Applied Psychological Measurement
2021, Vol. 0(0) 1–13

DOI: 10.1177/01466216211051719
journals.sagepub.com/home/apm
SAGE

Quantifying the Distorting Effect of Rapid Guessing on Estimates of Coefficient Alpha

Joseph A. Rios[1] and **Jiayi Deng**[1]

Abstract

An underlying threat to the validity of reliability measures is the introduction of systematic variance in examinee scores from unintended constructs that differ from those assessed. One construct-irrelevant behavior that has gained increased attention in the literature is rapid guessing (RG), which occurs when examinees answer quickly with intentional disregard for item content. To examine the degree of distortion in coefficient alpha due to RG, this study compared alpha estimates between conditions in which simulees engaged in full solution (i.e., do not engage in RG) versus partial RG behavior. This was done by conducting a simulation study in which the percentage and ability characteristics of rapid responders as well as the percentage and pattern of RG were manipulated. After controlling for test length and difficulty, the average degree of distortion in estimates of coefficient alpha due to RG ranged from −.04 to .02 across 144 conditions. Although slight differences were noted between conditions differing in RG pattern and RG responder ability, the findings from this study suggest that estimates of coefficient alpha are largely robust to the presence of RG due to cognitive fatigue and a low perceived probability of success.

Keywords

Reliability, coefficient alpha, rapid guessing, non-effortful responding, test-taking motivation

Introduction

Test reliability is a cornerstone of operational analyses, as it provides information about the consistency of examinee scores due to random measurement error and provides a precursor to the validity of score use and interpretation (American Educational Research Association, et al., 2014). Although multiple types of reliability exist to address various forms of measurement error, internal consistency, which provides a measure of item response homogeneity, has become the most popular in practice, given that it can be estimated based on one test form and one testing occasion. While there are multiple approaches to estimating internal consistency (e.g., split-half, person

[1]Department of Educational Psychology, University of Minnesota, Minneapolis, MN

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

separation), coefficient alpha (α; Cronbach, 1951) is employed most in research and operational settings. The popularity of this approach stems from its ability to handle both dichotomous and ordinal data for a composite score (see McNeish, 2018). One way to define α for a given test is:

$$\alpha = \frac{\left(N \times \underline{c}\right)}{\underline{v} + (N-1)\underline{c}}, \tag{1}$$

where N equals the number of test items, $\underline{c}$ equals the mean of all item covariances, and $\underline{v}$ equals the average variance of each item.

An underlying threat to the validity of reliability measures, such as α, is the introduction of systematic variance into examinee scores from unintended constructs that differ from those assessed. This validity threat can occur when examinees are provided with insufficient time to adequately solve an item (i.e., test speededness) and/or when administered an assessment in low-stakes contexts (i.e., test performance has minimal perceived personal consequences for examinees; Wise, 2017). In these circumstances, examinees may engage in construct-irrelevant behavior, such as rapid guessing (RG) or responding with intentional disregard for item content by providing a response in a time that would not allow one to read the entire item, solve the presented problem, and provide an answer. The focus of this paper is on investigating the difference in coefficient alpha when examinees employ full solution (i.e., examinees do not engage in RG) versus partial RG behavior in low-stakes testing contexts. The sections that follow discuss the influence of test speededness on α, reviews applied analyses that have examined the association between RG and α in low-stakes tests, and provides a rationale for the current simulation study.

The Influence of Test Speededness on Coefficient Alpha Estimates

The concern of the potential biasing effect of test speededness on α has been well documented for years. As an example, researchers, such as Crocker and Algina (1986, p. 145), have noted that when a time limit is not long enough for examinees, coefficient α will likely be artificially inflated, due to consistencies in performance on items towards the end of the test. However, such a conclusion assumes that examinees omit and/or incorrectly respond to all speeded items. In fact, when RG occurs in a context in which there are no penalties for guessing, Attali (2005) showed via analytic derivations that reliability can be deflated for a test that is speeded, with greater underestimation expected as the inter-item correlations between speeded item responses decrease. Based on data simulations, this finding was supported by Hong and Cheng (2019) who found that under RG conditions coefficient α is generally underestimated as the percentage of speeded items and examinees engaging in test speededness increases. The work conducted by Attali (2005) and Hong and Cheng (2019) has improved the field's thinking about the role of RG on estimates of coefficient α; however, their research has been only confined to conditions of RG due to inadequate time constraints (i.e., test speededness).

