Large-scale Assessments in Education

RESEARCH Open Access

Does the choice of response time threshold procedure substantially affect inferences concerning the identification and exclusion of rapid guessing responses? A meta-analysis

Joseph A. Rios* and Jiayi Deng

*Correspondence:
jrios@umn.edu
Department of Educational Psychology, University of Minnesota, Twin Cities, 56 E. River Road, 164 Education Sciences Building, Minneapolis, MN 55455, USA

Abstract

Background: In testing contexts that are predominately concerned with power, rapid guessing (RG) has the potential to undermine the validity of inferences made from educational assessments, as such responses are unreflective of the knowledge, skills, and abilities assessed. Given this concern, practitioners/researchers have utilized a multitude of response time threshold procedures that classify RG responses in these contexts based on either the use of no empirical data (e.g., an arbitrary time limit), response time distributions, and the combination of response time and accuracy information. As there is little understanding of how these procedures compare to each other, this meta-analysis sought to investigate whether threshold typology is related to differences in descriptive, measurement property, and performance outcomes in these contexts.

Methods: Studies were sampled that: (a) employed two or more response time (RT) threshold procedures to identify and exclude RG responses on the same computer-administered low-stakes power test; and (b) evaluated differences between procedures on the proportion of RG responses and responders, measurement properties, and test performance.

Results: Based on as many as 86 effect sizes, our findings indicated non-negligible differences between RT threshold procedures in the proportion of RG responses and responders. The largest differences for these outcomes were observed between procedures using no empirical data and those relying on response time and accuracy information. However, these differences were not related to variability in aggregate-level measurement properties and test performance.

Conclusions: When filtering RG responses to improve inferences concerning item properties and group score outcomes, the actual threshold procedure chosen may be of less importance than the act of identifying such deleterious responses. However, given the conservative nature of RT thresholds that use no empirical data, practitioners may look to avoid the use of these procedures when making inferences at the individual-level, given their potential for underclassifying RG.

Keywords: Rapid guessing, Test-taking effort, Response times, Meta-analysis

Background

An underlying assumption of all educational assessments is that examinee scores are reflective of the assessed knowledge, skills, and/or abilities (KSAs). However, such an assumption is put into question when examinees' fail to employ their full effort. Although there are multiple forms of noneffortful responding (e.g., see Meade & Craig, 2012; Wise, 2017), in the context of assessments concerned predominately with power, increased attention has been placed on rapid guessing (RG). RG occurs when a response is provided in so little time that an examinee would not be able to fully read the item stem or response options, solve its challenge, and select an answer (Wise & Kuhfeld, 2020a, 2020b).

Assuming that examinees have been administered items in which they are capable of effortfully engaging (i.e., they have had an opportunity to learn the content assessed, they are proficient in the test language), RG can occur due to two factors: (a) time limit constraints (i.e., test speededness); and (b) low test-taking effort (Wise, 2017). Concerning the former, examinees may not have the time to fully engage in all test items, and may employ RG in an effort to increase their score (assuming no penalty for incorrect responses). This form of RG has been documented in high-stakes tests, in which the personal consequences for examinee performance is significant (see Schnipke & Scrams, 1997). In contrast, examinees may disengage when they perceive the assessment results to have little value and/or personal consequences (Penk & Schipolowski, 2015). Thus, the cost of expending effort is seen to be too great when compared to the perceived benefits. Disengaged RG has been documented across a number of low-stakes (i.e., examinee performance his minimal to no personal consequences) assessments and a myriad of ages and cultures (see Rios, 2021a).

Regardless of the underlying reason for RG, such responses represent cases in which examinees are not effortfully engaging with items as expected (Wise, 2017). In testing contexts that are predominately concerned with power (i.e., the predominate goal is to assess examinee KSAs without considering processing time as an important factor and only include time limits due to practical constraints), RG violates the assumption that an item response is reflective of examinee KSAs.[1] As a consequence, such behavior is generally associated with underestimation of examinee ability (e.g., Rios et al., 2017), leading to inaccurate inferences of measurement properties (e.g., DeMars & Wise, 2010; Mittelhaëuser et al., 2015; van Barneveld, 2007; Wise & DeMars, 2009) and performance (e.g., Debeer et al., 2014; Osborne & Blanchard, 2011; Rios, 2021a; Wise & DeMars, 2010; Wise et al., 2013).[2]

To address this concern, the *Standards for Educational and Psychological Testing* has noted that it is important to both clearly document examinee test-taking effort and the decision criteria for including individual examinee scores with questionable degrees of effort (Standard 13.9; American Educational Research Association et al., 2014; p. 213).

[1] In speeded tests, fast responding is a part of the construct assessed. This context is not the focus of the current study.

[2] Although RG undermines valid interpretations of measurement properties and test performance, RG responses due to test speededness provide valuable information to test developers/users concerning the appropriateness of the employed time constraints. Similarly, examining item correlates of RG due to reasons unrelated to test speededness can provide test developers with useful guidance on how to modify item characteristics, layout, or administration procedures to mitigate RG (for an example, see Wise et al., 2009).

Although a laudable recommendation, this first presumes that disengaged behavior, such as RG, can be accurately identified. To this end, practitioners have largely relied on utilizing response time (RT) information, which provides a gauge of how long each examinee spent on answering an item, as a proxy of RG behavior. However, to date, there is no generally agreed upon threshold that establishes when a response is provided too quickly as to be deemed a RG response.

To assist in addressing this limitation, we conduct a meta-analysis that investigates the variability in documenting RG across RT threshold procedures on computer-administered low-stakes power tests.[3] Specifically, we include studies that: (a) employed two or more RT threshold procedures on the same empirical dataset; and (b) evaluated differences between procedures on descriptive (the proportions of RG responses and responders identified), measurement property (item properties), and test performance outcomes. In the sections that follow, we elaborate on why the use of RT has become the preferred proxy for measuring RG, describe the various RT threshold procedures that have been proposed in the literature, and discuss the rationale for the current meta-analysis.

Proxies of RG

Three proxies of RG behavior have been proposed in the literature: (a) self-reported effort; (b) aberrant response patterns; and (c) response times. Of these three, the latter has seen increased usage in the literature (e.g., Silm et al., 2020) and in large-scale formative (e.g., the Measure of Academic Progress; see Wise & Kuhfeld, 2020a, 2020b) and international operational assessments (e.g., Programme for the International Assessment of Adult Competencies; see Goldhammer et al., 2017). This approach classifies any response provided in less time than an established RT threshold as RG (more detail on these threshold procedures are discussed in the next section). Across threshold-setting procedures, the use of RT has numerous advantages.

First, due to the use of log file information, it is an unobtrusive approach, as examinees are presumably unaware that their test-taking behavior is observed, which limits concerns about observer effects (e.g., subject behavior changing due to their knowledge that they are being observed). Second, contrary to the other procedures, this approach can evaluate RG on an item-by-item basis, which addresses the concern of shifting test-taker behavior (e.g., Wise & Kingsbury, 2016). This capability allows for the investigation of item characteristic correlates associated with RG (e.g., Wise et al., 2009) and provides the scoring advantage of estimating ability for examinees that have engaged in limited RG.[4] In contrast, procedures that identify global-levels of motivation or aberrant behaviors, may utilize listwise deletion to remove data from examinees deemed to be unmotivated/aberrant. Such an approach has been found to lead to a loss of as much as 25% of

[3] We acknowledge that the distinction of power and speeded tests is predominately theoretical, as most tests in education primarily intend to measure the construct of interest, but also employ a time limit to address practical constraints. However, we make this distinction given the difficulties in disentangling RG from disengagement and test-wiseness (i.e., fast responding) in speeded tests. Our focus is on RG that is unreflective of examinee KSAs in contexts in which processing speed is not a primary component of the KSAs assessed.

[4] Researchers have also proposed mixture model approaches, which utilize response patterns and/or times, to simultaneously assign class (rapid guessers and non-rapid guessers) probabilities and estimate ability (e.g., Wang & Xu, 2015). Given the computational demands of such models and their minimal adoption in operational settings, the focus of this paper is on studies that identify fixed RT thresholds for classifying RG responses.

the sample and biased scores if RG is associated with examinees' underlying ability (see Rios et al., 2014). However, one of the major limitations of RT is that it cannot be applied to data collected from paper-and-pencil test administrations, as it relies on the collection of log file information at the item by examinee level. Regardless of this limitation, the use of RT as a proxy of RG is the most researched and applied approach in operational settings, to date.

