	0.08	[-0.93, 1.16]	0.04	[-0.98, 1.02]	0.01	[-0.93, 0.95]
	100	0.27	[-0.49, 0.99]	0.23	[-0.51, 0.89]	0.16	[-0.42, 0.78]	0.02	[-0.50, 0.55]
4	5	-0.05	[-1.97, 1.91]	-0.05	[-1.99, 1.90]	-0.04	[-2.04, 1.82]	-0.03	[-1.89, 1.95]
	10	-0.03	[-1.90, 1.92]	-0.03	[-1.87, 1.81]	-0.02	[-1.92, 1.82]	-0.02	[-1.92, 1.83]
	30	0.03	[-1.66, 1.67]	0.03	[-1.64, 1.64]	0.03	[-1.62, 1.66]	0.02	[-1.57, 1.57]
	50	0.12	[-1.44, 1.59]	0.09	[-1.47, 1.63]	0.06	[-1.40, 1.49]	0.01	[-1.27, 1.34]
	100	0.32	[-1.11, 1.72]	0.28	[-1.04, 1.52]	0.21	[-0.92, 1.28]	0.07	[-0.86, 0.95]

Note. $\overline{\theta^*}$ = mean posterior predicted difficulty estimate bias, HPDI = highest posterior density interval.

These results suggest that even very long tests cannot produce estimates at deep nodes with low enough estimate uncertainty to distinguish respondents one standard deviation apart on the latent ability continuum. Tasks such as personnel selection become impractical because the estimate uncertainty for the latent abilities at deep nodes is too great. Another task a researcher might conduct involves using the estimated latent ability score to predict some external criterion via correlations or linear regression. To investigate the viability of such a task with IRTree estimated latent ability predictors, I conducted a beta regression to measure the effects of the manipulated and observed factors on the correlation between the true and estimated ability parameters.

Person Ability True and Estimated Parameter Correlation

Regression Model Results

To quantify the effects the manipulated factors had on the correlations between the true and estimated person ability parameters, I conducted a regression analysis with a beta distribution likelihood. The likelihood was parameterized in “location-scale” form, such that $r_{\hat{\theta}\theta} \sim \text{Beta}(\mu\phi, (1-\mu)\phi)$, where the model predictors determine the location, $\mu = \frac{1}{1+e^{-Xb_{[\mu]}}}$, and scale, $\phi = e^{Xb_{[\phi]}}$. The expected distribution mean is $\text{E}[r_{\hat{\theta}\theta}] = \frac{\mu\phi}{\mu\phi+(1-\mu)\phi} = \frac{\mu}{\mu+(1-\mu)}\frac{\phi}{\phi} = \mu$, and the variance is $\text{V}[r_{\hat{\theta}\theta}] = \frac{\mu-\mu^2}{1+\phi}$ such that the uncertainty around the estimated mean is inversely proportional to ϕ. The regression model predictor parameters for μ and ϕ are on the log-odds and log scales, respectively, so I will again rely on posterior predictions of the correlations for interpretation.

I calculated the correlations between the true and estimated person ability parameters by condition, node, and iteration, producing a sample size of 1,600. All of the correlations for the

1PL model estimates were positive. Two correlations were below zero for the 2PL model estimates. One was $r_{\theta\hat{\theta}} = -.03$ and the other was $r_{\theta\hat{\theta}} = -.004$. Both were simulated with 10-item tests, 500 respondents, and were from the fourth node. The domain of the beta distribution is defined between zero and one. In order to incorporate these two observations, I would need to create a mixture of beta distributions and account for the probability of a correlation being negative. Because there are only two negative correlations, and both are from the same simulation conditions, there doesn't seem to be much to learn by using a more complex model. I chose instead to remove these observations from the analysis and use a standard beta distribution.

I used sample size, test length, node depth, and their two-way and three-way interactions as predictors. As before, I centered and standardized the sample size ($M = 1{,}250, SD = 750.23$) and test length ($M = 20.00, SD = 10.00$) factors to facilitate estimation. I attempted to fit several models with partially pooled intercept terms grouped by simulation iteration for the location and scale parameters. All models produced a high proportion of divergent transitions during estimation and resulted in poorly mixed chains. I instead used a model with completely pooled intercepts for the location and scale parameters with three chains and $2{,}000$ sampling iterations each. Neither model displayed issues with estimation, the chains appear to have mixed appropriately and all $\hat{R}$ values rounded to 1.00, all of which suggest adequate sampling from the posterior. Table 16 displays model predicted and observed correlation means and standard deviations and Figures 14 and 15 display the observed and model predicted correlation distributions. The model predicts the marginal statistics quite well and appears to approximate the observed correlation distributions adequately.

Table 16. *True and Estimated Person Ability Model Predicted and Observed Correlation Means and Standard Deviations.*

			1PL Model				2PL Model			
	Test Length	Node	Sample Size = 500		Sample Size = 2000		Sample Size = 500		Sample Size = 2000	
			Predicted	Observed	Predicted	Observed	Predicted	Observed	Predicted	Observed
Mean Correlation	10	1	.80	.80	.80	.80	.85	.85	.85	.86
		2	.68	.66	.68	.67	.72	.72	.74	.74
		3	.53	.53	.53	.53	.55	.54	.58	.57
		4	.37	.39	.38	.40	.36	.37	.41	.42
	30	1	.92	.92	.91	.92	.94	.94	.94	.94
		2	.84	.84	.84	.83	.88	.88	.88	.88
		3	.73	.72	.72	.72	.77	.77	.78	.77
		4	.57	.59	.56	.58	.61	.61	.63	.63
Standard Deviation of Correlation	10	1	.02	.02	.02	.01	.05	.06	.04	.04
		2	.03	.04	.02	.03	.08	.08	.06	.06
		3	.05	.05	.04	.04	.11	.11	.08	.09
		4	.06	.06	.05	.04	.14	.13	.10	.09
	30	1	.01	.01	.01	.00	.01	.01	.01	.01
		2	.01	.01	.01	.01	.02	.02	.02	.02
		3	.03	.03	.02	.02	.04	.04	.04	.04
		4	.05	.04	.04	.03	.07	.06	.06	.05

The location intercept was reliably positive, $\bar{b}_{\mu[1PL]} = 1.87, 95\%_{\text{HDPI}}[1.87,\ 1.87]$, $\bar{b}_{\mu[2PL]} = 2.24, 95\%_{\text{HDPI}}[2.22,\ 2.26]$, as was the scale intercept, $\bar{b}_{\phi[1PL]} = 6.93, 95\%_{\text{HDPI}}[6.93,\ 6.93]$, $\bar{b}_{\phi[2PL]} = 5.45, 95\%_{\text{HDPI}}[5.32,\ 5.57]$. The node depth factor had a strong, negative, and reliable relationship with the location, $\bar{b}_{\mu[1PL]} = -0.67, 95\%_{\text{HDPI}}[-0.67,\ -0.67]$, $\bar{b}_{\mu[2PL]} = -0.75, 95\%_{\text{HDPI}}[-0.76,\ -0.73]$, and scale parameters, $\bar{b}_{\phi[1PL]} = -0.79, 95\%_{\text{HDPI}}[-0.79,\ -0.79]$, $\bar{b}_{\phi[2PL]} = -0.69,\ 95\%_{\text{HDPI}}[-0.76,\ -0.62]$, such that the correlation declines rapidly and increases in uncertainty at greater depths. Test length had a strong, positive, and reliable effect on both the location, $\bar{b}_{\mu[1PL]} = 0.50, 95\%_{\text{HDPI}}[0.50,\ 0.51]$, $\bar{b}_{\mu[2PL]} = 0.51, 95\%_{\text{HDPI}}[0.48,\ 0.53]$, and scale parameters, $\bar{b}_{\phi[1PL]} = 0.68, 95\%_{\text{HDPI}}[0.67,\ 0.68]$, $\bar{b}_{\phi[2PL]} = 1.23, 95\%_{\text{HDPI}}[1.10,\ 1.35]$, such that longer test lengths provide stronger and more consistent correlations between the true and estimated ability parameters. Sample size reliably and positively affected the precision of the correlation, $\bar{b}_{\phi[1PL]} = 0.29, 95\%_{\text{HDPI}}[0.29,\ 0.29]$, $\bar{b}_{\phi[2PL]} = 0.20, 95\%_{\text{HDPI}}[0.08,\ 0.34]$, but the effect was small and did not produce practically different predictions across different sample sizes. The test length effects on the location of the correlations for the 1PL model, $\bar{b}_{\mu[1PL]} = -0.04, 95\%_{\text{HDPI}}[-0.04,\ -0.04]$, and the scale of the correlations for both 1PL, $\bar{b}_{\phi[1PL]} = -0.13, 95\%_{\text{HDPI}}[-0.13,\ -0.13]$, and 2PL models, $\bar{b}_{\phi[2PL]} = -0.23, 95\%_{\text{HDPI}}[-0.29,\ -0.16]$, were reliably moderated by node depth, but the moderating effects were negligible.

Table 18 displays posterior predictions of correlations for both the 1PL and 2PL regression models. The posterior predictions suggest that medium to long tests are required to achieve correlations of $r_{\theta\hat{\theta}} \geq .90$ for the first two nodes, and long to very long tests are required for nodes three or four. Sample size does not appear to affect the magnitude or the precision of the correlation to a meaningful degree.

Table 17. *Person Ability Correlations – Posterior Means and Credibility Intervals of Predictors of Estimate Location and Scale.*

	1PL Model				2PL Model			
	Estimate Bias		Estimate Variability†		Estimate Bias		Estimate Variability†	
	$\overline{b_{\mu}}$	$95\%HPDI$	$\overline{b_{\sigma}}$	$95\%HPDI$	$\overline{b_{\mu}}$	$95\%HPDI$	$\overline{b_{\sigma}}$	$95\%HPDI$
Intercept	1.87	[1.87, 1.87]	6.93	[6.93, 6.93]	2.24	[2.22, 2.26]	5.45	[5.32, 5.57]
S	0.00	[0.00, 0.00]	0.29	[0.29, 0.29]	0.02	[-0.01, 0.04]	0.20	[0.08, 0.34]
T	0.50	[0.50, 0.51]	0.68	[0.67, 0.68]	0.51	[0.48, 0.53]	1.23	[1.10, 1.35]
D	-0.67	[-0.67, -0.67]	-0.79	[-0.79, -0.79]	-0.75	[-0.76, -0.73]	-0.69	[-0.76, -0.62]
S x T	0.00	[0.00, 0.00]	-0.02	[-0.03, -0.02]	-0.02	[-0.04, 0.01]	-0.07	[-0.19, 0.06]
S x D	0.00	[0.00, 0.00]	-0.03	[-0.04, -0.03]	0.02	[0.00, 0.03]	0.01	[-0.06, 0.08]
T x D	-0.04	[-0.04, -0.04]	-0.13	[-0.13, -0.13]	-0.01	[-0.03, 0.00]	-0.23	[-0.29, -0.16]
S x T x D	0.00	[0.00, 0.00]	-0.01	[-0.01, -0.01]	0.00	[-0.02, 0.01]	-0.03	[-0.09, 0.04]

Note. N = 10,000 randomly drawn from total simulated cases of 4×10^{6} for each model. D = node depth, S = sample size (standardized, $M_{[1PL]}$ = 1,687.25, $SD_{[1PL]}$ = 609.38; $M_{[2PL]}$ = 1,706.75, $SD_{[2PL]}$ = 594.91), T = test length (standardized, $M_{[1PL]}$ = 19.86, $SD_{[1PL]}$ = 9.99; $M_{[2PL]}$ = 20.07, $SD_{[2PL]}$ = 10.00), HPDI = highest posterior density interval. † Model parameters are on the log scale.

Table 18. *Posterior Prediction of Correlations between Person Ability Parameter Estimates and True Values.*

		1PL Model				2PL Model			
Node	Test Length	Sample Size = 500		Sample Size = 2,000		Sample Size = 500		Sample Size = 2,000	
		$\overline{r_{\hat{\theta}\theta}}$	95% HPDI	$\overline{r_{\hat{\theta}\theta}}$	95% HPDI	$\overline{r_{\hat{\theta}\theta}}$	95% HPDI	$\overline{r_{\hat{\theta}\theta}}$	95% HPDI
1	5	.75	[.70, .80]	.75	[.72, .79]	.81	[.67, .95]	.82	[.71, .91]
	10	.80	[.76, .84]	.80	[.77, .83]	.85	[.75, .94]	.85	[.78, .92]
	30	.92	[.90, .93]	.91	[.90, .93]	.94	[.92, .96]	.94	[.92, .96]
	50	.97	[.96, .97]	.97	[.96, .97]	.98	[.97, .98]	.98	[.97, .98]
	100	1.00	[1.00, 1.00]	1.00	[1.00, 1.00]	1.00	[1.00, 1.00]	1.00	[1.00, 1.00]
2	5	.62	[.55, .70]	.63	[.57, .69]	.67	[.46, .87]	.69	[.55, .83]
	10	.68	[.61, .74]	.68	[.63, .73]	.72	[.56, .86]	.74	[.62, .85]
	30	.84	[.81, .87]	.84	[.82, .86]	.88	[.84, .92]	.88	[.85, .91]
	50	.93	[.92, .94]	.93	[.92, .94]	.95	[.94, .96]	.95	[.94, .96]
	100	.99	[.99, .99]	.99	[.99, .99]	.00	[1.00, 1.00]	1.00	[.99, 1.00]
3	5	.47	[.37, .58]	.48	[.40, .56]	.49	[.23, .75]	.53	[.35, .70]
	10	.53	[.43, .62]	.53	[.46, .61]	.55	[.33, .76]	.58	[.44, .73]
	30	.72	[.67, .78]	.72	[.68, .77]	.77	[.69, .84]	.78	[.71, .85]
	50	.86	[.84, .89]	.86	[.83, .88]	.90	[.88, .92]	.90	[.87, .92]
	100	.98	[.98, .99]	.98	[.98, .98]	.99	[.99, .99]	.99	[.99, .99]
4	5	.33	[.20, .45]	.34	[.24, .45]	.31	[.04, .60]	.36	[.16, .56]
	10	.37	[.25, .50]	.38	[.28, .48]	.36	[.11, .62]	.41	[.22, .59]
	30	.57	[.48, .66]	.56	[.48, .65]	.61	[.46, .75]	.63	[.49, .75]
	50	.74	[.68, .80]	.73	[.68, .79]	.81	[.75, .87]	.80	[.73, .87]
	100	.95	[.94, .97]	.95	[.93, .96]	.98	[.97, .99]	.97	[.96, .99]

Note. $\overline{r_{\hat{\theta}\theta}}$ = mean posterior predicted correlation, HPDI = highest posterior density interval.

To provide some perspective, the extent to which an estimated parameter correlates with its true data generating parameter provides an upper bound for valid correlations the estimate could have with external criteria. To demonstrate this issue in practical terms, consider a common task where a set of ability parameters estimated from an IRT model are used to correlate with and predict some external criteria, say job performance JP_θ. I want to use an IRTree model to measure some latent ability, A_θ that I believe predicts JP_θ. After collecting data in a validation study from a sample of respondents, I run the IRTree model and produce estimates of their latent ability $\hat{A}_\theta$. For simplicity, I assume (unrealistically) that JP_θ is known or measured without error and both variables are multivariate normally distributed. The data generating correlation between JP_θ and A_θ is $\rho_{JP_\theta A_\theta}$ and is estimated by correlating JP_θ and $\hat{A}_\theta$, producing a measure of the linear relationship of interest, $r_{JP_\theta \hat{A}_\theta}$. Although not known in practice, the difference (bias) between $\rho_{JP_\theta A_\theta}$ and $r_{JP_\theta \hat{A}_\theta}$, or $r^*_{JP_\theta \hat{A}_\theta}$, has practical implications for whether the ability estimates can be used for prediction. Moderate bias may prevent inferences about the strength of the relationship. Larger bias may prevent inferences about the direction of the relationship. The data generating correlation between A_θ and $\hat{A}_\theta$ is $\rho_{A_\theta \hat{A}_\theta}$.

For 6,000 iterations, I simulated values for $\rho_{JP_\theta A_\theta}$ from a beta distribution with $\alpha = 1$ and $\beta = 1$, ensuring a positive relationship uniformly distributed between zero and one to cover a wide range of possible criterion-related correlations. I simulated person ability parameters for JP_θ and A_θ from each true correlation $\rho_{JP_\theta A_\theta}$ with multivariate standard normal distributions. I drew the true and estimated ability correlation, $r_{A_\theta \hat{A}_\theta}$, from a beta distribution with mean and precision, μ and ϕ, using posterior draws of the regression model weights b_μ and b_ϕ, respectively, across a range of sample sizes, test lengths, and node depths. I then simulated $\hat{A}_\theta$

using model predictions for $r_{A_\theta \hat{A}_\theta}$ from a normal distribution with a mean of $A_\theta \cdot r_{A_\theta \hat{A}_\theta}$ and standard deviation of $\sqrt{1 - r^2_{A_\theta \hat{A}_\theta}}$ so that the resulting distribution had a standard deviation of approximately 1. Finally, I calculated the correlations between true job performance and the ability estimate, $r_{JP_\theta \hat{A}_\theta}$, as well as the bias (difference), $r^*_{JP_\theta \hat{A}_\theta}$, between the true and estimated correlation for each iteration. The simulation equations can be represented as,

$$\rho_{JP_\theta A_\theta} \sim Beta(1, 1)$$

$$\Sigma_{JP_\theta A_\theta} = \begin{bmatrix} 1 & \rho_{JP_\theta A_\theta} \\ \rho_{JP_\theta A_\theta} & 1 \end{bmatrix}$$

$$\begin{bmatrix} JP_\theta \\ A_\theta \end{bmatrix} \sim MVN\left(0, \Sigma_{JP_\theta A_\theta}\right)$$

$$\mu = \frac{1}{1 + e^{b_\mu X}}$$

$$\phi = e^{b_\phi X}$$

$$r_{A_\theta \hat{A}_\theta} \sim Beta(\mu\phi, (1 - \mu)\phi)$$

$$\hat{A}_\theta \sim Normal\left(A_\theta \cdot r_{A_\theta \hat{A}_\theta}, \sqrt{1 - r^2_{A_\theta \hat{A}_\theta}}\right)$$

$$r^*_{JP_\theta \hat{A}_\theta} = r_{JP_\theta \hat{A}_\theta} - \rho_{JP_\theta A_\theta}.$$

Tables 19 and 20 display the medians and 95% HPDI credibility intervals for $r^*_{JP_\theta \hat{A}_\theta}$ across various conditions for both 1PL and 2PL models. Table 21 displays correlation 95% confidence interval coverage rates using Fisher's *z* transformation. Figures 16 and 17 display the resultant distributions of $r^*_{JP_\theta \hat{A}_\theta}$ across various conditions. The dashed vertical lines at zero represent perfect estimation of the data generating correlation $\rho_{JP_\theta A_\theta}$, whereas lower values indicate estimated correlations below the true data generating correlation.

Table 19. *Posterior Prediction Medians and Credibility Intervals of Marginal Bias Between True and Estimated Correlations of 1PL Person Ability Parameter Estimates and Criterion Variable.*

Node	Test Length	Sample Size = 250		Sample Size = 500		Sample Size = 1,000		Sample Size = 2,000	
		$\overline{r^*_{\hat{\theta}\theta}}$	95% HPDI	$\overline{r^*_{\hat{\theta}\theta}}$	95% HPDI	$\overline{r^*_{\hat{\theta}\theta}}$	95% HPDI	$\overline{r^*_{\hat{\theta}\theta}}$	95% HPDI
1	5	-.12	[-.28, .05]	-.12	[-.27, .03]	-.12	[-.26, .01]	-.12	[-.25, .00]
	10	-.10	[-.24, .06]	-.10	[-.22, .03]	-.10	[-.21, .02]	-.10	[-.21, .00]
	30	-.04	[-.14, .07]	-.04	[-.11, .05]	-.04	[-.10, .03]	-.04	[-.09, .01]
	50	-.02	[-.11, .08]	-.02	[-.08, .06]	-.02	[-.06, .04]	-.02	[-.05, .02]
	100	.00	[-.10, .10]	.00	[-.07, .07]	.00	[-.05, .05]	.00	[-.04, .03]
2	5	-.19	[-.42, .04]	-.19	[-.41, .01]	-.19	[-.39, .01]	-.19	[-.37, .01]
	10	-.16	[-.36, .05]	-.16	[-.34, .03]	-.16	[-.34, .01]	-.16	[-.33, .00]
	30	-.08	[-.20, .06]	-.08	[-.18, .03]	-.08	[-.17, .02]	-.08	[-.16, .01]
	50	-.03	[-.13, .08]	-.04	[-.10, .05]	-.03	[-.09, .03]	-.04	[-.08, .02]
	100	.00	[-.10, .10]	.00	[-.07, .07]	.00	[-.05, .04]	.00	[-.04, .03]
3	5	-.27	[-.56, .03]	-.27	[-.54, .03]	-.26	[-.53, .02]	-.26	[-.51, .00]
	10	-.24	[-.52, .03]	-.24	[-.50, .02]	-.23	[-.48, .01]	-.24	[-.46, .00]
	30	-.14	[-.31, .05]	-.14	[-.29, .03]	-.14	[-.29, .02]	-.14	[-.28, .01]
	50	-.07	[-.18, .06]	-.07	[-.16, .04]	-.07	[-.15, .02]	-.07	[-.15, .01]
	100	-.01	[-.10, .09]	-.01	[-.07, .07]	-.01	[-.06, .04]	-.01	[-.04, .03]
4	5	-.34	[-.71, .03]	-.34	[-.69, .01]	-.33	[-.67, .00]	-.33	[-.66, .00]
	10	-.32	[-.67, .02]	-.31	[-.64, .02]	-.31	[-.63, .00]	-.31	[-.62, .00]
	30	-.22	[-.47, .04]	-.22	[-.45, .03]	-.22	[-.44, .01]	-.22	[-.44, .00]
	50	-.13	[-.30, .04]	-.13	[-.28, .02]	-.13	[-.28, .01]	-.14	[-.27, .01]
	100	-.02	[-.12, .09]	-.02	[-.09, .06]	-.02	[-.07, .03]	-.03	[-.07, .02]

Note. $\overline{r^*_{\hat{\theta}\theta}}$ = median posterior predicted bias of correlation, HPDI = highest posterior density interval.

Table 20. *Posterior Prediction Medians and Credibility Intervals of Marginal Bias Between True and Estimated Correlations of 2PL Person Ability Parameter Estimates and Criterion Variable.*

Node	Test Length	Sample Size = 250		Sample Size = 500		Sample Size = 1,000		Sample Size = 2,000	
		$\overline{r^*_{\hat{\theta}\theta}}$	95% HPDI	$\overline{r^*_{\hat{\theta}\theta}}$	95% HPDI	$\overline{r^*_{\hat{\theta}\theta}}$	95% HPDI	$\overline{r^*_{\hat{\theta}\theta}}$	95% HPDI
1	5	-.09	[-.28, .07]	-.09	[-.25, .05]	-.09	[-.23, .03]	-.09	[-.22, .01]
	10	-.08	[-.22, .08]	-.08	[-.20, .05]	-.07	[-.19, .03]	-.07	[-.17, .02]
	30	-.04	[-.12, .08]	-.04	[-.10, .05]	-.03	[-.08, .03]	-.03	[-.07, .02]
	50	-.01	[-.11, .09]	-.01	[-.08, .06]	-.01	[-.06, .04]	-.01	[-.04, .03]
	100	.00	[.10, .09]	.00	[-.07, .07]	.00	[-.05, .05]	.00	[-.03, .04]
2	5	-.16	[-.43, .06]	-.15	[-.40, .06]	-.15	[-.38, .02]	-.14	[-.34, .01]
	10	-.14	[-.36, .06]	-.14	[-.34, .04]	-.13	[-.31, .03]	-.12	[-.28, .02]
	30	-.07	[-.18, .06]	-.07	[-.15, .04]	-.06	[-.14, .02]	-.06	[-.13, .01]
	50	-.03	[-.12, .08]	-.03	[-.09, .05]	-.03	[-.07, .04]	-.03	[-.06, .02]
	100	.00	[-.10, .10]	.00	[-.07, .06]	.00	[-.05, .05]	.00	[-.04, .03]
3	5	-.24	[-.62, .05]	-.24	[-.59, .03]	-.24	[-.56, .02]	-.23	[-.50, .02]
	10	-.22	[-.54, .05]	-.21	[-.51, .03]	-.22	[-.47, .02]	-.20	[-.44, .02]
	30	-.12	[-.28, .04]	-.12	[-.26, .03]	-.11	[-.25, .02]	-.11	[-.24, .01]
	50	-.06	[-.15,.07]	-.06	[-.13, .04]	-.05	[-.12, .03]	-.05	[-.11, .02]
	100	-.01	[-.11, .09]	-.01	[-.08, .06]	-.01	[-.05, .05]	-.01	[-.04, .03]
4	5	-.33	[-.78, .03]	-.33	[-.77, .02]	-.32	[-.72, .01]	-.31	[-.67, .00]
	10	-.31	[-.73, .03]	-.31	[-.72, .01]	-.29	[-.67, .02]	-.29	[-.61, .01]
	30	-.20	[-.45, .06]	-.20	[-.43, .03]	-.19	[-.42, .01]	-.18	[-.40, .01]
	50	-.10	[-.24, .06]	-.10	[-.22, .04]	-.10	[-.22, .02]	-.10	[-.22, .01]
	100	-.01	[-.11, .09]	-.01	[-.07, .07]	-.01	[-.06, .04]	-.02	[-.05, .03]

Note. $\overline{r^*_{\hat{\theta}\theta}}$ = median posterior predicted bias of correlation, HPDI = highest posterior density interval.

Figure 16. *Posterior Predictions of Marginal Bias Between True and Estimated Correlations of 1PL Person Ability Parameter Estimates and Criterion Variable.*

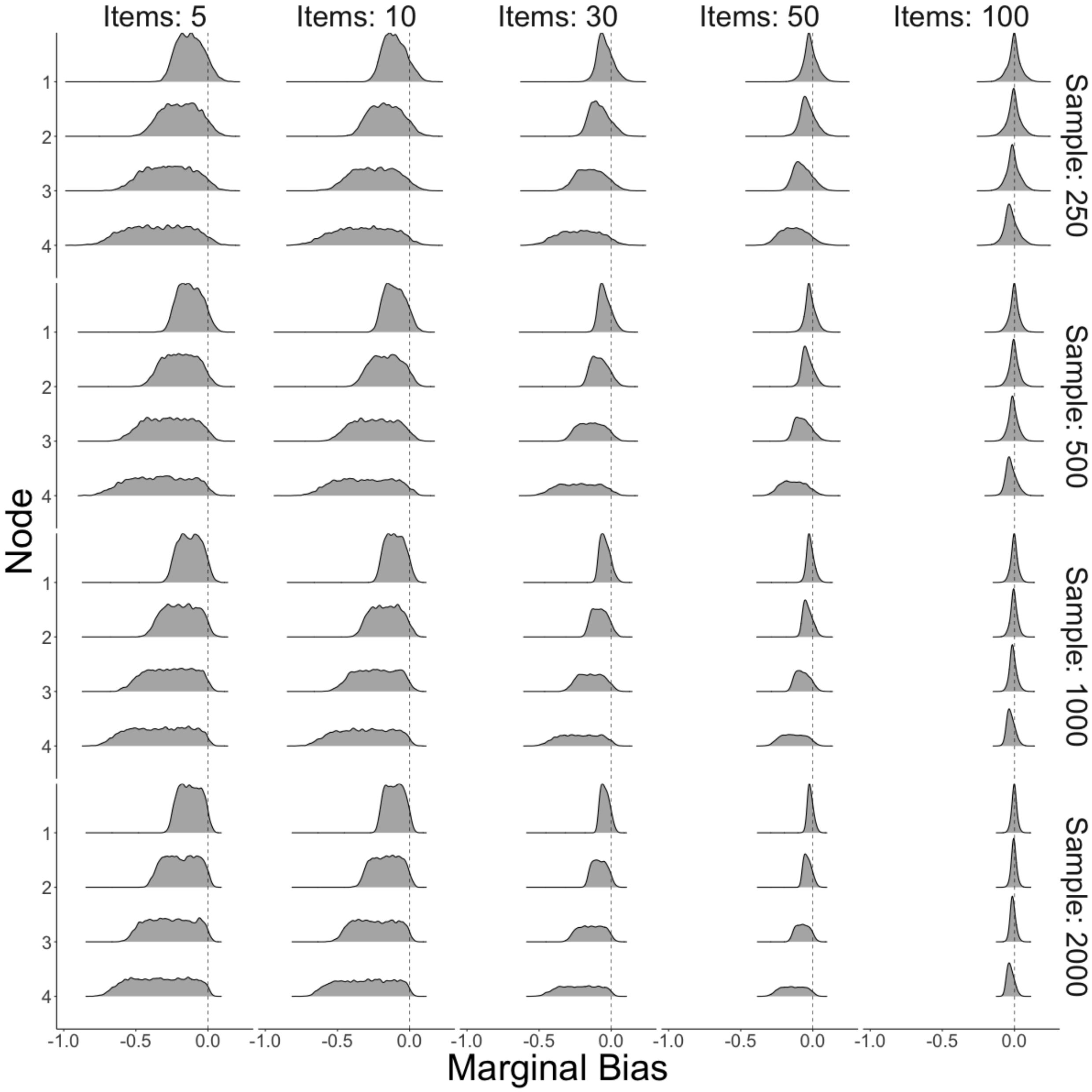


Figure 17. *Posterior Prediction Density Plots of Marginal Bias Between True and Estimated Correlations of 2PL Person Ability Parameter Estimates and Criterion Variable*.

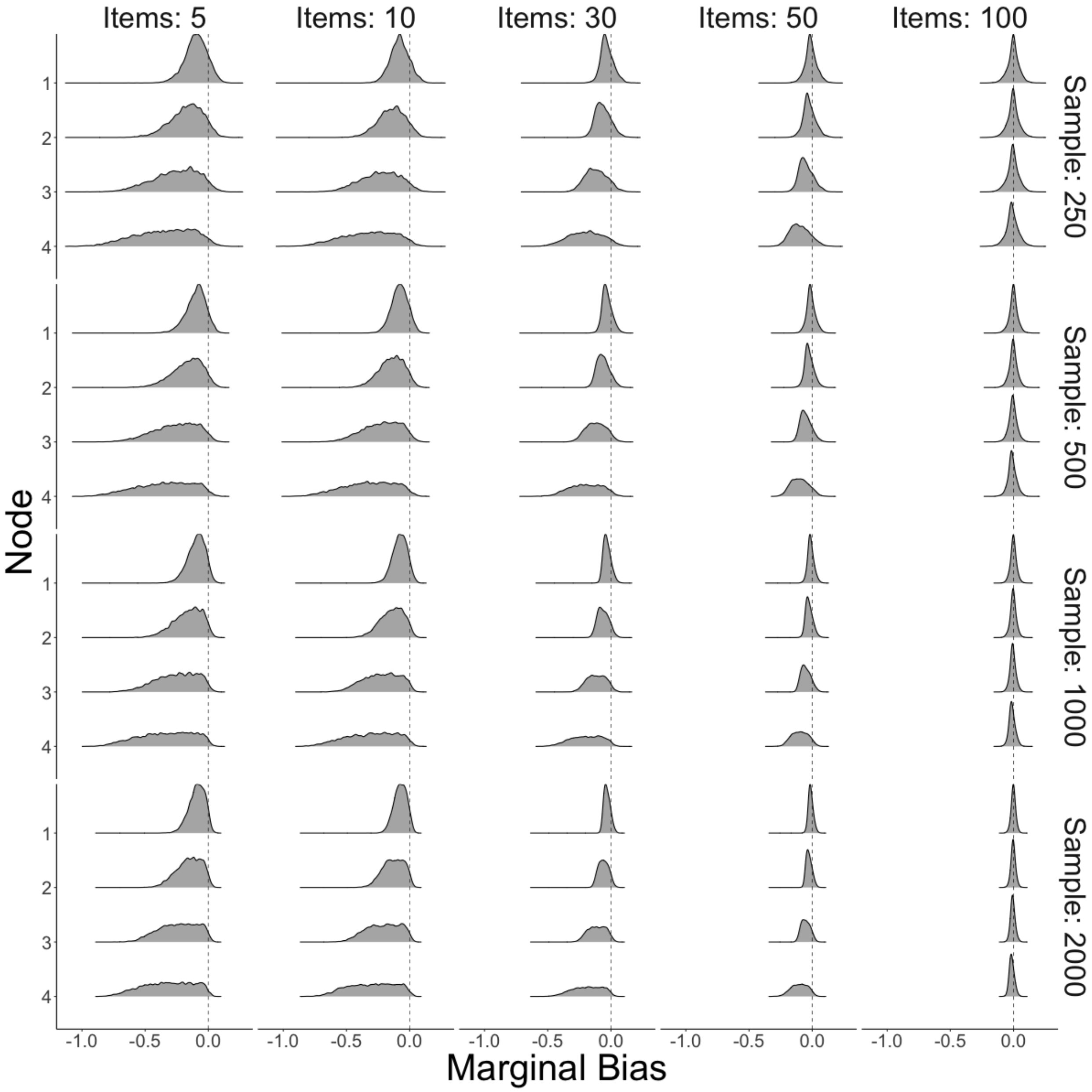


Table 21. Person Ability *Predictive Validity Correlation Estimate Confidence Interval Coverage Rates.*

Node	Test Length	1PL Model				2PL Model			
		Coverage by Sample Size				Coverage by Sample Size			
		250	500	1,000	2,000	250	500	1,000	2,000
1	5	.48	.40	.32	.24	.40	.32	.23	.17
	10	.53	.46	.36	.28	.46	.36	.27	.20
	30	.74	.66	.59	.49	.67	.58	.51	.39
	50	.86	.82	.78	.71	.82	.77	.72	.65
	100	.94	.94	.93	.92	.93	.93	.92	.92
2	5	.35	.26	.19	.14	.31	.22	.16	.11
	10	.39	.30	.24	.17	.34	.24	.19	.14
	30	.60	.50	.40	.32	.51	.43	.33	.25
	50	.78	.71	.64	.54	.72	.64	.55	.44
	100	.94	.92	.91	.89	.92	.91	.89	.85
3	5	.24	.19	.14	.10	.22	.17	.12	.08
	10	.28	.20	.15	.11	.25	.18	.13	.10
	30	.42	.33	.25	.18	.39	.29	.21	.16
	50	.63	.54	.47	.35	.58	.47	.37	.27
	100	.91	.89	.88	.81	.88	.84	.80	.74
4	5	.19	.14	.09	.07	.18	.14	.10	.07
	10	.20	.14	.10	.07	.20	.14	.10	.07
	30	.30	.21	.16	.12	.26	.20	.15	.10
	50	.48	.37	.28	.21	.39	.30	.22	.16
	100	.87	.84	.78	.68	.79	.72	.62	.52

Note. Coverage indicates the rate that the true latent correlation between JP_θ and A_θ lies between the lower and upper correlation estimate 95% confidence interval bounds. The correlation estimate confidence intervals were calculated using Fisher's *z* transformation.

The predictions suggest that the 2PL model produces slightly less biased and slightly more variable ability estimates for prediction, but both the 1PL and 2PL lead to qualitatively similar conclusions. Because $r^*_{JP_\theta \hat{A}_\theta}$ is attenuated, the positive predictive validity coefficients are downwardly biased in most cases. Short test lengths are severely downwardly biased for depths 1 and deeper, and short to medium test lengths are severely downwardly biased for depths 2 and deeper. Also note that the absolute value of the lower bound of the prediction credibility intervals serve as minimum values for the true relationship $\rho_{JP_\theta A_\theta}$ in order to establish directionality of the relationship with 95% confidence (assuming no estimation uncertainty for $r_{JP_\theta \hat{A}_\theta}$). Short test lengths prohibit using anything beyond the second node in an IRTree model for tests of directionality, let alone inference about the magnitude of the relationship. Directional tests for deeper nodes appear feasible for average test lengths if the underlying relationship is believed to be strong, but inferences involving the magnitudes is still questionable. Very long test lengths appear to provide adequate estimates for practical inference under all models and sample sizes. These are optimistic predictions as they do not incorporate measurement error for JP_θ and $r_{JP_\theta \hat{A}_\theta}$. If I incorporate estimate uncertainty for $r_{JP_\theta \hat{A}_\theta}$ (Table 21), coverage rates for 95% confidence intervals are abysmal for most practical testing conditions. Deeper nodes lead to lower coverage rates and longer tests lead to higher coverage rates. Counter-intuitively, larger sample sizes lead to lower coverage rates. This is because the correlation estimates confidence intervals narrow with larger sample sizes, but the correlation estimates are still systemically downwardly biased. The 1PL coverages rates are also slightly higher than the 2PL coverage rates.

