Educational Measurement
ISSUES AND PRACTICE

NCME national council on measurement in education

Educational Measurement: Issues and Practice
Winter 2017, Vol. 36, No. 4, pp. 52–61

Rapid-Guessing Behavior: Its Identification, Interpretation, and Implications

Steven L. Wise, *NWEA*

The rise of computer-based testing has brought with it the capability to measure more aspects of a test event than simply the answers selected or constructed by the test taker. One behavior that has drawn much research interest is the time test takers spend responding to individual multiple-choice items. In particular, very short response time—termed rapid guessing—has been shown to indicate disengaged test taking, regardless whether it occurs in high-stakes or low-stakes testing contexts. This article examines rapid-guessing behavior—its theoretical conceptualization and underlying assumptions, methods for identifying it, misconceptions regarding its dynamics, and the contextual requirements for its proper interpretation. It is argued that because it does not reflect what a test taker knows and can do, a rapid guess to an item represents a choice by the test taker to momentarily opt out of being measured. As a result, rapid guessing tends to negatively distort scores and thereby diminish validity. Therefore, because rapid guesses do not contribute to measurement, it makes little sense to include them in scoring.

Keywords: motivation, rapid guessing, score validity, test-taking engagement

For well over a century, the measurement of student achievement has been based on the administration of a set of items to a test taker, with the responses to those items combined to calculate the test taker's score. Although the response to a single item provides limited information about the test taker's achievement level, the collective information from a set of item responses can yield a reliable estimate of achievement whose precision increases with the number of items in the set. The validity of the resulting score, however, depends in part on the assumption that all of the item responses reflect what the test taker knows and can do. This requires that the test taker remains engaged throughout a test event, and applies his or her knowledge, skills, and abilities when responding to each item. This assumption of engaged test taking is fundamental to virtually all measurement models used in practice (Wise, 2015). Nevertheless, most test givers recognize that test takers do sometimes disengage from the test-taking process. When this happens, at least some of the item responses will not reflect what the test taker knows and can do, which threatens the validity of the resulting test score.

The rise of the computer-based test (CBT) has brought with it the capability to measure more aspects of a test event than simply the answers selected or constructed by the test taker. One variable that has drawn much research interest is the time that test takers spend responding to individual items. Of particular relevance to this article is research involving multiple-choice achievement items, which has found evidence of test takers giving responses whose response times were so short that the test takers could not have read and fully considered the item. This behavior, termed *rapid guessing*, has been shown to indicate disengaged test taking (Schnipke, 1995; Wise & Kong, 2005).

This idea that rapid-guessing behavior indicates disengaged test taking is important because it suggests that engagement can be assessed down to the level of individual item responses. Not only does such information promote a more fine-grained understanding of test-taking behavior than can be provided by person fit indices or self-reported engagement (Wise & Kong, 2005), but it can also be useful in assessing the validity of a variety of psychometric calculations. Because rapid guesses tend to be accurate at a rate markedly lower than the rate observed under engaged test taking, their presence in test data can distort estimates of test performance (Rios, Guo, Mao, & Liu, 2016; Wise, 2015; Wise & DeMars, 2006; Wise & Kingsbury, 2016), item parameters (Schnipke, 1995, 1999; Schnipke & Scrams, 2002), test speededness (Schnipke, 1995; Schnipke & Scrams, 1997), and test norms. Moreover, differences in rapid-guessing rates between test taker subgroups can lead to differential item functioning (DeMars & Wise, 2010) and biased comparisons of subgroup differences in test performance (Setzer, Wise, van den Heuvel, & Ling, 2013; Wise & DeMars, 2010). Hence, there has been a growing understanding of the psychometric consequences of rapid guessing in test data, and it is important that we be able to identify when, and to what degree, disengaged test taking occurs during test events.

After over 20 years of research on rapid-guessing behavior, however, a lack of consensus remains regarding the identification, interpretation, and implications of rapid guessing. In this article, I broadly examine rapid-guessing behavior—its theoretical conceptualization, underlying assumptions, and methods for identifying it. In addition, I present a theory of rapid guessing along with research evidence consistent with this theory. My general intent is to argue that rapid guesses

Steven L. Wise is a Senior Research Fellow, NWEA, 121 NW Everett St, Portland, OR 97209; steve.wise@nwea.org.

do not contribute to measurement, and that they should be excluded from scoring.

