Parameter Estimation Accuracy of the Effort-Moderated IRT Model Under Multiple Assumption Violations

Joseph A. Rios[a] & James Soland[bc]

[a]University of Minnesota

[b]University of Virginia

[c]NWEA

Author's Note

The first author holds departmental affiliation at the Department of Educational Psychology, University of Minnesota, Twin Cities. The second author is affiliated with the Department of Leadership, Foundations and Policy at the University of Virginia.

The authors would like to thank Michael Rodriguez from the University of Minnesota and Megan Kuhfeld from the Northwest Evaluation Association for their helpful comments on an earlier draft.

Author contribution statement: The first author conceived of the presented idea and wrote the R syntax for the simulation analyses. The second author identified the datasets and conducted the analyses for the applied analyses. All authors interpreted findings, drafted the article, and conducted critical revisions of the article throughout the review process. Final approval of the version to be published was made by all authors.

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., & 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

Abstract

As low-stakes testing contexts increase, low test-taking effort may serve as a serious validity threat. One common solution to this problem is to identify noneffortful responses and treat them as missing during parameter estimation via the Effort-Moderated IRT (EM-IRT) model. Although this model has been shown to outperform traditional IRT models (e.g., 2PL) in parameter estimation under simulated conditions, prior research has failed to examine its performance under violations to the model's assumptions. Therefore, the objective of this simulation study was to examine item and mean ability parameter recovery when violating the assumptions that noneffortful responding occurs randomly (assumption #1) and is unrelated to the underlying ability of examinees (assumption #2). Results demonstrated that, across conditions, the EM-IRT model provided robust item parameter estimates to violations of assumption #1. However, bias values greater than 0.20 *SDs* were observed for the EM-IRT model when violating assumption #2; nonetheless, these values were still lower than the 2PL model. In terms of mean ability estimates, model results indicated equal performance between the EM-IRT and 2PL models across conditions. Across both models, mean ability estimates were found to be biased by more than 0.25 *SDs* when violating assumption #2. However, our accompanying empirical study suggested that this biasing occurred under extreme conditions that may not be present in some operational settings. Overall, these results suggest that the EM-IRT model provides superior item and equal mean ability parameter estimates in the presence of model violations under realistic conditions when compared to the 2PL model.

Keywords. test-taking effort, noneffortful responding, parameter estimation, item response theory

Parameter Estimation Accuracy of the Effort-Moderated IRT Model Under Multiple Assumption Violations

The assumption underlying all educational and psychological assessments is that a score reflects the ability, skills, or proficiencies assessed. However, such an assumption is violated when students fail to employ maximal effort to correctly answer test items to the best of their ability (hereon referred to as low test-taking effort). In such instances, low test-taking effort can be viewed as a source of construct-irrelevant variance, as it typically leads to an underestimation of an examinee's true ability (e.g., Rios et al., 2017). The issue of low test-taking effort has been shown to be a particular threat for group-based accountability testing programs across the United States (e.g., the National Assessment of Educational Progress [NAEP]) and internationally (e.g., Programme for International Student Assessment [PISA]). In these assessment contexts, inferences are made at an aggregate-level (e.g., state or country) and thus have no personal consequences for examinees (i.e., the testing context is low-stakes; in most cases, scores are not even provided to individual examinees; Goldhammer et al., 2016). However, results from these assessments have real-world implications for educational policymakers, as they are used for prescribing reform related to teaching and student learning. Thus, there is the potential for low test-taking effort to lead to inaccurate inferences for stakeholders concerned with educational accountability (Wainer, 1993). This has led some researchers to suggest that low test-taking effort is "…one of the most vexing assessment problems that we face today" (Erwin & Wise, 2002, p. 71).

To assist practitioners in making valid inferences from assessments in which low test-taking effort is present among examinees, a small number of Item Response Theory (IRT) models have been proposed (for a discussion, see Wise & Kingsbury, 2016). To date, the

approach that has received the most attention in the literature consists of two stages. First, noneffortful responses (i.e., invalid item responses that are provided with disregard to the item content due to low-testing effort) are generally identified by relying on response time and/or accuracy data to detect examinee responses that do not reflect the examinees' underlying knowledge, abilities, or skills, due to a failure to expend full effort (for a greater discussion on classifying noneffortful response using response time and/or accuracy data see Wise, 2017).[1] Second, data are purified by either removing individual examinees found to engage in noneffortful responding (examinee-level filtering; see Hong et al., 2020) or noneffortful responses are treated as missing data and ability is essentially estimated based solely on item responses judged to be effortful (response-level filtering). Rios et al. (2017) compared these two approaches and found that the former led to greater bias in ability estimates when noneffortful responding was related to examinees' underlying ability. Thus, the latter approach has become increasingly popular as it both provides more robust estimates under certain conditions and avoids deleting a potentially large percentage of the sample.

In research, the most widely used and thoroughly validated model that follows this two-step approach is likely the Effort-Moderated IRT (EM-IRT) model proposed by Wise and DeMars (2006). The EM-IRT model is common in research for two reasons. First, it is less computationally demanding than some of the proposed mixture model approaches that can simultaneously identify noneffortful responses and estimate parameters (e.g., Lu et al., 2020; Wang & Xu, 2015; Wang et al., 2018). Second, it outperforms traditional IRT models (i.e.,

[1] This term is also referred to as "rapid guessing" (see Wise, 2017); however, we refrain from the use of this terminology as new response time threshold procedures have begun to classify noneffortful responses based on a combination of response time and accuracy data. As a result, item responses classified as invalid are no longer based on the assumption that responses are provided rapidly. In fact, the opposite may be true. Some examinees may spend an exorbitant amount of time in providing a noneffortful response, due to off-task behaviors (e.g., daydreaming). Thus, the term "rapid guessing" may be inappropriate in these contexts.

models that do not filter invalid responses) in terms of item parameter, person parameter, and test information estimation accuracy when its underlying assumptions are met (Kong, 2007; Wise & DeMars, 2006; Wise & Kingsbury, 2016).

Despite the prevalence of the EM-IRT model, major gaps in our understanding of its performance under different conditions remain. For example, its ability to recover measurement parameters is unknown when its basic assumptions are not satisfied. This gap in the literature is problematic given research calling into question the tenability of the assumptions that are fundamental to the EM-IRT model (e.g., Rios et al., 2017). To address this gap, we examine how well the EM-IRT model performs when its assumptions are violated.

The EM-IRT Model

In the EM-IRT model, it is assumed that any response classified as noneffortful is an invalid indicator of test taker ability. This assumption is operationalized through an extension of the two-parameter logistic (2PL) model[2]:

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

In this model, SB_{ij} is a dichotomous indicator specifying whether an item response is considered effortful or noneffortful and is defined as follows:

$$SB_{ij} = \{1\ if\ RT_{ij} \geq T_i,\ 0\ otherwise\}, \quad (2)$$

where RT_{ij} is the response time threshold[3] for item *i* and test-test taker *j*, and T_i is the threshold limit for item *i* indicating the maximum response time for classification of noneffortful responses. So, if $SB_{ij} = 1, P_{ij}(\theta) = \left(\frac{1}{1+\exp\{-1.7a_i(\theta_j - b_i)\}}\right)$, where a_i is the discrimination

[2] The EM-IRT model can be extended to other IRT models, such as the Rasch, one-parameter, and three parameter logistic models.
[3] For a review of threshold procedures, the reader is referred to Wise (2017).

parameter for item i, b_i is the difficulty parameter for item i, and θ_j is the ability parameter for person j. 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.[4] 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 would imply that when a noneffortful response is provided, the probability of correctly responding to item i does not depend on the examinee's ability. Therefore, scoring under the EM-IRT model excludes noneffortful responses because they are considered to be uninformative to parameter estimation (Wise & DeMars, 2006).

