

Is Differential Noneffortful Responding Associated With Type I Error in Measurement Invariance Testing?

Educational and Psychological
Measurement

2021, Vol. 81(5) 957–979

© The Author(s) 2021

Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/0013164421990429

journals.sagepub.com/home/epm



Joseph A. Rios¹ 

Abstract

Low test-taking effort as a validity threat is common when examinees perceive an assessment context to have minimal personal value. Prior research has shown that in such contexts, subgroups may differ in their effort, which raises two concerns when making subgroup mean comparisons. First, it is unclear how differential effort could influence evaluations of scale property equivalence. Second, if attaining full scalar invariance, the degree to which differential effort can bias subgroup mean comparisons is unknown. To address these issues, a simulation study was conducted to examine the influence of differential noneffortful responding (NER) on evaluations of measurement invariance and latent mean comparisons. Results showed that as differential rates of NER grew, increased Type I errors of measurement invariance were observed only at the metric invariance level, while no negative effects were apparent for configural or scalar invariance. When full scalar invariance was correctly attained, differential NER led to bias of mean score comparisons as large as 0.18 standard deviations with a differential NER rate of 7%. These findings suggest that test users should evaluate and document potential differential NER prior to both conducting measurement quality analyses and reporting disaggregated subgroup mean performance.

Keywords

test-taking effort, noneffortful responding, measurement invariance, subgroup comparisons, validity

¹University of Minnesota, Minneapolis, MN, USA

Corresponding Author:

Joseph A. Rios, University of Minnesota, 164 Education Sciences Building, 56 East River Road, Minneapolis, MN 55455, USA.

Email: jrrios@umn.edu

Low test-taking effort as a validity threat is common when examinees perceive an assessment context to have minimal personal value (low-stakes testing; Penk & Schipolowski, 2015). This can occur because examinees do not have (e.g., international comparative education studies, such as the Programme for International Student Assessment [PISA]) or are unaware of the individual consequences for their test performance (e.g., young children assessed for remediation). When an assessment context is perceived to be low stakes, some examinees have been found to engage in noneffortful responding (NER; i.e., providing a random guess without consideration for the item content due to low-test-taking effort), leading to random score error that is generally associated with significant underestimation of examinee ability (e.g., Rios et al., 2017; Wise & DeMars, 2005). Furthermore, NER has also been found to produce biased measurement properties, such as estimates of (a) item parameters (e.g., Rios & Soland, 2020; van Barneveld, 2007); (b) test information (e.g., van Barneveld, 2007); (c) classical test theory (CTT) score reliability (e.g., Wise & DeMars, 2009); (d) construct dimensionality (e.g., Kam & Meyer, 2015); and (e) linking coefficients (Mittelhaeuser et al., 2015). As a result of these findings, the *Standards for Educational and Psychological Testing* calls for test developers and users to document the potential role of low-test-taking effort as a source of construct-irrelevant variance prior to evaluating measurement quality and making score-based inferences (e.g., Standard 13.9; American Educational Research Association et al., 2014). The purpose of this article is to investigate the effect of NER on measurement invariance and mean score comparisons when examinee subgroups possess differential test-taking effort. In the sections that follow, prior literature documenting subgroup differences in test-taking effort, its documented effects on measurement invariance, and the rationale for the current study are discussed.

Differential NER and Measurement Invariance

A significant number of researchers have documented that test-taking effort differs across subgroups of examinees in low-stakes testing contexts. For instance, differences have been illustrated by (a) gender (DeMars et al., 2013; Schnipke, 1995; Soland, 2018; Wise & Cotton, 2009; Wise & DeMars, 2010; Wise et al., 2004); (b) age (DeMars, 2007; Goldhammer et al., 2016; Wise & DeMars, 2010); (c) ethnicity (Soland, 2018); (d) language group (Goldhammer et al., 2016; Setzer et al., 2013); (e) school-track (Penk et al., 2014); (f) educational attainment level (Goldhammer et al., 2016); and (g) nationality (Boe et al., 2002; Borghans & Schils, 2012; Debeer et al., 2014; Goldhammer et al., 2016; Rios & Guo, 2020; Zamarro et al., 2019). In these applied contexts, subgroup differences in NER have been found to be as high as 23% (Rios & Guo, 2020).¹ Thus, it is of little surprise that researchers have heeded concern that disparities in test-taking effort can lead to inaccurate inferences concerning subgroup comparisons (e.g., Soland, 2018).

However, subgroup comparisons first assume that the statistical property of measurement invariance holds for the given measure across subgroups of interest.

Measurement invariance is met when an examinee's group membership adds nothing to their observed score on measure X above and beyond our knowledge of their standing on the latent variable measured by X (Millsap, 2011). Although there are multiple levels of measurement invariance, practitioners concerned with making subgroup comparisons are most attentive to attaining full scalar invariance (e.g., Fischer & Karl, 2019). This form of invariance stipulates that the same data configurations or structures (i.e., the same number of factors and loading pattern) of the purported construct are present, the strength of the relationships between the indicators and latent construct(s) are equivalent (i.e., equal factor loadings), and the intercepts are equal across subgroups. Meeting these assumptions allows practitioners to make direct subgroup mean comparisons, as the measure of interest has been demonstrated to possess equal measurement units and the same origin values for all items across subgroups (Dimitrov, 2010).

Although establishing full scalar measurement invariance is a critical step to ensuring valid subgroup comparisons, there has been minimal research to date on the impact of differential NER in establishing this statistical property. One of the only studies to investigate the relationship between disparities in subgroup test-taking effort and invariance was conducted by DeMars and Wise (2010); however, the focus of their analysis was on item-level invariance or differential item functioning (DIF). In simulating a context in which the generating item parameters were the same, but differential NER differed between subgroups by upwards of 25%, DeMars and Wise assessed whether detectable levels of DIF were present using the Mantel-Haenszel procedure. Findings from this study illustrated that as many as 18% of items were incorrectly misclassified as possessing DIF across item characteristics. However, a closer examination of the item properties showed that misclassification rates as high as 100% were observed for very easy ($b = -2.5$) and discriminating ($a = 1-2$) items. This was due to overestimation of item difficulty for the unmotivated subgroup, as most of these simulees would have provided a correct response if full effort was given. Overall, the results of this study illustrated that differential NER can lead to inaccurate inferences concerning DIF; however, it is yet to be determined how differing effort across subgroups could affect scale-level invariance analyses (i.e., simultaneous invariance analyses for all items of a given measure).

Rationale for Current Study

Better understanding the impact of differential NER on evaluations of full-scalar invariance is of critical importance in establishing the robustness of inferences in the presence of disparate subgroup test-taking effort, which has been documented across multiple testing contexts and populations. This work is inspired by the important policy implications that low-stakes assessments, such as those used in test-based accountability systems (e.g., state mandated end-of-year assessments), can have for monitoring achievement gaps between subpopulations.

