Assessing the Accuracy of Parameter Estimates in the Presence of Rapid Guessing Misclassifications

Joseph A. Rios
University of Minnesota

Author's Note

Joseph A. Rios: https://orcid.org/0000-0002-1004-9946

The author holds departmental affiliation at the Department of Educational Psychology, University of Minnesota, Twin Cities.

The author would like to thank Samuel Ihlenfeldt from the University of Minnesota for his helpful comments on an earlier draft.

Correspondence concerning this article should be sent to Joseph A. Rios, University of Minnesota, 56 E. River Road, 164 Education Sciences Building, Minneapolis, MN 55455.

Email: jrios@umn.edu

Cite as:

Rios, J. A. (2021). Assessing the accuracy of parameter estimates in the presence of rapid guessing misclassifications. *Educational and Psychological Measurement.* Advanced online publication. doi: 10.1177/00131644211003640

Abstract

The presence of rapid guessing (RG) presents a challenge to practitioners in obtaining accurate estimates of measurement properties and examinee ability. In response to this concern, researchers have utilized response times as a proxy of RG, and have attempted to improve parameter estimation accuracy by filtering RG responses using popular scoring approaches, such as the Effort-moderated IRT (EM-IRT) model. However, such an approach assumes that RG can be correctly identified based on an indirect proxy of examinee behavior. A failure to meet this assumption leads to the inclusion of distortive and psychometrically uninformative information in parameter estimates. To address this issue, a simulation study was conducted to examine how violations to the assumption of correct RG classification influences EM-IRT item and ability parameter estimation accuracy and compares these results to parameter estimates from the three-parameter logistic (3PL) model, which includes RG responses in scoring. Two RG misclassification factors were manipulated: type (underclassification vs. overclassification) and rate (10%, 30%, and 50%). Results indicated that the EMIRT model provided improved item parameter estimation over the 3PL model regardless of misclassification type and rate. Furthermore, under most conditions, increased rates of RG underclassification were associated with the greatest bias in ability parameter estimates from the EM-IRT model. In spite of this, the EM-IRT model with RG misclassifications demonstrated more accurate ability parameter estimation than the 3PL model when the mean ability of RG subgroups did not differ. This suggests that in certain situations it may be better for practitioners to: (a) imperfectly identify RG than to ignore the presence of such invalid responses, and (b) select liberal over conservative response time thresholds to mitigate bias from underclassified RG.

Keywords: rapid guessing, noneffortful responding, response times, parameter estimation, IRT

Assessing the Accuracy of Parameter Estimates in the Presence of Rapid Guessing Misclassifications

For well over half a century, researchers have warned of noneffortful responding (i.e., responding without putting forth full effort) serving as a validity threat to score-based inferences (e.g., Cronbach, 1960). Although there are multiple forms of noneffortful responding (e.g., skipping items, random responding), one form that has received increased attention in the literature is rapid guessing (RG; see Wise, 2017). RG occurs when an examinee provides a response in so little time that they would not be able to fully read the item stem or response options, solve its challenge, and select an answer (Wise & Kuhfeld, 2020).

Assuming that examinees are administered items in which they are capable of effortfully engaging (e.g., administering items in a language that an examinee can comprehend, examinees have had an opportunity to learn, sufficient time is provided to adequately engage in all problems), RG can occur because of two interrelated factors: (a) low task value; and/or (b) a low perceived probability of success (Penk & Schipolowski, 2015).[1] Concerning the former, examinees may engage in RG owing to a belief that their test performance has little to no personal consequences or an unawareness of the consequences for their performance. Thus, the cost of expending effort is seen to be too great when compared to the perceived personal benefits. This has been shown to be a particular issue as cognitive fatigue sets in for examinees on low-stakes tests (see Wise & Kingsbury, 2016). In terms of the second factor, for a given item, examinees may engage in RG based on the perception that the probability of success is too low to warrant the cost of full effort (see Penk & Schipolowski, 2015). As an example, prior

[1] The assumption underlying RG is that it is uninformative. However, as noted by Wise (2017), when examinees engage in RG because they do not have the full capability to engage in an item or task, RG can be informative to better understanding examinee knowledge.

research has shown that examinees who perceive a test to be very difficult tend to rapid guess at a higher rate than those who perceive the test to be easier (Rios & Guo, 2020). Across these factors, RG has been documented to occur for a number of low-stakes assessment contexts (e.g., accountability and international education studies of student knowledge) and populations that range in age and nationality (e.g., Goldhammer et al., 2016; Rios & Guo, 2020; Wise, 2017).

RG has been shown to bias measurement properties, such as reliability estimates (e.g., Wise & DeMars, 2009), measurement invariance (e.g., DeMars & Wise, 2010), linking coefficients (e.g., Mittelhaëuser et al., 2015), and item and person parameter estimates (e.g., Rios & Soland, 2020; van Barneveld, 2007). Furthermore, as RG is generally associated with underestimation of examinee ability (Silm et al., 2020), it has been documented to bias treatment effects (e.g., Osborne & Blanchard, 2011; Liu et al., 2015), achievement gains (e.g., Wise & DeMars, 2010), value-added estimates of teacher effectiveness (Wise et al., 2013), and subgroup comparisons (e.g., Debeer et al., 2014). These results have prompted the measurement community to call for test users to document test engagement in contexts where it may be a concern (see above; American Education Research Association et al., 2014).

However, to mitigate the deleterious role of RG first presumes that it can be correctly identified.[2] Such an assumption may be untenable as one can only use proxies of test-taking behavior to make inferences concerning RG (more detail is provided below). This has two implications. Firstly, knowing whether RG has occurred can never be fully realized, and secondly, because of this, it is plausible that any attempt to identify RG may lead to inaccurate

[2] Recent mixture models have been proposed that estimate the probability of aberrant responding, and thus, identification of such behavior is not dichotomous (e.g., Wang et al., 2018). However, given the complexity of estimation procedures as well as the additional parameters required, these models have seen limited application in practice. Thus, the focus of this paper is on the dichotomous identification of RG using a response time threshold procedure, which is currently employed in operational settings (e.g., Wise & Kuhfeld, 2020).

classifications of this behavior. Focusing on the latter issue, the goal of this paper is to understand the impact of RG misclassifications on item and person parameter estimation accuracy. In the sections that follow, procedures for RG identification and modeling are reviewed, and a rationale for the current study is provided.

Response Times as a Proxy for Identifying RG

To date, the most popular proxy for identifying RG is to use response times. The use of response times provides a number of advantages over competing proxies such as self-report measures of effort and person-fit statistics (Silm et al., 2020; Wise, 2017). First, it is an unobtrusive approach, as examinees are unaware that their test-taking behavior is evaluated due to the use of log file information.[3] As such, concerns about observer effects, such as the Hawthorne effect (i.e., subject behavior changing due to their knowledge that they are being observed) and observer bias, are mitigated. Second, this approach can evaluate RG on an item-by-item basis, which is advantageous, as prior research has demonstrated that examinee behavior can change throughout a test administration (e.g., Wise & Kingsbury, 2016). This provides several scoring advantages as valid and invalid responses within each examinee can be distinguished; thus, allowing for the capacity to estimate ability for examinees that have engaged in a certain level of RG, as opposed to listwise deletion of unmotivated/aberrant examinee data (a common approach for self-report and person-fit statistic approaches). The latter approaches have

[3] One reviewer raised the concern of the ethical nature of using log file information to inform scoring when the examinee is unaware that their behavior is monitored. However, it is unclear whether operational testing programs notify examinees of this fact, potentially due to two factors. First, as noted, RG is generally a concern in testing contexts that are low-stakes (i.e., there is minimal to no personal consequences for examinees' test performance), and as a result, utilizing response time information to improve the validity of score-based inferences has negligible consequences for individual examinees. Second, by making examinees aware that their time of responding is monitored, examinees may mask their noneffortful responding via slower responses, and thus, may mitigate the utility of identifying RG from response time information. Regardless, it is argued that examinees should be made aware of all components that will inform their scoring, particularly in contexts in which there are high-stakes consequences for individual examinee performance.

been found to be associated with a loss of as much as 25% of the sample (see Rios et al., 2014). Building off these advantages, response times are used to establish a threshold in which any response provided in less time than the criterion is assumed to be RG. To this end, numerous response time threshold procedures have been developed (for a discussion, see Wise, 2017).

These procedures can be categorized into three distinct typologies, corresponding to methods that utilize: (a) no empirical data; (b) only response time information; and (c) a combination of response time and accuracy information. In the former class of procedures, a threshold can be arbitrarily set equal across all items (e.g., three seconds) or established based on taking the number of characters contained within a given item, and coupling this information with estimated reading speeds for a given test-taking population (see Wise & Kong, 2005). In the second category of procedures, researchers have utilized observed response time distributions, to establish criteria based on normative information (e.g., stipulating that a response provided below 10% of the mean item response time is RG; Wise & Ma, 2012) or to indicate the point at which examinees transition from RG to solution behavior based on the shape of a RT distribution (for more details, see Schnipke & Scrams, 1997; Rios & Guo, 2020). In addition, researchers have proposed employing both response accuracy and time information based on the assumption that RG responses possess accuracy rates that are approximately equal to chance (typically defined as the reciprocal of the number of response options), which has been supported by prior research (e.g., Wise & Kong, 2005). Thus, for a procedure, such as the Cumulative Proportion Correct (CUMP) method proposed by Guo et al. (2016), a threshold is established at the time point in which the correct response rate begins to be consistently greater than chance. As noted by Wise (2017), each procedure provides advantages and disadvantages, and to date, there is no clear consensus on the best threshold procedure to employ in practice.