The appropriate application of the EM-IRT model depends on fulfilling a number of assumptions. First, the approach to defining noneffortful responses in this model requires that the probability of a correct response from non-effortful responding is irrelevant to neither the characteristics of items nor the ability levels of examinees and that noneffortful responses can be correctly identified. Although prior literature has proposed methods for identifying noneffortful responses in survey data (e.g., Hong et al., 2020; Meade et al., 2012), the EM-IRT model has been predominately applied to computer-administered cognitive assessments comprised of multiple-choice questions, due to the availability of both keyed-responses and log file information (response time and response accuracy data).

Second, the conceptualization of noneffortful responding in the EM-IRT model relies on two important and, at times, unstated assumptions related to the characteristics of noneffortful responses. For one, the model assumes that noneffortful responding occurs randomly within test

[4] Prior research has shown that in certain contexts the 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, Ong & Strickman, 2019). For example, Attali and Bar-Hillel (2003) have shown that examinees tend to guess by choosing the middle response options. If this is the case and the correct answers were middle response options, the chance rate would be expected to be higher than 1 divided by the number of response options.

takers as the test progresses (said differently, noneffortful responding is independent from item parameters). Under this assumption, valid responses should be representative of the test's content and varied item difficulty; thus, providing an accurate estimate of how well a test taker would have performed on the entire test if giving full effort. However, if this assumption is not met, an estimate of ability would be biased based on the test content and item difficulties that induced an examinee to respond noneffortfully. An additional underlying assumption of the EM-IRT model when estimating mean ability is that noneffortful responding is unrelated to examinees' true ability (said differently, noneffortful responding is independent from person parameters). If this assumption is violated, the mean ability estimates will be either under/overestimated depending on the ability characteristics of the unmotivated examinees. The degree of under/overestimation is expected to increase as the difference between examinees' underlying ability and the item difficulties to which noneffortful responses were given increases.

Rationale for Current Study

As noted, prior research has demonstrated the superiority of the EM-IRT over traditional IRT models (e.g., 2PL model) in parameter recovery when noneffortful responding is present. However, all studies to date have assumed that noneffortful responding occurs at random within and between persons (i.e., adhering to the assumptions underlying the EM-IRT model). Yet, these assumptions may be untenable in operational testing contexts. As an example, Wise and Kingsbury (2016) demonstrated that test takers can employ a multitude of noneffortful responding patterns that are nonrandom (i.e., related to item parameters). Furthermore, there have been a number of studies that demonstrated nonnegligible ability differences between examinees who did and did not engage in noneffortful responding (e.g., Rios et al., 2017). Thus,

little is known about the performance of the EM-IRT model when noneffortful responding is related to item and person parameters.

The aims of this study were therefore twofold. First, we investigated how much bias is introduced into person and item parameter estimates using the EM-IRT model when its assumptions are violated. Second, under these violations, we compared the bias detected between the EM-IRT and 2PL models. In so doing, we looked to provide evidence on whether the EM-IRT model outperforms the 2PL model under less than ideal circumstances. We pursued these two aims using simulation and empirical studies. In the first study, data were simulated for a testing context in which ability inferences were made at the group-level (i.e., ability estimates were aggregated at the total sample level; hereon person/ability parameters refer to the mean sample ability), reflecting accountability testing programs, such as Smarter Balanced, in which low test-taking effort is a concern. In this simulation study, the following research questions were investigated:

1. How robust are EM-IRT item and person parameter estimates to violations of the underlying model assumptions that: (a) within persons, noneffortful responding occurs randomly over the course of a test; (b) between persons, noneffortful responding is unrelated to true ability?
2. Do results to the first question change dependent on the number of examinees who show low effort and the proportion of items on which they are unmotivated?
3. When the underlying assumptions of the EM-IRT model are violated, does it still outperform the 2PL model in terms of item and person parameter recovery?

To investigate the tenability of the EM-IRT model's assumptions in practice, Study 2 utilized multiple empirical datasets to examine the following research questions:

4. How prevalent is noneffortful responding? Specifically, what proportion of students noneffortfully respond on at least one item? Among those students, what is the distribution of noneffortful responses?
5. What patterns of noneffortful responding are observed for examinees over the course of a test?
6. Is noneffortful responding related to estimated ability in each sample?

Results from these studies have the potential to inform researchers and test developers about the effectiveness of the EM-IRT model in providing robust parameter estimates under model violations when compared to traditional approaches used extensively in operational settings that ignore low effort.

Study 1 - Simulation

Method

Data Generation

To evaluate the EM-IRT model under varying patterns of noneffortful responding, data were generated for a 50-item test using the unidimensional 2PL model. Simulee abilities were sampled from a normal distribution (more detail is provided in the next section) and generating item parameters were taken from an operational administration of a low-stakes college learning outcomes assessment of critical thinking and reading. The mean discrimination and difficulty parameters employed were 1.00 (SD = 0.35, min = 0.18, max = 1.78) and 0.02 (SD = 0.67, min = -1.50, max = 1.42), respectively, indicating that data were generated for a moderately difficult and adequately discriminating assessment, assuming a standard normal ability distribution. Data generation was performed in *R*, version 3.5.0 (R Development Core Team, 2018).

Simulation conditions

Five independent variables were examined: (a) sample size (500, 5,000); (b) percentage of unmotivated simulees in the total sample (10%, 30%, 50%); (c) percentage of noneffortful responses provided by each unmotivated simulee (10%, 30%, 50%, 70%); (d) the relationship between noneffortful responding and true ability (related [unmotivated mean ability = 0] and unrelated [unmotivated mean ability = -0.5 or -1]); (e) noneffortful responding pattern over the course of the test (random, difficulty-based, location-based, and decreasing effort). These five independent variables and their respective levels were fully crossed, which resulted in a 2 x 3 x 4 x 3 x 4 design for a total of 288 conditions. To minimize sampling error, each condition was replicated 200 times. Below, the rationale for each condition is discussed.

Sample Size. Sample size was included as an independent variable to investigate its impact on parameter estimation, particularly when considering the presence of noneffortful responses in the data matrix. Two sample sizes were inspected: 500 and 5,000. The former was included as it both represents a minimal sample size for obtaining stable parameter estimates for the 2PL model (Hulin et al., 1982) and is a realistic sample size for many small-scale testing programs in which low test-taking effort may be a concern (e.g., formative assessments used in schools). The latter level was incorporated into the design to represent a sample size that was ten times greater in magnitude, and thus, could be viewed as a comparatively large sample size.