To support these efforts, the objective of this study is to investigate the influence of differential NER on evaluations of full scalar invariance and latent mean comparisons. This objective was examined via a simulation analysis that represented a context in which an assessment measuring a unidimensional construct via keyed multiple-choice items was administered to a population that was divided into two subgroups. The generating item parameters were held constant across both subgroups reflecting full scalar invariance. However, these subgroups differed in their test-taking effort, with one subgroup far less motivated (hereon referred to as the focal subgroup) than the other (hereon referred to as the reference subgroup). Furthermore, as is common in practice, the presence of NERs was ignored (see Wise, 2017). Based on this context, the following research questions were addressed:

1. How does differential NER affect Type I error (i.e., incorrectly rejecting the true null hypothesis of full scalar invariance) rates of measurement invariance analyses?
2. When attaining full scalar invariance in the presence of differential NER, what is the degree of bias on latent subgroup mean comparisons?

Findings have the potential to inform testing programs about the importance of considering NER prior to conducting measurement quality evaluations and reporting disaggregated subgroup mean performance.

Method

Data Generation

Data were generated for a unidimensional test consisting of n (either 30 or 60 items) multiple-choice items that were administered to two subgroups (focal and reference) comprising a total of 5,000 simulees. A total sample size of 5,000 was chosen as it is expected to provide both stable parameter estimates and adequate power for model fit statistics (e.g., Kim, 2005; Wolf et al., 2013). Effortful item response probabilities were created in both subgroups based on the two-parameter logistic (2PL) model. This was done by first sampling item and person generating parameters. The former were taken from an operational administration of the NAEP reading assessment (for a full list of item parameters, see Appendices A and B of the online Supplemental Material). 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 2PL model to obtain effortful item response probabilities.

For unmotivated simulees, the next step consisted of replacing effortful probabilities with chance probabilities (assuming each item possessed four response options) to reflect progressive NER (i.e., decreasing examinee effort as the test proceeds), which has been observed in operational testing contexts (e.g., Wise & Kingsbury, 2016). This was done via a three-step process. First, the total test length was split into

five bins (for the 30-item condition, each bin consisted of six items, while for the 60-item condition, 12 items comprised each bin). Second, the number of NERs in each bin was specified. These numbers were determined based on the condition's specified within-simulee responding rate. As an example, when this rate was 50%, the number of NERs in each of the five bins for the 60-item condition was 0, 3, 6, 9, and 12.² Third, once this distribution was determined, NERs were randomly selected in each bin and the true item probability was replaced with the chance rate. Both effortful and noneffortful (i.e., chance) probabilities obtained were then compared with a random number sampled from a uniform distribution ranging from 0 to 1. For each simulee, if the random number was less than the probability, the item response was treated as correct. All data generation was conducted in *R*, version 3.5.0 (R Development Core Team, 2018).

Conditions

Below is a description of how NER was manipulated across five factors: (a) test length, (b) subgroup sample sizes, (c) group impact, (d) relationship between NER and true ability (NER–ability relationship), and (e) differential subgroup NER rate (hereon referred to as NER rate). These five variables were fully crossed producing 96 total conditions, with each condition replicated 100 times.

Test Length. Given that the number of items loading onto a latent factor can influence goodness of fit indices (Cheung & Rensvold, 2002), total test length was manipulated in the current study across two levels: 30 and 60 items. These two levels were chosen as they reflect the range of common test lengths of low-stakes assessments in which NER has been shown to be a concern (e.g., DeMars, 2007; Smith et al., 2013). For the 30-item condition, the mean item difficulty and discrimination were 1.07 ($SD = 0.39$; minimum = 0.4, maximum = 1.74) and 0.17 ($SD = 0.89$; minimum = -2.14 , maximum = 1.55), respectively, while the averages were nearly identical for the 60-item condition (item discrimination: $M = 1.12$; $SD = 0.41$; minimum = 0.4, maximum = 1.91; item difficulty: $M = 0.14$; $SD = 1.10$; minimum = -2.14 , maximum = 2.17) based on item parameters obtained from a NAEP assessment.

Subgroup Sample Sizes. In operational settings, there may be contexts in which there is interest in making comparative inferences between subgroups that differ in sample size (e.g., English learners vs. native English speakers). When there is such an imbalance, researchers have shown that factorial invariance tests can be affected (Yoon & Lai, 2018). To examine this issue under the current study context, subgroup sample sizes were manipulated. Specifically, the first level included equal sample sizes across subgroups (each consisted of 2,500 simulees), while the second reflected a scenario in which the focal subgroup ($n = 3,350$) outnumbered the reference ($n = 1,650$) by a 2:1 ratio. Across all conditions, the total sample size was constrained to 5,000 (this sample size provided stable parameter estimation).

Group Impact. Differences in subgroup latent mean ability (group impact) were manipulated for motivated simulees in both the reference and focal subgroups, as prior literature has suggested that group impact can increase measurement invariance Type I errors (Stark et al., 2006). This was done for two scenarios. In the first, referred to as the no group impact condition, the true latent mean ability was constrained equal for motivated simulees in both the focal and reference subgroups by sampling both subgroups' generating ability parameters from a standard normal distribution. In the second scenario, which is referred to as the group impact condition, the latent mean ability between motivated simulees from the two subgroups differed by 0.5 *SDs* (reference: $N[0, 1]$; focal: $N[-0.5, 1]$). This condition assumed that the focal subgroup was on average of lower ability than the reference subgroup, which is an assumption that has been examined in numerous studies (e.g., DeMars, 2010).

NER–Ability Relationship. There is some debate as to whether NER is related to examinees true underlying ability or whether such a relationship has a nonnegligible impact on ability parameter estimation accuracy (for a discussion, see Wise, 2015). To address this debate, two levels were manipulated in which unmotivated simulees were sampled from (a) across the ability continuum (unrelated); and (b) predominately below the mean ability (related). For this factor, the sampling procedure across unmotivated simulees in both subgroups was constrained equal. Specifically, for level (a), ability parameters for unmotivated simulees were sampled from the same distribution as their motivated counterparts (reference: $N[0, 1]$; focal: $N[0, 1]$ or $N[-0.5, 1]$ depending on whether group impact was present). As prior literature has demonstrated that in some contexts NER occurs more often for low-ability examinees when compared with their higher achieving counterparts (Goldhammer et al., 2016; Kuhfeld & Soland, 2020; Rios et al., 2017; Soland & Kuhfeld, 2019), unmotivated simulees' ability parameters for level (b) were sampled to be -0.5 *SDs* below the mean of the motivated simulees in their respective subgroup. This mean difference value was chosen because Rios et al. (2017) found an average prior ability difference of 0.5 *SDs* (favoring motivated examinees) between motivated and unmotivated test takers. Thus, for the reference subgroup, unmotivated simulees' ability parameters were sampled from $N(-0.5, 1)$. Depending on the presence of group impact, ability parameters for focal subgroup unmotivated simulees were sampled either from $N(-0.5, 1)$ or $N(-1, 1)$ for no impact and impact, respectively, when NER and ability were related.