To achieve adequate estimation for predictive validity, long test lengths are required for the second node in IRTree models, and very long test lengths are required (100+ items) for deeper nodes. The number of items needed is compounded if the study design requires explicit responses to each node rather than simply recoding rating-scale items to IRTree auxiliary-items. A three-node response survey design would require 60 items per node across 3 nodes for a total of 180 items. This poses a serious challenge to test administration and the practical use of IRTree models. I used orthogonal latent factors in the current study and estimated them as such. Correlations between IRTree person abilities and some external criterion may be better estimated if the criterion responses are incorporated into the model and the latent correlations estimated from the model rather than correlating posterior predictions of the latent ability scores separately. This is useful if a researcher's goal is to test hypotheses about the relationship between latent traits. This is not very useful for prediction which requires estimation of individual scores on the latent traits. Additionally, because the simulation assumed uncorrelated latent person abilities, the probability of a respondent having an inestimable latent ability score at a given node is uncorrelated with their latent abilities used for responding to ancestor nodes. This is not the case for correlated latent abilities. If the latent abilities are all positively correlated, persons with low standing on the latent abilities are more likely to exhibit inestimable latent ability scores for deeper nodes. This means that the true and estimated ability parameter correlations, and subsequent criterion-related correlations, will be further biased than what is presented here due to additional selection bias.

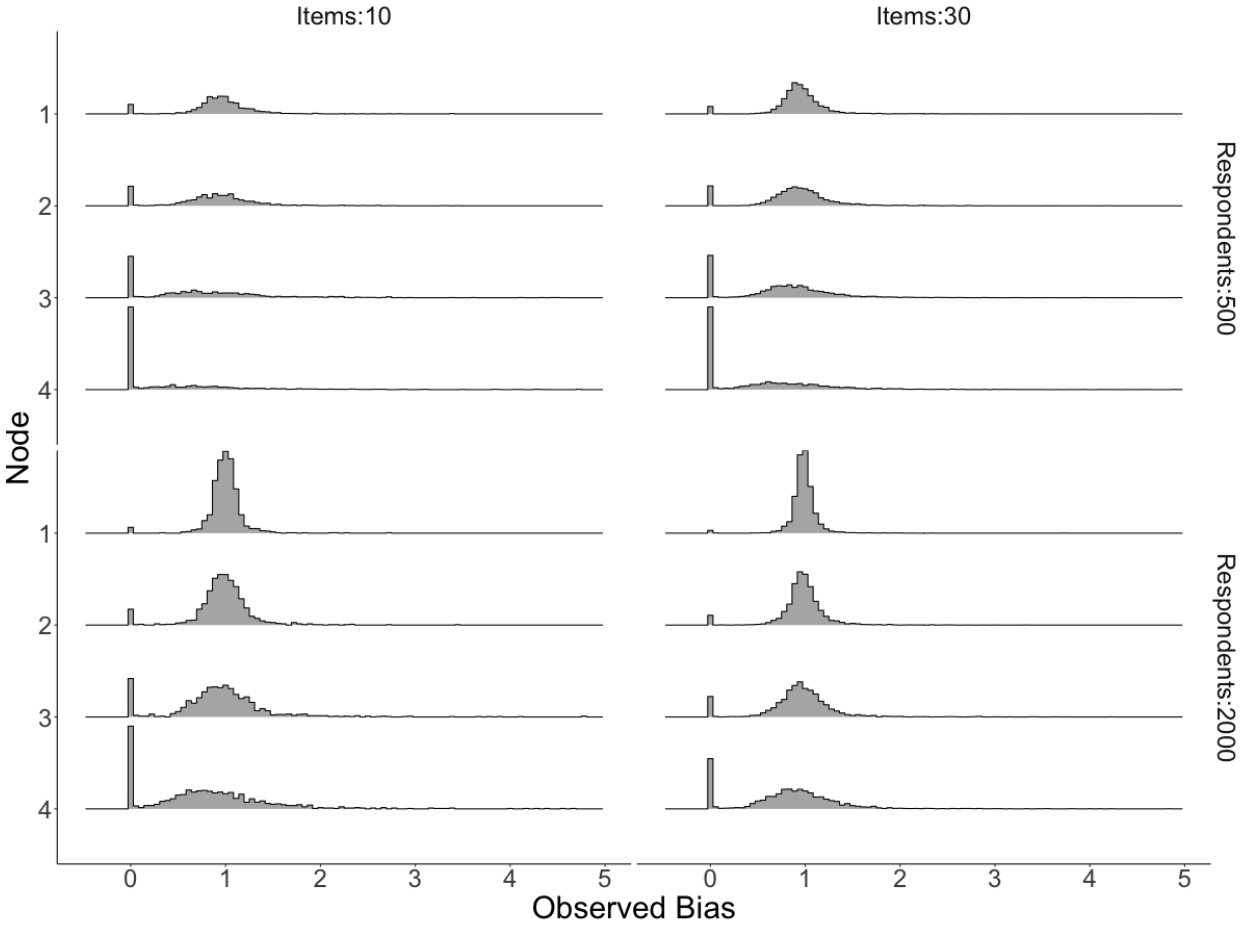
The results suggest that, although estimates of item difficulty and person ability parameters are unbiased, the parameter variability increases under conditions that reduce the number of observations to estimate these parameters. In many cases where tree lengths are long, sample sizes are small, and test lengths are short, the resulting estimates prohibit many tasks such as item selection or estimation of latent abilities for personnel selection or prediction.

Item Discrimination Parameters

Descriptive analysis

Modeling the bias of the discrimination parameters posed several problems. First, the true discrimination parameters are log-normally distributed. The mean and variance parameters of a log-normal distribution are on the log scale. Researchers and practitioners using the 2PL model would likely be more interested in the marginal bias, or the bias of the observed distribution, rather than the bias of the underlying parameters. Taking the difference between the true and estimated parameters to get the marginal bias would result in a distribution without an obvious functional form, exhibiting high variance and kurtosis. Fitting a normal distribution to the marginal bias would be inappropriate. Instead, I calculated bias as the ratio of estimated and true discrimination parameters. The ratio of two log-normally distributed variables is itself log-normally distributed. I chose to use a distributional model with a log-normal likelihood to model the bias. After that, I used posterior predictions to estimate the bias of the marginal distribution for a practical assessment. The second issue was that the discrimination estimates produced a mixture of what appears to be relatively unbiased log-normally distributed estimates and another set of estimates shrunk to near zero without a clear functional form. Calculating the bias results in a similar mixture (Figure 18). Seven estimates (0.02%) were too small for the precision of the computer I used and registered as zeros. Modeling this as if it were a regular log-normally distributed variable would likely severely bias inferences from the distributional regression model. The mean would be downwardly biased and the variance upwardly biased. Instead, I initially removed these seven estimates and tried to model this as a mixture of two lognormal distributions with the predictors predicting the mean, variance, and mixture probabilities. I was unable to produce an adequately estimated model after several attempts to reparametrize and simplify it.

Figure 18. *Histograms of Observed 2PL Item Discrimination Estimate Bias*.

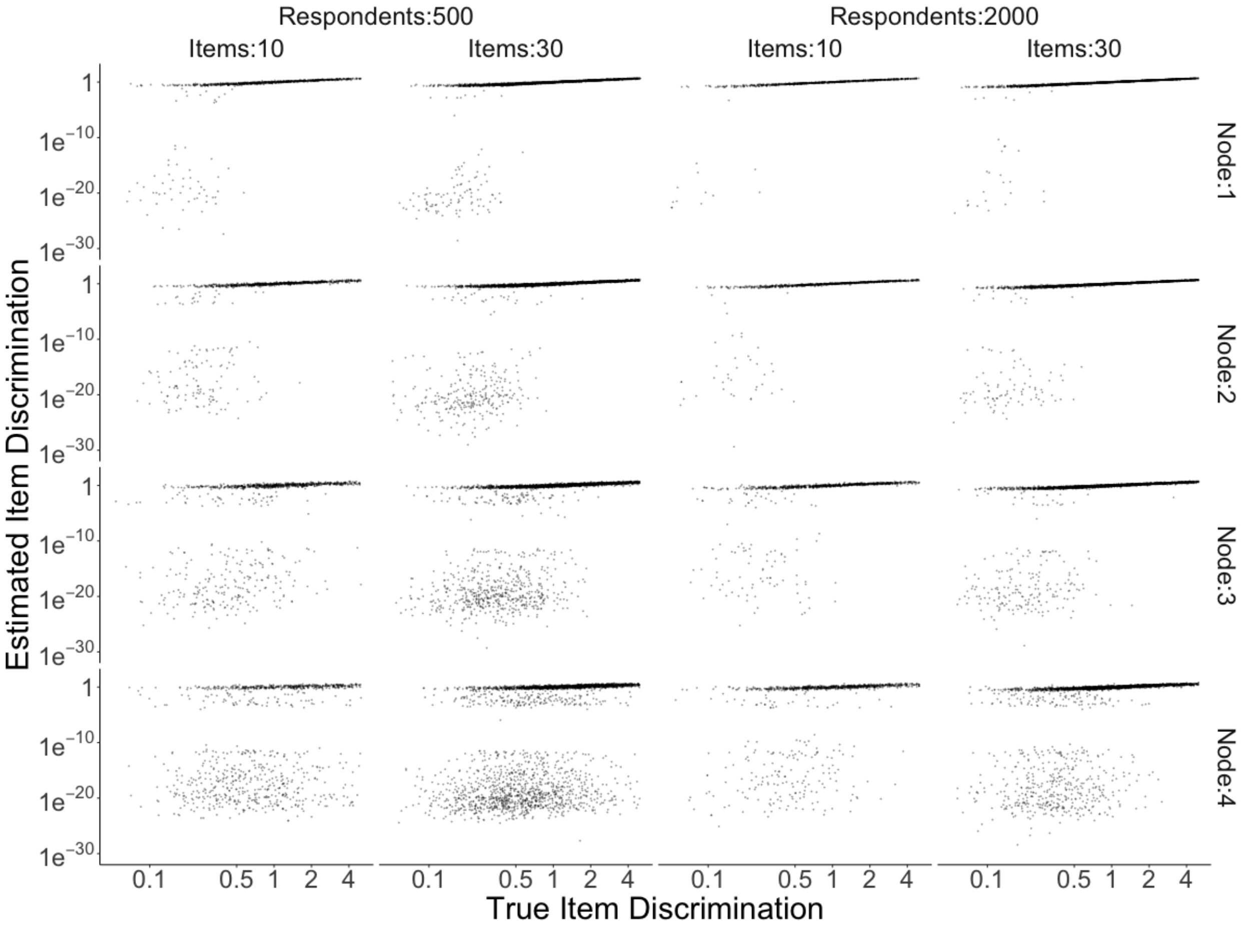


I decided to remove estimates that were shrunk to zero. I inspected a histogram of the log of the bias estimates (Figure 19) and determined that there was a distinct break in the values around e^{-5}. I chose to remove observations where the estimate fell below this threshold. This is an arbitrary cutoff but should afford less biased measurement of the estimates that were not shrunk to near zero. Table 22 displays the proportions of observations that fall below this criterion. A few things should be noted about these shrunken values. These shrunken values occurred under all conditions. Deeper nodes, lower observation propagation, smaller sample sizes, and shorter test lengths all exhibited greater rates of shrunken values. It seems likely that the mechanism that is causing these estimates to shrink is the same as that causing poor parameter estimation in other analyses, namely fewer observations available for a given parameter. Finally, e^{-5} is extremely small. This cutoff does not exclude other discrimination values that are too small for practical use (e.g., $e^{-4.61} = 0.01$).

Table 22. *Proportion of Discrimination Parameter Estimates Below* e^{-5}.

Test Length	Node	Sample Size = 500		Sample Size = 2,000	
		Below Average Propagation	Average or Above Propagation	Below Average Propagation	Average or Above Propagation
10 Items	1	52 (5.20%)	—	13 (1.30%)	—
	2	67 (13.24%)	42 (8.50%)	24 (4.71%)	11 (2.24%)
	3	165 (30.22%)	57 (12.56%)	62 (10.53%)	11 (2.68%)
	4	364 (60.47%)	103 (25.88%)	147 (23.71%)	20 (5.26%)
30 Items	1	100 (3.33%)	—	23 (0.77%)	—
	2	162 (10.95%)	105 (6.90%)	61 (4.07%)	26 (1.73%)
	3	396 (23.93%)	149 (11.08%)	127 (7.41%)	55 (4.28%)
	4	810 (44.46%)	246 (20.88%)	348 (18.83%)	69 (5.99%)

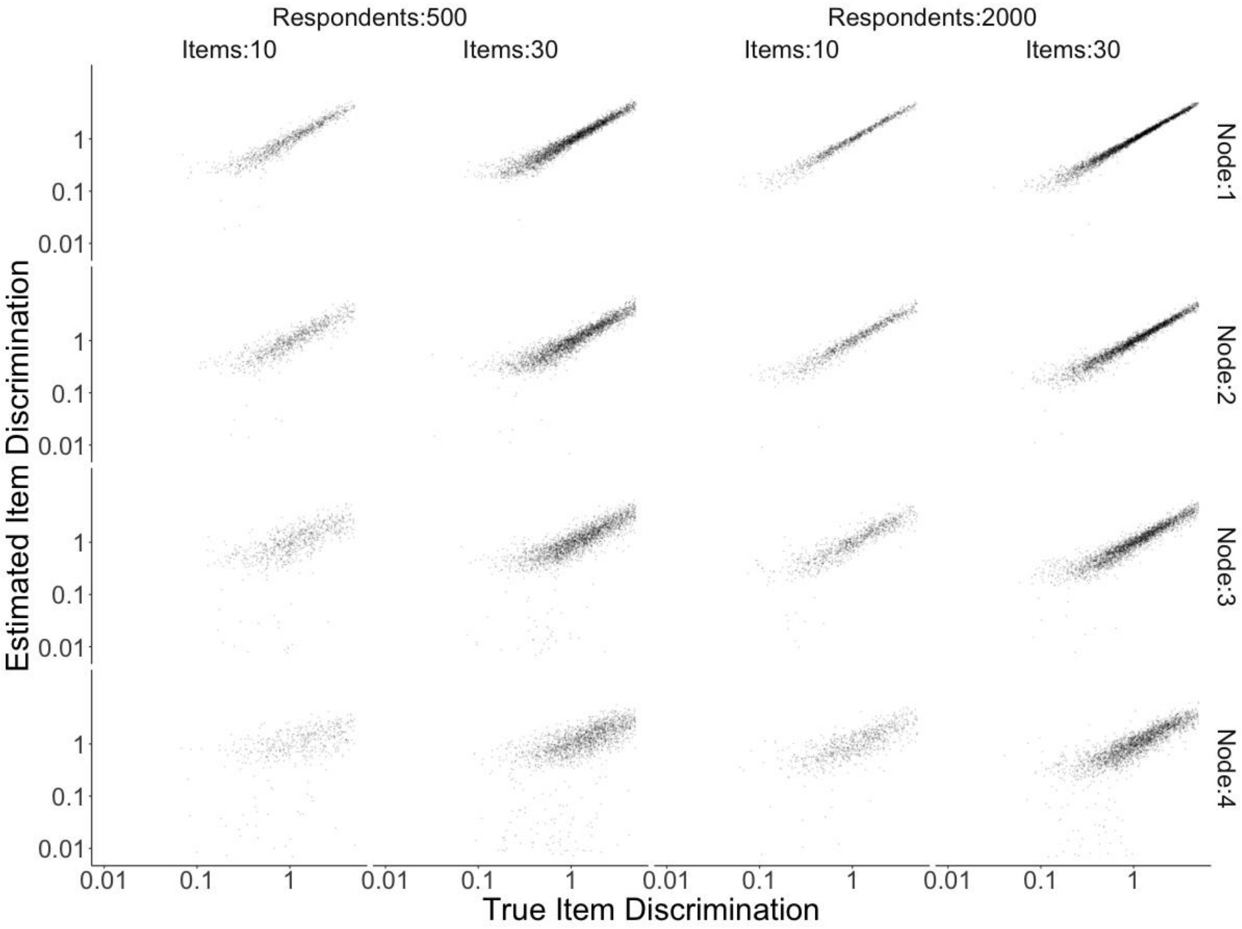
Figure 19*. Scatterplot of True and Estimated 2PL Discrimination Parameters.*



Note. The x-axis is on the natural log scale and y-axis is on the log$_{10}$ scale.

The remaining observations (Figure 20) appeared to also be composed of at least two distributions, one of them normal and centered around zero on the log scale (group 1), and the other appearing normal with a mean below zero (group 2). There was no obvious break between the two distributions. The group 2 distribution centered below zero was composed of relatively few observations. In order to diminish the influence of these observations on the measurement of the mean and variance of the group 1 distribution, I used a student's *t*-distribution which is more robust against outliers. The need for such a complicated analysis implies that recovery of the discrimination parameters is poor in general. The rationale for attempting to measure recovery for group 1 and not the other observations is that, in practice, most tasks would suggest that extremely small estimates would result in the removal of such items from further analysis. Estimating the bias and variability of extremely small discrimination parameter estimates is not practically meaningful.

Figure 20. *Scatterplot of True and Estimated 2PL Item Discrimination Parameters with Estimated Parameter Observations Falling Below e^{-5} Removed.*



Note. The x- and y-axes are on the natural log scale.

Regression Model Results

I used the location-scale parameterization of the t-distribution, with location $\hat{\mu}$, scale $\hat{\sigma}$, and degrees-of-freedom ν. The $\hat{\sigma}$ is not a direct estimate of the variance of the distribution which relies on ν and is calculated as $\sigma^2 = \hat{\sigma}^2 \frac{\nu}{\nu-2}$ when the distribution has a definite variance (i.e., $\nu > 2$). For a more direct interpretation of the predictor relationships with the variance, I sampled regression weights for the predictors predicting σ instead of $\hat{\sigma}$ and then calculated the scale parameter as $\hat{\sigma} = \sqrt{\sigma^2 \frac{\nu-2}{\nu}}$. I also constrained ν to have a lower bound of 2, ensuring finite variance. After multiple attempts to fit this model, I was required to exclude all predictors of the mean in order to get the model to sample within a reasonable timeframe.

The resulting model displayed adequately mixing chains and $\hat{R}$ values of 1.00 after rounding. The model parameters are presented in Table 23. Figure 21 displays both the observed and model posterior predicted distribution of marginal bias. Table 24 displays model predicted and observed log-means and log-standard deviations of discrimination estimate bias. The figure and table suggest that the model underestimates the locations of the distributions, and it severely under- and overestimates the variances of the distributions. The location parameter is uniform under all conditions and isn't very informative. The model parameters may be valid in direction but not in magnitude. I present the results and predictions of the model below for the sake of consistency, but I discourage the reader from using the results for making strong inferences of their own.

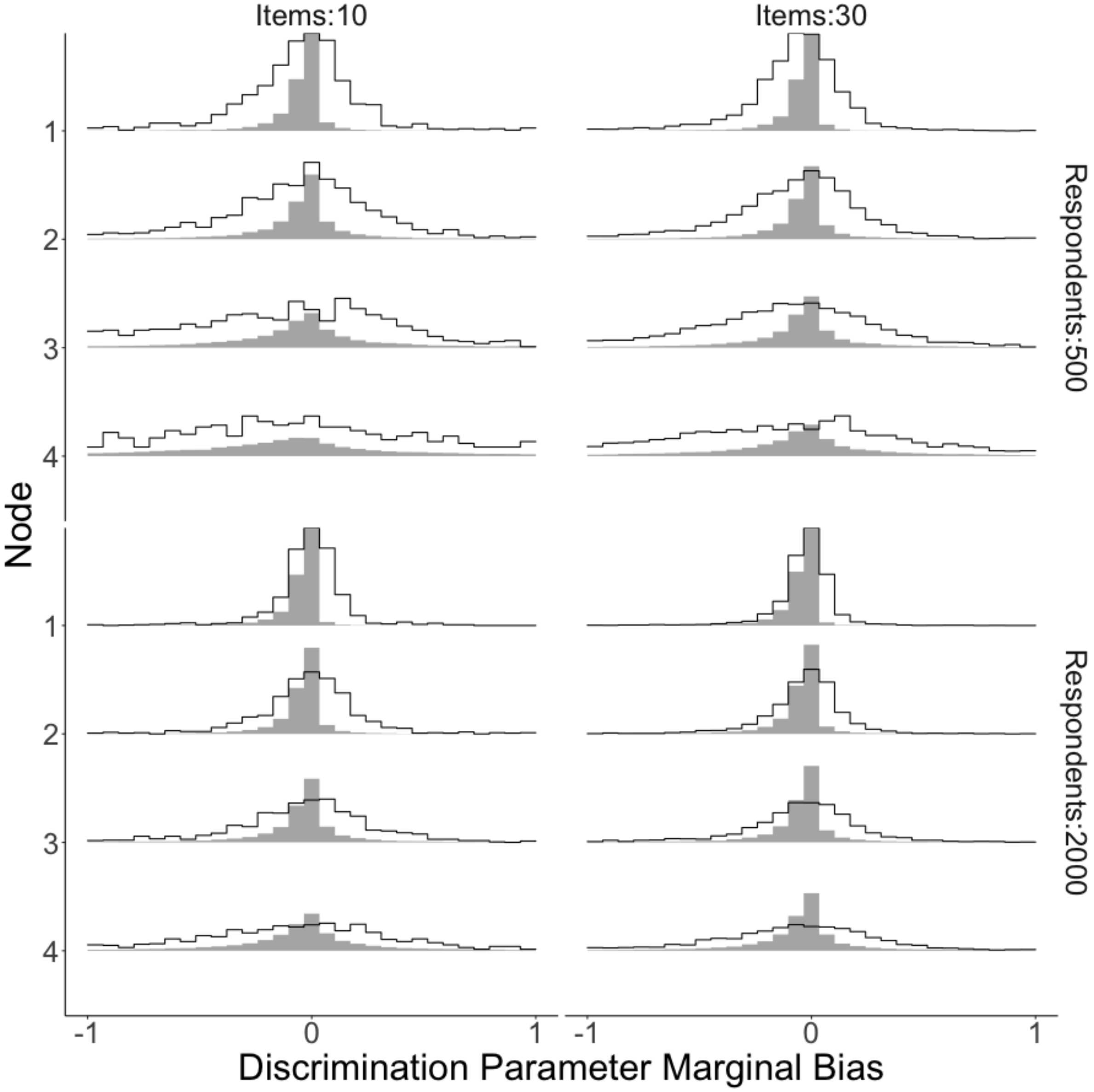
The estimates exhibited very little bias on average, $\overline{b_\mu} = -0.04,\ 95\%_{\text{HDPI}}[-0.04, -0.03]$. There was also very little iteration-specific variance associated with the mean $\overline{\tau_\mu} = 0.03, 95\%_{\text{HDPI}}[0.02, 0.03]$. For estimates that weren't shrunk to near zero, there appears to be little bias on average. The degrees-of-freedom of the *t*-distribution was $\nu = 2.09$, $95\%_{\mathrm{HDPI}}[2.03, 2.15]$. The intercept of the estimate variance was small, $\overline{b_\sigma} = -0.79$, $95\%_{\mathrm{HDPI}}[-1.08, -0.37]$. Sample size, $\overline{b_\sigma} = -0.31, 95\%_{\mathrm{HDPI}}[-0.33, -0.29]$, and relative propagation, $\overline{b_\sigma} = -0.25,\ 95\%_{\mathrm{HDPI}}[-0.32, -0.18]$, had modest negative effects on estimate variability. Test length, $\overline{b_\sigma} = -0.10, 95\%_{\mathrm{HDPI}}[-0.12, -0.08]$, had a small negative effect on estimate variability, and node depth, $\overline{b_\sigma} = 0.48, 95\%_{\mathrm{HDPI}}[0.47,0.49]$, had a moderate positive effect on variability. Again, iteration-specific intercept variance was small, $\overline{\tau_\sigma} = 0.04$, $95\%_{\mathrm{HDPI}}[0.00, 0.07]$. All interaction effects were small and unreliably different from zero.

Table 23. *2PL Item Discrimination Model Posterior Means and Credibility Intervals of Predictors of Estimate Bias and Variability.*

	Estimate Bias		Estimate Variability†	
	$\overline{b_{\mu}}$	$95\%HPDI$	$\overline{b_{\sigma}}$	$95\%HPDI$
Intercept	-0.04	[-0.04, -0.03]	-0.79	[-1.08, -0.37]
S			-0.31	[-0.33, -0.29]
T			-0.10	[-0.12, -0.08]
D			0.48	[0.47, 0.49]
P			-0.25	[-0.32, -0.18]
S x T			0.00	[-0.02, 0.02]
S x D			0.00	[-0.01, 0.01]
T x D			-0.03	[-0.05, -0.02]
S x P			-0.02	[-0.09, 0.05]
T x P			-0.06	[-0.13, 0.01]
D x P			-0.03	[-0.06, 0.00]
S x T x D			0.01	[-0.01, 0.02]
S x T x P			-0.01	[-0.08, 0.06]
S x D x P			0.02	[-0.01, 0.05]
T x D x P			0.02	[-0.01, 0.05]
S x T x D x P			-0.01	[-0.04, 0.03]
τ	0.03	[0.02, 0.03]	0.04	[0.00, 0.07]

Note. *N* = 32,000, degrees-of-freedom ν = 2.09 [2.03, 2.15]. D = node depth, S = sample size (standardized, *M* = 1,250, *SD* = 750.01), T = test length (standardized, *M* = 25, *SD* = 8.66), P = log-relative propagation, τ = between-simulation iteration intercept variance, HPDI = highest posterior density interval. † Model parameters are on the log scale.

Figure 21. *Observed and Posterior Predicted Marginal Bias for 2PL Discrimination Parameters.*



Note. Observed marginal bias distribution is indicated by the solid black line and the posterior predictions of the marginal bias distribution is depicted with the grey shaded regions.

Table 24. *True and Estimated Item Discrimination Model Predicted and Observed Bias Means and Standard Deviations.*

Test Length	Node	Log Mean				Log Standard Deviation			
		Sample Size = 500		Sample Size = 2000		Sample Size = 500		Sample Size = 2000	
		Predicted	Observed	Predicted	Observed	Predicted	Observed	Predicted	Observed
10	1	-0.04	-0.05	-0.04	0.00	0.07	0.37	0.03	0.15
	2	-0.04	-0.11	-0.03	-0.02	0.31	0.55	0.09	0.27
	3	-0.03	-0.28	-0.03	-0.08	1.21	0.94	0.24	0.58
	4	-0.03	-0.41	-0.03	-0.17	4.94	1.05	0.85	0.69
30	1	-0.03	-0.06	-0.04	-0.02	0.05	0.30	0.03	0.20
	2	-0.04	-0.07	-0.04	-0.03	0.22	0.41	0.07	0.27
	3	-0.04	-0.18	-0.04	-0.04	0.74	0.65	0.16	0.32
	4	-0.03	-0.33	-0.04	-0.15	1.70	0.94	0.45	0.64

Tables 25 and 26 display posterior predictions of differences between the marginal distributions of estimated and true discrimination parameters for 10- and 30-item tests, respectively. Under all conditions, the marginal estimate bias is very small, and the variance is large. Only under ideal conditions of large samples sizes and long test lengths with highly propagating and highly discriminating items does the estimate variability permit reliable inferences. For example, for the first node with a sample size of 2,000 and test length of 30 items, an estimate of $\alpha = 1$ has an uncertainty interval between $.75 < \alpha < 1.06$, or a range of 0.31 which is nearly a third the size of the estimate. In contrast, if $\alpha = 3$, the uncertainty interval is $2.75 < \alpha < 3.06$. The range is only a tenth of the size of the estimate. Several conditions, such as those with low propagation rates or deep nodes, exhibit extremely large uncertainty intervals. This poses very restrictive conditions and severely limits many practical inferences involving the item discrimination parameters such as item selection tasks.

There is still a large amount of uncertainty around the model estimates not expressed in the results. The bias and uncertainty of the estimates needs to be understood in the context of the poorly estimated distributional regression model and the data that were removed prior to estimating the model. Table 22 suggests that rates of extremely low discrimination parameter estimates increased for deeper nodes, smaller sample sizes, shorter test lengths, and lower observation propagation rates. The conditions that produce poor estimate recovery in the distributional model of estimate bias are the same conditions that produce high rates of extremely small discrimination estimates. In addition, the model does not account for a large portion of the data that were removed and indicates large uncertainty in the item parameter estimates.

Table 25. *Posterior Prediction Medians and Credibility Intervals of Differences Between True and Estimated 2PL Item Discrimination Parameters for Tests with 10 Items.*

Node	Propagation Ratio†	Sample Size = 250		Sample Size = 500		Sample Size = 1,000		Sample Size = 2,000	
		α^*_{Med}	95% HPDI	α^*_{Med}	95% HPDI	α^*_{Med}	95% HPDI	α^*_{Med}	95% HPDI
1	—	-0.03	[-0.37, 0.17]	-0.03	[-0.34, 0.16]	-0.03	[-0.30, 0.09]	-0.03	[-0.26, 0.06]
2	0.05	-0.02	[-2.93, 3.69]	-0.02	[-2.42, 2.62]	-0.02	[-1.36, 1.43]	-0.02	[-0.62, 0.44]
	0.5	-0.02	[-0.97, 0.86]	-0.02	[-0.81, 0.63]	-0.02	[-0.59, 0.40]	-0.03	[-0.39, 0.14]
	1	-0.02	[-0.78, 0.56]	-0.02	[-0.71, 0.49]	-0.02	[-0.52, 0.25]	-0.03	[-0.32, 0.13]
	2	-0.02	[-0.64, 0.41]	-0.02	[-0.53, 0.33]	-0.03	[-0.43, 0.22]	-0.03	[-0.31, 0.09]
3	0.05	0.00	[-26.90, 163.00]	-0.02	[-7.38, 38.32]	-0.03	[-6.15, 11.07]	-0.02	[-1.33, 1.85]
	0.5	-0.02	[-2.76, 4.26]	-0.02	[-2.43, 2.75]	-0.02	[-1.55, 1.48]	-0.02	[-0.69, 0.46]
	1	-0.02	[-1.88, 2.11]	-0.02	[-1.50, 1.84]	-0.02	[-1.20, 0.98]	-0.02	[-0.58, 0.37]
	2	-0.02	[-1.32, 1.23]	-0.02	[-1.03, 1.05]	-0.02	[-0.84, 0.60]	-0.03	[-0.46, 0.27]
4	0.05	0.03	[-29.75,2.98e+11]	-0.01	[-58.30,9.90e+7]	-0.01	[-57.68, 50.20e+2]	-0.02	[-5.91,13.13]
	0.5	-0.03	[-13.21, 57.77]	-0.03	[-8.07, 24.39]	-0.02	[-4.28, 8.38]	-0.02	[-1.46, 2.09]
	1	-0.02	[-6.77, 13.20]	-0.02	[-4.45, 7.54]	-0.02	[-2.74, 3.86]	-0.02	[-1.13, 1.20]
	2	-0.02	[-3.29, 4.69]	-0.02	[-2.48, 3.56]	-0.02	[-1.94, 2.13]	-0.02	[-1.03, 0.70]

Note. α^*_{Med} = median of the posterior predicted discrimination estimate marginal bias, HPDI = highest posterior density interval.
†Formula for approximate number of observed responses: $n = Sample\ Size \times Propagation\ Ratio \times 0.5^{Node-1}$. Bias is difference between the marginal distributions of estimated discrimination parameters and true discrimination parameters.

Table 26. *Posterior Prediction Medians and Credibility Intervals of Differences Between True and Estimated 2PL Item Discrimination Parameters for Tests with 30 Items.*

Node	Propagation Ratio†	Sample Size = 250		Sample Size = 500		Sample Size = 1,000		Sample Size = 2,000	
		α^*_{Med}	95% HPDI	α^*_{Med}	95% HPDI	α^*_{Med}	95% HPDI	α^*_{Med}	95% HPDI
1	—	-0.03	[-0.29, 0.13]	-0.03	[-0.33, 0.08]	-0.03	[-0.27, 0.07]	-0.03	[-0.25, 0.06]
2	0.05	-0.02	[-2.08, 2.20]	-0.02	[-1.78, 1.86]	-0.02	[-1.17, 1.18]	-0.02	[-0.65, 0.54]
	0.5	-0.02	[-0.64, 0.47]	-0.02	[-0.62, 0.35]	-0.03	[-0.40, 0.21]	-0.03	[-0.33, 0.08]
	1	-0.03	[-0.44, 0.31]	-0.03	[-0.44, 0.21]	-0.03	[-0.37, 0.17]	-0.03	[-0.26, 0.07]
	2	-0.03	[-0.42, 0.16]	-0.03	[-0.36, 0.15]	-0.03	[-0.33, 0.12]	-0.03	[-0.25, 0.07]
3	0.05	-0.02	[-4.55,13.05]	-0.02	[-3.97, 10.08]	-0.02	[-3.30, 4.47]	-0.02	[-1.38, 1.62]
	0.5	-0.02	[-1.31, 1.47]	-0.02	[-1.11, 1.11]	-0.02	[-0.82, 0.78]	-0.02	[-0.48, 0.26]
	1	-0.02	[-1.01, 0.77]	-0.03	[-0.76, 0.60]	-0.02	[-0.65, 0.38]	-0.03	[-0.36, 0.14]
	2	-0.02	[-0.69, 0.47]	-0.02	[-0.59, 0.34]	-0.03	[-0.47, 0.24]	-0.03	[-0.30, 0.13]
4	0.05	0.01	[-18.08,1148.53]	-0.02	[-11.71, 289.20]	-0.03	[-9.77, 35.09]	-0.02	[-4.15, 5.03]
	0.5	-0.02	[-3.44, 4.34]	-0.02	[-2.49, 4.09]	-0.02	[-1.75, 2.21]	-0.02	[-0.88, 0.79]
	1	-0.02	[-1.85, 2.49]	-0.02	[-1.56, 1.91]	-0.02	[-1.19, 1.14]	-0.02	[-0.60, 0.48]
	2	-0.02	[-1.47, 1.19]	-0.02	[-1.20, 1.04]	-0.02	[-0.84, 0.69]	-0.03	[-0.43, 0.30]

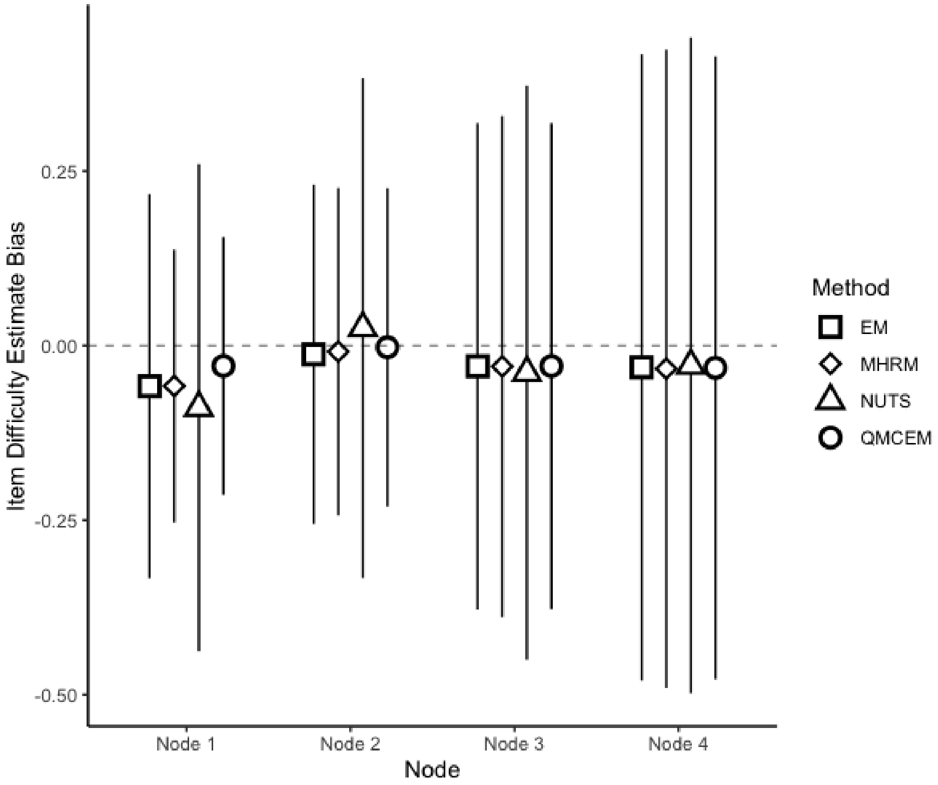
Note. α^*_{Med} = median of the posterior predicted discrimination estimate marginal bias, HPDI = highest posterior density interval.
†Formula for approximate number of observed responses: $n = Sample\ Size \times Propagation\ Ratio \times 0.5^{Node-1}$. Bias is difference between the marginal distributions of estimated discrimination parameters and true discrimination parameters.

