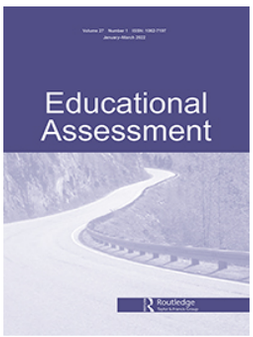

Routledge
Taylor & Francis Group

Educational Assessment

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/heda20

To What Degree Does Rapid Guessing Distort Aggregated Test Scores? A Meta-analytic Investigation

Joseph A. Rios, Jiayi Deng & Samuel D. Ihlenfeldt

To cite this article: Joseph A. Rios, Jiayi Deng & Samuel D. Ihlenfeldt (2022): To What Degree Does Rapid Guessing Distort Aggregated Test Scores? A Meta-analytic Investigation, Educational Assessment, DOI: 10.1080/10627197.2022.2110465

To link to this article: https://doi.org/10.1080/10627197.2022.2110465

EDUCATIONAL ASSESSMENT
https://doi.org/10.1080/10627197.2022.2110465

Routledge
Taylor & Francis Group

To What Degree Does Rapid Guessing Distort Aggregated Test Scores? A Meta-analytic Investigation

Joseph A. Rios, Jiayi Deng, and Samuel D. Ihlenfeldt

University of Minnesota, Twin Cities, Minnesota, USA

ABSTRACT

The present meta-analysis sought to quantify the average degree of aggregated test score distortion due to rapid guessing (RG). Included studies group-administered a low-stakes cognitive assessment, identified RG via response times, and reported the rate of examinees engaging in RG, the percentage of RG responses observed, and/or the degree of score distortion in aggregated test scores due to RG. The final sample consisted of 25 studies and 39 independent samples comprised of 443,264 unique examinees. Results demonstrated that an average of 28.3% of examinees engaged in RG (21% were deemed to engage in RG on a nonnegligible number of items) and 6.89% of item responses were classified as rapid guesses. Across 100 effect sizes, RG was found to negatively distort aggregated test scores by an average of 0.13 standard deviations; however, this relationship was moderated by both test content area and filtering procedure.

For over a century, researchers have warned that in certain circumstances test scores can be contaminated by item responses that are not reflective of examinees' maximal effort (see Wise & Smith, 2011). Such a concern has been raised as a validity threat in the *Standards for Educational and Psychological Testing* (American Educational Research Association et al., 2014) and has spurred extensive research into noneffortful responding (i.e., responding with intentional disregard for item content; e.g., Meade & Craig, 2012; Wise, 2017). One form of noneffortful responding that has gained exponential attention in the literature is rapid guessing (RG), which occurs when an examinee provides a response in a time that would not allow one to read the item stem or response options, solve the problem, and provide an answer. Assuming that sufficient time is provided to respond to an item, RG has largely been shown to occur in testing contexts in which assessment results have little to no personal consequences for examinees (i.e., low-stakes testing contexts; Wise & DeMars, 2005). One hypothesis for this consistent finding is that, for some examinees, test-taking effort may dissipate when an assessment activity is perceived to hold minimal personal value (Wise, 2017).

Given the potential for RG to undermine the validity of inferences made from such measurement contexts, it is critical for practitioners to become aware of the degree to which aggregated test scores can be distorted due to RG and how such a relationship is moderated by sample, assessment, and methodological factors.[1] To that end, we conduct a meta-analysis of primary studies that group-administered a low-stakes cognitive assessment, identified RG via response times (RT), and reported the rate of examinees engaging in RG, the percentage of RG responses observed, and/or the degree of score distortion in aggregated test scores due to RG. In the sections that follow, we elaborate on how

CONTACT Joseph A. Rios jrios@umn.edu Department of Educational Psychology, University of Minnesota, 56 E. River Road, 164 Education Sciences Building, Minneapolis, MN 55455, USA

[1]The term "score distortion" is used throughout this manuscript to reflect differences between scores based on data that have and have not been filtered for RG responses. A term often used in the simulation literature is "bias"; however, given that we examined operational data, true scores reflective of pure solution behavior are unknown, due to the use of test-taking effort proxies.

Supplemental data for this article can be accessed online at https://doi.org/10.1080/10627197.2022.2110465

RG is classified in practice, methods for dealing with RG in scoring test performance once identified, the impact of RG on measurement properties and test performance results, and the need for the current meta-analysis.

Identification of rapid guessing in practice

Response times have been the most popular source of information utilized to identify RG in both the literature (e.g., Rios et al., 2017) and operational testing contexts (e.g., Wise & Kuhfeld, 2020). In this approach, a RT threshold is defined in which any response provided in less time than the defined criterion is considered to be RG (for details, see Rios & Deng, 2021; Wise, 2017). The extensive adoption of RT as a proxy for RG is associated with its ability to be unobtrusive (i.e., examinees are unaware that their behavior is tracked) and identify RG on an item-by-item basis for each examinee (e.g., Silm et al., 2020).

This latter strength is of particular importance as prior research has demonstrated that examinees' employment of RG can change throughout a test administration, given cognitive fatigue and decreased interest in the assessment activity (e.g., Wise, 2017; Wise & Kingsbury, 2016). The capability of RT methods to identify RG at the item response level allows for the investigation of item characteristic correlates associated with RG, which can be used to modify test development practices to reduce RG. Additionally, it provides several scoring advantages, which are discussed in the next section.

Filtering approaches for dealing with RG in scoring test performance

Upon identifying RG, practitioners are next confronted with how to handle such responses in scoring. Given the ability to identify RG at the item response level for each examinee, two approaches have been widely adopted to report aggregated group scores (i.e., not individual scores; see Rios et al., 2014). The first, referred to as examinee-level filtering, employs listwise deletion of examinee data for those individuals employing RG on a percentage of items that exceeds an acceptable level predefined by practitioners/researchers. The rationale of such an approach is that data from these examinees is untrustworthy, and consequently, should be removed prior to data analyses. Although a simple approach, it assumes that examinees engaging in RG are representative of the entire ability continuum (i.e., RG is unrelated to examinees' underlying ability). If this assumption is untenable, listwise deletion will result in either an inflation or underestimation of the true mean depending on whether filtered examinees are predominately of low or high ability (Rios et al., 2017).

An alternative to examinee-level filtering is to filter individual item responses. This latter approach avoids the loss of extensive data that is seen in applications of the former (i.e., examinee-level filtering can exclude as much as 25% of sample data; Rios et al., 2014) by including data from all examinees engaging in RG when estimating model parameters (with the intention of making aggregate-level inferences). Similar to examinee-level filtering, this approach assumes that RG can be identified correctly and that any response not classified as RG is a valid indicator of examinee ability. Based on this latter assumption, proponents of this approach argue that although an examinee may engage in RG, their data may possess valid responses that can be used in estimating ability (e.g., Rios & Soland, 2021b; Wise & Kingsbury, 2016). To this end, a number of Item Response Theory (IRT) models have been proposed (for a review, see Deribo et al., 2021; Liu et al., 2019; Wise & Kingsbury, 2016), with some (e.g., Effort-Moderated IRT model) showing improved estimation accuracy over traditional models that ignore the presence of RG (e.g., Liu et al., 2019; Rios & Soland, 2021a). However, it should be noted that these scoring approaches cannot fully mitigate bias from non-effortful responding, particularly when RG is related to the underlying ability of examinees (i.e., data treated as missing/not-administered are non-ignorable; see Rios & Soland, 2021a).