NER Rate. Although the context simulated reflects a situation in which the reference subgroup is more motivated than the focal, NERs were generated in both subgroups. This was done to mirror the reality that in most low-stakes testing situations, not all examinees will be fully motivated, regardless of subgroup membership. To this end, for the reference subgroup, the percentage of NERs in the data matrix was constrained to 0.5% across all conditions, reflecting the NER rate observed for some of the more motivated subgroups found in DeMars (2007). This percentage was

produced by constraining the percentage of unmotivated simulees in the reference subgroup to 5% and the percentage of NERs for each unmotivated simulee to 10%.

In contrast, the percentage of NERs in the data matrix varied for the focal subgroup, with percentages of 2.5%, 5%, 7.5%, 15%, and 22.5%. To produce these NER rates, the percentage of unmotivated simulees (either 10% or 30%) and percentage of NERs (25%, 50%, or 75%) for each unmotivated simulee was manipulated to within acceptable levels observed in both applied and simulated research (DeMars & Wise, 2010; Rios et al., 2017; Wise & DeMars, 2006). Comparing these NER rates between subgroups reflects the differential NER percentages (ranging from 2% to 22%) observed in operational settings (1% to 23%; e.g., DeMars, 2007; Goldhammer et al., 2016; Rios & Guo, 2020).

Analyses

Evaluating Measurement Invariance. Measurement invariance can be evaluated in both confirmatory factor analytic and item response theory frameworks. In this study, measurement invariance was tested in the former framework for two reasons. First, a single-factor confirmatory factor analysis is equivalent to a unidimensional 2PL item response theory (IRT) model (this was the model used for data generation; e.g., Kamata & Bauer, 2008). Second, reviews of psychological literature have shown that the factor analytic approach is most popular among researchers when testing for measurement invariance (Putnick & Bornstein, 2016). Therefore, as the methodological approaches are identical in the current context, this study adopted the common approach in practice.

Using multiple group confirmatory factor analysis, measurement invariance was evaluated by testing (in order) for configural (equality of factor model configurations), metric (equality of factor loadings), and scalar invariance (equality of factor loadings and intercepts; partial invariance was not assessed) via the *lavaan R* package (version 0.6-5; Rosseel, 2012). Across all tests of measurement invariance, parameterization of the models occurred by setting the factor variance to one for both subgroups. Furthermore, the latent means were constrained to zero for the reference and focal subgroups at the configural and metric invariance levels, given that valid latent mean comparisons cannot be established at these levels (Putnick & Bornstein, 2016); however, once testing for scalar invariance, the focal subgroup latent factor mean was allowed to be freely estimated, which provided an approximation of the difference between latent means of the reference and focal subgroups. This parameterization approach was taken instead of the common tactic of fixing a referent item's factor loading and intercept to 1 and 0, respectively, as the presence of differential NER could have led to choosing a referent item that was noninvariant across groups. The consequence of doing so could lead to other items incorrectly appearing metric and/or scalar invariant due to differences in the latent factor scales across subgroups (Putnick & Bornstein, 2016). The weighted least squares with mean and variance adjustment (WLSMV) estimator was used for each model, as all indicators were dichotomous.³

Measurement Invariance Type I Error Rates. A primary interest of this study was to determine the conditions of NER that would lead to model fit deterioration, and ultimately, Type I error when assessing both metric and scalar invariance tests. For metric invariance tests, this was done by comparing model fit between the configural and metric invariance models, while the latter test evaluated fit between the metric and scalar invariance models. To investigate fit for these nested models, three indices commonly used in research and practice were evaluated: Δ comparative fit index (CFI), Δ root mean square error of approximation (RMSEA), and Δ standardized root mean square residual (SRMR) (Cheung & Rensvold, 2002; Joo & Kim, 2019; Putnick & Bornstein, 2016). Using the guidelines proposed by various researchers (Chen, 2007; Cheung & Rensvold, 2002; Rutkowski & Svetina, 2014), the null hypothesis that invariance should not be rejected was based on meeting two of the following three criteria: $\Delta\text{CFI} \leq -0.01$, $\Delta\text{RMSEA} \leq 0.01$, and $\Delta\text{SRMR} \leq 0.015$. Although most researchers rely on only meeting a single criterion (Putnick & Bornstein, 2016), multiple criteria were employed based on the recommendation that multiple fit indices should be examined prior to making conclusions concerning invariance tests (Cheung & Rensvold, 2002). Type I error was calculated for each replication when the true null hypothesis of full scalar invariance was incorrectly rejected.

Type I Error and Bias in Latent Mean Subgroup Differences. The second variable of interest examined estimated differences in latent means between subgroups. This was investigated only for replications that correctly failed to reject the null hypothesis of full scalar invariance. Replications that did not meet this criterion were dropped from the analysis, as it is recommended that direct mean subgroup comparisons should be avoided if full scalar invariance cannot be established (Putnick & Bornstein, 2016). For replications meeting the criterion, two dependent variables were of interest: (a) Type I error and (b) bias. The former outcome was included to determine whether the estimated focal subgroup latent mean difference was statistically different from its true value. This was done by calculating a one-sample z statistic:

$$z = \frac{\bar{x} - \mu}{se}, \quad (2)$$

where \bar{x} and se are, respectively, the estimated latent mean difference and standard error for the data possessing NERs, and μ was the known latent mean difference. The z -statistic was compared with the critical value for a two-tailed test at an alpha level of .05 (1.96) to determine statistical significance. This index was averaged across replications to compute a summary value for a given condition.

Although this test was informative, it did not provide an indication of the magnitude and direction of difference between estimated and known latent mean subgroup differences. To provide this information bias was calculated:

$$Bias = \left(\frac{1}{R}\right) \sum_{r=1}^R \hat{\phi}_r - \phi_r, \quad (3)$$

where $\hat{\phi}_r$ is the estimated difference in latent means between subgroups for replication r , ϕ_r is the known latent mean difference between subgroups, and R is the total number of replications.⁴ Given that the latent mean variances were set to one, bias was interpreted in *SD* units.

Results

Results are presented separately for measurement invariance and latent mean difference outcomes.

Measurement Invariance Type I Error Rates

Across all conditions, model convergence was met for every replication. As expected, when no NER was present in the baseline data, full scalar invariance was attained for all replications under impact (across conditions, the grand mean $\Delta CFI < .0001$) and no impact (across conditions, the grand mean $\Delta CFI < .0001$). Concerning the effect of NER on measurement invariance, approximately 26% of replications incorrectly rejected the true null hypothesis of full scalar invariance. A closer examination of factors demonstrated that Type I errors only occurred when testing for metric invariance. That is, configural invariance was met for every replication, while full scalar invariance was attained in all cases in which metric invariance was also attained. Given this finding, results are presented below for tests of metric invariance only.