Percentage of Unmotivated Simulees. The percentage of unmotivated examinees observed in operational settings has been found to range from 0-25% (e.g., Rios et al., 2014; Rios & Guo, 2020; Rios et al., 2017; Soland, 2018). To account for this variability, prior simulation studies have examined various percentages of unmotivated simulees that are well within the range observed in operational settings (10%, 25%, 30%), while also examining more extreme cases of low effort (e.g., 50%; Rios et al., 2017; Wise & DeMars, 2006). Following

previous simulation research, the levels examined in this study were: 10%, 30%, and 50%. Thus, in this simulation study, unmotivated simulees were operationalized based on noneffortfully responding on a minimum of 10% of items or more.[5]

Percentage of Noneffortful Responses. In operational analyses, the percentage of noneffortful responses that have been observed for unmotivated examinees has ranged from 1% to 100% (e.g., DeMars & Wise, 2010). Previous simulation analyses examining noneffortful responding have manipulated percentages to range from 2.3% to 50% within unmotivated simulees (DeMars & Wise, 2010; Rios et al., 2017; Wise & DeMars, 2006). The percentages of noneffortful responses (constrained equal across all unmotivated simulees to conform to prior simulation designs) investigated in this study were 10%, 30%, 50%, and 70% of the items seen by the simulee. These percentages reflect those of previous simulations, but also add a condition with extreme noneffortful responding (70%). Fully crossing this condition with the percentage of unmotivated simulees led to the percentage of noneffortful responses in the complete data matrix across all simulees to range from 1% to 35%.

The Relationship between Noneffortful Responding and True Ability. Studies have shown that individuals with low prior achievement and academic attainment are inclined to respond noneffortfully more often and on more items than their higher-achieving counterparts, suggesting that there may be a relationship between noneffortful responding and true ability (Kuhfeld & Soland, 2020; Rios et al., 2017; Soland et al., 2019). One theoretical argument for this finding is that low ability test takers engage in noneffortful responding to minimize negative self-perceptions by attributing their failure to not trying (Jagacinski & Nicholls, 1990; Thompson et al., 1995). However, some researchers maintain that such a relationship is minimal enough to

[5] Prior research has classified unmotivated examinees based on the criterion that they provide noneffortful responses on 10% of items or more (e.g., Wise & Kong, 2005).

be trivial when making assumptions about why students rapidly guess (see Wise, 2015, for a discussion).

Given this debate, we included simulation conditions in which noneffortful responding was, and was not, related to true ability. This condition was created by separately sampling ability parameters for motivated and unmotivated simulees. For the former group, ability parameters were randomly sampled from a standard normal distribution across all conditions. Similarly, in the condition where ability and effort were unrelated, unmotivated simulees' ability parameters were randomly sampled from a standard normal distribution. In the condition assuming a relationship between ability and effort, disengaged 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 because Rios et al. (2017) found an average performance difference on a prior ability measure between motivated and unmotivated simulees equal to approximately 0.50 standard deviations (favoring motivated examinees). The latter mean served as an extreme ability difference between motivated and unmotivated simulees. Drawing from a different normal distribution for unmotivated simulees in the related condition meant that most noneffort occurred for low ability simulees.

Noneffortful Responding Pattern. Wise and Kingsbury (2016) have suggested that there are four patterns of noneffortful responding for keyed multiple-choice items, all of which were examined in our own study: (a) random (examinees noneffortfully respond randomly across the test); (b) difficulty-based (noneffortful responses occur only when examinees perceive the item to be too difficult); (c) changing state (at some point on a test, examinees become unmotivated and then only provide noneffortful responses thereafter); (d) decreasing effort (examinees generally become less motivated as the test progresses). Across all four patterns,

noneffortful responses were given a probability of being correct equal to chance, or $P_i(\theta) = .25$ (assuming each item possessed four response options). Condition (a) was produced by randomly replacing known probabilities of correctly responding to an item obtained from the 2PL model for simulees identified as being unmotivated with the chance rate (Rios et al., 2017; Wise & DeMars, 2006).

To simulate difficulty-based noneffortful responding (option [b] above), known probabilities of successfully responding to an item for each unmotivated simulee were rank ordered (ties were randomly ordered). Based on the specified proportion of noneffortful item responses, the items with the lowest probability of success were replaced with the chance rate, meaning that noneffort occurred for the most difficult items.

Turning to (c), known probabilities of success for unmotivated simulees for the final *x* percentage of the items on the test were replaced with the chance rate depending on the level of noneffortful responding stipulated. This approach was taken to mimic an examinee switching from a motivated to unmotivated state. For example, an examinee who was noneffortful on 10% of the items would switch into a noneffortful state with 10% of the items on the test remaining, at which point probabilities of a correct response were replaced with the chance rate.

The final noneffortful responding pattern, (d), was simulated to reflect the occurrence of test takers engaging in less effortful responding as the test progresses, which has been demonstrated to occur across a number of operational testing contexts (e.g., Pastor et al., 2019; Penk & Richter, 2017; Wise & Kingsbury, 2016). Such a phenomenon might occur if, say, students become cognitively fatigued as the test progresses, but do not simply quit trying altogether at a point in the test as simulated in (c). To reflect this pattern of behavior, three steps were taken. First, the 50 items on the assessment were split into five bins of 10 items each (i.e.,

items 1-10, 11-20, etc.). Second, the number of noneffortful responses within each bin was specified. These numbers, which progressively increased across bins of items for most conditions, were driven by the overall percentage of within-simulee noneffortful responding (described above). As an example, when unmotivated simulees were specified to noneffortfully respond on 50% of items, the number of invalid responses in each of the five bins was 3, 4, 5, 6, and 7, respectively.[6] In this example, simulees were noneffortful on seven of the ten items in bin 5. Third, after identifying the number of low-effort responses within each item bin, items were randomly selected (in accordance with the number of disengaged responses in that bin) and invalid responses were then simulated. In the example, seven out of the ten items in bin five were randomly selected to be noneffortful responses.

Analyses

Item and ability parameters were estimated based on two unidimensional IRT models using the *mirt R* package (Chalmers, 2012): (a) a 2PL model that ignores noneffortful responses and (b) the EM-IRT model, which treats all noneffortful responses as missing. For the latter model, we assumed that noneffortful responses were known, which is in accordance with prior simulation studies (e.g., Rios et al., 2017; more detail on this choice is provided in the limitations section). In the *mirt* package, data were fit using a maximum likelihood 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-

[6] The number of noneffortful responses in each of the five bins for the 10% condition was 0, 0, 0, 2, and 3, while for the 30% condition it was 1, 2, 3, 4, and 5. Finally, for the 70% condition, the distribution of noneffortful responses was 5, 6, 7, 8, and 9 across the five item bins. We acknowledge that there was a multitude of ways to disperse noneffortful responses across the test, however, our decision was meant to reflect a progressive decrease in an examinee's test-taking effort.

Shannor 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 as a prior with 21 θ values and 61 quadrature points ranging from -4 to 4. EAP was chosen for ability estimation as it has been shown to be one of the most robust IRT estimators to atypical response behaviors (Kim & Moses, 2016). Upon estimating the model, item and ability parameter estimates were compared with known parameters based on standardized bias, which was calculated for each replication as:

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

where $\hat{y}$ is the estimated parameter, y is the known parameter, and n is the number of observations. Thus, bias values were standardized based on the standard deviations of their sampling distributions, which allowed for comparisons across parameters with different units. In addition, the averaged root mean square error (RMSE) was calculated for each replication to provide a standard deviation of the residuals:

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

In terms of ability parameter recovery, both bias and RMSE were calculated for the total sample as well as low (simulees with true ability below the 25th percentile), middle (simulees with true ability between the 25th and 75th percentiles), and high (simulees with true ability above the 75th percentile) ability simulees separately. We report on results for all and low ability simulees, with the latter being a focus based on prior research suggesting that unmotivated examinees are inclined to be below average students (e.g., Debeer et al., 2014; Rios et al., 2017). Furthermore, as the trends between bias and RMSE tended to be very similar, only the former is reported in

the results section; however, for interested readers, the latter index can be obtained via online supplementary information.