Utilizing response times as a proxy for identifying RG is not without its limitations. To begin with, the use of response times requires the collection of log file information, which means that it cannot be applied to data collected from paper-and-pencil test administrations. Similar to the other approaches, it is limited in that response times are used as a proxy of test-taking behavior, and thus, requires two assumptions to be made. The first is that a quick response is invalid. However, it has been argued that response times are associated with individual ability differences (see Goldhammer, 2015). This assertion is supported by research showing that higher ability examinees are more likely to correctly respond to an item at a quicker rate than their lower ability counterparts when taking into consideration item difficulty (De Boeck & Jeon, 2019; Loken & Beverly, 2020). Thus, concerns have been raised by some researchers that response time thresholds may be too liberal and incorrectly classify a valid quick response as RG (see Wise, 2017).

The second assumption is that RG, by definition, is associated with quick responding. Such an assumption ignores the possibility of RG behavior that occurs slowly. For example, an examinee could disengage from an item for a prolonged amount of time due to noneffortful behavior (e.g., daydreaming), and then return to the item by RG. In such an instance, the log file information would suggest that the examinee was engaged on the item for a long period, and consequently, their RG response would go undetected. Due to this, one of the potential limitations of this approach is that thresholds may be too conservative and fail to capture slow RG behavior (see Wise & Kuhfeld, 2020).[4]

[4] An alternative would be to establish two thresholds, which each capture fast and slow forms of noneffortful responding. Using the CUMP approach, the latter threshold type could be determined by examining cumulative proportion correct rates by time in reverse (max number of seconds to 0 seconds).

Taken together, these assumptions get at two sides of the same issue, which is that the use of response times, like all other procedures, leads to accurate identification of RG. However, as noted, complete accuracy may be untenable as it is impossible to be certain about examinees' behavior from process data. Regardless of these limitations, it is argued that the advantages of using response times as a proxy of RG provides the best available solution, due to it being unobtrusive and possessing the capability to identify RG for individual responses. Due to these advantages, the use of response times as a proxy of RG has seen increased usage in the literature and practice (e.g., Silm et al., 2020). Below, scoring models that incorporate response times to specifically mitigate the deleterious impact of RG are reviewed.

Scoring Approaches Accounting for RG that Use Response Times

Various Item Response Theory (IRT) models have been developed that incorporate response times to distinguish between non-RG and RG behavior. These models can be categorized into two classes, those that incorporate: (a) response times; and (b) both item responses and response times.

Scoring Approach Incorporating Response Times

Concerning the former, Wise and DeMars (2006) first proposed rescoring RG via the Effort Moderated IRT (EM-IRT) model:

$$P_{ij}(\theta) = \left(SB_{ij}\right)\left(\left(c_i + (1 - c_i)\left(\frac{e^{-1.7a_i\left(\theta_j - b_i\right)}}{1+e^{-1.7a_i\left(\theta_j - b_i\right)}}\right)\right)\right) + \left(1 - SB_{ij}\right)\left(\frac{1}{d_i}\right). \quad (1)$$

The EM-IRT model can be conceptualized as possessing two sub-models, which decompose the probability of correctly responding to an item separately for solution and RG behavior. These behaviors are distinguished at the item-by-examinee level using a binary indicator, SB_{ij}, determined based on the adoption of a response time threshold procedure. The use of this indicator allows an examinee to switch between solution and RG behavior throughout the test, which avoids making assumptions about how RG occurs across examinees (see Wise & Kingsbury, 2016). For solution behavior ($SB_{ij} = 1$), $P_{ij}(\theta) = c_i + (1 - c_i)\left(\frac{e^{-1.7a_i(\theta_j - b_i)}}{1+e^{-1.7a_i(\theta_j - b_i)}}\right)$, where a_i is the discrimination parameter for item i, b_i is the difficulty parameter for item i, c_i is the lower asymptote of item i, and θ_j is the ability parameter for person j (this is equivalent to the standard 3PL model). [5] However, if $SB_{ij} = 0$, $P_{ij}(\theta) = \frac{1}{d_i}$, where d_i is roughly equal to the number of response options for item i.[6] When $SB_{ij} = 0$, noneffortful responses add a constant across all levels of the theta continuum to the log-likelihood function, and thus, do not influence the maximum of the function. This implies that when an examinee employs RG, their probability of correctly responding to item *i* does not depend on their underlying ability. Due to the uninformative nature of such responses, scoring under the EM-IRT model downweights RG responses to have no influence on parameter estimation (Wise & DeMars, 2006).

There are two assumptions underlying this model: (a) responses classified as effortful are representative of the range of item characteristics and content on the test; and (b) the true abilities of rapid guessers are reflective of the sample distribution when estimating aggregate-level ability. Concerning the former assumption, prior research has suggested that it may be untenable in practice given that RG has been found to be associated with item position, length, difficulty, and depth of knowledge required (e.g., Wise, 2020). However, recent evidence indicates that aggregate-level ability estimates may be largely robust to such a violation (Rios & Soland, 2020). In terms of the latter assumption, aggregate-level ability estimates will be biased

[5] The EM-IRT model can be extended to other IRT models, such as the Rasch, one-parameter, and two parameter logistic models.

[6] Correct rates of noneffortful responses may be beyond the chance level due to a function of the location of the correct answers on the assessments (see Pastor et al., 2019).

either positively or negatively if noneffortful responders underlying ability is consistently higher or lower than the average ability of effortful responders. Evidence from operational tests have illustrated that noneffortful responding can occur among examinees with predominately low prior ability (e.g., Kuhfeld & Soland, 2020; Rios et al., 2017), and under such circumstances, biased ability parameter estimates can be obtained via the EM-IRT model (Rios & Soland, 2020).

Although meeting the assumptions underlying the EM-IRT model is a prerequisite to obtaining accurate item and ability parameter estimates, numerous simulation and applied studies have shown that the use of this model is associated with improved parameter estimation and convergent validity with external variables compared to naïve models that include RG responses in scoring (e.g., Liu et al., 2019; Rios et al., 2017; Rios & Soland, 2020; Wise & DeMars, 2006; Wise & Kingsbury, 2016). Furthermore, apart from establishing a response time threshold to classify RG, this model does not require additional parameter estimation. Due to this simplicity, it has become one of the most popular models used to study RG (e.g., Rios et al., 2017; Rios & Soland, 2020; Wise & Kingsbury, 2016), and is currently one of the only models to be employed in operational testing contexts to report examinees' RG behavior to score users (see Wise & Kuhfeld, 2020).

Scoring Approaches Utilizing both Item Responses and Response Times

A second class of models has been proposed based on the theory that classifying examinee RG behavior can best be accomplished by incorporating both item response and response time data. To this end, Meyer (2010) proposed a mixture model that assigns examinees to one of two latent groups based on speededness (speeded or non-speeded). Item parameters and population mean/variances of person parameters are estimated separately for each latent class.

The major limitation of this model is that it divides examinees into global-level RG classes, which limits the ability to identify RG at the item-by-examinee level and does not account for differences in RG rates for examinees within the same latent class.

To address this limitation, Wang and Xu (2015) developed a mixture hierarchical model that distinguishes between solution and RG behavior at the item-by-examinee level. Similar to the EM-IRT model, parameter estimates are calibrated based solely on the purified sample (i.e., examinees engaging in solution behavior) by employing a standard IRT model (e.g., 2PL model), while the probability of correctly answering an item for RG behavior is constrained equal to the item-specific guessing probability. Although conceptually similar to the EM-IRT model, this mixture model requires additional parameters to be estimated (e.g., a time intensity parameter for each item and examinee speed parameter), several constraints to be imposed (e.g., specification of either the mean of the speed parameter or the mean of the time intensity parameter), and a specialized Monte Carlo estimation procedure that is not readily available in most commercially available software. Given these disadvantages, Wang and Xu's (2015) model has seen limited application in the research literature and practice. Thus, the remainder of this paper focuses on the modeling of RG using the EM-IRT model.

Study Rationale

Parameter estimation accuracy is dependent on the correct identification of RG when employing scoring models that utilize RG information at the item-by-examinee level, such as the EM-IRT model. A failure to meet this assumption leads to the inclusion of distortive and psychometrically uninformative information in parameter estimates (Wise, 2017). However, it is argued that the tenability of this assumption can never be known as current procedures are only indirect proxies of RG, and thus, they may lead to false negative (i.e., failing to identify a RG response) and/or false positive (i.e., classifying a valid response as RG) identifications (Wise, 2017). Therefore, it is unclear how robust parameter estimates are to differing RG misclassification types and percentages. To address this gap in the literature, the objective of this simulation study is two-fold. First, to examine the degree of bias in both item and person parameters when employing the EM-IRT model under incorrect identification of RG, the degree and type of RG misclassification is manipulated. Second, the accuracy of item and ability parameter estimates from the EM-IRT model with misclassified RG are compared to those obtained from the naïve three-parameter logistic (3PL) model (including RG responses in scoring). This is done to determine whether the former model provides improved parameter estimation when violating the assumption that RG is correctly identified. Findings from this study have the potential to inform practitioners about the robustness of item and ability parameter estimates from the EM-IRT model in the presence of RG misclassifications.