The Impact of RG on Reliability Estimates in Low-Stakes Tests

An area of the reliability literature that has received considerably less attention is when RG occurs due to examinees' having little to no personal consequences associated with their test performance, a common circumstance in low-stakes assessment contexts (Wise, 2017). Unlike test speededness, RG on low-stakes assessments may not solely occur at the end of the test, but rather may take place throughout, with most items receiving RG responses (Wise, 2006). To investigate the association between RG and α, an empirical review of literature was conducted. This allowed

for us to summarize the descriptive characteristics and quantify the results and heterogeneity from individual studies in this area.

To be included in the review, prior empirical studies had to compare the difference in reliability estimates between datasets that included (i.e., contaminated reliability) and excluded (i.e., filtered reliability) RG responses in low-stakes multiple-choice tests (i.e., studies investigating non-effortful responding on surveys were excluded, such as Steedle et al., 2019). Although there are multiple approaches to identifying aberrant responding, we only reviewed studies that used a response time (RT) threshold to classify RG responses. This was done because the use of RT thresholds allows for the identification of RG at the individual item response level, which is important because potential distortions in reliability estimates are associated with construct-irrelevant noise introduced into item covariances (Wise & DeMars, 2009). Thus, studies identifying aberrant responders using self-report measures of test-taking effort were excluded (e.g., Sundre & Wise, 2003; Wise & DeMars, 2005; Wise et al., 2006). Furthermore, papers that investigated differences between contaminated and filtered data using item response theory measures of reliability (e.g., marginal reliability) were not analyzed (e.g., DeMars, 2007). This literature review was completed in September 2020.

Table 1 shows effect sizes for reliability estimate differences between contaminated and filtered data as well as descriptive information across seven empirical studies and 26 effect sizes. Multiple reliability estimates listed for one study indicate that: (a) multiple tests or scales were used in the study; (b) multiple response time thresholds were used to identify RG; or (c) multiple examinee-level filtering criteria (i.e., criteria established to listwise delete data for an examinee that engaged in RG for a predefined number of items) were employed to address RG. Across the included studies, the average sample size was 368 examinees (ranging from 103 to 586), and the examined tests included between 35 and 108 items, with all tests composed solely of selected-response items. For the 26 effect sizes, researchers filtered RG responses using primarily examinee-level filtering ($n = 18$), while response-level filtering (i.e., treating RG responses as missing data) was utilized for eight effect sizes. Among those studies employing examinee-level filtering, six criteria were used to determine the removal of examinees based on engaging in RG on 10% ($n = 13$), 20% ($n = 1$), 25% ($n = 1$), 30% ($n = 1$), 40% ($n = 1$), and 50% ($n = 1$) or more of items.

As shown in Table 1, the percentage of RG responses in the included sample ranged from 6% to 13% (mean of 9%) and the percentage of examinees engaging in RG ranged from 2.1% to 58% (mean of 12%). Of the 26 effect sizes, 21 showed a positive distortion (i.e., the difference between contaminated and filtered reliability estimates was positive). Of the remaining five effect sizes, two exhibited no change, while the other three demonstrated negative distortion. Overall, the average difference in reliability estimates between contaminated and filtered data was .06 ($SD = .06$); however, a large degree of variation between studies was observed with distortion in α values ranging from −.05 to .18. This variation was not attributable to filtering procedure, as examinee- and response-level filtering produced nearly identical effect sizes (the average difference was .01 favoring the former procedure).

A closer inspection of Table 1 shows that for three studies (Abdelfattah, 2007; Smith et al., 2013; Wise & Kong, 2005), the reliability estimate differences tended to be consistently within one *SD* of the mean (ranging from −.05 to .06) across different filtering criteria. In contrast, a number of studies conducted by Wise and colleagues (Wise, 2006; Wise & DeMars, 2009, 2010) produced distortion effects that were greater than two *SD*s from the mean, with differences ranging from .13 to .18. Although six of seven studies showed consistency in α values, the work conducted by Liu et al. (2015) did not. Specifically, the varied distortion effects (−.01 to .13) could be attributed to differences in α across different subtests of the ETS Proficiency Profile. For instance, for the overall test and the math subdomain, observed differences in alpha values were minimal

Table 1. Effect Sizes of Reliability Estimate Differences between Unfiltered and Filtered RG Data.