To assist in making sense of the results for the large number of conditions, the effects of study factors on bias were estimated via a linear regression model. In this model, bias served as the dependent variable, while the five factors investigated (sample size, percentage of unmotivated simulees, percentage of noneffortful responses, relationship between noneffortful responding and ability, and responding pattern) were included as independent variables. Each main effect was treated as a categorical variable, with a sample size of 5,000, 10% unmotivated simulees, 10% noneffortful responses, no relationship between noneffortful responding and true ability, and random noneffortful responding serving as the reference groups. 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. To control for familywise error rate, the Benjamini-Hochburg procedure was employed to test for statistical significance based on a false discovery rate of 10% (for details of this procedure, readers are referred to Benjamini & Hochberg, 1995). Variance-explained was evaluated based on the multiple R^2 value, and the analysis was conducted using the *lm* function in *R*.

Results

Table 1 presents model results for the regression of study factors on item and ability parameter estimation recovery. Across conditions, results demonstrated that neither sample size nor noneffortful responding pattern had a large influence on the outcomes under investigation. As an example, across dependent variables, the absolute difference between sample size conditions was less than 0.05 *SD*s, while the same difference was noted for nearly every comparison between responding patterns. Therefore, we focus on the regression results for the

other main and interaction effects and present descriptive results aggregated by sample size and pattern conditions separately by item and person parameter recovery below.

Item Parameter Recovery

A Parameter Estimates

The main effects model accounted for 60% of variance in *a* parameter estimate bias. All two-way interactions accounted for less than 1% of variance, except for the interactions between IRT model and percentages of unmotivated simulees and noneffortful responses. Including these interactions into the final model led to an additional 21% of variance explained (the combined main and two-way interactions accounted for 81% of variance). As our analysis excludes discussion of sample size and responding pattern due to their limited impact, we focus on the main effect for the relationship between noneffortful responding and true ability as well as the interaction effects. In terms of the former, *a* parameter bias was found to be greater for conditions in which the unmotivated simulees possessed a mean ability that was lower than motivated simulees. Specifically, compared to conditions with equal mean ability between motivation groups, the average bias for mean ability differences of 0.5 and 1 *SDs* was greater by 0.07 and 0.20 *SDs*, respectively. As the main effects of the percentages of unmotivated simulees and noneffortful responses were found to be dependent on their association with the IRT models, we turn to the interaction effects. Specifically, in comparing models, the regression results indicated that, across all conditions, the average bias in *a* parameter estimates was greater for the EM-IRT model by as much as 0.44 and 0.57 *SD*s when the percentage of noneffortful responses and simulees was equal to 70% and 50% (both compared to a reference of 10%), respectively (Table 1). Though the regression results suggested increased bias for the EM-IRT model, the descriptive results illustrated in Figure 1 provide a more nuanced understanding of the results.

Figure 1 shows a plot for the *a* parameter estimates, with standardized bias on the vertical axis and different combinations of the percentages of unmotivated simulees and noneffortful responses on the horizontal. Whereas the top panel is for the condition in which ability and effort were uncorrelated, the bottom two panels are for conditions in which they were related. As the figure shows, while 2PL estimates were biased when effort and ability were unrelated (estimates showed bias between -0.13 and -0.54 *SD*s when 30% of simulees disengaged), the EM-IRT estimates had virtually no bias across all conditions in which the mean ability was equal between motivated and unmotivated simulees. Similarly, in conditions where the unmotivated mean was -0.5 *SDs*, the EM-IRT model still outperformed the 2PL model, but EM-IRT estimates also displayed overestimation of the *a* parameter. However, the degree of bias for the EM-IRT estimates across all conditions was less than 0.10 *SDs* (Figure 1).

Under the extreme condition of a 1 *SD* mean ability difference between motivation groups, mixed findings were observed. For instance, the EM-IRT model outperformed the 2PL model under large degrees of noneffortful responding in the data matrix (e.g., when the percentage of noneffortful responses ranged from 15% to 35%). Though, under conditions with less noneffortful responses, the degree of overestimation detected for the EM-IRT model surpassed the degree of underestimation for the 2PL model. One reason for this finding is that, for the EM-IRT model, bias in the *a* parameter was found to decrease as the percentage of noneffortful responses increased. Although counterintuitive, this finding reflects an artifact of our sampling procedure. Specifically, as the unmotivated simulees were predominately of lower ability, noneffortful responding may have been beneficial in some cases, as it increased the probability of a correct item response, particularly at lower rates of noneffortful responding. However, as the percentage of noneffortful responses increased, across all items, the average

probability of success for these responses approached chance, and thus, more accurately reflected unmotivated simulees' true probability and provided less biased *a* parameter estimates.

B Parameter Estimates

Similar to the *a* parameter estimates, the inclusion of the two-way interaction effects between IRT models and percentages of unmotivated simulees and noneffortful responses along with the main effects accounted for 81% of variance in *b* parameter estimate bias (all other two-way interaction effects were not included due to their accounting for less than 1% of variance). Focusing on the main effect for the relationship between noneffortful responding and true ability, significant differences were noted between levels. Specifically, across conditions, the average amount of bias was greater for conditions in which the unmotivated simulees were of lower ability than their motivated counterparts by 0.18 and 0.35 *SD*s for ability differences of 0.5 and 1 *SD*s. Turning to the interaction effects, when compared to the 2PL model, the EM-IRT model was found to be associated with a decrease in *b* parameter bias by an average of 0.31 and 0.38 *SD*s in conditions in which the percentages of unmotivated simulees and noneffortful responses were equal to 50% and 70% (both compared to a reference of 10%), respectively. Descriptive results shown in Figure 2 further indicate that the EM-IRT model outperformed the 2PL model across all conditions. In general, the former model showed little bias when ability and effort were unrelated.

However, when ability and effort were related (unmotivated mean abilities of -0.5 and -1 *SDs*), the EM-IRT estimates showed bias that increased as the proportion of low-effort simulees increased. For instance, when the mean ability difference between motivation groups was 0.5 *SDs*, the EM-IRT model overestimated the *b* parameter by nearly 0.20 and 0.33 *SDs* for unmotivated simulee percentages of 30% and 50%, respectively. Meanwhile, for the 2PL model,

bias for these conditions was equal to 0.41 and 0.72 *SDs*. As noted, bias in *b* parameter estimates was found to increase as the disparity in mean ability between motivation groups grew. As an example, for the EM-IRT model, when ability disparities between motivation groups was 1 *SD*, the average bias for conditions with 30% unmotivated simulees was 0.40 compared to 0.20 *SDs* for a mean ability difference of 0.5 *SDs*. In the same condition, bias was 0.58 compared to 0.41 *SDs* for the 2PL model. Overall, these results suggested that the EM-IRT model led to reduced overestimation of difficulty parameters; however, both models were susceptible to increased bias as the percentage of unmotivated simulees and the disparity in mean ability between motivation groups increased.

Ability Parameter Recovery

All Simulees

The main effects model accounted for 61% of variance in ability parameter bias. Only a single two-way interaction was found to account for variance in the dependent variable beyond 1%, which was the interaction between the percentage of unmotivated simulees and the relationship between noneffortful responding and ability. Including this interaction into the final model with the main effects led to an R^2 equal to .93, indicating that the interaction effect accounted for an additional 32% of variance. Examining the main effect for the percentage of noneffortful responses suggested that there was a negligible increase in bias across levels. For instance, across conditions, the average difference in bias between 10% and 70% noneffortful response levels was <0.01 *SD*s. Similarly, across all conditions, model results indicated that the difference in mean ability parameter bias between the 2PL and EM-IRT models was near zero (Table 1). This is further corroborated in the boxplots found in Figure 3.