Method

Data Generation of Effortful Response Probabilities

Data were generated for a unidimensional test consisting of 30 items administered to 5,000 simulees in *R*, version 4.0.0 (R Core Team, 2020). The number of simulees was chosen based on prior researchers demonstrating that a sample size of 5,000 is expected to provide stable parameter estimates for the three-parameter logistic (3PL) model with 30 items (e.g., Hulin et al., 1982), which was the model used to create effortful item response probabilities in this simulation:

$$P_{ij}(\theta) = c_i + (1 - c_i)\left(\frac{e^{-1.7a_i(\theta_j - b_i)}}{1+e^{-1.7a_i(\theta_j - b_i)}}\right). \tag{2}$$

This was done by first sampling item and person generating parameters. The former were taken from an operational administration of the NAEP math assessment (for a full list of item parameters, see Appendix A of the supplementary file). Across all 30 items, the mean discrimination, difficulty, and guessing parameters were 1.15 (*SD* = 0.49), -0.08 (*SD* = 1.29), and 0.21 (*SD* = 0.07), respectively. Generating ability parameters were sampled from a normal distribution (more detail is provided in the next section). Both the item and ability generating parameters were then entered into the 3PL model to obtain effortful item response probabilities.

Substitution of RG Response Probabilities

To simulate RG for simulees, the next step consisted of replacing effortful probabilities with the chance probability (.25; assuming each item possessed four response options). This was done by manipulating three independent variables: (a) group impact between RG and non-RG simulees (no impact, moderate impact, large impact); (b) the percentage of RG in the data matrix (10%, 20%); and (c) RG pattern (difficulty-based, progressive).

Group Impact between RG and non-RG Simulees

RG was simulated for conditions in which group impact was and was not present. To accomplish this, ability parameters were sampled separately by RG group (simulees engaging in RG and those that did not). Specifically, for non-RG simulees, ability parameters across all conditions were randomly sampled from a standard normal distribution. In contrast, RG simulees were sampled differently based on the presence of group impact. For instance, to mimic no impact, RG simulees' ability parameters were randomly sampled from a standard normal distribution to parallel the ability distribution of the non-RG simulees. In conditions in which group impact was present, RG simulees' ability parameters were randomly sampled from a normal distribution with a variance of 1 and a mean that was either -0.50 or -1. The former mean was chosen based on findings by Rios et al. (2017), who found an average performance difference on a prior ability measure between RG and non-RG examinees equal to approximately

0.50 standard deviations (*SD*s; favoring non-RG examinees). The latter mean served as an extreme ability difference between non-RG and RG simulees.

Percentage of RG in Data Matrix

Two percentages of RG were examined: 10% and 20%. Although previous researchers have constrained the number of rapid guesses (RGs) equal across RG simulees (e.g., Rios et al., 2014), it is argued that such an approach is not reflective of operational settings, as prior research has shown that examinees can engage in a wide range of RGs across a test administration (e.g., Wise & Kingsbury, 2016). To better reflect this reality, the number of RGs for simulees in the RG subgroup was allowed to vary from 1 to 30. Across RG simulees, the number of RGs added up to either 10% or 20% of item responses in the data matrix (refer to Appendix B in online supplementary information for a distribution of RGs by condition). These two percentages are within the range of RG observed in operational settings (e.g., Goldhammer et al., 2016; Rios & Guo, 2020). Across conditions, the percentage of RG simulees was constrained to 30%. This percentage is within the range of examinees engaging in RG observed in operational settings (DeMars, 2007; DeMars & Wise, 2010) and of those examined in prior simulation studies (Rios et al., 2017; Wise & DeMars, 2006).

RG Pattern

Two RG patterns were manipulated: (a) difficulty-based; and (b) progressive RG. The former reflects the situation in which examinees who perceive an item to be too difficult engage in RG based on the perception that the probability of success is too low to warrant the cost of full effort (see Penk & Schipolowski, 2015; hereon referred to as difficulty-based RG). To simulate this form of RG, the true item probabilities for each RG simulee were rank ordered in descending order (ties were randomly ordered). Based on the number of RGs within each RG simulee, the items with the lowest probabilities of success were replaced with the chance rate (.25). The second RG pattern reflected simulees progressively engaging in RG as the test progresses, due to cognitive fatigue (hereon referred to as progressive RG; see Wise & Kingsbury, 2016). To reflect effort progressively decreasing, for each RG simulee, the number of known RGs were assigned to items based on descending item order. For instance, if a simulee possessed five RGs, their true probabilities would be replaced with the chance rate for the last five of 30 items on the assessment. As the number of RGs within each RG simulee was allowed to vary from 1 to 30, this assignment process resulted in a greater number of RGs toward the end of the assessment, thus, simulating cognitive fatigue.

Manipulation of RG Misclassifications

Once RG behavior was generated, misclassifications were generated in the data by treating a response probability (could be either effortful or RG based on the misclassification type noted below) as missing, which reflects the approach taken by the EM-IRT model (i.e., all RG responses are indicative of uninformative information). Two independent variables were used to manipulate misclassifications: (a) misclassification type (underclassification and overclassification); and (b) rate of misclassifications (10%, 30%, 50%). These variables were fully crossed with those used to create noneffortful responses (i.e., group impact, percentage of RG, and RG pattern) producing 72 total conditions, with each condition replicated 100 times.

Misclassification Type

Two types of RG misclassifications were investigated: underclassification and overclassification.[7] Although both reflected misidentification issues, the former was included to imitate the occurrence of false negative classifications of true RGs. As noted by Wise (2017),

[7] Although it is possible that current RG classification methods can both over- and underclassify RG, such a condition was not simulated in this study, as it is argued that it would largely lead to a cancellation effect.

this misclassification type may be common in practice as many of the current response time threshold procedures for identifying RG tend to adopt conservative thresholds. To simulate this condition, a percentage of known RGs for RG simulees were randomly selected and treated as observed RGs, while the remaining percentage of true RGs were included in parameter estimation.

In contrast, overclassification reflected false positive RG classifications. This misclassification type may occur in practice due to a threshold procedure that is overly liberal (i.e., setting a high item response time threshold), which results in capturing both RGs and valid responses. To mimic this scenario, two approaches were taken. First, 100% of true RGs from RG simulees were identified based on the assumption that true RGs would be represented with short response times, and thus, would be detected by a procedure using a liberal threshold. Second, a proportion of valid responses (based on the misclassification rate noted below) were incorrectly identified as RGs. This was done by misclassifying responses for the easiest items answered by simulees with ability parameters in the top 30th percentile. The latter decision was employed based on prior literature, which has suggested that the time intensity (i.e., the time required to correctly solve a problem) for easy items is reduced for higher ability examinees (see Goldhammer, 2015; Kyllonen & Zu, 2016), which has been supported by recent research (Loken & Beverly, 2020). In this simulation study, the number of misclassified item responses for each high ability simulee (i.e., simulees in the top 30th percentile) was allowed to vary (ranging from 0 to 10).

Misclassification Rate

Three levels of RG misclassification rates were examined. These rates were meant to reflect small (10%), moderate (30%), and large (50%) degrees of misclassification. In the context of this simulation, the percentage of misclassifications were in relation to the overall percentage of RG in the data. For example, when there were 20% true RGs, a 50% underclassification rate would suggest that only 10% of the true RGs were identified, while a 50% overclassification rate would mean that all 20% of the true RGs were identified, with an additional 10% of valid responses classified as RGs.

Data Analysis

From these manipulations, two datasets were created. The first consisted of item response probabilities of combined effortful and noneffortful responses (i.e., no inclusion of RG misclassifications; hereon referred to as the true RG dataset), while the second consisted of the same, with the addition of RG misclassifications based on manipulating misclassification type and rate (hereon referred to as RG misclassification dataset). The creation of the two datasets allowed for examining the biasing effects due to known RG and RG misclassifications. Across datasets, item response probabilities were compared to a random number sampled from a uniform distribution ranging from 0 to 1. If the random number was less than a given item response probability, a correct response was assigned, otherwise, the response was treated as incorrect.

Item and Ability Parameter Estimation

Item and ability parameters were estimated for the true RG and RG misclassification datasets based on the 3PL and EM-IRT models respectively using the *mirt R* package (Chalmers, 2012). Both datasets were fit using a maximum likelihood confirmatory item factor analysis model for dichotomous data under the IRT paradigm, with an expectation-maximization (EM) algorithm. The EM convergence threshold was .0001 using the Broyden-Fletcher-Goldfarb-Shannon optimization algorithm with the maximum number of cycles set to 500. Ability parameters were obtained via expected a posteriori (EAP) proficiency estimation using the standard normal distribution.[8] This estimator was chosen based on the work of Kim and Moses (2016), who showed EAP to be one of the most robust IRT estimators to atypical response behaviors (Kim & Moses, 2016).

Outcome Variables

Upon estimating model parameters (both item and person parameters were estimated simultaneously), accuracy for both datasets was evaluated based on bias and root mean square error (RMSE). Bias was calculated as:

$$\text{Bias} = \sum_{i=1}^{n}\left(\hat{y}_i \; - y\right), \tag{3}$$

where $\hat{y}$ is the estimated parameter, y is the known parameter, and n is the number of replications. The root mean square error (RMSE) was calculated as:

$$\text{RMSE:}\sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y)^2}{n-1}}. \tag{4}$$

In terms of item parameter recovery, both bias and RMSE were calculated for the *a*, *b*, and *c* parameters. Similarly, ability parameter recovery was evaluated by calculating both bias and RMSE for the mean theta of the total sample as well as RG and non-RG subgroups. As the trends between bias and RMSE tended to be very similar, only the former is reported in the results section; however, the RMSE results are available upon request from the author.