Study	Assessment	Test length	Grade	Filtering approach	EF criteria (% rapid guesses)	% rapid guesses in data	% examinees with RG	Unfiltered α	Filtered α	α difference[a]
Abdelfattah (2007)	TIMMS	35	K-12	RF	—	10%	—	.78	.74	.04
				RF	—	10%	—	.78	.78	.00
				RF	—	10%	—	.78	.83	−.05
Smith et al. (2013)	ILT	60	K-12	RF	—	—	—	.89	.94	−.05
				EF	50%	—	7%	.89	.88	.01
				EF	25%	—	20%	.89	.86	.03
				EF	10%	—	29%	.89	.84	.05
Wise (2006)	ILT	60	Postsecondary	RF	—	—	26%	.88	.75	.13
Wise and Kong (2005)	ILT	80	Postsecondary	EF	40%	—	2.1%	.83	.80	.03
				EF	30%	—	3.8%	.83	.79	.04
				EF	20%	—	4.2%	.83	.79	.04
				EF	10%	—	7.4%	.83	.77	.06
Wise and DeMars (2009)	ILT	60	Postsecondary	RF	—	6%	—	.88	.75	.13
	SRT	49	Postsecondary	RF	—	11%	—	.77	.64	.13
	IST	NA	Postsecondary	RF	—	13%	—	.76	.61	.15
Wise and DeMars (2010)	OCK	40	Postsecondary	EF	10	6%	11%	.84	.66	.18
Liu et al. (2015)	EPP	108	Postsecondary	EF	10	—	5%	.93	.92	.01
				EF	10	—	58%	.95	.93	.02
	EPP-R	NA	Postsecondary	EF	10	—	—	.82	.79	.03
				EF	10	—	—	.90	.79	.11
	EPP-W	NA	Postsecondary	EF	10	—	—	.66	.61	.05
				EF	10	—	—	.81	.68	.13
	EPP-M	NA	Postsecondary	EF	10	—	—	.84	.84	.00
				EF	10	—	—	.83	.84	−.01
	EPP-CT	NA	Postsecondary	EF	10	—	—	.78	.72	.06
				EF	10	—	—	.84	.72	.12

Note. All studies with multiple effect sizes, except for Wise and DeMars (2009), employed a single sample. EF = examinee-level filtering; RF = response-level filtering; TIMMS = trends in international mathematics and science study; ILT = information literacy test; SRT = science reasoning test; IST = information systems test; OCK = oral communication knowledge test; EPP = ETS proficiency profile (EPP); EPP-R = reading subtest in EPP; EPP-W = writing subtest in EPP; EPP-M = mathematic subtest in EPP; EPP-CT = critical thinking subtest in EPP; NA = missing information.

[a]Reliability estimate difference is equal to reliability estimates with RG included subtracted by reliability estimates with RG excluded.

(−.01 to .02); however, for the reading, writing, and critical thinking subdomains, distortion in α varied from .03 to .13.

The results from this review suggest that RG can be associated with both inflated/deflated and small/large differences in α values. However, when α was negatively distorted, the difference in values tended to be within one *SD* of the sample mean, supporting the work by Attali (2005) and Hong and Cheng (2019). In contrast, much larger absolute differences were observed in cases of positive α distortions (α could be more than two *SD*s greater than the mean). One observation of interest was that α inflation may have been related to the examinee-level filtering criteria employed by primary researchers. For instance, when using the same sample and comparing filtered and unfiltered datasets, Wise and Kong (2005) found that the alpha difference increased from .03 to .06 when respectively removing examinees engaging in RG for 40% and 10% of items. Similarly, the alpha difference between a RG response criterion of 50% and 10% was .01 and .05 in work conducted by Smith et al. (2013). These results were likely related to deleting data from a higher percentage of examinees for the stricter examinee-level filtering criteria, which thus, was associated with the removal of score heterogeneity, leading to lower filtered alpha estimates and greater alpha distortion. This may have been the case for the large positive distortion values in Liu et al. (2015), given that these researchers also utilized a 10% RG response criterion for removal of examinee data.