Post Hoc Analyses for 2PL Model

After conducting the initial analyses, I performed some post hoc analyses to try and further diagnose the estimation issues I experienced with the 2PL models. I simulated 50 datasets with 500 respondents and 10-items per node for the 4-node resilience tree. I estimated the 2PL model parameters with the MH-RM, expectation maximization, and quasi-Monte Carlo expectation maximization routines offered by the mirt package (Chalmers, 2012). I also estimated the parameters using a similar model in Stan (Stan Development Team, 2019) that uses a Hamiltonian Monte Carlo No-U-Turn Sampler (NUTS; Hoffman & Gelman, 2014). None of the other estimators in the mirt package produced standard errors. Stan produces posterior draws that can be summarized with standard deviations for each parameter, which is an estimate of uncertainty similar to normal theory standard errors. Figures 22, 23, and 24 and Table 27 display the means and standard deviations of estimate bias for the point estimates of the mirt estimation routines and the posterior medians from Stan for item discrimination, item difficulty, and person ability parameters. The results for the difficulty parameters suggest that all estimators displayed very little bias with moderate and similarly sized estimate uncertainty. The NUTS routine in Stan produced larger estimate uncertainty compared to the other methods, particularly for node 1. This difference diminished with deeper nodes. The results for the discrimination parameters suggest that the mirt estimation methods are systemically downwardly biased, increasingly so with deeper nodes, whereas the NUTS estimates are unbiased. The MH-RM estimates were severely downwardly biased because large portions of the discrimination parameter estimates were shrunk towards zero like I found in the previous 2PL simulation. None of the other estimation methods produced severely shrunken discrimination parameter estimates. Finally, all estimation methods produced adequate person ability estimates that were not appreciably different in mean bias or variance.

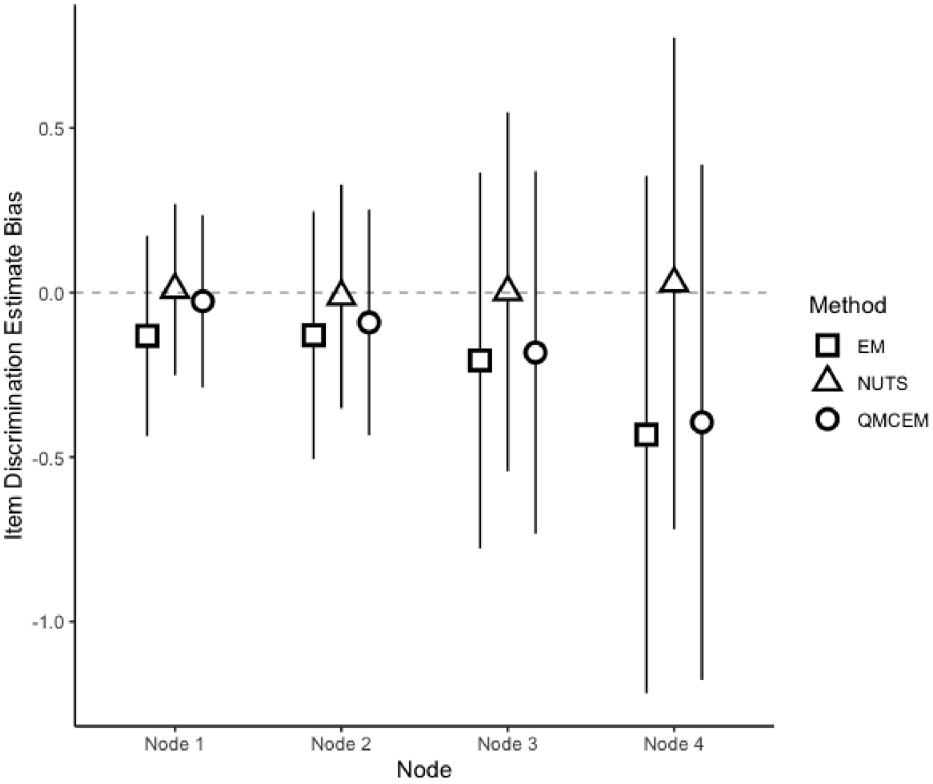
Although this is just a small sample of a single combination of conditions, it does suggest that the poor recovery of the discrimination parameters was largely the result of the MH-RM estimation routine. The discrimination parameter regression results presented for the initial 2PL simulation should not be trusted. The results from the 2PL models measuring recovery of the difficulty and person ability parameters are not likely to be substantially different from the other estimation methods and may still afford valid inferences.

Figure 22. *Sample Means and Standard Deviations of Item Difficulty Parameters by Estimation Method.*



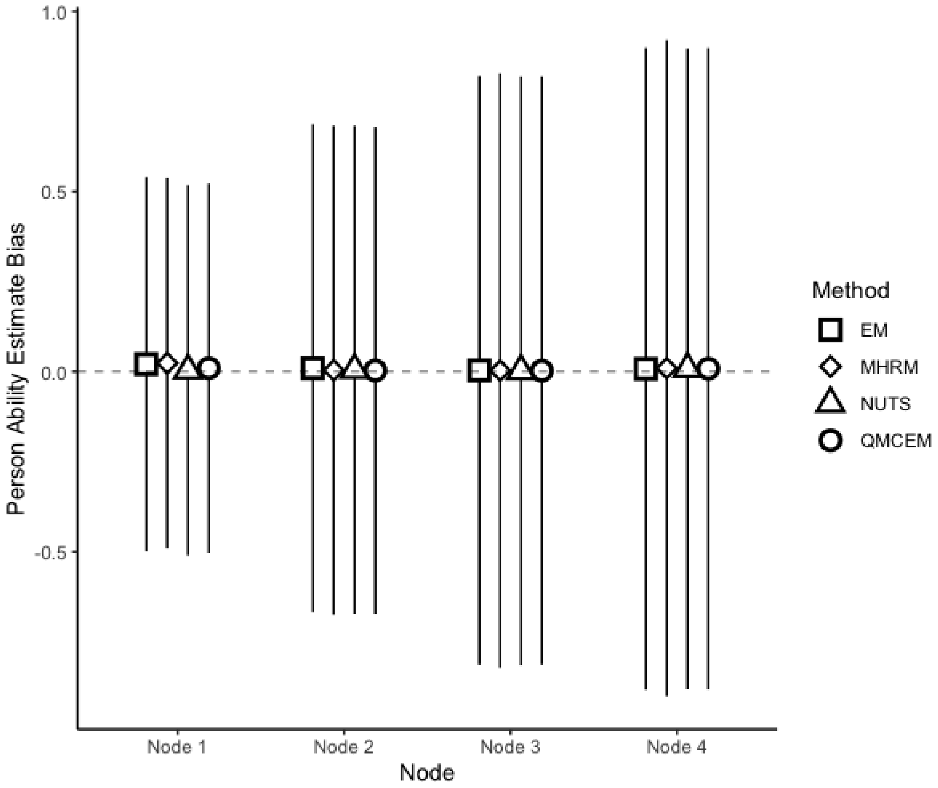
Note. EM = Expectation Maximization, MHRM = Metropolis-Hastings Robbins-Monro, NUTS = Hamiltonian Monte-Carlo No-U-Turn Sampler, and QMCEM = Quasi-Monte Carlo Expectation Maximization.

Figure 23. *Sample Means and Standard Deviations of Item Discrimination Parameters by Estimation Method.*



Note. EM = Expectation Maximization, NUTS = Hamiltonian Monte-Carlo No-U-Turn Sampler, and QMCEM = Quasi-Monte Carlo Expectation Maximization. The MH-RM estimation method is not included because the mean bias and standard deviations were too large in magnitude compared to the other methods to appropriately display on the same scale. Refer to Table 27 for MH-RM results. Estimate bias is the log of the ratio between the estimated and true discrimination parameters.

Figure 24. *Sample Means and Standard Deviations of Person Ability Parameters by Estimation Method.*



Note. EM = Expectation Maximization, MHRM = Metropolis-Hastings Robbins-Monro, NUTS = Hamiltonian Monte-Carlo No-U-Turn Sampler, and QMCEM = Quasi-Monte Carlo Expectation Maximization.

Table 27. *Post Hoc Sample Statistics for Item and Person Parameters*.

Node	Method	$\hat{\alpha}$		$\hat{\beta}$		$\hat{\theta}$	
		M	*SD*	*M*	*SD*	*M*	*SD*
1	EM	-0.13	0.30	-0.06	0.27	0.02	0.52
	MH-RM	-1.94	8.74	-0.06	0.20	0.02	0.51
	NUTS	0.01	0.26	-0.09	0.35	0.00	0.51
	QMCEM	-0.03	0.26	-0.03	0.18	0.01	0.51
2	EM	-0.13	0.38	-0.01	0.24	0.01	0.68
	MH-RM	-4.69	12.79	-0.01	0.23	0.00	0.68
	NUTS	-0.01	0.34	0.03	0.36	0.01	0.68
	QMCEM	-0.09	0.34	0.00	0.23	0.00	0.68
3	EM	-0.21	0.57	-0.03	0.35	0.00	0.82
	MH-RM	-8.56	16.21	-0.03	0.36	0.00	0.83
	NUTS	0.00	0.55	-0.04	0.41	0.00	0.82
	QMCEM	-0.18	0.55	-0.03	0.35	0.00	0.82
4	EM	-0.43	0.79	-0.03	0.45	0.01	0.89
	MH-RM	-20.39	20.33	-0.03	0.46	0.01	0.91
	NUTS	0.03	0.75	-0.03	0.47	0.01	0.89
	QMCEM	-0.39	0.78	-0.03	0.45	0.01	0.89

Note. *n* per cell = 500 for item parameters and 25,000 for person parameters. EM = Expectation Maximization, MH-RM = Metropolis-Hastings Robbins-Monro, NUTS = Hamiltonian Monte-Carlo No-U-Turn Sampler, and QMCEM = Quasi-Monte Carlo Expectation Maximization. $\hat{\alpha}$= item discrimination parameter, $\hat{\beta}$ = item difficulty parameter, $\hat{\theta}$= person ability parameter.

Discussion

As with any IRT model, IRTrees are useful insofar as they adequately measure or summarize the underlying data generating item and person properties. Simulation studies are a standard method for investigating the validity of a model. Recovery of the data generating parameters is often of interest. Large estimate bias threatens valid inferences with a model. Large estimate variability threatens reliable inferences. In the present study, I simulated data generated from 1PL and 2PL IRTree models with varying sample sizes and test lengths, estimated the model parameters, and then investigated parameter recovery for item and person ability parameters. A benefit of the IRTree framework is that it provides a large amount of flexibility and potential to test novel research hypotheses. However, the results of this project indicate the need for limitations on their practical use. The IRTree framework is novel in how it may be applied but is essentially the same as any other IRT model at its core. IRTree models require enough observations for each item to estimate item parameters. They also require enough observations for each person to estimate person parameters, as any other IRT model does. This means there must be enough respondents in the sample and enough items per node in the test. The issues presented in this study are not novel themselves. Rather, they are the same issues studied by other researchers for decades, but these issues have presented in IRTrees in a unique way.

Item Difficulty Parameters

The results from the 1PL and 2PL simulations were very similar and suggest that the item difficulty parameters are unbiased under most conditions. From a frequentist perspective, under the assumption of infinite repeated sampling or sample size, the item difficulty estimates are consistent and should converge to their true data generating values at the limit where $n \rightarrow \infty$.

Usage of IRTrees, or any other IRT model, involves finite samples, often with sample sizes that are not adequate approximations of this limit. Estimate variability is important to quantify the range or distribution of plausible values. In the present study, the sampling variability or measurement uncertainty is large under many practical testing and modeling conditions. Although large sample sizes provide smaller uncertainty intervals, the propagation of observations to the deepest node is more influential. If a set of auxiliary-items each have low endorsement rates, they will not propagate many observations to descendent nodes. Large sample sizes ($N \geq 2{,}000$) can diminish the chance of obtaining too few observations for a given node due to a set of sampled auxiliary-items with low propagation. Researchers should be mindful of the number of endorsements of each category for each rating-scale item or set of auxiliary-items and the IRTree structure they plan to use for analysis. If a category for a rating-scale item is indicative of an implicit response that occurs several nodes deep into a set of responses and that category has a low endorsement rate, the researcher should lower their expectations for how precise the difficulty estimates for that item can be.

For the 1PL model, the results suggest that a minimum of around 100 observations are required to adequately estimate a given item at a given node if a researcher wants to be able to distinguish items that are one standard deviation apart on the ability scale. This means, on average, researchers need a minimum of $N = 100 \times 2^{depth}$ total respondents for adequate measurement at a given node depth. This requires average to easy items at each ancestor node, a requirement that becomes less likely with each depth level. This is paradoxical because a well calibrated test typically requires items that target a wide range of the latent trait levels. Requiring only or mostly average to easy items suggests that the test items must target a specific area of the latent trait in the population of interest. In order to achieve sufficiently small uncertainty intervals for the difficulty parameter estimates, a researcher may need to sacrifice adequate estimation of the targeted trait in the population of interest. The 2PL regression model suggests that the difficulty parameters for 2PL IRTrees comes with greater estimate variability. The posterior predictions indicated that node-specific sample sizes increase with greater depths. The root-node must propagate at least 200 observations on average for nodes at a depth of 1. This required sample size increases up to around 300 to 400 respondents depending on the total sample size and test length. Under the assumption of normally distributed item difficulty parameters, the predictive simulation suggests that deep nodes for either model under any set of conditions are unlikely to be propagated enough observations for adequate estimate reliability. For the same reason, the 2PL model is unlikely to produce estimate reliability for nodes deeper than 1. As I said before, the criterion of differentiating items 1 standard deviation apart is arbitrary, but it provides a starting point for understanding the practical implications of the study results. Some researchers will need a more stringent criterion whereas others may need a less stringent criterion.

The standard errors for the 1PL model provided adequate coverage over the data generating values. These are useful for informing the researcher about the limits of inference, but do not resolve the issue of high variability. The 2PL models did not produce standard errors due to estimation issues. The post-hoc simulation suggests that several estimation routines offered by the mirt package are not able to produce standard errors for the item parameters.

Person Ability Parameters

The person ability parameters for the 1PL and 2PL models were, on average, unbiased. However, the estimate variance was large. Deeper nodes and shorter tests produced greater estimate variability. Deeper nodes produce greater missingness for a given person, providing less information about that person’s ability parameter. Shorter tests produce fewer opportunities for a given respondent to respond to an auxiliary-item at a given descendent node resulting in greater missingness per node. Both contribute to respondents with completely missing responses on descendent nodes. Both models produced ability estimates that are shrunk to zero, increasing in number with greater depth and shorter tests. I believe the cause is the lack of observed item responses for a given person at a given node. EAP estimation is the expected value of the posterior distribution of a person’s ability parameter. If there are no items to provide information on where a person is located on the ability continuum, the most likely location is the population mean (zero when the population is assumed to come from a standard normal distribution). The shrunken estimates are symmetrical around the mean of the ability distribution, so the aggregated estimate bias and variance are largely unaffected. The shrunken estimates attenuate the correlation between the true and estimated ability parameters. A consequence is that all relationships with external criteria are also attenuated. Unless the true relationship between the IRTree ability and the external criterion is very strong, there are few practical conditions where mere directional tests of the relationship are reliable.

A researcher could possibly remove respondents with sparse response patterns for a given node to try and limit attenuation of the correlation. However, the ability of a researcher or practitioner to utilize an IRTree model for assessment or selection purposes is then conditional on a person’s measured attribute for latent factors utilized at the beginning of a branch. Respondents essentially select themselves out as a result. A very large number of items, or items tailored specifically to the target range of abilities, would be required in order to minimize this selecting out process. These results make tasks such as using IRTrees to produce latent ability scores parted from response styles doubtful.

Item Discrimination Parameters

Recovery of the item discrimination parameters in the 2PL model was not accurately assessed in this study. The simulated dataset contained a substantial number of estimates shrunk to near zero, diverging substantially from the log-normal distribution that generated the true values. This made analysis of estimate bias and variability challenging. The model that I constructed required removal of most of these shrunken estimates and still could not provide adequate posterior predictions of the observed data. The discrimination parameter regression model results are not trustworthy. The post hoc analyses are more informative. The MH-RM estimation routine is the likely culprit of the shrunken discrimination parameter estimates. The expectation maximization routine and quasi-Monte-Carlo expectation maximization routine both produced downwardly biased estimates, but they did not produce estimates shrunk to zero. The NUTS routine in Stan adequately estimated the discrimination parameters.

The results for the person ability parameters did not seem to suffer from the poor estimation of the discrimination parameters. One explanation is that when discrimination parameters were severely biased, they tended to be downwardly biased or shrunk all the way to zero. Estimates that small provide little information across the latent ability dimension. The likelihood of an ability score $\theta_{[k]}$ given an observed response pattern $X_{[i]}$ is the product of the observed response probabilities to each item given item parameters $\xi_{[j]}$,

$$\mathcal{L}\left(\theta_{[k]}|X_{[i]}\right) = \prod_{j=1}^{J} \Pr\left(1|\theta_{[k]}, \xi_{[j]}\right)^{x_{[i,j]}} \times \Pr\left(0|\theta_{[k]}, \xi_{[j]}\right)^{1-x_{[i,j]}}.$$

If one of the items had a near zero discrimination parameter, the probability function would be nearly flat across the latent ability dimension. Likewise, there would be little difference in the likelihood across ability levels for a given response pattern. The EAP score would be estimated as if the item weren't included in the first place. This is more desirable than an upwardly biased discrimination parameter wherein the item has much greater influence over the EAP scores.

Limitations

Estimation Method and 2PL IRTree Models

A potential threat to validity is the estimation method I used. I used the MH-RM method for estimation (Cai, 2010) implemented in the mirt package (Chalmers, 2012). Chalmers and Flora (2014) found that the MH-RM method produces high parameter estimate variances for 2PL non-compensatory models, especially with short tests and strong correlations between the latent factors. The IRTrees that I used are a form of non-compensatory IRT model because some rating scale responses are conditional on successfully responding to multiple items on multiple dimensions. All auxiliary-items except the root-node auxiliary-item are non-compensatory in nature. This may partly explain why the parameters were so poorly recovered for deeper nodes. Wang and Nydick (2015) found similar results to Chalmers and Flora (2014), and also found that MCMC methods provide better recovery of non-compensatory IRT model parameters. The results from the NUTS method used in the post hoc analyses for the 2PL models agree with these findings. Many researchers in the past have used the lme4 package (Bates et al., 2015) for estimation of IRTree models (e.g., De Boeck & Partchev, 2012), which uses Laplace approximation or adaptive Gauss-Hermite quadrature and non-linear optimizers. These estimation methods may produce estimates with greater or lesser bias and variance. The purpose of the post hoc analyses was largely to explore whether other estimation methods would produce shrunken discrimination parameter estimates, so I did not explore other estimation methods such as those used in the lme4 package. Future research should investigate differences in IRTree parameter recovery between parameter estimation methods.

This study does not provide much insight into the estimation of the discrimination parameter for 2PL models. I believe the estimation issues that produced item discrimination estimates near-zero were the result of the MH-RM estimation routine. The post-hoc analyses suggest that several other routines, although they do not produce near-zero estimates, do produce downwardly biased estimates. Future research should more rigorously investigate which estimation routines provide adequate recovery of the discrimination parameter. Despite this limitation, recovery of the difficulty and person ability parameters are not substantially different from the 1PL model. The 2PL results may be used for inferences about those parameters, although I recommend further research be conducted.

Previous Research and Higher Order Tests

The study results are neither consistent with nor contradictory to previous simulation studies performed by other researchers. I argue that the design of this study is more thorough and systematic in its approach to understanding parameter recovery than previous IRTree simulation studies. Some researchers have conducted limited simulation studies with IRTrees before and found that the models adequately estimate the quantities of interest. This study suggests the need for caution for those wishing to use IRTree models or use previous literature on IRTree models. Researchers conducting IRTree simulation studies have not thoroughly presented their findings, have previously used aggregated estimates of bias and variance, have used a single sample size and test size, or have investigated criteria that are a level too far removed from the item and person parameter estimates to make a direct comparison with this study.

This simulation study focused on parameter recovery, particularly in the context of tasks requiring precise measurement such as item or personnel selection or for predicting criteria using the measured person abilities. These are common tasks for other IRT models, but many tasks in the IRTree literature focus on hypothesis testing by means of (primarily likelihood-based) model comparisons (e.g., Jeon & De Boeck, 2019; Partchev & De Boeck, 2012) or measuring latent correlations between the person abilities (e.g., Cho et al., 2020; Debeer et al., 2017). Other simulation studies using IRTrees have shown that inferences using model comparisons are valid under many practical testing and measurement conditions (Debeer et al., 2017; DiTrapani, 2019; Jin et al., 2019; Tijmstra et al., 2018). This may speak to the fungibility of parameter estimates for higher levels of analysis in which precise measurement of person or item characteristics is secondary to identifying a model to best characterize the data. A wide range of estimates to quantify relationships in the data may produce near identical inferences when comparing models using criteria at a higher level of analysis that are less sensitive to measurement error. Researchers should investigate the limits of these inferences using methods similar to those found in the SEM literature such as fungible parameter contours (Pek & Wu, 2018) which can characterize how sensitive model inferences are to small changes in the parameter estimates.

For IRTrees, because of the coding scheme, deeper nodes have fewer data points. As a result, the overall model likelihood accounts for fewer observations for deep nodes versus shallow nodes. This means that the model likelihood may be less affected by model differences that occur at deeper nodes than at shallower nodes. It may be the case that many likelihood-based inferences, such as using AIC, BIC, or likelihood ratio tests for model comparisons, are largely dependent on model differences occurring at the first or second node. Model differences occurring at descendent nodes become less influential on the likelihood and the overall likelihood-based comparison may be increasingly influenced by the differences in model degrees-of-freedom (or number of parameters) rather than fit to the data, especially when sample sizes are small and test lengths are short. Further research is needed to understand how variable parameter estimates can be and still allow researchers to conduct higher order tests and inferences.

Generalizability

There are several threats to the generalizability of this study. The most obvious is that this is a highly controlled simulation study. “Real world” data includes measurement error due to a variety of sources that further threaten valid inferences using IRTrees. This study represents a best-case scenario. Researchers should plan for greater estimate bias and uncertainty than what is presented here.

I only used two levels for the sample size and test length conditions. The distributional regression models considered in this study use simple linear approximations for the effects of these factors. The factors likely have some logarithmically diminishing effect as either sample size or test length grow in number, each have a logical lower bound at one, and each probably have a practical lower bound where estimation issues become prohibitive. Predictions beyond the limits of the chosen experimental levels should be considered with some caution.

I used an orthogonal latent factor structure, but in practice the latent factors would likely be correlated some degree. Imposing uncorrelated latent factors afforded me greater control over what caused bias and variability, particularly regarding the propagation effect. Correlated latent factors may provide additional information for estimation of the latent factors in the face of high missingness. A related limitation is that this study does not investigate the adequacy of IRTrees in recovering the latent covariance structure, which may be of interest to some researchers. The adequacy of estimating the covariance between latent factors may not be the same as estimation of the item and person ability parameters. Future studies should investigate recovery of the latent covariance structure.

I did not include the true value of a parameter as a predictor of estimate bias or variance. The common finding in past simulation studies is that person and item parameters located further away from the average tend to exhibit greater bias and variance (Thissen & Wainer, 1982). The presented models should be understood as explaining the average item or average person. Items and persons further away from the average will likely produce estimates with greater uncertainty than those predicted here.

Readers that want to use the results of these studies to inform a frequentist analysis of IRTree models should proceed with some caution. I used a Bayesian approach for estimating the IRTree models. When I estimated each IRTree model using the mirt package, I specified weakly informative priors on the item parameters. This contrasts with non-informative (e.g., uniform) priors or frequentist estimation techniques that do not explicitly use priors. Given many observations, the influence of these priors on the posterior would be overwhelmed by the data and point estimates would not likely differ much from frequentist estimation (Gelman et al., 2014). However, when observations are sparse, such as when samples sizes are small, test lengths are short, and deep auxiliary-items contain large amounts of missingness, the weakly informative priors would serve a regularizing function on the estimated parameters. For example, for the difficulty parameters, I used a normal distribution with a mean of zero and standard deviation of one. When information from the data was sparse, estimates of the difficulty parameters should have been regularized towards zero and should have made extreme negative or positive estimates less likely, providing a more conservative point estimate in the posterior. The lack of such a regularizing prior may result in substantially different difficulty estimates when there are few observations for a given item. If this is the case for the present study, the results for the item parameter estimates may be overly optimistic if a researcher wanted to use frequentist analysis methods. This is also suggestive that, if a researcher still wanted to use a frequentist interpretation, using at least weakly informative priors has a practical advantage as regularizing functions.

Other Recommendations and Thoughts for Future Research

IRTree Designs and Depth

The node depth and relative observation propagation factors in this study have implications for the types of IRTree models researchers use. The structural diagrams used to represent the IRTree of interest are not always representative of the depth of each node involved. For example, to model a 4-point agreement rating scale, I could specify a two-node IRTree. The first decision is whether to provide an agreeable response or disagreeable response (agreement node), and the second decision is whether to provide an extreme response (extreme node). The response process terminates only after responding to both nodes, and there is a single set of item parameters for the extreme node so that extreme agreement and extreme disagreement parameters are the same. Recoding rating scale responses into auxiliary-items would not require the introduction of missingness, and the resulting child node probabilities would not be conditional on parent nodes. It is perhaps debatable whether this qualifies as an IRTree at all. Similar caveats may be found in IRTrees such as the MPP model with a single extreme node where the agreement and extreme nodes are conditional on the midpoint node but are not conditional on one another. The commutativity of the order of the agreement and extreme nodes makes the MPP model have a maximum depth of 1. This is not the case when the MPP model is specified as having independent parameters for extreme agreement and extreme disagreement nodes, which would then have a maximum depth of 2. In sum, researchers should be cognizant of the number of observations available for estimating item parameters, which is directly related to the configuration of the IRTree model and its maximum depth.

A broader discussion about the usage of IRTrees is also warranted. Many researchers have used IRTrees to measure respondent response biases such as midpoint or extreme response styles (Böckenholt, 2017; Jeon & De Boeck, 2019; LaHuis et al., 2019). Using an IRTree model implies the need to account for some response conditional on another response. In the case of an MPP model, there is some hypothesized response process where a respondent poses a series of questions, each conditional on the preceding question. For the MPP model, the questions may be about the relevance (midpoint), valance (agreement), and intensity (extreme) of the respondent’s opinion towards the item content. The hypothesized structure is integral to the research question at hand and an IRTree model provides one potential solution. However, in some cases researchers purely interested in measuring and separating response styles from a substantive latent factor do not need to incorporate the conditional response structure. Incorporating a conditional response structure will, in the case of IRTrees, unnecessarily introduce missingness. Without a theory of conditional responding, it would be equally valid to recode a rating scale item into zeros and ones without introducing missingness and then estimate a regular multidimensional IRT model, a technique already common in the response style literature. Researchers should investigate differences between these two methods of measuring response styles and consider whether a conditional response structure is necessary or valid prior to using an IRTree model.

Reporting of Item Parameters

It is not yet custom to report the item parameters in published articles. For long tests, this seems somewhat reasonable due to space concerns in a journal (though a growingly unconvincing reason with the convenience of online journals and repositories for making research materials publicly available). Some IRTree studies are focused primarily on criteria that are of consequence to item estimation. The validity (and credibility) of the results of some study utilizing IRTrees may be conditional on adequate item parameter recovery. This cannot be ascertained by a reader if item parameters and standard errors are not reported. Response counts for each response option should also be tabulated and reported for each item. Depending on the design of the study and the goals of the researcher, some IRTree models may be precluded from analysis if a sufficient number of responses aren't observed for each response option.

Researchers have not reported issues with estimating standard errors for 2PL IRTree models in previous studies. The post hoc analyses for the 2PL models that I performed suggest that this is an issue for several estimation routines. One possibility is that there was some issue with the code I wrote for generating the simulated data. This is unlikely as there were no other signs of this possible issue in the 1PL model results or in the recovery of the 2PL difficulty and ability parameters. Another possibility is that item parameter estimates are not commonly reported, let alone their standard errors. Regarding previous simulation studies, many uses of IRTree models involve tasks that may not require precise estimation of the item parameters. The item parameter estimates may be highly fungible in the sense that they can take on a wide range of values that deviate from their data generating values and still produce valid inferences. Null hypothesis testing using a sensitive test, such as the likelihood-ratio test for model fit, may produce similar results with a wide range of item parameter estimate bias.

Bayesian Methods for Estimation

A fully Bayesian approach may aid estimation of the item parameters. In many contexts, the item difficulty parameters are likely correlated across nodes. De Boeck and Partchev (2012) provide an example IRTree where the first node indicates whether a respondent chose to omit a response to an item, and the second node indicates agreement to the item conditional on providing a response. The difficulty parameters of the two auxiliary-items are likely correlated, such that items with difficult item content are also more likely to be omitted by a respondent. A hierarchical multivariate normal prior on the difficulty parameters could account for this latent correlation if the correlation is strong. Auxiliary-items in the second node that have few observations could "borrow information" via partial pooling from other second node auxiliary-items that have many observations and the latent correlation for more reliable estimation. Another strategy that might prove useful for lowering estimate variability is setting stronger priors on the item parameters. Incorporating beliefs about the difficulty or relevance of the item content into the prior distributions could be especially informative for nodes that reside deep within a tree or that have ancestor nodes with low endorsement probabilities and likely suffer from having few observations.

Adaptive Testing as a Solution

Adaptive testing may provide a solution to some of these concerns. Adaptive testing often involves estimation of item parameters beforehand and then selectively administering items to optimize some criterion such as minimizing person ability estimate standard errors. In this case, instead of using pre-estimated item parameters, marginal sums of item endorsements/non-endorsements could be used to determine whether to administer additional items to a person. For example, if a person answered 20 items and their observed responses would not produce any response to recoded auxiliary-items at the deepest node, additional items could be administered. This could prevent instances of respondents with zero observations for a given node and therefore an inestimable latent ability for that node.

Conclusion

IRTrees provide a flexible framework for testing unique hypotheses and measuring latent abilities with complex survey structures. Many IRTree applications involve the introduction of missingness due to the recoding procedure. Prior to this study, simulation studies involving IRTree models were either lacking thorough investigation or focused on specific applications. None have given attention to the unique conditions that result in greater amounts of missing data. My aim in this study was to measure the adequacy of parameter recovery across several conditions and provide some perspective with selected estimation criteria. In particular, I wanted to highlight the conditional structure of the model can produce insufficient estimates under many common testing conditions. Like all item response models, conditions that result in fewer observations per item result in greater item parameter estimate uncertainty, such as smaller sample sizes, lower propagation rates due to lower item endorsement, and deeper node depths. Conditions that result in fewer observations per person for a given node result in greater estimate uncertainty, such as shorter test lengths and deeper node depths. Researchers should be cognizant of the number of observations available to a given auxiliary-item and person to ensure adequate estimation. For tasks such as item selection or prediction of external criteria with the person ability parameters, IRTrees may require large sample sizes, long test lengths, and short tree depths.

Appendix

1PL IRTrees

Item Difficulty Regression Model Results

Table 28. *1PL Item Difficulty Distributional Regression Model Predictor Parameters of the Estimate Bias Mean.*

Predictor	*M*	*SD*	Highest Posterior Density Intervals						$\hat{R}$	Bulk ESS	Tail ESS
			2.5%	12.5%	25%	75%	87.5%	97.5%			
Intercept	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.000	3611.17	4568.28
D	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.001	6229.07	4592.93
S	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.000	3636.62	4252.22
T	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.000	4704.59	5465.16
P	0.00	0.01	-0.01	-0.01	-0.01	0.00	0.00	0.01	1.000	9508.78	4760.59
D x S	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.001	5971.04	4019.62
D x T	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.002	6010.29	4669.91
S x T	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.000	5024.52	5101.49
D x P	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	1.001	10079.25	5023.35
S x P	0.00	0.01	-0.01	0.00	0.00	0.01	0.01	0.01	1.000	10477.38	4444.23
T x P	0.00	0.01	-0.01	0.00	0.00	0.01	0.01	0.01	1.000	10246.41	4852.82
D x S x T	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.001	6273.96	4835.46
D x S x P	0.00	0.00	-0.01	0.00	0.00	0.00	0.00	0.00	1.000	10088.07	4478.37
D x T x P	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.000	9791.61	5083.19
S x T x P	0.00	0.01	-0.01	-0.01	0.00	0.00	0.00	0.01	1.000	10094.16	4490.09
D x S x T x P	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.000	10090.12	4976.06
τ	0.02	0.00	0.02	0.02	0.02	0.02	0.02	0.02	1.000	2138.83	3760.54

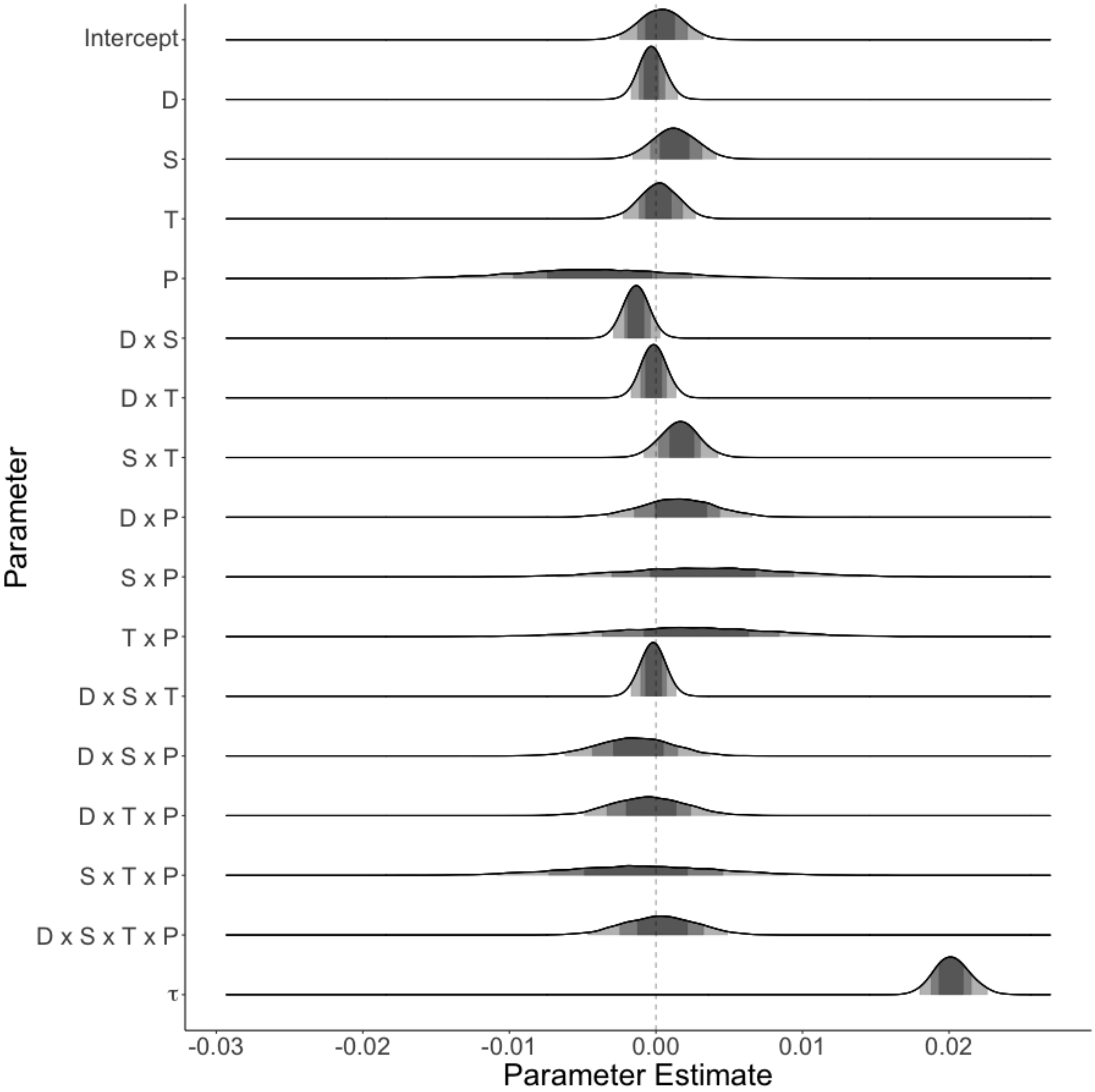
Note. D = node depth, S = sample size, T = test length, P = log-relative propagation, τ = between-simulation iteration intercept variance, ESS = Effective Sample Size.

