Large-scale Assessments in Education
a SpringerOpen Journal

RESEARCH Open Access

Using response time to investigate students' test-taking behaviors in a NAEP computer-based study

Yi-Hsuan Lee[1*] and Yue Jia[2]

* Correspondence: ylee@ets.org
[1]Educational Testing Service, MS 12T, Rosedale Road, Princeton, NJ 08541, USA
Full list of author information is available at the end of the article

Abstract

Background: Large-scale survey assessments have been used for decades to monitor what students know and can do. Such assessments aim at providing group-level scores for various populations, with little or no consequence to individual students for their test performance. Students' test-taking behaviors in survey assessments, particularly the level of test-taking effort, and their effects on performance have been a long-standing question. This paper presents a procedure to examine test-taking behaviors using response time collected from a National Assessment of Educational Progress (NAEP) computer-based study, referred to as MCBS.

Methods: A five-step procedure was proposed to identify rapid-guessing behavior in a more systematic manner. It involves a non-model-based approach that classifies student-item pairs as reflecting either solution behavior or rapid-guessing behavior. Three validity checks were incorporated in the validation step to ensure reasonableness of the time boundaries before further investigation. Results of behavior classification were summarized by three measures to investigate whether and how students' test-taking behaviors related to student characteristics, item characteristics, or both.

Results: In the MCBS, the validity checks offered compelling evidence that the recommended threshold-identification method was effective in separating rapid-guessing behavior from solution behavior. A very low percent of rapid-guessing behavior was identified, as compared to existing results for different assessments. For this dataset, rapid-guessing behavior had minimum impact on parameter estimation in the IRT modeling. However, the students clearly exhibited different behaviors when they received items that did not match their performance level. We also found disagreement between students' response-time effort and self reports, but based on the observed data, it is unclear whether the disagreement was related to how the students interpreted the background questions.

Conclusions: The paper provides a way to address the issue of identifying rapid-guessing behavior, and sheds light on the question about students' extent of engagement in NAEP and the impact, without relying on students' self evaluation or additional costs in test design. It reveals useful information about test-taking behaviors in a NAEP assessment setting that has not been available in the literature. The procedure is applicable to future standard NAEP assessments, as well as other tests, when timing data are available.

Keywords: Response time; Low-stakes assessment; Engagement; Motivation; Solution behavior; Rapid-guessing behavior; Speededness

Background

Large-scale national and international survey assessments, such as the National Assessment of Educational Progress (NAEP), the Programme for International Student Assessment (PISA), and the Trends in International Mathematics and Science Study (TIMSS), have been used for decades to monitor what students know and can do. Those survey assessments are often referred to as low-stakes assessments as they aim to provide group-level scores for various populations, and students taking the assessments receive no academic credit and bear little or no consequences for their test performance. One long-standing question for low-stakes assessments is the level of student's engagement with the test and its effect on performance (Braun, Kirsch, and Yamamoto 2011; O'Neil et al. 1995), particularly for students at higher grade levels. Braun et al. (2011) further suggested that differences in engagement with the test might be confounded with differences in the cognitive abilities to be assessed. From a measurement perspective, responses associated with disengagement could lead to model misfit and biased parameter estimation in item response theory (IRT) calibration and scoring (Wise and DeMars 2006; Wise and Kong 2005). To improve the quality of parameter estimation in measurement models, and therefore the validity of group score estimates, one solution is to identify responses from individual students showing disengagement with the test/items and remove them from the analysis.

A common approach to evaluating students' engagement with the test in low-stakes assessments is to employ self-report questionnaires. As noted by Wise and Kong (2005), one limitation of self-reporting is that it is difficult to ascertain how truthfully the students respond to the questionnaires. It is also doubtful how the students interpret the questions about their engagement with the test. For example, NAEP collects evidence from the questionnaires on whether students are trying hard on NAEP or are engaged. Twelfth graders who said they did not try hard on NAEP scored higher than those who said they tried harder or much harder compared to most other tests taken in the same year in school (National Assessment Governing Board 2005). Another approach is to conduct experimental studies to examine whether certain practices, such as offering monetary incentives, can effectively motivate students (Braun et al. 2011). This type of approach is often implemented on a small scale and not feasible to be employed on a regular basis.

A different school of thought in measuring student engagement is to distinguish students who demonstrate solution behavior from those showing rapid-guessing behavior. Generally, rapid-guessing behavior is represented as responses occurring so rapidly that students either do not have time to fully consider the item (i.e., test speededness) or do not give their effort. The accuracy of such rapid guesses is typically at or near the chance level, as the responses are essentially random. The response time (RT) of rapid guesses is usually very short relative to the amount of time required for the items. A few researchers have proposed IRT-based approaches to model item responses in considering the above mentioned test-taking behaviors (e.g., Bolt et al. 2002; Yamamoto and Everson 1997). Their main concern was how test speededness affected the estimation of IRT model parameters. (Hadadi and Luecht 1998) indicated that a better set of detection methods for identifying rapid-guessing behavior may be developed given the availability of detailed data of responses and RTs.

In the literature, item-level RT distributions involving both solution behavior and rapid-guessing behavior usually have a bimodal shape (Lee and Chen 2011). Such RT distributions have been observed in both high-stakes and low-stakes assessments as a result of lack of time (e.g., Schnipke and Scrams 1997), lack of effort (e.g., Wise and Kong 2005), or other reasons. This line of research makes the assumption that students with rapid-guessing behavior would show very short RTs. For example, Figure 1 shows the RT distribution of an item in the data set examined in this paper. The gap, which clearly separates the two groups of RTs at around 20 seconds, could be considered a threshold for the item that represents the RT boundary between solution behavior and rapid-guessing behavior (Wise and Kong 2005). A number of research studies have shown valuable gain in improving the estimation of measurement consequences (such as item difficulty, student ability, test reliability, and test validity) by filtering out rapid guesses based on some predefined RT thresholds (Kong et al. 2006; Wise et al. 2006a; Wise and DeMars 2006; Wise et al. 2006b).

Identifying the RT boundary between solution behavior and rapid-guessing behavior for each item is a crucial step in the analysis, and several methods for threshold identification have been suggested. For instance, Kong et al. (2007) recommended four different ways to decide RT thresholds. One of them uses a common three-second threshold for all items, which is easiest to implement. Another is to visually inspect the RT distribution of each item, which is a more widely used approach (e.g., DeMars 2007; Setzer et al. 2013). A gap in the distribution that clearly separates two groups of RTs suggests a possible threshold for the item. The third suggested approach is to set up the threshold based on the amount of reading required for an item. The last is to apply mixture models to fit the RT distribution of each item (also see Meyer 2010;

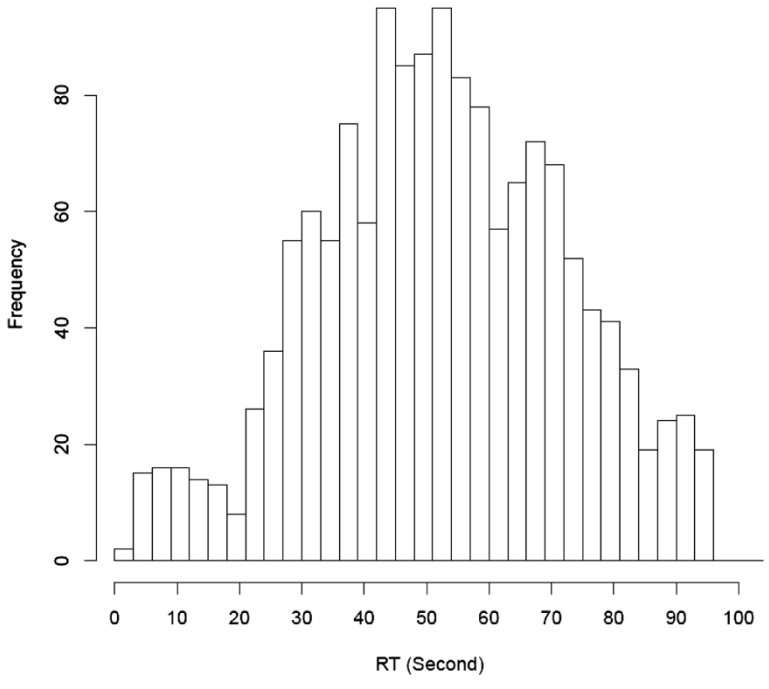

Figure 1 The RT distribution for an MCBS item that shows a bimodal pattern (the gap around 20 seconds could be used as the threshold that separates the two clusters of RTs).