Two likely reasons for the negligible difference between IRT models was that bias was both robust to high rates of noneffortful responses and largely influenced by the relationship between noneffortful responding and ability. The former factor suggests that the 2PL model, which does not account for noneffortful responding, can perform well when unmotivated simulees disengage on a large percentage of items, while the latter factor clearly violates one of the major assumptions underlying the EM-IRT model. As shown in Figure 3, both models performed well when motivated and unmotivated simulees possessed equal mean abilities; however, as the percentage of unmotivated simulees in the sample who possessed below-average ability grew, the bias was found to increase across both models (Table 1; Figure 3). As an example, across IRT models and under conditions where the ability difference between motivation groups was 0.5 *SDs*, the average degree of bias increased by 0.10 and 0.21 *SDs* when 30% and 50% of simulees were unmotivated compared to no difference in mean ability between motivation groups (Table 1). For these same percentages of unmotivated simulees, the degree of bias grew to 0.21 and 0.43 *SDs* when increasing the group mean ability difference to 1 *SD*. In short, when ability and effort were related, both models performed equally poorly in estimating mean ability for all simulees, particularly as the percentage of unmotivated simulees increased.

Low Ability Simulees

Focusing on disaggregating bias for low ability simulees (i.e., simulees with true ability below the 25th percentile) produced a different finding. Specifically, when regressing study factors on ability bias using the same model as for all simulees, significant differences between models was observed, with the EM-IRT model possessing an average bias across all conditions that was 0.20 *SDs* lower than the 2PL model (Table 1). As can be seen in Figure 4, the EM-IRT model generally overestimated ability to a lesser degree across all conditions. However, test

users should be aware that ability estimates for this simulee group can be greatly influenced by the association between noneffortful responding and true ability as well as the percentages of unmotivated simulees and noneffortful responses (Table 1). For instance, contrasted to conditions in which motivation groups were matched on their mean theta, ability was overestimated by an average of 0.15 and 0.36 *SD*s more when unmotivated simulees possessed a mean ability that was 0.5 and 1 *SD* lower than their motivated counterparts. Furthermore, compared to conditions in which 10% of the sample was comprised of unmotivated simulees, bias increased by 0.08 and 0.15 *SD*s as unmotivated simulees made up 30% and 50% of the sample. Finally, low ability bias was found to significantly increase as the percentage of noneffortful responses increased. Specifically, bias grew by 0.08, 0.18, and 0.32 *SD*s for conditions with 30%, 50%, and 70% noneffortful responses (compared to 10%). Overall, these findings suggest that the EM-IRT model generally outperformed the 2PL model in estimating ability for simulees that fall below the 25th percentile; however, across both models, bias was amplified as the difference in mean ability between motivation groups and the percentages of unmotivated simulees and noneffortful responses increased. As can be seen in Figure 4, the degree of bias in many contexts is so great that it would be unadvisable to report disaggregated ability estimates for this group.

Study 2 - Empirical

Simulation results suggest that the EM-IRT model provides robust estimates of item and ability parameters under all conditions, except when noneffortful responding was related to ability and the percentage of unmotivated simulees in the sample was ≥ 30%. Thus, in the empirical study, we provide evidence on the percentage of unmotivated examinees and the association between estimated ability and noneffort for two operational tests.

Methods

Sample and Measures

The two operational tests we used were the Measures of Academic Progress (MAP) Growth test of reading and the Office for Economic Cooperation and Development (OECD) Test for Schools, which are both described below.

MAP Growth

The MAP Growth assessments are low-stakes item-level computer-adaptive tests typically administered three times a year in the fall, winter, and spring. MAP Growth tests begin with a question appropriate for the student's grade level, and then adapt throughout the test in response to the student's performance. Test scores are reported on the Rasch Unit (RIT) scale, which is $200 + 10 \times \theta$ (θ refers to the logit scale units of the Rasch item response theory model). The sample consisted of 854,437 students in 3rd through 8th grade from the United States who took the test as many as three times (fall, winter, and spring) in the 2017-18 school year, resulting in over two million test records.

OECD Test for Schools

The OECD Test for Schools is an assessment used by schools to support research, benchmarking, and academic improvement. The test measures the knowledge of 15-year-old students in reading and mathematics on scales comparable to the main PISA scales. All item parameters and scoring methods were drawn directly from PISA. The OECD Test for Schools is fixed-form, which means students often receive items that are difficult relative to their estimated achievement. Our sample consisted of approximately 4,400 10th grade students from the United States who took the test in the 2016-17 school year. For this assessment, research questions were examined separately for math and reading.

Analytic Approach

The methodological approach generally consisted of descriptive statistics. Analyses were conducted using all test events for a student (e.g., including both scores from a student who tested in fall and spring), as well as limiting results only to the spring administration. For Question 1, the proportion/percentage of students who produced noneffortful responses on one or more items is reported by test along with the proportion/percentage of items on which those students were noneffortful. Across tests, noneffortful responses were identified using the response time threshold procedure proposed by Wise and Ma (2012). For an item of interest, this procedure classifies any response provided in less than 10% of the average response time as a noneffortful response with a maximum threshold of 10 seconds.[7]

For question 2, the proportion of noneffortful responses was plotted against the student's estimated ability. One problem with such plots is that noneffort likely leads to an understatement of estimated achievement (Rios et al., 2017). As a result, estimated ability is likely to be more correlated with noneffortful responding than with true ability. While there is no perfect solution to this confounding of estimated ability with noneffortful responding, we take an approach used by Kuhfeld and Soland (2020), who examined the association between rates of noneffortful responding and EM-IRT estimates of ability. In so doing, one can investigate the relationship

[7] Perhaps the most difficult technical challenge associated with identifying noneffortful is setting a response time threshold separating effortful and noneffortful responses (Guo et al., 2016; Wise, 2015; Wise & Kong, 2005). In plain terms, how fast does a student need to respond in order to conclude that the content of the item was not fully understood? Approaches to setting thresholds include visually inspecting response time distributions (Wise & Kong, 2005), comparing response times for a single item response to the overall distribution of response times for that item (Wise & Ma, 2012), and setting the threshold at the response time below which students get the item right at a rate no better than chance (Guo et al., 2016). We compared a subset of items deemed rapid under the Wise and Ma (2012) method to the one developed by Guo et al. (2016), but found very high overlap, likely because the frequency with which students answer an item correctly was used as a criterion to validate the thresholds set using the Wise and Ma (2012) approach. These results suggest that the inferences made in the empirical study are largely robust to the response time threshold applied to classify noneffortful responding.

between noneffortful responding and an estimate of ability that accounts for noneffortful responding (if imperfectly, as our simulation study shows).

Results

Results are presented for all test events (not just spring test events) because the conclusions drawn were not affected by including students in the sample more than once.

What is the Percentage of Unmotivated Examinees for Each Test?

Table 2 presents counts of examinees by proportion of responses that were noneffortful. As the table indicates, rates of examinees who were deemed unmotivated on one or more items were about 9% for OECD math, 11% for OECD reading, and 20% for MAP Growth reading. Most of those examinees were unmotivated on relatively few items. For example, across all three tests, about half of the students who rapidly guessed at least once did so on 5% or fewer of the items. Further, while some students did show low effort on half or more of the items, the percentages of students who did so was only 3% on OECD math, 5% on OECD reading, and 1% on MAP Growth reading.