Response time threshold procedures

Given the popularity of using RT, researchers have proposed a number of procedures for establishing thresholds to classify RG behavior. These procedures can be categorized into three distinct typologies, corresponding to methods that utilize: (a) no empirical data (NED); (b) only response time information (RT); and (c) a combination of response time and accuracy information (RTRA). Below we provide a greater description of each typology.

Procedures utilizing no empirical data

Two distinctive procedures fall under this typology, which establish thresholds without utilizing either RT or response accuracy data from examinees: surface feature and common *k* second methods.

Surface feature thresholds One of the first approaches proposed to establish RT thresholds is to consider the impact of an item's characteristics on responding time, with the rationale that items requiring more reading should on average take longer to solve. To this end, thresholds for RG could be established based on taking the number of characters contained within a given item, and coupling this information with estimated reading speeds for a given test-taking population. This would provide estimates on how long an item may take at minimum to read, which could be used to set thresholds (see Wise & Kong, 2005). One limitation of this procedure is that it does not consider empirical response time distributions, and thus, estimates may be inaccurate. Furthermore, this procedure is limited in that surface feature information is required for each item, which may be a limiting factor for practitioners that do not have access to this material or those applying this procedure to a large number of items (e.g., items contained within an item pool for a computer adaptive test [CAT]).

Common k second thresholds This procedure sets the criterion for RG based on a common second threshold (e.g., 3 s) that is applied across all items (see Wise et al., 2004). Though a simple procedure, it has been utilized operationally by the PIAAC testing program to identify RG (see Goldhammer et al., 2016). This procedure is advantageous in that it does not require item surface feature information nor response time and/or accuracy information. Thus, it can be applied to a large number of items in an item pool with limited demands on practitioners. A major disadvantage of this approach is that it ignores variability in items. For example, the same threshold is applied to items requiring a heavy reading load (e.g., testlet-based items) and simple computation (e.g., math addition items). Although it has been argued that the response time distribution is nearly identical for all items under RG due to disengagement (Wang & Xu, 2015), it is possible

that variability in RG response times may be observed if examinees engage in an item, perceive a low probability of success, and then employ RG. If this were the case, this procedure may be too conservative and lead to increased false negatives, considering its inability to adjust to item and population characteristics (Wise, 2017).

Procedures utilizing only response time information

This typology relies on sample response time distributions to set thresholds, and is comprised of two distinctive procedures: percentile and bimodal distribution threshold methods.

Percentile thresholds Wise and Ma (2012) proposed the Normative Threshold (NT) procedure to provide an approach that adapts to potential differences in items and testing populations, but can simultaneously be applied to a large number of items. This procedure sets the threshold for RG as the percentile for a given item's response time distribution. Considerable research has been conducted to establish the best percentile (see Wise & Kuhfeld, 2020a, 2020b; Wise & Ma, 2012); however, many applied studies have set the threshold at the 10th percentile (e.g., Wise & Ma, 2012). Although an easily applied procedure, one of its major disadvantages is that it is devoid of a strong theoretical rationale for its classification of RG, similar to the common k second procedure.

Bimodal distribution thresholds Using mixture modeling, Schnipke (1995) was one of the first researchers to demonstrate that in the presence of RG, a RT distribution may be bimodal in which the lower and upper modes represent RG and solution behavior respectively. Using this rationale, Schnipke proposed setting RT thresholds at the lowest point where the two distributions meet, which conceptually is the time at which examinees transition from RG to solution behavior. Setting this threshold in practice can be done by visually inspecting response time distributions; however, such a process is both time consuming and can lead to disagreements between observers. To remedy this issue, Rios and Guo (2020) proposed the mixed log-normal distribution (MLN) procedure, which essentially automates the process of finding the lowest point between the two thresholds.

Although this general approach to setting thresholds based on a RT distribution provides a strong theoretical rationale, in practice, an item's RT distribution may not always be bimodal. This may occur when an item requires relatively a short amount of time to solve when utilizing solution behavior (Wise, 2017). As a consequence, it may be impossible to set thresholds for some items, which is a drawback of this procedure. In addition, some researchers have questioned whether this procedure actually captures RG behavior, as it has the tendency to set higher thresholds (e.g., above 30 s) than other procedures (see Wise & Kuhfeld, 2020a, 2020b).

Procedures utilizing a combination of response time and accuracy information

A newer category of response time threshold procedures have been developed by incorporating response accuracy information with response times (e.g., Goldhammer et al., 2016; Guo et al., 2016; Lee & Jia, 2014). The underlying assumption of these procedures is that RG responses possess accuracy rates that are approximately equal to chance (typically defined as the reciprocal of the number of response options), which has been

supported by prior research (e.g., Wise & Kong, 2005). Thus, for a procedure, such as the Cumulative Proportion Correct (CUMP) method proposed by Guo et al. (2016), a threshold is established at the time point in which the correct response rate begins to be consistently greater than chance.

The advantage of this approach is that it combines multiple sources of empirical information and its rationale is supported by prior research. However, such an approach may require substantial item-level data to accurately detect an increase in accuracy rates by second (Wise, 2017). Further, setting thresholds using this approach may be impossible for items that are either very difficult (i.e., the proportion correct never exceeds the chance rate across response times) or easy (i.e., the proportion correct consistently exceeds the chance rate across response times; Wise & Kuhfeld, 2020a, 2020b).

Study rationale

Given the variety of RT threshold procedures, practitioners are often confronted with selecting the best approach to adopt. However, to date, there has been limited guidance provided to practitioners on how RT threshold typologies differ from each other across samples and assessment contexts in terms of descriptive, measurement property, and performance outcomes. Such an understanding would require a synthesis of previous research that has compared various RT threshold procedures on the same datasets, which at the time of this writing, has not been conducted. As a consequence, it is unknown how many studies have compared disparate RT threshold procedures, how the choice of threshold procedure is associated with differences in outcomes, and the role that contextual variables play in impacting the association between RT threshold procedure and the outcomes under investigation.

Therefore, to fill the gap in the literature, the purpose of this paper is to conduct a meta-analysis of studies that compare two or more RT threshold procedures on the same empirical dataset obtained from computer-administration of a low-stakes power assessment (i.e., assessments that design time limits to ensure that examinees have sufficient time to respond to all items). Such a project is vital as extensive research efforts have been placed on controlling for low test-taking effort via post-hoc analyses, such as filtering RG responses (e.g., Rios et al., 2014, 2017) or incorporating measures of RG into IRT ability estimation (e.g., Wise & Kingsbury, 2016). However, in doing so, it is assumed that RG can be accurately identified. Although accuracy can never be truly known, given that RT is used as a proxy of actual behavior, this meta-analysis seeks to compare differences in descriptive, measurement property, and test performance outcomes (once filtering responses classified as RG) across RT threshold procedures that utilize NED, RT, or RTRA. In addressing these study objectives, the following research questions are evaluated:

1. When comparing RG threshold procedure pairings, what is the average difference in:

    a. The number of RG responses and examinees engaging in RG identified?
    b. Measures of item properties (i.e., item difficulty and discrimination), and performance outcomes upon filtering RG responses?

2. If non-negligible differences between RG threshold procedure pairings are observed:

   a. What contextual variables (e.g., participant, assessment characteristics, and RG threshold attributes) are associated with such differences?
   b. How do threshold types differ?

Findings from these analyses have the potential to inform practitioners about the RG threshold procedures that have been most extensively studied and how the choice of such procedures are associated with potentially differing outcomes.

Method

Search strategy

To conduct a thorough literature search, four strategies were employed: (a) academic database, (b) Internet browsing, (c) backward and forward citation; and (d) expert consultation searches. Data collection was conducted by the second author and completed on July 31, 2020. A full description of each search strategy is provided below, presented in the order conducted.

Academic database search

The first approach employed to locate relevant articles consisted of searching the following academic databases: PsycINFO (via Ovid); ERIC (via EBSCOhost); Education Source (via EBSCOhost); and Academic Search Premier (via EBSCOhost). These databases covered journal articles across multiple fields, such as psychology (PsycINFO), education (ERIC and Education Source), and statistics (Academic Search Premier). The key terms employed across databases were "rapid guess" and "response time". As there were multiple formats of the term "guess" (e.g., guess, guesses, guessing), the term "rapid guess" was entered with the Boolean modifier Asterisk ("*") to be used as a root word. To narrow the search and improve accuracy, the key terms "rapid guess*" and "response time" were entered with the Boolean operator "AND". Additionally, only studies published in the English language were included. No other initial restrictions were placed on the search.

