Soland *et al. Large-scale Assess Educ* (2021) 9:8
https://doi.org/10.1186/s40536-021-00100-w

Large-scale Assessments in Education

RESEARCH Open Access

Comparing different response time threshold setting methods to detect low effort on a large-scale assessment

James Soland[1,2*], Megan Kuhfeld[2] and Joseph Rios[3]

*Correspondence:
jgs8e@virginia.edu
[1] School of Education and Human Development, University of Virginia, 405 Emmet Street, Charlottesville, VA 22904, USA
Full list of author information is available at the end of the article

Abstract

Low examinee effort is a major threat to valid uses of many test scores. Fortunately, several methods have been developed to detect noneffortful item responses, most of which use response times. To accurately identify noneffortful responses, one must set response time thresholds separating those responses from effortful ones. While other studies have compared the efficacy of different threshold-setting methods, they typically do so using simulated or small-scale data. When large-scale data are used in such studies, they often are not from a computer-adaptive test (CAT), use only a handful of items, or do not comprehensively examine different threshold-setting methods. In this study, we use reading test scores from over 728,923 3rd–8th-grade students in 2056 schools across the United States taking a CAT consisting of nearly 12,000 items to compare threshold-setting methods. In so doing, we help provide guidance to developers and administrators of large-scale assessments on the tradeoffs involved in using a given method to identify noneffortful responses.

Keywords: Test effort, Rapid guessing, Large-scale assessments, Validity, Computer adaptive testing

Introduction

An assumption fundamental to the validity of most intended uses of achievement tests is that examinees are providing maximal effort on the test (AERA et al., 2015). Unfortunately, this assumption is often violated, especially on tests with few or minimal stakes for students (Jensen et al., 2018; Rios et al. 2016; Wise & Kuhfeld, 2020; Wise & Kong, 2005). In some cases, more than 15% of examinees in middle school grades have demonstrated low effort on items sufficient to potentially undermine the validity of their observed score (Soland, 2018b). Low effort of that magnitude is not uncommon and can downwardly bias observed test scores, oftentimes by as much as 0.2 standard deviations (Rios et al., 2016). Further, low effort occurs differentially by subgroup, which can bias achievement gap estimates (Soland, 2018a), and oftentimes affects students who are disengaging from school (Soland & Kuhfeld, 2019; Soland, Jensen, et al., 2019)—in short, measurement is often most impacted among the students for whom it is arguably most important. Given the prevalence of low examinee effort and its effect on estimated test

scores, it is now considered a fundamental threat to uses of test scores, including from international tests like the Program for International Student Assessment (PISA; Debeer et al., 2014; Goldhammer et al., 2014; Rios & Guo, 2020; Wise et al., 2019).

Fortunately, a number of approaches have been developed to identify item responses that were provided without full effort (e.g., Wise & Kong, 2005). Many of these approaches, especially those used in operational testing, rely on item response times. Item response times are the seconds that elapse between when a question is presented and answered (Schnipke & Scrams, 1997). These response times can be used to identify item responses provided in a normative amount of time versus provided much quicker (Schnipke & Scrams, 1997; Wise & Kong, 2005). For example, if a student responds to an item with a lengthy reading passage in, say, five seconds, one can be fairly sure full effort was not given. Response times have been used to detect noneffortful responses in research using a bevy of large-scale domestic (Demars, 2007; Rios et al., 2014; Soland & Kuhfeld, 2019; Wise, 2015) and international tests including the PISA (Debeer et al., 2014; Goldhammer et al., 2014; Soland et al., 2018; Zamarro et al., 2016). One reason for the popularity of using response times is that some related approaches are supported by decades of validity evidence chronicled by Wise (2015) and approaches to re-scoring achievement tests that account for low effort are now used in practice with large-scale assessments like the ones used in our own study (Wise & DeMars, 2006; Wise & Kuhfeld, 2020).

Given the importance of accurately detecting low examinee effort, a number of methods have been proposed to set item response time thresholds. Each method separates effortful and noneffortful responses by identifying a response time threshold below which responses are deemed to be provided with less than maximal effort (Kong et al., 2007). These methods differ in the ways they operationalize low effort, and their usefulness may differ dependent on the intended use of the test score (an issue we discuss more fully later). While a handful of studies have presented threshold-setting methods and, in some cases, compared them (e.g., Guo et al., 2016; Kroehne et al., 2020; Rios & Guo, 2020; Sahin & Colvin, 2020; Wise, Kuhfeld, & Soland, 2019), those studies have a few limitations. First, they often involve modest samples from relatively small-scale assessments. Second, even when large-scale assessments are used, most related studies use only a few items (<30), such as when examining released items from PISA. Using such a small set of items raises questions about whether results are generalizable to a larger item pool and/or to other tests.

Third, they often do not use computer-adaptive tests (CATs), which are employed increasingly in operational settings and pose particular challenges for threshold-setting. For example, CATs typically produce sparse item response matrices, which can make it hard to use normative threshold-setting approaches (Wise & Ma, 2012). Further, CATs target items to examinee ability, which likely changes how frequently students get items correct at or below the chance rate, a criterion used to validate many thresholds. However, a CAT probably would not influence the proportion of correct item responses when a student is not providing full effort. Regardless of whether the items were administered with CAT, rapid guesses should have proportions correct around the chance level. Thus, one might expect that item responses correctly classified as rapid guesses should have proportions correct around the chance level and this criterion for validating thresholds

should be applicable for CAT. All told, the relative performance of threshold-setting methods in large-scale operational CAT settings—including how often they actually produce viable thresholds and how much their results differ—has not been compared. As discussed at more length in the methods section, we define viability as producing a threshold that is not indeterminate (i.e., actually produces a threshold) and that suggests a two-group solution (effortful and non-effortful) fits better than a one-group solution.

In this study, we use item-level data from almost 3 million test records to set response-time thresholds using different approaches and compare results. While a number of complex threshold-setting approaches have been developed (e.g., Lu et al., 2020; Ulitzsch et al., 2020), including ones that employ mixture models to jointly estimate ability and effort, we focus on those that either have been used in operational testing contexts, or that were designed for that purpose. Beyond tailoring our study to large-scale operational testing contexts, the decision was also made because any challenges associated with implementing more complex models are likely to be compounded by the sparseness and local independence violations observed in CAT item response data. The test we use, MAP Growth reading, is a CAT administered in approximately one in four US public schools, as well as widely internationally. We apply commonly used but conceptually distinct threshold setting methods to our data in order to investigate three research questions:

1. How often do various methods for estimating thresholds produce viable estimates with a large-scale operational assessment?
2. Which method most consistently identifies responses that are correct at rates no better than chance?
3. How often is there consistency in the item responses identified as noneffortful across the different threshold-setting methods?

We conclude with a discussion of the results, including providing guidance for large-scale operational test developers and administrators on the tradeoffs involved in selecting a given threshold-setting method. As we detail in the background section and expand on in the discussion section, the tradeoffs associated with a particular method often cannot be disentangled from the intended use of the measure.