Table 29. *1PL Item Difficulty Distributional Regression Model Predictor Parameters of the Estimate Bias Variability.*

Predictor	*M*	*SD*	Highest Posterior Density Intervals						$\hat{R}$	Bulk ESS	Tail ESS
			2.5%	12.5%	25%	75%	87.5%	97.5%			
Intercept	-2.54	0.01	-2.56	-2.55	-2.55	-2.54	-2.53	-2.53	1.002	10800.21	4938.87
D	0.34	0.00	0.34	0.34	0.34	0.35	0.35	0.35	1.003	10791.49	5060.76
S	-0.36	0.01	-0.37	-0.37	-0.36	-0.35	-0.35	-0.34	1.000	11885.23	4740.79
T	0.00	0.01	-0.02	-0.01	-0.01	0.00	0.00	0.01	1.000	11887.12	4291.98
P	-0.42	0.02	-0.46	-0.44	-0.43	-0.40	-0.39	-0.38	1.000	10690.90	4219.32
D x S	0.01	0.00	0.00	0.00	0.01	0.01	0.01	0.02	1.000	10535.88	4600.26
D x T	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	1.000	10476.93	4340.42
S x T	0.00	0.01	-0.02	-0.01	-0.01	0.00	0.00	0.01	1.002	11986.14	4400.92
D x P	0.02	0.01	0.00	0.01	0.01	0.02	0.03	0.03	1.000	10795.45	4425.64
S x P	0.04	0.02	0.00	0.02	0.03	0.05	0.06	0.08	1.001	9718.84	4370.00
T x P	-0.01	0.02	-0.05	-0.03	-0.02	0.01	0.02	0.04	1.000	10450.08	4334.58
D x S x T	0.00	0.00	-0.01	0.00	0.00	0.00	0.01	0.01	1.001	9956.08	4659.77
D x S x P	-0.02	0.01	-0.04	-0.03	-0.02	-0.01	-0.01	0.00	1.000	9820.51	4245.21
D x T x P	0.00	0.01	-0.02	-0.01	0.00	0.01	0.01	0.02	1.000	10101.15	4442.13
S x T x P	-0.02	0.02	-0.06	-0.04	-0.04	-0.01	0.00	0.02	1.000	9593.08	4351.82
D x S x T x P	0.01	0.01	-0.01	0.00	0.00	0.02	0.02	0.03	1.001	9190.03	4182.62
τ	0.02	0.01	0.00	0.01	0.02	0.03	0.03	0.04	1.002	1268.63	1937.88

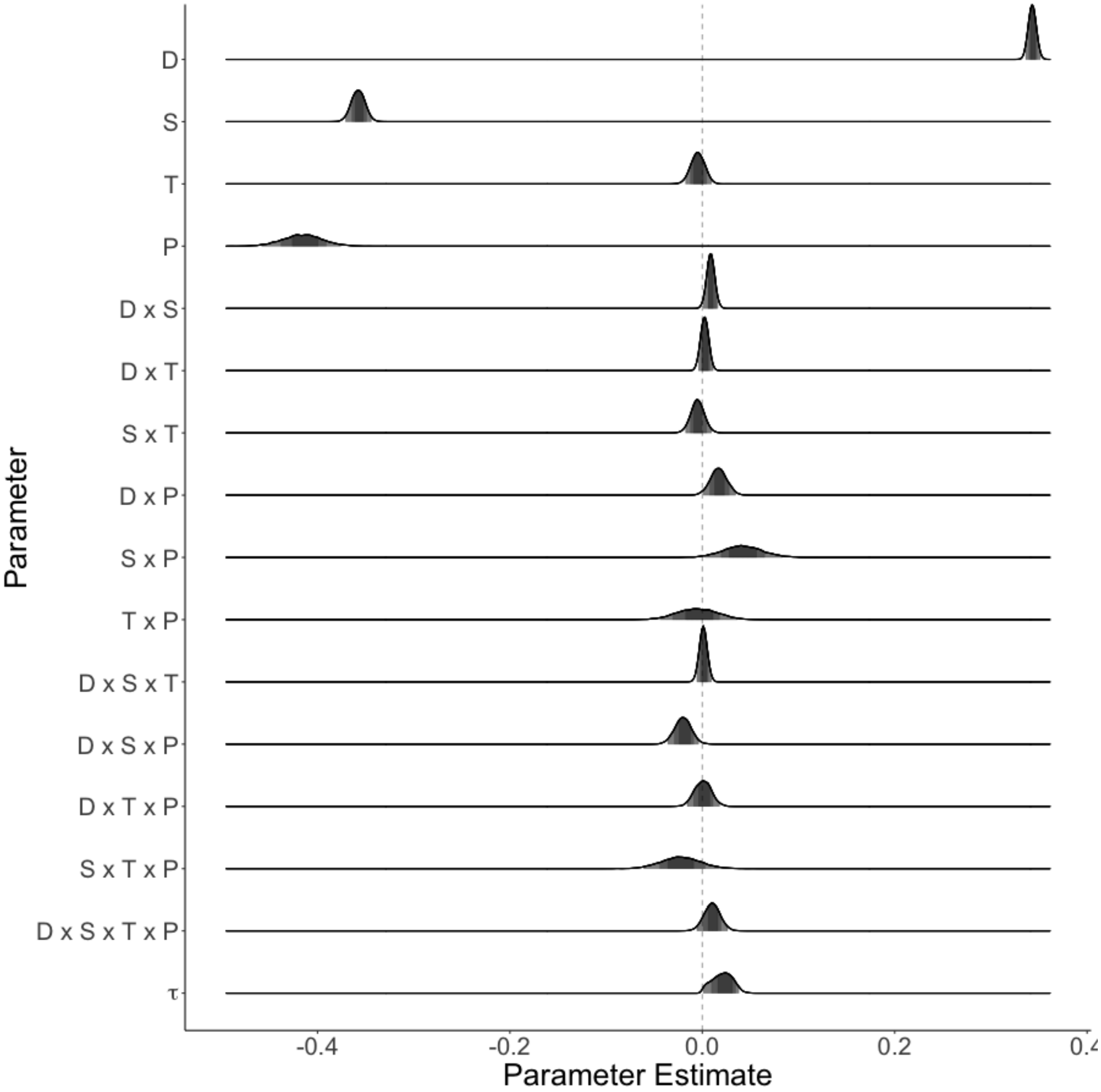
Note. D = node depth, S = sample size, T = test length, P = log-relative propagation, τ = between-simulation iteration intercept variance, ESS = Effective Sample Size. Estimates are on the log scale.

Figure 25. *1PL Item Difficulty Estimate Bias, Model Parameter Posterior Distributions for Estimate Mean Bias.*



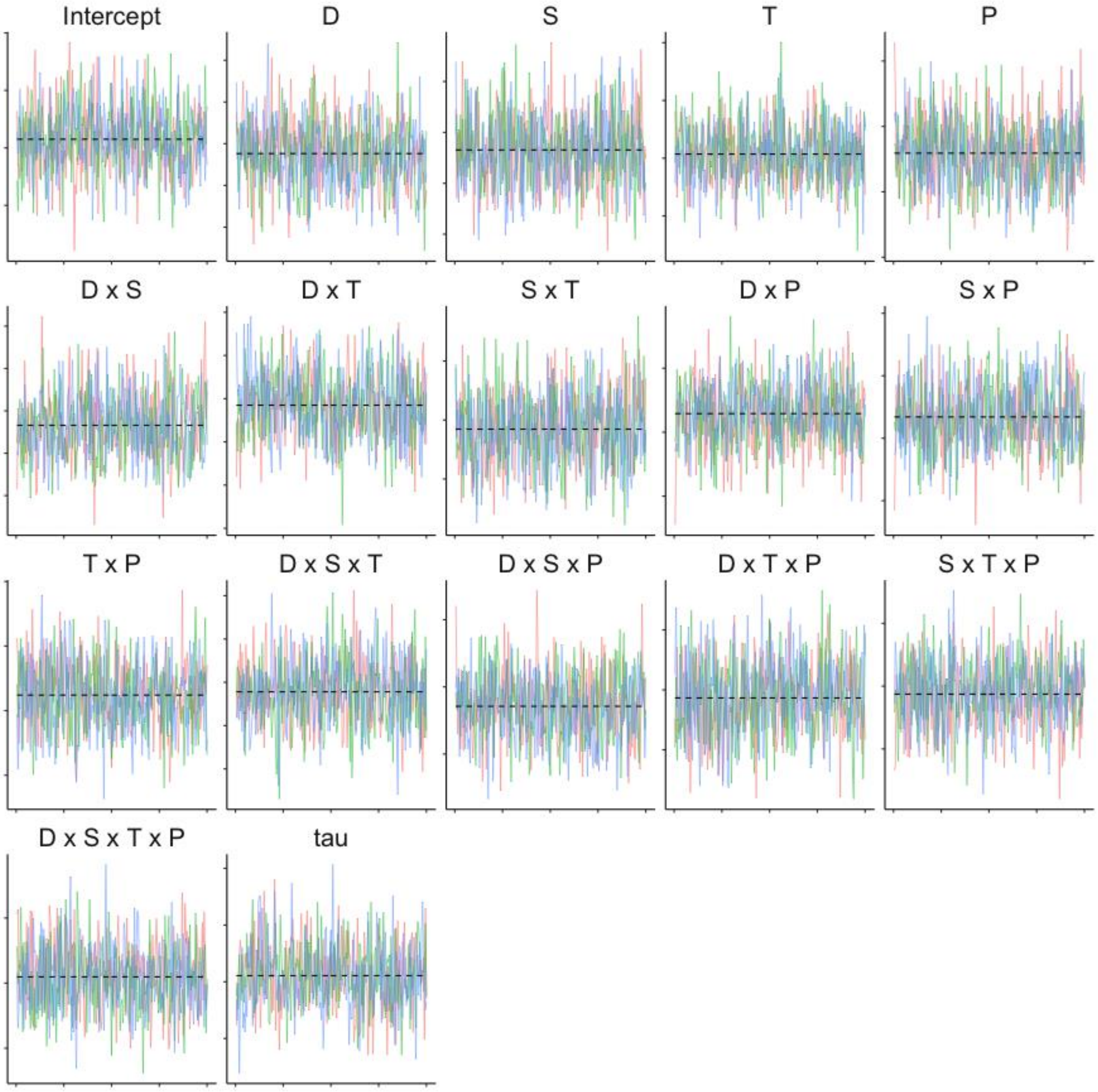
Note. 95%, 75%, and 50% highest posterior density intervals are represented with light, medium, and dark shades of grey, respectively.

Figure 26. *1PL Item Difficulty Estimate Bias, Model Parameter Posterior Distributions for Estimate Variability.*



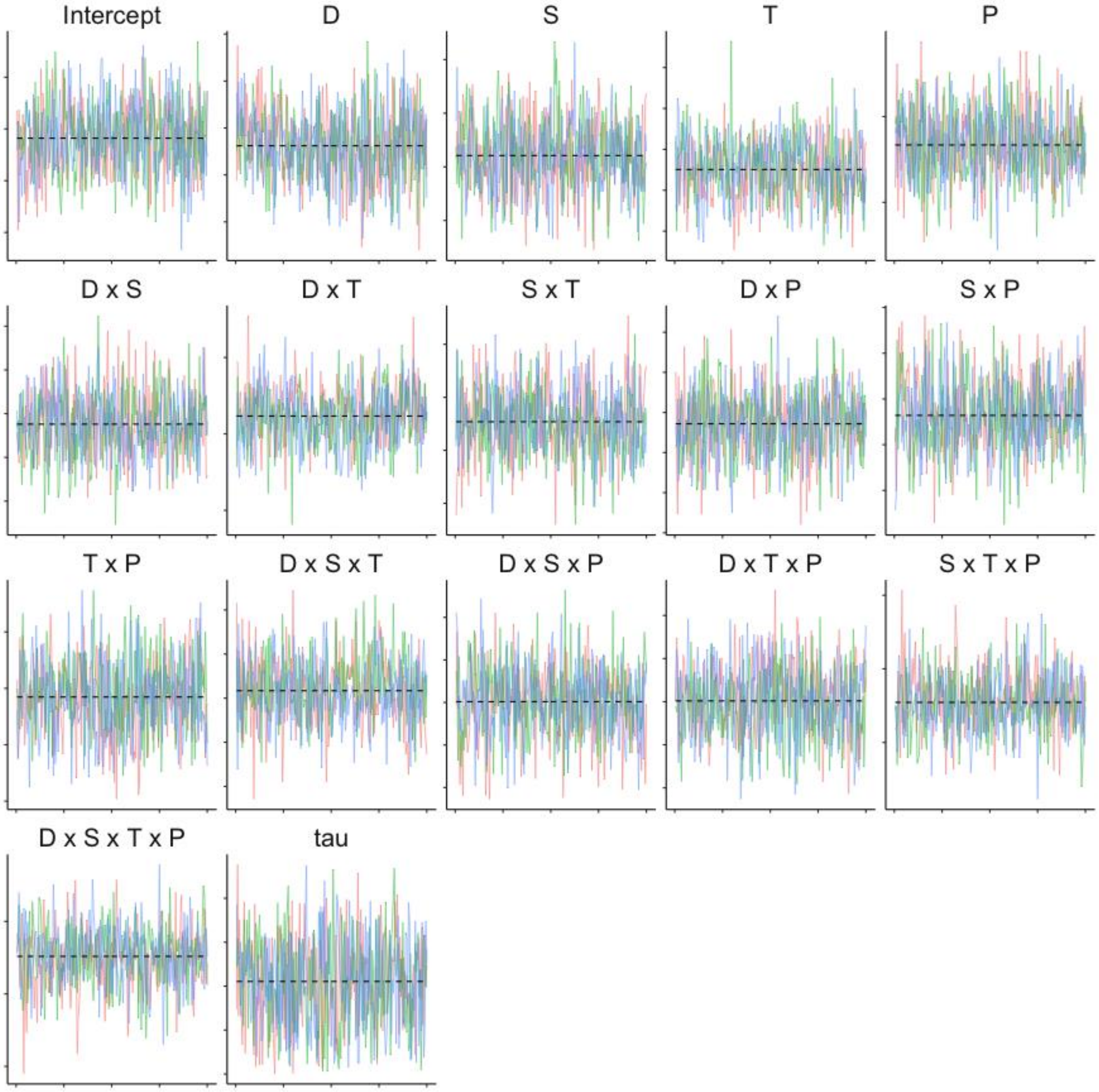
Note. 95%, 75%, and 50% highest posterior density intervals are represented with light, medium, and dark shades of grey, respectively. The intercept is not included because it was too distant from the other parameter to display properly.

Figure 27. *1PL Item Difficulty Estimate Bias, Model Parameter Trace Plot for Estimate Mean Bias.*



Note. Each chain was thinned by using every 10th draw to facilitate visualization.

Figure 28. *1PL Item Difficulty Estimate Bias, Model Parameter Trace Plot for Estimate Variance.*



Note. Each chain was thinned by using every 10th draw to facilitate visualization.

Figure 29. *1PL Item Difficulty Estimate Bias, Model Parameter Scatter Plot for Estimate Mean Bias.*

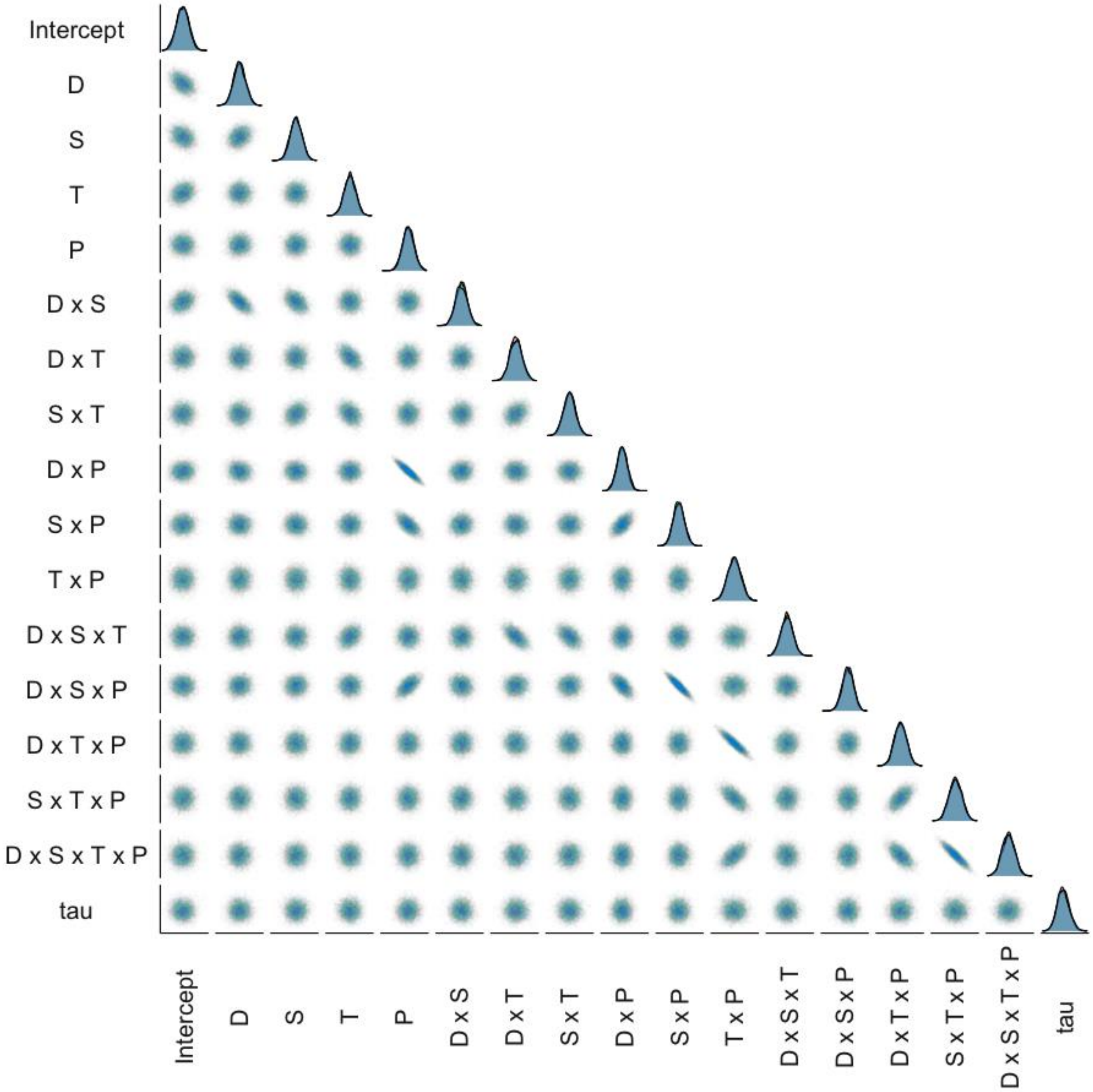
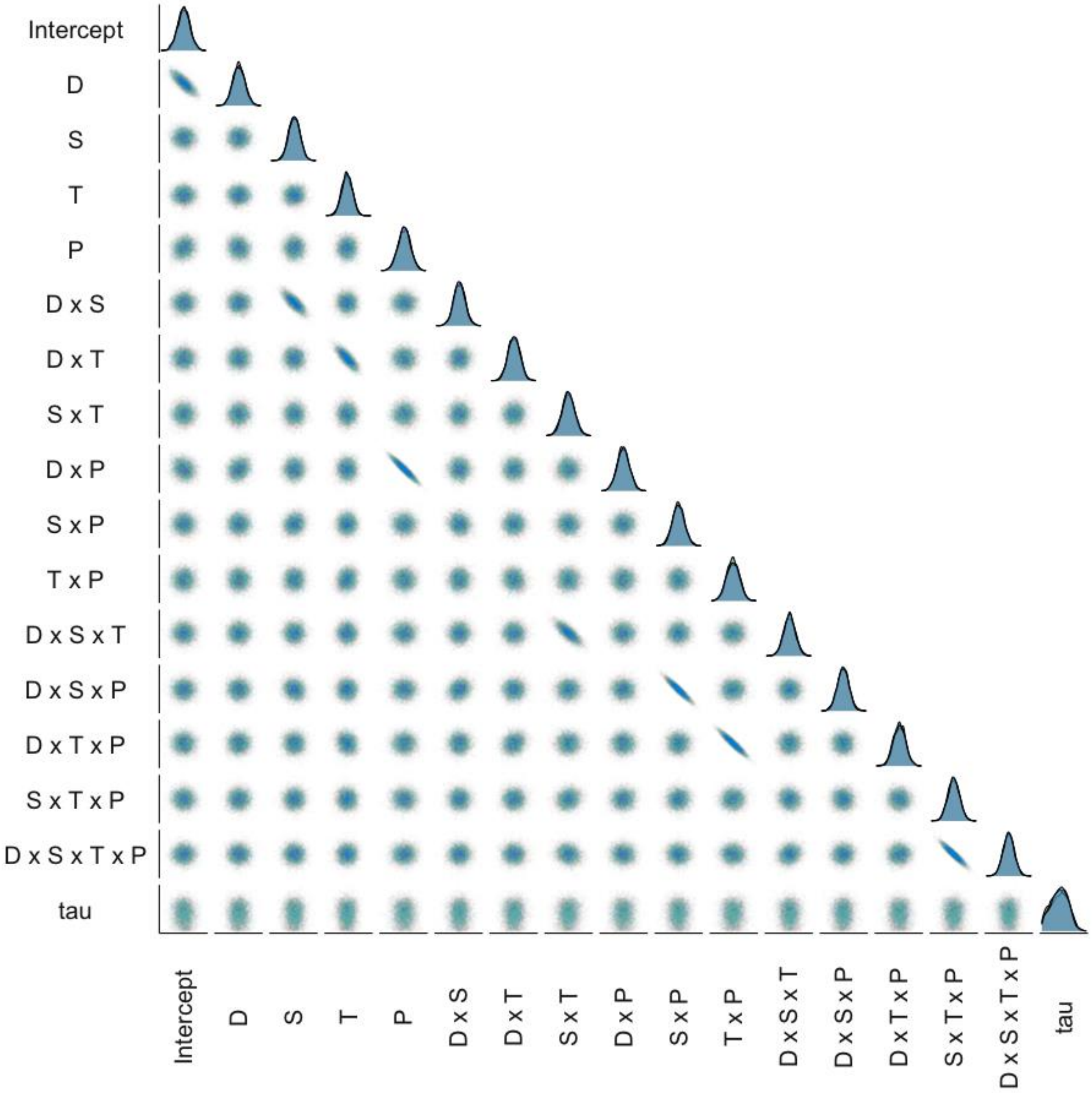


Figure 30. *1PL Item Difficulty Estimate Bias, Model Parameter Scatter Plot for Estimate Variability.*



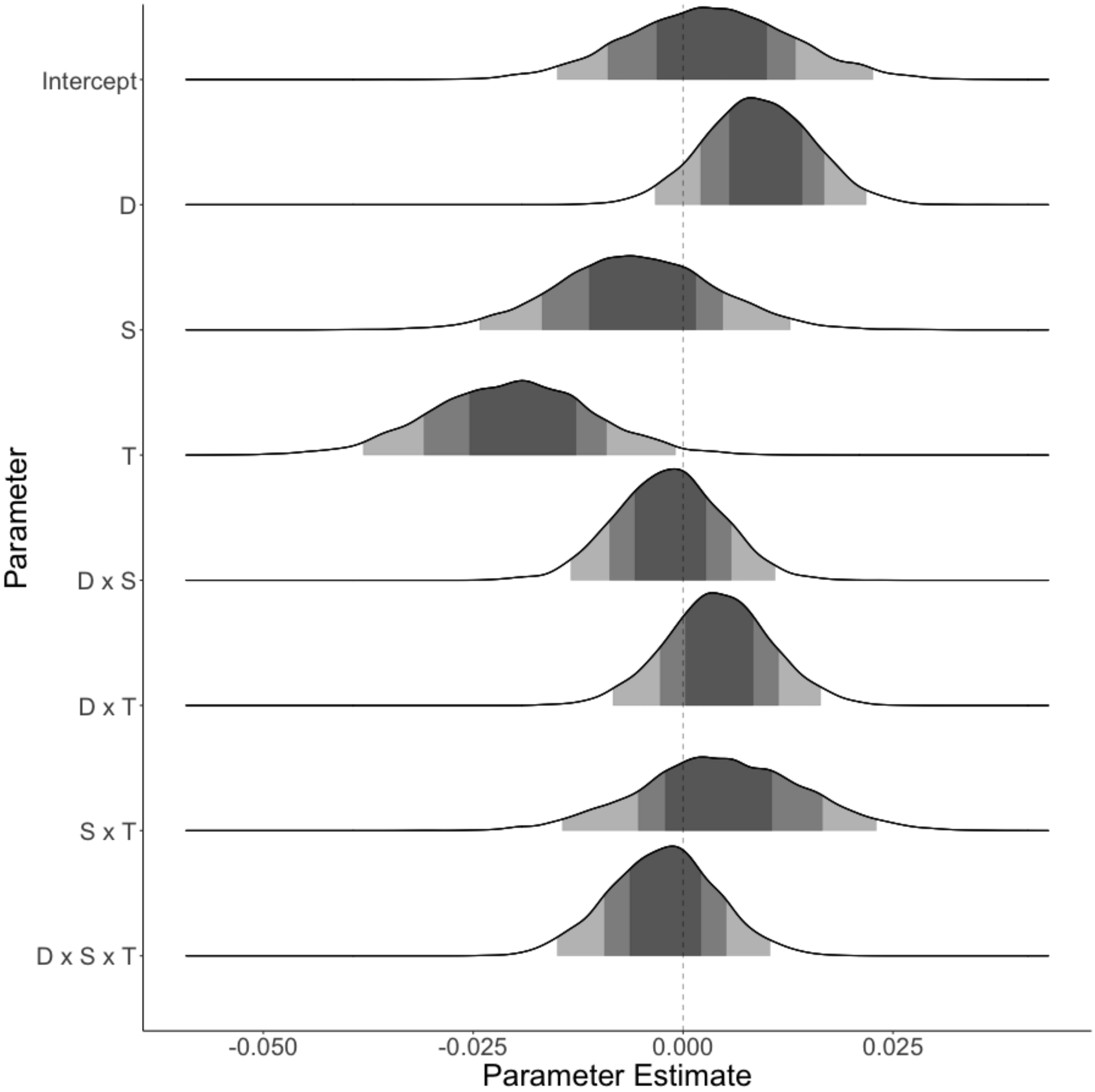
Person Ability Regression Model Results – Bias

Table 30. *1PL Person Ability Distributional Regression Model Predictor Parameters of the Estimate Bias Standard Deviation.*

Parameter	Predictor	*M*	*SD*	Highest Posterior Density Intervals						$\hat{R}$	Bulk ESS	Tail ESS
				2.5%	12.5%	25%	75%	87.5%	97.5%			
μ	Intercept	0.00	0.01	-0.02	-0.01	0.00	0.01	0.01	0.02	1.000	6441.01	4486.33
	D	0.01	0.01	0.00	0.00	0.01	0.01	0.02	0.02	1.000	6544.17	4741.71
	S	-0.01	0.01	-0.02	-0.02	-0.01	0.00	0.00	0.01	1.000	6819.87	5009.85
	T	-0.02	0.01	-0.04	-0.03	-0.03	-0.01	-0.01	0.00	1.000	6632.29	4808.03
	D x S	0.00	0.01	-0.01	-0.01	-0.01	0.00	0.01	0.01	1.000	7574.19	5168.98
	D x T	0.00	0.01	-0.01	0.00	0.00	0.01	0.01	0.02	1.000	6566.62	4030.88
	S x T	0.00	0.01	-0.01	-0.01	0.00	0.01	0.02	0.02	1.000	6968.38	4850.27
	D x S x T	0.00	0.01	-0.02	-0.01	-0.01	0.00	0.01	0.01	1.001	6859.05	4067.15
σ	Intercept	-0.67	0.01	-0.70	-0.69	-0.68	-0.66	-0.66	-0.65	1.000	8542.05	4296.49
	D	0.19	0.01	0.17	0.18	0.18	0.19	0.19	0.20	1.000	8071.87	4900.24
	S	-0.01	0.01	-0.03	-0.02	-0.01	0.00	0.01	0.02	1.000	8169.55	4670.44
	T	-0.21	0.01	-0.23	-0.22	-0.22	-0.20	-0.20	-0.19	1.000	8254.62	3984.04
	D x S	0.00	0.01	-0.01	-0.01	-0.01	0.00	0.01	0.01	1.000	8146.85	4383.94
	D x T	0.05	0.01	0.04	0.04	0.05	0.05	0.06	0.06	1.000	8224.66	4548.96
	S x T	0.01	0.01	-0.01	0.00	0.00	0.02	0.02	0.03	1.000	7898.63	4495.05
	D x S x T	0.00	0.01	-0.01	-0.01	-0.01	0.00	0.01	0.01	1.001	8225.72	3677.96

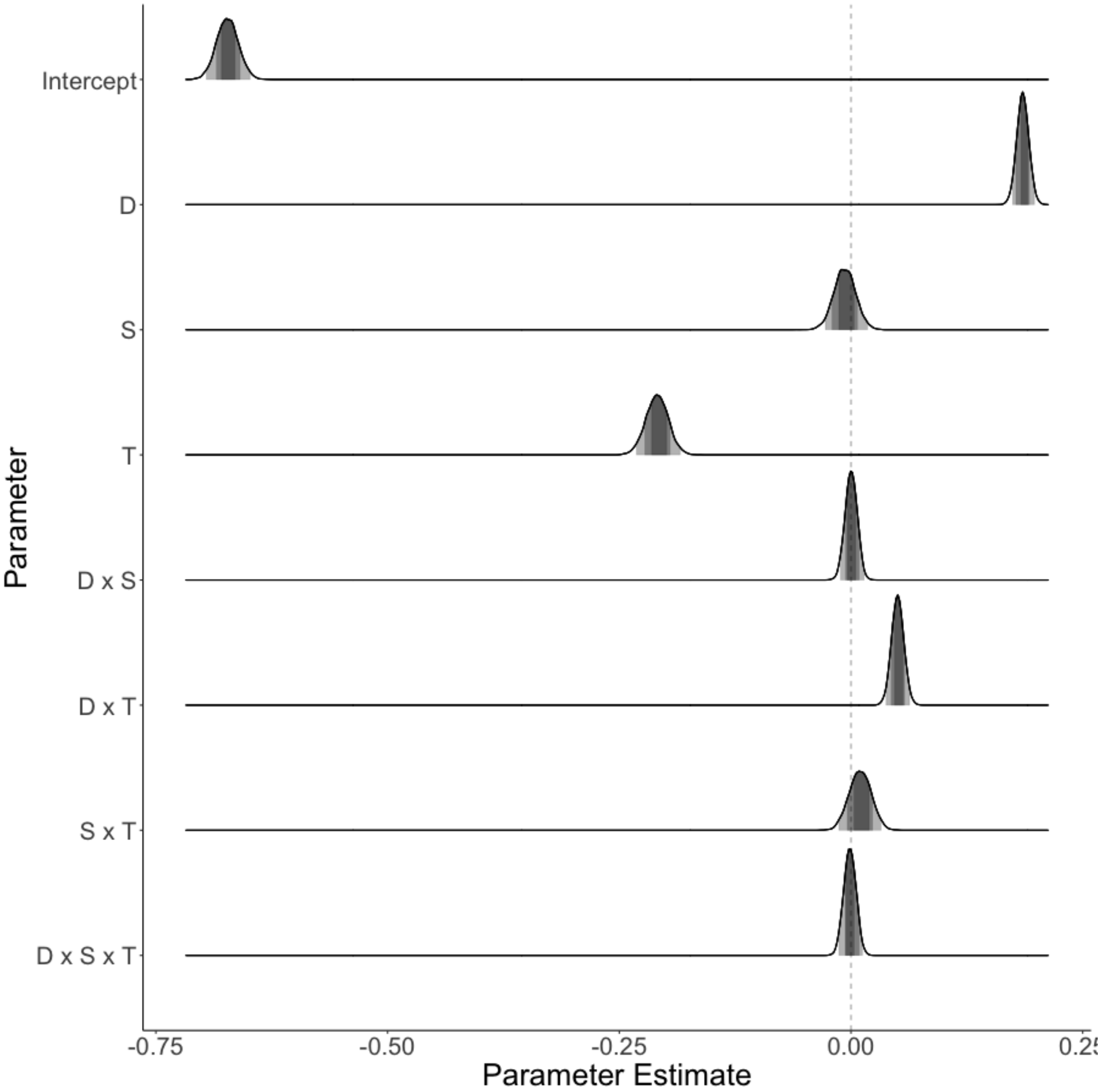
Note. D = node depth, S = sample size, T = test length, ESS = Effective Sample Size. σ predictor estimates are on the log scale.