Internet browsing

An Internet search was conducted via Google Scholar to strengthen the coverage of grey literature not included in the academic databases noted above. The key terms used in the Internet search were identical to those used in the academic database search. Results produced in Google Scholar were sorted by relevance (little information is known on how Google sorts its hits; Haddaway et al., 2015). Although the first 1000 results were accessible, the return rate of relevant articles continued to decrease as we progressed through the results. 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.

Expert consultation

This search strategy was conducted by directly contacting the following researchers known to have conducted work and/or published extensively on the topic of RG: Steve Wise (Northwest Evaluation Association, USA), Megan Kuhfeld (Northwest Evaluation Association, USA), Sara Finney (James Madison University, USA), Dena Pastor (James Madison University, USA), Jim Soland (University of Virginia), and Brandi Weiss (George Washington University, USA). Each individual was contacted via email to ascertain whether they had conducted unpublished research that met our inclusion criteria and/or knew of such research authored by others. All communication and article retrieval was completed by July 1, 2020.

Citation searching

Beyond the search strategies described above, backward citation searching was also included. This was done by searching the reference lists of two pertinent review articles (Silm et al., 2020; Wise, 2017) identified by the first author as well as all articles found to meet our eligibility criteria (described below) from the academic database, Internet, and expert consultation searches. This search was conducted using both Social Sciences Citation Index and Google Scholar. All studies found using backward citation searching were then evaluated based on the eligibility criteria, and if met, 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. This was done by typing in the title for the study of interest into Google Scholar. Upon finding the article of interest, the *cited by* link was clicked on, which allowed for the ability to search studies that cited the article of interest. Any study included from this strategy 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 July 31, 2020.

Eligibility criteria

To be included in this meta-analysis, studies had to meet the eligibility criteria set forth along three dimensions: (a) data type; (b) RG identification methodology; and (c) outcomes.

Data type

Only studies that utilized empirical data to study RG threshold procedures were included. Empirical data could have been obtained from either primary or secondary data collections from an unspeeded, group-administered classroom, formative, or accountability test that was low-stakes and computer-administered. The choice of only including low-stakes power or unspeeded tests was to ensure that RG largely reflected test disengagement rather than test speededness.[5] No further restrictions were placed

[5] As noted by Wise (2017), it is possible that examinees in low-stakes test contexts may engage in RG due to time constraints, however, this issue is likely to occur less frequently than RG due to disengagement.

on examinee (e.g., age, country of origin, ethnicity, language) nor assessment (e.g., content area, length, item types) characteristics. However, data obtained from simulation studies were excluded.

RG identification methodology

Although there are multiple proxies for identifying RG (e.g., self-report measures, person-fit statistics), this meta-analysis only included studies that utilized RT threshold procedures. To be included, studies had to investigate RG based on two or more threshold approaches. These thresholds could either be a variant (e.g., using different percentiles) of one or two of the following procedures: (a) surface feature; (b) common k-second; (c) percentile; (d) bimodal distribution; and (e) response time and accuracy information procedures.

Outcomes

The outcomes of interest were differences between RT thresholds on three categories of variables: (a) descriptives; (b) measurement properties; and (c) performance. For a study to be included, it must have presented quantitative results on one or more of these outcomes. A quantitative result was defined as any test statistic (e.g., χ^2, Z, t, F, $\widehat{p}$) necessary for computing a Cohen's *d* (standardized difference), Cohen's *h* (difference in proportions), or a correlation effect size (more detail on these effect size calculations are included below).

Within the descriptive category of variables, we coded for differences in the proportion/percentage of RG responses identified and proportion/percentage of examinees engaging in RG. Concerning the latter, if not explicitly classified in the original article, 0.9 was set as the cut-score utilizing the response time effort index (RTE; i.e., the proportion of responses not identified as RG) to classify motivated and unmotivated test takers. This cut-off (i.e., an unmotivated examinee employed RG on 10% or more of items administered) was first proposed by Wise and Kong (2005) and has since been used extensively in applied research (e.g., Rios et al., 2014). Consequently, the RG examinee proportion rate was calculated by subtracting the percentage of the participants with RTE equal to or greater than 0.9 from 1.

In regard to measurement properties, we were interested in how the choice of RT threshold procedure was associated with differences in: (a) average item difficulty (measured using proportion correct and/or IRT calibration estimates) after removing RG responses or examinees; and (b) average item discrimination (measured using an item-total correlation and/or IRT calibration estimates) after removing RG responses or examinees.

The last outcome variable of interest was the difference in the average sample performance between RT thresholds. This variable could be reported as the mean raw, scale, or theta estimate score for the total sample, after removing RG responses or examinees. This level of aggregation was chosen given that most low-stakes educational tests report group-level performance for monitoring and accountability efforts.

Variable coding

Five variables were identified as potential factors that could account for differences in the outcomes of interest: (a) examinee age; (b) test subject; (c) test length; (d) threshold typology pairing; and (e) threshold pairing variability. The first three variables were included as they have been found to moderate the extent of RG observed in operational testing (meaning more potential variability in the number of RG responses), while the last two variables were the main independent variables of interest. A detailed description of each variable is presented below.

Age

RG has been shown to vary across age groups, with older examinees more likely to engage in unmotivated behaviors (Goldhammer et al., 2016). As a result, the average age of the sample was coded. Operationally, if a primary author did not provide the sample's average age, examinee grade-level was used as a proxy. As an example, age 6 was used for 1st-year primary school students, 13 was coded for 8th-grade students, and 18 was utilized for college freshmen. Additionally, if examinees were reported to come from a range of grades, the midpoint was used for the group. For example, age 20 was imputed for a sample of undergraduate college students. If neither age nor grade were provided, this variable was coded as missing.

Test subject

Prior research has suggested that test subject can moderate participants' test taking effort. As an example, Kiplinger and Linn (1994) found that more than half of students in both grade 8 and grade 12 reported significantly more effort in taking math tests than in other subjects. To account for this potential moderation, we dichotomously coded for a test's subject as either a "STEM or mixed subject" or "non-STEM" subject. Mixed subject tests were defined as those that included both STEM and non-STEM content.

Test length

Test takers have been shown to exhibit more disengaged responses as a test grows in length, potentially due to issues of cognitive fatigue (e.g., Wise & Kingsbury, 2016). As a result, test length was considered as a moderator that might impact the identification of RG responses, given that longer tests may be associated with greater variability in RG.

Threshold typology pairing

As described earlier, there are three main typologies of RT thresholds, based on the utilization of: (a) NED; (b) RT; and (c) RTRA. In the present study, this led to six different comparison pairings, three for variants of RT thresholds found within the same typology (e.g., comparison of variants within the NED typology) and the remaining three for RT thresholds that differed between typologies. Specifically, each RT threshold fell into one of the following pairings: (a) NED–NED; (b) RT–RT; (c) RTRA–RTRA; (d) NED–RT; (e) NED–RTRA; and (f) RT–RTRA.

Threshold pairing variability

To account for variance in RT threshold comparisons, a dummy-coded variable was created to signify whether a comparison was within or between threshold typology/typologies (within served as the reference).

Interrater agreement

Interrater reliability was computed for three distinctive stages: title and abstract screening, full-text reviewing, and variable coding. Rayyan (https://rayyan.qcri.org) was employed for interrater reliability coding of the title and abstract and full-text review phases, while Excel was used for variable coding. Prior to coding for each stage, the principal investigator provided training to the second author (a Ph.D. student in educational measurement) that was comprised of discussing the objectives and evaluation criteria of the stage, reviewing each variable's operational definition, and engaging in joint coding of a small percentage of articles. Upon completion of training, the second author was responsible for all coding across stages, while the first author coded 20% of articles at each stage to evaluate interrater reliability. An interrater agreement value of 0.80 was set as the criterion for establishing rater consistency. Any inconsistent decisions across raters were resolved through discussion and consensus. For the first stage (i.e. title and abstract screening), the two authors were in high agreement on article inclusion with an interrater agreement of 0.91. For the other two review stages, no conflicts were presented, with the interrater agreement equal to 1.

Statistical methods

The sections that follow describe the procedures for: (a) calculating effect sizes; (b) evaluating publication bias; (c) identifying outliers; (d) estimating average effect sizes and effect size heterogeneity; and (e) performing moderator analyses.