In the discussion that follows, several new empirical findings about characteristics and dynamics of rapid guessing will be presented and discussed. These findings are based on analyses of data from NWEA's MAP® Growth™ multiple-choice testing system, which administers computerized adaptive tests (CATs). MAP Growth assessments are not considered high-stakes, as test performance typically does not count toward a student's grade. Moreover, MAP Growth is administered using liberal time limits, yielding test events that are essentially unspeeded. In the current study, two data sets were used. The first, which I will call the Test Event Data Set, consisted of 415,437 MAP Growth test events in Reading (Grades 2–11), collected from a single U.S. state during the 2013–2014 school year. Each MAP Growth Reading item has four response options. The second data set, called the Item Data Set, focused on six items chosen arbitrarily from the MAP Growth Reading item pool (which contains several thousand items). The only selection requirement for these items was that each item had to have been administered during at least 15,000 test events.

Research Origins

Research on rapid-guessing behavior has two primary origins, each of which reflects an innovative way to deal with a particular measurement problem. The first point of origin came from Deborah Schnipke, who focused her dissertation research on the evaluation of test speededness in high-stakes tests (Schnipke, 1995). She observed that, as time was expiring during timed tests, some test takers would quickly fill in answers to remaining items, apparently in hopes of getting some correct by lucky guessing. Schnipke asserted that the test event should be considered speeded for these test takers, because they would have been unlikely to engage in such behavior if they had had enough time to read and fully consider each item. Such rapid-guessing behavior, however, was not recognized by traditional measures of speededness—which are based on the percentage of test takers completing a certain number of items within the time limit—because all items had been answered. Thus, Schnipke concluded, traditional indices will tend to underestimate the degree to which a test was speeded.

Schnipke (1995) hypothesized that item response time could provide a more effective way of assessing speededness. She suggested classifying each item response as indicating one of two types of response behaviors:

> In *rapid-guessing behavior*, the examinee responds rapidly to items as time expires; accuracy will be at or near chance because the examinee is not fully considering the items. The examinee may skim the item briefly for keywords, but the examinee does not completely read the item when engaging in rapid-guessing behavior. . . . In contrast, in *solution behavior*, the examinee actively tries to determine the correct answer to every item. The examinee reads each item carefully and fully considers the answer. Accuracy will depend on item difficulty and other item characteristics and on the examinee's ability. (p. 5)

Analyzing data from the CBT version of the Graduate Record Exam, Schnipke showed that the presence of rapid-guessing behavior for a particular test taker indicated that a test was speeded. This inference was supported by the finding that rapid-guessing behavior tended to be clustered toward the end of test events. In subsequent studies, she proposed a new method for measuring speededness (Schnipke & Scrams, 1997) and showed how rapid guessing can distort item parameter estimation (Schnipke, 1996, 1999; Schnipke & Scrams, 2002).

A second point of origin emerged several years later from my own efforts to address the problem of low test-taking effort on my university's general education outcomes assessments. These assessments, some of which were administered as CBTs, measured a variety of constructs such as information literacy or scientific reasoning. Testing occurred in special sessions outside of the normal classroom context, the tests were relatively unspeeded, and there were no performance-based consequences for test takers. Having read some of Schnipke's research, I hypothesized that unmotivated test takers might also exhibit rapid-guessing behavior. Analyses of the CBT data confirmed this, and we demonstrated that in this low-stakes testing context, rapid guessing indicated low test taker motivation[1] (Wise & Kong, 2005). Subsequent to this discovery, my colleagues and I developed (a) measures of both the effort exhibited by test takers (Wise & Kong, 2005) and received by items (Wise, 2006), (b) a method for adjusting scores for the effects of rapid guessing (Wise & DeMars, 2006), and (c) a CBT that could detect, in real time, when a test taker begins to exhibit rapid guessing, and attempt to preempt this behavior by displaying messages of encouragement or warning (Kong, Wise, Harmes, & Yang, 2006; Wise, Bhola, & Yang, 2006).