Figure 31. *1PL Person Ability Estimate Bias, Model Parameter Posterior Distributions for Estimate Mean Bias.*



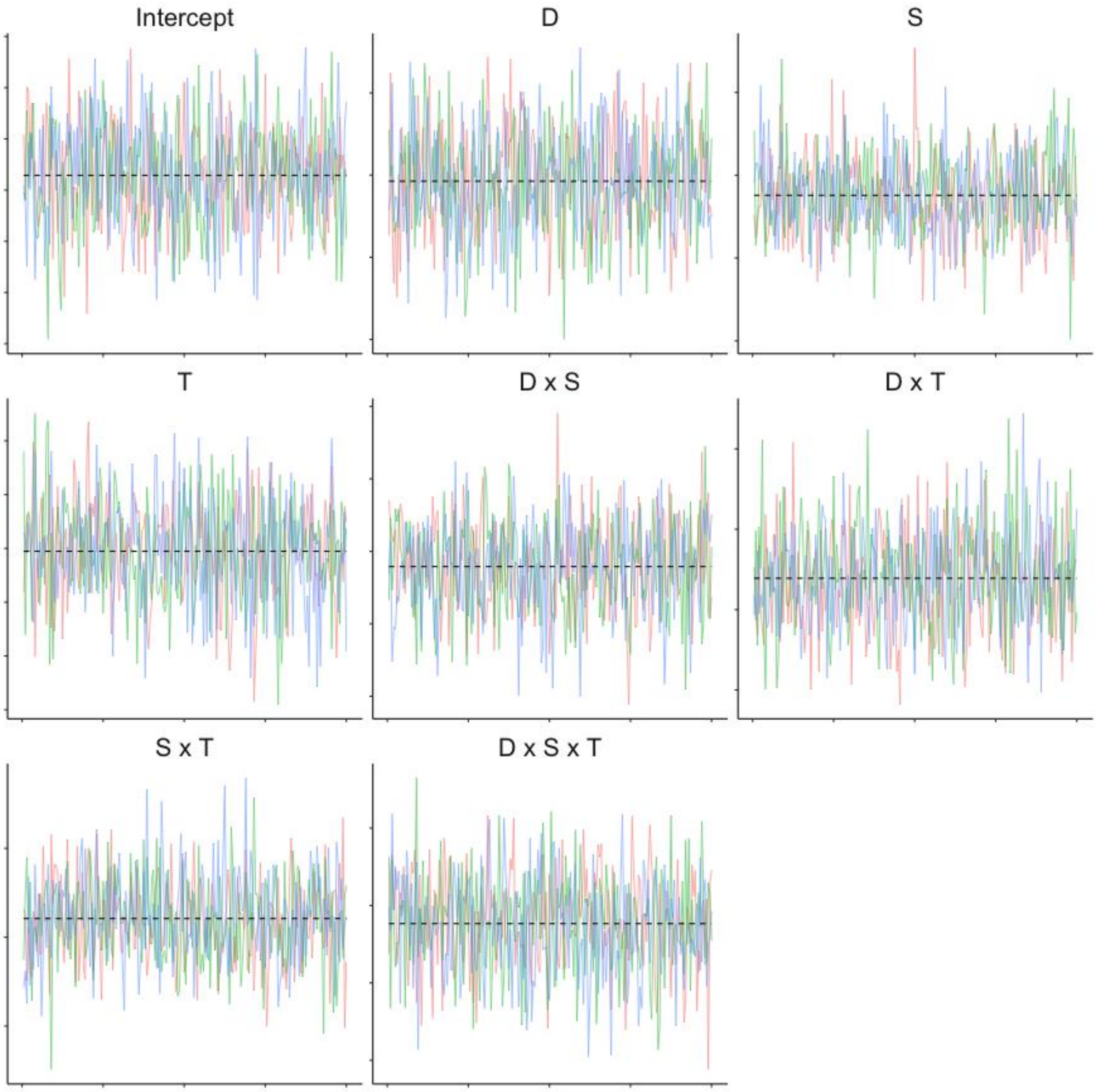
Note. 95%, 75%, and 50% highest posterior density intervals are represented with light, medium, and dark shades of grey, respectively.

Figure 32. *1PL Person Ability Estimate Bias, Model Parameter Posterior Distributions for Estimate Variability.*



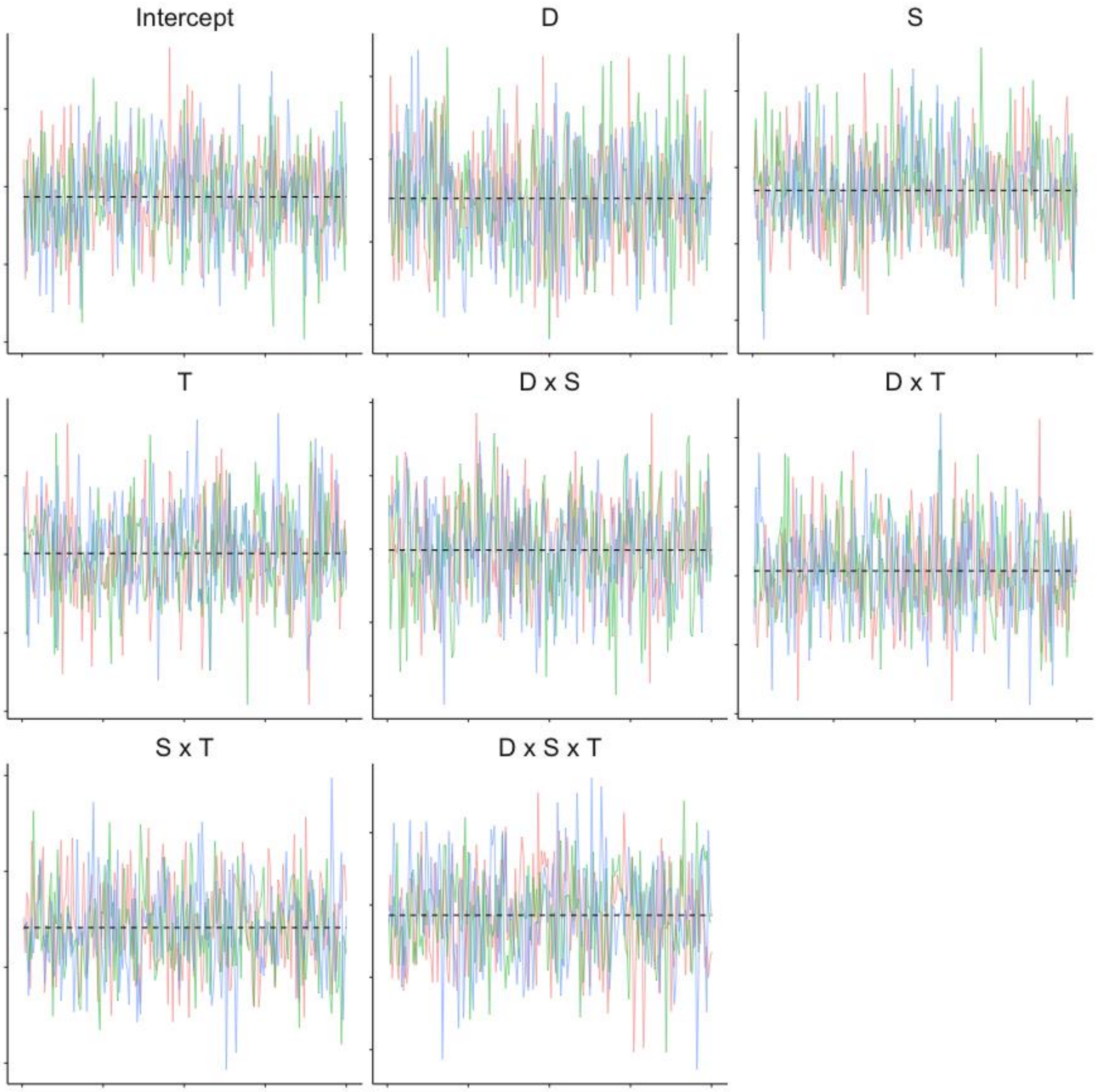
Note. 95%, 75%, and 50% highest posterior density intervals are represented with light, medium, and dark shades of grey, respectively.

Figure 33. *1PL Person Ability Estimate Bias, Model Parameter Trace Plot for Mean.*



Note. Each chain was thinned by using every 10th draw to facilitate visualization.

Figure 34. *1PL Person Ability Estimate Bias, Model Parameter Trace Plot for Variance.*



Note. Each chain was thinned by using every 10th draw to facilitate visualization.

Figure 35. *1PL Person Ability Estimate Bias, Model Parameter Scatter Plot for Estimate Mean Bias.*

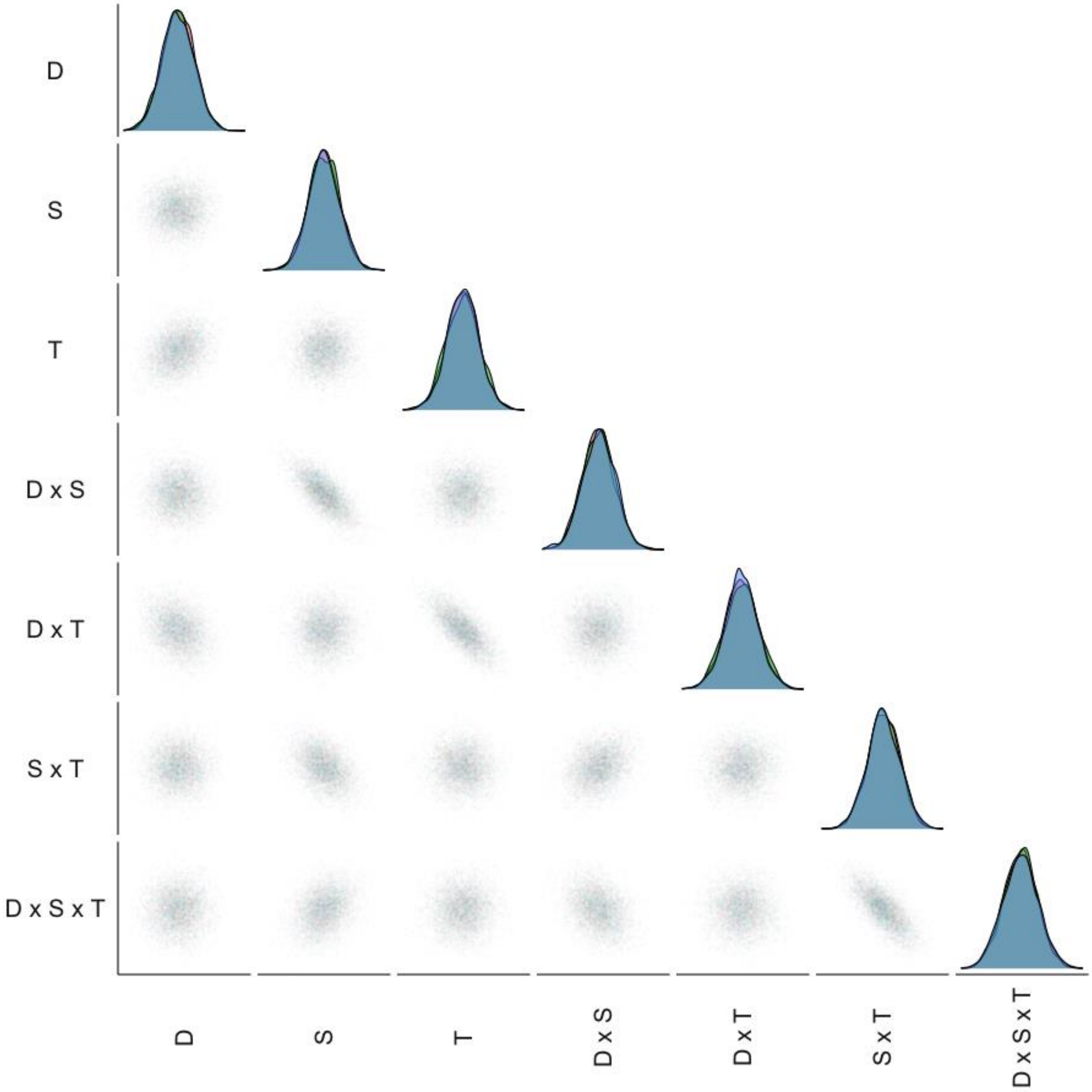
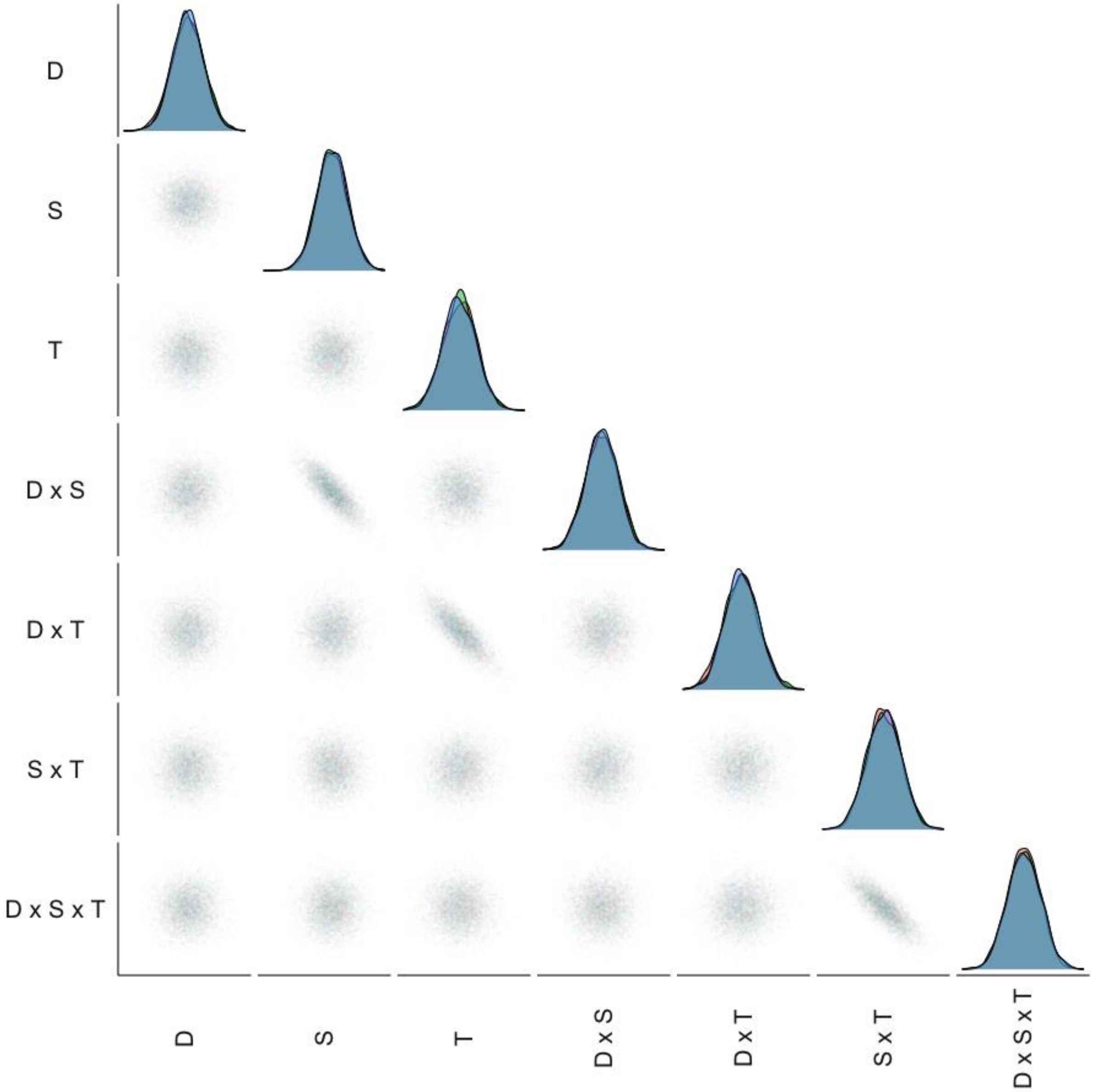


Figure 36*. Person Ability Estimate Bias, Model Parameter Scatter Plot for Variance.*



Person Ability Regression Model Results – True and Estimated Parameter Correlation

Table 31. *1PL Person Ability Distributional Beta Regression Model Predictor Parameters of the True and Estimated Parameter Correlations.*

Parameter	Predictor	*M*	*SD*	Highest Posterior Density Intervals						$\hat{R}$	Bulk ESS	Tail ESS
				2.5%	12.5%	25%	75%	87.5%	97.5%			
μ	Intercept	1.87	0.00	1.87	1.87	1.87	1.87	1.87	1.87	1.000	5672.74	4945.07
	S	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.000	5352.62	5029.42
	T	0.50	0.00	0.50	0.50	0.50	0.50	0.50	0.51	1.000	6003.49	5480.81
	D	-0.67	0.00	-0.67	-0.67	-0.67	-0.67	-0.67	-0.67	1.000	5888.16	5311.47
	S x T	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.000	5909.04	5174.76
	S x D	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.000	5412.41	4915.21
	T x D	-0.04	0.00	-0.04	-0.04	-0.04	-0.04	-0.04	-0.04	1.000	5901.73	5470.90
	S x T x D	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.000	5698.79	5403.82
ϕ	Intercept	6.93	0.00	6.93	6.93	6.93	6.93	6.93	6.93	1.000	3750.93	3108.99
	S	0.29	0.00	0.29	0.29	0.29	0.29	0.29	0.29	1.000	3243.49	3654.19
	T	0.67	0.00	0.67	0.67	0.67	0.68	0.68	0.68	1.000	3407.59	3025.25
	D	-0.79	0.00	-0.80	-0.80	-0.79	-0.79	-0.79	-0.79	1.000	3981.42	3765.37
	S x T	-0.02	0.00	-0.03	-0.03	-0.02	-0.02	-0.02	-0.02	1.000	3713.94	4074.13
	S x D	-0.04	0.00	-0.04	-0.04	-0.04	-0.04	-0.04	-0.04	1.000	3252.63	4052.21
	T x D	-0.13	0.00	-0.13	-0.13	-0.13	-0.13	-0.13	-0.13	1.000	3777.78	3602.84
	S x T x D	-0.01	0.00	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	1.000	3904.52	4521.92

Note. D = node depth, S = sample size, T = test length, ESS = Effective Sample Size. μ predictor estimates are on the logit scale, and ϕ predictor estimates are on the log scale.

Figure 37. 1PL Person Ability True and Estimated Parameter Correlations, Model Parameter Posterior Distributions for Predictors of Correlation Location.

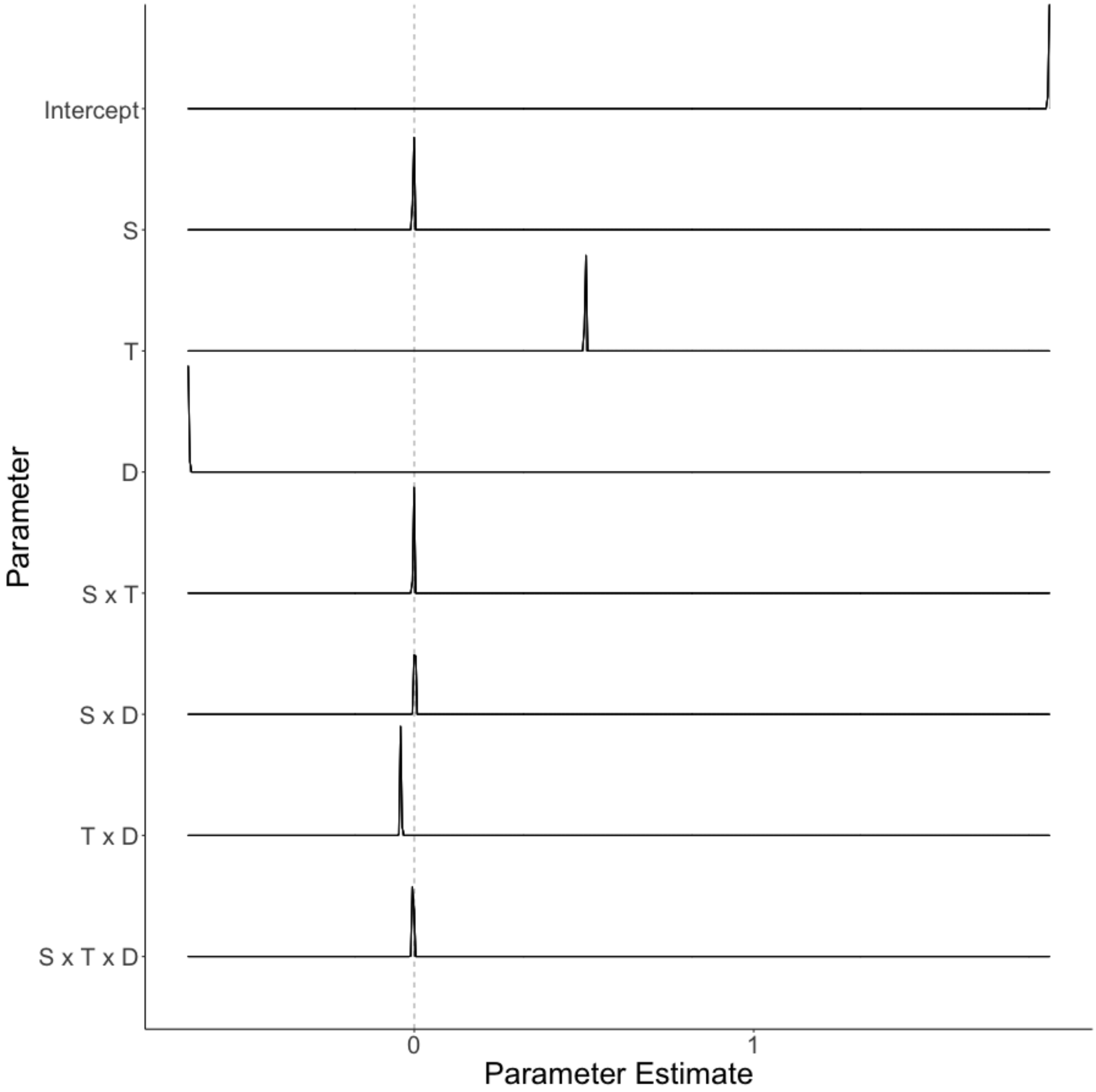


Figure 38. *1PL Person Ability True and Estimated Parameter Correlations, Model Parameter Posterior Distributions for Predictors of Correlation Scale.*

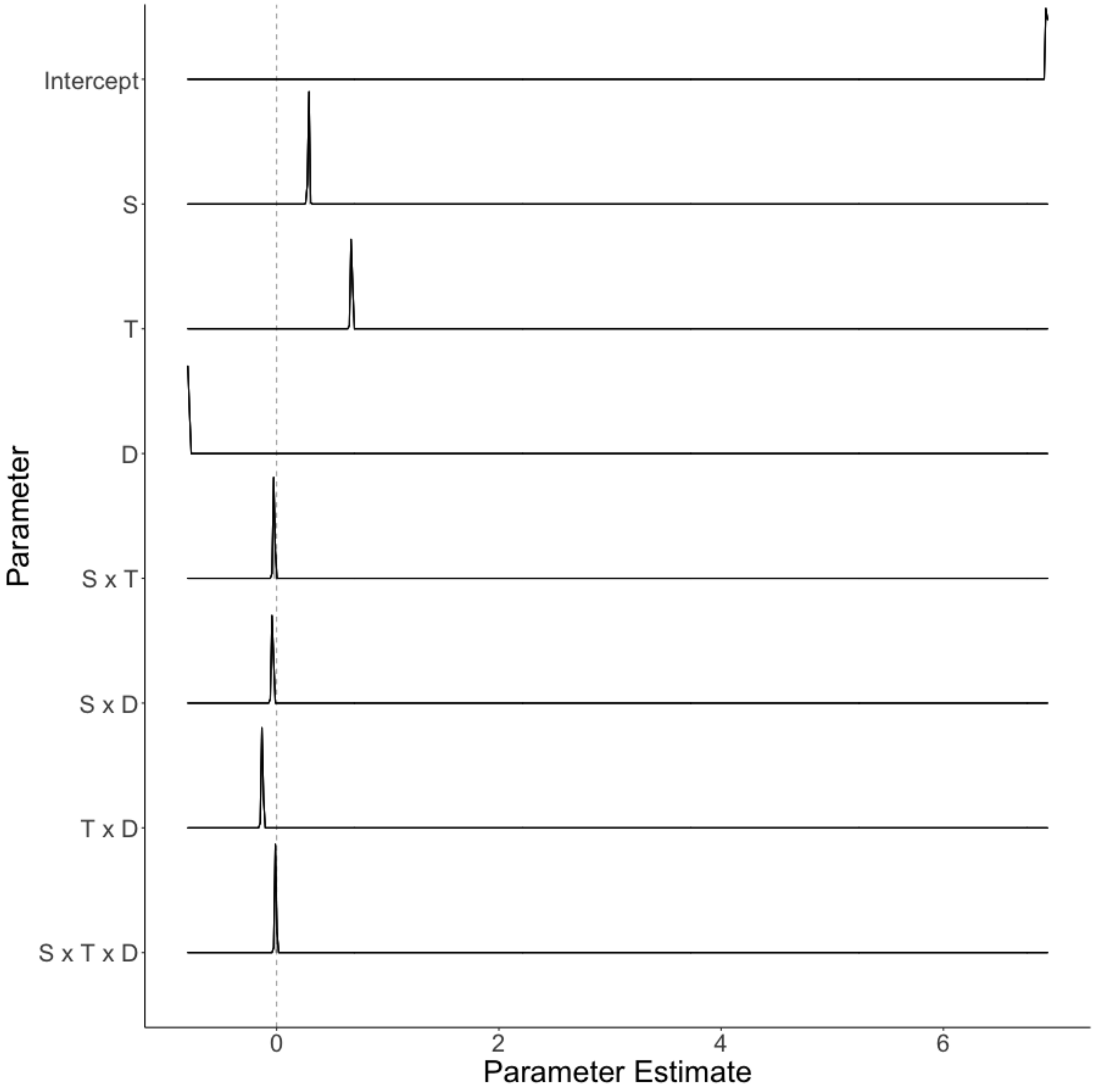
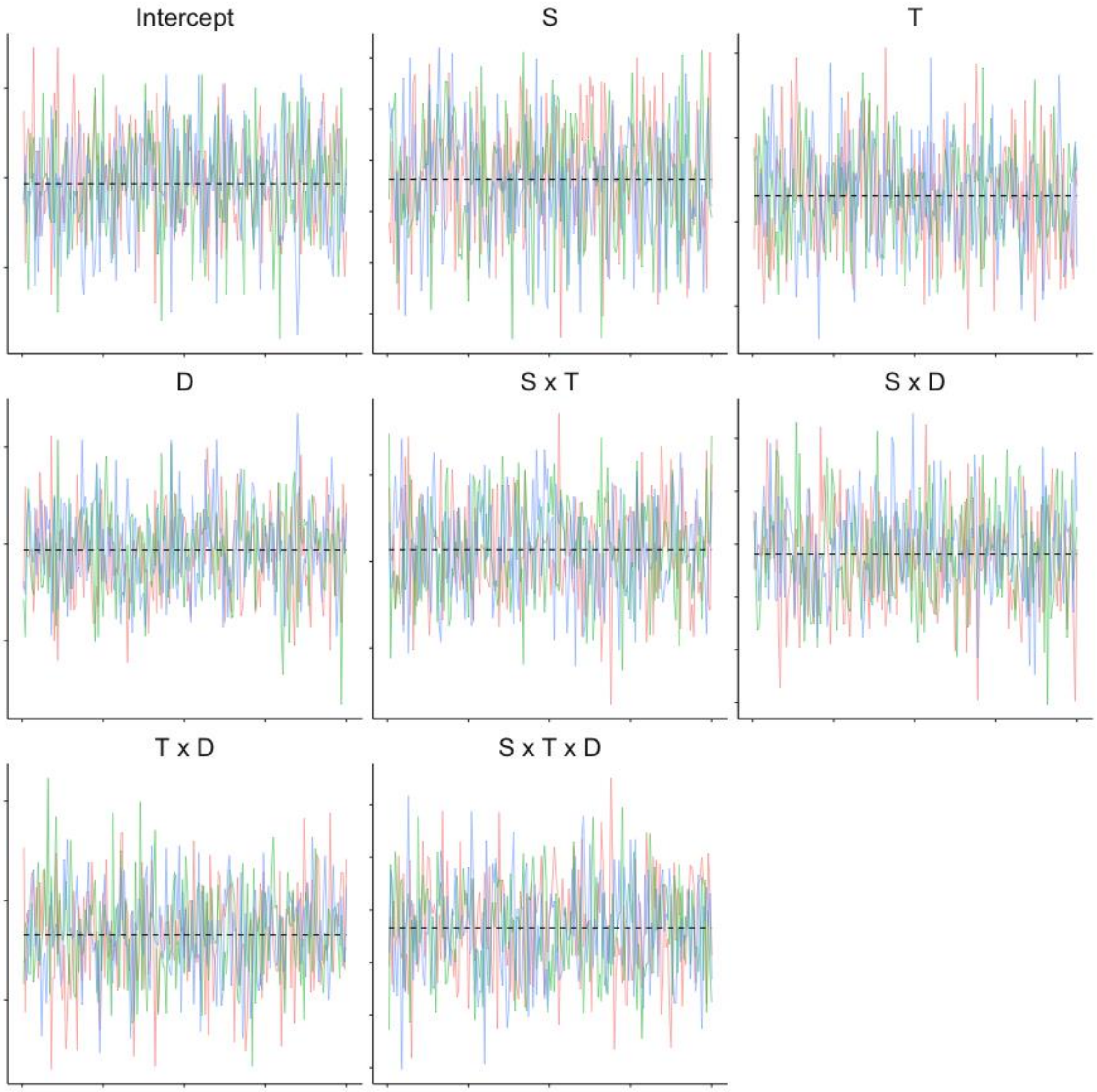
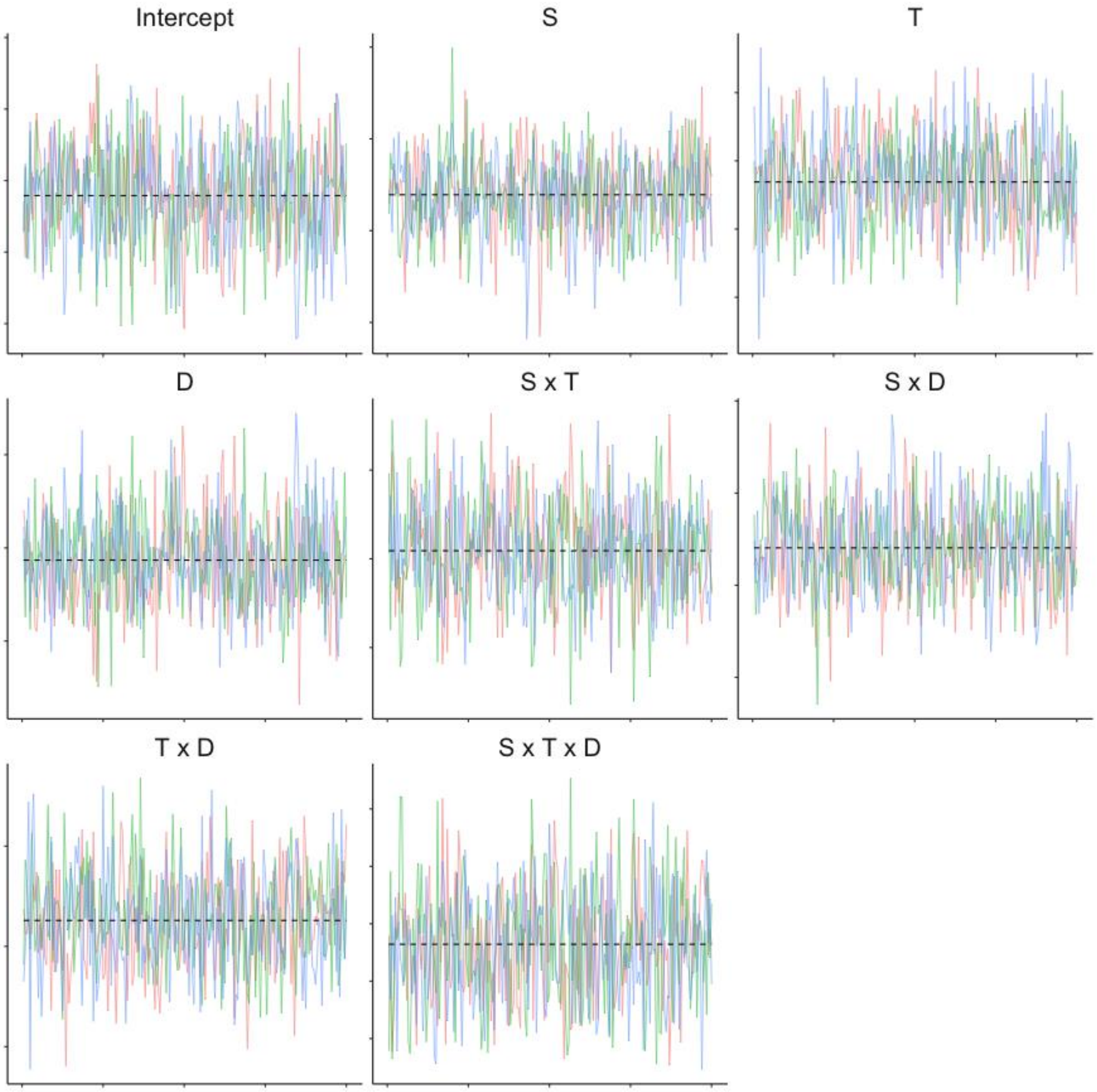


Figure 39. *1PL Person Ability True and Estimated Parameter Correlations, Model Parameter Trace Plot for Location.*



Note. Each chain was thinned by using every 10th draw to facilitate visualization.

Figure 40. *1PL Person Ability True and Estimated Parameter Correlations, Model Parameter Trace Plot for Scale.*



Note. Each chain was thinned by using every 10th draw to facilitate visualization.