Schnipke and Scrams 1997). Kong et al. (2007) analyzed data from an Information Literacy Test using the common three-second rule and several variable-threshold methods, concluding that the common three-second rule performed slightly worse than the others. In a separate study, Hauser and Kingsbury (2009) found it rather conservative to assign a common three-second threshold to all items when some items required more time than the others. They also noted that item thresholds should be unique to each item, depending on the time demand of an item. Ma et al. (2011) further examined, among others, the use of mixture models and a non-parametric model with mathematics items of a computerized adaptive test, finding that neither model was practically useful. Wise and Ma (2012) proposed a new variable-threshold method, called the normative threshold method, which defines the RT threshold of an item as a certain percent of the average RT. This method can be easily applied to assessments with large item pools. Conversely, it is possible to detect aberrant behavior of students using person-fit statistics (e.g., Karabatsos 2003; Meijer and Sijtsma 2001), many of which are designed to identify general types of aberrances in behaviors by studying patterns of responses (Lee et al. 2014). The research studies mentioned above used RTs to classify test-taking behaviors. There have also been studies that modeled responses and RTs jointly (e.g., van der Linden 2007), assuming a unimodal distribution for RTs with elaborate parameterization. Such models are appropriate for application in a single latent class (e.g., representing solution behavior) rather than identifying different test-taking behaviors. Readers may refer to Schnipke and Scrams (2002) and Lee and Chen (2011) for more exhaustive reviews of existing methods.

Large-scale survey assessments have been commonly administered in a paper-and-pencil (P&P) format. More recently, the programs such as NAEP and PISA have begun to consider computer-delivered testing formats. A special NAEP study, referred to as Mathematics Computer-Based Study (MCBS) herein, was conducted in 2011 to assess the benefit of multistage testing in the NAEP administration setting (Xu et al. 2012). The MCBS provided detailed timing data, which permits the investigation of student behaviors in terms of RT, as the students move through the test. It also helped answer a practical, long-term question—can we evaluate NAEP students' extent of engagement when timing data are available? Because the MCBS had time limits at both stages of the test (see the Methods section for more descriptions), test speededness should not be completely precluded, even though disengagement with the test is of primary concern in low-stakes assessments. Thus, this study focuses on general rapid-guessing behavior, possibly a result of disengagement with the test and test speededness. The observed effects of rapid-guessing behavior serve as an upper bound of the effects of disengagement with the test.

Because this research is the very first attempt to investigate student behaviors using RTs in NAEP assessments and in low-stakes educational survey assessments at large, we adopted a non-model-based approach to examining the properties of RTs and classifying behaviors. The paper reveals useful information about test-taking behaviors in a NAEP assessment setting that has not been available in the literature. To our best knowledge, the existing studies on this topic have typically focused on timing distributions with clear bimodal patterns and their methods may be less effective when the bimodal patterns are less obvious. As will be discussed, the recommended five-step procedure in this paper aims to (a) strengthen the existing approaches to address the

circumstance where rapid-guessing behavior is of concern and to be assessed, but not all items have clear bimodal RT distributions, and to (b) identify rapid-guessing behavior in a more systematic manner. Step 2 of the procedure intends to identify time thresholds between rapid-guessing and solution behaviors with different threshold-identification methods. The method we recommended has been explored in the literature but was tailored in this paper to accommodate the new application to the MCBS data. In addition, new and existing validity checks were incorporated in the validation step to ensure reasonableness of the time boundaries before further investigation. The validity checks are important to achieve the objective of identifying rapid-guessing behavior in an effective manner while attempting to avoid misclassifying responses representing solution behavior as rapid guesses.

Methods

Data collection in the MCBS

The MCBS was administered under similar testing conditions (e.g., time limit, test length, and consequences to the students) as regular NAEP assessments, with the only major difference being the test format.

Items and test forms

Six test forms were administered in a two-stage design for the MCBS. The items were taken from an existing item pool to form five 25-minute blocks. The existing item pool was composed of 8^{th} grade mathematics items from the 2011 NAEP P&P operational assessment.

Among the five blocks of items, two of them served as routing blocks at the first stage (Router 1 and Router 2), and the other three were used at the second stage with items of high (Hard block), medium (Medium block), and low difficulties (Easy block). The test form for each student consisted of two separately timed blocks, one routing block and one second-stage block. Either routing block consisted of 17 multiple-choice items, while each second-stage block had 16 items that were either multiple-choice or constructed-response questions. All of the multiple-choice items had five options, except for two items with only four options (one in Router 2 and the other in the Easy block). In this study, item P + was defined as the proportion of correct responses for a dichotomous item, or the average observed score for a polytomous item. Table 1 summarizes the average item P + for the five blocks. Within each timed block, the students could move freely among items, edit or change their answers, and skip questions and return to them later. However, the students were not allowed to move across the two blocks in a test form.

Table 1 Average P + for the blocks in the MCBS

Stage	Block	Number of items	Average P+
Stage 1	Router 1	17	0.57
	Router 2	17	0.59
Stage 2	Easy	16	0.77
	Medium	16	0.56
	Hard	16	0.36

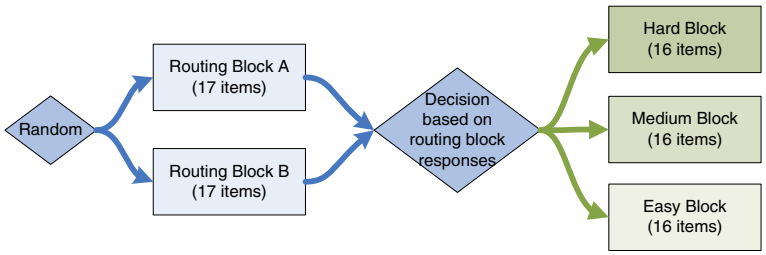

Figure 2 Flowchart of block assignment for the MST sample.

Design

The experimental study involved two testing conditions, which led to two student samples: Multistage Test (MST) sample and Control sample. The test given to the MST sample was adaptive at the block level, while the test for the Control sample was linear. In both the MST sample and the Control sample, the students were classified into three performance levels--low, medium, and high--based on their performance on the routing blocks[a] (see Xu et al. 2012, for more discussion about the classification rule in the MCBS). We name the three subgroups in the MST sample as MST_Low, MST_Medium, and MST_High. Similarly, Control_Low, Control_Medium, and Control_High refer to the three subgroups in the Control sample.

As illustrated in Figure 2, all the students in the MST sample were randomly assigned at the first stage to one of the two routing blocks. At the second stage, the students received a block that best matched their estimated performance level. That is, MST_Low took the Easy block, MST_Medium took the Medium block, and MST_High took the Hard block. By contrast, the students in the Control sample were randomly assigned a routing block and a second-stage block, regardless of their performance on the routing block (Figure 3). Thus, the students in each of the Control_Low, Control_Medium, and Control_High had equal probabilities to receive the Easy, Medium, and Hard blocks. Disaggregating the two student samples by performance level made it possible to examine whether the students exhibited different behaviors when there was discrepancy between students' performance level and block difficulty.

Students

A national representative sample of 8,400 students participated in this study, where about 40% of the students were placed in the MST sample and 60% in the Control sample[b].

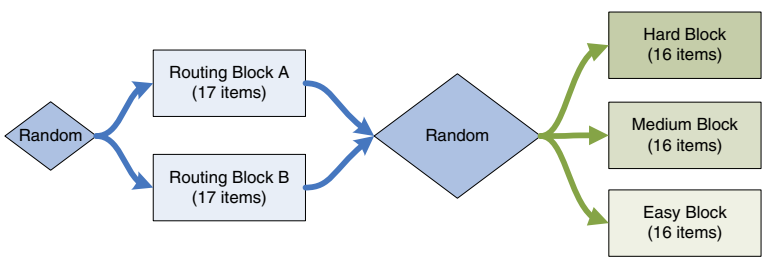

Figure 3 Flowchart of block assignment for the Control sample.

Analyses

Defining RT thresholds

The RT for an item was defined as the total time spent on the item[c]. As mentioned previously, we used RTs to classify student-item pairs as reflecting either solution behavior or rapid-guessing behavior and to investigate whether student behaviors were connected to person characteristics, item characteristics, or both. Accomplishing the objectives requires a threshold to be defined for each item that represents the RT boundary between the two test-taking behaviors. Following the convention of the literature, the RT threshold per item is defined as the lower bound for the amount of time an average student (in terms of speed) needs to fully consider the item and answer it. In other words, the RTs classified as representing rapid-guessing behavior may include the time spent browsing through items for some individuals, as well as the time spent reading and processing the materials for others (yet insufficient to answer the items). Clearly, the former may be equally short for all items, but the latter depends on the time demand of individual items.

One complication for the study is that the MCBS was a multistage test and there were two student samples (MST and Control) assigned to the second-stage blocks based on different rules. To determine if different RT thresholds are necessary for an item given to the students from different samples and performance groups, it requires knowledge about how item-level RTs distributed in different student samples and in the further disaggregated performance groups. Thus, we first examined the RT distributions for all items by block and by student sample, based on descriptive statistics of the RT distributions and histograms. This analysis showed whether and how the RT distribution of each item differed across student samples in each of the blocks. Second, we focused on data collected from the Control sample and the second stage blocks. Within each of these blocks, the Control sample was disaggregated into three performance groups and the associated RT distributions were compared numerically and graphically. This analysis suggested if the students of varying performance levels revealed different RT patterns when taking a block that may not match their performance level.