Background

Schnipke and Scrams (2002) divide test examinees into two categories: those exhibiting "solution behavior" and those exhibiting "rapid-guessing behavior." Students in the latter category, who respond to a test item without sufficient time to have understood the question, are not engaged with the test during that item (Schnipke & Scrams, 2002; Wise & Kong, 2005). Assignment of examinees to these categories is based on response time. Response time has several advantages as a measure of examinee engagement. Importantly, because the examinee is unaware the data are being collected, response time does not suffer from self-report biases like on surveys of student engagement (Kyllonen, 2012; Wise & Kong, 2005). Response times also allow one to identify low effort at the item-level, unlike other approaches such as person fit statistics (Wise, 2015).

The main difficulty inherent in using response time lies in determining what item responses constitute effortful verses noneffortful behavior. Originally, Wise and Kong (2005) used an empirical visualization approach to identify rapid-guessing behavior. Specifically, solution behavior thresholds were identified using histograms of response times for items, virtually all of which were bimodal with one local maximum somewhere below 10 s. Any items responded to in an amount of time that corresponded with the local maximum below 10 s was deemed rapid.

This approach (and its eventual descendants) is supported by several forms of validity evidence chronicled by Wise (2015). First, measures of rapid guessing that aggregate individual responses to the examinee level for a given test[1] are reliable with coefficient alpha estimates around 0.97. Second, rapid guessing is likely to be associated with other measures of test-taking effort, including self-reports of effort and person fit statistics. Third, rapid guessing is not highly correlated with academic ability, which is to say effort measures are not simply proxying academic proficiency (though there is debate on this issue). Fourth, rapid guesses are associated with item scores that are correct with a consistency not much better than chance. That is, the students, on average, appear to have guessed on the item. Finally, test statistics like coefficient alpha increase when students with frequent rapid guesses are dropped from the analysis.

Since the initial visual inspection approach to threshold setting, a number of more sophisticated methods have been developed. In the remainder of this background section, we describe the three threshold-setting methods we use in this study: normative threshold, cumulative proportion correct, and mixture log normal approaches. We selected these three because they were specifically designed to be simple enough to use in conjunction with large-scale operational tests (Guo et al., 2016; Rios & Guo, 2020; Wise & Ma, 2012).

Normative threshold method

The normative threshold (NT) method was developed by Wise and Ma (2012), based on the concern that the visual inspection approach was too time-consuming and not standardized enough to use on a large-scale assessment, especially a CAT consisting of thousands of items. Thus, they developed an approach that bases the item duration cutoff separating solution behavior from rapid-guessing behavior on the mean duration of responses to individual items. Specifically, their findings supported the use of item thresholds set at 10% of the mean item duration with a maximum threshold of ten seconds (dubbed the NT10 for normative threshold with a maximum of 10 s). They selected these thresholds by showing that items flagged as rapid using the NT10 approach tended to have responses that were correct at rates comparable to chance (e.g., 25% of responses were correct on multiple choice items with four response categories). Wise and Ma (2012) also showed that the NT10 approach tended to identify item responses as rapid that highly overlapped with those identified using the visual inspection approach. Given the validity evidence provided by Wise and Ma (2012) for this threshold setting method,

[1] Wise and Kong (2005) use an empirical approach to identify rapid-guessing behavior and generate an overall measure of a student's test-taking engagement, which they term response-time effort or RTE. RTE scores range from zero to one and represent the proportion of test items on which the student exhibited solution behavior.

it is now used in operational settings to identify rapid guesses, including in operational administrations of MAP Growth.

However, more recent research indicates that NT10 may be too conservative, i.e. it tends to under-identify noneffortful responses. Specifically, Wise and Kuhfeld (2020) used achievement test data composed of test takers who were quickly retested after an initial test on which they showed low effort. Two primary findings emerged. First, rescoring tests accounting for low effort identified using NT10 thresholds accounted for roughly one-third of the score distortion due to differential effort. Second, a modified scoring method that accounted for low effort using more liberal time thresholds (such as using 20% or 30% of the mean response time, deemed NT20 and NT30, respectively) performed better, accounting for upwards of two-thirds of the distortion. Thus, their results simultaneously provided validity evidence in support of the general NT approach while suggesting that both the 10% of the response time distribution cutoff and the 10-s maximum might be too restrictive under certain circumstances.

Beyond the study's main finding, Wise and Kuhfeld (2020) raised two important considerations when evaluating other threshold setting methods. First, they acknowledge that using more liberal thresholds (e.g., NT30 using 30% of the response time distribution) may do a better job of correcting for low effort in a test/re-test context, but that these longer response times might not be capturing rapid guessing, per se. That is, low effort might not be synonymous with only partially understanding the content, but with not spending as much time on the item as is needed (a subtle difference). Second, they acknowledge that the appropriate approach to setting thresholds may depend on the particular use of the score. For example, in the context of proctor notification or invalidating tests for low effort for operational purposes, one likely wants to avoid misclassifying engaged and highly efficient students as noneffortful. In such a case, one may wish to use a more conservative approach like NT10. However, if one is using the scores for research purposes, or the primary interest involves using scores in the aggregate (e.g. PISA), then more liberal methods like NT30 may be preferred. The final two threshold-setting methods we consider both use a broader definition of noneffort than just classifying responses as rapid guesses.

Cumulative proportion method

Another recent method was proposed by Guo et al. (2016) that used validation criteria similar to those utilized by Wise and Ma (2012), but that directly incorporated response accuracy into the threshold-setting procedure. Their method expanded work by Lee and Jia (2014), making the method better able to handle small sample sizes or sparse response-time frequencies. Specifically, they defined noneffortful behavior as providing responses that are correct at or below the chance rate (which might or might not be synonymous with rapid guessing). Thus, a threshold could be set at the response time where response accuracy rose above the chance rate.

Their study defined the cumulative proportion, denoted as CUMP(t), for an item at time t as the proportion correct of all those students who spend t seconds or fewer on the item. For example, if students on a given item who spent 12 s or under on the item tended to get that item correct 40% of the time, then CUMP($t = 12$) would equal 0.40. Guo et al. (2016) further showed that, as the response time increased, it eventually

converged on the mean proportion correct. They then used this strategy to set thresholds by identifying a response time threshold at which CUMP(t) equaled the chance rate. For instance, if an item had five response categories, then the threshold would be set at the time where the cumulative proportion correct reached 0.20. Thus, the response time threshold was set where the probability of a correct response exceeded chance.

Guo et al. (2016) concluded that the CUMP method has several advantages, especially for large-scale assessments. According to their findings, the biggest advantage of the CUMP method is that thresholds can be computed easily from data using a mathematical definition (unlike visual inspection). Further, like the NT approach, a maximum response time can be set. Their results also showed that the CUMP method appears to recover rapid-guessing behavior better than the other methods they studied, including a variation on NT10. They hypothesized that the result occurred because the CUMP method assumes that low effort will produce item responses that are correct at the chance rate, comparable to random guessing. In short, the primary validity criterion used to support other methods—proportion of correct item responses among those deemed to be noneffortful—is built right into the threshold-setting method. However, the performance of CUMP has never been studied when using a CAT.