Influence of RG on measurement properties and test performance results

Regardless of the identification and filtering approaches utilized, RG has been documented via both simulation and applied analyses to bias estimates of item and measurement properties, such as: (a) item difficulty and discrimination (e.g., Rios & Soland, 2021b); (b) item and test information (e.g., van Barnevald, 2007); (c) measurement invariance (e.g., Rios, 2021a); and (d) linking coefficients (Mittelhaëuser et al., 2015). In terms of aggregate-level score-based inferences, simulation research has demonstrated that mean performance (based on sum scoring) can be underestimated by more than 0.20 standard deviations (*SD*s) when RG responses comprise as little as 6% of total item responses (Rios et al., 2017). Given that effort (as measured by the absence of RG) has been shown to be highly related to achievement (r = .72; Silm et al., 2020), some researchers have found RG to underestimate test performance results to a non-negligible extent across a multitude of testing contexts (e.g., DeMars et al., 2013; Osborne & Blanchard, 2011; Wise & DeMars, 2010). However, others, such as Wise et al. (2020), have documented and argued that RG has minimal impact on estimates of group-based performance.

These contradictory findings can be attributed to a number of factors. First, RG is sample dependent, as prior research has found that RG behavior differs based on examinee demographic characteristics (e.g., Rios & Guo, 2020). Second, RG has been shown to be associated with various item characteristics (e.g., item length, complexity, position), suggesting that the degree of RG observed may be dependent on assessment features (e.g., Guo et al., 2022; Rios & Guo, 2020; Wise et al., 2009). Third, some researchers have proposed that RG is ability dependent – primarily lower ability examinees engage in RG (e.g., Rios et al., 2017). This would suggest minimal bias in test scores given that examinees engaging in RG possess a low true probability of successfully answering items in which RG is employed. Finally, differences between studies may reflect the filtering approaches employed and the degree to which observed data meet the assumption that RG is unrelated to examinee ability. The contradictory findings in the literature concerning the influence of RG on test score distortion suggests that such a relationship is not automatic (i.e., RG is not always associated with significant test underperformance) and instead may be based on a number of contextual factors.

Study rationale and objective

The need to better understand the association between RG and test performance results, as well as the contextual factors that influence this relationship calls for a meta-analytic investigation; however, to date, such an analysis has not been conducted. While few would be surprised to find that RG distorts score estimates in operational settings, the typical degree of distortion is less certain and has varied widely across prior research. In fact, the only meta-analysis that has attempted to quantify the influence of motivational issues on test scores was conducted by Wise and DeMars (2005). In their study, the authors compared performance between examinee subgroups administered the same test under high- and low-stakes testing contexts. Across 12 empirical studies and 25 effect sizes, they found that examinees in the high-stakes condition scored 0.59 standard deviations higher than their low-stakes counterparts.

This finding is often miscited as the average degree of underperformance expected due to low test-taking effort. However, such an interpretation is limited in that the meta-analysis compared performance by test-stakes, which does not generalize to estimated test underperformance due to RG in practice. This lack of generalization is present because examinees: (a) cannot be randomly assigned to test-stake conditions in most operational settings; and (b) in naturalistic settings, may be unaware of the test-stakes prior to engaging with the assessment (e.g., when repurposing accountability test results to make individual instructional decisions; see Rios & Soland, 2021a). Furthermore, given that examinee employment of RG may fluctuate throughout a test, due to item (e.g., item difficulty) or test (e.g., test length) characteristics (Wise, 2017), it is unclear how assessment type, examinee

characteristics, or incentive type in the high-stakes testing contexts impacted the results. This is particularly an issue as Wise and DeMars (2005) did not investigate moderating factors. Due to these limitations, it is unclear the extent to which RG distorts aggregated test scores.

To address this limitation, the objective of this study is to conduct a meta-analytic review of primary studies that investigated the influence of RG on test performance results for low-stakes group-administered cognitive assessments. Although limited in that the true influence of RG is unknowable based on the use of an indirect proxy (RT) of examinee behavior, such an approach provides the best estimate of score distortion in practice. In saying that, the intended objective of this study is addressed via the following research questions:

(1) What is the average percentage of examinees that engage in and responses classified as RG?
(2) What is the mean distortion of aggregated scores associated with RG?
(3) How is the association between RG and test performance results moderated by publication type, participant age, test content, test length, and RG classification and filtering procedures?

Results from this meta-analysis have the potential to inform practitioners of both the extent to which RG can distort group level test performance results and the contextual factors that drive this relationship.

Method

Search strategy

Four search strategies were employed to identify potential studies for inclusion: (a) academic database searches; (b) Internet browsing; (c) backward and forward citation searches; and (d) expert consultation. Data collection was completed on September 4, 2020 by the second and third authors. A detailed description of each search strategy, in the order conducted, is provided below.

Academic database search

The following academic databases were searched to identify relevant studies: PsycINFO (via Ovid); ERIC (via EBSCOhost); Education Source (via EBSCOhost); and Academic Search Premier (via EBSCOhost). "Rapid guess" and "response time" were entered as key terms with the Boolean operator "AND" and the Boolean modifier asterisk ("*"). The "AND" was applied between the two keywords to help narrow the search and improve precision. The asterisk ("*") was utilized right after the "Rapid guess" term to include multiple formats of the word "guess" (e.g., guess, guesses, guessing, and guessed). Additionally, only studies published in the English language were included. No other initial restrictions were placed on this search.

Internet browsing

The next approach employed was an Internet search using Google Scholar. The key terms employed in the Internet search were identical to those used in the academic database search. Results produced in Google Scholar were sorted by relevance. However, only the first 660 of the 1,000 accessible articles were included for review, given the decreasing positive hit rate.[2]

Citation searching

Beyond academic database and Internet browsing, backward and forward citation searching were also employed. The backward citation search was composed of two parts. The first consisted of searching the reference lists of two pertinent review articles (Silm et al., 2020; Wise, 2017) identified by the first

[2]For example, only 12 items (approximately 3.3%) were potentially relevant to the topic of current review from the 301st to the 660th result. Due to this low hit rate, our search consisted of the first 660 results.

author. The latter part involved searching the reference list of all articles found to meet our eligibility criteria (described below) from the academic database and Internet searches. Both backward citation searches were conducted using the Social Sciences Citation Index and Google Scholar. All studies found using backward citation searching were then evaluated based on the eligibility criteria, and if accepted, were included for additional referencing.

Upon completing the backward citation search, forward citation searching (i.e., searching for studies that cited the manuscript of interest) was employed to examine studies that met the eligibility criteria from the search strategies noted above. To this end, the title for the study of interest was typed into Google Scholar, and the *cited by* link was clicked on. Any study included from this strategy (based on meeting the eligibility criteria) also underwent backward and forward citation searching. This process was repeated until no new articles met the eligibility criteria. Both citation search strategies were completed by September 4, 2020.

Expert consultation

This search strategy was conducted by directly emailing researchers known to have conducted work and/or published extensively on the topic of rapid guessing. Each individual was contacted to ascertain whether they had conducted unpublished research that met our inclusion criteria and/or knew of such research authored by others. A total of six researchers, with academic and industry affiliations within the United States and Europe, were emailed.

Eligibility criteria

To further elucidate the rationale for article inclusion, we describe the eligibility criteria along three dimensions: (a) data type; (b) RG identification methodology; and (c) outcome measures.

Data type

To be included in this meta-analysis, studies had to use empirical data in studying RG responses. Either primary or secondary data collected from a group-administered, low-stakes cognitive assessment (i.e., assessments of maximum performance) were acceptable. No further restrictions were placed on examinee (e.g., age, country of origin, ethnicity, language) nor assessment (e.g., length, item types) characteristics.