To examine the primary research objectives of the simulation study, two linear regression models were estimated for each dependent variable (*a*, *b*, *c*, total sample theta, RG subsample

[8] The choice to assume a standard normal distribution was based on the fact that free estimation of the latent mean may be largely inaccurate for the EM-IRT model due to the combination of large standard errors caused by missing data (due to treating RG responses as missing) and inaccurate response patterns (due to misclassifying RG responses), To mitigate this bias, a standard normal distribution was assumed; however, this likely led to bias under impact conditions, given that the total sample's ability distribution was not standard normal, but instead had a slightly lower mean. Given that practitioners will not know the underlying ability trait distribution of examinees, this degree of bias introduced is likely reflective of what would be seen in operational settings.

theta bias) using the *lm* function in *R*. The first model was implemented to understand the robustness of item and ability parameter estimates in the presence of RG misclassifications when using the EM-IRT model. Therefore, bias for the parameter estimate of interest from the EM-IRT model served as the dependent variable (five dependent variables), while the five factors investigated (group impact, percentage of RG in the data matrix, RG pattern, misclassification type, and rate of misclassifications) were included as main effects. Specifically, each main effect was treated as a categorical variable, with the no impact, 10% RG, progressive RG pattern, underclassification, and 10% misclassification rate levels serving as the reference groups. In addition, the following interaction effects that consisted of RG item response and misclassification characteristics (the main effect for impact between RG and non-RG subsamples was controlled for as a covariate) were added to the model: (a) percentage of RG x misclassification type; (b) percentage of RG x misclassification rate; (c) RG pattern x misclassification type; (d) RG pattern x misclassification rate; and (e) misclassification type x misclassification rate.

The purpose of the second model was to investigate whether there were differences in bias between the EM-IRT model in the presence of RG misclassifications and the naïve 3PL model, which includes RG responses in scoring. To do this, the above model with the addition of a dichotomous variable for IRT model (the EM-IRT served as the reference) was fit separately for each outcome (*a*, *b*, *c*, total sample theta, and RG subsample theta bias).

Across models, variance-explained was evaluated based on the adjusted R^2 value. Further, statistical significance for factors with more than two levels was evaluated based on the Wald Test, and post-hoc comparisons between levels was investigated using multiple contrasts. Familywise error rate was controlled using the Benjamini-Hochburg procedure by testing for statistical significance based on a false discovery rate of 10% (for details of this procedure, readers are referred to Benjamini & Hochberg, 1995).

Results

Across all conditions, model convergence was met for the 3PL model; however, non-convergence issues were noted for the EM-IRT model when RG overclassifications were equal to 50%, with convergence rates ranging from 89% to 100%. Replications that did not converge were removed from the final analyses.[9]

Parameter Estimation Bias for the EM-IRT Model in the Presence of RG Misclassifications

Table 1 provides the model results of regressing item and ability parameter bias when employing the EM-IRT model on study factors. Below the findings are presented independently for *a*, *b*, and *c* item parameters, while outcomes related to ability parameter bias are described for the total sample and RG subsamples respectively.

Item Parameter Estimates

Descriptive results indicated that the degree of bias observed for the *c* parameter across conditions was near zero for every condition, indicating that the investigated factors had no significant impact. Therefore, results are reported only for the *a* and *b* item parameters.

A Parameter. After controlling for the significant main effect of ability differences between RG and non-RG subsamples (i.e., bias was slightly greater when the RG subsample possessed a mean ability lower than the non-RG subsample), significant interaction effects were observed between: (a) RG percentage and percentage of misclassifications; and (b) misclassification type and percentage (see Table 1). The former finding indicated that bias in the *a* parameter increased as the percentage of misclassifications as a function of RG in the data

[9] Replications that did not converge were re-run to ensure that all conditions were based on the same total number of replications (n =100).

matrix increased. For instance, bias in the *a* parameter rose by 0.16 units when the percentage of misclassifications in relation to the total number of responses increased from 1% to 10%. The latter interaction effect suggested that the degree of bias grew at a higher rate when the percentage of overclassifications increased relative to underclassifications. Examination of Figure 1, which provides a ridgeline plot of *a* parameter bias by misclassification type and percentage separately across impact conditions, further elucidates this relationship. Specifically, this figure shows that the average biases for the overclassification conditions were slightly higher by 0.03, 0.04, and 0.07 units respectively for 10%, 30%, and 50% misclassification rates compared to the underclassification conditions. Taken together, these interaction effects partly accounted for an additional 13% of variance over the main effects model (i.e., the interactions model accounted for 69% of variance in *a* parameter bias).

B Parameter. The investigated main and interaction effects were found to only explain 8% of variance in the *b* parameter bias, with two significant effects. Specifically, when the RG subsample possessed a mean ability that was 0.5 and 1 *SD* lower than the non-RG subsample, *b* parameter estimate bias was on average greater by 0.11 and 0.18 units than under no impact. Furthermore, biased estimates of item difficulty were also associated with a significant interaction between misclassification type and percentage. A closer inspection of Figure 2 shows that compared to overclassification conditions, when RG responses were underclassified by 10%, 30%, and 50%, average bias was greater respectively by 0.05, 0.16, and 0.15 units across impact conditions. Overall, it is clear from Figure 2 that as the percentage of misclassifications increased, the average bias also increased across misclassification types, with greater pronouncements observed as impact grew; however, across conditions, overclassifying RGs was associated with less bias. With that said, model results should be interpreted with caution given the large residual standard error of the regression (0.45).

Ability Parameter Estimates

Below results of ability parameter estimation bias are reported separately for the total sample and RG subsamples.

Total Sample. The main effects model accounted for 99% of variance in ability parameter estimate bias for the total sample. As shown in Table 1, ability differences between RG and non-RG subsamples were found to be associated with the greatest amount of differences in bias. Specifically, when there were no ability differences, the average theta estimate was underestimated by 0.45 *SD*s; however, this negative bias decreased by 0.15 and 0.30 standard deviations (*SD*s) when the RG subsample possessed a mean ability that was 0.5 and 1 *SD* lower than the non-RG subsample, respectively. The percentage and patterns of RG were not associated with bias. Similarly, neither misclassification type nor rate were found to be significantly associated with bias, suggesting that mean ability parameter estimates were robust to RG misclassifications in the data.

RG Subsample. The main and interaction effects investigated accounted for a total of 98% of variance in ability bias for the RG subsample. The main effect for group impact was found to be associated with significant levels of bias. Specifically, compared to conditions in which no impact was present, the mean ability for the RG subsample was overestimated by 0.30 and 0.62 *SD*s respectively for impact conditions of 0.5 and 1 *SD*. After controlling for this variable, significant interaction effects were observed for: (a) RG percent and misclassification type; and (b) misclassification percentage and type. Concerning the former, bias was on average greater by 0.22 *SD*s across conditions when the RG percent was equal to 20% and RG responses were overclassified versus the condition in which RG responses comprised 10% of the data matrix and RG responses were underclassified.

This result is further corroborated by the significant interaction between misclassification type and percentage; however, as shown in Figure 3, this result was largely moderated by group impact. Specifically, under no impact, the observed influence of misclassification type on ability bias was as expected. For instance, for RG overclassification, under no impact, bias was approximately zero across RG misclassification rate conditions, as RG was correctly identified with 100% accuracy. Concerning RG underclassification, Figure 3 shows bimodal distributions representing descriptive differences between RG response patterns, with greater negative bias observed for progressive RG responding. As item difficulty was randomly assigned across the test form, greater negative bias was observed as simulees engaged in RG on items in which they had a high probability of success. Once aggregating results across conditions by RG percentage and pattern, under no impact, negative bias increased from -0.08 *SD*s to -0.33 *SD*s for misclassification percentages of 10% and 50%, respectively. The increase observed was associated with a greater rate of RGs from above average simulees in the RG subsample, which led to incorrect responses to items in which they possessed a high probability of success.

As the difference in mean ability between the two subsamples increased, the RG subsample's ability bias was overestimated across nearly every misclassification percentage and type. For example, for overclassification conditions, bias increased to 0.63 *SD*s for an average theta difference of 1 *SD* between RG subgroups, while bias decreased from 0.54 *SD*s to 0.35 *SD*s as the percentage of underclassifications increased from 10% to 50%, respectively (see Figure 3). This latter result indicated that an RG subsample predominately comprised of below average ability simulees generally benefited from RG correctly across items in which the item difficulty parameter estimates were overestimated; however, as the rate of RG that went undetected increased, ability parameter estimates were less overestimated, due to the overall proportion correct for RGs approaching the chance level.

Non-RG Subsample. The main and interaction effects investigated accounted for 95% of variance in the ability bias for the non-RG subsample. A significant main effect for group impact was noted. Although smaller than observed for the RG subsample, compared to no impact, non-RG ability was overestimated by an average of 0.09 and 0.16 *SD*s when the non-RG subsample possessed a true ability that was 0.5 and 1 *SD* higher than the RG subgroup. Accounting for this main effect, a significant interaction between misclassification type and percentage was observed. As is shown in Figure 4, the degree of bias differed between misclassification types. In particular, average bias was lower when overclassifying RG, with a consistent overestimation of 0.08 *SD*s across impact conditions, regardless of the misclassification percentage. This result is to be expected as overclassifications occurred for simulees with true ability in the top 30th percentile receiving the easiest items. Thus, misclassifying valid responses to such items provided little bias, as these items were largely uninformative to estimating these simulees' true ability.

Comparatively, non-RG subsample bias was greater when underclassifications occurred (i.e., misclassifications were only present for the RG subsample). For instance, the degree of non-RG subsample ability bias averaged across group impact conditions was 0.12, 0.17, and 0.21 *SD*s with underclassification rates of 10%, 30%, and 50%, respectively. This overestimation in non-RG subsample ability was likely due to biased item difficulty parameter estimates in which items appeared to be more difficult than they were. Consequently, correct responses to such items were deemed to represent a higher proficiency, which led to high ability simulees appearing to be more proficient than they were in truth.

Comparison between the 3PL and EM-IRT Models

Results comparing the two models are presented separately for item and ability parameter estimates. Table 2 presents the regression results comparing the 3PL and EM-IRT models on item and ability parameter estimate bias after controlling for the main and interaction effects.