As shown in Table 1, a logistic regression model demonstrated that increased NER rates were significantly associated with a rise in Type I errors of metric invariance; however, after controlling for test length, this relationship was found to be moderated by both group impact and focal subgroup percent. This finding is illustrated in Figure 1, which NER rate on the x -axis and Type I error rate on the y -axis for test lengths of 30 and 60 items. Within each plot, results are presented separately by subgroups differing in their interaction between sample balance and impact. Across both test lengths, Type I errors were not observed when NER rates were 2% and 4.5%. Therefore, results are discussed for the remaining NER rates in which Type I error rates were observed to be as high as 100% under certain conditions.

Across NER percentages of 7%, 14.5%, and 22%, conditions in which samples were unbalanced (i.e., focal group simulees comprised 67% of the sample) were found to possess Type I error rates that were consistently lower than conditions where focal and reference simulees were equal. For instance, when averaging across impact conditions for a 60-item test, the mean Type I error rate for unbalanced conditions was lower by 19% and 22% for NER rates of 14.5% and 22%, respectively. In addition, Type I error rates were greater for conditions in which subgroup impact was present (i.e., the focal subgroup possessed an average ability that was 0.5 *SDs* lower

Table 1. Results of Regressing Measurement Invariance Type I Error on Study Factors.

Predictor	Estimate	SE
Intercept	-9.77***	0.46
Test Length	0.21*	0.09
Group Impact	4.08***	0.50
Ability Relationship	-0.11	0.09
Focal Percent	2.40***	0.53
NER Rate	93.65***	5.02
NER Rate × Group Impact	-58.79***	5.21
NER Rate × Focal Percent	-28.33***	5.67
Group Impact × Focal Percent	-2.19***	0.62
NER Rate × Group Impact × Focal Percent	18.01**	5.98

Note. A logistic regression analysis was conducted for the Type I error dependent variable ($N = 4,800$).

Estimates are on a logit scale. NER = noneffortful responding.

* $p < .05$. ** $p < .01$. *** $p < .0001$.

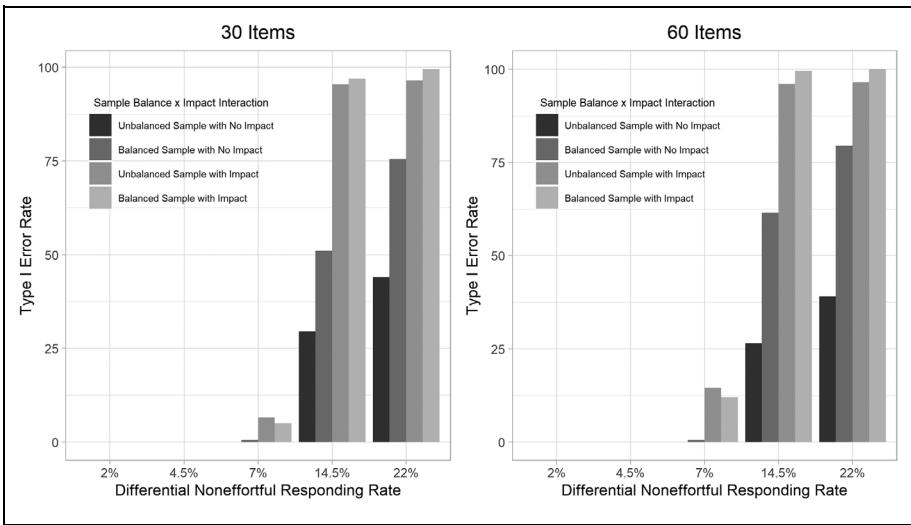


Figure 1. Measurement invariance Type I error rates.

Note. Type I error rates shown are aggregated across relationship with ability conditions.

than the reference subgroup). As an example, across focal subgroup percentages and test length conditions, average Type I error rates ranged from 96% to 98% for impact conditions, while they were as low as 40% to 60% under no impact. Taken together, Figure 1 clearly shows that when examining the interactions between sample balance and group impact for NER rates ranging from 7% to 22%, the lowest Type I error

Table 2. Latent Variable Mean Difference Type I Error Rates and Bias.

Condition		Differential NER rate		
Test length	Group impact	NER-ability relationship	2%	7%
30 items	No impact	Related	19% (-0.04)	27.5% (-0.05)
		Unrelated	7.5% (-0.02)	14% (-0.03)
	Impact	Related	19.5% (-0.04)	24.5% (-0.04)
		Unrelated	9.5% (-0.02)	11.5% (-0.03)
60 items	No impact	Related	32% (-0.04)	52% (-0.06)
		Unrelated	6% (-0.01)	19% (-0.04)
	Impact	Related	25% (-0.05)	38% (-0.06)
		Unrelated	13% (-0.02)	20% (-0.04)

Note. The average latent mean difference bias is presented in parentheses. Results are averaged across focal simulee percentage. For the 2% and 4.5% conditions, sample sizes were equal to 100, while the 7% condition is based on sample sizes of 93 and 85 for test lengths of 30 and 60 items, respectively (replications failing to attain full scalar invariance were removed). Differential NER rates of 14.5% and 22% were excluded due to the high measurement invariance Type I errors observed for these conditions. NER = noneffortful responding.

Table 3. Results of Regressing Latent Mean Difference Type I Error on Study Factors.

Predictor	Estimate	SE
Intercept	0.52***	0.02
No. items	-0.03	0.02
Group Impact	-0.10***	0.01
Ability Relationship	0.19***	0.02
Focal Percent	0.03*	0.02
NER Rate: 4.5%	-0.38***	0.03
NER Rate: 7%	-0.22**	0.03
No. items × NER Rate: 4.5%	0.17***	0.03
No. items × NER Rate: 7%	0.18***	0.03
Ability Relationship × NER Rate: 4.5%	-0.01	0.03
Ability Relationship × NER Rate: 7%	-0.11***	0.03

^aDue to the extensive number of replications that failed to attain full scalar invariance, the 14.5% and 22% NER rate conditions were excluded. Estimates are on a logit scale. NER = noneffortful responding.

* $p < .05$. ** $p < .01$. *** $p < .0001$.

rates were observed for conditions in which there was an unbalanced sample with no impact, while the highest rates occurred for conditions with a balanced sample and group impact (see Appendix C of online Supplemental Material for descriptive results).

Latent Mean Type I Error and Bias When Attaining Full Scalar Invariance

Table 2 presents the Type I error rates and bias for estimated latent mean subgroup differences. Results are only presented for replications that attained full scalar invariance (allowing for direct mean comparisons) when the subgroups differed in NER by 2%, 4.5%, and 7% (NER rates of 14.5% and 22% were excluded due to their high Type I errors). Across these conditions, this led to the inclusion of between 85% and 93% of replications for a NER rate of 7% (dependent on test length) and 100% of replications for NER rates of 2% and 4.5%.