Calculating effect sizes

Effect sizes were calculated separately for continuous, proportional, and correlational data. Concerning the former, continuous data were presented for the following outcome variables: IRT item discrimination parameter estimates, IRT item difficulty parameter estimates, and group test performance. To calculate the effect sizes for these variables, Cohen's d formula was used:

$$d = \left| \frac{\overline{M}_2 - \overline{M}_1}{\sqrt{\frac{(n_2-1)S_2^2+(n_1-1)S_1^2}{n_1+n_2-2}}} \right|, \tag{1}$$

where M_1 and M_2 are sample means for threshold 1 and threshold 2 respectively, n_1 and n_2 are sample sizes for threshold 1 and threshold 2 respectively, S_1 and S_2 are the standard deviations of outcomes for threshold 1 and threshold 2 respectively. As the direction of outcome was not of interest, an absolute value of the effect size was computed. Furthermore, the variance of Cohen's d was calculated as:

$$v_d = \frac{n_1 + n_2}{n_1 n_2} + \frac{d^2}{2(n_1 + n_2)}, \tag{2}$$

where d is the absolute value of Cohen's d calculated from formula (1) above. The computation of Cohen's d was completed in the *R* package *compute.es* (Del Re, 2020).

Proportional data were reported for the following dependent variables: (a) proportion of examinees engaging in RG; (b) proportion of responses identified as RG; and (c) CTT item difficulty values (i.e., proportion of correct responses) for effortful test takers. As the power to detect differences in proportions is dissimilar across studies due to unequal sample sizes (Cohen, 1988), a nonlinear transformation, defined as φ, was applied to provide equal detectability of outcomes. Given the estimated proportions ($\widehat{p}$) of two thresholds on an outcome of interest, φ is computed via the formula:

$$\varphi = 2arcsin\sqrt{\hat{p}}. \tag{3}$$

This transformation was then used to calculate an effect size for differences between proportions for a given threshold pairing using Cohen's h formula:

$$h = |\varphi_1 - \varphi_2|. \tag{4}$$

Similar to Cohen's d, the absolute value was computed for Cohen's h as no directional assumptions were made. However, unlike Cohen's d formula, a variance estimate is not readily available for Cohen's h. Thus, to obtain some measure of variability, heterogeneity was demonstrated via the standard deviation of effect sizes. For interpretation purposes, Cohen's (1988) guidelines were adopted in which an h value between 0.2 and 0.5 indicates a small effect size, an h value between 0.5 and 0.8 represents a medium effect size, and an h value greater than 0.8 reflects a large effect size. Across proportional outcome variables, Cohen's h was calculated using the *pwr* package in *R* (Champely, 2020).

Finally, correlation coefficients were reported for the average CTT item discrimination (item-total correlations). Although the correlation coefficient can serve as an effect size on its own, Fisher's z transformation was applied to every correlation to normalize the sampling distribution using the *metafor* package in R (Viechtbauer, 2020). This transformation was applied as:

$$z = 0.5 * \ln\left(\frac{1 + \mathrm{r}}{1 - \mathrm{r}}\right), \tag{5}$$

and the variance of the transformation was calculated as:

$$v_z = \frac{1}{n - 3}. \tag{6}$$

Then the effect size difference for each threshold pair was calculated using Cohen's q index (Cohen, 1988):

$$q = |z_1 - z_2|, \tag{7}$$

while the variance for this index was calculated as:

$$var(q) = \frac{1}{N_1 - 3} + \frac{1}{N_2 - 3}, \tag{8}$$

where N_1 and N_2 are the sample sizes based on the correlation for threshold 1 and threshold 2, respectively. For this effect size, Cohen (1988) proposed the following categories for interpreting *q values*: no effect: < 0.10; small effect: 0.10–0.29; moderate effect: 0.30–0.50; and large effect > 0.50.

Estimating average effect sizes and evaluating effect size heterogeneity

Prior to estimating average effect sizes and effect size heterogeneity, the effect sizes of each outcome were diagnosed for potential outliers. Outliers were defined as any estimated effect size greater than three standard deviations (based on the absolute value) from the mean effect size of the given outcome. To avoid the loss of data, any identified outlier was down-weighted to a value equal to three standard deviations from the mean. A sensitivity analysis was then conducted to evaluate the impact of the identified outliers on the estimation of the mean effect sizes for each dependent variable. If any inflation or deflation of the mean effect size under study was observed, the adjusted effect size estimates were used for all subsequent analyses.

For continuous and correlational outcome variables, an intercept-only random-effects meta-regression model was run in the *robumeta* package in *R* (Fisher et al., 2017) to calculate the average effect size and effect size heterogeneity. To avoid artificially reducing variance estimates and inflating Type I error due to effect size dependencies (i.e., multiple effect sizes are produced by comparing various response time thresholds from the same study; Borenstein et al., 2009), the robust variance estimation (RVE) procedure developed by Hedges et al. (2010) was employed. The heterogeneity of effect sizes was investigated using the I^2 statistic:

$$I^2 = \left(\frac{Q - k}{Q}\right) \times 100\%, \tag{9}$$

where Q is a homogeneity statistic that represents the degree that random-effect variance is significantly different from 0, and k is the number of studies. Higgins and Thompson (2002) proposed effect size guidelines for this statistic, with I^2 values less than 50% indicating small heterogeneity, values between 50% and 75% representing medium heterogeneity, and values greater than 75% reflecting large heterogeneity.

As variance estimates for dependent variables that reported only proportional data were not available, classical approaches to calculating average effect sizes and heterogeneity were taken. These consisted of respectively computing the mean and standard deviation of the effect sizes for the outcome under investigation.

Moderator analyses

For continuous and correlational data, moderator analyses were conducted for any outcome that was found to have a large degree of heterogeneity.[6] This was done by estimating the following random-effects meta-regression model:

[6] A meta-regression model could not be calculated for proportional data, as a measure of variance for each effect size was not available.

$$\begin{aligned}\widehat{y} = {} & b_0 + b_1(age) + b_2(test\ subject) + b_3(test\ length) \\ & + b_4(RT - RT) + b_5(RTRA - RTRA) \\ & + b_6(NED - RT) + b_7(NED - RTRA) + b_8(RT - RTRA) \\ & + b_9(threshold\ pairing\ variability) + e,\end{aligned} \tag{10}$$

where $\widehat{y}$ was equal to one of the continuous or correlational outcome variables of interest (test performance, IRT item discrimination parameter, IRT item difficulty parameter, item-total correlation), b_0 was equal to the average effect size for the outcome variable holding all included variables constant, *age* and *test length* were continuous variables, *test subject* was coded dichotomously as "non-STEM subject" or "STEM or mixed subject" (reference group), b_4 through b_8 were dummy-coded variables for RT–RT, RTRA–RTRA, NED–RT, NED–RTRA and RT–RTRA (NED–NED served as the reference group), *threshold pairing variability* was coded as between threshold typologies or within a threshold typology (reference group), and e was the residual term. The moderator analyses were conducted in the *R* package *metafor* (Viechtbauer, 2020).

Results

Overall, 958 studies were reviewed based on academic database, Internet browsing, expert consultation, and backward and forward citation searching. Out of these 958 studies, 19 were found to meet the eligibility criteria (2% hit rate; see Fig. 1), which produced between 16 and 87 effect sizes across nearly every outcome of interest, except for differences in item discrimination (both CTT and IRT estimates) and item difficulty using IRT calibration. These latter outcomes were not found to be investigated in our sample, and thus were not included in the analyses noted below.

Regarding the characteristics of the studies included, 89% (17 out of 19) were published on or after 2010, while only two (11%) were written before 2000. Nearly half (47%) of the included studies were grey literature, representing dissertations (2 studies), research reports (2 studies), conference papers (3 studies), and works in progress (2 studies). In terms of sample characteristics, the mean sample size was 126,876 (ranging from 213 to 728,923) and 10 studies used participants in postsecondary education. For 10 out of 19 (53%) studies, data were collected within the United States, while the rest were either comprised of non-US or mixed (i.e., U.S. and non-U.S.) populations.