Item Parameter Estimates

Due to the minimal bias observed for the *c* parameter, the description of findings focuses on all other item parameter outcomes. Overall, results indicated significant differences between the two models for the *a* and *b* item parameters, with the 3PL respectively overestimating each parameter by an average of 0.21 and 0.23 units compared to the EM-IRT model. Figures 1 and 2 further elucidate these findings by showing the bias for the 3PL model. Across these figures, the ridgeline plots show a multimodal distribution for the 3PL model reflecting descriptive differences between RG pattern, with greater bias observed for progressive RG responding. Regardless, for both item parameters, bias was consistently lower for the EM-IRT model irrespective of misclassification type, misclassification percentage, and the presence of group impact.

Ability Parameter Estimates

In terms of ability parameter estimates for the total sample, no significant model differences were found; however, across conditions, significant model differences were observed for RG and non-RG subsample ability estimates (Table 2). Concerning the former, compared to the EM-IRT model, the average bias was lower for the 3PL model by 0.37 *SD*s. A closer examination of Figure 3 shows that the absolute degree of bias for the EM-IRT model was smaller across conditions when RG was overclassified (regardless of percent misclassified) under no impact. Furthermore, although bias was smaller for the EM-IRT model when underclassifying RG by 10%, the 3PL model provided similar degrees of bias as underclassification rates grew under no impact conditions. This was likely due to higher ability simulees RG on items to which they would be expected to correctly answer based on their true ability.

Across conditions in which group impact was present, the findings painted a more complex picture. Specifically, when impact was equal to 0.5 *SD*s, underclassifying RGs with the EM-IRT model largely led to better estimates of RG responder mean ability than the 3PL model, particularly as the percentage of misclassifications increased. This result reflects the fact that the benefits of RG dissipate as simulees engage in higher rates of RG (i.e., as RGs increase, the average probability of success more closely reflects chance), and thus, underclassifying such behaviors leads to relatively accurate theta estimates for low ability simulees. However, this applies only to a certain degree. That is, when RG responders were of very low ability (i.e., group impact was equal to 1 *SD*), attempting to identify RGs and apply the EM-IRT model was associated with overestimation of ability, regardless of misclassification type. In such circumstances, results indicated that including RGs in scoring by applying the 3PL model led to the least degree of overall bias. The reason for this is that many of the simulees engaging in RG possessed low probabilities of correctly responding to items, and thus, their ability estimates were positively biased (i.e., RG improved their scores), but the degree of bias was less than the EM-IRT model. However, as reflected in the multimodal distributions of the 3PL model shown in Figure 3, the degree of bias for this model was associated with the overall percentage of RGs in the data matrix (i.e., larger percentages of RG were associated with greater bias) and RG pattern (i.e., greater bias was observed for progressive RG).

In addition, the models differed in ability parameter estimate bias of the non-RG subsample. Specifically, the 3PL model on average possessed a bias that was 0.16 *SD*s higher than the EM-IRT model across conditions. Figure 4 illustrates this finding by showing that the EM-IRT model produced bias that was lower for all conditions, including those in which a high degree of group impact was present and misclassification rates were as large as 50%.

Discussion

The objective of this study was to examine the impact of RG misclassifications on item and ability parameter estimates when employing the EM-IRT model. Overall, the findings from the simulation analysis indicate that the EM-IRT model is susceptible to non-convergence issues when the degree of overclassified RG reaches rates of 30% or higher. For those models that did converge, *a* and *b* parameter estimates from the EM-IRT model were found to be susceptible to bias in the presence of RG misclassifications, while the *c* parameter was not observed to be impacted. Although the parameter estimates from the EM-IRT model were negatively influenced by RG misclassifications, this model outperformed the naïve 3PL model across all conditions examined.

Concerning ability parameter estimates, the EM-IRT model produced unbiased theta estimates for the total sample similar to the 3PL model. The lack of difference in models is likely because total sample estimates of ability have been shown to be robust to high RG rates (e.g., Rios & Soland, 2020; Wise et al., 2020). As a result, misclassifications of RG were shown to have no impact. However, when examining the non-RG subgroup's ability parameter estimate bias, the EM-IRT model was found to outperform the 3PL model across misclassification type, misclassification rates, and subgroup impact conditions.

In terms of the RG subsample, ability estimates for the EM-IRT model were more accurate than the 3PL model when RG was correctly classified with 100% accuracy (i.e., RG was overclassified) or underclassified by 10% and no group impact was present. As expected, under no group impact, bias for the EM-IRT model was found to increase as the degree of underclassification increased, as many of the item responses to which above average simulees guessed on had a high probability of a correct response. Furthermore, although group impact was also associated with increased bias for the non-RG subsample, its effect on the RG subgroup's ability estimates was large. This finding supports prior research in showing that aggregate-level ability parameter estimation accuracy is greatly undermined when violating the EMIRT model's assumption that the true abilities of rapid guessers are reflective of the sample distribution (e.g., Rios et al., 2017; Rios & Soland, 2020). Thus, in most contexts where group impact is present, the 3PL model can provide more accurate estimates for examinees that have engaged in RG compared to the EM-IRT model.

Limitations and Future Research

In interpreting the findings from this study, several limitations should be noted. First, this study examined only one approach to handling RG. As noted, a number of models have been developed, however, this study focused solely on the EM-IRT model due to its popularity in the literature as well as its current use in operational settings. Future research should examine the comparative utility between the EM-IRT model and other response time models (e.g., Wang and Xu's [2015] mixture hierarchical model). In addition, there may be alternative scoring procedures that could be employed to handle RG that do not rely on response times, which may be of particular importance for assessments that are administered via paper-and-pencil. For instance, one could employ a modification to maximum likelihood estimation equations by downweighting observations that are prone to response disturbances using Huber-Type weights (see Schuster & Yuan, 2011). This approach would allow for practitioners to handle potential RG when log-file information is not readily available, and could also loosen the restrictions on sample size requirements observed for complex mixture models when response times are accessible. Future research should evaluate such an approach compared to alternatives.

Second, in this study, ability estimation for both the EM-IRT and 3PL models relied on EAP scoring. As a consequence, the results may have been biased as ability estimates using this scoring procedure are shrunk toward the mean. That is, in general, greater shrinkage is observed as the degree of missing data increases. This may have been one reason why RG misclassification rates had little impact on the total sample mean ability estimate, particularly for conditions in which RG misclassification rates were large. Future research should examine this issue by comparing model results when using EAP and maximum likelihood scoring. An additional limitation related to estimation was that both item and person parameters were jointly estimated. As a result, bias in item parameters were associated with increased bias in ability estimates, over and above the bias associated with RG misclassifications. This may have been a particular concern under group impact conditions as the assumption that the trait levels followed a standard normal distribution was violated. To mitigate this issue, future research should examine the utility of fixed item parameter ability estimation in the presence of RG. This could be done by first estimating item parameters based on a filtered sample (i.e., either removing RGs or down-weighting potential aberrant responses as noted above), and then treating these item parameter estimates as fixed to estimate ability for the entire sample.

Implications

In light of the limitations, this study provides implications for practitioners confronted with the decision of choosing a response time threshold. As noted by Wise (2017), when practitioners are making this decision, they must often weigh whether to adopt a conservative approach to limit false positives (i.e., identifying a valid response as RG), while increasing the potential for false negatives (i.e., classifying RG as a valid response), or adopt a liberal approach that has the opposite effect. Assuming that false positives of RG occur due to high ability examinees responding very quickly to easy items, the findings from this simulation analysis indicate that adopting conservative response time threshold procedures may be associated with greater bias in item difficulty and ability parameter estimates than procedures that overclassify RG when employing the EM-IRT model for scoring. Therefore, it may be preferable in many cases to adopt liberal response time threshold procedures to limit false negative classifications of RG.

An additional implication is that due to the overestimation of item difficulty due to RG misclassifications, practitioners may consider replacing joint estimation of item and ability parameters with fixed item parameter ability estimation to mitigate the influence of biased item parameters on ability estimates. As noted, further research is needed to evaluate the improvements of this latter approach over the traditional joint estimation procedure. Lastly, although the EM-IRT model was found to provide improved estimation of item, and in many cases, ability parameters, increased bias was generally observed when subgroup impact was present. As a consequence, it is recommended that practitioners make every effort to evaluate the tenability of the assumption that RG is unrelated to examinees' true ability prior to using the EM-IRT model. This may be done by utilizing prior achievement measures to investigate ability differences between RG and non-RG subsamples (see Rios et al., 2017). If such information is unavailable, the multidimensional approach to estimating the covariance between RG behavior and performance estimates proposed by Liu et al. (2019) may be helpful.

References

American Educational Research Association, American Psychological Association, & National Council for Assessment in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.

Attali, Y., & Bar-Hillel, M. (2003). Guess where: The position of correct answers in multiple-choice test items as a psychometric variable. *Journal of Educational Measurement*, *40*(2), 109-128.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289-300.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1-29.

Cronbach, L. J. (1960). Essentials of psychological testing (2nd ed.). New York, NY: Harper & Row.

De Boeck, P., & Jeon, M. (2019). An overview of models for response times and processes in cognitive tests. *Frontiers in Psychology*, 10, 102.

Debeer, D., Buchholz, J., Hartig, J., & Janssen, R. (2014). Student, school, and country differences in sustained test-taking effort in the 2009 PISA reading assessment. *Journal of Educational and Behavioral Statistics*, *39*, 502-523.

DeMars, C. E. (2007). Changes in rapid-guessing behavior over a series of assessments. *Educational Assessment*, *12*(1), 23-45.