RG identification methodology

Only studies utilizing RT as a proxy of RG were included, given the fact that this procedure can identify RG on an item-by-item basis. Primary studies could employ any one of the RT threshold methods outlined by Wise (2017) as they are the most popular in practice (Silm et al., 2020).[3]

Outcomes

The outcome measures of interest were: (a) percentage of examinees engaging in RG; (b) percentage of item responses classified as RG; and (c) differences between unfiltered and filtered data on group test performance results. The first two variables were included for descriptive purposes to answer our first

[3]We acknowledge that there are other approaches for identifying and accounting for RG. As an example, research has investigated the utility of employing mixture models to mitigating bias from RG via simulations and empirical examples. Although these models have a similar purpose to the methods covered in this meta-analysis (i.e., to identify and downweight RG responses), they have received minimal applications in operational settings. This is largely due to their computational demands, sample size requirements, model convergence issues, and assumptions concerning underlying response time distributions, which if untenable, can lead to incorrect inferences (Molenaar, Bolsinova, & Vermunt, 2018). With that said, the methods examined in this paper possess their own assumptions. For example, similar to mixture models, they assume that RG can be effectively distinguished based on response time distributions, which if untenable, can lead to biased item and ability estimates (Rios, 2022a). Although both approaches have their own limitations, the focus in this study is on threshold methods that are more widely utilized in applied contexts (see Rios & Deng, 2021). For those interested in learning more about flexible mixture modeling approaches that employ computationally efficient parameter estimation, see Nagy and Ulitzsch (2021).

research question, and were not included in the moderator analyses. In regards to (a), we first coded for the proportion of examinees who engaged in any amount of RG, and the proportion who were identified as rapid guessers (i.e., employing RG on a non-negligible number of items) by the primary authors. Concerning the latter, we noted the response time effort criteria (RTE; i.e., the proportion of non-rapid guessing responses) used for classifying rapid guessers. Next, we coded for the proportion of item responses classified as RG in the total sample as well as the average RTE across examinees. Outcome (c) was operationalized as a mean raw, scale, or theta estimate reported for a total or subsample. For this variable, a study must have presented any test statistic (e.g., χ^2, Z, t, F, $\hat{p}$, r) necessary for computing a standardized difference effect size for group-level test performance. Studies that only compared participants' performance between unfiltered and filtered data at the item-level were excluded (e.g., Şahin, 2017).

Variable coding

Within each study, four levels of variables were coded: (a) study; (b) sample; (c) assessment; and (d) threshold and filtering procedures. Below we discuss the variables coded and the rationale for their inclusion.

Study variables

For each study, we coded for the publication year and noted whether the study was published in a peer-reviewed journal or in some other venue (e.g., a dissertation; hereon referred to as gray literature). Publication type is an important moderator for any meta-analysis, as publication bias, or the tendency for journals to more likely publish significant findings, has been shown to influence meta-analytic results if not properly addressed (Lin & Chu, 2018).

Sample variables

Five demographic characteristics were considered: (a) sample age; (b) grade; (c) sex; (d) ethnicity; and (e) nationality. Of these, only sample grade was considered for the moderator analysis, as older examinees have been shown to employ RG at higher rates than their younger counterparts (e.g., Rios & Guo, 2020). The remaining variables are presented descriptively to provide a summary of the sample characteristics examined in the literature. If a paper included (a) multiple studies or (b) multiple samples within the same study, the demographic variables were coded separately for each sample/study.

Assessment variables

Three assessment variables, drawn from Rios (2021b), were coded for the moderator analysis. First, test content was coded dichotomously as science, technology, math, and engineering (STEM) or non-STEM (the latter served as the reference); assessments that included both content areas were coded as STEM. This variable was included as prior research has shown that students reported lower effort for STEM assessments when compared to non-STEM assessments (Sundre, 1997).

Next, we coded for item type categorically as (a) selected response; (b) open-ended; or (c) a combination of the prior. In addition, both the language of the test (dichotomously coded as English or other languages or mixed) and the name of the assessment investigated were coded for descriptive purposes. Lastly, test length was coded as the total number of items on the assessment. This variable was included because prior research has shown that RG increases as the length of an assessment grows (Wise & Kingsbury, 2016). Of these variables, only content area and test length were included as moderator variables.

Threshold and filtering variables

Two variables related to the identification of RG were included. The first investigated the RT threshold procedure employed to classify RG. Specifically, procedures were dichotomized as either (a) empirically-based (i.e., utilizing item response accuracy and/or time data to establish RT thresholds) or (b) non-empirically-based (i.e., establishing thresholds without the use of empirical data; for more detailed information on these procedures, refer to Wise, 2017). It is hypothesized that the degree of distortion observed may on average be lower for non-empirically-based procedures, given that these procedures tend to classify RG more conservatively (Wise, 2017).

The second moderator variable included was based on the type of filtering procedure employed, which included either examinee- or response-level filtering. The former approach listwise deletes data for examinees engaging in RG at a rate that exceeds an a-priori criterion, while the latter treats any RG response as missing (i.e., downweights the contribution of a RG to zero for ability estimation; the former served as the reference). Rios et al. (2017) have shown that observed differences between filtering approaches may be present when RG is related to the underlying ability of examinees.

Interrater reliability

To examine interrater reliability, the second and third authors each coded 20% of the total studies included. For each study, interrater agreement was calculated separately for over 50 descriptive, moderator, and effect size variables. An interrater agreement value of 0.80 was set as the criterion for establishing rater consistency. Across studies, the overall average percent agreement was 93.7%, with agreement equal to 90.25% for the effect size variables. Any rater disagreements were settled through discussion between the second and third authors.

Effect size calculations

Upon coding the means and standard deviations of group test performance or theta parameter estimates for each study, standardized mean difference effect sizes were first computed for the test performance variable based on Cohen's *d* formula. These effect sizes were then converted using Hedges' *g* formula to account for potential overestimation of effect sizes due to small samples. These calculations were completed in *R* version 4.0.0 (R Core Team, 2020).

Publication bias and outlier analysis

Diagnostic analyses of publication bias and outliers were conducted prior to proceeding with the main effect and moderator analyses. Given that studies with statistically significant results are more likely to be published, one concern of meta-analytic research is that results may be biased by failing to include unpublished literature (Greenhouse & Iyengar, 2009). Although a significant effort was made in searching for gray literature that met the eligibility criteria of the current study, it was still necessary to evaluate the potential presence of publication bias. Two methods, Egger's test and the trim-and-fill method, were employed to test for publication bias. Beyond publication bias, the funnel plot was used to identify outliers (i.e., any effect size more than three standard deviations above or below the mean effect size). Any effect size deemed to be an outlier was downweighted to be equal to three standard deviations from the mean effect size.

Average effect size and heterogeneity

To calculate the average effect size and heterogeneity, an intercept-only meta-regression model was fit in the *robumeta* package in *R* (Fisher et al., 2016). However, as some studies produced multiple effect sizes (e.g., Rios et al., 2017), there was a concern of effect size dependencies, which are associated with

biased variance estimates (Matt & Cook, 2009). To mitigate this potential bias, the robust variance estimation (RVE) procedure was implemented, which accounts for nested effect sizes in standard error calculations.

The heterogeneity of effect sizes was evaluated based on the I^2 statistic, which represents the proportion of variation in an effect size estimate not due to chance. This statistic is interpreted as follows: if $I^2 > 75\%$, the sample is said to have high heterogeneity, $75\% > I^2 > 50\%$ reflects medium heterogeneity, and a value below 50% is considered low heterogeneity (Higgins & Thompson, 2002). Before proceeding with a moderator analysis, we first checked to see if the sample possessed medium to high heterogeneity, as this would indicate that there is significant true variability across effect sizes that should be further investigated. If not, the moderator analysis was deemed unwarranted, due to a lack of heterogeneity in the sample.