Across test length, group impact, and NER–ability relationship conditions, Type I errors increased as the differential NER rate between subgroups increased, with error rates ranging from 6% to 100% (Table 2). However, as is shown in Table 3, the association between Type I error and NER rate was moderated by both test length and the NER–ability relationship. Concerning the former moderator, higher Type I error rates of latent mean subgroup differences were observed for the longer of the two test length conditions. For instance, aggregating across group impact and NER–ability relationship conditions, Type I error was greater for the 60-item condition by 5%, 13%, and 5% for NER rates of 2%, 4.5%, and 7%, respectively; though, this result may be associated with the greater statistical power obtained in the longer test

condition (i.e., more statistically significant differences were observed because the standards errors of the latent subgroup mean differences estimates were smaller). This is supported by an examination of the bias results, which showed nearly identical magnitude of negative bias between test lengths.

Turning to the NER–ability relationship moderating effect, results demonstrated that when simulees engaging in NER were predominately of lower ability a greater degree of Type I errors was observed. As an example, for the unrelated NER–ability condition, estimated latent mean differences across test lengths were found to be statistically different from their known values by 9%, 16%, and 74% for NER rates of 2%, 4.5%, and 7%, respectively. In comparison, for the same NER rates, when NER was related to simulees’ underlying ability, Type I error rates increased to 24%, 36% and 100%. Concerning the extent and direction of estimation distortion, the bias results demonstrated that across conditions estimated latent mean differences were always lower than the true difference; however, the degree of negative bias was consistently smaller for the unrelated NER–ability condition. For instance, when NER and ability were related, the average bias for NER rates of 2%, 4.5%, and 7% were equal to -0.04 , -0.05 , and -0.19 *SDs* across test length and impact conditions, which was two times larger than the values observed in the unrelated condition.

Applied Analysis

An applied analysis is included to examine how differential NER may influence decisions around measurement invariance analyses in practice. To do this, data were examined for examinees sampled from two countries (mirroring the simulation study design) who were administered the PISA. Details of the methodology for this analysis are described below.

Methodology

Sample. Data were sampled from examinees administered the PISA science domain (more detail provided below) from the United Arab Emirates (UAE; $n = 763$) and China ($n = 452$). These countries were selected as they provided some of the largest sample sizes compared with all other countries and represented distinctive cultures from the Middle East and Asia that have been shown to display differential levels of test endurance (OECD, 2019). In each country, examinees were sampled (using a matrix sampling design) from 5,000 nationally representative students attending 150 schools or more. Examinees were excluded if possessing: (a) a moderate to severe permanent physical, cognitive, behavioral, or emotional disability that would not allow them to participate in testing; (b) and/or limited proficiency in the assessment language.

Measure. PISA is an international assessment measuring 15-year-old’s knowledge and skills in reading, mathematics, and science. The focus of this study is on Form 18 of the 2018 administration of the science literacy (i.e., knowledge of science and of

science-based technology) domain, which comprised item Clusters 1 and 6 (each cluster was expected to take 30 minutes to complete). As response times were utilized as a proxy of NER (more detailed is provided below), only the 30 keyed selected-response option items were utilized, due to the limitations associated with current methods to evaluate test-taking effort for items with open-ended response options (see Wise, 2017).

Analysis. The analysis consisted of two distinctive activities. First, NER was identified via response times. Specifically, a response time threshold was established in which any response provided in less time than the criterion was classified as a noneffortful response. Although there are multiple approaches to choosing a threshold (for details, see Wise, 2017), this study adopted the same approach taken by Wise and Kuhfeld (2020) in which any response provided in less than 30% of the sample's average response time for the given item of interest was deemed to be a noneffortful response. Due to differences in reading load introduced by the separate testing languages employed for the Brazilian and Chinese samples, a criterion threshold for each item was established separately by country. On identifying noneffortful responses, a filtered data set was created in which noneffortful responses were treated as missing based on the assumption that such responses are uninformative in reflecting an examinee's underlying science knowledge (see Wise & DeMars, 2006).

To examine the potential impact of noneffortful responses on inferences related to measurement equivalence, nested invariance analyses were conducted separately for unfiltered (including noneffortful responses) and filtered (treating noneffortful responses as missing) data sets. These invariance analyses were conducted in the exact manner described in the analysis subsection of the simulation study. Following the recommendations of Kline (2005), the chi-square statistic, CFI, RMSEA, and SRMR fit indices were reported. Adequate model fit was supported by meeting the following criteria for at minimum two of the three fit indices: $CFI \geq .90$, $RMSEA \leq .06$, and $SRMR \leq .08$ (Hu & Bentler, 1999). Similar to the simulation study, the following criteria were used to evaluate the fit of nested models: $\Delta CFI < -.01$, $\Delta RMSEA < .01$, and $\Delta SRMR < .015$. If the fit of a constrained model was found to exceed two of the three criteria, it was determined that the additional equalities specified led to significant model deterioration.

Results

Invariance analyses are first presented for the unfiltered data. The first step of this analysis was to establish a baseline model by fitting each country's data separately to test for unidimensionality. Across both UAE ($\chi^2 = 501.34$, $df = 405$, $p = .001$; $CFI = .988$; $RMSEA = .018$; $SRMR = .052$) and Chinese samples ($\chi^2 = 426.68$, $df = 405$, $p = .24$; $CFI = .991$; $RMSEA = .010$; $SRMR = .079$), the unidimensional model was found to provide adequate fit to the sample data. Next, configural invariance was evaluated across countries based on the unidimensional model and found to be

supported in the sample data based on CFI, RMSEA, and SRMR model fit statistics (CFI and RMSEA; $\chi^2 = 974.68$, $df = 868$, $p = .007$; CFI = .989; RMSEA = .014; SRMR = .063), suggesting that the overall factor structure stipulated fit equally well across UAE and Chinese examinees. However, constraining the factor loadings equal across countries ($\chi^2 = 1527.85$, $df = 899$, $p < .001$; CFI = .937; RMSEA = .034; SRMR = .090) led to significant model deterioration according to the Δ CFI (−.052), Δ RMSEA (.020), and Δ SRMR (.027) indices. This result indicates that the magnitudes of the factor loadings across countries were not equivalent (i.e., the scale origin differed by sample). As a consequence, there is a lack of evidence to support the validity of making direct mean comparisons across UAE and Chinese samples on the science assessment examined based on the unfiltered data.

Turning next to the invariance analyses based on filtered data, a comparison between the two countries sampled demonstrated large differences in NERs. Specifically, the percentage of noneffortful responses in the UAE data matrix was approximately 20% compared with only 8% for the Chinese sample. Although, the percentage of noneffortful responders (i.e., examinees engaging in at least one noneffortful response) was only 9% higher in the UAE sample (85% compared with 76% in the Chinese sample), nearly 50% of Emirati examinees noneffortfully responded on more than five of 31 items and 15% provided a disengaged response on 50% or more of items. In comparison, almost 70% of Chinese noneffortful responders provided a disengaged response for 15% or less of items, while no examinees provided noneffortful responses on 50% or more of items. Furthermore, the average number of noneffortful responses per examinee was higher by 0.70 *SDs* for Emirati examinees ($M = 6.08$; $SD = 7.12$) when compared with the Chinese ($M = 2.32$; $SD = 2.39$).