This diversity in population nationality was most likely associated with more than half (63%) of studies sampling data from large-scale assessments (LSAs), with five utilizing international data from either the Programme for International Student Assessment (PISA; three studies) or the Programme for the International Assessment of Adult Competencies (PIAAC; two studies). Other LSAs examined were NWEA's Measures of Academic Progress (MAP) Growth (16%), the ETS Proficiency Profile (EPP, 5%) and HEIghten® Critical Thinking Assessment (5%). The remaining seven studies (37%) utilized data from locally developed performance tests. Across these assessments, test length ranged from 7 to 275 items, with an average of 57. Among all 19 studies, 14 (74%) used assessments consisting solely of selected-response items, 12 (63%) utilized tests only delivered in English, and 13 (68%) employed measures of STEM or a mixture of STEM and non-STEM subjects.

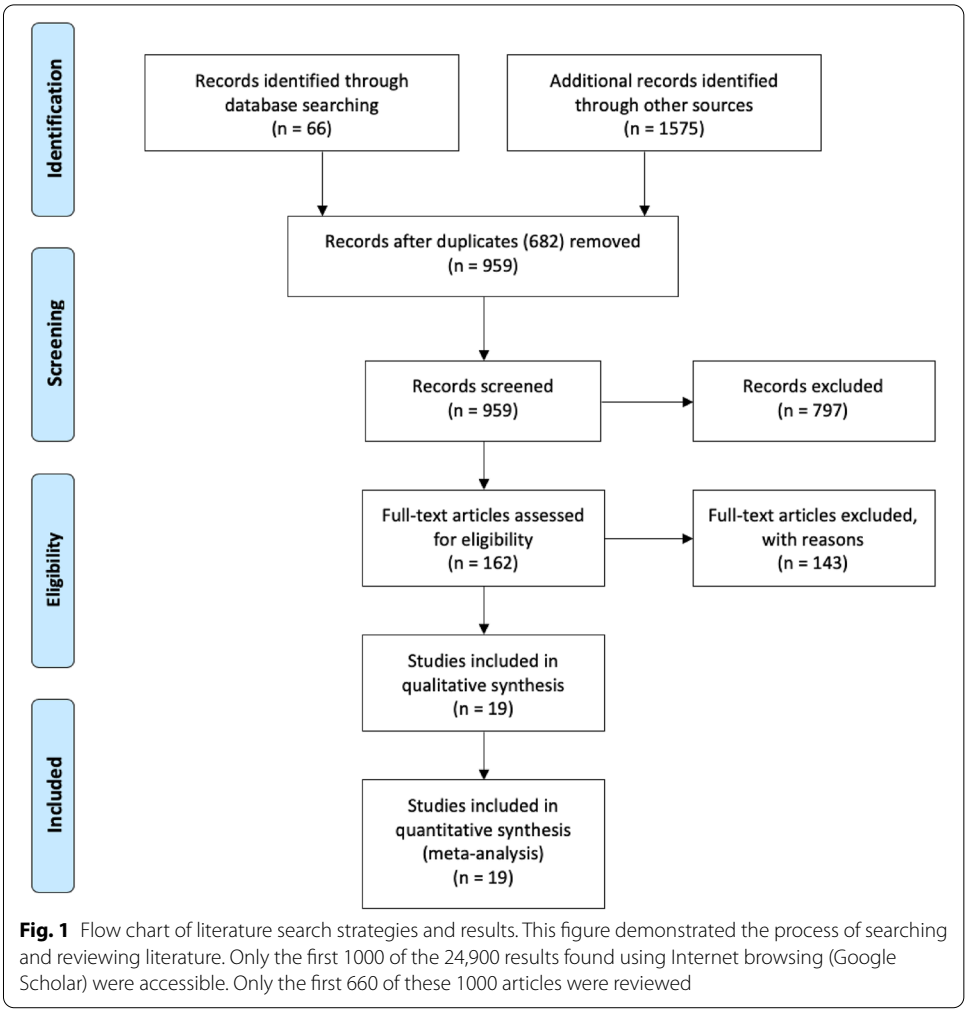

Fig. 1 Flow chart of literature search strategies and results. This figure demonstrated the process of searching and reviewing literature. Only the first 1000 of the 24,900 results found using Internet browsing (Google Scholar) were accessible. Only the first 660 of these 1000 articles were reviewed

Average effect sizes and heterogeneity

Across outcomes, no effect sizes were found to lie outside three standard deviations of the mean. Figure 2 presents the average effect size and variability for each outcome. All subgroup comparisons made below are limited to RT threshold pairings with a minimum of 20 effect sizes. This is done to mitigate spurious inferences due to small sample sizes.

Descriptive outcomes

Results for each threshold type are included in Table 1 for the descriptive outcomes.[7] Furthermore, each effect size based on RT typology pairing for the proportion of RG responses (left plot) and responders (right plot) identified is shown in Fig. 3.

[7] Providing performance outcomes made little sense, given the different characteristics of the assessment examined for each threshold type across studies. Therefore, these variables were excluded from the table. For the variables presented, the reader should be cautioned from overgeneralizing the results due to small sample sizes for some cells.

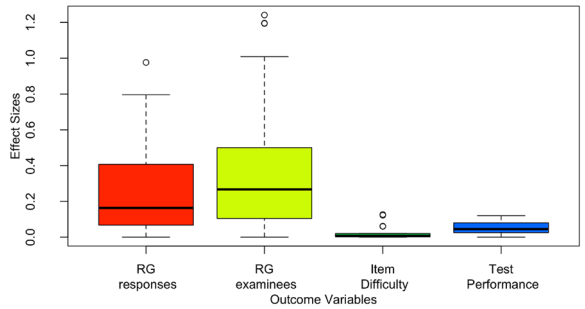

Fig. 2 Average effect size and variability of each outcome variables. *RG responses* = proportion of RG responses identified, *RG examinees* = proportion of examinees engaging in RG

Table 1 Means and standard deviations for descriptive outcomes by threshold type

Threshold typology	Threshold procedure	Proportion of RG responses identified	Proportion of examinees engaging in RG
No empirical data	Surface feature thresholds	0.22 (–) n = 1 k = 1	0.14 (0.10) n = 5 k = 3
	Common-*k* second thresholds	0.06 (0.05) n = 7 k = 2	0.19 (0.21) n = 16 k = 6
Response time	Response time distribution thresholds	0.11 (0.06) n = 6 k = 3	0.13 (0.06) n = 11 k = 5
	Percentile thresholds	0.06 (0.07) n = 17 k = 5	0.22 (0.21) n = 14 k = 5
Response time and accuracy	Response time and accuracy thresholds	0.29 (0.18) n = 4 k = 1	0.36 (0.30) n = 5 k = 2

The descriptive values provided should not be directly compared across procedures because they are based on different samples of examinees and tests. Standard deviations are provided in parentheses. *n* is the number of unique effect sizes for the cell; *k* is the number of unique studies for the cell

Proportion of RG responses identified A total of 54 effect sizes examining the difference in the proportion of RG responses were obtained in our sample. Although a total of four threshold typology pairings were investigated in the literature (RT–RT, NED–RT, NED–RTRA, RT–RTRA), the two most studied were the RT–RT and NED–RT procedures, each respectively contributing 24 and 22 effect sizes. Across all pairings, the average difference observed was a Cohen's *h* of 0.26 ($SD = 0.24$), indicating a small-moderate effect of threshold type.