DeMars, C. E., & Wise, S. L. (2010). Can differential rapid-guessing behavior lead to differential item functioning? *International Journal of Testing*, *10*(3), 207-229.

Goldhammer, F. (2015). Measuring ability, speed, or both? Challenges, psychometric solutions, and what can be gained from experimental control. *Measurement: interdisciplinary research and perspectives*, *13*(3-4), 133-164.

Goldhammer, F., Martens, T., Christoph, G., & Lüdtke, O. (2016). *Test-taking engagement in PIAAC* (OECD Education Working Papers, No. 133). OECD Publishing.

Guo, H., Rios, J. A., Haberman, S., Liu, O. L., Wang, J., & Paek, I. (2016). A new procedure for detection of students' rapid guessing responses using response time. *Applied Measurement in Education*, *29*(3), 173-183.

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.

Kim, S., & Moses, T. (2016). *Investigating robustness of item response theory proficiency estimators to atypical response behaviors under two-stage multistage testing* (ETS RR-16-22). Educational Testing Service.

Kyllonen, P. C., & Zu, J. (2016). Use of response time for measuring cognitive ability. *Journal of Intelligence*, *4*(14), 1-29.

Liu, Y., Li, Z., Liu, H., & Luo, F. (2019). Modeling test-taking non-effort in MIRT models. *Frontiers in Psychology*, *10*, 145.

Liu, O. L., Rios, J. A., & Borden, V. (2015). The effects of motivational instruction on college students' performance on low-stakes assessment. *Educational Assessment*, *20*(2), 79-94.

Loken, E., & Beverly, T. (2020, July). *The covariance structure of response time and accuracy during a test.* Paper presented at the annual conference of the Psychometric Society, College Park, MD.

Meyer, J. P. (2010). A mixture Rasch model with item response time components. *Applied Psychological Measurement*, *34*(7), 521-538.

Mittelhaëuser, M. A., Béguin, A. A., & Sijtsma, K. (2015). The effect of differential motivation on IRT linking. *Journal of Educational Measurement*, *52*, 339-358.

Osborne, J. W., & Blanchard, M. R. (2011). Random responding from participants is a threat to the validity of social science research results. *Frontiers in Psychology*, *1*, 220.

Pastor, D. A., Ong, T. Q., & Strickman, S. N. (2019). Patterns of solution behavior across items in low-stakes assessments. *Educational Assessment*, *24*(3), 189-212.

Penk, C., & Schipolowski, S. (2015). Is it all about value? Bringing back the expectancy component to the assessment of test-taking motivation. *Learning and Individual Differences*, *42*, 27-35.

R Core Team (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rios, J. A., & Guo, H. (2020). Can culture be a salient predictor of test-taking engagement? An analysis of differential noneffortful responding on an international college-level assessment of critical thinking. *Applied Measurement in Education*. Advance online publication. DOI: 10.1080/08957347.2020.1789141

Rios, J. A., Guo, H., Mao, L., & Liu, O. L. (2017). Evaluating the impact of noneffortful responses on aggregated scores: To filter unmotivated examinees or not? *International Journal of Testing*, *17*, 74-104.

Rios, J. A., Liu, O. L., & Bridgeman, B. (2014). Identifying low-effort examinees on student learning outcomes assessment: A comparison of two approaches. *New Directions for Institutional Research*, *2014*(161), 69-82.

Rios, J. A., & Soland, J. (2020). Parameter estimation accuracy of the Effort-Moderated IRT model under multiple assumption violations. *Educational and Psychological Measurement*. Advance online publication. doi: 10.1177/0013164420949896

Schnipke, D. L., & Scrams, D. J. (1997). Modeling response times with a two-state mixture model: A new method of measuring speededness. *Journal of Educational Measurement*, *34*, 213-232.

Schuster, C., & Yuan, K. H. (2011). Robust estimation of latent ability in item response models. *Journal of Educational and Behavioral Statistics*, *36*(6), 720-735.

Silm, G., Pedaste, M., & Täht, K. (2020). The relationship between performance and test-taking effort when measured with self-report or time-based instruments: A meta-analytic review. *Educational Research Review*, *31*, 100335.

van Barnevald, C. (2007). The effect of examinee motivation on test construction within an IRT framework. *Applied Psychological Measurement*, *31*, 31-46.

Wang, C., & Xu, G. (2015). A mixture hierarchical model for response times and response accuracy. *British Journal of Mathematical and Statistical Psychology*, *68*(3), 456-477.

Wang, C., Xu, G., Shang, Z., & Kuncel, N. (2018). Detecting aberrant behavior and item preknowledge: A comparison of mixture modeling method and residual method. *Journal of Educational and Behavioral Statistics*, *43*(4), 469-501.

Wise, S. L. (2017). Rapid-guessing behavior: Its identification, interpretation, and implications. *Educational Measurement: Issues and Practice*, *36*(4), 52-61.

Wise, S. L. (2020). The impact of test-taking disengagement on item content representation. *Applied Measurement in Education*, *33*(2), 83-94.

Wise, S. L., & DeMars, C. E. (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement*, *43*(1), 19-38.

Wise, S. L., & DeMars, C. E. (2009). A clarification of the effects of rapid guessing on coefficient α: A note on Attali's ''Reliability of speeded number-right multiple-choice tests''. *Applied Psychological Measurement*, *33*(6), 488-490.

Wise, S. L., & DeMars, C. E. (2010). Examinee noneffort and the validity of program assessment results. *Educational Assessment*, *15*, 27-41.

Wise, S. L., & Kingsbury, G. G. (2016). Modeling student test-taking motivation in the context of an adaptive achievement test. *Journal of Educational Measurement*, *53*(1), 86-105.

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, *18*(2), 163-183.

Wise, S. L., & Kuhfeld, M. R. (2020). Using retest data to evaluate and improve effort-moderated scoring. *Journal of Educational Measurement*. Advance online publication. DOI: 10.1111/jedm.12275

Wise, S. L., & Ma, L. (2012, April 14–16). *Setting response time thresholds for a CAT item pool: The normative threshold method* [Paper presentation]. National Council on Measurement in Education 74th Annual Meeting, Vancouver, BC, Canada.

Wise, S. L., Ma, L., Cronin, J., & Theaker, R. A. (2013, April). *Student test-taking effort and the assessment of student growth in evaluating teacher effectiveness*. Presented at the annual conference of the American Educational Research Association, San Francisco, CA.

Wise, S. L., Soland, J., & Bo, Y. (2020). The (non) impact of differential test taker engagement on aggregated scores. *International Journal of Testing*, *20*(1), 57-77.

Table 1

Model Results for Regressing Item and Ability Parameter Bias for the EM-IRT Model on Study Factors

	Item Parameter Estimate			**Ability Parameter Estimate**		
Factor	**A Parameter** **R^2 = .69**	**B Parameter** **R^2 = .08**	**C Parameter** **R^2 = .26**	**Total Sample** **R^2 = .99**	**RG Subsample** **R^2 = .98**	**Non-RG Subsample** **R^2 = .95**
Intercept	.14* (.02)	-.69* (.18)	-.14* (.01)	-.45* (.00)	-.17* (.02)	-.57* (.01)
Impact: -0.5	.04* (.00)	.11* (.01)	.02* (.00)	.15* (.00)	.30* (.00)	.09* (.00)
Impact: -1	.09* (.00)	.18* (.01)	.02* (.00)	.30* (.00)	.62* (.00)	.16* (.00)
RG %	-.26* (.01)	-.07 (.07)	.03* (.00)	-.00 (.00)	-.16* (.00)	.06* (.00)
RG Pattern	.02 (.01)	.07 (.07)	-.00 (.00)	-.00 (.00)	-.00 (.00)	.00* (.00)
Misclassify Type	-.07* (.01)	.28* (.08)	.01 (.00)	-.00 (.00)	-.52* (.00)	.22* (.00)
Misclassify %: 30%	-.06* (.01)	.18 (.07)	.05* (.00)	-.00 (.00)	-.17* (.00)	.07* (.00)
Misclassify %: 50%	-.20* (.01)	.18 (.07)	.08* (.00)	-.00 (.00)	-.32* (.00)	.13* (.00)
RG % x Misclassify Type	.02* (.00)	-.09* (.02)	-.01* (.00)	.00 (.00)	.22* (.00)	-.09* (.00)
RG % x Misclassify %: 30%	.06* (.00)	.07 (.03)	.00 (.00)	.00 (.00)	-.05* (.00)	.03* (.00)
RG % x Misclassify %: 50%	.16* (.00)	.14* (.03)	.00 (.00)	.00 (.00)	-.09* (.00)	.04* (.00)
RG Pattern x Misclassify Type	.00 (.00)	.02 (.02)	-.01* (.00)	.00 (.00)	-.04* (.00)	.02* (.00)
RG Pattern x Misclassify %: 30%	-.01 (.00)	-.03 (.03)	-.00* (.00)	.00 (.00)	.01 (.00)	-.00 (.00)
RG Pattern x Misclassify %: 50%	-.00 (.00)	-.05 (.03)	-.00 (.00)	.00 (.00)	.01* (.00)	-.00* (.00)
Misclassify Type x Misclassify %: 30%	.02* (.00)	-.11* (.03)	-.01* (.00)	-.00 (.00)	.12* (.00)	-.05* (.00)
Misclassify Type x Misclassify %: 50%	.05* (.00)	-.11* (.03)	-.02* (.00)	-.00 (.00)	.22* (.00)	-.09* (.00)