Figure 41. *1PL Person Ability True and Estimated Parameter Correlations, Model Parameter Scatter Plot for Location.*

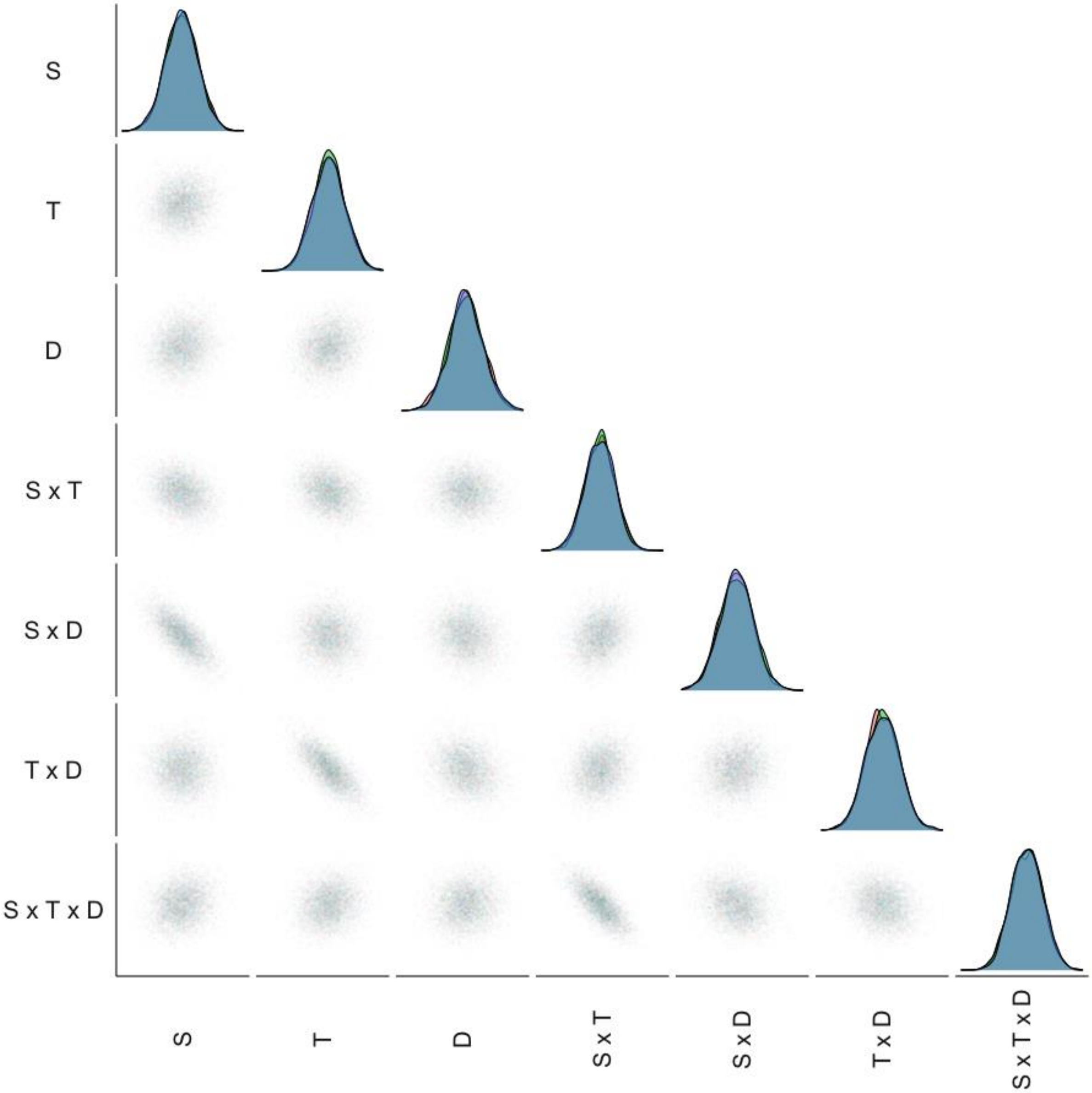
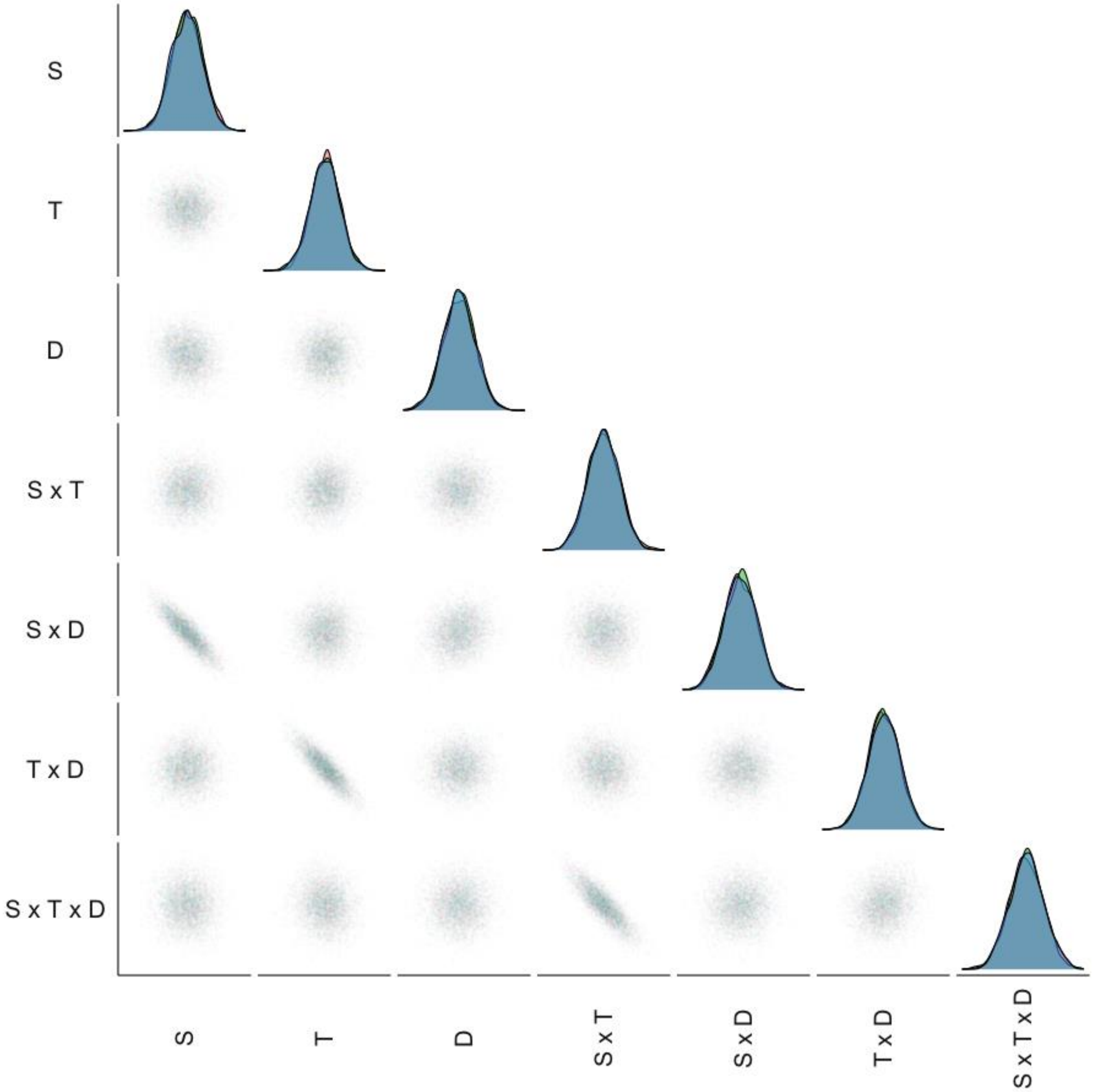


Figure 42. *1PL Person Ability True and Estimated Parameter Correlations, Model Parameter Scatter Plot for Scale.*



2PL IRTrees

Item Difficulty Regression Model Results

Table 32. *2PL Item Difficulty Distributional Regression Model Predictor Parameters of the Estimate Bias.*

Predictor	*M*	*SD*	Highest Posterior Density Intervals						$\hat{R}$	Bulk ESS	Tail ESS
			2.5%	12.5%	25%	75%	87.5%	97.5%			
Intercept	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.000	6604.99	5072.26
D	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.000	5995.59	5222.66
S	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.000	6374.65	5075.96
T	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.000	6587.43	5482.53
P	0.01	0.00	0.00	0.00	0.00	0.01	0.01	0.01	1.000	10831.35	4988.38
D x S	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.000	6356.90	5607.97
D x T	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.000	6224.39	4845.27
S x T	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.000	6074.45	4886.75
D x P	0.00	0.00	-0.01	0.00	0.00	0.00	0.00	0.00	1.001	9850.38	5073.56
S x P	-0.01	0.00	-0.01	-0.01	-0.01	0.00	0.00	0.00	1.000	10865.64	5123.97
T x P	0.00	0.00	-0.01	0.00	0.00	0.00	0.00	0.01	1.000	10517.02	4690.75
D x S x T	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.000	6293.31	4925.87
D x S x P	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	1.002	9631.66	5099.55
D x T x P	0.00	0.00	-0.01	0.00	0.00	0.00	0.00	0.01	1.000	10302.68	4666.86
S x T x P	0.00	0.00	-0.01	-0.01	-0.01	0.00	0.00	0.00	1.000	11065.92	4878.89
D x S x T x P	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.01	1.000	10197.66	5247.92
τ	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.001	2231.61	2725.68

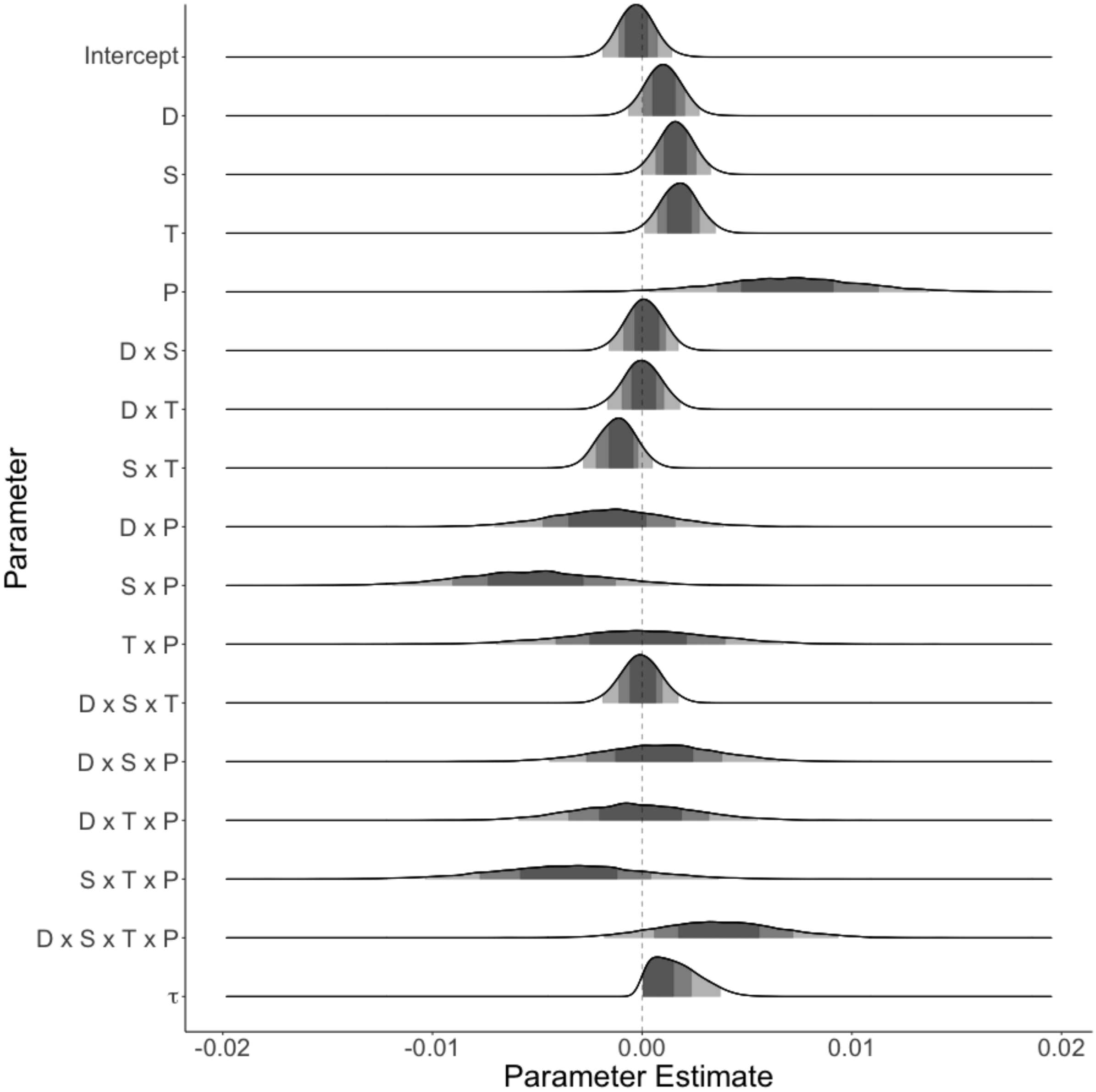
Note. D = node depth, S = sample size, T = test length, P = log-relative propagation, τ = between-simulation iteration intercept variance, ESS = Effective Sample Size.

Table 33. *2PL Item Difficulty Distributional Regression Model Predictor Parameters of the Estimate Variance.*

Predictor	*M*	*SD*	Highest Posterior Density Intervals						$\hat{R}$	Bulk ESS	Tail ESS
			2.5%	12.5%	25%	75%	87.5%	97.5%			
Intercept	-2.10	0.01	-2.11	-2.10	-2.10	-2.09	-2.09	-2.09	1.000	7073.59	4886.97
D	0.35	0.00	0.34	0.34	0.35	0.35	0.35	0.36	1.001	10733.97	4508.22
S	-0.33	0.01	-0.34	-0.33	-0.33	-0.32	-0.32	-0.31	1.001	7004.38	4941.10
T	-0.03	0.01	-0.04	-0.04	-0.04	-0.03	-0.03	-0.02	1.000	8780.94	5217.56
P	-0.42	0.01	-0.44	-0.43	-0.42	-0.41	-0.40	-0.39	1.000	10171.98	4394.39
D x S	0.01	0.00	0.01	0.01	0.01	0.02	0.02	0.02	1.000	11118.98	4597.97
D x T	0.00	0.00	-0.01	-0.01	0.00	0.00	0.00	0.00	1.000	10115.33	4792.00
S x T	-0.01	0.01	-0.02	-0.02	-0.01	-0.01	0.00	0.00	1.000	7709.26	5475.35
D x P	0.04	0.01	0.02	0.03	0.03	0.04	0.05	0.06	1.000	9708.34	4448.12
S x P	-0.03	0.01	-0.06	-0.05	-0.04	-0.02	-0.02	0.00	1.001	11018.92	4783.91
T x P	0.00	0.01	-0.03	-0.02	-0.01	0.01	0.01	0.02	1.000	11463.21	4558.29
D x S x T	-0.01	0.00	-0.02	-0.01	-0.01	-0.01	0.00	0.00	1.000	9991.51	4749.80
D x S x P	0.01	0.01	-0.01	0.00	0.00	0.01	0.02	0.02	1.001	10119.24	4814.44
D x T x P	-0.01	0.01	-0.02	-0.02	-0.01	0.00	0.00	0.01	1.000	11363.50	4386.81
S x T x P	-0.02	0.01	-0.05	-0.04	-0.03	-0.01	-0.01	0.00	1.000	10652.12	4150.54
D x S x T x P	0.01	0.01	-0.01	0.00	0.00	0.01	0.02	0.03	1.000	10436.58	3894.67
τ	0.06	0.01	0.05	0.06	0.06	0.07	0.07	0.08	1.000	2039.47	3542.99

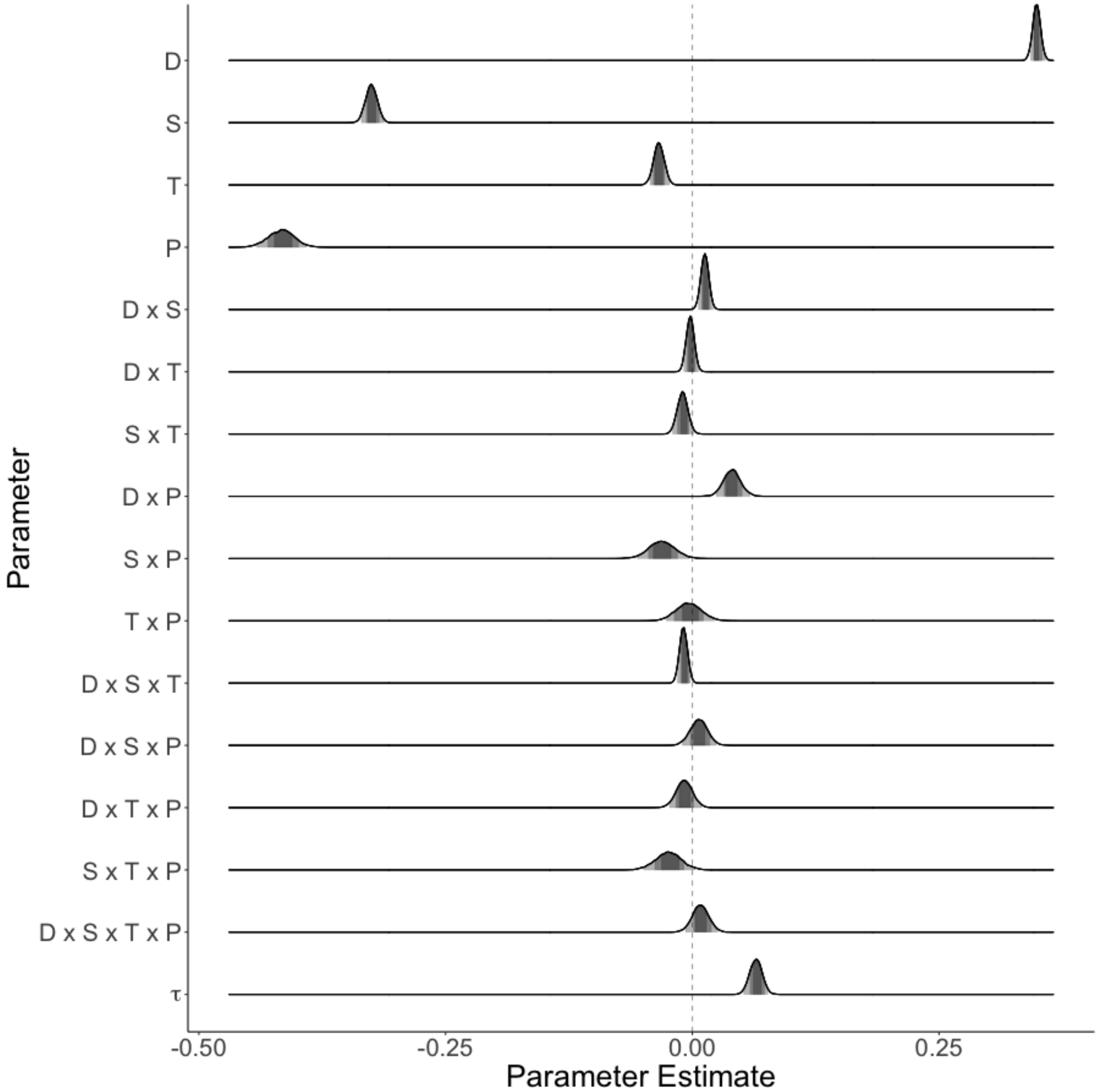
Note. D = node depth, S = sample size, T = test length, P = log-relative propagation, τ = between-simulation iteration intercept variance, ESS = Effective Sample Size. Estimates are on the log scale.

Figure 43. *2PL Item Difficulty Estimate Bias, Model Parameter Posterior Distributions for Mean Bias.*



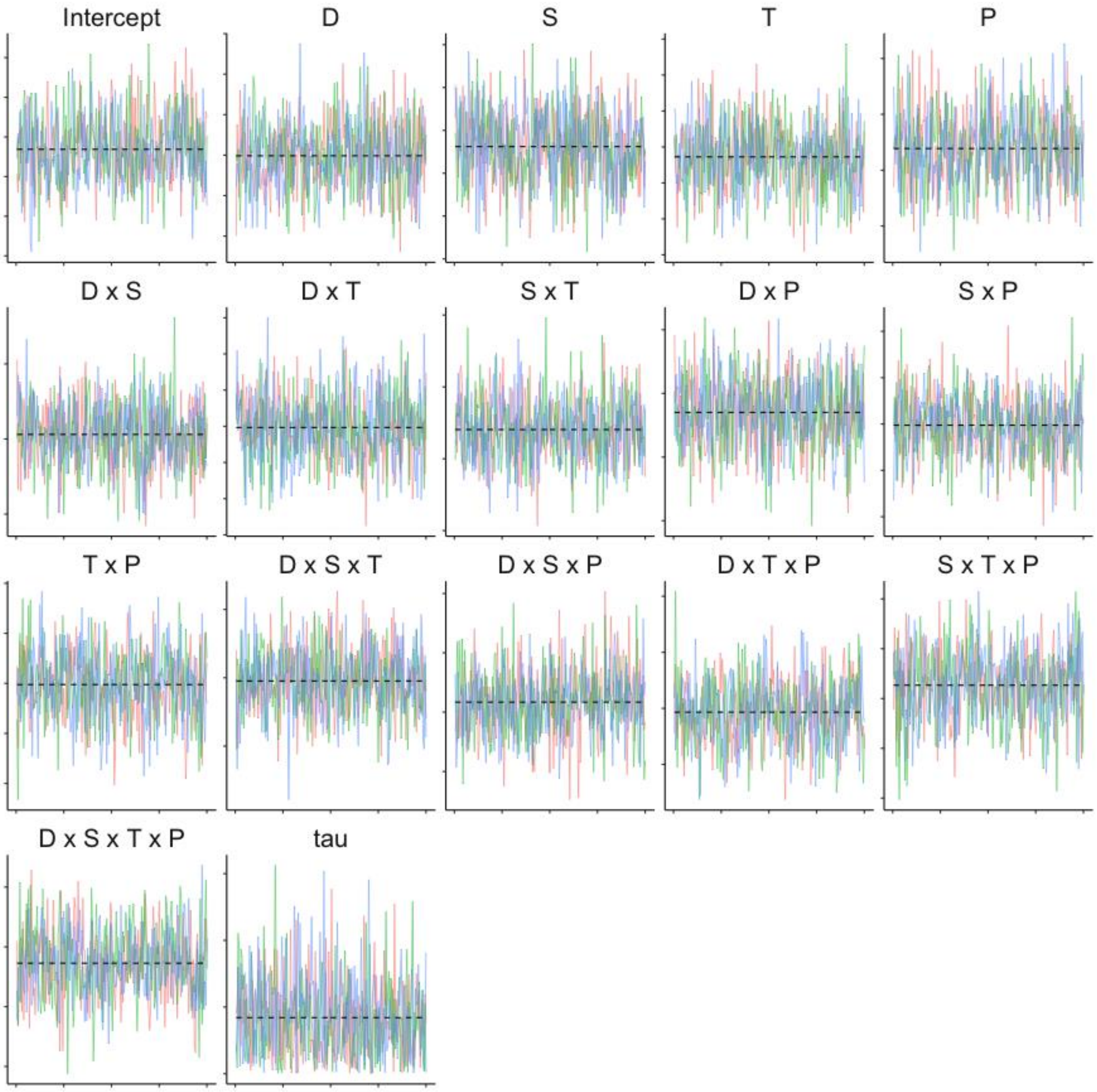
Note. 95%, 75%, and 50% highest posterior density intervals are represented with light, medium, and dark shades of grey, respectively.

Figure 44. *2PL Item Difficulty Estimate Bias, Model Parameter Posterior Distributions for Estimate Variance.*



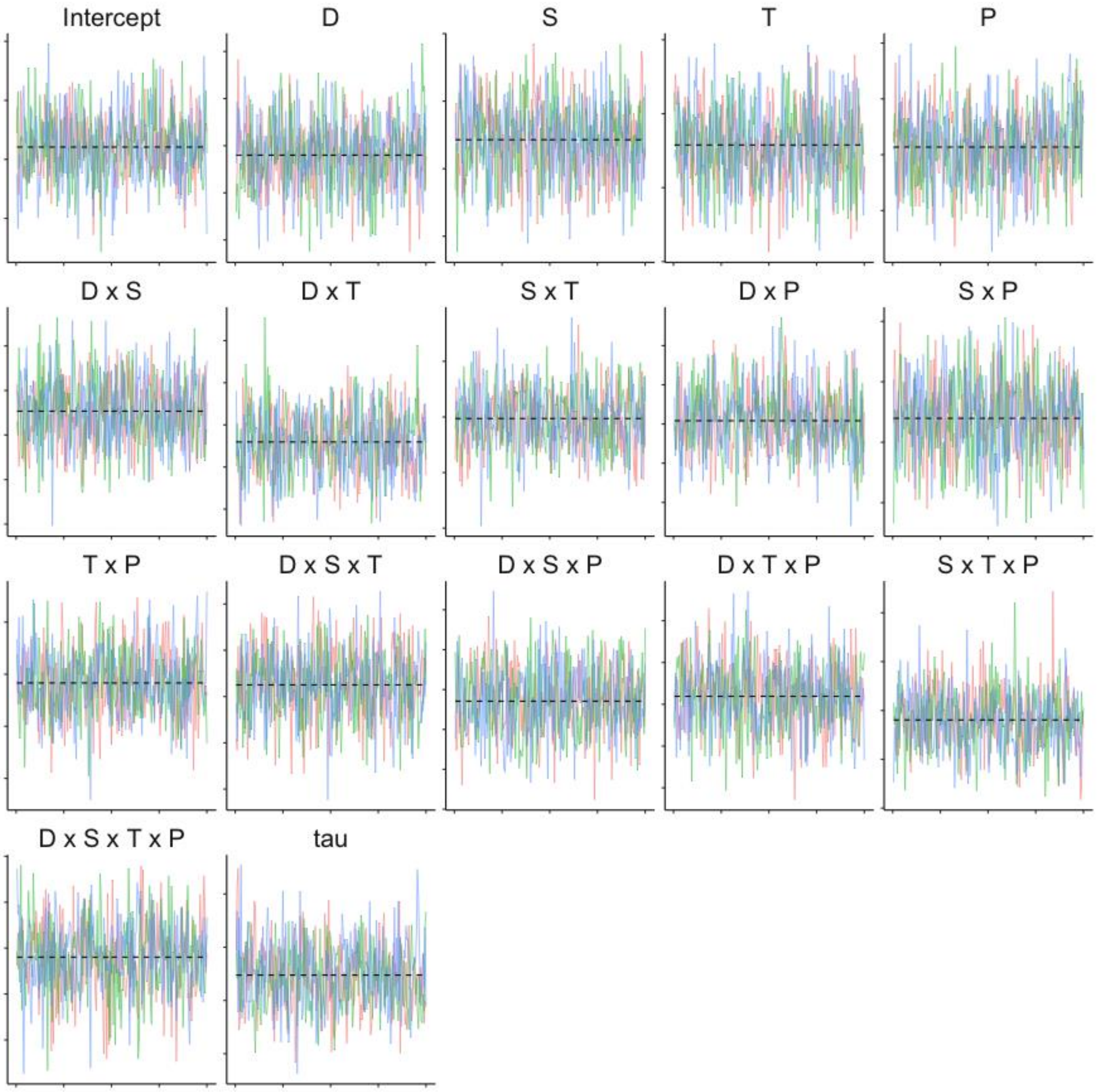
Note. 95%, 75%, and 50% highest posterior density intervals are represented with light, medium, and dark shades of grey, respectively.

Figure 45. *2PL Item Difficulty Estimate Bias, Model Parameter Trace Plot for Estimate Mean Bias.*



Note. Each chain was thinned by using every 10th draw to facilitate visualization.

Figure 46. *2PL Item Difficulty Estimate Bias, Model Parameter Trace Plot for Estimate Variance.*



Note. Each chain was thinned by using every 10th draw to facilitate visualization.

Figure 47. *2PL Item Difficulty Estimate Bias, Model Parameter Scatter Plot for Estimate Mean Bias.*

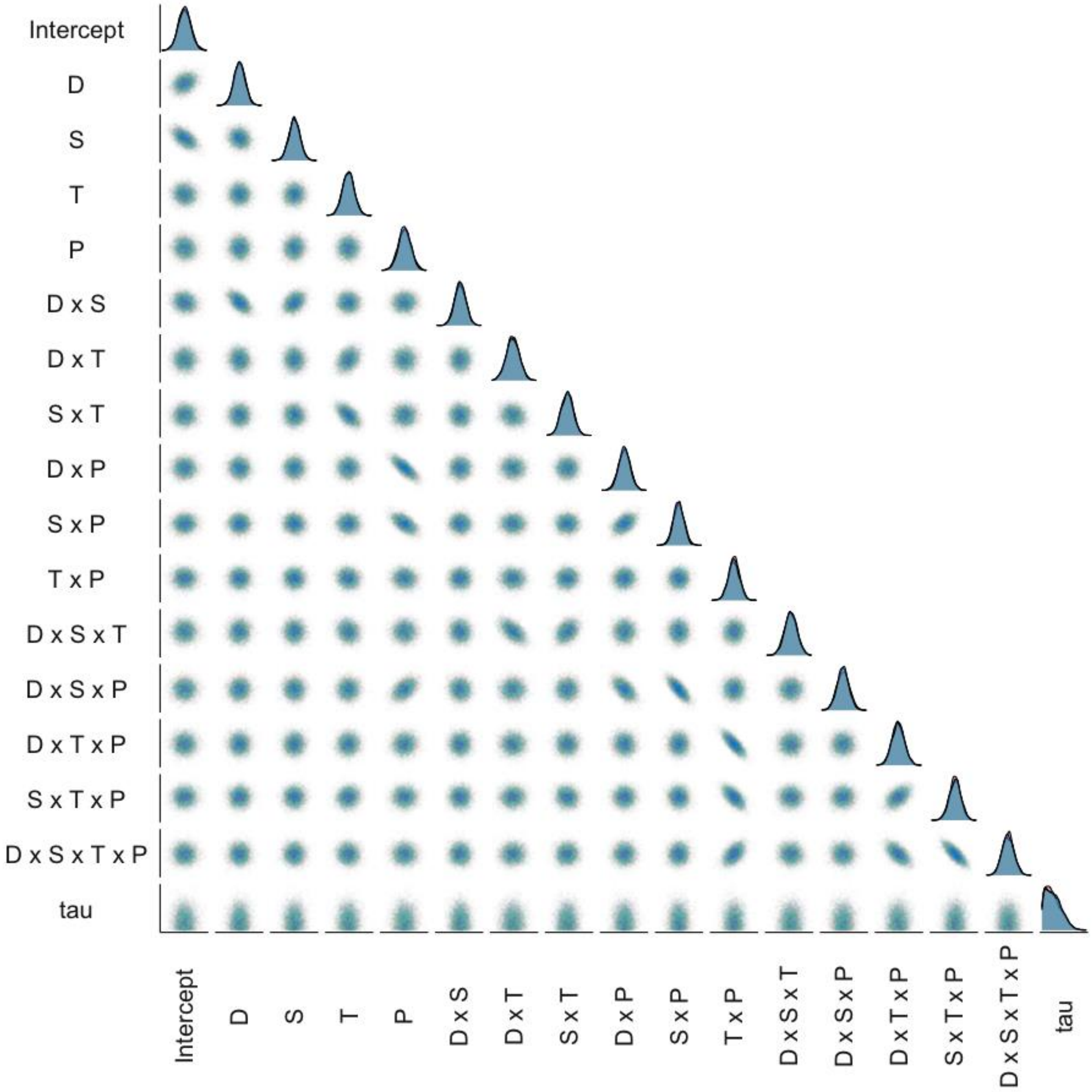
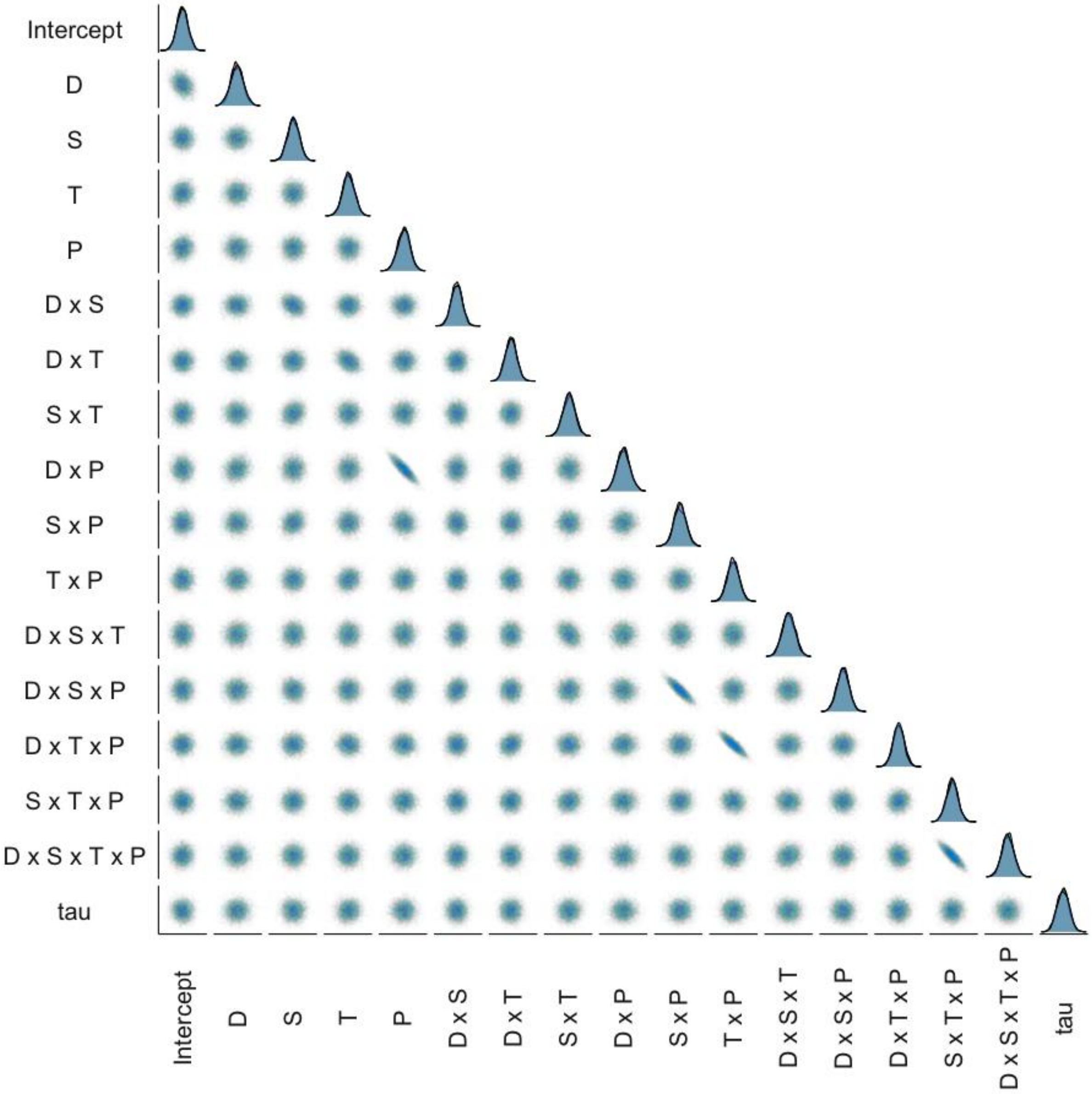


Figure 48. *2PL Item Difficulty Estimate Bias, Model Parameter Scatter Plot for Estimate Variance.*



Item Discrimination Regression Model Results

Table 34. *2PL Item Discrimination Distributional Regression Model Predictor Parameters.*

Parameter	Predictor	*M*	*SD*	Highest Posterior Density Intervals						$\hat{R}$	Bulk ESS	Tail ESS
				2.5%	12.5%	25%	75%	87.5%	97.5%			
μ	Intercept	-0.04	0.00	-0.04	-0.04	-0.04	-0.03	-0.03	-0.03	1.001	3123.64	3941.72
	τ	0.03	0.00	0.02	0.03	0.03	0.03	0.03	0.03	1.001	1914.80	3422.06
σ	Intercept	-0.79	0.20	-1.08	-1.02	-0.97	-0.76	-0.65	-0.37	1.000	3415.29	1778.30
	S	-0.31	0.01	-0.33	-0.33	-0.32	-0.31	-0.30	-0.29	1.000	6955.52	4849.59
	T	-0.10	0.01	-0.12	-0.11	-0.11	-0.09	-0.09	-0.08	1.000	6636.15	4370.84
	D	0.48	0.01	0.47	0.47	0.48	0.48	0.49	0.49	1.000	12764.97	4099.65
	P	-0.25	0.03	-0.32	-0.29	-0.27	-0.23	-0.21	-0.18	1.000	6079.79	4324.62
	S x T	0.00	0.01	-0.02	-0.01	-0.01	0.01	0.01	0.02	1.001	6743.19	4951.16
	S x D	0.00	0.01	-0.01	-0.01	0.00	0.01	0.01	0.01	1.000	7348.84	4614.89
	T x D	-0.03	0.01	-0.05	-0.04	-0.04	-0.03	-0.03	-0.02	1.000	6843.70	5090.66
	S x P	-0.02	0.03	-0.09	-0.06	-0.05	0.00	0.01	0.05	1.000	6259.65	4050.02
	T x P	-0.06	0.04	-0.13	-0.10	-0.08	-0.04	-0.02	0.01	1.000	6122.30	4208.51
	D x P	-0.03	0.01	-0.06	-0.05	-0.04	-0.02	-0.01	0.00	1.000	6137.54	4442.77
	S x T x D	0.01	0.01	-0.01	0.00	0.00	0.01	0.01	0.02	1.001	7691.86	4811.33
	S x T x P	-0.01	0.04	-0.08	-0.05	-0.03	0.02	0.03	0.06	1.000	5594.71	4346.47
	S x D x P	0.02	0.01	-0.01	0.00	0.01	0.03	0.03	0.05	1.001	5904.11	4205.33
	T x D x P	0.02	0.02	-0.01	0.00	0.01	0.03	0.04	0.05	1.000	5965.89	4046.62
	S x T x D x P	-0.01	0.02	-0.04	-0.02	-0.02	0.00	0.01	0.03	1.000	5284.48	4309.68
	τ	0.04	0.02	0.00	0.02	0.03	0.05	0.06	0.07	1.000	708.53	1287.99
ν		2.09	0.03	2.03	2.05	2.07	2.11	2.13	2.15	1.001	3418.99	1735.00

Note. D = node depth, S = sample size, T = test length, P = log-relative propagation, τ = between-simulation iteration intercept variance, ESS = Effective Sample Size. Estimates are on the log scale.