As is shown in Fig. 3, there was variability around the mean depending on RT threshold pairing. Specifically, in comparing the mean Cohen's *h* for the RT–RT ($M = 0.17$, $SD = 0.14$) and NED–RT ($M = 0.26$, $SD = 0.29$), the latter pairing was found to produce a

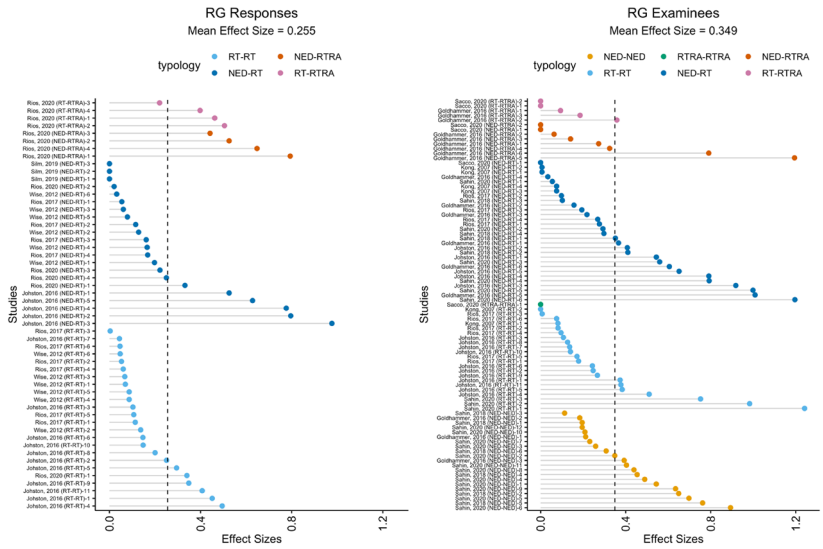

Fig. 3 Effect size based on RT typology pairing for descriptive outcomes. *RG responses* = proportion of RG responses identified, *RG examinees* proportion of examinees engaging in RG

larger difference by 0.38 *SDs*. A closer examination of studies examining this latter pairing, suggests that the rate of RG responses were lower when employing a RT threshold without the use of empirical data when compared to utilizing response times for three of five studies (60%). Furthermore, across studies, within the RT threshold typology, response time distribution thresholds identified a larger proportion of RG responses ($M = 0.11$, SD = 0.06; n = 6; k = 3) compared to percentile thresholds ($M = 0.06$, SD = 0.07; n = 17; k = 5; Table 1). However, of all threshold types, RTRA thresholds identified the largest proportion of RG responses ($M = 0.29$, SD = 0.18; n = 4; k = 1).

Proportion of examinees engaging in RG One of the most studied outcome variables in our meta-analysis was the comparison of proportion of examinees engaging in RG (defined as an examinee RG on 10% or more of the administered items), with a total of 85 effect sizes found based on a comparison of all six RT typology pairings. Across these effect sizes, the average Cohen's *h* was 0.35 ($SD = 0.31$), suggesting a small-moderate effect. A closer examination of each threshold typology pairing showed that three contributed a minimum of 21 effect sizes. Specifically, the most represented were the NED–RT pairing (n = 28), followed by the RT–RT (n = 22) and NED–NED (n = 21) pairings. The subgroup comparisons provided two interesting findings.

First, threshold pairings within the NED typology ($M = 0.41$, $SD = 0.22$) were found to produce larger differences on this outcome compared to those thresholds within the RT typology ($M = 0.23$, $SD = 0.32$) by an average of 0.66 *SD*s. This suggests that procedures utilizing RT, on average, produced more similar outcomes than those procedures not employing empirical data. Second, when comparing thresholds between the RT and NED typologies, a moderate difference in the proportion of RG examinees identified was found ($h = 0.42$, $SD = 0.34$), with the former generally classifying a higher proportion in six out of seven studies (86%). Concerning RTRA procedures,

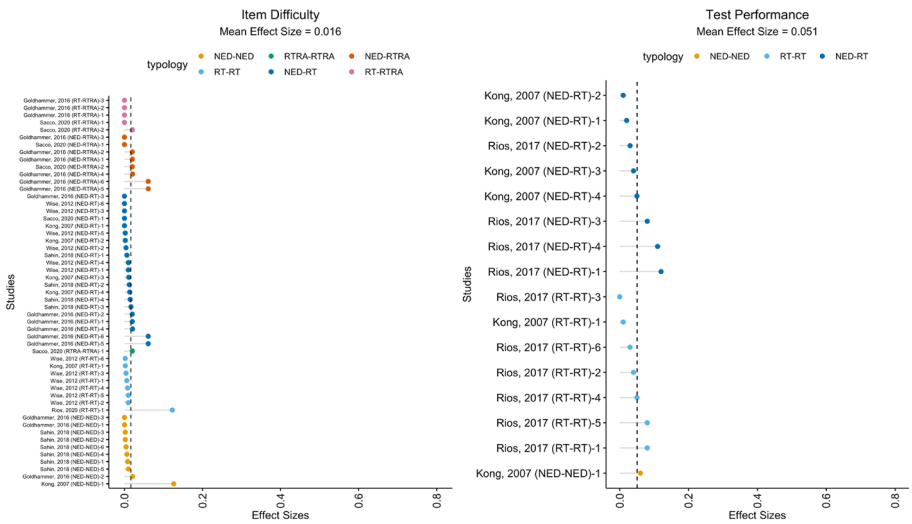

Fig. 4 Effect size based on RT typology pairing for measurement property and performance outcomes

minimal comparisons were made, given that only two studies examined this typology and outcome. With that said, descriptive results indicate that this typology identified the largest proportion of examinees engaging in RG ($M = 0.36$, SD = 0.30; n = 5; k = 2), followed by percentile ($M = 0.22$, SD = 0.21; n = 14; k = 5) and common-k second ($M = 0.19$, SD = 0.21; n = 16; k = 6) thresholds (Table 1).

Measurement property outcomes

As noted, none of the included studies investigated differences in average item discrimination between thresholds. Therefore, we focus solely on differences in item difficulty (based only on CTT, as IRT difficult parameter estimates were not reported in our sample). The effect sizes for this outcome by RT threshold typology pairing is shown in the left-panel of Fig. 4.

Differences in item difficulty based on RT threshold was investigated across six unique studies, which led to a total sample size of 53 effect sizes. These 53 effect sizes were unequally distributed across all six typology pairings, with the NED-RT pairing the most studied (n = 21), while all other pairings possessed 10 or less effect sizes. Across this sample, the average Cohen's h was equal to 0.02 ($SD = 0.03$), indicating that the choice of RT threshold was associated with practically no difference in item difficulty estimates. This is further supported in examining comparison differences between the NED-RT typology pairing, which produced an average Cohen's h of 0.01 ($SD = 0.02$).

Performance outcomes

Test performance effect size differences for each RT threshold typology pairing are provided in the right-panel of Fig. 4. A total of two studies examined this outcome, producing 16 effect sizes. These effect sizes came primarily from NED–NED (n = 8) and RT–RT

(n = 7) comparisons, with one based on a NED–RT comparison. Across effect sizes, no practical difference for the overall effect of threshold procedures on test performance was observed. Specifically, the average difference was equal to 0.05 *SD* ($p = 0.16$, 95% CI [−0.12, 0.23]). Furthermore, the effect size estimates of test performance were found to be homogeneous across all studies ($I^2 = 0$), suggesting no need for a moderator analysis.

Discussion

Given the increased attention to identifying RG behavior in the literature, this study looked to investigate how the choice of RT threshold procedure influences descriptive, measurement property, and test performance outcome variables. To accomplish this, a meta-analysis was conducted in which studies comparing two or more RT threshold procedures on the same empirical dataset were sampled. A number of key takeaways can be summarized.

First, the choice of RT threshold procedure was found to be associated with non-negligible differences in the proportions of RG responses and responders identified. In particular, the largest differences were observed when comparing NED to RT and RTRA typologies, with NED thresholds generally producing smaller proportions (h was as large as 1.2). Considering that the NED typology was the most investigated in our sample, there is some evidence to indicate that researchers/practitioners may generally employ conservative thresholds to avoid false positive (type I error) RG classifications, supporting Wise's (2017) hypothesis. Although minimizing type I errors is desirable, adopting conservative RT thresholds may lead to underclassifying RG (i.e., failing to identify all RG responses; Wise, 2017). When this is the case, recent simulation work shows that there is greater potential for increased bias on item and ability parameter estimates than overclassifying (i.e., identifying all true RG responses, but including type II errors) such responses (Rios, 2021b). Thus, conservative approaches, such as those observed within the NED typology, may be less effective in reducing bias.

Although non-negligible divergences were observed in RG responses and examinees, these differences were not found to be associated with statistically significant differences in average item difficulty and test performance once filtering RG responses. One likely reason for this finding is that simulation research has shown negligible effects of RG on both item difficulty estimates and aggregate-level performance, even when rates of RG exceed 10% of all item responses (see Rios & Soland, 2020a, 2020b). Thus, the differences in RG identification observed across RT threshold procedures would be expected to have minimal practical impact on the outcomes examined. Though no differences were found for aggregate-level inferences, the choice of RT threshold may still impact individual-level estimates of ability and classification decisions (see Rios & Soland, 2021). Furthermore, it is possible that if differential RG is present between subgroups, variability in RG identification may be associated with incorrect inferences of measurement properties and subgroup test performance differences (see Rios, 2021a). Future research is needed in these areas.