Mixture log normal

Another recent method based almost wholesale on the visual inspection strategy is called the Mixture Log Normal (MLN) method (Rios & Guo, 2020). Specifically, MLN assumes that, in the presence of low effort, a bimodal response time distribution should be observed in which the lower mode represents noneffortful responding and the upper mode indicates effortful responding. While MLN is a parameterized version of the visual inspection method, it does not limit noneffort to what one would traditionally define as rapid guessing. In fact, Rios and Guo (2020) expressly state that low effort could take many forms other than rapid guessing. For example, one could imagine slow responding due to a host of reasons (e.g., an examinee initially engaging with an item, realizing it is too difficult, and then providing a haphazard answer). Conservative definitions of effort like NT10 would not define such a behavior as noneffortful. In their own empirical analyses, Rios and Guo (2020) found that many of the thresholds identified far exceeded the 10 s maximum imposed by NT10.

The MLN method employs an automated process that utilizes an empirical response time distribution, fits a mixed log normal distribution, and then locates the lowest point between the two modes of the distribution, which is set as the threshold. In mathematical terms, let $y = \log(x)$, where x is the response time on an item. The log is taken of this response time to make the distribution better approximate a normal distribution (van der Linden, 2007). Further, assume $y_1 \sim N(\mu_1, \sigma_1) = f_1(y)$ and $y_2 \sim N(\mu_2, \sigma_2) = f_2(y)$ represent two normal distributions within the overall response time distribution where $\mu_1 < \mu_2$. Thus, the distribution of y is equal to

$$f(y) = \pi_1 f_1(y) + \pi_2 f_2(y).$$

Here, π_1 and π_2 are proportions of the two normal density functions. The threshold is defined as the time point $x \epsilon [\mu_1, \mu_2]$ where $g(x)$ reaches the minimum value (which is also the intercept of f_1 and f_2).

The validity of this threshold-setting method was supported by two primary pieces of evidence. First, it performed well relative to other methods in setting the response time threshold such that examinees got the item correct at a rate no better than chance. The second piece of validity evidence involved descriptive statistics showing that, as effortful responding increased, so too did total testing time and test performance when using MLN.

MLN is just one example of a preponderance of recent research on test effort that employs mixture models of various kinds (e.g., Meyer, 2010; Molenaar & de Boeck, 2018; Schnipke & Scrams, 1997; Wang & Xu, 2015). However, unlike many other mixture modeling approaches, MLN is not especially complicated or computationally intensive because it is not directly incorporated into the scoring process (thresholds can be set a priori; Wang & Xu, 2015). By contrast, other methods that jointly estimate response time and ability (e.g., Ulitzsch et al., 2020), and in particular mixture-based methods, can pose challenges for operational use. For example, such methods create estimation complexities (many of the models call for estimation methods that are not readily available in operational software) and need to estimate additional parameters that require additional assumptions to be made (among other issues; Wang & Xu, 2015). Given these challenges, MLN is much better suited to actual use in an operational large-scale tests than many other mixture methods (e.g., Goldhammer et al., 2014).

Relative advantages and disadvantages of the methods

Beyond the advantages and disadvantages of the methods already discussed, such methods have strengths and weaknesses relative to each other that are partially the object of our own study. For example, when applying the MLN method, there could be no mixture, with the data adequately represented by a single normal distribution. Similarly, there could be cases where a mixture of more than two normal distributions fits optimally, in which case identifying and interpreting thresholds becomes complicated. One could also imagine scenarios in which thresholds are identified after considerable time has passed. Even if one abandons the definition of low effort as rapid guessing, item responses in the short response time group might still have lengthy response times. While such an issue could be addressed by setting a cap on the maximum response time (as with NT methods), the question of what that cap should be and how criteria can be developed for setting it has not been discussed in any great detail.

Meanwhile, CUMP's primary limitation is that thresholds cannot be defined on difficult items for which the total proportion correct is above or below the chance rate. That is, if the proportion correct conditional on response time never crosses the chance rate, then CUMP would fail to define a threshold. Partially to address such issues, Rios and Guo (2020) also employed a "hybrid" approach that essentially combines CUMP and MLN. When CUMP failed to generate a viable threshold, Rios and Guo (2020) used MLN to set the threshold.

By way of contrast with the CUMP and MLN methods, the NT method will virtually always provide a solution given it requires only the first two moments of the response time distribution to calculate a threshold. However, one could argue the NT method overly standardizes the process, applying criteria (percent of the response time

Table 1 Details on studies focused on comparing threshold-setting methods

Study	Thresholds compared	Examinees	N. items	CAT?
Guo et al. (2016)	CUMP; visual inspection	1422	27	No
Kroehne et al. (2020)	CUMP; change in information; NT10; NT20; visual inspection	8612	59	No
Rios and Guo (2020)	CUMP; MLN; NT10; uniform 3s cutoff	19,879	26	No
Sahin and Colvin (2020)	NT10; uniform cutoffs at 5, 10, 20, 30, 40, 50, and 60 s	1951	7	No
Wise et al. (2019)	Change in information; CUMP; NT10; visual inspection	23,000	30	Yes
Current study	*CUMP; MLN; NT10; NT30*	*728,923*	*11,968*	*Yes*

distribution and maximum time) based on a sample from a single study (albeit a large one). By comparison, MLN and CUMP set thresholds in much less constrained ways.

Setting thresholds in a CAT context

Many of these advantages and disadvantages can have different implications in a CAT context (Wise & Ma, 2012). For example, a CAT matches item difficulty to examinees' estimated ability. As a consequence, examinees are less likely to disengage due to a mismatch between item difficulty and ability. This aspect of a CAT may have some consequences for setting thresholds. Specifically, students may be less likely to show low effort because an item is too difficult given most items should be of an appropriate difficulty. Therefore, there is less chance that true ability and low effort are correlated, which has large implications for the bias introduced into person and item parameters (Rios & Soland, 2021). Further, this aspect of CATs may reduce the likelihood that methods like CUMP identify a threshold because there are fewer items on which the proportion of correct responses dips below the chance rate. While threshold methods like NT10 were specifically developed for use with CAT item pools (Wise & Ma, 2012), the relative performance of such methods has not been examined in a large-scale CAT setting.

Prior comparisons of threshold-setting methods

Table 1 presents details on studies that have explicitly compared thresholds by threshold-setting method. While the list in Table 1 is not exhaustive, it is intended to at least highlight the most recent and comprehensive studies on this topic. On the last line, details from our own study are included as a point of comparison. As the table makes clear, while a handful of studies have compared thresholds by method as we do, most of them use fairly small numbers of examinees and very few items, ranging from 7 to 59 items (by comparison, our sample includes nearly 12,000 items). Further, only one other study used a CAT as its mode of testing. Perhaps the study most relevant to ours was conducted by Kroehne et al. (2020), who used a large-scale assessment and showed that rapid guessing rates can differ across methods. However, comparing thresholds was not their primary purpose, nor did they use a CAT.