On filtering noneffortful responses, a unidimensional model was fit separately to each country's data to establish a baseline model. Across UAE ($\chi^2 = 472.40$, $df = 405$, $p = .01$; CFI = .987; RMSEA = .015; SRMR = .063) and Chinese samples ($\chi^2 = 414.51$, $df = 405$, $p = .36$; CFI = .994; RMSEA = .007; SRMR = .088), the data were found to support a unidimensional factor structure. Fitting the multiple group configural invariance model showed excellent fit to the sample data across the CFI, RMSEA, and SRMR indices ($\chi^2 = 942.07$, $df = 868$, $p = .04$; CFI = .989; RMSEA = .012; SRMR = .073). Turning to the stricter metric invariance model, the analysis showed significant model fit deterioration ($\chi^2 = 1281.98$, $df = 899$, $p < .01$; CFI = .944; RMSEA = .027; SRMR = .092) across all three criteria (Δ CFI = −.055; Δ RMSEA = .015; Δ SRMR = .019).

This result has two implications. First, based on filtered data, there is no evidence to support direct mean comparisons between Emirati and Chinese samples, given a failure to attain metric invariance. Second, although the measurement invariance inferences between the filtered and unfiltered data sets were similar, the results demonstrated improved model fit at the metric invariance level when filtering noneffortful responses for two of the three indices (filtered—unfiltered; Δ CFI = −.003; Δ RMSEA = −.005; Δ SRMR = −.008).

Discussion

The objective of this study was to examine the impact of differential NER on measurement invariance analyses and latent mean subgroup comparisons. Overall, results demonstrated that Type I errors of measurement invariance were found to occur as differential NER rates between subgroups grew, with Type I errors observed under certain conditions with as little as a 7% difference in NER, which is well within the range observed in prior applied analyses (e.g., Goldhammer et al., 2016; Rios & Guo, 2020). When invariance was incorrectly rejected, which was observed for 26% of replications investigated, it was done consistently at the metric invariance level, with no negative impacts on either configural or scalar invariance. This finding indicates that NER significantly led to biased factor loading estimates, which supports prior literature that has demonstrated that NER, when ignored, typically leads to significant underestimation of item discrimination (an equivalent of factor loadings in the IRT framework; e.g., Rios & Soland, 2020). Another potential cause for the incorrect rejection of measurement invariance is that differential NER could be associated with misspecified correlated errors, which have been shown to lead to Type I errors of metric invariance tests (but not scalar invariance tests; see Joo & Kim, 2019).

The relationship between NER rate and Type I error was found to be moderated by subgroup sample sizes and group impact. Concerning the former, when the subgroups were unbalanced (i.e., focal group simulees comprised 67% of the sample), the percentage of Type I error was lower. One potential reason for the lower Type I errors is that imbalanced subgroup sample sizes can mask violations of measurement invariance (Yoon & Lai, 2018). In addition, across conditions, group impact was associated with higher Type I error rates. Prior research conducted by Stark et al. (2006) supports this finding, as these authors showed that when testing for measurement invariance within a confirmatory factor analytic framework, Type I errors can increase in the presence of group impact, particularly when sample sizes are large ($N = 1,000$). This is likely due to differential stability of model parameter estimates based on a shifting of the ability distribution, leading to less available data for estimation. When this is coupled with differential NER, measurement invariance model parameter estimates and Type I errors can be biased. Taken together, findings from this study suggest that Type I error rates may be quite high (as high as 100%) when testing for measurement invariance in the presence of differential NER between subgroups that are unbalanced in sample size and differ in their underlying mean abilities; however, this is largely dependent on the NER rate. As demonstrated via data from PISA and other operational testing contexts, subgroups in practice can differ in NER by as much as 22%, which at minimum can lead to model fit deterioration, and potentially incorrect measurement invariance inferences under the certain contexts.

Although minimal Type I errors were observed for NER rates less than or equal to 7% when testing for measurement invariance, differential NER still led to bias in

estimates of latent mean differences. Specifically, the relationship between latent mean difference Type I error (and bias) and NER rate was moderated by whether simulees engaging in NER were predominately of lower ability or were representative of the entire ability continuum. As expected, greater Type I error and bias were observed for the former condition, due to the tendency of overestimating group ability when NER is related to the underlying ability of examinees (see Rios & Soland, 2020). As a consequence, for a NER rate of 7%, Type I error rates reached 100%, while latent mean differences were biased by an average of -0.18 *SDs*. To put this magnitude into context, the observed degree of negative bias is nearly equivalent to two-thirds a year reduction in the average annual growth in science for K-12 students in the United States (0.29 *SDs*; Bloom et al., 2008). Such a degree of bias has the potential to negatively affect the validity of inferences around subgroup inferences concerning achievement gains (e.g., Wise & DeMars, 2010), treatment effects (e.g., Osborne & Blanchard, 2011), and international comparisons (e.g., Debeer et al., 2014), to name a few. Taken together, findings from this study suggest that even when differential NER does not lead to Type I errors of measurement invariance, under certain conditions, its presence is still linked to making potentially incorrect inferences concerning subgroup comparisons given the tendency of NER to mask true differences.

Limitations and Directions for Future Research

A number of study limitations should be noted. First, although the simulation design in this study included factors that considered underlying sample size, ability, and NER rate differences between subgroups, additional variables should be explored in future research. As an example, across simulation conditions, the number of subgroups in the invariance analyses was constrained to two. Although this reflects the most common number of subgroups included in most simulation research on measurement invariance topics, given the limited research, it is unclear what the influence on fit indices and measurement invariance inferences would be if increasing the number of subgroups (Putnick & Bornstein, 2016). Clearly, more research is needed in this area, and as a result, readers should limit the generalizability of findings to the two-group context. Similarly, while this study assumed normal ability distributions, prior research has found that skewed latent trait distributions can influence measurement invariance testing (Finch et al., 2018). As such, future research should investigate the dependent variables examined in this study under conditions with skewed ability distributions and group impact.

Second, there is a need to research the practical effect of differential NER on other measurement contexts. One area with serious potential consequences is score linking. Though Mittelhaeuser et al. (2015) examined the role of differential test-taking effort for linking under external anchor and pretest designs, many international testing

programs, such as PISA, use an internal anchor design with IRT concurrent calibration linking. In such an approach, item parameters are tested for invariance across forms (or countries) to identify an internal anchor for which to link scores. However, as countries in international testing contexts, such as PISA, have been found to show large variation in test-taking effort (e.g., Debeer et al., 2014), it would be of interest to examine how differential NER could affect the accuracy of identifying anchor items, and ultimately, bias in linking coefficients.