Moderator analysis

The following model was selected to include the most relevant moderating variables:

$$Y = \beta_0 + \beta_1\ (Sample\,Grade) + \beta_2\ (Assessment\,Content) + \beta_3(Assessment\,Content) + \beta_4\ (Threshold\,Type) + \beta_5\ (Filter\,Type) + \beta_6\ (Publication\,Type) + e, \quad (1)$$

where Y is the predicted Hedges' g for test performance, β_0 is the average effect size after controlling for all other moderator variables, and e is the residual error term. Each of the other coefficients are based on the coding procedure previously described in the *variable coding* section.

Results

Overall, 959 studies were reviewed based on academic database, Internet browsing, expert consultation, and backward and forward citation searching. Among these 959 studies, 25 were found to meet the eligibility criteria (2.5% hit rate; see Appendix A of the supplementary file). Across outcomes, all studies included in the sample were written after 2002 and more than half appeared in the last decade (2010–2020; 15 out of 25; 60%). Most of the included studies were published in peer-reviewed journals (18 out of 25; 72%), with the two most popular outlets, Applied Measurement in Education (6 out of 18, 33%) and Educational Assessment (3 out of 18, 17%), accounting for 50% of the non-gray literature publications.[4] The remaining manuscripts did not appear in peer-reviewed journals (i.e., were gray literature) at the time of coding and represented works in progress (n = 2), dissertations or theses (n = 3), and book chapters (n = 2). All first authors had U.S. institutional affiliations at the time of publishing, with Dr. Steven Wise contributing 36% of the articles included in the sample. Appendix B of the supplementary file presents the descriptive statistics for sample, methodological, assessment, and publication characteristics of all included studies.

Average percentage of rapid guessers and rapid guessing responses

The 25 included studies contained effect sizes for 39 distinct samples based on a total sample size of 443,264 examinees (the average size for each distinct sample was 11,366). Across studies, an average of 28.3% of examinees were found to engage in RG for at least one item, while 6.9% of all item responses were identified as RG responses. RG was found to be classified primarily by empirically based methods (79% of classification procedures sampled), with the most common threshold procedures being RT distribution (n = 24), percentile (n = 20), and a mixture of multiple thresholds (n = 20). As expected, the percentage of RG responses identified varied greatly by response time threshold procedures. For

[4]The remaining publications were from the Journal of Educational Measurement (n = 2), International Journal of Testing (n = 2), Educational and Psychological Measurement (n =1), New Directions for Institutional Research (n =1), Library & Information Science Research (n =1), Journal of Research on Educational Effectiveness (n =1), and Large-scale Assessments in Education (n =1).

instance, percentile procedures (i.e., establishing the RG response time threshold based on a percentile of the response time distribution; 6.24%) classified RG responses at three times the rate of common-k second threshold methods (i.e., a response provided in less than three seconds is reflective of RG; 2.68%). However, minimal differences were noted across grade (K–12 = 5.14%; postsecondary = 6.69%) and test content (STEM/mixed = 5.68%; non-STEM = 6.83%) variables. Additionally, of examinees that were found to engage in RG, 21% (*SD* = 29%) were deemed to rapid guess on a nonnegligible number of items as determined by a-priori guidelines established by primary authors. For the 10 studies that provided information on the criterion utilized, 60% classified the extent of RG as nonnegligible if an examinee employed RG for a minimum of 10% of items, while 20% of studies utilized a criterion of 20%.

Average effect size and heterogeneity for test performance outcome

Twenty of the 25 included studies examined test performance as an outcome, resulting in 100 effect sizes (with each study producing 5 effect sizes on average; *SD* = 6) based on a total sample size of 131,808 examinees. Prior to calculating the average effect size and heterogeneity for this outcome, outlier and publication bias analyses were conducted. Concerning the former analysis, only one test performance effect size was classified as an outlier. To mitigate its influence, this effect size was downweighted to three standard deviations from the mean. In regard to publication bias, Eggers' test yielded a significant *p*-value ($p > .9$) indicating asymmetry of the effect size distribution, which was potentially attributable to publication bias. However, a visual inspection of the distribution suggested that this asymmetry was likely not indicative of publication bias, which is typically characterized by studies with low/negative effect sizes and high variance (the funnel plot is provided in Appendix C of the supplementary file).

Upon downweighting outliers and ensuring that publication bias was minimal, we next calculated the mean effect size and heterogeneity across all studies via an intercept-only meta-regression model. Results from this model indicated that on average RG distorted aggregated test scores by −0.13 *SD*s ($p < .001$; 95% CI [0.08, 0.18]). Descriptive analyses showed that 95% of effect sizes were negatively distorted, while the remaining 5% showed cases of positive score distortion (i.e., test performance results were inflated due to RG). Across studies that investigated the test performance outcome, on average 18.8% of all examinees were categorized as engaging in non-negligible degrees of RG by primary authors based on a-priori criteria, and 7.4% of all item responses were identified as RG responses. The degree of heterogeneity across the effect size distribution was deemed to be large via the I^2 statistic (92.25%), suggesting the need for a moderator analysis to explain the heterogeneity (Table 1).

Moderator analysis

Below we describe the difference in effect sizes attributable to moderators by sample, assessment, and filtering characteristics, once controlling for publication type, which was a non-significant moderator ($p = .45$). The moderators included accounted for over 21% of variance in effect sizes as shown in the model results provided in Table 1.

Sample characteristics

The 20 included studies contained effect sizes for 33 distinct samples, with an average sample size of 4,436 examinees. Although only reported 75.8% of the time, samples were on average fairly even in terms of gender representation (50.8% female). The median participant age was 19 years old, with the majority of samples comprised of post-secondary examinees (22 out of 33; 66.67%). After controlling for all other moderators, our analysis indicated that the influence of RG on test performance results did not differ between higher education and K–12 examinees ($p = .28$).

Table 1. Differences in Effect Sizes Attributable to Moderators.

	Null Model (k = 20; n = 100)I^2 = 92.25			Moderator Model (k =20; n = 100)I^2 = 70.96		
Moderator	Estimate	S.E.	95% CI	Estimate	S.E.	95% CI
Intercept[a]	0.13	0.02	0.08, 0.18	0.16	0.04	0.06, 0.26
Participant Characteristics						
Age[b]	–	–	–	0.04	0.03	−0.04, 0.11
Assessment Design						
Assessment Length	–	–	–	−0.00	0.00	−0.00, 0.00
Test Content[c]	–	–	–	0.08*	0.03	0.02, 0.13
Methodology						
Filter Type[d]	–	–	–	−0.16**	0.03	−0.22, −0.09
Threshold Type[e]	–	–	–	−0.05	0.03	−0.13, 0.03
Publication Type[f]						
Grey Literature	–	–	–	0.02	0.03	−0.05, 0.09

*$p < .05$; **$p < .01$

Robust standard errors are provided in parentheses.

[a]The intercept is interpreted as the average effect size for independent samples that are published as gray literature, included a K-12 sample, a non-stem assessment with 0 items, employed a non-empirical threshold procedure and examinee-level filtering.

[b]*Age* was dichotomously coded (reference was K-12).

[c]*Test content* was dichotomously coded (reference was STEM or mixed).

[d]Filter type was dichotomously coded (reference was examinee-level).

[e]Threshold type was dichotomously coded (reference was non-empirical).

[f]*Gray literature* was dichotomously coded (reference was peer-reviewed).