Thus, there have been two distinct lines of research on rapid-guessing behavior. One looked at test taker behavior on speeded, high-stakes tests, whereas the other focused on unspeeded, low-stakes tests. In the first case, rapid guessing reflects a strategic choice made by test takers who are motivated to try to maximize their test scores. In the second case, rapid guessing indicates unmotivated test takers who are not trying to maximize test performance. Hence, the same test taker behavior can have two very different antecedents and interpretations, a point discussed more fully in a later section.

Identifying Rapid Guessing

Test takers vary in the amount of time they spend responding to a given test item, due to individual differences on a variety of factors such as achievement level, reading speed, cognitive speed, or fatigue. Moreover, there are additional episodic factors that can influence response time (e.g., distracting noise). Collectively, these multiple influences render it difficult to explain clearly why one test taker took, say, 30 seconds to respond to the item, while another took 35 seconds. There is, however, a range of potential response times that is far less ambiguous. If a response occurs rapidly—by which I mean much faster than the time required by a test taker to read, understand, and select a response—one can reasonably infer that the test taker did not interact with the item in an engaged fashion. Such disengaged responses reflect a separate response process from that used when the test taker is engaged, and do not provide information about the test taker's achievement level (Schnipke, 1995; Wise & Kingsbury, 2016).

Figure 1 depicts a hypothetical example of the respective response time distributions for a particular item under rapid-guessing and solution behaviors. In this example, the solution behavior response times follow a positively skewed

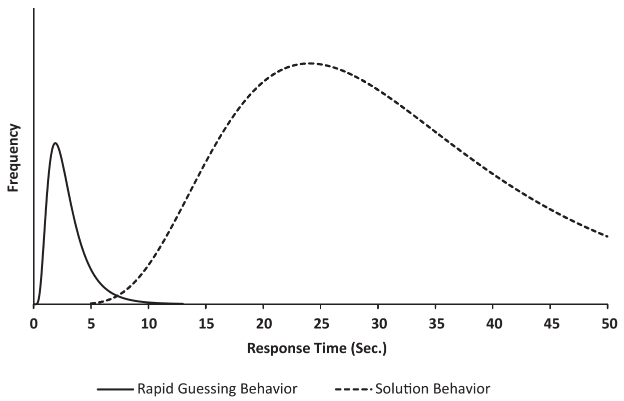

FIGURE 1. Conceptual distributions of response times for rapid-guessing and solution behaviors.

distribution, with the shortest response times beginning at around 5 seconds. This indicates that engaged test takers require, at a minimum, at least 5 seconds to read, understand, and answer the item. In contrast, the rapid-guessing response times follow a distribution in which most of the responses occur within 5 seconds, but a small percentage exceeds 5 (or even 10) seconds. Under this conceptualization, if the item receives rapid guesses from some test takers and solution behaviors from others, the resulting overall distribution of response times will be bimodal, with one mode corresponding to the rapid-guessing frequency "spike" occurring during the first few seconds, and another occurring later at the modal response time for solution behavior. Bimodal frequency distributions of this type were observed by both Schnipke (1995) and Wise and Kong (2005). Presumably, for each administered item, a test taker makes either an engaged response (i.e., exhibits solution behavior) or a disengaged response (i.e., exhibits rapid guessing). That is, for each item response, response time will follow one of the two distributions in Figure 1.

Threshold Identification Methods

In practice, the differentiation of rapid-guessing behavior from solution behavior requires the establishment of a time threshold, which is used to classify individual item responses as reflecting one of the two response behaviors. However, as depicted in Figure 1, whenever the response time distributions overlap—as will typically be the case—classifications based on any threshold value will inevitably result in misclassifications. False positives will occur when effortful responses are classified as rapid guesses, with false negatives occurring when noneffortful responses are classified as solution behaviors. Reducing the number of false positives will increase the frequency of false negatives, and vice versa. This suggests that the test giver's judgments regarding the relative importance of the two types of classification errors should be an important consideration in choosing a time threshold.

Several methods have been proposed for choosing time thresholds for identifying rapid-guessing behavior. These methods have used a variety of information about item characteristics and the responses given to items. While these methods have usually been used singly, they have sometimes been used in combination.