Table 2 Descriptive characteristics of the sample

Grade	N	Male (%)	White (%)	Black (%)	Asian (%)	Native American (%)	Hispanic (%)	Other Race/ ethnicity (%)
3	192,197	50.7	35.9	18.7	4.9	0.8	26.9	12.7
4	195,806	50.6	36.7	18.0	5.0	0.9	27.0	12.5
5	196,992	50.5	37.4	17.5	4.9	0.8	27.1	12.2
6	194,461	50.8	37.8	17.0	4.9	0.8	27.5	12.0
7	195,512	51.0	38.7	16.6	5.1	0.9	27.1	11.7
8	197,749	51.0	39.4	16.9	5.1	0.9	26.3	11.4
All grades	728,923	50.8	37.7	17.5	5.0	0.8	27.0	12.1

Methods

Analytic sample

The data for this study are from one U.S. state within the Growth Research Database (GRD) at NWEA. School districts partner with NWEA to monitor elementary and secondary students' reading and math growth throughout the school year, with assessments typically administered in the fall, winter, and spring. We use the reading test scores of over 728,923 3rd–8th grade students in 2056 schools from the 2016–17 through the 2017–18 school year. The GRD also includes demographic information, including student race/ethnicity, gender, and age at assessment, though student-level socioeconomic status is not available. Table 2 provides descriptive statistics for the sample by grade. In each grade, we observe between 192,000 and 198,000 unique students. Overall, the sample is 51% male, 38% White, 18% Black, 5% Asian, and 27% Hispanic.

Measures of achievement

Student test scores from NWEA's MAP Growth Common Core-aligned reading assessments are used in this study. MAP Growth is a CAT—which means measurement is precise even for students above or below grade level—and is vertically scaled to allow for the estimation of gains across time. Each test begins with a question appropriate for the student's achievement level (either based on a student's past performance or grade-level expectations), and then adapts throughout the test in response to student performance. The MAP Growth assessments are typically administered three times a year (fall, winter, and spring) and are aligned to state content standards. Test scores are reported on the RIT (Rasch unIT) scale, which is a linear transformation of the logit scale units from the Rasch item response theory model.

The full dataset used in this study contained over 115 million item responses within 2.9 million unique test events. To ensure sufficient sample size to conduct each threshold-setting method, we limited analyses to items with more than 100 responses (this response rate was not guaranteed even with a very large dataset due to the combination of an extremely large item bank and the computer-adaptive nature of the test). While having only 100 item responses could be problematic for producing distributions of response time and accuracy that are consistent across samples (Wise &

Ma, 2012), we nonetheless made this choice because being more limiting would have considerably reduced our sample size (and therefore represented a problem in the context of our operational CAT). Of the 11,968 items in our sample, 7,826 items had 100 or more responses between the 2016–17 and 2017–18 school years. These are the items we focus on in this study. Ninety percent of the items included in our study had over 1,000 observed responses (mean = 14,690, SD = 16,116).

Analytic strategy

Using our achievement measure and sample, we identified response time thresholds using three approaches: NT, CUMP, and MLN. Given the findings of Wise and Kuhfeld (2020), we used two different criteria for the NT approach: NT10 (10% of the mean of the response time distribution) and NT30 (30% of the mean of the response time distribution). Unlike what is often done in practice, for our NT thresholds, we did not set a maximum of 10 s for NT10. Instead, we used a uniform maximum of 100 s across all methods, even for the NT methods. While such a maximum is longer than those used in prior research (e.g., Rios & Guo, 2020), we made this decision to better compare thresholds across methods when those thresholds are not being driven primarily by very short maxima. However, most of our analyses allow for the reader to evaluate what would have happened if the 10 s maximum (or some other limit) had been used for the NT methods.

Question 1. How often do various methods for estimating thresholds produce viable estimates with a large-scale operational assessment?

To answer Question 1, we tried to identify cases in which various methods produced viable thresholds. We examined (and defined) viability in several ways. First, we tabulated how often a given method actually produced a threshold. For the NT approaches, we were always able to set thresholds, which is one reason they are already used operationally with MAP Growth (Wise, Kuhfeld, & Soland, 2019; Wise & Kuhfeld, 2020). As long as the mean of a response time distribution can be estimated, the NT method will work (leaving aside whether it actually identifies random guesses effectively). Thus, the NT method always produces a threshold.

By contrast, as previously noted, CUMP will not produce a threshold when the cumulative proportion correct never strays above or below chance. Thus, we tabulate how often that phenomenon occurs using our items responses. For MLN, thresholds cannot be estimated if: (a) the two-group mixture model does not converge or (b) the two-group solution does not show adequate fit to the data relative to the one-group solution, indicating that the response time distribution is not bimodal. Thus, we examined model fit for one- and two-group MLN solutions, and deemed a threshold viable when a likelihood ratio test indicated that the two-group model showed improved fit to the response time data relative to the one-group model. As a point of comparison, we also fit a three-group model to see how often that solution fit the data best. As we discuss later, identification of a three-group solution would not rule out identifying the fastest response time distribution as containing low effort item responses, but would complicate the traditional bifurcation of item responses into effortful versus noneffortful.

Finally, we examined how often a threshold-setting method produced a threshold that exceeded 100 s. While 100 s is somewhat arbitrary, it has been used in prior studies as

a maximum for setting response time cutoffs (e.g., Guo et al., 2016). Such results are meant to show how often viable thresholds are produced, but involve such long response times they may lack face validity if the argument is that examinees are not spending enough time on the item to have given full effort.

One should note that, in general, a potential explanation for lacking a viable threshold is that no students showed low effort. However, such a result does not mean that no threshold is required. Even if no students in our sample rapidly guessed, that does not mean none will going forward. This point is especially important given the intent when setting thresholds is often to use them in subsequent administrations. Further, the issue is particularly germane in CAT assessments, which have sparse item response matrices. In the context of CATs, threshold-setting approaches may be best seen as an iterative process where items without viable thresholds are revisited again after additional item responses are collected.

Question 2. Which method most consistently identifies responses that are correct at rates no better than chance?

In this question, we examined the proportion of correct responses for items deemed rapid by threshold method. Ideally, the proportion would equal the chance rate. Since MAP Growth items typically have four response categories, 0.25 was used for most items as the chance rate. To answer this question, we produced the proportion of item responses that were correct among those flagged as noneffortful. We examined these proportions for each item separately, then summarized how often they exceeded the chance rate.

Given the CUMP method explicitly uses the proportion of correct responses in setting thresholds, one would expect this method to perform best. The main value in this question lies in comparing CUMP to MLN and NT. Examining how well NT performs is especially important given it uses such generalized rules for identifying rapid guesses. In short, this question helps us consider method tradeoffs discussed in the background section. Whereas NT methods will always produce viable thresholds, they may over- or under-identify noneffortful responses. By contrast, whereas the CUMP approach may fail to produce viable thresholds, it is built to ensure the method identifies individuals getting items correct at the chance rate. In tandem, Question 1 and Question 2 mean we can compare these tradeoffs directly.