Although our review produced a number of interesting findings, no clear pattern was observed. Given the small sample size, moderator analyses beyond filtering procedure could not be conducted to evaluate how aspects of the datasets were related to the varied differences. Furthermore, due to the reliance on applied analyses, a major limitation of the current literature is that it is difficult to evaluate the expected degree and underlying mechanisms of distortion in α estimates due to RG. In operational contexts, the influence of RG on estimates of α are confounded with the accuracy of correctly identifying and the criteria for filtering RG responses. As an example, when valid item responses are incorrectly classified as RG, artificial deflation of reliability estimates can occur, due to the false removal of valid score variance. Since the identification of RG behavior is based on proxy information (e.g., item response times), score users can never be sure of the accuracy of RG classifications in operational contexts (Rios, 2021a). Given these limitations, causal investigations using simulated data are needed to better investigate the influence of RG on estimates of α.

Research Objectives

Systematically investigating the effect of RG on coefficient α for low-stakes testing contexts is of critical importance in bringing attention to the deleterious effects that ignoring RG responses can have on inferences concerning both measurement properties and test scores. This is of particular importance as low-stakes testing contexts are increasingly employed in educational accountability efforts (e.g., Every Student Succeeds Act) and international education studies (e.g., Programme for International Student Assessment). Given that most operational testing programs rarely collect validity evidence to address potential issues of RG (Hubley & Zumbo, 2017; Wise & Kuhfeld, 2020), the degree to which reliability estimates may be systematically distorted is unknown. This is particularly a concern given that the only evidence to date is based on disparate results from applied analyses. The findings from these studies are confounded by variations in sample and assessment characteristics, which have been shown to be associated with differing rates of RG behavior (Wise, 2017). Thus, a more systematic analysis is needed that goes beyond the context of test speededness by examining the occurrence of RG throughout a test as opposed to primarily at the end.

To address these issues, this paper examines the magnitude and direction of difference in estimates of coefficient alpha under conditions in which examinees engage in full solution (i.e., do not engage in RG) and partial RG behavior. This is accomplished via a simulated context in which a low-stakes assessment measuring a single composite score via multiple-choice items is administered to a sample of simulees with no penalty for guessing. Two previously unexamined RG patterns found to occur frequently in low-stakes testing contexts are generated—RG associated with: (a) a low perceived probability of success; and (b) cognitive fatigue (see Wise, 2017). Distortion in coefficient α estimates is assessed when comparing clean (i.e., all simulees employed full effort and did not engage in RG) and contaminated (i.e., RG responses are included) data. Results from this study have the potential to inform practitioners about the possible consequences of ignoring the presence of RG when making inferences about the reliability of test scores.

Method

Data Generation

To evaluate the influence of RG on estimates of test reliability, item response data for 5000 simulees were generated for an assessment consisting of n multiple-choice items with four-response options. This was done via the unidimensional three-parameter logistic (3PL) model

$$P_i(\theta) = c_i + \frac{1 - c_i}{1 + exp\{-1.7a_i(\theta - b_i)\}}, \tag{2}$$

where $P_i(\theta)$ indicates the probability of answering item i correctly; a_i is the item discrimination parameter for item i; b_i is the item difficulty parameter for item i; and c_i is the pseudo-guessing parameter for item i. This was done via a three-step process.

First, based on this model, the probabilities of answering each item correctly were calculated for each simulee based on sampled item and ability parameters. Concerning the former, generating item parameters were sampled from NAEP math assessments (more information on form creation is described below). For non-rapid guessers, simulee ability parameters (θ) were randomly sampled from $N(0, 1)$ (more information on ability sampling for rapid guessers is provided in the next section). Second, for RG responses, the true probability of correctly answering an item was replaced with the chance probability (.25) to reflect random guessing. Third, for each simulee by item interaction, the probability of a correct response was compared to a random number taken from a uniform distribution ranging from 0 to 1. If the random number was greater than the probability, the simulee was given an incorrect response for that item; otherwise, a correct response was generated.

Conditions

Six factors were examined as potential moderating effects of the association between RG and α estimates. Two of these factors were incorporated to influence the test characteristics under investigation: (a) test length (40 and 80 items) and (b) test difficulty (easy, moderate, and difficult). The remaining four factors were included to manipulate RG: (a) percentage of rapid guessers in the sample (20% and 40%); (b) ability characteristics of rapid guessers (low and average ability); (c) percentage of RG responses (5%, 10%, and 15%); and (d) rapid guessing pattern (difficulty-based and progressive). These six variables were fully crossed producing 144 conditions, with each condition replicated 100 times. Below, we provide a description of each factor and their respective levels.