Figure 49. *2PL Item Discrimination Estimate Bias, Model Parameter Posterior Distributions for Predictors of Estimate Mean Bias.*

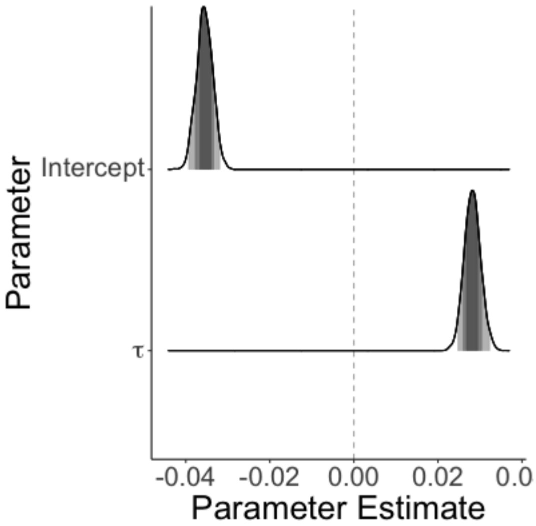


Figure 50. *2PL Item Discrimination Estimate Bias, Model Parameter Posterior Distributions for Predictors of Estimate Variability.*

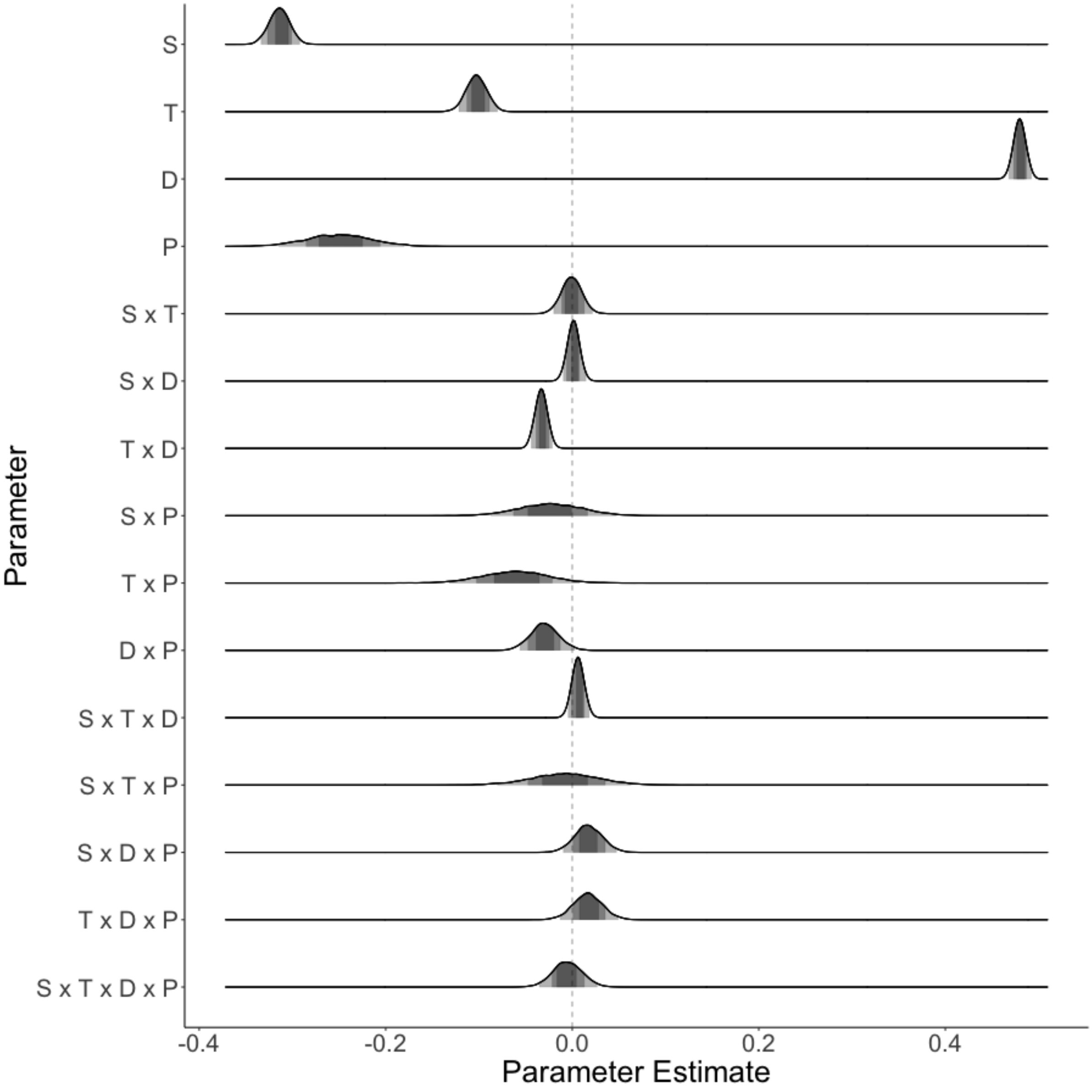


Figure 51. *2PL Item Discrimination Estimate Bias, Model Parameter Posterior Distribution of t-distribution Degrees-of-Freedom.*

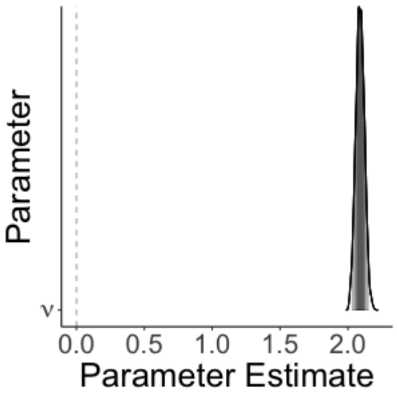


Figure 52. *2PL Item Discrimination Estimate Bias, Model Parameter Scatter Plot for Estiamte Mean Bias.*

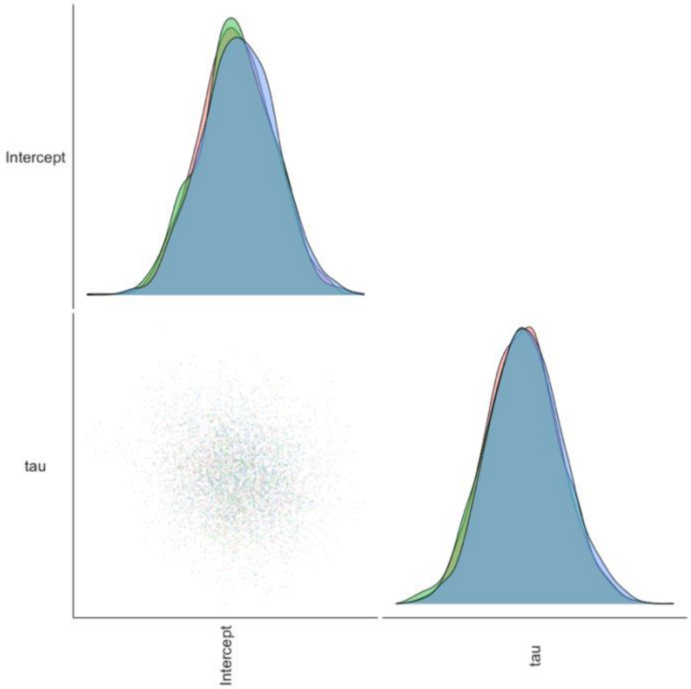
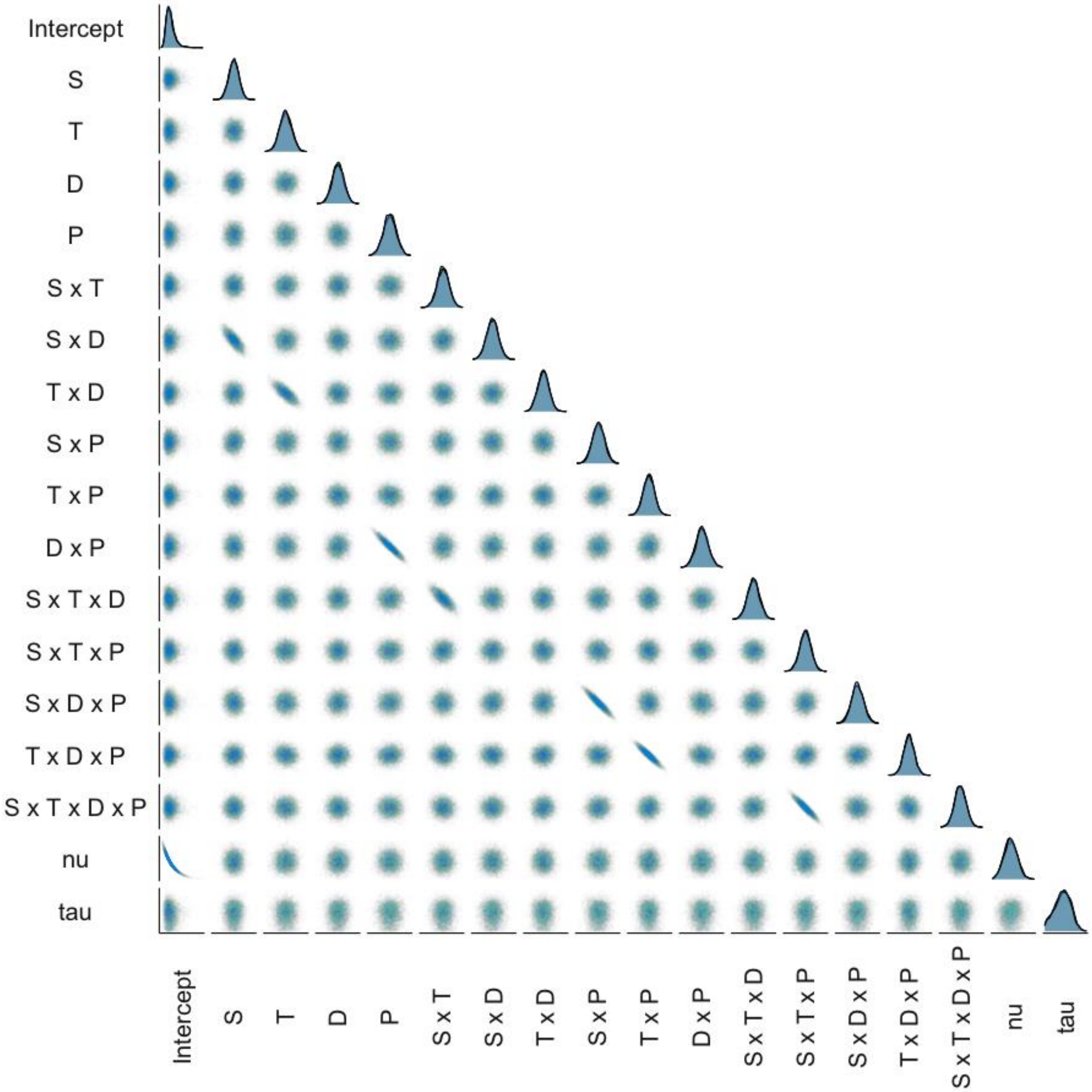


Figure 53. *2PL Item Discrimination Estimate Bias, Model Parameter Scatter Plot for Variability.*



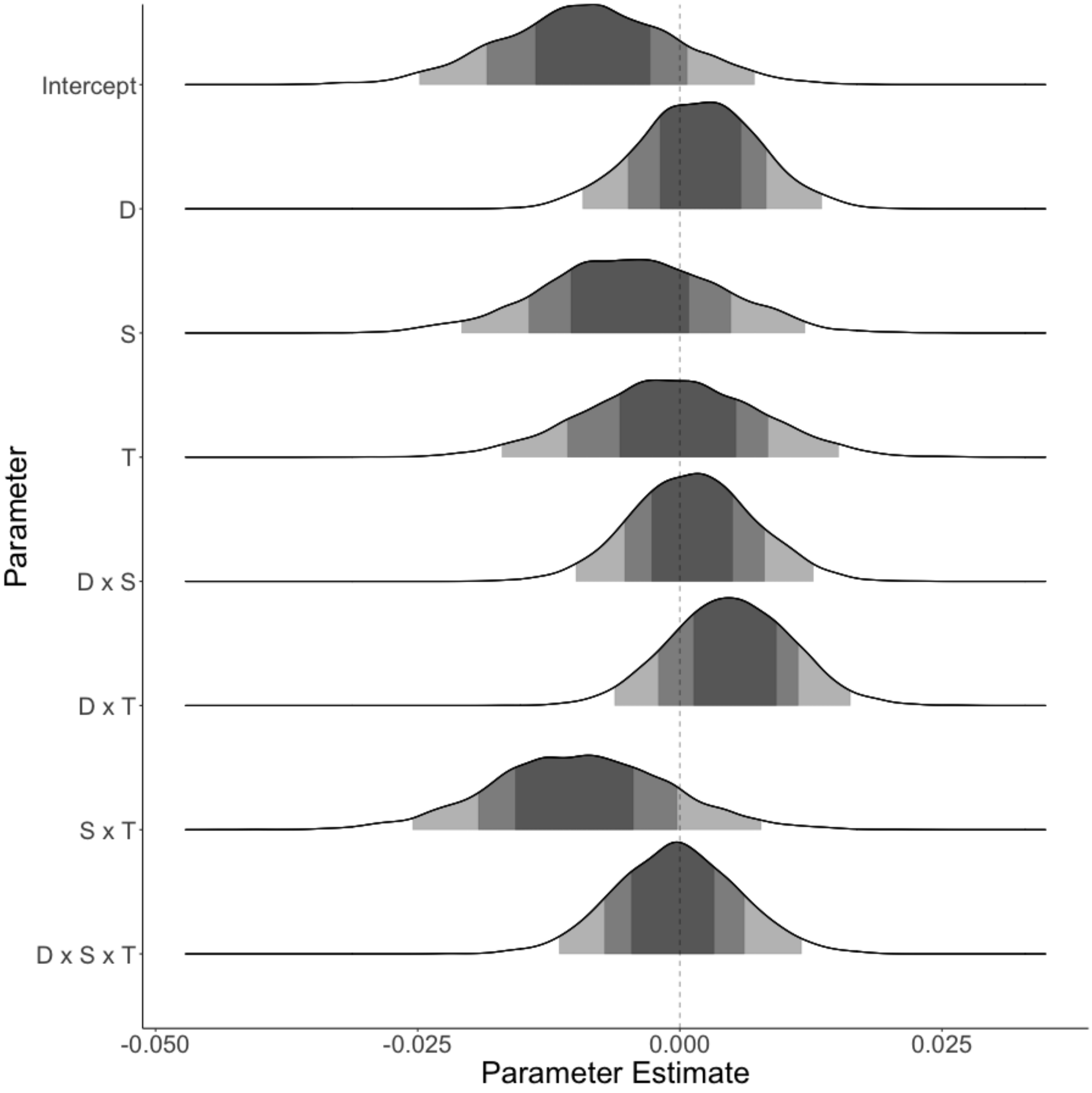
Person Ability Regression Model Results – Bias

Table 35. *2PL Person Ability Distributional Regression Model Predictor Parameters of the Estimate Mean Bias.*

Parameter	Predictor	*M*	*SD*	Highest Posterior Density Intervals						$\hat{R}$	Bulk ESS	Tail ESS
				2.5%	12.5%	25%	75%	87.5%	97.5%			
μ	Intercept	-0.01	0.01	-0.02	-0.02	-0.01	0.00	0.00	0.01	1.000	6648.01	4845.28
	D	0.00	0.01	-0.01	0.00	0.00	0.01	0.01	0.01	1.000	6488.97	5169.86
	S	0.00	0.01	-0.02	-0.01	-0.01	0.00	0.00	0.01	1.001	7073.42	4881.37
	T	0.00	0.01	-0.02	-0.01	-0.01	0.01	0.01	0.02	1.000	6668.27	4519.90
	D x S	0.00	0.01	-0.01	-0.01	0.00	0.00	0.01	0.01	1.001	7629.42	5227.36
	D x T	0.00	0.01	-0.01	0.00	0.00	0.01	0.01	0.02	1.000	7099.74	4389.68
	S x T	-0.01	0.01	-0.03	-0.02	-0.02	0.00	0.00	0.01	1.000	7439.80	4879.22
	D x S x T	0.00	0.01	-0.01	-0.01	0.00	0.00	0.01	0.01	1.001	8035.84	5080.82
σ	Intercept	-0.84	0.01	-0.86	-0.85	-0.84	-0.83	-0.82	-0.81	1.000	5370.13	3777.54
	D	0.24	0.01	0.22	0.23	0.23	0.24	0.24	0.25	1.000	6353.98	4212.60
	S	-0.01	0.01	-0.04	-0.03	-0.02	-0.01	0.00	0.01	1.000	6812.04	4724.06
	T	-0.20	0.01	-0.23	-0.22	-0.21	-0.19	-0.19	-0.18	1.000	5941.80	3927.52
	D x S	0.00	0.01	-0.01	-0.01	0.00	0.01	0.01	0.01	1.000	7243.57	4881.31
	D x T	0.04	0.01	0.03	0.03	0.04	0.05	0.05	0.05	1.000	6618.94	4436.13
	S x T	0.00	0.01	-0.03	-0.02	-0.01	0.01	0.01	0.02	1.001	5846.17	4067.86
	D x S x T	0.00	0.01	-0.02	-0.01	-0.01	0.00	0.00	0.01	1.001	6315.90	4046.52

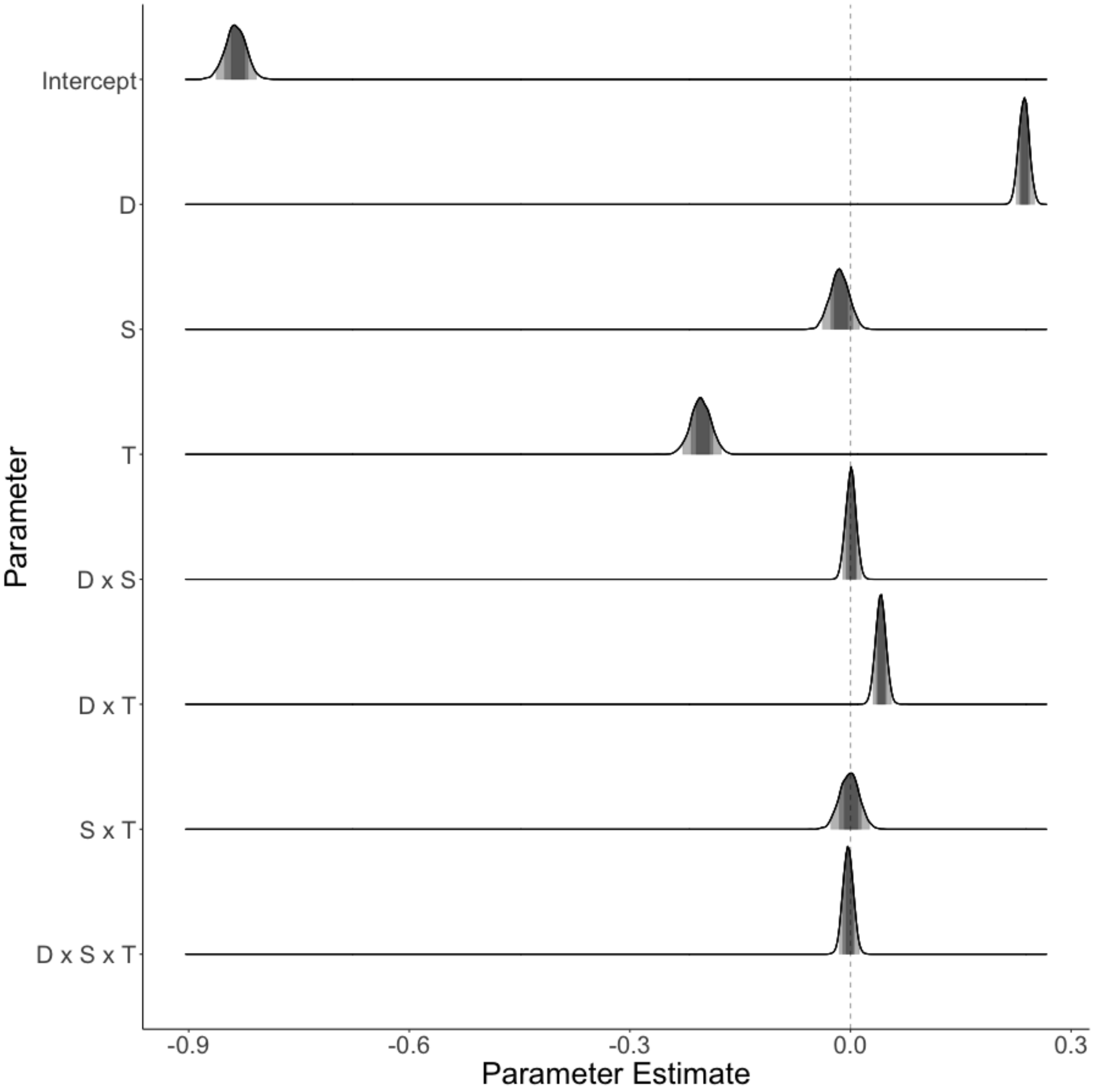
Note. D = node depth, S = sample size, T = test length, ESS = Effective Sample Size. σ predictor estimates are on the log scale.

Figure 54. *2PL Person Ability Estimate Bias, Model Parameter Posterior Distributions for Estimate Mean Bias.*



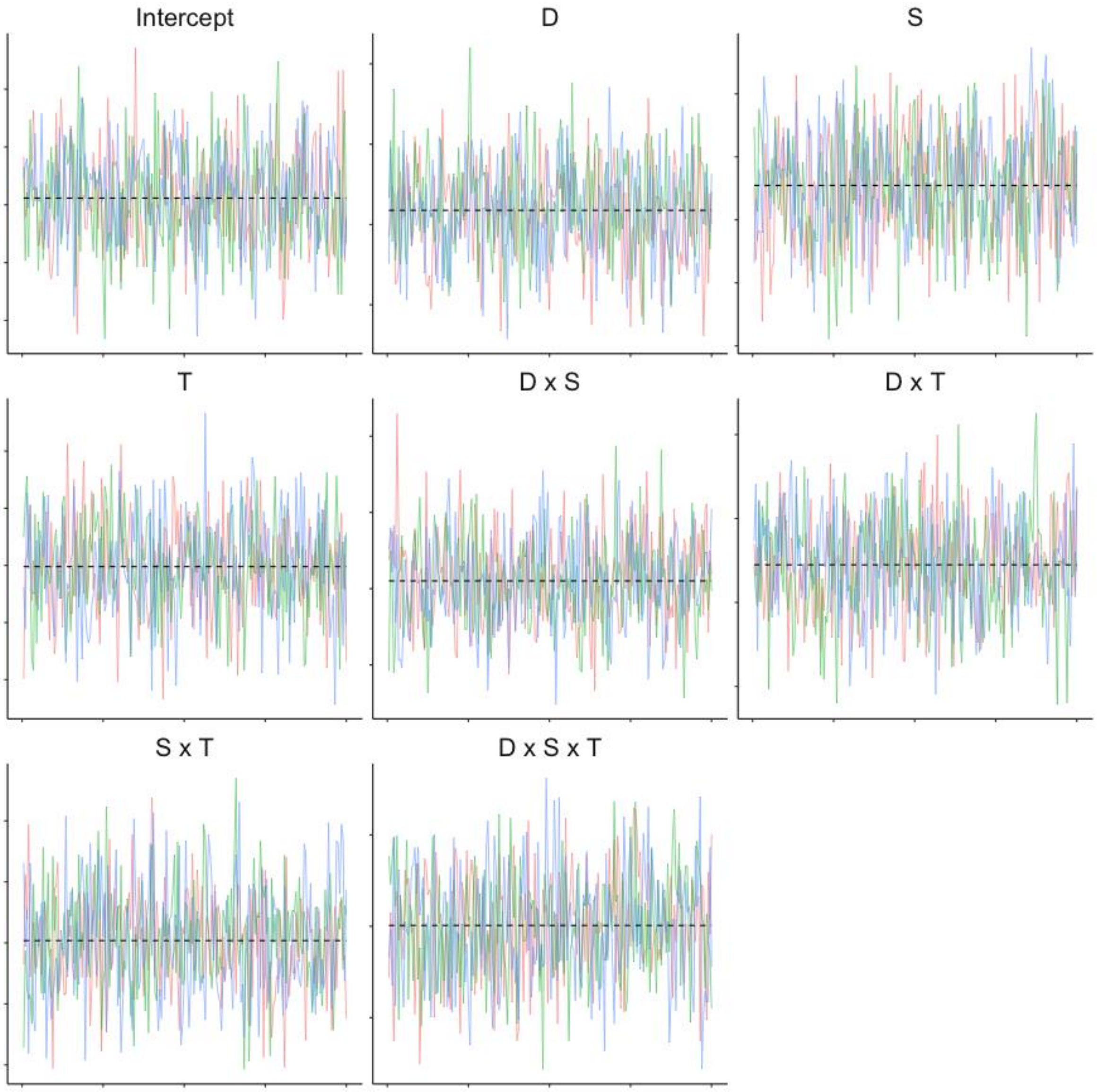
Note. 95%, 75%, and 50% highest posterior density intervals are represented with light, medium, and dark shades of grey, respectively.

Figure 55. *2PL Person Ability Estimate Bias, Model Parameter Posterior Distributions for Estimate Variability.*



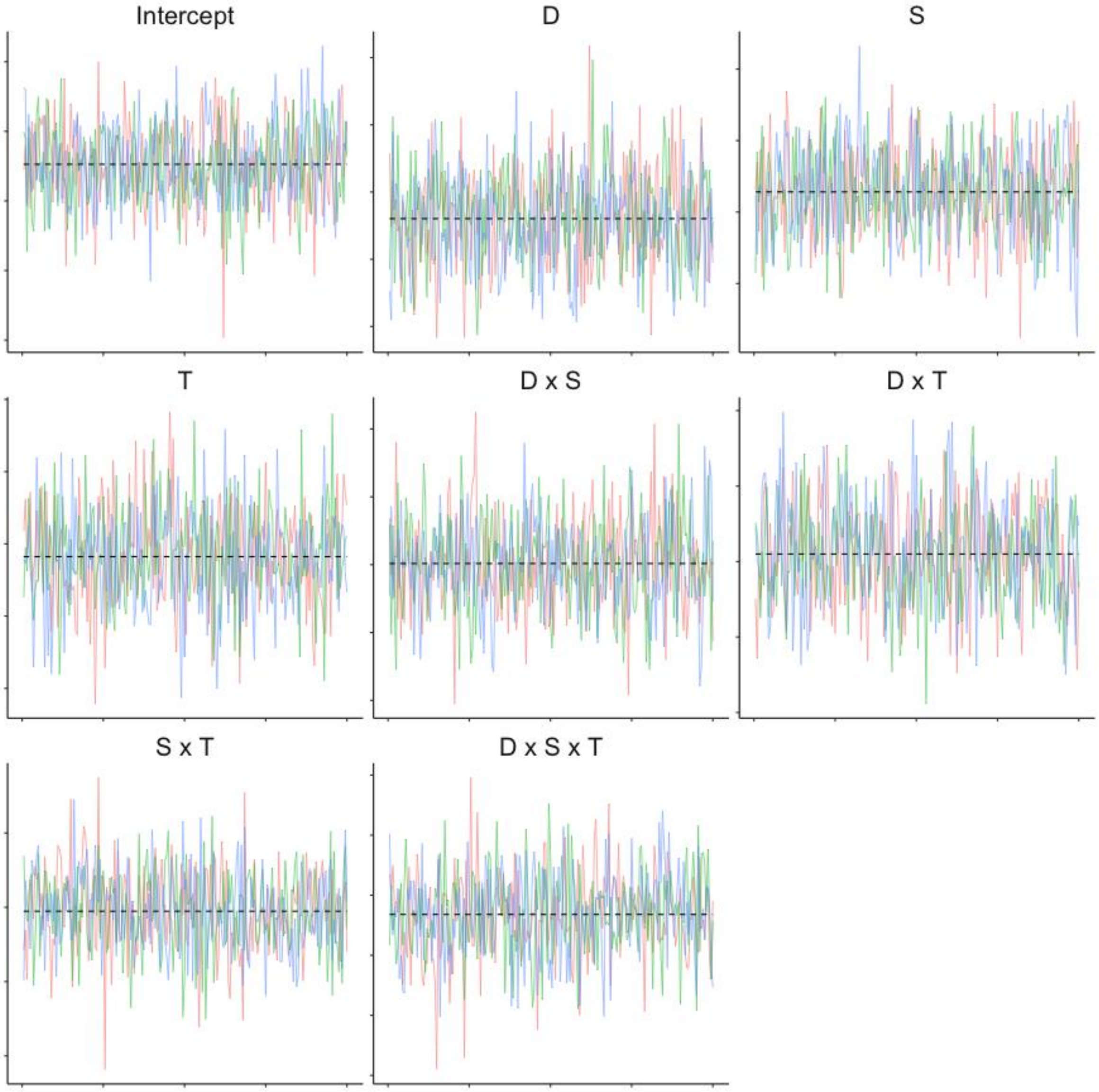
Note. 95%, 75%, and 50% highest posterior density intervals are represented with light, medium, and dark shades of grey, respectively.

Figure 56. *2PL Person Ability Estimate Bias, Model Parameter Trace Plot for Estimate Mean Bias.*



Note. Each chain was thinned by using every 10th draw to facilitate visualization.

Figure 57. *2PL Person Ability Estimate Bias, Model Parameter Trace Plot for Estimate Variability.*



Note. Each chain was thinned by using every 10th draw to facilitate visualization.

Figure 58. *2PL Person Ability Estimate Bias, Model Parameter Scatter Plot for Estimate Mean Bias.*

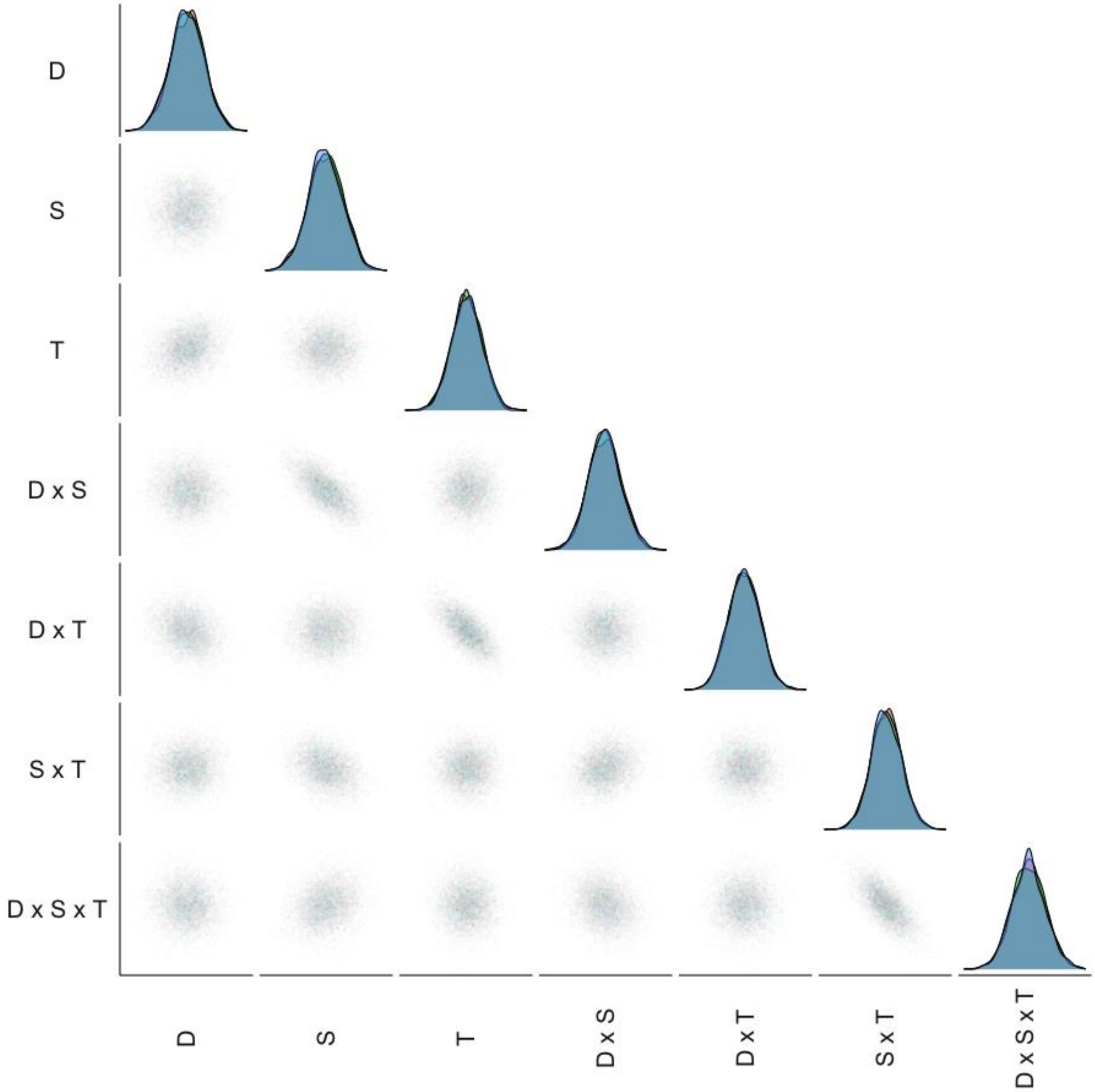
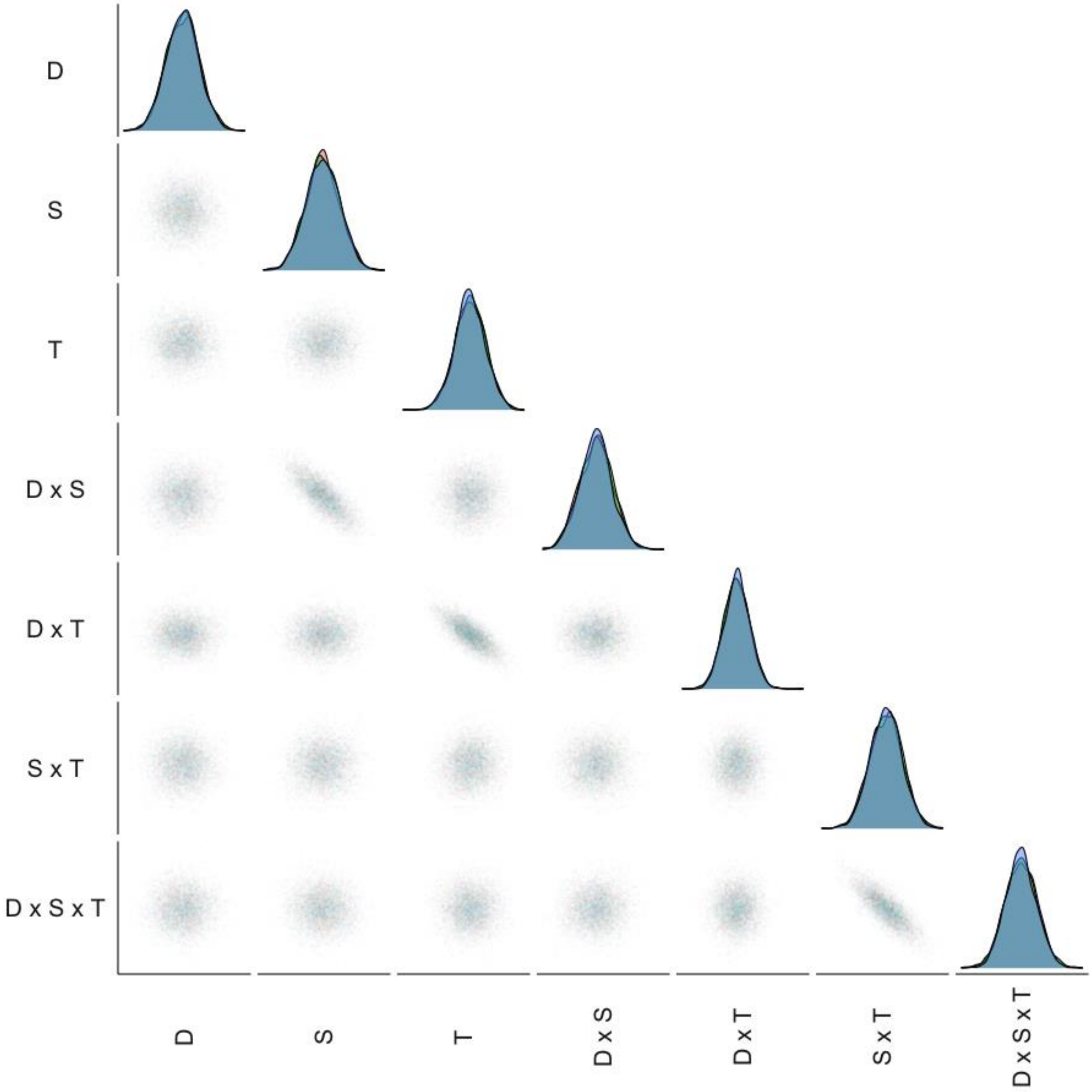


Figure 59. *2PL Person Ability Estimate Bias, Model Parameter Scatter Plot for Estimate Variability.*



Person Ability Regression Model Results – True and Estimated Parameter Correlation

Table 36. *2PL Person Ability Distributional Beta Regression Model Predictor Parameters of the True and Estimated Parameter Correlation.*

Parameter	Predictor	*M*	*SD*	Highest Posterior Density Intervals						$\hat{R}$	Bulk ESS	Tail ESS
				2.5%	12.5%	25%	75%	87.5%	97.5%			
μ	Intercept	2.24	0.01	2.22	2.23	2.23	2.25	2.25	2.26	1.001	5042.73	4402.45
	S	0.02	0.01	-0.01	0.00	0.01	0.02	0.03	0.04	1.000	4019.28	4411.39
	T	0.51	0.01	0.48	0.49	0.50	0.51	0.52	0.53	1.001	4650.67	4374.21
	D	-0.75	0.01	-0.76	-0.75	-0.75	-0.74	-0.74	-0.73	1.000	4954.44	4409.81
	S x T	-0.02	0.01	-0.04	-0.03	-0.02	-0.01	0.00	0.01	1.001	3873.86	4034.09
	S x D	0.02	0.01	0.00	0.01	0.01	0.02	0.03	0.03	1.000	4371.41	4091.02
	T x D	-0.01	0.01	-0.03	-0.02	-0.02	-0.01	0.00	0.00	1.000	4602.63	4410.79
	S x T x D	0.00	0.01	-0.02	-0.01	-0.01	0.00	0.00	0.01	1.001	4149.67	3842.89
ϕ	Intercept	5.45	0.07	5.32	5.38	5.41	5.49	5.53	5.57	1.000	4571.23	3700.65
	S	0.20	0.07	0.08	0.13	0.16	0.25	0.28	0.34	1.000	4694.04	3697.10
	T	1.23	0.06	1.10	1.16	1.18	1.26	1.31	1.35	1.001	5907.16	4222.52
	D	-0.69	0.04	-0.76	-0.73	-0.71	-0.67	-0.65	-0.62	1.000	4546.78	3754.87
	S x T	-0.07	0.06	-0.19	-0.14	-0.11	-0.03	0.01	0.06	1.001	5607.98	4527.26
	S x D	0.01	0.04	-0.06	-0.03	-0.01	0.03	0.05	0.08	1.000	4818.34	3949.72
	T x D	-0.23	0.03	-0.29	-0.27	-0.25	-0.20	-0.19	-0.16	1.000	5804.48	4409.57
	S x T x D	-0.03	0.03	-0.09	-0.07	-0.05	0.00	0.01	0.04	1.001	5690.73	4467.44

Note. D = node depth, S = sample size, T = test length, ESS = Effective Sample Size. μ predictor estimates are on the logit scale, and ϕ predictor estimates are on the log scale.