As noted in the Background section, several methods for setting RT thresholds have been considered. The literature also suggests that RT thresholds are likely to be item-specific (e.g., Hauser and Kingsbury 2009)—for example, an item with more text, with a table and/or figure, or with more complicated stems may take more time to read and process. Preliminary analysis of our timing data led to consistent findings, and it appeared unreasonable to use a common threshold for all items. In particular, the three-second rule proposed by Kong et al. (2007) did not work in the MCBS since almost no item had any RTs shorter than three seconds. In addition, our exploratory analysis revealed that about one third of the MCBS items had a unimodal timing distribution with a heavy left tail at short RTs, instead of the clear bimodal pattern shown in Figure 1 (see the Results section for more discussion). Existing methods that rely on the bimodal pattern, such as visual inspection of RT distributions and model-based approaches (i.e., mixture models and non-parametric models), did not work in this situation.

Two approaches can be considered in this circumstance. One is the normative threshold method (Wise and Ma 2012). The authors recommended the RT threshold of an item as 10 percent of the average RT for that item, up to a maximum threshold value of 10 seconds. The 10-second ceiling helped prevent extremely large thresholds, as sample mean is sensitive to outliers in RT values.

The other approach is visual inspection of RT distributions with conditional P + information (VITP). This approach was considered in Ma et al. (2011), among other methods. More precisely, for each multiple-choice item, we overlaid the frequency distribution (i.e., histogram) of RTs with the conditional P + evaluated at every possible RT value. The idea is that, for each multiple-choice item, short RTs that are connected with a P + near the chance level (i.e., 0.2 for five-option items and 0.25 for four-option items) are likely to involve rapid-guessing behavior. Thus, an RT threshold was defined for each multiple-choice item as the right end of a cluster of short RTs whose P + fluctuated around 0.2 or 0.25, depending on the number of options. In general, the following rules were applied to the multiple-choice items[d]: For each item with a bimodal RT distribution, the threshold was defined as the RT value at which there was a gap in the RT distribution and the conditional item P + stopped fluctuating around the chance level. For an item with a heavy left tail (rather than bimodal patterns) in the RT distribution and the conditional P + fluctuated around the chance level in the left tail, then the RT threshold was determined as the maximum RT value in the left tail where the conditional P + started to stay consistently above the chance level. For any item without clear bimodal RT patterns and the conditional P + was always higher than the chance level, no RT threshold was set for that item (this case occurred to the 10 items classified under "NA" in Table 2). Because there was no expected item P + under rapid-guessing behavior for constructed-response items, no RT thresholds were set for those items.

The timing and P + patterns in Figure 4 were used to decide on a threshold. Some items, like this one, had a bimodal RT histogram, and the gap between the faster response mode (assumed to indicate guessing) and the slower response mode (assumed to indicate solution behavior) was an indication of the time boundary between the two behaviors. The visual inspection of RT distribution suggested a 20-second threshold for this item. A second indication of the presence of rapid-guessing behavior was the conditional P + for each RT bin, which is shown as the connected circles in the figure. On this item, the students were not consistently responding above the chance level, which is indicated by the horizontal dashed line at conditional P + equal to 0.2, until about 20 seconds (the conditional P + fluctuated greatly before 20 seconds due to the small sample size per second). The timing and conditional P + information for this item converged on a plausible indication of the presence of rapid-guessing behavior if less than 20 seconds was spent. On the other hand, when applying the normative threshold method to that item, 10 percent of the average RT was equal to 8.49 seconds.

Table 2 A summary of response time thresholds in the MCBS

Stage	Block	Number of total/MC items	Number of items for each RT threshold						
			VITP				Normative Threshold		
			5 sec	10 sec	20 sec	NA	1-5sec	6-9 sec	10 sec
Stage 1	Router 1	17/17	0	10	5	2	3	9	5
	Router 2	17/17	0	14	3	0	5	9	3
Stage 2	Easy	16/14	1	7	3	3	6	8	2
	Medium	16/13	1	6	3	3	6	5	5
	Hard	16/13	0	7	4	2	5	7	4

Note: RT thresholds were only defined for the multiple-choice (MC) items in the VITP approach.

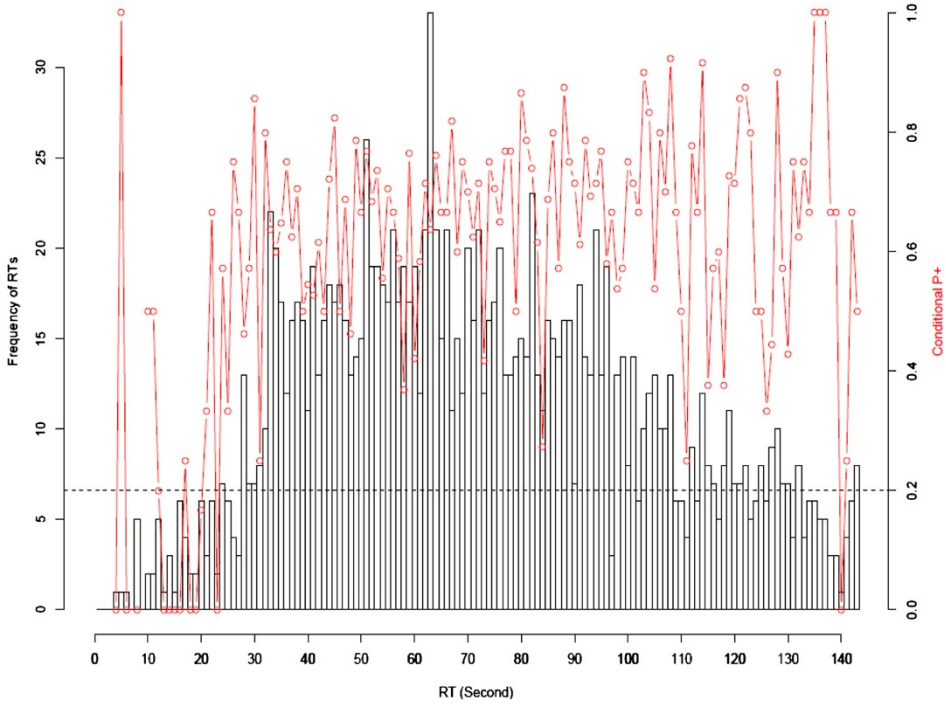

Figure 4 Histogram of RTs and conditional P + for an item exhibiting identifiable patterns (truncated at the 90th percentile of RTs and based on the Control sample). The bars represent the frequency of RTs at every possible RT, the circles joined by line represent the conditional P+, and the horizontal dashed line at conditional P + equal to 0.2 represents the chance level.

Because of the MST design, the RT distribution for some items, especially those in the second-stage blocks, may differ between the student samples (MST and Control) and among the further disaggregated performance groups. For each item with comparable RT distributions among the two student samples, the student samples were combined to produce one set of RT distributions and conditional P+, which led to one RT threshold defined. Alternatively, for items with somewhat different RT distributions for the two student samples, the RT distributions and conditional P + were examined separately for the student samples, and the plausible RT thresholds based on either sample were compared.

Validity checks

After the thresholds were determined, three validity checks were performed. First, we reviewed the actual items to make sure the thresholds corresponded to the reading load and complexity of individual items. For this validation step, the amounts of time that the majority of the students spent on an item (i.e., mode and median of the students' RTs) were used to quantify item complexity, and whether the item had tables or figures was taken into consideration. Second, we compared the P + associated with solution behavior with the P + associated with rapid-guessing behavior. Ideally, the former should be much higher than the chance level, while the latter should be close to the chance level. This validity check was first proposed by Wise and Kong (2005) and employed in Wise and Ma (2012) in developing their normative threshold method. Third, we evaluated the relationship between students' overall scores and their performance on each item conditioning on their behaviors on that item: For each item, we

divided the students who took the item into 10 equal-sized (score) groups ordered by their scores, and then calculated the conditional P + for each score group. For students judged to be engaged in solution behavior on an item, their performance on the item should be positively related to their overall scores; however, such a relationship is not expected for students judged to be disengaged in solution behavior on that item. It is worth noting that, although the effort-moderated IRT model proposed in Wise and DeMars (2006) was based on the idea of the third approach, it was first utilized in this study as a validity check for the identified RT thresholds.

Classifying behaviors

For items with an RT threshold identified, a dichotomous index of item solution behavior (SB) was computed for each student-item pair (see, e.g., Wise and Kong 2005) as follows: SB = 1 if the RT was greater than the threshold; SB = 0 otherwise. This index was used as an indicator for solution behavior at the person-item level. We summarized the SB index for all students and items as a student by item two-way table, and examined the table marginally and conditionally to form three different measures, aiming to investigate whether and how students' test-taking behaviors related to student characteristics, item characteristics, or both:

- Response-time effort (RTE; Wise and Kong 2005): Aggregate the two-way SB table marginally by student across items. This person-level measure represents the proportion of test items for which a student exhibited solution behavior. RTE can be used to categorize behaviors. One can then study whether RTE correlates with students' performance on the test or with different measures of engagement available for the students.
- Response-time fidelity (RTF; Wise 2006): Aggregate the two-way SB table marginally by item across students. This item-level measure represents the proportion of students exhibiting solution behavior to an item. RTF can be used to examine whether items from a particular block or items with certain characteristics (e.g., in terms of IRT item parameters) tend to evoke rapid-guessing behavior. One can also study whether RTF correlates with item position and content area.
- Conditional RTF: Compute the RTF conditional on students' performance level, rather than based on all students as is accomplished in the marginal analysis above. This is a new measure at the item-by-group level. It shows whether and how item RTF relates to students' performance level. (Conditioning RTF on performance was a natural choice with data from a multistage design. Other variables, such as demographic subgroups, may be considered in different applications).