Table 2. Average Item Parameters by Test Form.

	40 items			80 items		
Test difficulty	a	b	c	a	b	c
Easy	1.00 (.53)	−1.01 (.98)	.16 (.06)	.92 (.45)	−.94 (.81)	.17 (.07)
Moderate	.97 (.42)	.01 (1.01)	.16 (.07)	1.00 (.57)	.01 (.96)	.18 (.08)
Hard	1.01 (.41)	.99 (.99)	.20 (.1)	1.08 (.56)	1.00 (.86)	.20 (.09)

Note. Item parameter standard deviations are provided in parentheses.

Test length. Given that estimates of internal consistency are influenced by the number of items on a test (Attali, 2005), test length was included as a moderating variable with two levels: 40 and 80 items. These levels reflect common test lengths observed in low-stakes assessment contexts (e.g., DeMars, 2007; Smith et al., 2013).

Test difficulty. Prior work conducted by Hong and Cheng (2019) found that coefficient α was influenced by the composition of items when test speediness was present, with highly discriminating and easy items associated with inflated reliability estimates. This result may largely be influenced by the decreased score variance observed for easy items in the presence of RG (Rios et al., 2017). To account for this influence, three levels of test difficulty were included in the simulation design: easy (b: $M \sim -1$, $SD = 1$), moderate (b: $M \sim 0$, $SD = 1$), and difficult (b: $M \sim 1$, $SD = 1$). This was done by building six test forms (2 test lengths × 3 item difficulty levels) that reflected the targeted difficulty levels based on sampling estimated item parameters from NAEP math assessments (see Table 2 for item parameter descriptives).

Percentage of rapid guessers. The percentage of simulees engaging in RG was constrained to one of two levels: 20% or 40% of the sample. These percentages reflect the range of examinees found to employ RG in operational low-stakes test administrations (e.g., Wise, 2006; Wise & DeMars, 2006; Wise & DeMars, 2010).

Ability characteristics of rapid guessers. There is some debate as to whether RG is related to examinees true underlying ability or whether such a relationship has a non-negligible impact on ability parameter estimation accuracy. Proponents of the former point to literature that has found a negative association between RG behavior and indicators of academic performance, such as grade point average (see Soland et al., 2019). In contrast, others have made the point that RG can occur due to low perceived task value, which may occur irrespective of examinee ability (Wise, 2015). To reflect this debate, two levels were manipulated in which simulees engaging in RG were sampled to possess average (N [0, 1]) or low ability (N [−.5, 1]). The latter condition reflects empirical work conducted by Rios et al. (2017), which found that rapid guessers scored approximately 0.5 *SD*s lower than their non-rapid guesser counterparts on prior high-stakes measures of ability (SAT/ACT).

Percentage of RG responses. Four percentages of RG responses in the item response matrix were simulated: 0%, 5%, 10%, and 15%. These percentages were created via a two-step process. For each rapid guesser, a random number of RG responses ranging from one to the test length were sampled from a Bernoulli distribution. As each RG responder engaged in RG on a different number of items, we ensured that the sum of RG responses across the sample was equal to the specified percentage of RG responses in the data matrix. Then, to mimic RG, the true probability

of success for any response deemed to be a rapid guess was replaced with the chance probability (.25). The overall percentages of rapid guesses in the data matrix are reflective of empirical rates observed in operational testing contexts and those simulated in prior research studies (e.g., DeMars & Wise, 2010; Rios et al., 2017; Wise & DeMars, 2006); however, the 0% condition was included as a baseline to obtain an estimate of α under the context of full solution behavior by all simulees.

Rapid guessing pattern. Prior research suggests that RG may be associated with item-level content or cognitive fatigue (Hong & Cheng, 2019; Wise, 2017). As such, this simulation study considers two RG patterns previously not investigated in the context of α. The first of which is referred to as difficulty-based RG, which reflects the situation of examinees engaging in RG due to a low perceived probability of success. That is, examinees may believe that they lack the requisite knowledge, skills, or abilities to correctly answer an item, and therefore, engage in RG rather than expend maximal effort (Wise, 2017). To reflect this interaction between RG and item difficulty, across all items, known probabilities of successful responses were rank ordered (ties were randomly ordered), and the items with the lowest probability of success (based on the specified proportion of RG item responses) were replaced with the chance rate.