Question 3. How often is there consistency in the item responses identified as noneffortful across the different threshold setting methods?

To answer this question, we took two approaches. First, we compared response time thresholds across the four approaches by item. Specifically, we plotted the thresholds for each item with a different method on each axis. Second, we calculated the percent of responses by item classified as noneffortful for each threshold method and compared. Thus, we could see whether rates of noneffortful responding differed considerably by method, and rank the methods from most to least conservative (with the most conservative being the method least likely to call a response noneffortful).

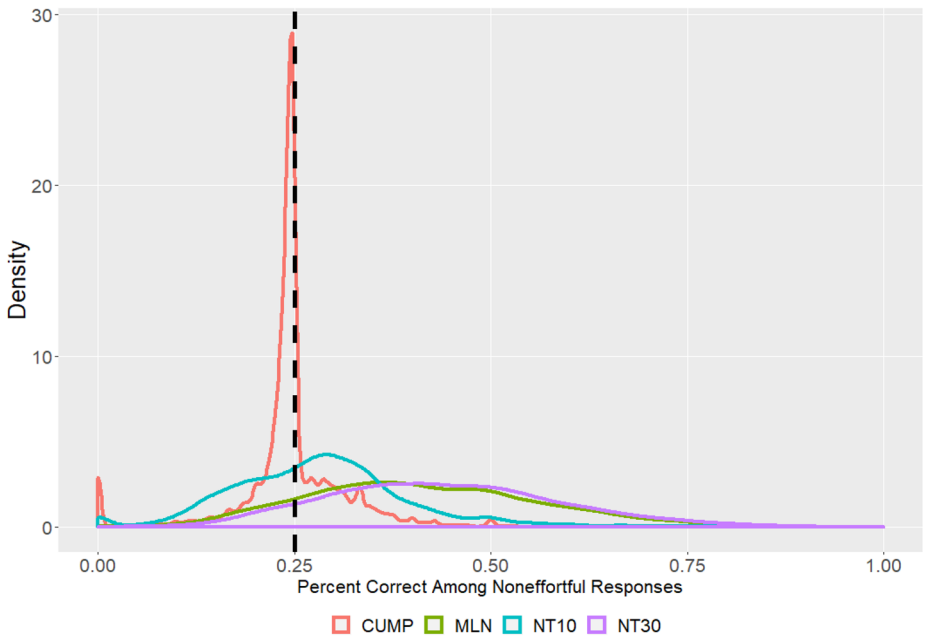

Fig. 1 Percent of correct responses among by item responses classified as noneffortful by threshold approach

Results

Question 1. How often do various methods for estimating thresholds produce viable estimates?

In our sample, the number of viable thresholds differed meaningfully by threshold-setting method. As expected, both NT methods produced thresholds for all 7826 items. Meanwhile, the CUMP approach produced thresholds for only 5987 items because nearly 2000 items had response accuracies that never dipped below the chance rate conditional on response time. This phenomenon likely occurred because the test we used is a CAT, with items targeted at an examinee's estimated ability level. The MLN method produced thresholds for nearly all items (7597). Thus, for 227 items, the one-group mixture model solution fit the data best.

While the MLN method technically produced thresholds for virtually all items, there was a wrinkle worthy of mention. In our analyses, the one-group solution was often not the best fit. Rather, a three-group solution frequently fit the data better than a two-group solution. Among the 7,597 items for which the one-group solution did not fit best, only 807 items demonstrated the best fit with two-group solution.

Further complicating the use of MLN, 13.3% of the identified thresholds were above 100 s. Therefore, even students spending 1.5 min on an item could often have responses deemed noneffortful. By contrast, the numbers for the other methods were CUMP at 3%, NT30 at 0.3%, and NT10 at 0%.

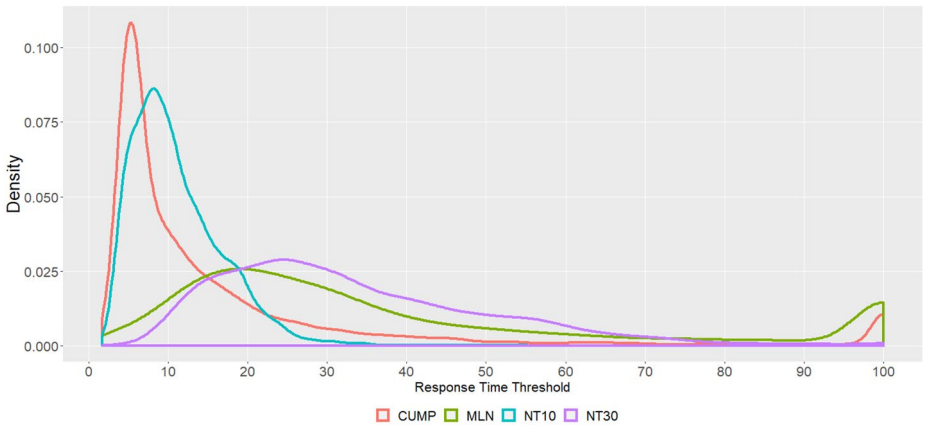

Fig. 2 Comparison of response time thresholds across four approaches

Question 2. Which method most consistently identifies responses that are correct at rates no better than chance?

Figure 1 shows density plots of the proportion of correct responses across items by method for item responses identified as noneffortful. For example, if an item had a value of, say, 0.40, that means item responses flagged as noneffortful were correct about 40% of the time. Recall, MAP Growth items almost always have four response categories, which means students who get an item correct one quarter of the time are essentially doing no better than guessing. As expected given its incorporation of response accuracy rates into the threshold-setting algorithm, CUMP identifies item responses as noneffortful such that virtually all of those responses are correct around the chance rate. By comparison, item responses flagged as noneffortful by both NT30 and MLN tended to be correct at rates well above chance (responses were correct ~ 45% of the time), though there was considerable variability across items. Interestingly, NT10 frequently flagged items that were correct near the chance rate, though with substantially more variability than for CUMP.

Question 3. How often is there consistency in the item responses identified as noneffortful across the different threshold setting methods?

Figure 2 shows density plots for response time thresholds across items by method. As the figure makes clear, thresholds differ considerably by method. For example, MLN and NT30 have mean thresholds somewhere between 20 and 30 s and the thresholds are highly variable, with values above one minute occurring. By contrast, NT10 and CUMP tend to have thresholds below 10 s (for CUMP, well below 10 s), and variability was much lower.