Some analyses described above involve IRT model parameters. Typically, NAEP uses the three-parameter logistic (3PL) model for dichotomously scored items and the generalized partial credit model for polytomously scored items. In our study, item parameters were estimated by the maximum likelihood estimates, and the expected a posteriori (EAP) estimate was computed for each student as their MCBS score.

Response-time effort (RTE) for students

We first examined whether the classification of behaviors was reliable. Because the SB index is a binary variable for each student-item pair, reliability of the RTE index (or,

equivalently, the sum of the SB index across items for each student) was estimated using Cronbach's alpha (Cronbach 1951). In addition, we examined the correlation between the RTE index and the students' responses to the background questions on engagement. Two background questions and possible answers available in the MCBS are as follows:

- Question 1: How important was it to you to do well on this test? (Not very important, somewhat important, important, or very important)
- Question 2: How hard did you try on test compared to other tests? (Tried not as hard, tried about as hard, tried harder, or tried much harder)

There were only 0.6% of missing responses to these questions, so they were excluded from the computation of the following correlation measures. As will be shown in the Results section, there was very high concentration on RTE = 1 in the MCBS, which may constrain some measures of association from having high values. To take this issue into account, we considered three measures of association that were defined differently: (a) Pearson correlation, (b) Spearman's rho, and (c) Goodman and Kruskal's gamma. Pearson correlation and Spearman's rho were based on continuous RTE values (i.e., higher for students more consistently showing solution behavior during the test) and scores of the response categories (1 = not very important/tried not as hard to 4 = very important/tried much harder). On the other hand, Goodman and Kruskal's gamma was based only on the number of concordant and discordant pairs of observations on an ordinal scale and should not be largely affected by the skewed RTE distribution. To compute Goodman and Kruskal's gamma, we classified the RTE values into three categories—"High RTE" if RTE was equal to 1, "Medium RTE" if RTE was between 0.8 and 1, and "Low RTE" if RTE was no greater than 0.8, and then treated the RTE categories and the responses to each background question as ordinal variables. The frequency distribution of the response categories was also tabulated. The cutoffs for classifying RTE into high, medium, and low are arbitrary and mainly for purposes of demonstration. Varying the cutoffs did not change the observed relationship between RTE and the background questions in the MCBS. In addition, we looked at the distribution of the responses to each background question per RTE category using clustered bar charts. The (overall) relationship between RTE and EAP score for the students was evaluated through Pearson correlation of the two variables; their relationship conditioning on student sample and performance level was further examined using scatter plots.

Response-time fidelity (RTF) for items

Because the SB index was defined only for items with an RT threshold, RTF was not available for all items in the MCBS. We first correlated item RTF with item position in each block. There were five content areas assessed in the MCBS—algebra, data analysis statistics and probability (abbreviated as data), geometry, measurement, and number properties and operations (abbreviated as number), and it was of interest to examine whether the students' test-taking behavior was associated with the items' content area. Thus the box plot of item RTF by content area was made. The relationship between RTF and item characteristics was further explored by plotting the RTF

values against the estimated item discrimination and difficulty parameters obtained from the IRT models[e]; Pearson correlation was also computed.

Conditional RTF by student's performance level

We evaluated conditional RTF, or RTF by students' performance level. Recall that each of the two samples (MST and Control) was disaggregated into low, medium and high performance levels, yielding a total of 6 subgroups (i.e., MST_Low, MST_Medium, MST_High, Control_Low, Control_Medium, and Control_High). Given the experimental design, each routing block was taken by students from all 6 subgroups, while the three second-stage blocks were each taken by students from 4 subgroups—all 3 subgroups from the Control sample and one subgroup from the MST sample that matched the block difficulty. Because student performance was confounded with their test-taking behavior, it would not be surprising if RTF based on the low performing groups were consistently lower than those based on the higher performing groups. However, conditional RTF was computed and plotted to see if, for example, any item had particularly low RTF values for some or all subgroups.

Data filtering

The last part of the analysis concerns the effects of data filtering on parameter estimation based on the results of behavior classification. We calibrated the items under two conditions: one included all responses, while the other only included responses connected with solution behavior (SB = 1). The two sets of item parameter estimates were then compared.

Steps involved in analysis of a data set

In summary, our procedure to determine if using RT to evaluate test-taking behaviors is likely to provide useful information for an assessment involves the following steps:

1. Conduct exploratory analysis with RT data to examine the properties of RTs.
2. Define RT thresholds. We considered two approaches in this study:
   - Visual inspection of RT distributions with conditional P + information (VITP)
   - Normative threshold method
3. Conduct three validity checks to ensure reasonable RT thresholds. (The following two steps are performed with the validated RT thresholds).
4. Classify behaviors by defining the SB index, an indicator for solution behavior at the person-item level. Evaluate the two-way SB table at the person level (RTE), the item level (RTF), and the item-by-group level (conditional RTF), and relate these measures to student and item characteristics.
5. Perform data filtering by excluding responses with SB = 0 from IRT calibration. Examine the effects of data filtering on the calibration results.

Results

By examining how RTs distributed in the two student samples (MST and Control) and in the further disaggregated performance groups, we had the following observations: First, the RT distributions for all routing-block items were similar between the student

samples. This observation is expected given that the two student samples were randomly equivalent and the students within each sample were also randomly assigned to either routing block. Second, the RT distributions for items in each second-stage block were somewhat different between the student samples. The differences were mainly at the lower tail of the RT distributions and most noticeable for the Hard-block items, resulting from differences among the performance groups. In particular, the RT distributions for higher-performing students had a much shorter left tail than did the RT distributions for lower-performing students. These observations are not surprising as by design, each block at the second stage was taken by students of varying performance levels in the MST and the Control samples. Finally, only about two thirds of the items revealed an apparent bimodal shape as in Figure 1, and the rest had a unimodal RT distribution with a heavy left tail at short RTs. When the bimodal RT pattern appeared, it was primarily for items closer to the end of a block or for items taken by students whose performance level was low relative to the difficulty of the items/block.

Defining RT thresholds

Recall that two approaches were applied to determine the RT threshold for each item: the VITP approach and the normative threshold method. As discussed earlier in this section, the RT distributions for the routing-block items were similar between the MST and Control samples, yet different for items in the second-stage blocks (especially at short RTs). Thus, we combined the two samples and produced one set of results (i.e., RT distributions, conditional P+, and average RTs) for items in the routing blocks, yielding only one set of RT thresholds for either the VITP approach or the normative threshold method for each item in the routing blocks. Conversely, for each of the Easy, Medium, and Hard blocks at Stage 2, the results of RT distributions, conditional P+, and average RTs were examined separately for the two student samples and then compared. It turns out that for items in the Easy and Medium blocks, the plausible RT thresholds were quite close between the Control and MST samples. For students receiving the Easy block, those in the Control sample tended to provide clearer cutoffs than did those in the MST sample (i.e., MST_Low), a result anticipated because the contrast between the chance level and the conditional P + above the threshold was more distinct for the Control sample than for MST_Low. The same phenomenon was observed for the Medium block. For all items in the Hard block, the RT distributions for MST_High had a much shorter left tail than did the RT distributions for the Control sample; the extended left tail for the Control sample also had chance-level conditional P+. Thus, for both threshold-identification methods, we only defined one set of RT thresholds for items in the second-stage blocks, mostly based on the RT distribution and P + information for the Control sample.

In the end, under the VITP approach, we followed a 5-, 10-, and 20-second rule to define thresholds for the multiple-choice items. Table 2 provides a summary. Among the 74 multiple-choice items, 64 had an RT threshold identified, and about 42 (i.e., two thirds of the 64 items) had a bimodal RT distribution as in Figure 1. Ten of the 74 items did not have an RT threshold identified based on the VITP approach, suggesting no clear rapid-guessing behavior observed on those items. A 10-second threshold was found adequate for most of the items. In terms of item context and reading load, the two items with a 5-second threshold had a very short stem and the calculation was straightforward—e.g., solving a linear

equation or polynomial given the value of the unknown variable. Compared to items with a 10-second threshold, the items with a 20-second threshold tended to include tables and/or figures, involve more text and more complex stems, or require the students to compare the options with the materials provided in the stems.

For comparison, Table 2 also shows the thresholds identified by the normative threshold method. For some of the RT distributions in the MCBS, the 10-second ceiling appeared too conservative, as the gap that clearly separated the bimodal distributions appeared beyond 10 seconds. In different assessments in which the items have varying time demands, the 10-second ceiling may need to be modified to identify rapid-guessing behavior in an effective manner while maintaining the accurate meaning of rapid guessing. In addition, this method assigned a threshold automatically to every item, even if there was no evident rapid-guessing behavior on some items according to the timing and response data.