Based on a prior hypotheses, a one-tailed statistical test was employed for the following moderators: (a) age, (b) assessment length, (c) test content, (d) threshold type, (e) filter type, and (f) gray literature.

Assessment characteristics

The measurement instruments included in the sample ranged greatly in terms of the number of items included (ranging from 26 to 108 items, M = 40.75 items, SD = 16.63) and content area assessed (52.83% STEM or mixed); however, no variability in item type was observed, given that every measure contained only selected response items. The Measures of Academic progress assessment was the most commonly studied assessment across samples (12 out of 53; 22.64%), although many other instruments, such as PISA (n = 5), HEIghten Critical Thinking (n = 4), the Natural World Test (n = 3), and the Information Literacy Test (n = 3), among others were included. Nearly all test administrations were specified to be in English (93.6%).

Model results did not show a significant decrease in test performance results as the number of items on an assessment increased (p= .29). However, when comparing STEM and non-STEM assessments, the latter type was found to be associated with greater distortion of test performance results by −0.08 *SD*s (p= .02, 95% CI [0.02, 0.13]; Figure 1).

Filtering procedure

Across studies, the majority of effect sizes came from thresholds relying on empirical methods (n = 78). Specifically, the most common RT threshold approach was based on inspecting RT distributions (n = 26). Common k-second (n = 14), percentile (n = 11), a combination of RT and accuracy (n = 9), and surface feature methods (n = 8) were less common; however, it should be noted that 32 effect sizes came from employing a combination of multiple procedures. In investigating whether threshold type impacted the distortion of aggregated test scores, no significant differences were found between empirical and non-empirical threshold types (p = .13).

Once defining RG responses based on a given threshold approach, practitioners are next confronted with how to filter such responses. Our results showed that of the 100 effect sizes in the sample, 38% were based on examinee-level filtering, while the remaining were calculated using response-level filtering. An examination of filtering type as a moderator demonstrated

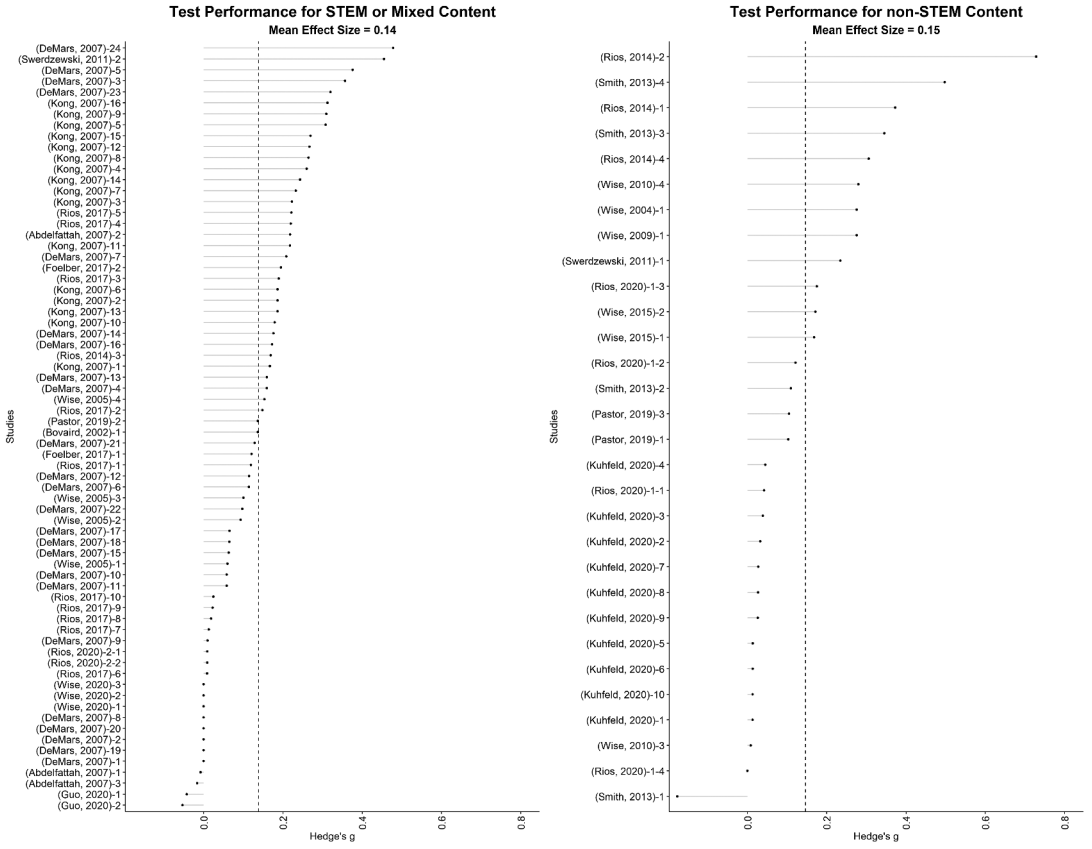

Figure 1. Effect Sizes for Test Performance Distortion Grouped by Content Area. Note.The figures above provide the reader with the effect size distributions disaggregated by test content area. The presented average effect sizes do not account for other moderators.

that after controlling for all other variables, the influence of examinee-level filtering was associated with an effect size that was 0.16 *SD*s higher than response-level filtering (p= .001, 95% CI [−.22, −.09]; Figure 2).

Discussion

The objective of the present meta-analysis was to quantify the average degree of aggregated test score distortion related to RG. Overall, across the studies sampled, on average approximately 28% of examinees were found to employ RG on 7% of responses. These percentages were associated with an average distortion of −0.13 *SD*s, which is nearly equivalent to one-half a year reduction in the average annual change score in science for K–12 students in the United States (0.29 *SD*s; Bloom et al., 2008). Furthermore, it is greater than the performance differences in achievement shown between students whose family receives social welfare resources/government aid and those that do not (−0.12; Visible Learning, 2018). The magnitude of score distortion observed supports concerns raised by prior researchers in how RG can undermine the validity of inferences in areas such as achievement gains (e.g., Wise & DeMars, 2010), treatment effects (e.g., Osborne & Blanchard, 2011), teacher quality evaluations (e.g., Williams, 2015), and subgroup differences (e.g., Rios, 2021a).