The second pattern simulated reflects examinees engaging in less effortful responding as the test progresses, due to cognitive fatigue, which has been demonstrated to occur across a number of operational testing contexts (e.g., Pastor et al., 2019; Penk & Richter, 2017; Wise & Kingsbury, 2016). This was accomplished in three steps. First, the test length was split into four equal sized bins (e.g., for the 40-item condition, each bin consisted of 10 items). Second, to reflect cognitive fatigue, the percentage of rapid guesses in each bin was specified to increase from 10% to 40% in 10% increments. Third, for each simulee, items were randomly selected to replace the true probability of success with the chance rate within each item bin.

Analyses

To examine the degree of difference in reliability due to RG, we computed the average difference in coefficient α estimates between clean (i.e., RG % = 0) and contaminated data (i.e., RG % > 0). This was done by first calculating reliability estimates for each dataset (clean and contaminated) separately using the *psych* package in *R* (Revelle, 2020), taking the difference in estimates, and averaging across replications. Furthermore, to provide descriptive results that could elucidate these findings, the mean inter-item correlation, the total score variance, and the ratio of number of inflated to deflated inter-item correlations were calculated. This latter value was included as Wise and DeMars (2009) suggested that potential distortion of coefficient α estimates is related to this ratio, with increased positive ratios representative of greater inflation in reliability.

Results

Across all conditions, the average degree of bias in coefficient α was −.004 (SD = .01). A linear regression model was conducted in which the difference in α values between clean and contaminated data were regressed onto the six independent variables under investigation. As can be seen in Table 3, results from this model showed that the ability of rapid guessers and RG pattern respectively accounted for 62% and 17% of the overall variance explained in the model (R^2 = .47), based on the R^2 contribution averaged over orderings among regressors.[1] Given the relative importance of these two variables, an interaction effect (ability × RG pattern) was added to the regression formula. This interaction accounted for an additional 18% of variance beyond the main effects model (R^2 = .67; Table 3). In this final model, test length, test difficulty, and RG response

Table 3. Model Results of Regressing Coefficient Alpha Distortion on Study Factors.

Variable	Main effects model		Interaction effects model	
	Estimate	R^2 explained[a]	Estimate	R^2 explained[a]
Intercept	−.003*	—	.000	—
Test length[b]	.002*	.04	.002*	.03
Moderately difficult test[c]	−.002*	.01	−.002*	.01
Hard test[c]	−.004*	.05	−.004*	.04
Rapid guesser percent[d]	−.004*	.10	−.004*	.07
Ability[e]	.009*	.62	.002*	.45
RG response percent: 10%[f]	−.000	.00	−.000	.00
RG response percent: 15%[f]	.000	.00	.000	.00
RG pattern[g]	−.005*	.17	−.012*	.12
Ability × RG pattern	—	—	.014*	.28

Note. *p <.001. The adjusted R^2 for the main and interaction effect models was .47 and .65, respectively.
[a]Based on the R^2 contribution averaged over all orderings among regressors.
[b]A test length of 40 items served as the reference group.
[c]Easy test difficulty served as the reference group.
[d]Conditions with 20% RG simulees (reference) were compared to those with 40%.
[e]Average (reference) versus low ability RG simulee conditions were contrasted.
[f]A RG response percent of 5 was the reference group.
[g]Difficulty-based RG (reference) was compared to progressive RG.

percent were shown to provide negligible prediction of the outcome by each explaining less than 5% of variance. Thus, we turn our attention to the interaction between RG pattern and simulee ability.

Table 4 presents the degree of bias (aggregated by test length, test difficulty, and RG response percentage) for this interaction by RG simulee percentage. Results show that the difficulty-based RG pattern possessed differences in α values between clean and contaminated datasets that were approximately equal to zero across conditions. This result is supported by the average ratio of inflated to deflated inter-item correlations, which ranged from .98 to 1.10, indicating that simulated RG had negligible distorting effects on the associations between items.