Classifying behaviors

Next, for items with an RT threshold identified, the SB index was computed for each student-item pair. Among the student-item pairs with non-missing RTs, 98.9% were identified as exhibiting solution behavior for the VITP approach and 99.7% were identified for the normative threshold method. Both percentages are much higher than the percentages of solution behavior reported for other low-stakes assessments (e.g., 94.2% in Wise and DeMars 2006; 89.7% in Wise et al. 2009). Recall that this study concerns general rapid-guessing behavior, which can result from disengagement with the test/items and test speededness. Although long RTs are usually connected with high engagement with the items, they may also be a result of low engagement with the items as well as distraction by unrelated activities—which is less likely to be identified by the behavior-classification methods based on RTs and responses. Thus, the percentage of low engagement with the test/items with short RTs cannot exceed the percentage of rapid-guessing behavior.

Validity checks

We first reviewed the actual items and related the identified thresholds to the reading load and complexity of individual items with respect to the mode and median of the RT distributions and whether there were additional materials (tables or figures) to process (see the discussion in "defining RT thresholds"). For the second validity check, we evaluated the overall response accuracy rate conditional on the SB index: For the VITP approach, the P + for SB = 1 and the P + for SB = 0 were equal to 0.576 and 0.207, respectively−-the P + under solution behavior was almost three times the P + under rapid-guessing behavior, with the latter close to the chance level. The values for the normative threshold method were 0.568 for SB = 1 and 0.176 for SB = 0, both of which were smaller than the respective value under the VITP approach and the value for SB = 0 was farther away from 0.2.

The third validity check was accomplished by correlating students' overall scores (i.e., the EAP scores) with their performance on each item for students with SB = 1 and those with SB = 0. Figure 5 shows the results for the same item displayed in Figure 4. About 96.4% and 99.3% of the students taking this item had SB = 1 for the VITP approach and the normative threshold method, respectively. For students judged to be engaged in

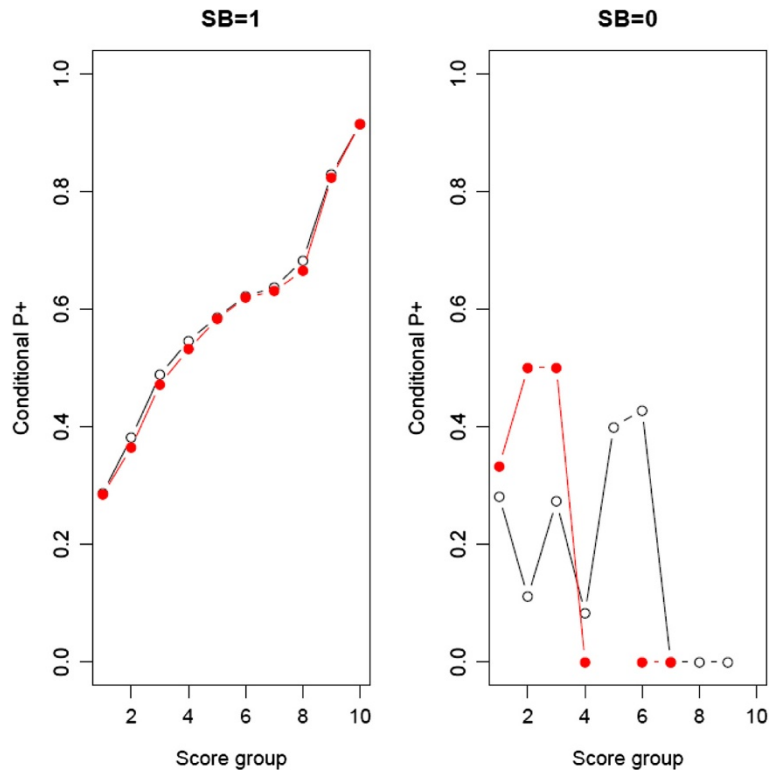

Figure 5 Relationship between students' EAP scores (score group) and their performance on a selected item (conditional P+). The connected filled circles are for the normative threshold method, and the connected open circles are for the VITP approach.

solution behavior (SB = 1) on this item, there was positive correlation between their overall scores (score group) and their performance on this item (conditional P+). However, for students judged to be disengaged in solution behavior (SB = 0) on this item, there was no relationship observed. Such finding was consistent for all items, except for some items with very few observations with SB = 0. For SB = 1, the P + conditioning on EAP scores for the normative threshold method was slightly smaller than that for VITP for several score groups. This finding, together with those from the other validity checks and the overall percent of SB = 1, suggested that the normative threshold method identified more responses as representing solution behavior than did the VITP approach, but those responses are more likely to result from rapid-guessing behavior. The results are not surprising given that the normative threshold method tended to produce smaller RT thresholds than did the VITP approach in the MCBS. Thus, the VITP approach was deemed to outperform the normative threshold method in our study. The validity checks also offered compelling evidence that the RT thresholds given by the VITP approach were effective in separating rapid-guessing behavior from solution behavior in the MCBS. It is hence meaningful to further discuss the results of behavior classification based on the VITP approach at different levels as follows.

Response-time effort (RTE) for students

The RTE index for a student was defined as the average of the SB index across items taken by the student. Among the 8,401 students, about 85% had an RTE value equal to

1, which means those students were classified as showing solution behavior consistently throughout the test. Roughly 14% of the students were found to engage in solution behavior for at least 80% (but less than 100%) of the items they received. There were merely 0.9% of the students with RTE smaller than 0.8; the minimal RTE was equal to 0 (this student, who was in the MST group, spent 1.7 minutes on the routing block and 2.3 minutes on the Easy block). The reliability estimate of the RTE index was roughly 0.76, an acceptable level of reliability for the classification of behaviors.

Results of all three measures of association indicated a very weak positive correlation between RTE and the responses to either survey questions listed in the Methods section. For Question 1, the estimated Pearson correlation, Spearman's rho, and Goodman and Kruskal's gamma were equal to 0.082, 0.076, and 0.161, respectively. For Question 2, the corresponding results were 0.006, 0.026, and 0.061. For a strong positive correlation to be observed for either question given that the RTE values were so skewed to the right, the marginal distribution of the responses should also have a clear decreasing trend from 4 (very important/tried much harder) to 1 (not very important/tried not as hard), possibly with a distinct proportion of students choosing 4. However, it is not the case in the MCBS: For Question 1, the percentages were 27.9% for "very important", 38.4% for "important", 26.8% for "somewhat important", and 6.3% for "not very important". For Question 2, the percentages were 6.4% for "tried much harder", 16.8% for "tried harder", 57.7% for "tried about as hard", and 18.6% for "tried not as hard".

In addition, we examined the distribution of the responses to either background question in each RTE category. Figure 6 presents the clustered bar chart for Question 1. In the figure, the horizontal axis shows the three RTE categories, and the vertical axis

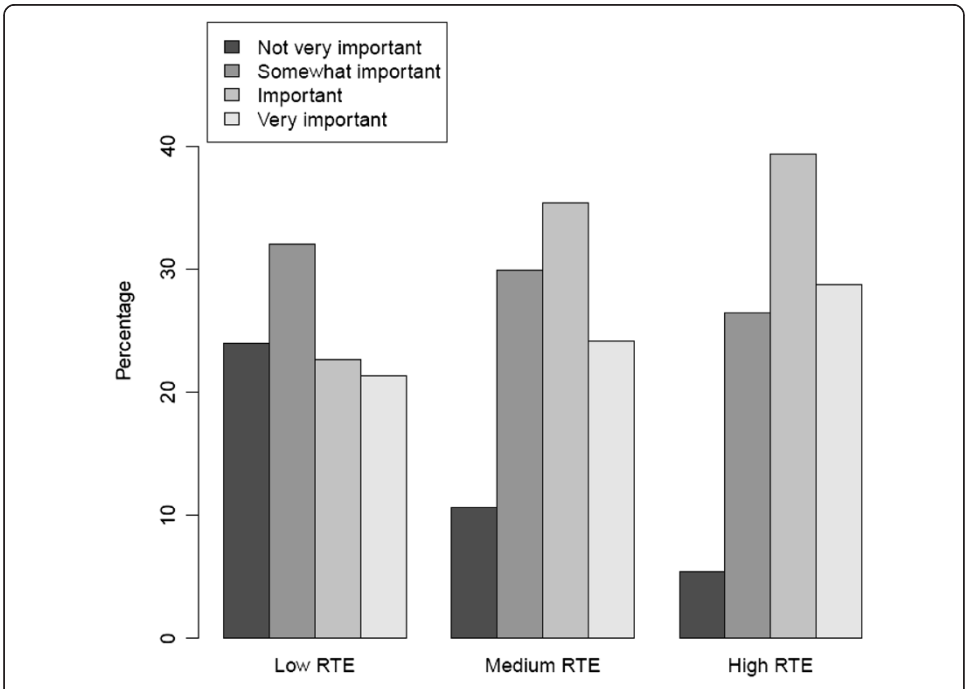

Figure 6 Clustered bar charts for the background question "How important was it to you to do well on this test?" (conditional on RTE levels).

shows the percentage of each answer in a given RTE category. It is clear that the fraction of students answering "not very important" decreased as RTE increased, while the fraction for "very important" was higher for higher RTE. The conditional patterns supported the low overall correlation between RTE and the background questions on engagement. Because the same patterns were observed for Question 2, the plot is omitted. In summary, the RTE results were reliable to an acceptable extent, while the RTE results and the students' self reports disagreed. Whether the disagreement related to how the students interpreted the background questions or related to unidentified low engagement with long RTs cannot be easily answered based on response and timing data (this issue is discussed in the end of the paper).