Recommendations for Practice

As the presence of differential NER can negatively influence both evaluations of scale property equivalence and latent mean comparisons across subgroups, it is vital that practitioners document evidence that subgroups put forth equal effort when disaggregating data. To provide this evidence, a number of procedures have been proposed that rely on survey data, item responses, and/or the availability of response time data (for a review, see Wise & Kong, 2005). Furthermore, a number of filtering procedures and IRT models have been developed to improve ability estimation in the presence of NER (e.g., Liu et al., 2019; Rios et al., 2017; Rios & Soland, 2020; Wise & Kingsbury, 2016). Although further research is needed to continually improve the accuracy of both identification and ability estimation procedures for NER, many options are readily available for practitioners to document and attempt to improve ability estimation in the presence of differential NER.

Beyond documenting this information post hoc, practitioners can attempt to increase test-taking effort either before and/or during test administration to mitigate NER. To this end, Rios (in press) has documented several interventions that researchers have developed to improve test-taking effort, which include increasing test relevance, providing feedback, altering test design and administration procedures, and offering contingency-based external incentives. Although Rios found the latter two intervention types to be most successful on average, there has been minimal research that has investigated whether the utility of interventions is equivalent across subpopulations. Clearly, further research is needed on this topic; however, practitioners can still attempt to mitigate NER by addressing low test-taking effort via some of these interventions.

Therefore, it is recommended that prior to evaluating measurement invariance and making subgroup comparisons from low-stakes testing contexts, practitioners should (a) employ interventions to improve test-taking effort, (b) document potential differential NER, and (c) filter (i.e., remove) or model NERs to improve item and ability parameter estimation accuracy. A failure to do so may lead to incorrect inferences concerning scale property equivalence and subgroup mean comparisons when differential NER is present.

Acknowledgments

The author would like to thank Michael Rodriguez and Samuel Ihlenfeldt from the University of Minnesota for providing comments on an earlier draft of the manuscript.


Declaration of Conflicting Interests

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

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

Supplemental Material

Supplemental material for this article is available online.

Notes

1. This research has primarily relied on the use of response times to identify one type of noneffortful responding referred to as rapid guessing (i.e., a respondent provides a response in so little time that they would be able to read the item stem and response options; for more detail, the reader is referred to Wise, 2017).
2. The number of noneffortful responses in each of the five bins for a within-simulee NER rate of 25% was 1, 2, 3, 4, and 5, while for 75% it was 6, 6, 9, 12, and 12. It is acknowledged that there is a multitude of ways to disperse noneffortful responses across the test, however, the approach taken was meant to reflect a progressive decrease in an examinee's test-taking effort.
3. One reviewer suggested employing maximum likelihood estimation with robust standard errors. Although this estimation procedure has been found to perform similarly to the WLMSV (weighted least squares with mean and variance adjustment) estimator under certain conditions (Bandalos, 2014), it was not employed in this study due to its unavailability in the *lavaan* R package at the time of this writing.
4. It should be noted that the baseline data captured latent mean differences under completely effortful responding in both subgroups, while for the estimated data the latent mean differences were captured based on the inclusion of NER in both subgroups, including the reference subgroup (0.5%). As such, it is expected that the degree of bias is underestimated; however, the degree of underestimation is likely negligible.

References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing* (6th ed.). American Educational Research Association.
- Bandalos, D. L. (2014). Relative performance of categorical diagonally weighted least squares and robust maximum likelihood estimation. *Structural Equation Modeling: A Multidisciplinary Journal*, 21(1), 102-116. <https://doi.org/10.1080/10705511.2014.859510>

- Bloom, H. S., Hill, C. J., Black, A. B., & Lipsey, M. W. (2008). Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions. *Journal of Research on Educational Effectiveness*, 1(4), 289-328. <https://doi.org/10.1080/19345740802400072>
- Boe, E. E., May, H., & Boruch, R. F. (2002). *Student task persistence in the Third International Mathematics and Science Study: A major source of achievement differences at the national, classroom, and student levels* (Research Report 2002-TIMSS1). Center for Research and Evaluation in Social Policy, University of Pennsylvania.
- Borghans, L., & Schils, T. (2012). *The leaning tower of PISA: Decomposing achievement test scores into cognitive and noncognitive components* [Unpublished manuscript].
- Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, 14(3), 464-504. <https://doi.org/10.1080/10705510701301834>
- Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, 9(2), 233-255. https://doi.org/10.1207/S15328007SEM0902_5
- 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(6), 502-523. <https://doi.org/10.3102/1076998614558485>
- DeMars, C. E. (2007). Changes in rapid-guessing behavior over a series of assessments. *Educational Assessment*, 12(1), 23-45. <https://doi.org/10.1080/10627190709336946>
- DeMars, C. E. (2010). Type I error inflation for detecting DIF in the presence of impact. *Educational and Psychological Measurement*, 70(6), 961-972. <https://doi.org/10.1177/0013164410366691>
- DeMars, C. E., Bashkov, B. M., & Socha, A. B. (2013). The role of gender in test-taking motivation under low-stakes conditions. *Research & Practice in Assessment*, 8, 69-82.
- 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. <https://doi.org/10.1080/15305058.2010.496347>
- Dimitrov, D. M. (2010). Testing for factorial invariance in the context of construct validation. *Measurement and Evaluation in Counseling and Development*, 43(2), 121-149. <https://doi.org/10.1177/0748175610373459>
- Finch, H. W., French, B. F., & Hernández Finch, M. E. (2018). Comparison of methods for factor invariance testing of a 1-factor model with small samples and skewed latent traits. *Frontiers in Psychology*, 9(332), 1-12. <https://doi.org/10.3389/fpsyg.2018.00332>
- Fischer, R., & Karl, J. A. (2019). A primer to (cross-cultural) multi-group invariance testing possibilities in R. *Frontiers in Psychology*, 10, Article 1507. <https://doi.org/10.3389/fpsyg.2019.01507>
- Goldhammer, F., Martens, T., Christoph, G., & Lüdtke, O. (2016). *Test-taking engagement in PIAAC*. OECD.
- Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1), 1-55. <https://doi.org/10.1080/10705519909540118>
- Joo, S. H., & Kim, E. S. (2019). Impact of error structure misspecification when testing measurement invariance and latent-factor mean difference using MIMIC and multiple-