Turning to the progressive RG pattern, distinctive differences were noted between ability conditions. Specifically, when RG responders were predominately of low ability, differences in α values between clean and contaminated data were approximately zero for all conditions; however, when RG responders possessed average ability, a consistent pattern of deflated α values was observed. That is, as the percentage of RG simulees in the sample increased from 20% to 40%, the average ratio of inflated to deflated inter-item correlations decreased from approximately .90 to .42. This was associated with a respective reduction of α by .01 and .03. Although the conditions noted above showed some differences, the overall extent of distortion in reliability due to RG only ranged from −.04 to .02 across all 144 conditions investigated.

Discussion

The objective of the present study was to investigate the magnitude of difference in coefficient α estimates under circumstances in which simulees engage in full solution and partial RG behavior due to a low perceived probability of success and cognitive fatigue—two patterns previously unexamined in the literature for this context. Results of the simulation study suggest that RG had a practically negligible impact on estimates of coefficient α once controlling for test length and

Table 4. Influence of RG Contamination on Coefficient Alpha Estimates by Test Difficulty, RG pattern, and Ability Characteristics of RG Responders.

		Difficulty-based RG		Progressive RG	
RG simulee percent, %	RG ability	I/D ratio	α diff.	I/D ratio	α diff.
20	Low	1.07	.00	1.52	.00
	Average	.98	.00	.90	−.01
40	Low	1.10	.00	1.70	.00
	Average	.98	.00	.42	−.03

Note. Low = conditions in which the RG responders possessed predominately lower ability than non-RG responders; Representative = conditions in which the RG responders possessed the same mean ability as the non-RG responders; I/D Ratio = ratio of inflated to deflated item correlations; α diff. = difference in α values between clean and contaminated data.

difficulty. For instance, under extreme conditions in which RG was employed on 15% of item responses for a given sample, the average degree of bias was found to only range from −.04 to .02. This negligible effect was also noted for a number of applied studies that compared contaminated and filtered RG data in our review of the literature (Abdelfattah, 2007; Smith et al., 2013; Wise & Kong, 2005). One potential explanation for the muted association is that the ratio between inflated and deflated item covariances was close to one in most contexts investigated. This is likely to occur when examinees engage in RG on one item and employ solution behavior for most of the remaining items (Attali, 2005). Such a pattern of behavior has been observed in operational contexts (DeMars, 2007).

With that said, our review of applied analyses showed large inflations in α for a number of studies when comparing contaminated and filtered data (see Liu et al., 2015; Wise, 2006; Wise & DeMars, 2009; Wise & DeMars, 2010). These contrary findings to the simulation analysis may be associated with the accuracy of identifying and filtering RG responses in applied contexts. Specifically, our review found that when strict examinee-level filtering criteria were employed, the difference between unfiltered and filtered datasets tended to be larger when compared to more liberal criteria. As noted, removing examinee data based on these strict criteria may have led to eliminating heterogeneity in the item response matrix, which likely produced the large observed differences in alpha estimates. This is supported by Rios et al.'s (2017) work, which concluded that listwise deletion of examinees that engaged in small percentages of RG can bias score inferences to a greater extent than the inclusion of RG responses. Thus, it is recommended that practitioners avoid the use of examinee-level filtering with strict RG criteria when possible.

In addition, the variability observed between some of the applied analyses and our simulation study may be associated with differences in the mechanisms underlying RG. For instance, in the simulated context, simulees employed RG on items that were inconsistent across the sample, which is likely the reason for the minimal distortion detected. However, it is possible that the inflation of coefficient α estimates seen in a number of the applied analyses was due to a sizable percentage of examinees engaging in RG on the same items (assuming that the probability of success for RG is equivalent across all items in which RG is employed; Crocker & Algina, 1986, p. 145). This may have been the case for assessments that included a large number of items with characteristics suggested to be associated with RG, such as those that (a) require extensive reading (i.e., possess long character/word lengths; e.g., Wise et al., 2009); (b) are difficult (assuming that most RG responders perceive the same items to be difficult; e.g., Pintrich & Schunk, 2002; Rios & Guo, 2020); (c) reflect certain content areas (e.g., Liu et al., 2015); and (d) fail to include multimedia content (e.g., Wise, 2006; Wise et al., 2009). Therefore, future simulation research is needed that investigates the influence of RG on coefficient alpha when RG behavior is driven by the same item-level characteristics across examinees.