The overall Pearson correlation between RTE and EAP score was equal to 0.24. This is because high EAP scores corresponded to high RTE values, but conversely, low EAP scores associated with a wide range of RTE values. This result nicely coincides with Wise and DeMars' (2006) finding in a different educational assessment. Figure 7 shows the scatter plots of RTE and EAP score conditioning on student sample and performance level. Different symbols were used for the observations for students assigned the Easy block (circle), the Medium block (triangle), and the Hard block (plus). Due to the experimental design, the plots for the MST sample included only one symbol for each performance level; whereas each of the plots for the Control sample involved three symbols. Generally, performance level and RTE were closely related within each student sample. MST_High (6) and Control_High (3) had high scores and high RTE, even though some students in Control_High were given easier (less challenging) blocks. The RTE values for MST_Medium (5) and Control_Medium (2) were quite variable, and only about one third of them seemed to be running short of time (i.e., using more

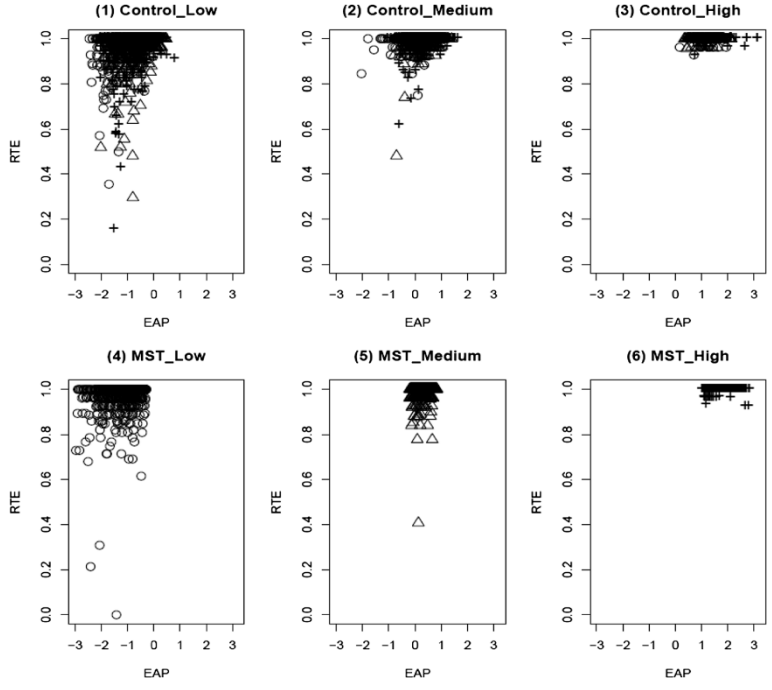

Figure 7 Scatter plot of score and RTE by student sample and performance level (different symbols for different second-stage blocks: circle for the Easy block, triangle for the Medium block, and plus for the Hard block).

than 20 out of 25 minutes on both stages). Some students in these subgroups probably were less proficient but highly motivated, and the others were somewhat disengaged on part of or the entire test. In addition, there were more observations with low RTE (e.g., less than 0.8) in Control_Low (1) than in MST_Low (4). Recall that the Control_Low students were randomly routed to the Easy, Medium, or the Hard block. Roughly 1.5% of those taking the Easy block had low RTE, but 2.6% (2.8%) of those taking the Medium (Hard) block had low RTE. It appears that low performing students in the routers were less likely to show rapid-guessing behavior when routed to the Easy second block than when routed to the Medium or Hard block. This is evidence that the multistage testing procedure may reduce the presence of rapid-guessing behavior (including disengagement with the test and test speededness) in low-stakes assessments.

Response-time fidelity (RTF) for items

The RTF index for an item was defined as the average of the SB index across students receiving the item, and its value ranged from 0 to 1. An item with higher RTF implies that more students showed solution behavior on that item. Recall that 64 of the 74 items had an RT threshold identified. Figure 8 shows the item RTF, arranged by item position in each block. Among the total of 64 items, only two had an RTF smaller than 0.94 (i.e., about 6% of the students taking either item exhibited rapid-guess behavior), ten had an RTF lying between 0.96 and 0.98, and the rest had an RTF greater than 0.98. Both items with RTF less than 0.94 had a 20-second RT threshold. The figure reveals a clear pattern that the items appearing later in each block tended to evoke more short RTs (Pearson correlation of -0.43), an outcome also observed in other educational assessments (e.g., Setzer et al. 2013; Wise et al. 2009). The average RTF for Router 1 or Router 2 was higher than the average RTF for any of the second-stage blocks.

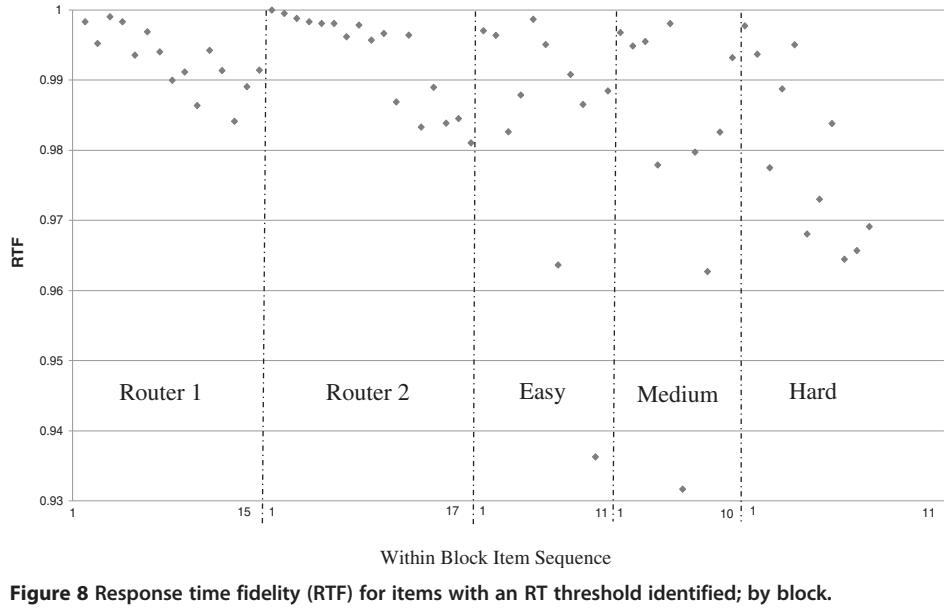

Figure 8 Response time fidelity (RTF) for items with an RT threshold identified; by block.

Next, we examined whether test-taking behaviors related to the items' content area. The box plot of item RTF by content area (Figure 9) suggests that none of the content areas evoked especially low RTF. Conversely, there was no strong correlation between item discrimination and RTF (see Figure 10; Pearson correlation of -0.21), or between item difficulty and RTF (see Figure 11; Pearson correlation of -0.29). The parameter estimates were obtained with filtering.

Conditional RTF by Students' Performance Level

The RTF discussed above was computed based on all students and represents the overall trends. To see if any item had particularly low RTF for some or all subgroups, we further evaluated conditional RTF by student's performance level. Figure 12 shows the results for the Medium block items. The RTF values for the low performing groups were consistently lower than those for the higher performing groups, an anticipated outcome due to the confounding between performance and test-taking behaviors. The RTF values for Control_High were always above 0.98, with slightly lower RTF for the latter half of the items. Conversely, Control_Low had noticeably variable RTF values across items, with the lowest being 0.82. The curves for MST_Medium and Control_Medium had similar patterns as the curve for Control_Low, but the magnitude was relatively stable. As with the marginal RTF, the same pattern of particular items evoking more rapid guesses was also observed in the conditional RTF, especially in the low performing groups. An example of items showing low RTF was item 12, which is probably not an item position effect as item 16 in the same block had a higher conditional RTF than did item 12 for all four groups. The estimated discrimination parameter for item 12 was highest among all the items (estimated a parameter = 2.49), with medium difficulty (estimated b parameter = 0.08). There were two more items (items 8 and 14) in

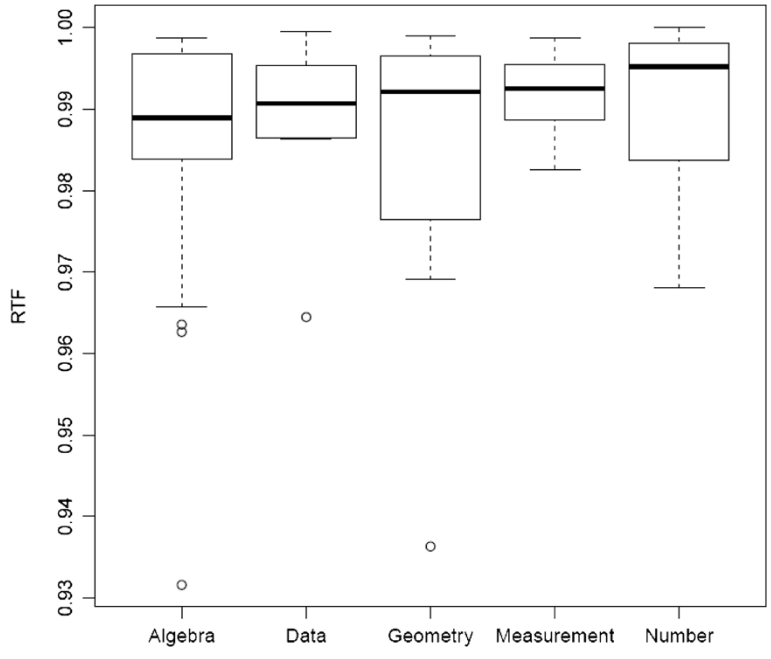

Figure 9 Box plot of item RTF by content area.