Is There a Relationship between Estimated Ability and Noneffortful Responding?

Figure 5 plots the proportion of item responses that were noneffortful responses on the vertical axis and estimates of ability on the horizontal axis for OECD and MAP Growth reading tests. Note that the MAP Growth results are presented in RITs and the OECD results in logits. On both tests, there is a steady decline in items that were responded to rapidly as ability increases. Note that the proportion of items that were responded to rapidly reaches about .20 for low ability examinees on the OECD test, but only .08 on MAP Growth. These differences could be due to the fact that MAP Growth is computer adaptive, which means students are less likely to

see items that are very difficult relative to their ability. Results were similar when EM-IRT scores were used on the horizontal axis.

Discussion

The EM-IRT model has been shown to better recover item and person parameters relative to the 2PL model when its assumptions are met. As a result, the EM-IRT model is used often in research to address noneffortful responding. Yet, there is a dearth of evidence regarding its performance when its underlying assumptions are unmet, particularly when compared to extensively used models, such as the 2PL model. The objective of this study was to investigate the item and ability parameter estimation accuracy of the EM-IRT model under violations of its underlying assumptions. Overall, in regard to item parameter recovery, results demonstrated that this model outperformed the 2PL model across nearly all conditions and outcomes. However, though still an improvement over the 2PL model, the EM-IRT model produced bias in item parameter estimates when violating the assumption that noneffortful responding is unrelated to ability. This result was especially pronounced when the percentage of unmotivated simulees in the sample was greater than or equal to 30%. Furthermore, when estimating the *a* parameter under conditions in which the mean ability of the unmotivated simulees was lower by 1 *SD* when compared to motivated simulees, the EM-IRT model actually performed worse for certain percentages of unmotivated simulees and noneffortful responses.

In regard to estimation accuracy of ability parameters, the 2PL and EM-IRT models were found to perform nearly identically. Across all conditions, approximately no bias in ability estimates was observed when noneffortful responding was unrelated to ability. However, when unmotivated simulees were predominantly of low ability, both models tended to overestimate ability parameters, with greater bias observed as the percentage of unmotivated simulees in the

sample increased. In general, when this percentage was ≤ 30%, bias was ≤ 0.16 *SDs*. This finding of bias on ability estimates has been supported by a number of applied analyses (e.g., Kuhfeld & Soland, 2020; Rios et al., 2017; Soland, 2018). Although the two models performed equally well when estimating ability for the total sample, the EM-IRT model was found to produce less bias for low ability simulees (simulees with ability below the 25th percentile), though this advantage largely disappeared for other ability groupings (e.g., middle ability). However, it should be noted that these findings are in the context of simultaneous estimation of both item and person parameters. Thus, the degree of bias that we observed in ability estimates is likely greater than employing fixed item parameter ability estimation in which the item parameters are first estimated based on a filtered sample (i.e., all noneffortful responses are treating as missing). Examining the utility of employing such an approach to improve ability parameter estimation accuracy in the EM-IRT model is a direction for future research.

To examine if the conditions associated with increased bias in item and person parameters are present in practice, two operational testing programs were examined. Our applied analyses demonstrated that examinees with lower ability were found to noneffortfully respond at increased rates when compared to their higher ability counterparts. However, the percentages of unmotivated examinees (i.e., examinees engaging in one or more noneffortful responses) in our samples that noneffortfully responded on more than 5% of items did not exceed 10%. Thus, our results suggest that, although there is the potential for estimation bias of item parameters when employing the EM-IRT model under violations to its underlying assumption that noneffortful responding and ability are unrelated, the degree of bias in practice may be small.

Limitations and Future Research Directions

A limitation of the simulation analysis was our treatment of noneffortful responses as known. Although this was an approach taken in prior simulation studies (e.g., Rios et al., 2017), the accuracy of the EM-IRT model parameter estimates assumes that noneffortful responses can be correctly classified. Consequently, our simulation analysis ignored potential violations to this assumption, and the findings in this paper should be interpreted as the best-case performance for the EM-IRT model. It is unknown how misclassifying noneffortful responses will impact parameter estimates for this model. This is a clear area for future research.

An additional set of limitations was related to our manipulation of two factors in the simulation study. First, we constrained all unmotivated simulees to noneffortfully respond on the same percentage of items. Such a scenario is unlikely given that prior research has demonstrated a large variance in the number of noneffortful responses that examinees provide in operational settings (Wise & Kingsbury, 2016). Although the intended aim of this approach was to control for this factor in an effort to isolate its underlying effect on the study outcomes, it may have led to simulation contexts in which the percentage of noneffortful responses observed on each item was higher than what is seen in practice. As a consequence, when examining item parameter recovery, the results may reflect a greater degree of bias than what may be present in operational settings. It is recommended that future research examining this factor and others in this study include sampling methods that create greater degrees of randomness in simulees' noneffortful responding patterns, which may better reflect reality (e.g., see Wang et al., 2018).

Finally, our simulation study only compared two IRT models that were selected based on their computational simplicity and popularity in the literature. However, as noted earlier, there have been recent applications of mixture modeling to simultaneously account for noneffortful responding while also estimating item and ability parameters (e.g., Lu et al., 2020; Wang & Xu,

2015; Wang et al., 2018). As a result, there is a need for a study comparing the two-stage approach (i.e., first identifying noneffortful responding, and then accounting for those responses prior to parameter estimation) to the mixture-model approach in terms of parameter recovery and computation time. As our findings showed that there is a clear need to improve estimation accuracy of ability parameters when unmotivated simulees are predominantly of low ability, this comparative study should also investigate if the mixture-modeling approach provides less bias in parameter recovery compared to the EM-IRT when motivation groups differ in their underlying mean ability.

Turning to the empirical analysis, as with the use of any applied data, our study is limited in the degree of generalizability that is permissible based on our evaluation of noneffortful responding in two testing contexts. Although we included a large-scale sample that consisted of over two million testing records for the MAP Growth assessment, limiting sampling error, both testing programs in this study consisted of students from the United States in grades K-12. Thus, the degree of noneffortful responding observed does not generalize to other testing contexts and populations. In particular, MAP Growth is an adaptive test, which tends to be more robust to noneffortful testing issues (Wise, 2014). Due to the limitations associated with the empirical samples and tests used, we would recommend that future research continue to examine characteristics of noneffortful responding patterns and rates, as well as characteristics of unmotivated examinees across different testing contexts.

Conclusion

This study provides two main takeaways. First, failing to handle noneffortful responses may affect item calibration by making items appear to be more difficult and less discriminating than they are. These inaccuracies can be largely mitigated by applying the EM-IRT model, which

has been shown to provide robust item parameter estimates under violations to its underlying assumptions for realistic conditions in practice. Therefore, it is recommended that when evaluating item properties during the pilot test phase, practitioners should apply the EM-IRT model to mitigate incorrect inferences about item quality, particularly as these data are often collected from examinees in low-stakes contexts. This strategy may be of particular benefit during new test or item pool construction.

Second, our simulation analyses demonstrated that mean ability estimates are robust to large percentages of noneffortful responding and unmotivated simulees, particularly when noneffortful responding is unrelated to ability. However, we did find a tendency for both the 2PL and EM-IRT models to overestimate mean ability when noneffortful responding and ability are related. Nonetheless, bias > 0.25 *SD*s was only detected when the percentage of unmotivated simulees was 50% and mean ability differences between motivation groups was 0.5 *SD* or more, which may occur somewhat infrequently in practice. Therefore, if estimating mean ability for a total sample, employing either the 2PL or EM-IRT models will often produce comparable results, including similar degrees of bias. However, if the focus of score reporting is to disaggregate data for low ability examinees, potentially to help make remediation decisions or subgroup comparisons, practitioners may benefit from employing the EM-IRT model, particularly in conditions in which motivation groups are matched in their underlying ability and the percentages of unmotivated simulees and noneffortful responses are low.