Figure 3 shows the proportion of item responses classified as noneffortful by threshold method. Each point on the plot represents an item with proportions on the axes. A contour plot is also overlaid in each panel to identify areas of high and low density. The figure includes an identity line: a point falling directly on the line would indicate an item for which the percent of items deemed noneffortful matched exactly between the two methods. As the figure indicates, with the exception of comparing MLN and NT30,

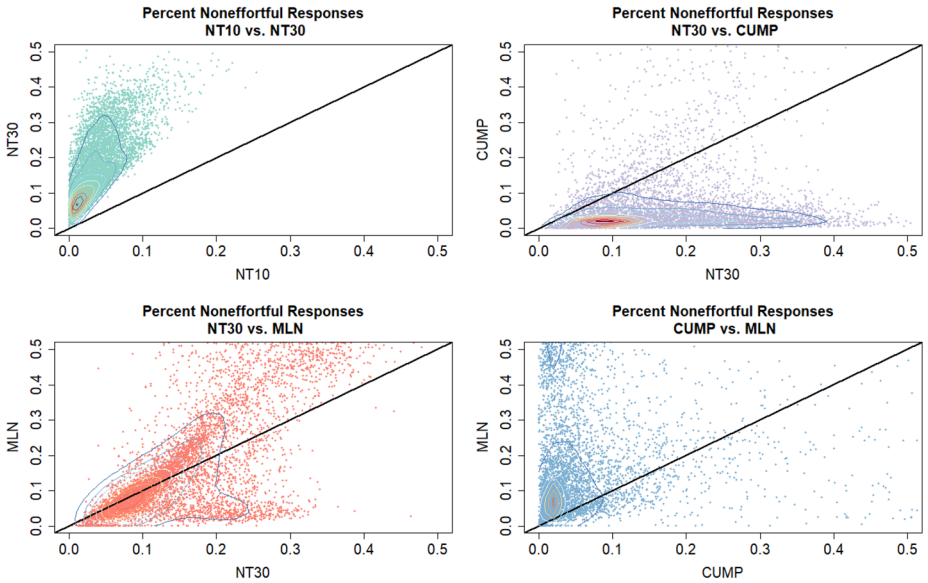

Fig. 3 Percent of item responses classified as noneffortful by threshold method

items rarely fall on the identity line. For example, NT30 and MLN tend to identify a much higher proportion of items as rapid compared to CUMP.

Discussion

Most valid uses of test scores for an intended purpose assume that examinees are providing full effort. Unfortunately, this assumption is frequently violated, with as high as 15% of examinees showing low effort sufficient to potentially bias estimated scores (Soland, 2018a), and the consequences of those intended uses can be dire. When using aggregate scores, low effort can affect policy-relevant metrics like achievement gap estimates (Kuhfeld & Soland, 2020; Soland, 2018a, b) and rank orderings of countries (Brozo et al., 2007; Debeer et al., 2014; Wise, Soland, & Bo, 2019). At the individual level, low effort can downwardly bias estimated scores by over 0.2 standard deviations (Rios et al., 2016). In short, the issue of low effort on large-scale tests, especially those that are low stakes for the examinees, can be nontrivial. Fortunately, response times can be used to identify noneffortful item responses in ways supported by decades of validity evidence chronicled by Wise (2015). In this study, we apply several threshold setting methods to large-scale CAT data and compare their performance. Our results generated several findings that can help large-scale test developers and administrators weigh the tradeoffs involved in selecting a threshold method.

First, we show that the various methods we compared produced viable thresholds for our data at very different rates. In particular, the CUMP approach often could not identify a threshold because all response accuracies conditional on response time were above the chance rate. This result likely occurred because we used a CAT, which means examinees typically get roughly half of the items right given the administered items are targeted to students' estimated ability levels. Thus, for roughly 2000 out of 8000 items, CUMP could not produce a threshold. As Rios and Guo (2020) suggested,

a possible solution is to use MLN when CUMP cannot produce a viable solution (a hybrid approach). However, we also found that MLN occasionally identifies a single group solution (essentially failing to divide the responses into effortful and noneffortful) as the best fitting, and produced thresholds that exceeded 100 s 13% of the time. NT approaches always produced viable thresholds, as one would expect.

Further, MLN frequently identified a three-group solution as fitting better than a two-group solution, and there is not much guidance available on how to interpret those three response time distributions in the context of low effort. While there is some emerging research from the mixture literature on how to treat solutions favoring more than two groups, interpretation of those various groups is not uniform across studies, nor is the number of groups identified (e.g., Wang & Xu, 2015). All told, little research has considered what three response time groups would mean for low effort. Could one still call the lowest group noneffortful? Is the group with the longest set of response times being especially diligent, or are they unfocussed and therefore providing less effort? While one could argue that the group with the fastest response times should be deemed noneffortful, the answers are unclear.

Second, we also showed that having the highest proportion of items with viable thresholds is not synonymous with being supported by the strongest validity evidence for those thresholds. In particular, while NT30 always produced viable thresholds, the proportion of item responses that were correct among those deemed noneffortful was at roughly 45%—well above the chance rate of 25%. By contrast, while CUMP often did not produce a viable threshold, its threshold-setting algorithm is designed to optimize identification of item responses that are correct at or below the chance rate. As we show in our sample, virtually all item responses were correct at around the chance rate when using CUMP. Notably, NT10 also tended to flag item responses that were correct around the chance rate, though with more variability than CUMP.

Third, we found that there are often inconsistencies in the thresholds and, therefore, the item responses identified as noneffortful by method. Notably, there tended to be some consistency between MLN and NT30, as well as between CUMP and NT10. However, those two groupings produced quite different results compared to each other. For instance, MLN and NT10 often produced very different thresholds, just as CUMP and NT30 tended to identify very different proportions of items as noneffortful.

While probably stating the obvious, these different rates of identifying responses as unmotivated has implications for how to address low effort, especially related to scoring. One approach used in large-scale operational tests to address low motivation is effort-moderated scoring (Wise & DeMars, 2006), which essentially treats item responses flagged as noneffortful as missing, then rescores the assessment. This scoring method has been shown to effectively recover item and person parameters, especially when its assumptions are met (Wise & DeMars, 2006; Rios & Soland, 2020). Thus, increasing the number of item responses flagged as noneffortful could potentially reduce the bias due to low motivation, but would also reduce the precision of the estimated scores. Emerging research (Authors, under review) indicates that bias in person parameter estimates is reduced when noneffortful responses are overidentified rather than under identified. Given those findings, assessment developers and administrators may prefer threshold-setting methods that are less conservative.

Limitations and future directions

One limitation of the present study is that, while we draw attention to issues in identifying low effort specific to CATs, we cannot make a direct empirical comparison between CAT and non-CAT item responses for the same set of students. Such a comparison might be germane to a host of relevant issues. For instance, response times might differ in general based on the test administration approach. Specifically, one could imagine CAT response times being less variable, which, in turn, may impact threshold methods. Such comparisons are worthy of additional study. Related to making comparisons (and generalizability), one should also note that results may depend on the test setting (which ability is measured, how many response options there are, whether the test is high stakes/low stakes, etc.). Thus, one cannot be sure our results will generalize to other samples and tests, though our sample is quite large relative to that used in other studies (per Table 1).