Implications

The findings from this study have a number of implications for practice. To begin with, if examinee motivation is a potential concern, practitioners should document RG to

support the validity of score-based inferences. Regardless of the procedure chosen to identify RG responses, the very act of collecting validity evidence based on response processes can assist in improving the credibility of results. This holds particularly true for contexts in which the aim of a test is to make inferences about scores at the aggregate-level, as our results showed minimal differences across RT typologies for this outcome. Thus, if filtering RG responses to improve inferences concerning group test performance, the actual threshold procedure chosen may be of less importance than communicating to stakeholders that such deleterious responses have been identified and trimmed prior to score reporting.

Although differences in RT typologies were shown to be of little consequence when making inferences about total sample performance, such differences can impact individual examinee score inferences (Rios & Soland, 2021). Rios and Soland (2021) highlighted how some states in the U.S. utilize scores obtained from low-stakes end-of-year assessments to make remediation decisions for individual students. When this is the case, they showed that inaccuracies in RG identification, particularly type II errors, can have deleterious effects on cut-score classifications. Thus, in such contexts (i.e., making individual examinee inferences), it is recommended that practitioners avoid employing NED typologies, given their conservative nature and potential for false negative classifications. Instead, utilizing threshold procedures that leverage empirical information, such as those within the RT and RTRA typologies, may be preferable. However, the decision to employ either RT or RTRA approaches will be driven by a number of factors, such as sample size as well as RT distribution and test characteristics.

Concerning sample size, if the number of examinees for a given sample is small, the RT approach may be most preferable. Within this typology, the decision to employ percentile versus bimodal distribution procedures will be determined based on item RT distribution characteristics. In many cases, RT distributions will be unimodal, which makes the employment of percentile procedures most appropriate (Wise, 2017).

Assuming a large sample size and variability across a response time distribution, RTRA approaches may be employed for maximal RG response identification; however, this still requires that items are neither too difficult nor easy. If the latter is the case, one alternative is to employ a hybrid approach, such as that proposed by Rios and Guo (2020). These authors suggested that a RTRA procedure can be utilized for all possible items, and when ineffective (e.g., when an item is too easy or difficult), the RTRA procedure can be replaced by a RT approach. Such a hybrid method allows for maximum RG response identification, while adhering to the basic assumptions underlying the RTRA procedures.

Further investigations on effective validation efforts are needed to provide more concrete recommendations on the best RT threshold approaches for practice. Where possible, practitioners are encouraged to not rely solely on RT procedures for identifying noneffortful responding. A number of alternative non-response time methods have been employed to gauge test-taking effort, such as item skipping (e.g., Liu & Hau, 2020), item multimedia interactions (e.g., Harmes & Wise, 2016), eye-tracking (e.g., Lindner et al., 2014; Toth & Campbell, 2019), electroencephalography (EEG; e.g., Halderman et al., 2021), retroactive video evaluations of emotional ratings (e.g., Lehman & Zapata-Rivera, 2018), and cognitive interviewing (e.g., Hopfenbeck & Kjærnsli, 2016). Although not all

of these procedures will be readily available, practitioners may benefit from utilizing multiple methods to strengthen claims concerning examining engagement.

Limitations

The findings from this study should be interpreted in light of a number of limitations. First, although a concerted effort was made to include a diverse set of search strategies, with particular focus on identifying grey literature, some pertinent papers may have been missed. This might have been due to only including English language papers and failing to consult professional research organization listservs. Second, although some of the dependent variables examined were based on 80 or more effect sizes, others possessed as few as 16. For the outcomes with small samples, this limited the: (a) evaluation of publication bias, which require large sample sizes to evaluate the symmetry of an effect size distribution (Lau et al., 2006); (b) ability to conduct moderator analyses; and (c) generalizability of the findings. Given these limitations, it is recommended that readers consider our findings to be descriptive and preliminary in nature (as most meta-analyses are).

Directions for future research

Given that true RG is unknowable in operational testing contexts, an important area of future research is to develop improved validation efforts of RG identification methods. To date, two popular indices for collecting validity evidence of RG classifications have been proposed by Wise and Kong (2005): the proportion correct rate for RG responses identified, and the correlation between RTE and test performance. The rationale for these indices is that RG responses should on average have a proportion correct rate approximately equal to chance as they largely reflect random responding, while increased effort should be positively correlated with test performance (see Silm et al., 2020; Wise, 2017).

Although theoretically sound arguments, these criteria are inherently flawed in operational settings. Concerning the proportion correct rate, this form of evidence is of little utility for methods that set RT thresholds in part based on response accuracy metrics. In such cases, reporting the proportion correct rate does little to provide validity evidence, given that it is simply a descriptive by-product of the procedure, and thus, leads to circular reasoning concerning its validity.

Turning to the association between RTE and total score as a source of validity evidence, the assumption is that a higher correlation is associated with improved validity. However, this approach is inherently limited in two ways. First, the size of the correlation is largely dependent on both the number of examinees engaging in RG and the number of RG responses. If these rates are low, the response time effort distribution will be negatively skewed, leading to artificial attenuation of the true correlation coefficients (see Rios et al., 2017). Closely related, the second limitation is that RT thresholds that identify a high percentage of RG responses and responders, while maintaining low proportion correct rates for RG responses will generally have higher correlations. This naturally benefits RTRA thresholds, due largely in part to the utilization of response accuracy metrics.

Given these limitations, it is difficult to support the use of proportion correct rates and correlations between RTE and test performance as sources of validity evidence; and thus, this is the reason why these indices were excluded from the present study. With that said, future approaches should consider alternative criteria for validity evidence that are external to the test. According to the argument-based approach to validation, the inference that response latencies are reflective of the claim that an item response is noneffortful requires both a warrant and backing. In this case, the warrant is the rule that assigns RG classifications to observed response latencies (i.e., RT thresholds), while the backing of this warrant is the appropriateness of the RT threshold. The current empirical approaches adopted to evaluate the appropriateness of this claim have a number of limitations, as noted above; however, there is the potential for future researchers to consider additional sources not yet employed. As an example, subject matter experts, examinees, and test users could evaluate the appropriateness of established RT thresholds for a given testing program by providing their expectations on the minimum time needed for an average examinee to read the item stem and response options, solve its challenge, and select an answer. This qualitative approach would require a much more involved process than the current heuristic methods, but could provide judgmental evidence that may be sufficient to support claims around RG. Further research in identifying effective validation evidence for RT threshold procedures is needed. With that said, it is our hope that this study has aided practitioners in better understanding how RT threshold procedures differ to date and laid the foundation for future research.

Abbreviations
NED: No empirical data; RG: Rapid guessing; RT: Response time(s); RTE: Response time effort; RTRA: Response time and response accuracy.

Acknowledgements
The authors would like to thank Samuel Ihlenfeldt from the Improving Educational Measurement Practice Lab at the University of Minnesota for his helpful comments on an earlier draft.

Authors' contributions
The presented idea was conceived of by JR. JD identified, coded, and analyzed articles. Both authors contributed to writing, conducted critical revisions, and read and approved the final manuscript.

Funding
Not applicable.

Availability of data and materials
The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 4 September 2020 Accepted: 13 July 2021
Published online: 17 August 2021

References

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

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.

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. (2009). *Introduction to meta-analysis* (2nd ed.). Wiley.

Champely, S. (2020). *Package "pwr"* (Package version 1.3-0) [Computer software]. Retrieved from https://cran.r-project.org/web/packages/pwr/pwr.pdf.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Lawrence Erlbaum Associates.

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

Del Re, A. C. (2020). *Package "Compute.es"* (Package version 0.2-5) [Computer software]. Retrieved from https://cran.r-project.org/web/packages/compute.es/compute.es.pdf.

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.

Fisher, Z., Tipton, E., & Zhipeng, H. (2017). *Package "robumeta"* (Package version 2.0) [Computer Software]. Retrieved from https://cran.r-project.org/web/packages/robumeta/robumeta.pdf.

*Goldhammer, F., Martens, T., Christoph, G., & Lüdtke, O. (2016). *Test-taking engagement in PIAAC*. (OECD Education Working Paper Series, No. 133). Organisation for Economic Co-operation and Development. https://doi.org/10.1787/5jlzfl6fhxs2-en.

Goldhammer, F., Naumann, J., Rölke, H., Stelter, A., & Tóth, K. (2017). Relating product data to process data from computer-based competency assessment. In D. Leutner, J. Fleischer, J. Grünkorn, & E. Klieme (Eds.), *Competence assessment in education. Methodology of educational measurement and assessment* (pp. 407–425). Cham: Springer. https://doi.org/10.1007/978-3-319-50030-0_24

*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.