References

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.

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

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., 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.

Hong, M., Steedle, J. T., & Cheng, Y. (2020). Methods of detecting insufficient effort responding: Comparisons and practical recommendations. *Educational and Psychological Measurement*, *80*(2), 312-345.

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.

Jagacinski, C. M., & Nicholls, J. G. (1990). Reducing effort to protect perceived ability: "They'd do it, but I wouldn't." *Journal of Educational Psychology*, *82*(1), 15-21.

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.

Kong, X. (2007). *Using response time and the effort-moderated model to investigate the effects of rapid guessing on estimation of item and person parameters* (Unpublished doctoral dissertation). James Madison University.

Kuhfeld, M., & Soland, J. (2020). Using assessment metadata to quantify the impact of test disengagement on estimates of educational effectiveness. *Journal of Research on Educational Effectiveness*, *13*(1), 147-175.

Lu, J., Wang, C., Zhang, J., & Tao, J. (2020). A mixture model for responses and response times with a higher-order ability structure to detect rapid guessing behaviour. *British Journal of Mathematical and Statistical Psychology*, *73*(2), 261-288.

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological methods*, *17*(3), 437-455.

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., & Richter, D. (2017). Change in test-taking motivation and its relationship to test performance in low-stakes assessments. *Educational Assessment, Evaluation and Accountability*, *29*(1), 55-79.

R Development Core Team (2018). R: A language and environment for statistical computing [Computer software]. R Foundation for Statistical Computing. Retrieved from 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*. Advanced online publication.

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*(1), 74-104.

Soland, J. (2018). Are achievement gap estimates biased by differential student test effort? Putting an important policy metric to the test. *Teachers College Record*, *120*(12).

Soland, J., Jensen, N., Keys, T. D., Bi, S. Z., & Wolk, E. (2019). Are test and academic disengagement related? Implications for measurement and practice. *Educational Assessment*, *24*(2), 119-134.

Wainer, H. (1993). Measurement problems. *Journal of Educational Measurement*, *30*(1), 1-21.

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. (2018). A two-stage approach to differentiating normal and aberrant behavior in computer based testing. *Psychometrika*, *83*(1), 223-254.

Wise, S. L. (2015). Effort analysis: Individual score validation of achievement test data. *Applied Measurement in Education*, *28*(3), 237-252.

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

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment*, *10*(1), 1-17.

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., & 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., & Ma, L. (2012, April). *Setting response time thresholds for a CAT item pool: The normative threshold method*. Presented at the annual meeting of the National Council on Measurement in Education, Vancouver, Canada.

Table 1

Model Results for Regressing Study Factors on Item and Ability Parameter Bias

Factor	A Parameter R^2 = .81 Estimate	B Parameter R^2 = .81 Estimate	Theta (All) R^2 = .93 Estimate	Theta (Low Ability) R^2 = .93 Estimate
Intercept	0.10***	-0.17***	-0.00***	0.25***
Sample Size	0.04***	-0.02***	-0.00***	-0.01***
30% Unmotivated Simulees	-0.20***	0.29***	0.00***	0.08***
50% Unmotivated Simulees	-0.49***	0.58***	0.00*	0.15***
NER Pattern: Difficulty	-0.04***	-0.09***	0.00	-0.04***
NER Pattern: Changing State	0.03***	-0.01***	0.00	0.01***
NER Pattern: Progressive	<0.01	0.01***	0.00	0.01***
30% NER	-0.20***	0.11***	0.00*	0.08***
50% NER	-0.37***	0.22***	0.00***	0.18***
70% NER	-0.47***	0.34***	0.00***	0.32***
Unmotivated Mean: -0.5	0.07***	0.18***	0.05***	0.15***
Unmotivated Mean: -1	0.20***	0.35***	0.10***	0.36***
30% Unmotivated Simulee x Unmotivated Mean: -0.5	----	----	0.10***	0.28***
50% Unmotivated Simulee x Unmotivated Mean: -0.5	----	----	0.21***	0.55***
30% Unmotivated Simulee x Unmotivated Mean: -1	----	----	0.21***	0.68***
50% Unmotivated Simulee x Unmotivated Mean: -1	----	----	0.43***	1.26***
Model	-0.14***	0.11***	0.00**	-0.20***
30% NER x Model	0.19***	-0.11***	----	----
50% NER x Model	0.36***	-0.24***	----	----
70% NER x Model	0.44***	-0.38***	----	----
30% Unmotivated Simulee x Model	0.25***	-0.16***	----	----
50% Unmotivated Simulee x Model	0.57***	-0.31***	----	----

Note. NER = noneffortful responding or noneffortful responses. All parameters had standard errors <.01. ***p<.001; ** p<.01; *p<.05

Table 2

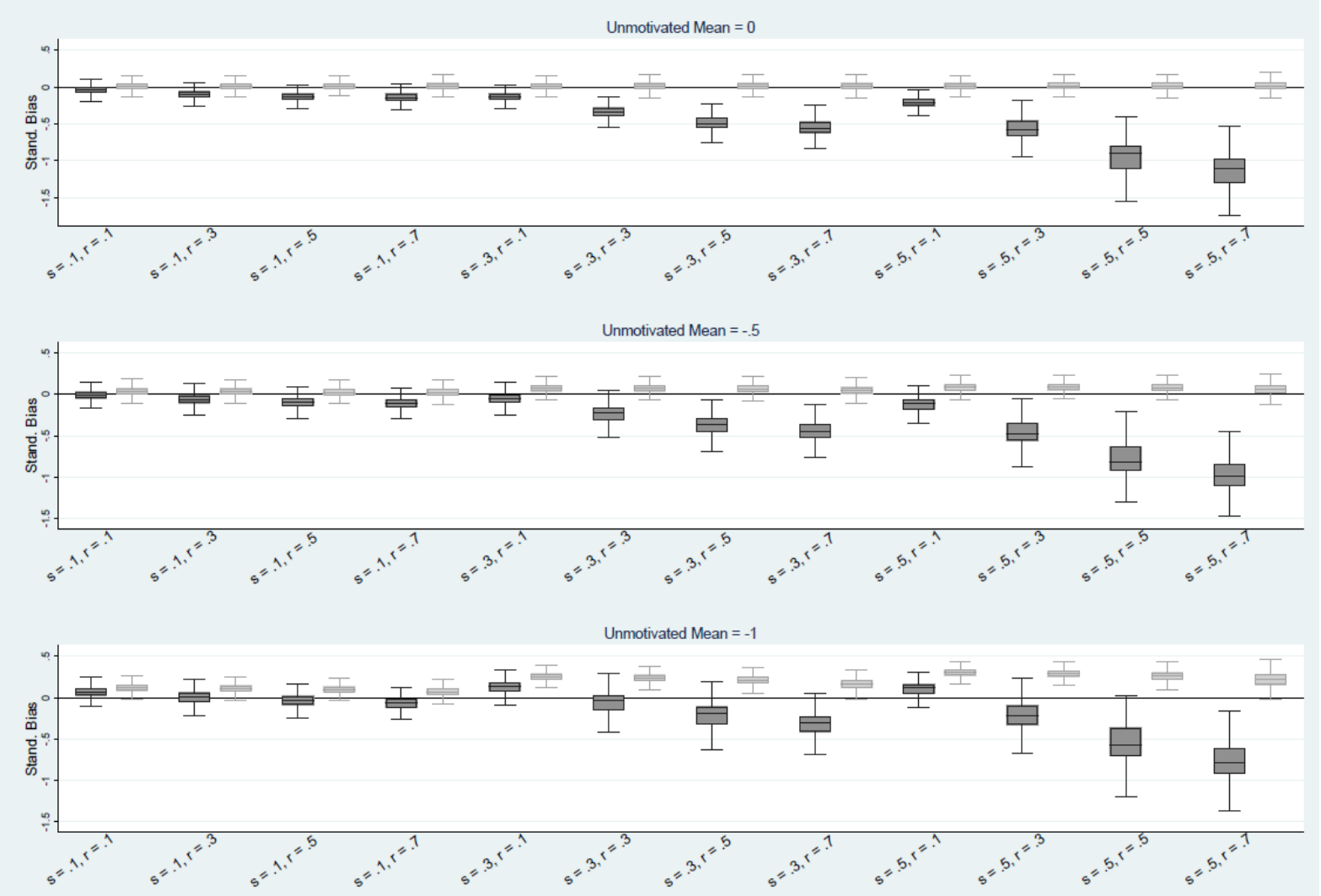
Counts of Examinees by Proportion of Responses that were Noneffortful