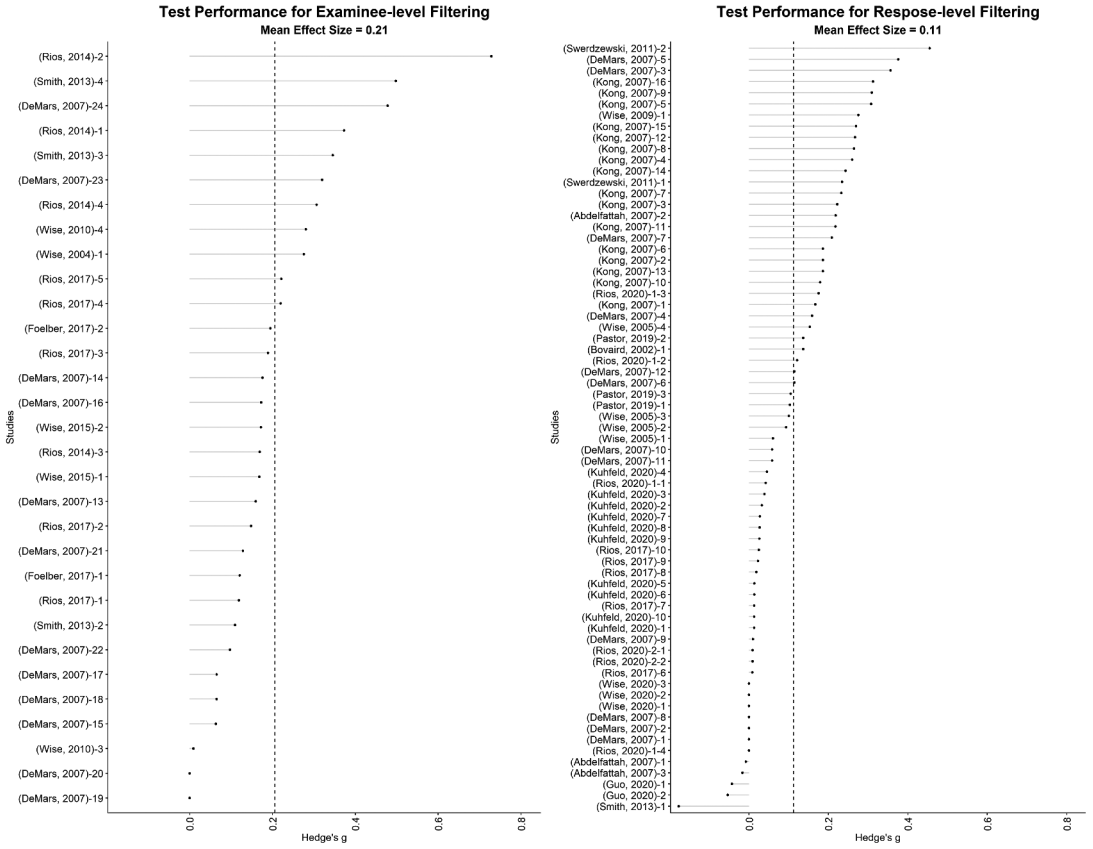

Figure 2. Effect Sizes for Test Performance Distortion Grouped by Filtering Procedure. Note. The figures above provide the reader with the effect size distributions disaggregated by test filtering procedure. The presented average effect sizes represented by the vertical dotted lines do not account for other moderators.

Both test content and filtering procedure moderated the association between RG and score distortion after controlling for other sample, assessment, and methodological characteristics. In terms of test content, RG on non-STEM assessments was associated with greater score distortion. This finding was surprising as students have been shown to engage in higher rates of RG for STEM assessments, due to a perceived lack of content knowledge and interest (Sundre, 1997). One potential hypothesis for this counterintuitive finding is that test content may be confounded with other assessment characteristics, such as item difficulty. Specifically, the degree of score distortion is driven by the difference between an item's weighted difficulty and the true ability of an examinee engaging in RG (i.e., distortion will be greater when examinees engage in RG for items in which they have a high probability of successfully answering; Rios et al., 2017; Rios & Soland, 2021b). Thus, if non-STEM tests were on average easier, lower rates of RG could have higher potential score distortion compared to STEM assessments. Unfortunately, this hypothesis could not be investigated in the current meta-analysis, as primary authors generally did not include assessment details beyond item type.

A second significant moderating effect indicated that the filtering procedure employed influenced the degree of score distortion observed. Specifically, the average degree of distortion was higher by 0.16 *SD*s for examinee- versus response-level filtering. According to simulation research conducted by Rios et al. (2017), differences in filtering procedures are only expected when examinees with predominantly low true abilities engage in RG. The reason for this finding is that listwise removal of data from low

ability rapid guessers artificially inflates group performance, whereas response-level filtering is more robust to such score distortions (Rios et al., 2017). Although minimal information was provided by primary authors on the ability characteristics of rapid guessers, it is possible that the observed differences in filtering procedures is a proxy for ability differences between effortful and RG subgroups.

Limitations and directions for future research

A number of limitations associated with this study should be noted. First, although a concerted effort was made to include a diverse set of literature search strategies, some pertinent papers may have been missed based on only including English language manuscripts and failing to consult professional research organization listservs.

Second, the reported degree of test score distortion is an estimate limited by the procedures employed by primary authors to identify and classify RG behavior. For instance, if primary studies employed response time thresholds that were too strict, RG may have been underidentified (Wise & Kuhfeld, 2020), leading to potentially understated score distortion. Our descriptive results did show large variations in the percentage of RG responses identified by response time threshold procedures, supporting prior research (Rios & Deng, 2021) and suggesting that classification procedures may have influenced observations of score distortion. Although researchers have utilized multiple proxies to identify RG (e.g., Harmes & Wise, 2016; Şahin & Colvin, 2020), they are all indirect indicators of examinee behaviors that make strong assumptions to infer the occurrence of RG behavior. Given that one can never truly know whether an examinee is engaging in full or partial effortful responding, the results from this paper should be interpreted as an estimate of score distortion that is dependent on how primary authors classified RG, and thus, may not accurately reflect true bias.

In addition, given a large degree of missing information, we were unable to include all moderators of interest in our analyses. Two unavailable moderators that should be examined in future investigations are test difficulty and administration procedures. Concerning the former, prior research has shown that the influence of RG on aggregated test scores is generally greater when average item difficulty is low (Rios et al., 2017). Thus, it is likely that the variability observed is associated with not only the rate of RG, but also test difficulty. This should be considered when interpreting findings in operational settings.

One context of test administration procedures that could be important to investigate in the future is whether individual feedback is promised as an incentive for examinees. Previous research findings suggest that individuals' intrinsic motivation can be increased by providing performance-contingent feedback that is perceived to be informative of competence (Ryan et al.,1983). Therefore, although a testing context may be low-stakes for individual examinees, their desire to obtain accurate performance feedback may increase test-taking effort. In such contexts, we may observe that RG's influence on test performance results may be greatly mitigated, which may explain variance in the effect sizes observed in our sample; however, as noted, we were unable to investigate this hypothesis due to limited information on assessment contexts provided by primary authors.

Implications

Although the *Standards for Educational and Psychological Testing* stipulates that "the degree of motivation of test takers" should be investigated to determine the validity of test score interpretations (American Educational Research Association et al., 2014, p. 213), there is little evidence that operational testing programs heed this advice (see Wise & Kuhfeld, 2020). Our results provide clear empirical evidence that RG is a concern for a variety of low-stakes testing contexts that differ by content area, examinee population, and settings. Specifically, we show that across 25 empirical studies the average percentage of examinees engaging in some form of RG is over one-quarter of test-takers, resulting in an approximate average of 7% RG responses and a negative score distortion of 0.13 *SD*s. These results have important implications for both research and practice.

From a research perspective, the descriptive findings presented can influence methodological work on the topic of RG. As an example, past simulation studies have investigated RG in contexts that vary the total percentage of simulees engaging in RG to be as high as 40% with RG responses comprising as much as 20% of all responses (e.g., Rios et al., 2017). Such simulation work perhaps has overstated the influence of RG on aggregated score inferences based on simulating contexts that may not best reflect actual examinee behavior in practice. Thus, results from this meta-analysis may not only assist researchers in better interpreting the plausibility of past simulation work, but it can also guide the future development of simulation studies by assisting researchers in choosing appropriate parameters that better reflect RG in reality.

In addition, the observed average rate of rapid guessers suggests that this construct-irrelevant behavior is a non-trivial concern amongst examinees and points out the need for assessment and testing specialists to consider employing interventions to improve examinee effort in such testing contexts. Wise and DeMars (2005) have noted that there are four broad categories of strategies to do so, which include: (a) increasing test relevance; (b) altering the test administration process (e.g., having proctors remind students to employ their maximal effort); (c) promising feedback; (d) offering external incentives. In a recent meta-analysis investigating these strategies, Rios (2021b) found that across 60 effect sizes, interventions improved test-taking effort (both self-reported and based on RT) by an average of 0.13 *SD*s (the same magnitude of negative score distortion found in this study), with external incentive interventions positively improving effort by an average of 0.37 *SD*s. Further work is needed to understand the generalizability of these interventions across varying examinee populations and assessment contexts; however, initial evidence suggests that they may be an effective solution to improving test-taking effort.