Note. Standard errors are noted in parentheses. * *p* <.001

Table 2

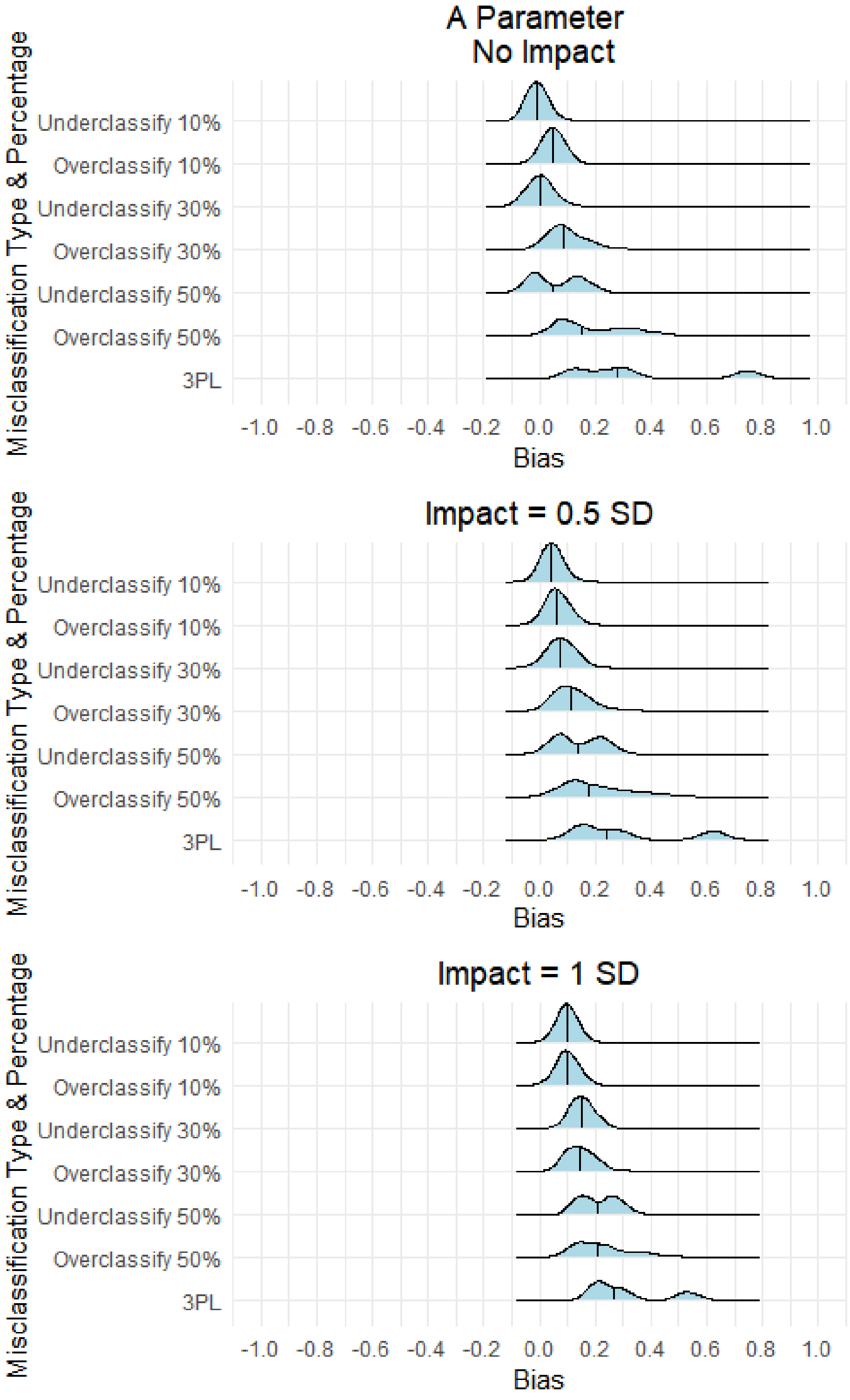
Comparison of Parameter Estimation Bias between EM-IRT and 3PL Models

	Item Parameter Estimate			**Ability Parameter Estimate**		
Factor	**A Parameter**	**B Parameter**	**C Parameter**	**Total**	**RG**	**Non-RG**
	***R^2* = .63**	***R^2* = .19**	***R^2* = .31**	**Sample**	**Subsample**	**Subsample**
				***R^2* = .99**	***R^2* = .87**	***R^2* = .77**
Intercept	-.18* (.03)	-.58* (.09)	.01 (.01)	-.45* (.00)	-.20* (.04)	-.54* (.02)
Impact: -0.5	-.00 (.00)	.07* (.01)	-.01* (.00)	.15* (.00)	.32* (.00)	.08* (.00)
Impact: -1	.03 (.00)	.09* (.01)	.01* (.00)	.30* (.00)	.69* (.00)	.14* (.00)
RG %	.02 (.01)	.12* (.04)	-.01* (.00)	-.00 (.00)	-.26* (.02)	.11* (.00)
RG Pattern	.10* (.01)	.10 (.04)	.02* (.00)	-.00 (.00)	-.16* (.02)	.07* (.00)
Misclassify Type	-.03 (.01)	.15* (.04)	.02* (.00)	-.00 (.00)	-.26* (.02)	.11* (.00)
Misclassify %: 30%	-.03 (.01)	.10 (.04)	.02* (.00)	-.00 (.00)	-.08* (.02)	.03* (.00)
Misclassify %: 50%	-.09* (.01)	.10 (.04)	.01* (.00)	-.00 (.00)	-.16* (.02)	.06* (.00)
RG % x Misclassify Type	.01 (.00)	-.05* (.01)	-.00* (.00)	.00 (.00)	.11* (.00)	-.04* (.00)
RG % x Misclassify %: 30%	.03* (.00)	.03 (.01)	.00 (.00)	.00 (.00)	-.03* (.00)	.01* (.00)
RG % x Misclassify %: 50%	.08* (.00)	.07* (.01)	.00 (.00)	.00 (.00)	-.05* (.01)	.02* (.00)
RG Pattern x Misclassify Type	.00 (.00)	.02 (.01)	.00* (.00)	.00 (.00)	-.02* (.01)	.01* (.00)
RG Pattern x Misclassify %: 30%	-.00 (.00)	-.02 (.01)	.00* (.00)	.00 (.00)	.00 (.01)	-.00 (.00)
RG Pattern x Misclassify %: 50%	-.00 (.00)	-.03 (.01)	-.00 (.00)	.00 (.00)	.01 (.01)	-.00 (.00)
Misclassify Type x Misclassify %: 30%	.01 (.00)	-.06* (.01)	-.00* (.00)	-.00 (.00)	.06* (.01)	-.03* (.00)
Misclassify Type x Misclassify %: 50%	.02 (.00)	-.06* (.01)	-.01* (.00)	-.00 (.00)	.11* (.01)	-.05* (.00)
Model	.21* (.00)	.23* (.01)	.01* (.00)	-.00 (.00)	-.37* (.00)	0.16* (.00)

Note. Standard errors are noted in parentheses. * *p* <.001

Figure 1

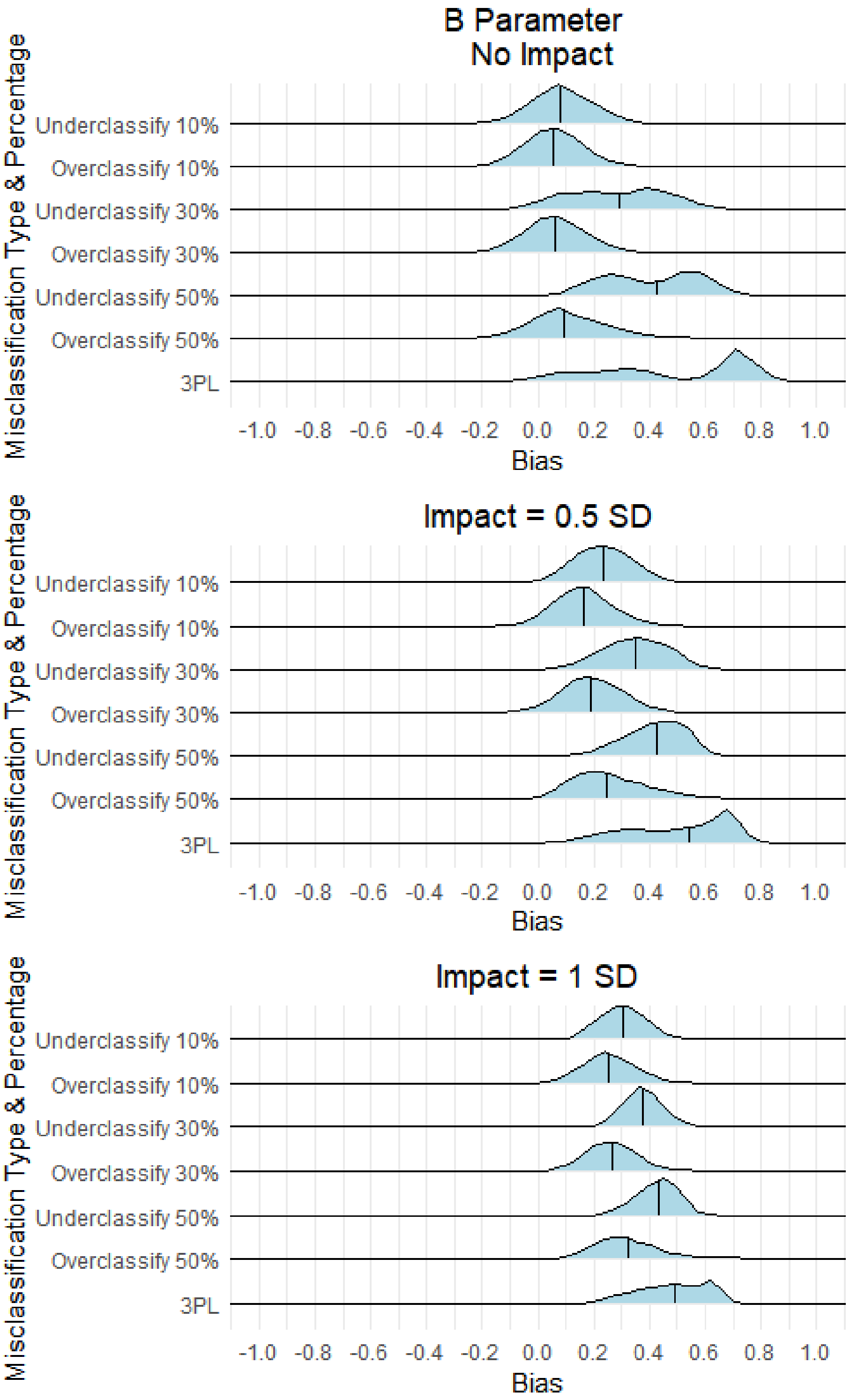
A Parameter Bias based on Misclassification Type and Percentage



Note. Results presented are based on aggregating conditions across group RG percentage and RG pattern. The vertical line in each distribution represents the median value.