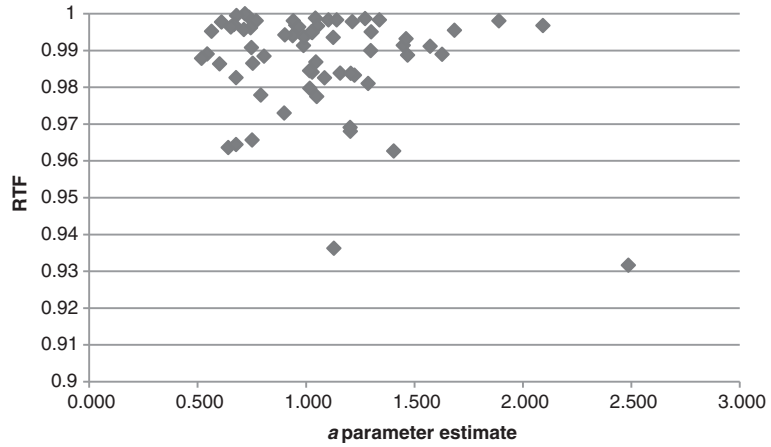

Figure 10 Plot of response-time fidelity (RTF) against IRT discrimination (*a*) parameter estimates (items with an RT threshold identified only).

the Medium block with RTF below 0.94 for the low performing group; their parameter estimates were 0.79 and 1.40 for the discrimination parameter and 0.35 and 0.54 for the difficulty parameter, respectively. The low RTF for those items may be attributable to the inclusion of tables, figures, or more complex stems.

Effects of data filtering based on the results of behavior classification

Given that only 1.1% of the responses in the entire data set were identified with rapid-guessing behavior, excluding those rapid guesses from IRT modeling should not have a substantial impact on item calibration. This expectation was confirmed. The estimated discrimination and guessing parameters were almost identical with and without removing the rapid guesses from IRT modeling. The difference between the estimated difficulty parameters was less than 0.06 for all items but one. Although the resulting differences in item parameter estimates were not substantial in magnitude, the data filtering appeared to

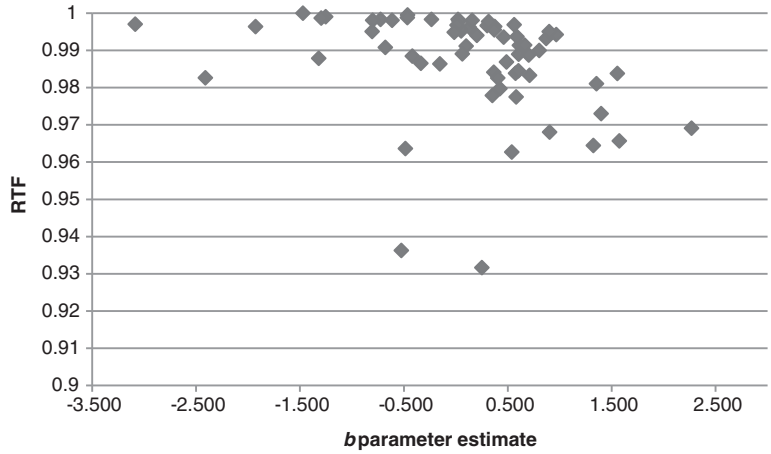

Figure 11 Plot of response-time fidelity (RTF) against IRT difficulty (*b*) parameter estimates (items with an RT threshold identified only).

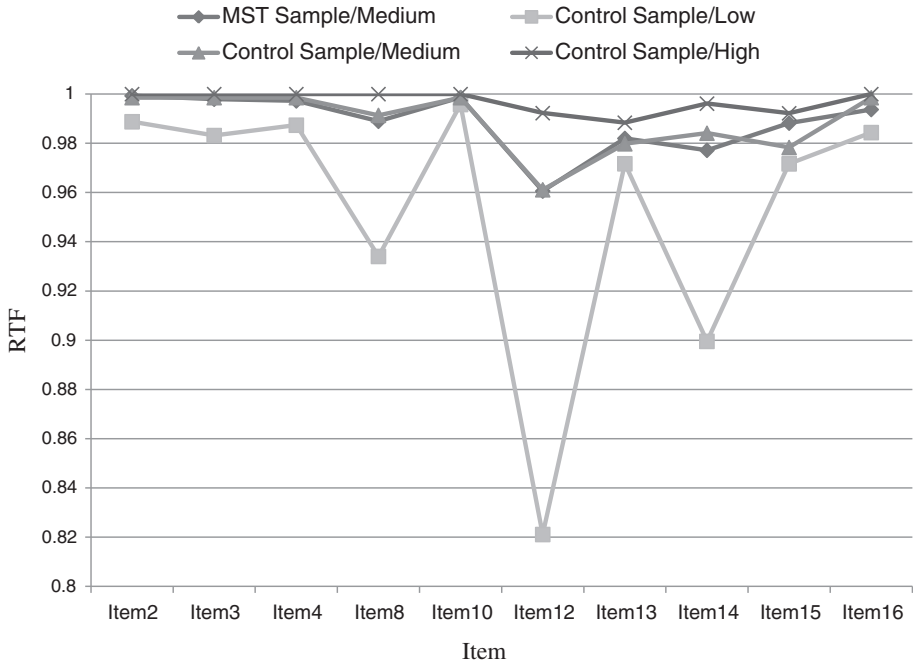

Figure 12 Response-time fidelity (RTF) for items in the Medium block, by performance group (items with an RT threshold identified only).

have the greatest impact on the difficulty parameter estimates for items in the Easy block among the five blocks in the MCBS (see Wise and DeMars 2006, for similar findings with a different test format).

Conclusions and discussion

This paper presents a five-step procedure to examine students' test-taking behaviors using RTs collected from a standard NAEP assessment setting. Instead of relying on students' self evaluation or additional costs in test design, the paper provides a way to address the issue of rapid-guessing behavior, including the long-term question about NAEP students' extent of engagement, and the impact. Further, establishing the analysis procedure based on the NAEP program data ensures that the procedure is readily applicable to future standard operational NAEP assessments when timing data become available. The procedure is also applicable to other tests with detailed timing data.

In the MCBS, the validity checks offered compelling evidence that the VITP approach was effective in separating rapid-guessing behavior from solution behavior and outperformed the normative threshold method. A very low percent of rapid-guessing behavior was identified in our data, as compared to existing results for different assessments. For this particular dataset, rapid-guessing behavior had minimum impact on parameter estimation in the IRT modeling. However, the students clearly exhibited different behaviors when there were discrepancies between their performance level and block difficulty. Especially, the students performing worse in the routers were less likely to show rapid-guessing behavior when routed to the Easy block than when routed to the Medium or the Hard block. We also found a disagreement between RTE and the students' self reports, but whether the

disagreement was related to how the students interpreted the background questions is not an easy question to answer based on the observed data.

Test-taking behaviors are tied to specific testing conditions. It is known that in low-stakes assessments, students from higher grades tend to exhibit more rapid-guessing behavior than those from lower grades (Ma et al. 2011; Wise et al. 2010), and mathematics items tend to solicit less rapid-guessing behavior compared to reading items (Wise et al. 2010). We expect to obtain similar findings in other NAEP mathematics assessments delivered to representative 8^{th} graders on computers with the same test settings, although there are limitations in generalizing the findings to NAEP assessments for different grades and subjects and to different assessments. The purpose of classifying behaviors is to identify responses associated with rapid-guessing behavior. Steps 4 and 5 of the procedure examine two possible applications of the classification results. One is to relate the three measures to student and item characteristics (step 4), which can inform test design and item development. For instance, items with low RTF may be reviewed and modified by test developers for future use. The other is to filter out responses associated with SB = 0 to mitigate the impact of rapid-guessing behavior on the estimation of IRT model parameters (step 5), which is important when the parameter estimates with and without filtering differ considerably. They are feasible applications for testing programs, including NAEP, to consider in practice.

In contrast to existing methods on this topic, the VITP approach is likely to show more advantages when more items have unclear bimodal patterns, primarily because incorporating response accuracy into the classification process can be quite informative. For testing programs with large item pools, implementing the VITP approach could be practically challenging, as the RT thresholds need to be identified item by item. One possible solution is to start with an item pool of feasible size with representative items to establish a baseline for threshold identification using the VITP approach, and then scale up to operational item pools using more automated methods.

There are other promising applications beyond the scope of discussion in this paper. For example, Setzer et al. (2013) investigated the variation in log-odds of solution behavior using a three-level random intercept hierarchical generalized linear model, and examined how RTF correlated with factors such as item difficulty, item position, etc. using regression. In the field of survey research, Couper and Kreuter (2013) modeled the RT of a question as a function of item characteristics, respondent characteristics, and interviewer characteristics.