Haddaway, N. R., Collins, A. M., Coughlin, D., & Kirk, S. (2015). The role of Google Scholar in evidence reviews and its applicability to grey literature searching. *PLoS ONE, 10*(9), e013823. https://doi.org/10.1371/journal.pone.013823

Halderman, L. K., Finn, B., Lockwood, J. R., Long, N. M., & Kahana, M. J. (2021). EEG correlates of engagement during assessment. *Educational Testing Service*. https://doi.org/10.1002/ets2.12312

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: IGI Global.

Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods, 1*(1), 39–65.

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

Hopfenbeck, T. N., & Kjærnsli, M. (2016). Students' test motivation in PISA: The case of Norway. *The Curriculum Journal, 27*(3), 406–422.

*Johnston, M. M. (2016). *Applying solution behavior thresholds to a noncognitive measure to identify rapid responders: An empirical investigation* (Unpublished doctoral dissertation). James Madison University.

Kiplinger, V. L., & Linn, R. L. (1994 April 4–8). *Linking statewide tests to the National Assessment of Educational Progress: Stability of results*. (Paper presentation). American Educational Research Association annual meeting, New Orleans, Louisiana, United States.

*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.

*Kong, X. J., Wise, S. L., Harmes, J. C., & Yang, S. (2006, April 8–10). *Motivational effects of praise in response-time based feedback: A follow-up study of the effort-monitoring CBT*. (Paper presentation). National Council on Measurement in Education 68th annual meeting, San Francisco, CA, United States.

*Kroehne, U., Deribo, T., & Goldhammer, F. (2020). Rapid guessing rates across administration mode and test setting. *Psychological Test and Assessment Modeling, 62*(2), 147–177.

Lau, J., Ioannidis, J. P., Terrin, N., Schmid, C. H., & Olkin, I. (2006). The case of the misleading funnel plot. *BMJ, 333*(7568), 597–600.

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.

Lehman, B. A., & Zapata-Rivera, D. (2018). Student emotions in conversation-based assessments. *IEEE Transactions on Learning Technologies, 11*(1), 41–53.

Lindner, M. A., Eitel, A., Thoma, G. B., Dalehefte, I. M., Ihme, J. M., & Köller, O. (2014). Tracking the decision-making process in multiple-choice assessment: Evidence from eye movements. *Applied Cognitive Psychology, 28*(5), 738–752.

*Lindner, M. A., Lüdtke, O., Grund, S., & Köller, O. (2017). The merits of representational pictures in educational assessment: Evidence for cognitive and motivational effects in a time-on-task analysis. *Contemporary Educational Psychology, 51*, 482–492.

Liu, Y., & Hau, K. T. (2020). Measuring motivation to take low-stakes large-scale test: New model based on analyses of "Participant-Own-Defined" missingness. *Educational and Psychological Measurement, 80*(6), 1115–1144.

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

Mittelhaëuser, M. A., Béguin, A. A., & Sijtsma, K. (2015). The effect of differential motivation on irt linking. *Journal of Educational Measurement, 52*(3), 339–358.

Osborne, J. W., & Blanchard, M. R. (2011). Random responding from participants is a threat to the validity of social science research results. *Frontiers in Psychology, 1*, 220. https://doi.org/10.3389/fpsyg.2010.00220

*Pastor, D. A., Ong, T. Q., & Strickman, S. N. (2015 April 16–20). *Patterns of solution behavior across items in low-stakes assessments*. (Paper presentation). American Educational Research Association annual meeting, Chicago, IL, United States.

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.
Rios, J. A. (2021a). Is differential noneffortful responding associated with type I error in measurement invariance testing? *Educational and Psychological Measurement*. https://doi.org/10.1177/0013164421990429 Advanced online publication.
Rios, J. A. (2021b). Assessing the accuracy of parameter estimates in the presence of rapid guessing misclassifications. *Educational and Psychological Measurement*. https://doi.org/10.1177/00131644211003640 Advanced online publication.
Rios, J. A. (2021c). Improving test-taking motivation on low-stakes educational assessments: A meta-analysis of interventions. *Applied Measurement in Education*. https://doi.org/10.1080/08957347.2021.1890741 Advanced online publication.
*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*. 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.
Rios, J. A., Liu, O. L., & Bridgeman, B. (2014). Identifying low-effort examinees on student learning outcomes assessment: A comparison of two approaches. *New Directions for Institutional Research, 2014*(161), 69–82.
*Rios, J. A., & Soland, J. (2020a). *Correlates of test-taking effort on PISA: An examination of item, examinee, and country characteristics*. (Manuscript in Preparation). Department of Educational Psychology, University of Minnesota.
Rios, J. A., & Soland, J. (2020b). Parameter estimation accuracy of the Effort-Moderated IRT model under multiple assumption violations. *Educational and Psychological Measurement*. https://doi.org/10.1177/0013164420949896 Advanced online publication.
Rios, J. A., & Soland, J. (2021). Investigating the impact of noneffortful responses on individual-level scores: Can the Effort-Moderated IRT model serve as a solution? *Applied Psychological Measurement, 42*, 359–375.
*Sacco, C. (2020). *Estimation of test-taking effort on INVALSI computer-based tests*. INVALSI. Retrieved from https://www.invalsi.it/download2/wp/wp50_Sacco.pdf.
*Şahin, F. (2017). *Exploring validity of computer-based test scores with examinees' response behaviors and response times*. (Unpublished doctoral dissertation). State University of New York at Albany.
*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.
Schnipke, D. L. (1995, April 19–21). *Assessing speededness in computer-based tests using item response times*. (Paper presentation). National Council on Measurement in Education annual meeting, San Francisco, CA, United States.
Schnipke, D. L., & Scrams, D. J. (1997). Modeling item response times with a two-state mixture model: A new method of measuring speededness. *Journal of Educational Measurement, 34*(3), 213–232.
*Silm, G., Must, O., & Täht, K. (2019). Predicting performance in a low-stakes test using self-reported and time-based measures of effort. *Trames, 23*(3), 353–376.
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. https://doi.org/10.1016/j.edurev.2020.100335
*Soland, J., Kuhfeld, M., & Rios, J. A. (in press). Comparing different response time threshold setting methods to detect low effort on a large-scale assessment. *Large-Scale Assessments in Education*.
Toth, A. J., & Campbell, M. J. (2019). Investigating sex differences, cognitive effort, strategy, and performance on a computerised version of the mental rotations test via eye tracking. *Scientific Reports, 9*(1), 1–11.
van Barnevald, 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
Viechtbauer, W. (2020). *Package "Metafor"* (Version 2.4-0) [Computer software]. Retrieved from https://cran.r-project.org/web/packages/metafor/metafor.pdf.
Wang, C., & Xu, G. (2015). A mixture hierarchical model for response times and response accuracy. *British Journal of Mathematical and Statistical Psychology, 68*(3), 456–477.
Wise, S. L. (2017). Rapid-guessing behavior: Its identification, interpretation, and implications. *Educational Measurement: Issues and Practice, 36*(4), 52–61.
*Wise, S. L. (2019). An information-based approach to identifying rapid-guessing thresholds. *Applied Measurement in Education, 32*(4), 325–336.
Wise, S. L., & DeMars, C. E. (2009). A clarification of the effects of rapid guessing on coefficient α: A note on Attali's "Reliability of speeded number-right multiple-choice tests." *Applied Psychological Measurement, 33*(6), 488–490. 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.
Wise, S. L., & Kingsbury, G. G. (2016). Modeling student test-taking motivation in the context of an adaptive achievement test. *Journal of Educational Measurement, 53*(1), 86–105.
Wise, S. L., 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.
Wise, S. L., & Kuhfeld, M. R. (2020a). 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). Routledge.
*Wise, S. L., & Kuhfeld, M. R. (2020b). Using retest data to evaluate and improve effort-moderated scoring. *Journal of Educational Measurement*. https://doi.org/10.1111/jedm.12275

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

Wise, S. L., Ma, L., Cronin, J., & Theaker, R. A. (2013 April 27–May 1). *Student test-taking effort and the assessment of student growth in evaluating teacher effectiveness*. (Paper presentation). American Educational Research Association annual meeting, San Francisco, CA, United States.

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.

Publisher's Note