Figure 2

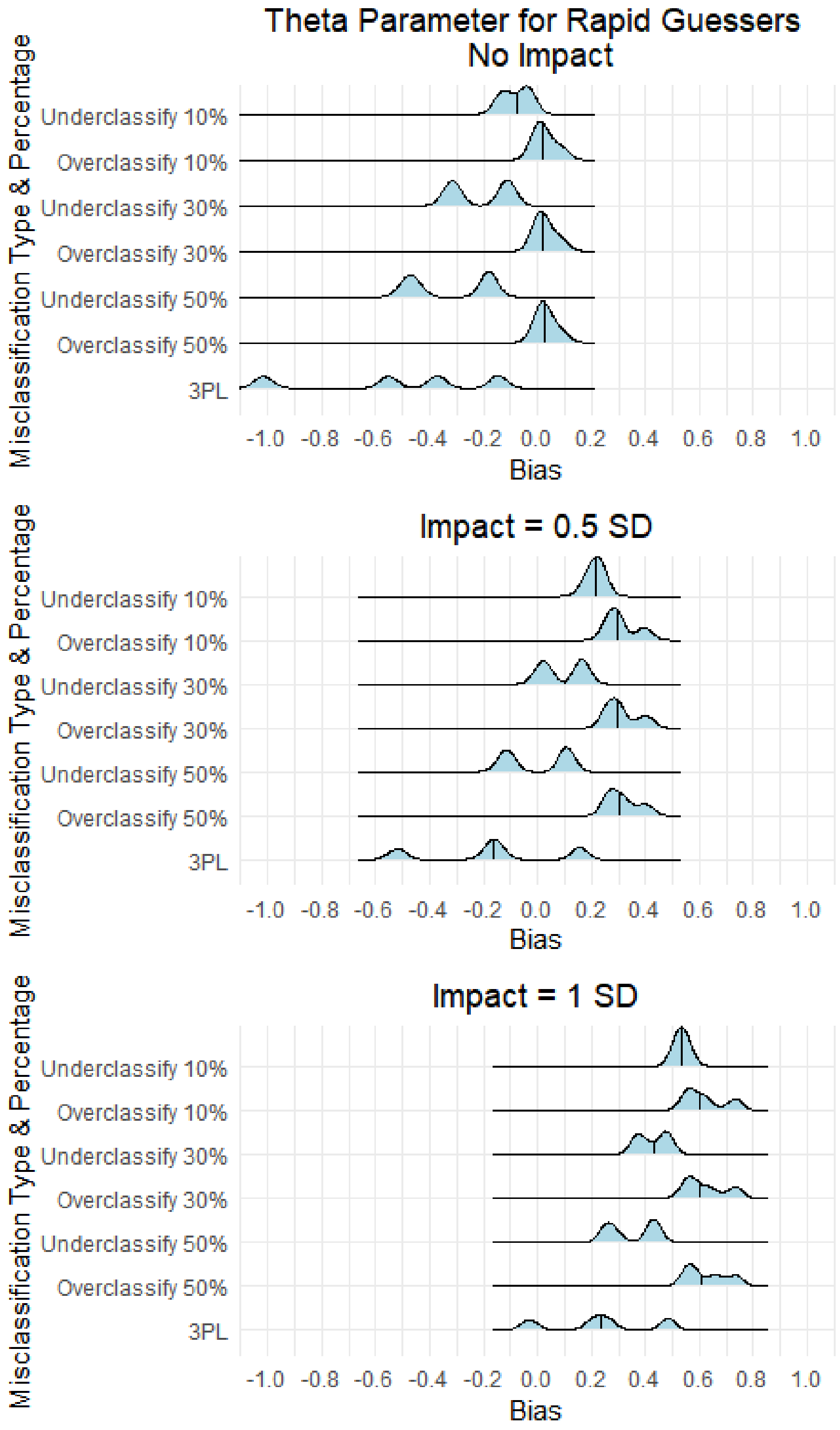
B Parameter Bias based on Misclassification Type and Percentage



Note. Results presented are based on aggregating conditions across group RG percentage and RG pattern. The vertical line in each distribution represents the median value.

Figure 3

Ability Parameter Bias for Rapid Guesser Subgroup by Misclassification Type and Percentage



Theta Parameter for Rapid Guessers
No Impact

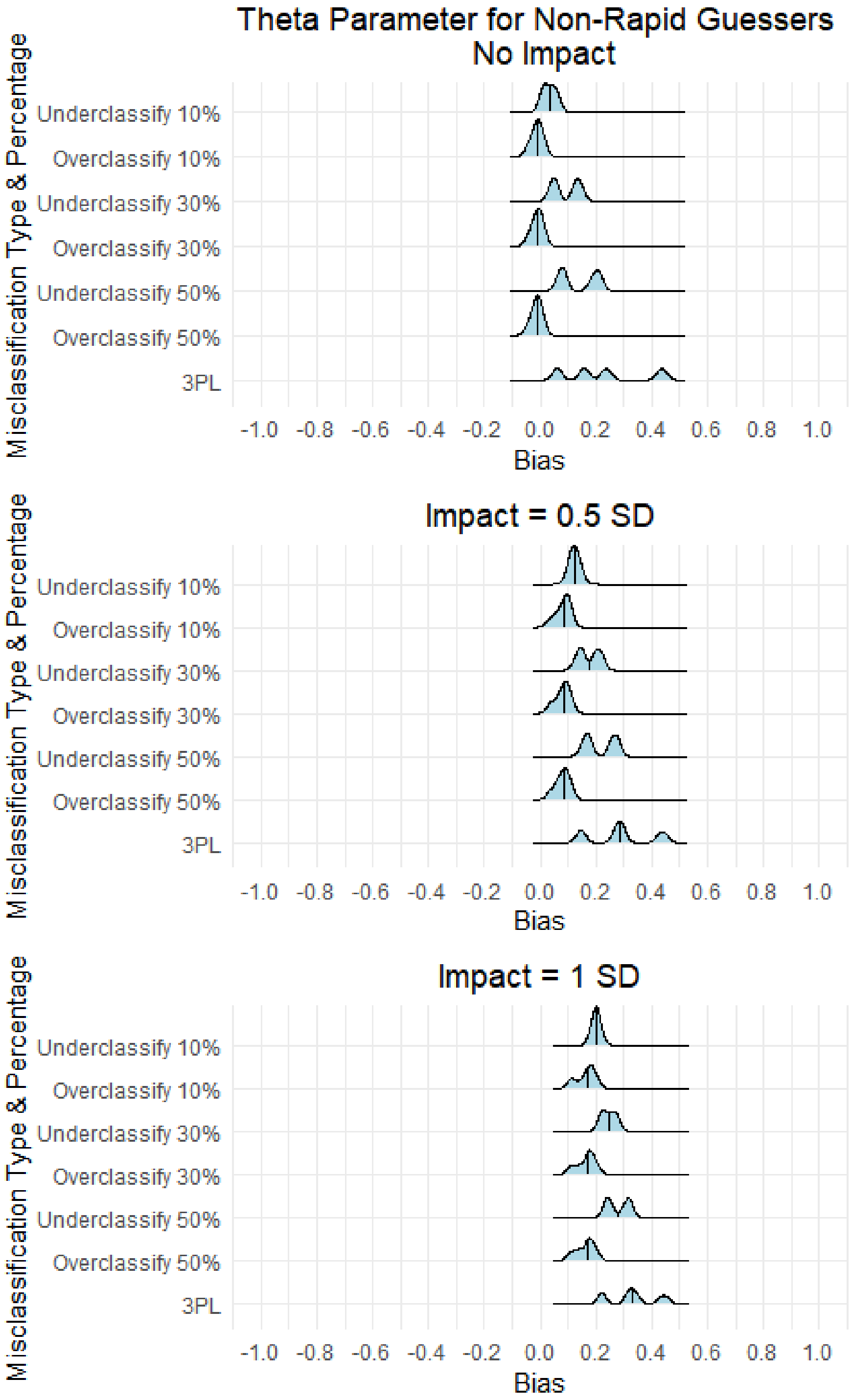
Impact = 0.5 SD

Impact = 1 SD

Note. Results presented are based on aggregating conditions across group RG percentage and RG pattern. The vertical line in each distribution represents the median value.

Figure 4

Ability Parameter Bias for Non-Rapid Guesser Subgroup by Misclassification Type and Percentage



Note. Results presented are based on aggregating conditions across group RG percentage and RG pattern. The vertical line in each distribution represents the median value.