As interventions may not fully mitigate RG for all examinees, effective approaches to identifying RG and filtering/modeling such responses is needed. Given that computer-based test administrations continue to increase, one of the most popular proxies of RG is response time. However, to date, it is assumed that response time threshold methods can accurately classify RG responses. A failure to meet this assumption will lead to the inclusion of psychometrically uninformative and biased information (Rios, 2022b). Thus, foundational research that incorporates multiple sources of evidence, such as eye-tracking, cognitive interviews, self-report measures, facial emotions, etc., is needed to support the validity of employing response times as a proxy of effort. There is also an opportunity for practitioners to utilize log-file information (e.g., RT, number of mouse clicks) that goes beyond response times as proxies of test-taking engagement. For instance, these proxies could be leveraged by establishing solution behavior progressions for each item in which typical behaviors of engaged examinees would be evaluated. For examples of utilizing multiple sources of log-file information to gauge examinee effort, the reader is referred to Harmes and Wise (2016) and Şahin and Colvin (2020).

Once identifying RG responses, deciding on how to best filter or model such responses is needed. To date, many researchers are still utilizing listwise deletion of data from examinees employing RG. The use of listwise deletion in practice is highly discouraged as it can lead to significant bias in estimation when RG responders are typically of low ability – a situation that can occur in some applied contexts (e.g., Deribo et al., 2021; Rios et al., 2017). Although response-level filtering may be a more robust option, it stills assumes that RG responses are ignorable (i.e., non-RG responses are a random sample of all true item responses). Under high rates of RG, large bias can be observed when treating RG responses that are related to examinee ability as missing/not-administered (see Rios & Soland, 2021b). Thus, researchers/practitioners should evaluate the ignorability assumption prior to employing either examinee- or response-level filtering. If this assumption is untenable, practitioners may benefit from adopting alternative approaches that leverage IRT to downweight RG responses, such as robust likelihood estimation (Rios, 2022a) and multidimensional IRT models that account for the covariance between examinee proficiency and engagement propensity (e.g., Deribo et al., 2021; Liu et al., 2019). However, more research is needed to examine the robustness of these procedures under varying RG

contexts. Viable solutions will need to consider the readily available capabilities and resources that practitioners have at hand. A failure to consider these factors will likely lead to minimal operational adoption.

Regardless of the filtering/modeling approach employed, there have been minimal investigations exploring how best to incorporate RG information into score reporting. A likely reason for this is that few operational testing programs evaluate examinee test-taking effort. However, once doing so, ensuring that score users are aware that measurement specialists have made efforts to mitigate the potential deleterious effects of RG is critical to strengthening the perceived credibility of the assessment results. Identifying effective communication methods about employment of this construct-irrelevant behavior will require piloting score report designs with score users via cognitive interviewing to confirm that the information is both accessible and comprehensible. Effective communication is of particular importance given concerns about the lack of assessment literacy amongst educational practitioners and stakeholders (Popham, 2009). Although there is much work to be done to mitigate, identify, filter/model, and communicate issues around RG behavior to score users, it is our hope that this meta-analysis has helped to both quantify the potential distortion that RG presents and laid the foundation for future research.

Author contribution statement

The first author conceived of the presented idea, designed the sampling and analytic approaches employed, interpreted findings, and wrote the majority of the article. The second and third authors conducted the literature searches, extracted and coded variable information, conducted analyses, and contributed to writing. All authors conducted critical revisions of the article throughout the review process and approved of the final version to be published.

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

Joseph A. Rios http://orcid.org/0000-0002-1004-9946
Jiayi Deng http://orcid.org/0000-0002-1962-2956
Samuel D. Ihlenfeldt http://orcid.org/0000-0003-3642-8407

References

References marked with an asterisk indicate studies included in the meta-analysis

*Abdelfattah, F. A. (2007). *Response latency effects on classical and item response theory parameters using different scoring procedures* [Unpublished doctoral dissertation]. Ohio University.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing* (6th ed.). Washington D.C.: American Educational Research Association.

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. doi:10.1080/19345740802400072

*Bovaird, J. A. (2002). *New applications in testing: Using response time to increase the construct validity of a latent trait estimate* [Unpublished doctoral dissertation]. The University of Kansas.

*DeMars, C. E. (2007). Changes in rapid-guessing behavior over a series of assessments. *Educational Assessment*, *12*(1), 23–45. doi:10.1080/10627190709336946

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.

Deribo, T., Kroehne, U., & Goldhammer, F. (2021). Model-based treatment of rapid guessing. *Journal of Educational Measurement*, *58*(2), 281–303. doi:10.1111/jedm.12290

Fisher, Z., Tipton, E., & Hou, Z. (2016). Package "Robumeta: Robust variance meta-regression" (R package version 1.8) [Computer software]. Accessed 10 September, 2020. Retrieved from https://cran.r-project.org/web/packages/robumeta/robumeta.pdf

*Foelber, K. J. (2017). *Using multiple imputation to mitigate the effects of low examinee motivation on estimates of student learning* [Unpublished doctoral dissertation]. JamesMadison University

Greenhouse, J. B., & Iyengar, S. (2009). Sensitivity analysis and diagnostics. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 103–126). New York, NY: Russell Sage Foundation.

*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. doi:10.1080/08957347.2016.1171766

*Guo, H., Rios, J. A., Ling, G., Wang, Z., Gu, L., & Liu, L. O. (2022). *Influence of selected-response variants on test characteristics and test-taking effort: An empirical study (ETS RR-22-01)*. Princeton, NJ: Educational Testing Service. https://onlinelibrary.wiley.com/doi/pdf/10.1002/ets2.12345

Harmes, J. C., & Wise, S. L. (2016). Assessing engagement during the online assessment of real-world skills. In Y. Rosen, S. Ferrara, & M. Mosharraf (Eds.), *Handbook of research on technology tools for real-world skill development* (pp. 804–823). Hershey, PA: IGI Global.

Higgins, J. P., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, *21*(11), 1539–1558. doi:10.1002/sim.1186

*Kong, X. J., Wise, S. L., & Bhola, D. S. (2007). Setting the response time threshold parameter to differentiate solution behavior from rapid-guessing behavior. *Educational and Psychological Measurement*, *67*(4), 606–619. doi:10.1177/0013164406294779

*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. doi:10.1080/19345747.2019.1636437

*Lee, Y. H., & Jia, Y. (2014). Using response time to investigate students' test-taking behaviors in a NAEP computer-based study. *Large-scale Assessments in Education*, *2*(1), 8. doi:10.1186/s40536-014-0008-1

Lin, L., & Chu, H. (2018). Quantifying publication bias in meta-analysis. *Biometrics*, *74*(3), 785–794. doi:10.1111/biom.12817

Liu, Y., Li, Z., Liu, H., & Luo, F. (2019). Modeling test-taking non-effort in MIRT models. *Frontiers in Psychology*. Retrieved 1 September, 2020, https://www.frontiersin.org/articles/10.3389/fpsyg.2019.00145/full

Matt, G. E., & Cook, T. D. (2009). Threats to the validity of generalized inferences. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 537–560). New York, NY: Russell Sage Foundation.