Implications

The findings from our study coupled with those from Hong and Cheng's (2019) work suggest that estimates of coefficient α are largely robust to the contamination of construct-irrelevant variance introduced by RG responses due to test speediness, cognitive fatigue, and low perceptions of probability of success. However, these findings reflect situations in which RG is idiosyncratic to the individual. Given that coefficient α can be inflated when a large percentage of examinees engage in RG on the same items, it is recommended that practitioners examine the item characteristic correlates of RG for their assessment context. This should be performed during the piloting phase, as it would allow for the revision of items that receive high rates of RG. Additionally, test developers could attempt to mitigate RG prior to its occurrence by reducing the cognitive demands required for a given item or set of items (assuming no alteration to the underlying construct). This could be accomplished by limiting the number of open-ended response items, tailoring the difficulty of an item to an examinee's ability level, so that the item is neither too easy or difficult, and writing item content that is likely to be of interest to examinees (Wise & DeMars, 2005). Such approaches may be helpful in limiting the deleterious influence that RG has on other measurement properties as well, such as tests of measurement invariance and equating error (Mittelhaëuser et al., 2015; Rios, 2021b). By engaging in further research that investigates examinee- and item-level correlates of RG, we may continue to build our knowledge of how RG can undermine measurement quality and valid score-based inferences.

Author Contributions

The first author conceived of the presented idea, conducted the simulation study, and wrote the manuscript. The second author conducted the literature search, extracted and coded variable information, and interpreted findings from the meta-analysis.
All authors conducted critical revisions of the article throughout the review process and approved of the final version to be published.

Declaration of Conflicting Interests

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

Funding

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

ORCID iD

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

Supplemental Material

Supplemental material for this article is available online.

Notes

1. The R^2 contribution was calculated using the *relaimpo* package in *R* (see Grömping, 2006).

References

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.). American Educational Research Association.

Attali, Y. (2005). Reliability of speeded number-right multiple-choice tests. *Applied Psychological Measurement*, *29*(5), 357–368.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Holt, Rinehart and Winston.

Cronbach, L. J. (1951). Coefficient α and the internal structure of tests. *Psychometrika*, *16*(3), 297–334.

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

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

Grömping, U. (2006). Relative importance for linear regression in R: The package relaimpo. *Journal of Statistical Software*, *17*(1), 1–27.

Hong, M. R., & Cheng, Y. (2019). Clarifying the effect of test speededness. *Applied Psychological Measurement*, *43*(8), 611–623.

Hubley, A. M., & Zumbo, B. D. (2017). Response processes in the context of validity: Setting the stage. In B. D. Zumbo, & A. M. Hubley (Eds.), *Understanding and investigating response processes in validation research* (pp. 1–12). Springer.

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

McNeish, D. (2018). Thanks coefficient α, we'll take it from here. *Psychological Methods*, *23*(3), 412–433.

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.

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

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

Pintrich, P. R., & Schunk, D. H. (2002). *Motivation in education: Theory, research, and applications*. Prentice Hall.

Revelle, W. (2020). *Package "psych"* (package version 2.0.9) Retrieved from https://cran.r-project.org/web/packages/psych/psych.pdf

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

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

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.

Rios, J. A., Guo, H., Mao, L., & Liu, O. L. (2017). Evaluating the impact of noneffortful responses on aggregated scores: To filter unmotivated examinees or not? *International Journal of Testing*, *17*(1), 74–104.

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.

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

Steedle, J. T., Hong, M., & Cheng, Y. (2019). The effects of inattentive responding on construct validity evidence when measuring social–emotional learning competencies. *Educational Measurement: Issues and Practice*, *38*(2), 101–111.

Sundre, D. L., & Wise, S. L. (2003). *"Motivation filtering": An exploration of the impact of low examinee motivation on the psychometric quality of tests*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

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.

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

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

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

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

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

Wise, S. L., & DeMars, C. E. (2010). Examinee noneffort and the validity of program assessment results. *Educational Assessment*, *15*(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., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, *18*(2), 163–183.

Wise, S. L., & Kuhfeld, M. R. (2020). 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., 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.

Wise, V. L., Wise, S. L., & Bhola, D. S. (2006). The generalizability of motivation filtering in improving test score validity. *Educational Assessment*, *11*(1), 65–83.