Caution is needed in employing the procedure and interpreting the results. First, the results of behavior classification rely on the RT thresholds and are not meaningful unless the identified thresholds are validated. Second, the procedure is employed to identify general rapid-guessing behavior, not specific to rapid guessing due to disengagement with the test or due to test speededness. We did not further differentiate the two types of rapid guessing because of the very low percent of SB = 0, implying an even smaller percent due to either type. In applications in which it is necessary to distinguish one type from the other, one may examine the total section/test time of the students with lower RTE as a rough estimate of time pressure. Last but not least, the line of research we followed assumes that students with rapid-guessing behavior show very short RTs. Short RTs alone do not indicate rapid guesses because students with pre-knowledge on items may have short RTs but high accuracy. However, long RTs might result from no engagement with

the items and distraction by unrelated activities, rather than high engagement. Our analysis may be supplemented with process data gathered from log files to gain more comprehensive understanding about individual test-taking behaviors.

Endnotes

[a]We use the term "performance level" rather than "ability level" because the MCBS was low-stakes, and hence the students' performance on the MCBS does not necessarily reflect their true ability (depending on whether they give the best effort).

[b]The MCBS received OMB approval (OMB# 1850-0790 v.28).

[c]If a student never clicked on an item, the RT was missing. Our analysis was based on non-missing RTs.

[d]One may argue that defining RT thresholds based on visual inspection is subjective to some extent. Different raters may arrive at RT thresholds that have low exact agreement but are close (e.g., Ma et al. 2011). Our experience indicates that the patterns of conditional P + are typically quite different under the two behaviors–-staying consistently above the chance level under solution behavior versus fluctuating around the chance level under rapid-guessing behavior. So visually inspecting the RT distributions in conjunction with conditional P + information may not be as subjective as one might think. Conversely, as noted in the Background section, existing studies have shown little gain in classifying behaviors with operational data using more objective but sophisticated methods such as mixture models and non-parametric models (Kong et al. 2007; Ma et al. 2011). Thus, our approach is reasonable, although not fully objective, as a new application to NAEP data.

[e]The responses classified as rapid guesses were excluded from the estimation of the IRT item parameters.

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

Authors' contributions
YHL and YJ reviewed the literature, designed and conducted the analyses, and prepared the manuscript and revisions. Both authors read and approved the final manuscript.

Acknowledgments
The authors gratefully acknowledge the significant contributions of the NAEP Design and Analysis Committee, a subgroup of whom reviewed results from the studies and offered helpful recommendations in February 2011, October 2011, and June 2012.
Foremost, we would like to thank Ruopei Sun and Jonathan Guglielmon from the Center for Data Analysis Research at ETS for conducting a large portion of the analyses presented herein. In addition, this research benefited from comments from Andreas Oranje, Shelby Haberman, Matthias von Davier, Tim Moses, Joanna Gorin, and two anonymous referees.
This research was funded by the National Center for Education Statistics (NCES) within the Institute of Education Sciences (IES) of the U.S. Department of Education under Contract Award No. ED-07-CO-0107. The opinions expressed in this paper are solely those of the authors and not of ETS, NCES, or any of their affiliates.

Author details
[1]Educational Testing Service, MS 12T, Rosedale Road, Princeton, NJ 08541, USA. [2]Educational Testing Service, MS 02T, Rosedale Road, Princeton, NJ 08541, USA.

Received: 18 April 2014 Accepted: 13 August 2014
Published online: 17 September 2014

References

Bolt, DM, Cohen, AS, & Wollack, JA (2002). Item parameter estimation under conditions of test speededness: Application of a mixture Rasch model with ordinal constraints. *Journal of Educational Measurement, 39*,331–348.

Braun, H, Kirsch, I, & Yamamoto, K. (2011). An experimental study of the effects of monetary incentives on performance on the 12th-grade NAEP Reading assessment. *Teachers College Record, 113*(11), 2309–2344.

Couper, M, & Kreuter, F. (2013). Using paradata to explore item level response times in surveys. *Journal of the Royal Statistical Society, A, 176*(Part 1), 271–286.
Cronbach, LJ. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297–334.
DeMars, CE. (2007). Changes in rapid-guessing behavior over a series of assessments. *Educational Assessment, 12*(1), 23–45.
Hadadi, A, & Luecht, RM (1998). Some methods for detecting and understanding test speededness on timed multiple-choice tests. *Academic Medicine, 73*(10), 47–50.
Hauser, C, & Kingsbury, GG. (2009). *Individual score validity in a modest-stakes adaptive educational testing setting*. San Diego, CA: Paper presented at the annual meeting of the National Council on Measurement in Education.
Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education, 16*, 277–298.
Kong, X, Wise, SL, Harmes, JC, & Yang, S. (2006). *Motivational effects of praise in response-time-based feedback: A follow-up study of the effort-monitoring CBT*. San Francisco, CA: Paper presented at the Annual Meeting of the National Council on Measurement in Education.
Kong, X, Wise, SL, & Bhola, DS. (2007). Setting the response time threshold parameter to differentiate solution behavior from rapid-guessing behavior. *Educational and Psychological Measurement, 67*, 606–619.
Lee, Y-H, & Chen, H. (2011). A review of recent response-time analyses in educational testing. *Psychological Test and Assessment Modeling, 53*(3), 359–379.
Lee, Y-H, Lewis, C, & von Davier, AA. (2014). Monitoring the quality and security of multistage tests. In D Yan, AA von Davier, & C Lewis (Eds.), *Computerized multistage testing: theory and applications* (pp. 285–300). New York: CRC Press.
Ma, L, Wise, SL, Thum, YM, & Kingsbury, G. (2011). *Detecting response time threshold under the computer adaptive testing environment*. New Orleans, LA: Paper presented at the annual meeting of the National Council of Measurement in Education.
Meijer, RR, & Sijtsma, K. (2001). Methodology review: evaluating person fit. *Applied Psychological Measurement, 25*, 107–135.
Meyer, PJ. (2010). A mixture Rasch model with response time components. *Applied Psychological Measurement, 34*, 521–538.
National Assessment Governing Board. (2005). NAEP 12th grade participation and motivation: Preliminary recommendations. http://www.nagb.org/content/nagb/assets/documents/policies/NAEP%2012th%20Grade%20Participation%20%20Motivation.pdf. Last accessed 8 July 2014.
O'Neil, HF, Sugrue, B, & Baker, EL. (1995). Effects of motivational interventions on the National Assessment of Educational Progress Mathematics performance. *Educational Assessment, 3*(2), 135–157.
Schnipke, DL, & Scrams, DJ. (1997). Modeling item response times with a two-state mixture model: a new method of measuring speededness. *Journal of Educational Measurement, 34*, 213–232.
Schnipke, DL, & Scrams, DJ. (2002). Exploring issues of examinee behavior: insights gained from response-time analyses. In CN Mills, M Potenza, JJ Fremer, & W Ward (Eds.), *Computer-based testing: building the foundation for future assessments* (pp. 237–266). Hillsdale, NJ: Lawrence Erlbaum Associates.
Setzer, JC, Wise, S, van den Heuvel, J, & Ling, G. (2013). An investigation of examinee test-taking effort on a large-scale assessment. *Applied Measurement in Education, 26*(1), 34–49.
van der Linden, WJ. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika, 72*, 287–308.
Wise, S. (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, & DeMars, C. (2006). An application of item response time: the effort-moderated IRT model. *Journal of Educational Measurement, 43*(1), 19–38.
Wise, S, & 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, SL, & Ma, L. (2012). *Setting response time thresholds for a CAT item pool: the normative threshold method*. Vancouver, Canada: Paper presented at the annual meeting of the National Council on Measurement in Education.
Wise, S, Bhola, DS, & Yang, S. (2006a). Taking the time to improve the validity of low-stakes tests: the effort-monitoring CBT. *Educational Measurement: Issues and Practice, 25*(2), 21–30.
Wise, V, Wise, S, & Bhola, D. (2006b). The generalizability of motivation filtering in improving test score validity. *Educational Assessment, 11*(1), 65–83.
Wise, S, Pastor, DA, & Kong, X. (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, SL, Ma, L, Kingsbury, GG, & Houser, C. (2010). *An investigation of the relationship between time of testing and test-taking effort*. Denver, CO: Paper presented at the annual meeting of the National Council on Measurement in Education.
Xu, X, Oranje, A, Mazzeo, J, & Kulick, E. (2012). *An adaptive approach for group-score assessments*. Vancouver, British Columbia, Canada: Paper presented at the annual meeting of the National Council on Measurement in Education.
Yamamoto, K, & Everson, H (1997). Modeling the effects of test length and test time on parameter estimation using the hybrid model. In J. Rost (Ed.), *Applications of latent trait and latent class models in the social sciences* (pp. 89–98). Münster, Germany: Waxmann.

doi:10.1186/s40536-014-0008-1
Cite this article as: Lee and Jia: **Using response time to investigate students' test-taking behaviors in a NAEP computer-based study.** *Large-scale Assessments in Education* 2014 **2**:8.