Figure 60. *2PL Person Ability True and Estimated Parameter Correlations, Model Parameter Posterior Distributions for Predictors of Correlation Location.*

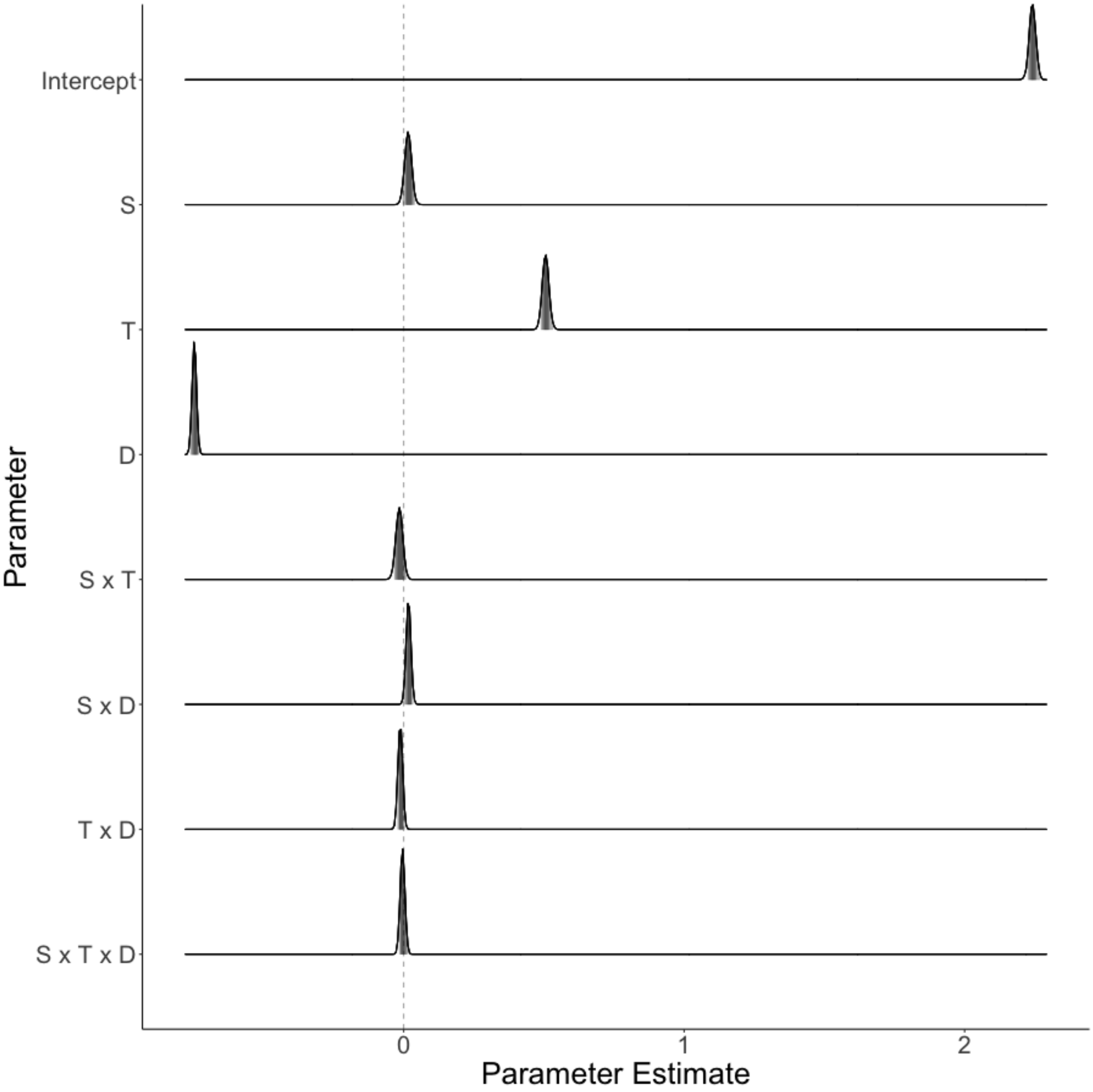


Figure 61. *2PL Person Ability True and Estimated Parameter Correlations, Model Parameter Posterior Distributions for Predictors of Correlation Scale.*

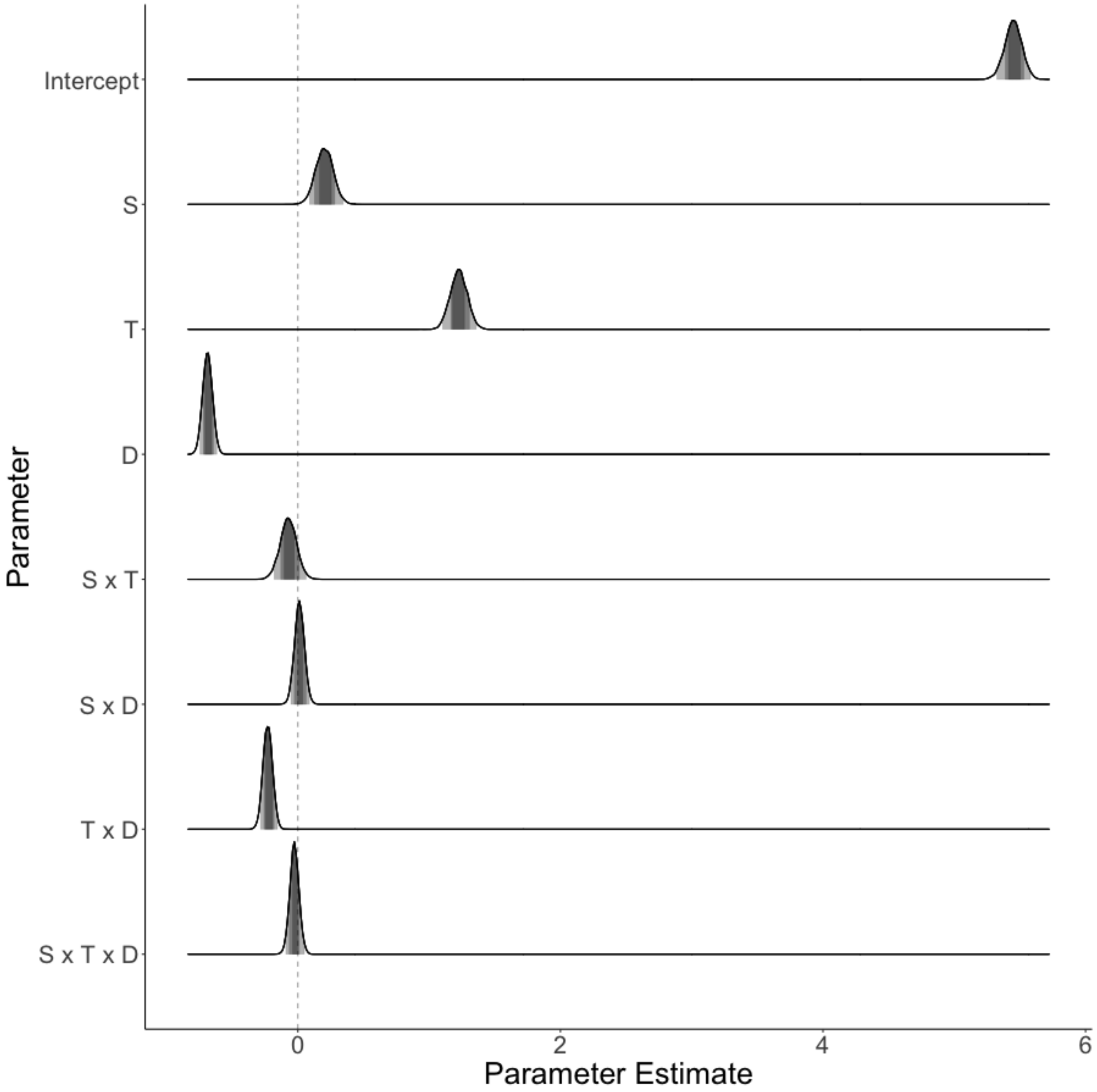
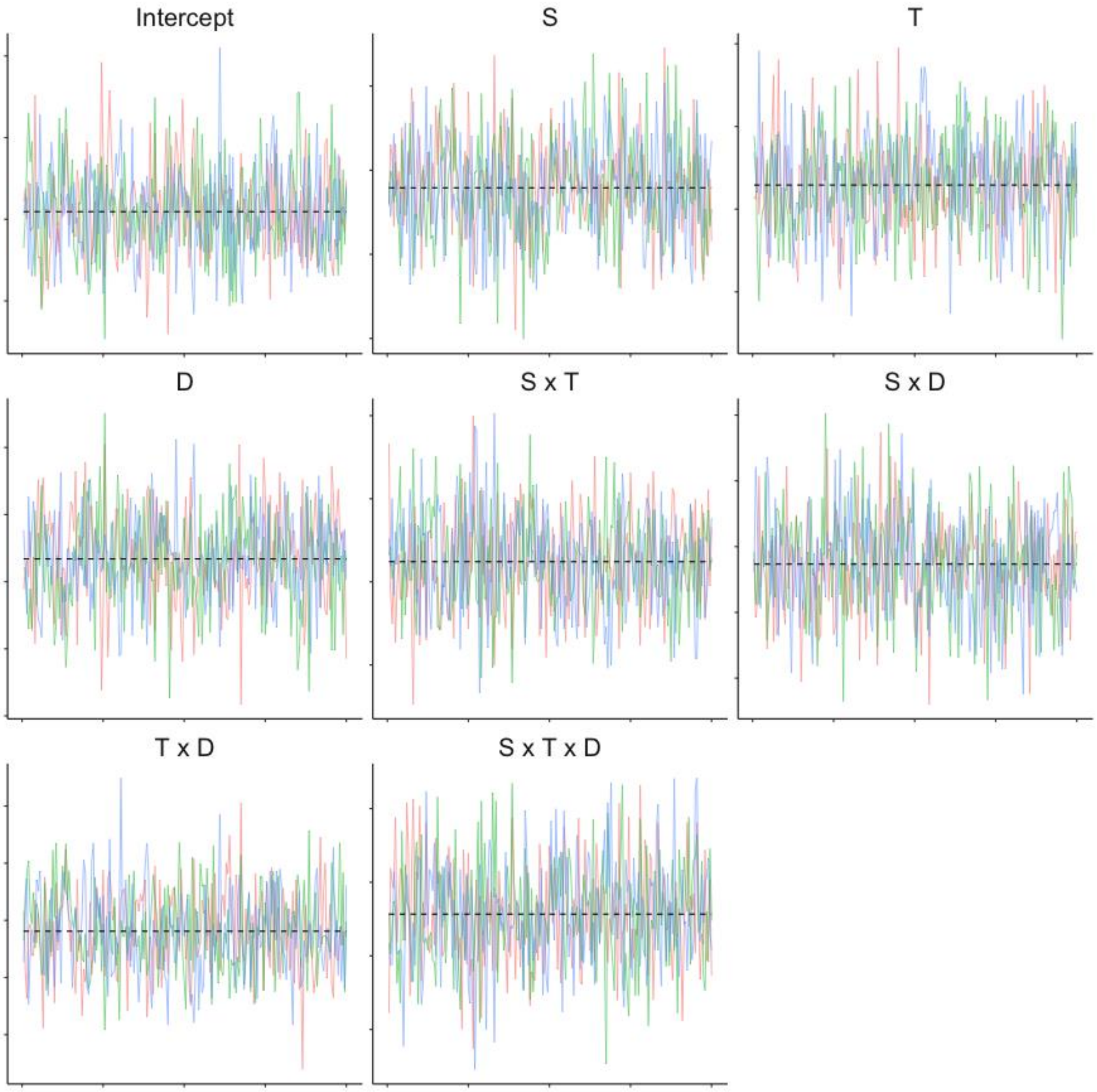
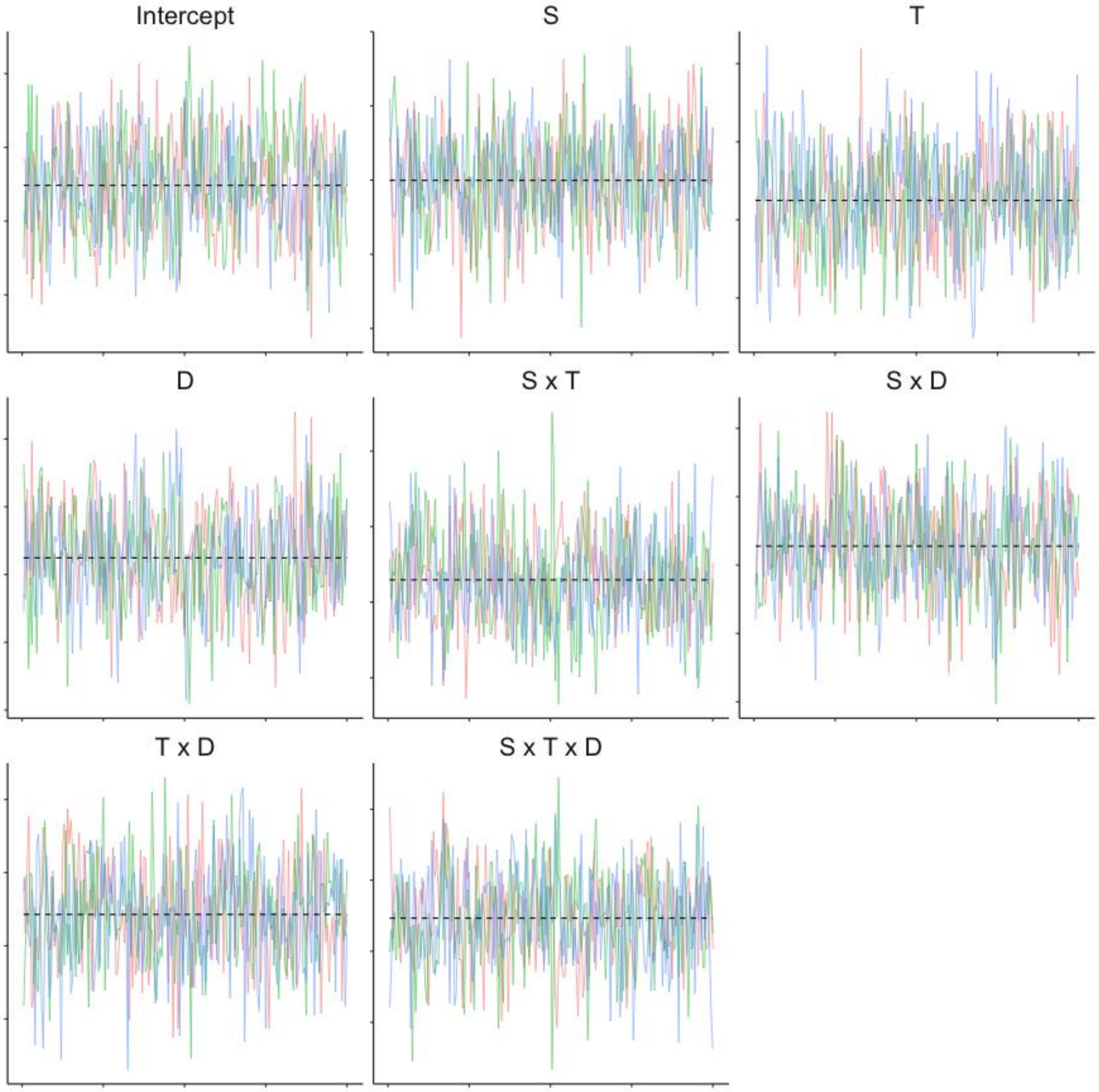


Figure 62. *2PL Person Ability True and Estimated Parameter Correlations, Model Parameter Trace Plot for Location.*



Note. Each chain was thinned by using every 10th draw to facilitate visualization.

Figure 63. *2PL Person Ability True and Estimated Parameter Correlations, Model Parameter Trace Plot for Scale.*



Note. Each chain was thinned by using every 10th draw to facilitate visualization.

Figure 64. *2PL Person Ability True and Estimated Parameter Correlations, Model Parameter Scatter Plot for Location*

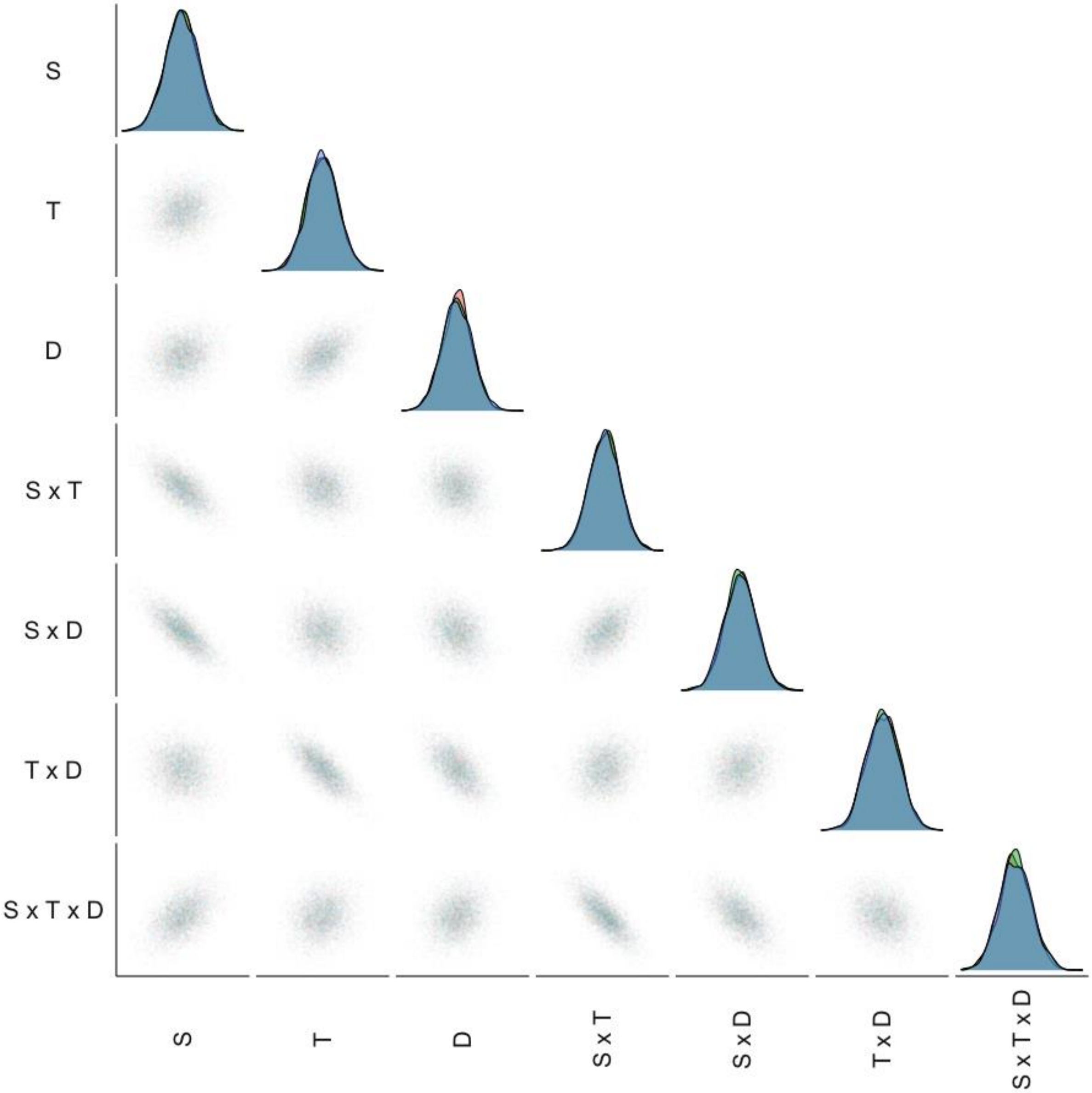
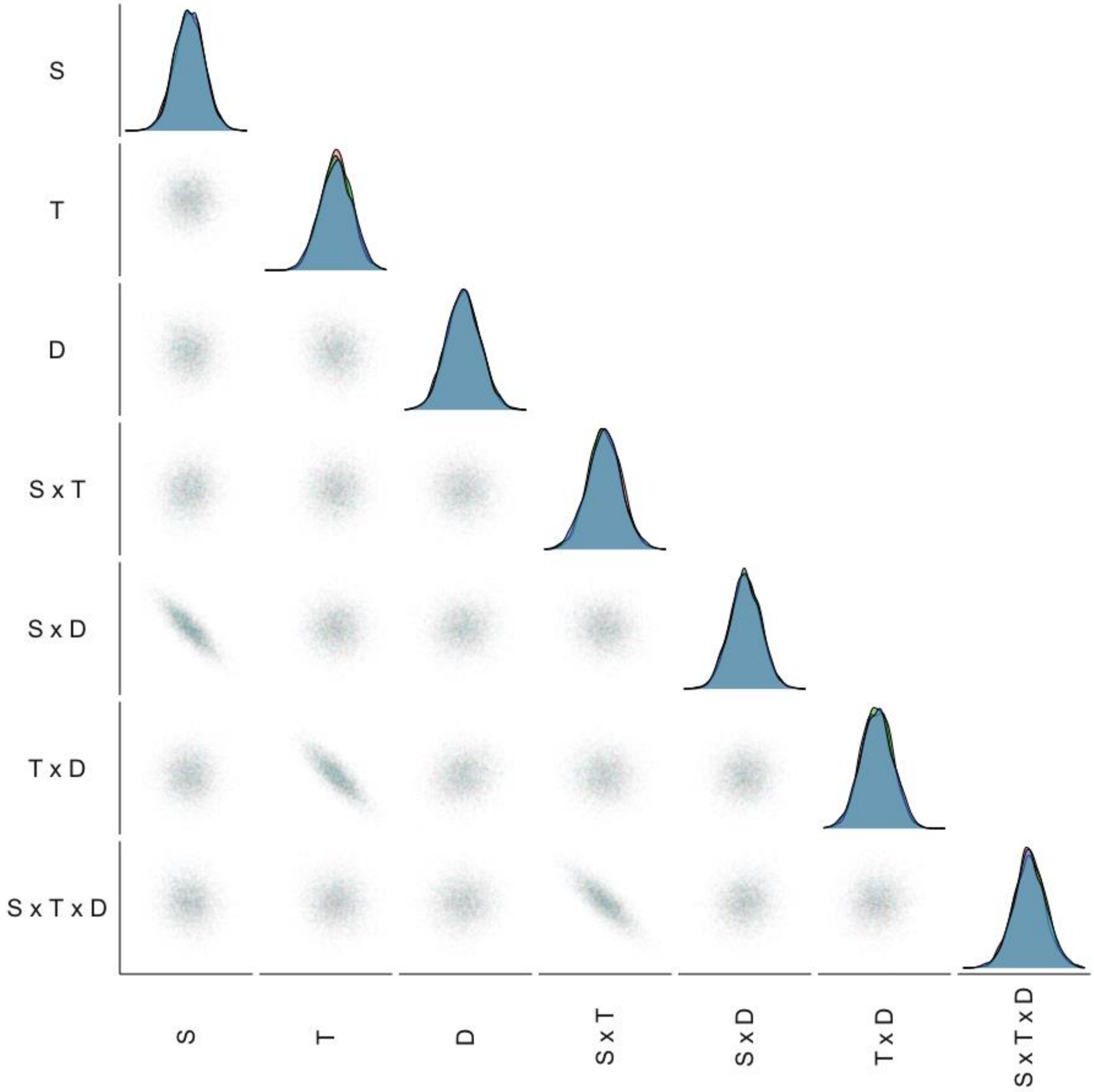


Figure 65. *2PL Person Ability True and Estimated Parameter Correlations, Model Parameter Scatter Plot for Scale*



References

Andrich, D. (2004). Controversy and the Rasch Model: A characteristic of incompatible paradigms? *Medical Care*, *42*(1), I7–I16. https://doi.org/10.1097/01.mlr.0000103528.48582.7c

Blacksmith, N., Yang, Y., Behrend, T. S., & Ruark, G. A. (2019). Assessing the validity of inferences from scores on the cognitive reflection test. *Journal of Behavioral Decision Making*, *32*(5), 599–612. https://doi.org/10.1002/bdm.2133

Bock, D. R. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*(1), 29–51. https://doi.org/10.1007/BF02291411

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

Böckenholt, U. (2012). The Cognitive-Miser Response Model: Testing for intuitive and deliberate reasoning. *Psychometrika*, *77*(2), 388–399. https://doi.org/10.1007/s11336-012-9251-y

Böckenholt, U., & Meiser, T. (2017). Response style analysis with threshold and multi-process IRT models: A review and tutorial. *British Journal of Mathematical and Statistical Psychology*, *70*(1), 159–181. https://doi.org/10.1111/bmsp.12086

Bürkner, P.-C., & Charpentier, E. (2020). Modelling monotonic effects of ordinal predictors in Bayesian regression models. *British Journal of Mathematical and Statistical Psychology*, *73*(3), 420–451. https://doi.org/10.1111/bmsp.12195

Cai, L. (2010). Metropolis-Hastings Robbins-Monro Algorithm for confirmatory item factor analysis. *Journal of Educational and Behavioral Statistics*, *35*(3), 307–335. https://doi.org/10.3102/1076998609353115

Chalmers, R. P. (2012). Mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software*, *48*(1), 1–29. https://doi.org/10.18637/jss.v048.i06

Cho, S.-J., Brown-Schmidt, S., Boeck, P. D., & Shen, J. (2020). Modeling intensive polytomous time-series eye-tracking data: A dynamic tree-based item response model. *Psychometrika*, *85*(1), 154–184. https://doi.org/10.1007/s11336-020-09694-6

De Boeck, P., & Partchev, I. (2012). IRTrees: Tree-based item response models of the GLMM family. *Journal of Statistical Software*, *48*. https://doi.org/10.18637/jss.v048.c01

Debeer, D., Janssen, R., & De Boeck, P. (2017). Modeling skipped and not-reached items using IRTrees. *Journal of Educational Measurement*, *54*(3), 333–363. https://doi.org/10.1111/jedm.12147

DiTrapani, J. B. (2019). *Assessing the absolute and relative performance of IRTrees using cross-validation and the RORME index* [The Ohio State University]. http://etd.ohiolink.edu/pg_10?0::NO:10:P10_ACCESSION_NUM:osu1555328378474406

DiTrapani, J., Jeon, M., De Boeck, P., & Partchev, I. (2016). Attempting to differentiate fast and slow intelligence: Using generalized item response trees to examine the role of speed on intelligence tests. *Intelligence*, *56*, 82–92. https://doi.org/10.1016/j.intell.2016.02.012

Drasgow, F. (1989). An evaluation of marginal maximum likelihood estimation for the two-parameter logistic model. *Applied Psychological Measurement*, *13*(1), 77–90. https://doi.org/10.1177/014662168901300108

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. L. Erlbaum Associates.

Harwell, Michael R. (1997). Analyzing the results of Monte Carlo studies in Item Response Theory. *Educational and Psychological Measurement*, *57*(2), 266–279. https://journals-sagepub-com.ezproxy.libraries.wright.edu/doi/pdf/10.1177/0013164497057002006

Harwell, Michael R., Baker, F. B., & Zwarts, M. (1988). Item parameter estimation via marginal maximum likelihood and an EM algorithm: A didactic. *Journal of Educational Statistics*, *13*(3), 243-271.

Harwell, Michael R., & Janosky, J. E. (1991). An empirical study of the effects of small datasets and varying prior variances on item parameter estimation in BILOG. *Applied Psychological Measurement*, *15*(3), 279–291. https://doi.org/10.1177/014662169101500308

Harwell, Michael R., Stone, C. A., Hsu, T.-C., & Kirisci, L. (1996). Monte Carlo studies in Item Response Theory. *Applied Psychological Measurement*, *20*(2), 101–125. https://doi.org/10.1177/014662169602000201

Huang, H.-Y. (2020). A mixture IRTree model for performance decline and nonignorable missing data. *Educational and Psychological Measurement*, *80*(6), 1168-1195. https://doi.org/10.1177/0013164420914711

Hulin, C. L., Lissak, R. I., & Drasgow, F. (1982). Recovery of two- and three-parameter logistic item characteristic curves: A Monte Carlo study. *Applied Psychological Measurement*, *6*(3), 249–260. https://doi.org/10.1177/014662168200600301

Jeon, M., & De Boeck, P. (2016). A generalized item response tree model for psychological assessments. *Behavior Research Methods*, *48*(3), 1070–1085. https://doi.org/10.3758/s13428-015-0631-y

Jeon, M., & De Boeck, P. (2019). Evaluation on types of invariance in studying extreme response bias with an IRTree approach. *British Journal of Mathematical and Statistical Psychology*, *72*(3), 517–537. https://doi.org/10.1111/bmsp.12182

Jin, K.-Y., Wu, Y.-J., & Chen, H.-F. (2019). Adopting the multi-process approach to detect differential item functioning in Likert scales. In M. Wiberg, S. Culpepper, R. Janssen, J. González, & D. Molenaar (Eds.), *Quantitative Psychology* (pp. 307–317). Springer International Publishing. https://doi.org/10.1007/978-3-030-01310-3_27

LaHuis, D. M., Blackmore, C. E., Bryant-Lees, K. B., & Delgado, K. (2019). Applying Item Response Trees to personality data in the selection context. *Organizational Research Methods*, *22*(4), 1007–1018. https://doi.org/10.1177/1094428118780310

Leventhal, B. C. (2019). Extreme response style: A simulation study comparison of three multidimensional item response models. *Applied Psychological Measurement*, *43*(4), 322–335. https://doi.org/10.1177/0146621618789392

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

McCauley, C. D., & Mendoza, J. (1985). A simulation study of item bias using a two-parameter item response model. *Applied Psychological Measurement*, *9*(4), 389–400. https://doi.org/10.1177/014662168500900408

Meiser, T., Plieninger, H., & Henninger, M. (2019). IRTree models with ordinal and multidimensional decision nodes for response styles and trait-based rating responses. *British Journal of Mathematical and Statistical Psychology*, *72*(3), 501–516. https://doi.org/10.1111/bmsp.12158

Plieninger, H. (2017). Mountain or Molehill? A simulation study on the impact of response styles. *Educational and Psychological Measurement*, *77*(1), 32–53. https://doi.org/10.1177/0013164416636655

Plieninger, H., & Meiser, T. (2014). Validity of Multiprocess IRT models for separating content and response styles. *Educational and Psychological Measurement*, *74*(5), 875–899. https://doi.org/10.1177/0013164413514998

R Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, *34*, 1–97.

Stan Development Team. (2019). *Stan modeling language users guide and reference manual* (Version 2.27) [Computer software].

Stone, C. A. (1992). Recovery of marginal maximum likelihood estimates in the two-parameter logistic response model: An evaluation of MULTILOG. *Applied Psychological Measurement*, *16*(1), 1–16. https://doi.org/10.1177/014662169201600101

Thissen, D., & Wainer, H. (1982). Some standard errors in item response theory. *Psychometrika*, *47*(4), 397–412. https://doi.org/10.1007/BF02293705

Thissen-Roe, A., & Thissen, D. (2013). A two-decision model for responses to Likert-type items. *Journal of Educational and Behavioral Statistics*, *38*(5), 522–547. https://doi.org/10.3102/1076998613481500

Tijmstra, J., Bolsinova, M., & Jeon, M. (2018). General mixture item response models with different item response structures: Exposition with an application to Likert scales. *Behavior Research Methods*, *50*(6), 2325–2344. https://doi.org/10.3758/s13428-017-0997-0

Tutz, G. (1990). Sequential item response models with an ordered response. *British Journal of Mathematical and Statistical Psychology*, *43*(1), 39–55. https://doi.org/10.1111/j.2044-8317.1990.tb00925.x

Tutz, G., & Draxler, C. (2019). *A common framework for classical and tree-based item response models including extended hierarchically structured models* (No. 227; p. 20).

Van Vaerenbergh, Y., & Thomas, T. D. (2013). Response styles in survey research: A literature review of antecedents, consequences, and remedies. *International Journal of Public Opinion Research*, *25*(2), 195–217. https://doi.org/10.1093/ijpor/eds021

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

Zettler, I., Lang, J. W. B., Hülsheger, U. R., & Hilbig, B. E. (2016). Dissociating indifferent, directional, and extreme responding in personality data: Applying the three-process model to self- and observer reports: Response processes in personality data. *Journal of Personality*, *84*(4), 461–472. https://doi.org/10.1111/jopy.12172