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*(3), 437–455. doi:10.1037/a0028085

Mittelhaëuser, M. A., Béguin, A. A., & Sijtsma, K. (2015). Selecting a data collection design for linking in educational measurement: Taking differential motivation into account. In R. Millsap, D. Bolt, L. van der Ark, & W. C. Wang (Eds.), *Quantitative psychology research. Springer proceedings in mathematics & statistics* (Vol. 89, pp. 181-193). Cham, Switzerland: Springer, Cham.

Molenaar, D., Bolsinova, M., & Vermunt, J. K. (2018). A semi-parametric within-subject mixture approach to the analyses of responses and response times. *British Journal of Mathematical and Statistical Psychology*, *71*(2), 205–228. doi:10.1111/bmsp.12117

Nagy, G., & Ulitzsch, E. (2021). A multilevel mixture IRT framework for modeling response times as predictors or indicators of response engagement in IRT models. *Educational and Psychological Measurement*. Advanced online publication. doi: 10.01316/44211045351.

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*, 1–7. doi:10.3389/fpsyg.2010.00220

*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. doi:10.1080/10627197.2019.1615373

Popham, W. J. (2009). Assessment literacy for teachers: Faddish or fundamental? *Theory into Practice*, *48*(1), 4–11. doi:10.1080/00405840802577536

R Core Team. (2020). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Accessed 12 September, 2020. https://www.R-project.org/

Rios, J. A. (2021a). Is differential noneffortful responding associated with type I error in measurement invariance testing? *Educational and Psychological Measurement*, *81*(5), 957–979. doi:10.1177/0013164421990429

*Rios, J. A. (2021b). Improving test-taking motivation in low-stakes group-based educational testing: A meta-analysis of interventions. *Applied Measurement in Education*, *34*(2), 85–106. doi:10.1080/08957347.2021.1890741

Rios, J. A. (2022a). A comparison of robust likelihood estimators to mitigate bias from rapid guessing. *Applied Psychological Measurement*, *46*(3) Advanced online publication, 236–249. 10.1177/01466216221084371.

Rios, J. A. (2022b). Assessing the accuracy of parameter estimates in the presence of rapid guessing misclassifications. *Educational and Psychological Measurement*, *82*(1), 122–125. doi:10.1177/00131644211003640

Rios, J. A., & Deng, J. (2021). Does the choice of response time threshold procedure substantially affect inferences concerning the identification and exclusion of rapid guessing responses? *A meta-analysis. Large-scale Assessments in Education*, *9*(1), 1–25.

Rios, J. A., & Soland, J. (2021a). Investigating the impact of noneffortful responses on individual-level scores: Can the effort-moderated IRT model serve as a solution? *Applied Psychological Measurement*, *45*(6), 391–406. doi:10.1177/01466216211013896

Rios, J. A., & Soland, J. (2021b). Parameter estimation accuracy of the effort-moderated item response theory model under multiple assumption violations. *Educational and Psychological Measurement*, *81*(3), 569–594. doi:10.1177/0013164420949896

Ryan, R. M., Mims, V., & Koestner, R. (1983). Relation of reward contingency and interpersonal context to intrinsic motivation: A review and test using cognitive evaluation theory. *Journal of Personality and Social Psychology*, *45*(4), 736–750. doi:10.1037/0022-3514.45.4.736

*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*, *33*(4), 263–279. doi: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. doi:10.1080/15305058.2016.1231193

*Rios, J. A., Liu, O. L., & Bridgeman, B. (2014). Identifying low-effort examinees on student learning outcomes assessment: A comparison of two approaches. *New Directions for Institutional Research*, *2014*(161), 69–82. doi:10.1002/ir.20068

Şahin, F. (2017). *Exploring validity of computer-based test scores with examinees' response behaviors and response times* [Unpublished doctoral dissertation]. University at Albany, State University of New York.

Şahin, F., & Colvin, K. F. (2020). Enhancing response time thresholds with response behaviors for detecting disengaged examinees. *Large-scale Assessments in Education*, *8*(5), 1–24. doi:10.1186/s40536-020-00082-1

Silm, G., Pedaste, M., & Täht, K. (2020). The relationship between performance and test-taking effort when measured with self-report or time-based instruments: A meta-analytic review. *Educational Research Review*, *31*, 100355. doi:10.1016/j.edurev.2020.100335

*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. doi:10.1016/j.lisr.2012.12.001

*Swerdzewski, P. J., Harmes, J. C., & Finney, S. J. (2011). Two approaches for identifying low- motivated students in a low-stakes assessment *context*. *Applied Measurement in Education*, *24*(2), 162–188. doi:10.1080/08957347.2011.555217

Sundre, D. L. (1997, March 24-28). *Differential examinee motivation and validity: A dangerous combination* [Paper presentation]. American Educational Research Association Annual Meeting, Chicago, IL, United States.

van Barnevald, C. (2007). The effect of examinee motivation on test construction within an IRT framework. *Applied Psychological Measurement*, *31*(1), 31–46. doi:10.1177/0146621606286206

Visible Learning (2018, March 28). *Hattie ranking: 252 influences and effect sizes related to student achievement*. https://visible-learning.org/hattie-ranking-influences-effect-sizes-learning-achievement/

Williams, L. M. (2015). *The effect of examinee motivation on value-added estimates* [unpublished doctoral dissertation]. James Madison University.

*Wise, S. L. (2006). An investigation of the differential effort received by items on a low-stakes computer-based test. *Applied Measurement in Education*, *19*(2), 95–114. doi:10.1207/s15324818ame1902_2

*Wise, S. L. (2015). Effort analysis: Individual score validation of achievement test data. *Applied Measurement in Education*, *28*(3), 237–252. doi:10.1080/08957347.2015.1042155

*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. Arief, & D. Liem (Eds.), *Student perspectives on assessment: What students can tell us about assessment for learning* (pp. 187-205).Charlotte, NC: Information Age Publishing.

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment*, *10*(1), 1–17. doi: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. doi:10.1111/j.1745-3984.2006.00002.x

*Wise, S. L., & DeMars, C. E. (2010). Examinee noneffort and the validity of program assessment results. *Educational Assessment*, *15*(1), 27–41. doi: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. doi:10.1111/jedm.12102

*Wise, S. L., Kingsbury, G. G., Thomason, J., & Kong, X. J. (2004, April 13–15). *An investigation of motivation filtering in a statewide achievement testing program* [Paper presentation]. National Council on Measurement in Education 67th Annual Meeting, 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. doi:10.1207/s15324818ame1802_2

Wise, S. L. (2017). Rapid-guessing behavior: Its identification, interpretation, and implications. *Educational Measurement: Issues and Practice*, *36*(4), 52–61. doi:10.1111/emip.12165

Wise, S. L., & Kuhfeld, M. R. (2020). A cessation of measurement: Identifying test taker disengagement using response time. In M. J. Margolis & R. A. Feinberg (Eds.), *Integrating timing considerations to improve testing practices* (1st ed., pp. 150–164). Philadelphia, PA: Routledge.

Wise, S. L., Pastor, D. A., & Kong, X. J. (2009). Correlates of rapid-guessing behavior in low-stakes testing: Implications for test development and measurement practice. *Applied Measurement in Education*, *22*(2), 185–205. doi:10.1080/08957340902754650

Wise, S. L., & Smith, L. F. (2011). A model of examinee test-taking effort. In J. A. Bovaird, K. F. Geisinger, & C. W. Buckendahl (Eds.), *High-stakes testing in education: Science and practice in K–12 settings* (pp. 139–153). Washington, D.C.: American Psychological Association.

*Wise, S. L., Soland, J., & Bo, Y. (2020). The (Non) impact of differential test taker engagement on aggregated scores. *International Journal of Testing*, *20*(1), 57–77. doi:10.1080/15305058.2019.1605999