	OECD - Math		OECD - Reading		MAP Growth - Reading	
Prop. Rapid	Freq.	Proportion	Freq.	Proportion	Freq.	Proportion
0	4,033	.913	3,887	.886	1,621,229	.793
.05	176	.040	214	.049	226,313	.111
.10	72	.016	84	.019	76,535	.037
.15	50	.011	63	.014	44,151	.022
.20	34	.008	39	.009	24,718	.012
.25	18	.004	22	.005	14,752	.007
.30	9	.002	18	.004	12,527	.006
.35	10	.002	20	.005	7,050	.003
.40	5	.001	8	.002	5,565	.003
.45	0	.000	8	.002	4,077	.002
.50	4	.001	13	.003	2,482	.001
.55	1	.000	4	.001	1,713	.001
.60	0	.000	3	.001	984	.000
.65	2	.000	2	.000	722	.000
.70	0	.000	2	.000	419	.000
.75	0	.000	1	.000	265	.000
.80	0	.000	0	.000	101	.000
.85	0	.000	0	.000	59	.000
.90	1	.000	0	.000	26	.000
.95	1	.000	0	.000	6	.000
1	3	.001	1	.000	2	.000
	4,419	1.000	4,389		2,043,696	

Note. Test lengths varied but were typically between 20-30 items.

Figure 1

A Parameter Bias



Note. S = proportion of unmotivated simulees in sample; R = proportion of noneffortful responses. The lighter color boxplots represent the results for the EM-IRT model.

Figure 2

B Parameter Bias

Note. S = proportion of unmotivated simulees in sample; R = proportion of noneffortful responses. The lighter color boxplots represent the results for the EM-IRT model.

Figure 3

Ability Parameter Bias (All Simulees)

Note. S = proportion of unmotivated simulees in sample; R = proportion of noneffortful responses. The lighter color boxplots represent the results for the EM-IRT model.

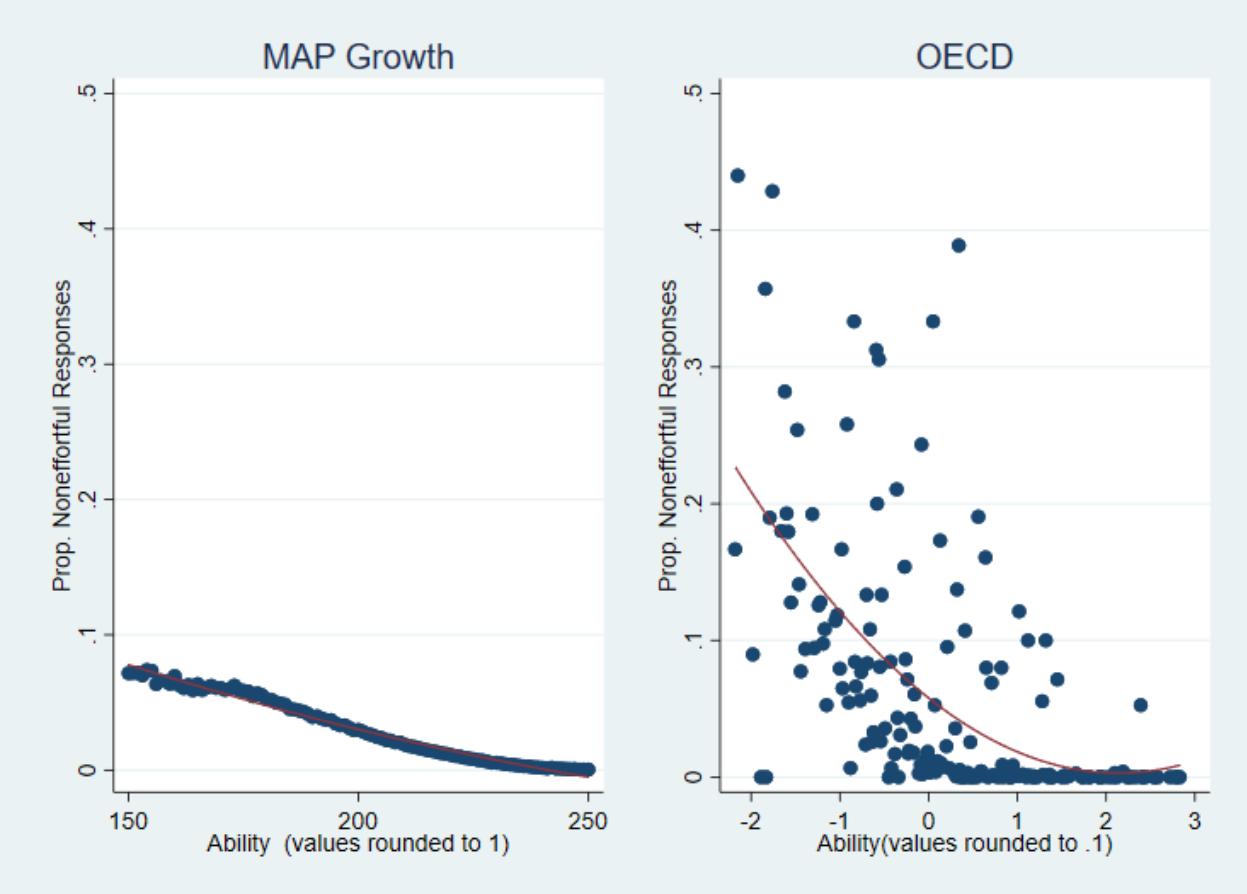
Figure 4

Ability Parameter Bias (Low Ability Simulees)

Note. Low ability simulees were classified based on possessing a true ability < 25th Percentile. S = proportion of unmotivated simulees in sample; R = proportion of noneffortful responses. The lighter color boxplots represent the results for the EM-IRT model.

Figure 5

Plots of Proportion of Noneffortful Responses versus Estimated Ability



MAP Growth Reading **OECD Reading**

Note. MAP Growth estimates fall much closer to the fitted function due to substantively larger sample sizes compared to the OECD sample.