- group confirmatory factor analysis. *Behavior Research Methods*, 51(6), 2688-2699. <https://doi.org/10.3758/s13428-018-1124-6>
- Kam, C. C. S., & Meyer, J. P. (2015). How careless responding and acquiescence response bias can influence construct dimensionality: The case of job satisfaction. *Organizational Research Methods*, 18(3), 512-541. <https://doi.org/10.1177/1094428115571894>
- Kamata, A., & Bauer, D. J. (2008). A note on the relation between factor analytic and item response theory models. *Structural Equation Modeling: A Multidisciplinary Journal*, 15(1), 136-153. <https://doi.org/10.1080/10705510701758406>
- Kim, K. H. (2005). The relation among fit indexes, power, and sample size in structural equation modeling. *Structural Equation Modeling*, 12(3), 368-390. https://doi.org/10.1207/s15328007sem1203_2
- Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). Guilford Press.
- 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. <https://doi.org/10.1080/19345747.2019.1636437>
- Liu, Y., Li, Z., Liu, H., & Luo, F. (2019). Modeling test-taking non-effort in MIRT models. *Frontiers in Psychology*, 10, 145. <https://doi.org/10.3389/fpsyg.2019.00145>
- Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. Routledge.
- Mittelhaeuser, M. A., Béguin, A. A., & Sijtsma, K. (2015). The effect of differential motivation on IRT linking. *Journal of Educational Measurement*, 52(3), 339-358. <https://doi.org/10.1111/jedm.12080>
- OECD. (2019). *PISA 2018 Results (Volume 1): What students know and can do*. <https://doi.org/10.1787/5f07c754-en>
- 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, Article 220. <https://doi.org/10.3389/fpsyg.2010.00220>
- Penk, C., Pöhlmann, C., & Roppelt, A. (2014). The role of test-taking motivation for students' performance in low-stakes assessments: An investigation of school-track-specific differences. *Large-Scale Assessments in Education*, 2(1), 1-17. <https://doi.org/10.1186/s40536-014-0005-4>
- 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. <https://doi.org/10.1016/j.lindif.2015.08.002>
- Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review*, 41, 71-90. <https://doi.org/10.1016/j.dr.2016.06.004>
- R Development Core Team. (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Rios, J. A. (in press). Improving test-taking motivation on low-stakes educational assessments: A meta-analysis of interventions. *Applied Measurement in Education*.
- Rios, J. A., & Guo, H. (2020). Can culture be a salient predictor of test-taking engagement? An analysis of differential NER on an international college-level assessment of critical thinking. *Applied Measurement in Education*, 33(4), 263-279. <https://doi.org/10.1080/08957347.2020.1789141>

- Rios, J. A., Guo, H., Mao, L., & Liu, O. L. (2017). Evaluating the impact of careless responding on aggregated-scores: To filter unmotivated examinees or not? *International Journal of Testing, 17*(1), 74-104. <https://doi.org/10.1080/15305058.2016.1231193>
- Rios, J. A., & Soland, J. (2020). Parameter estimation accuracy of the Effort-Moderated Item Response Theory Model under multiple assumption violations. *Educational and Psychological Measurement*. Advance online publication. <https://doi.org/10.1177/0013164420949896>
- Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1-36. <https://doi.org/10.18637/jss.v048.i02>
- Rutkowski, L., & Svetina, D. (2014). Assessing the hypothesis of measurement invariance in the context of large-scale international surveys. *Educational and Psychological Measurement, 74*(1), 31-57.
- Schnipke, D. L. (1995, April). *Assessing speededness in computer-based tests using item response times* [Paper presentation]. Annual meeting of the National Council on Measurement in Education, San Francisco, CA, United States.
- Setzer, J. C., Wise, S. L., van den Heuvel, J. R., & Ling, G. (2013). An investigation of examinee test-taking effort on a large-scale assessment. *Applied Measurement in Education, 26*(1), 34-49. <https://doi.org/10.1080/08957347.2013.739453>
- Smith, J. K., Given, L. M., Julien, H., Ouellette, D., & DeLong, K. (2013). Information literacy proficiency: Assessing the gap in high school students' readiness for undergraduate academic work. *Library & Information Science Research, 35*(2), 88-96. <https://doi.org/10.1016/j.lisr.2012.12.001>
- 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), 1-26.
- Soland, J., & Kuhfeld, M. (2019). Do students rapidly guess repeatedly over time? A longitudinal analysis of student test disengagement, background, and attitudes. *Educational Assessment, 24*(4), 327-342. <https://doi.org/10.1080/10627197.2019.1645592>
- Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology, 91*(6), 1292-1306. <https://doi.org/10.1037/0021-9010.91.6.1292>
- van Barneveld, C. (2007). The effect of examinee motivation on test construction within an IRT framework. *Applied Psychological Measurement, 31*(1), 31-46. <https://doi.org/10.1177/0146621606286206>
- Wise, S. L. (2015). Effort analysis: Individual score validation of achievement test data. *Applied Measurement in Education, 28*(3), 237-252. <https://doi.org/10.1080/08957347.2015.1042155>
- Wise, S. L. (2017). Rapid-guessing behavior: Its identification, interpretation, and implications. *Educational Measurement: Issues and Practice, 36*(4), 52-61. <https://doi.org/10.1111/emip.12165>
- Wise, S. L., & Cotton, M. R. (2009). Test-taking effort and score validity: The influence of student conceptions of assessment. In D. M. McInerney, G. T. L. Brown, & G. A. D. Liem (Eds.), *Student perspectives on assessment: What students can tell us about assessment for learning* (pp. 187-206). Information Age.
- Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment, 10*(1), 1-18. https://doi.org/10.1207/s15326977ea1001_1

- 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. <https://doi.org/10.1111/j.1745-3984.2006.00002.x>
- Wise, S. L., & DeMars, C. E. (2009). A clarification of the effects of rapid guessing on coefficient alpha: A note on Attali's reliability of speeded number-right multiple-choice tests. *Applied Psychological Measurement*, 33(6), 488-490. <https://doi.org/10.1177/0146621607304655>
- Wise, S. L., & DeMars, C. E. (2010). Examinee noneffort and the validity of program assessment results. *Educational Assessment*, 15(1), 27-41. <https://doi.org/10.1080/10627191003673216>
- 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. <https://doi.org/10.1111/jedm.12102>
- Wise, S. L., Kingsbury, G. G., Thomason, J., & Kong, X. (2004, April). *An investigation of motivation filtering in a statewide achievement testing program* [Paper presentation]. Annual meeting of the National Council on Measurement in Education, San Diego, CA, United States.
- 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. https://doi.org/10.1207/s15324818ame1802_2
- 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: <https://doi.org/10.1111/jedm.1227510.1111/jedm.12275>
- Wolf, E. J., Harrington, K. M., Clark, S. L., & Miller, M. W. (2013). Sample size requirements for structural equation models: An evaluation of power, bias, and solution propriety. *Educational and Psychological Measurement*, 73(6), 913-934. <https://doi.org/10.1177/0013164413495237>
- Yoon, M., & Lai, H. C. (2018). Testing factorial invariance with unbalanced samples. *Structural Equation Modeling: A Multidisciplinary Journal*, 25(2), 201-213. <https://doi.org/10.1080/10705511.2017.1387859>
- Zamarro, G., Hitt, C., & Mendez, I. (2019). When students don't care: Reexamining international differences in achievement and student effort. *Journal of Human Capital*, 13(4), 519-552. <https://doi.org/10.1086/705799>