Another limitation is that, given our emphasis on large-scale operational testing, we do not compare results from more complex models designed to identify low effort. Many such approaches use mixture models to jointly model responses and response times. For example, Lun et al. (2020) propose a new mixture model for responses and response times with a hierarchical ability structure, which incorporates auxiliary information from other subtests and the correlation structure of the abilities to detect rapid guessing behavior. Similarly, Ulitzsch et al. (2020) propose a hierarchical latent response model for identifying and modeling the processes associated with examinee disengagement jointly with the processes associated with engaged responses. They outline a mixture model that identifies disengagement at the item-by-examinee level by assuming different data-generating processes underlying item responses and omissions, respectively, as well as response times associated with engaged and disengaged behavior. While such models show great promise, they are likely to be impractical in a CAT context with potentially thousands of items where local independence assumptions are violated. Nonetheless, there is a need for future research that compares the response time procedures investigated in this study with more complex IRT mixture models to determine whether there are non-negligible differences in practical outcomes, including those in large-scale assessment contexts.

Third, we narrowed our analyses only to items with 100 or more responses. While this is a limitation of the study, it also provides useful information to large-scale operational CAT researchers. We examined how many item responses we did have for the items with under 100 responses, and found that most of them had under 20. One should note that this sparseness occurred even with a sample size of over 700,000 examinees. Further, MAP Growth may be unique in the size of its item pool due to its cross-grade nature and how many years it has been administered. On one hand, this issue may be less severe for tests with smaller item banks. On the other, many tests do not have as many item responses given fewer examinees take the test.

While we considered conducting analyses regardless of how many item responses were collected, we decided against it because using only 20 items is not justifiable, we would argue, even for the simple threshold-setting approaches we used. For example, there is evidence that mixture models like the ones we used require large sample sizes to achieve accurate parameter estimates, sometimes requiring sample sizes greater

than 1000 examinees (Kim, 2012; Peugh & Fan, 2012), which is why mixture models are considered large sample techniques. Even some of the other methods are likely to run into issues. For example, CUMP cannot accurately determine the point at which the proportion of correct responses rises above the chance rate if there are no or few observations at a given response time. Further, as Wise and Ma (2012) pointed out, having only 100 item responses could be problematic for producing distributions of response time and accuracy that are consistent across samples under the NT approaches (Wise & Ma, 2012). In short, a problem that precedes selecting an appropriate threshold-setting approach is having sufficient sample sizes in sparse CAT item response matrices to use those approaches.

Considerations for developers and administrators of large-scale assessments

In summary, we find that the performance of the threshold setting procedure tends to differ dependent on the criterion used. Thus, choosing a threshold-setting method involves a series of tradeoffs. Given these tradeoffs, how should test developers and administrators make an informed choice? Below, we mention a few factors that should probably be part of the decision. Embedded in this discussion are considerations of the limitations of our own study, as well as areas where future work is needed.

Type of assessment being used

In some cases, the decision could be dependent on the type of test being used. For our study, we used a CAT. We decided on an adaptive test because they are increasingly common, and because they pose particular challenges for identification of thresholds given they (a) tend to have sparse item response matrices (a particular issue for MLN and CUMP) and (b) examinees are often given items they should get correct roughly 50% of the time. Thus, the frequency with which CUMP failed to produce viable thresholds might be a much more minor issue on fixed-form tests. As another example of how the type of test might matter, the method used could depend on whether there are speededness issues (e.g., Schnipke & Scrams, 1997). The performance of these various threshold-setting methods in the presence of speededness is a topic worthy of additional study.

Decisions related to administration

The selection of a threshold-setting method could also relate to how low effort is treated operationally. For example, some testing companies notify proctors in real time when examinees are disengaging from a test so that the proctor can intervene and, hopefully, get the examinee back on task (Wise, Kuhfeld, & Soland, 2019). Under such a scenario, using a fairly liberal threshold-setting procedure like MLN could result in proctors being notified with great frequency, potentially impeding their ability to intervene. Further, after the test is completed, some testing companies will invalidate a test if the examinee shows low effort on a certain proportion of items. An approach like NT10 would likely lead to far fewer invalidated tests than CUMP or MLN. Obviously, tradeoffs between ensuring the validity of the score produced and the potential frustration to the users of a test (e.g., students and teachers) associated with having a test invalidated would need to be weighed. Finally, research shows

that item calibrations can be biased by low effort (e.g., Wise & DeMars, 2006; Rios & Soland 2020 in EPM). As related research suggest, test developers may wish to use less conservative methods if they are still calibrating the item parameters.

Intended use of the test scores

The effect of using different thresholds on test scores is not the focus of this study. However, there is evidence that the threshold-setting approach can impact scores (e.g., Wise & Kuhfeld, 2020). Thus, potential uses of scores should be considered when selecting a threshold. At heart, the decision is either to use a conservative method, which increases false negatives (wrongly assuming an item response is effortful), or use liberal methods, which lead to the opposite.

In particular, determining whether the scores are going to be used for individual decision-making or in the aggregate (e.g., for research purposes or comparing countries) is important. If the purpose is to use aggregate scores such as when estimating achievement gaps, or comparing the performance of countries using tests like PISA, one might be more willing to potentially overidentify noneffortul responses to help avoid bias. Further, if low motivation is addressed using effort-moderated scoring, then increased standard errors of measurement for individual scores is less consequential when examining only aggregate-level scores. Therefore, in the aggregate, methods like MLN and NT30 may be preferred.

By contrast, when individual-level inferences are desired—and especially when the stakes are high for the examinee—the tradeoff between bias and precision of individual scores matters much more. On one hand, discarding noneffortful items before scoring decreases the precision of the final estimate. Thus, examinees could be misclassified with greater frequency due to measurement error. For instance, a student might be more or less likely to be placed in a gifted program or targeted for remediation. On the other, using noneffortful items in the score can produce bias, and that bias is often downward relative to the true score (Rios et al., 2016). Further, bias can be particularly impactful when ability and effort are correlated, such as when students with discrepant ability levels differ in the level of effort they display (Rios & Soland, 2020). For example, low-ability examinees might be more likely to fall below a cut score (Soland & Kuhfeld, 2019). Returning to the gifted program hypothetical, low-income students, who are often lower performing on average, may be less likely to be placed in gifted education due to low effort.

Definition of low effort

Finally, the decision could come down to how much clarity one wishes to have over how low effort is defined. For example, NT10 clearly (if implicitly) defines low effort as rapid guessing—a response should only be discounted if it was provided so quickly, the content could not have been understood. CUMP also uses a very clear definition: the response time threshold should separate examinees getting items correct below the chance rate from those getting them correct above the chance rate.

However, the definitions for NT30 and MLN are less clear. While the latter is a parameterized version of the visual inspection method (which was focused on rapid guessing),

the thresholds for MLN are often above 100 s, which does not seem to pass a face validity test for being a rapid guess. Further, the three-group solution often fits best for MLN. Thus, MLN may effectively identify different response time distributions, but what it means, exactly, to fall in each distribution is unclear (and is a topic meriting additional study). Similarly, NT30 appears to produce effort-moderates scores that better close the gap between tests invalidated due to low effort and those re-taken under higher effort (Wise & Kuhfeld, 2020), but the exact definition of low effort is unclear. For test developers and administrators, if effective communication of how low effort is treated is of importance, then some methods may be preferable to others.

Conclusion

On one hand, there are a variety of increasingly sophisticated methods available for setting thresholds. On the other, most of these methods have not been compared, nor have they been tried using data from a large-scale test. In this study, we address both limitations of the literature and, in so doing, produce findings that can help developers and administrators of large-scale assessments make informed decisions about threshold setting. Specifically, we show that, using our large-scale operational CAT data, both the frequency with which threshold-setting methods identify viable thresholds and the response time associated with those thresholds differs considerably by method. Through these analyses, we try to help large-scale test developers and administrators understand the tradeoffs inherent in selecting a method, including providing criteria for that selection.

Acknowledgements
We would like to thank Steve Wise at NWEA for his feedback in the conceptualization phase of the study.

Authors' contributions
All three authors made contributions to drafting the manuscript. The analytical work was conducted by Dr. Soland and Dr. Kuhfeld.

Funding
None.

Availability of data and materials
Data are proprietary and not publicly available.

Declarations

Competing interests
None.

Author details
[1] School of Education and Human Development, University of Virginia, 405 Emmet Street, Charlottesville, VA 22904, USA. [2] NWEA, 121 NW Everett Street, Portland, OR 97209, USA. [3] College of Education and Human Development, University of Minnesota, 56 E River Rd, 164 Education Sciences Building, Minneapolis, MN 55455-0364, USA.

Received: 14 August 2020 Accepted: 7 April 2021
Published online: 27 April 2021

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (Eds.). (2014). Standards for educational and psychological testing. American Educational Research Association.

Brozo, W. G., Shiel, G., & Topping, K. (2007). Engagement in reading: Lessons learned from three PISA countries. *Journal of Adolescent and Adult Literacy, 51*(4), 304–315.

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.

Demars, C. E. (2007). Changes in rapid-guessing behavior over a series of assessments. *Educational Assessment, 12*(1), 23–45.
Goldhammer, F., Naumann, J., Stelter, A., Tóth, K., Rölke, H., & Klieme, E. (2014). The time on task effect in reading and problem solving is moderated by task difficulty and skill: Insights from a computer-based large-scale assessment. *Journal of Educational Psychology, 106*(3), 608.
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.
Jensen, N., Rice, A., & Soland, J. (2018). The influence of rapidly guessed item responses on teacher value-added estimates: Implications for policy and practice. *Educational Evaluation and Policy Analysis, 20*(1), 90–98.
Kim, S.-Y. (2012). Sample size requirements in single- and multiphase growth mixture models: A Monte Carlo simulation study. *Structural Equation Modeling: A Multidisciplinary Journal, 19*(3), 457–476.
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*(1), 606–619.
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.
Kuhfeld, M., & Soland, J. (2020). Using assessment metadata to quantify the impact of test disengagement on estimates of educational effectiveness. *Journal of Research on Educational Effectiveness, 13*(1), 147–175.
Kyllonen, P. C. (2012). *Measurement of 21st century skills within the common core state standards*. . New Jersey: Educational Testing Service (ETS).
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.
Lu, J., Wang, C., Zhang, J., & Tao, J. (2020). A mixture model for responses and response times with a higher-order ability structure to detect rapid guessing behaviour. *British Journal of Mathematical and Statistical Psychology, 73*(2), 261–288.
Meyer, J. P. (2010). A mixture Rasch model with item response time components. *Applied Psychological Measurement, 34*(7), 521–538.
Molenaar, D., & de Boeck, P. (2018). Response mixture modeling: Accounting for heterogeneity in item characteristics across response times. *Psychometrika, 83*(2), 1–19.
Peugh, J., & Fan, X. (2012). How well does growth mixture modeling identify heterogeneous growth trajectories? A simulation study examining GMM's performance characteristics. *Structural Equation Modeling: A Multidisciplinary Journal, 19*(2), 204–226.
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., Soland, J. (2020). Parameter Estimation Accuracy of the Effort-Moderated Item Response Theory Model Under Multiple Assumption Violations. *Educational and Psychological Measurement* 0013164420949896.
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., Guo, H., Mao, L., & Liu, O. L. (2016). Evaluating the impact of careless responding on aggregated-scores: To filter unmotivated examinees or not? *International Journal of Testing, 17*(1), 1–31.
Sahin, F., & Colvin, K. F. (2020). Enhancing response time thresholds with response behaviors for detecting disengaged examinees. *Large-scale Assessments in Education, 8*, 1–24.
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.
Schnipke, D. L., & Scrams, D. J. (2002). Exploring issues of examinee behavior: Insights gained from response-time analyses. *Computer-Based Testing: Building the Foundation for Future Assessments, 25*(1), 237–266.
Soland, J. (2018b). The achievement gap or the engagement gap? Investigating the sensitivity of gaps estimates to test motivation. *Applied Measurement in Education, 31*(4), 312–323.
Soland, J. (2018a). Are achievement gap estimates biased by differential student test effort? Putting an important policy metric to the test. *Teachers College Record*, *120*(12).
Soland, J., & Kuhfeld, M. (2019). Do students rapidly guess repeatedly over time? A longitudinal analysis of student test disengagement, background, and attitudes. *Educational Assessment, 24*(4), 327–342.
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), 1–16.
Soland, J., Zamarro, G., Cheng, A., & Hitt, C. (2019). Identifying naturally occurring direct assessments of social-emotional competencies: The promise and limitations of survey and assessment disengagement metadata. *Educational Researcher, 48*(7), 466–478.
Ulitzsch, E., von Davier, M., & Pohl, S. (2020). A hierarchical latent response model for inferences about examinee engagement in terms of guessing and item-level non-response. *British Journal of Mathematical and Statistical Psychology, 73*(1), 83–112.
van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika, 72*(3), 287–308.
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. (2015). Effort analysis: Individual score validation of achievement test data. *Applied Measurement in Education, 28*(3), 237–252.
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., & 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). Using retest data to evaluate and improve effort-moderated scoring. *Journal of Educational Measurement*.
Wise, S. L., & Ma, L. (2012). Setting response time thresholds for a CAT item pool: The normative threshold method. *Annual Meeting of the National Council on Measurement in Education, Vancouver, Canada*. https://nwea.org/content/uploads/2012/04/Setting-Response-Time-Thresholds-for-a-CAT-Item-Pool.pdf.
Wise, S. L., Kuhfeld, M. R., & Soland, J. (2019). The effects of effort monitoring with proctor notification on test-taking engagement, test performance, and validity. *Applied Measurement in Education, 32*(2), 183–192.
Wise, S. L., Soland, J., & Bo, Y. (2019). The (non) impact of differential test taker engagement on aggregated scores. *International Journal of Testing, 20*(1), 57–77.
Zamarro, G., Hitt, C., & Mendez, I. (2016). When students don't care: Reexamining international differences in achievement and non-cognitive skills. *Journal of Human Capital, 13*(4), 519–552.

Publisher's Note