Visual inspection of response time distributions. Schnipke (1995) relied on visually inspecting the distribution of observed response times to identify an item's threshold. In this method, one tries to identify a low-frequency "dip" in early part of the distribution, as it presumably indicates the point at which the rapid guessing and solution behavior distributions intersect. In Figure 1, for example, such a dip occurs at around 7 seconds. Many subsequent studies have used this method (DeMars, 2007; Pastor, Strickman, & Ong, 2015; Setzer & Allspach, 2007; Setzer et al., 2013; Wise et al., 2006; Wise & Cotten, 2009; Wise & DeMars, 2006, 2010; Wise, Pastor, & Kong, 2009). Note that because the visual inspection method focuses on the intersection between the two distributions, both false positive and false negative classifications will typically occur.

An advantage of the visual inspection method is that it is intuitively congruent with the theoretical conceptualization illustrated in Figure 1. In practice, however, the method can sometimes be problematic, because the observed response time distribution will not always be bimodal. This problem will occur with items that require relatively short amounts of time under solution behavior. For example, consider that in Figure 1 the modal solution behavior response time is about 25 seconds. For items with a much shorter modal response time (e.g., 10 seconds), the solution behavior distribution would lie so close to its rapid-guessing counterpart that the two distributions will largely overlap. In these types of instances, the response time distribution will be unimodal (or even J-shaped) and the visual inspection method will be challenging to apply.

Surface features. Wise and Kong (2005) based their thresholds on the number of characters in the item text, along with whether or not the item contained tables or figures. The rationale for this was that items requiring more reading should have longer time thresholds. This method, which was also used by Wise (2006) and Silm, Must, and Täht (2013), does not explicitly reference the empirical response time distribution and therefore does not have clear implications regarding classification errors.

Common k-second thresholds. Some research has used a common time threshold (usually 3–5 seconds) for all items (Bowe, Wise, & Kingsbury, 2011; Wise, Kingsbury, Thomason, & Kong, 2004; Wise, Ma, Kingsbury, & Hauser, 2010). This is the simplest threshold method to use, as it does not require information about each item's surface features or response time distributions, and it is particularly useful with large item pools used with adaptive tests. Its one-size-fits-all nature, however, will often produce variation in classification errors across items. Consider, for example, a multiple-choice test in reading that uses a common 3-second threshold. For simple vocabulary items, many test takers may respond effortfully in less than 3 seconds, yielding a preponderance of false positive classifications. Conversely, for items involving reading passages, noneffortful responses may occur after the threshold, yielding a preponderance of false negatives. Overall, the relative frequencies of the two types of errors may vary substantially across items, which is a disadvantage of this approach.

Mixture models. Several studies have used mixture modeling applied to response times to represent rapid-guessing and solution behavior distributions (Bovaird, 2002; Meyer, 2010; Pastor et al., 2015; Schnipke, 1999). These methods choose time thresholds corresponding to the estimated point of intersection between the distributions. As with the visual inspection approaches, these thresholds will yield both false positive and false negative classifications.

Combining response time and accuracy. A newer set of threshold identification methods has emerged that incorporates response time with the accuracy of rapid guesses. These methods are based on the idea that rapid guesses tend to be markedly less accurate than solution behaviors. Wise and Ma (2012) examined the accuracy of responses from a CAT in which solution behaviors were correct about 50% of the time, and rapid guesses were correct at a lower rate close to the rate expected under random responding. Wise and Ma computed accuracy rates across a series of successive time intervals, finding that at a certain point the accuracy would begin to markedly increase from a baseline rapid-guessing rate. The point of increase indicates the earliest time point at which both solution behaviors and rapid guesses had begun to occur (e.g., in Figure 1 this point would begin at 5 seconds). This logic underlies Wise and Ma's (2012) normative threshold method. Wise and Ma found, moreover, that the point of increase could be reasonably approximated using a threshold value equal to 10% of the mean response time given to the item (with a maximum threshold value of 10 seconds). An advantage of this approximation-based normative threshold method is that it can be readily applied to the larger sized item banks typically used with CATs, because its computations would require only the mean response times given to each item.

Several additional threshold identification methods that combine response time with accuracy have been proposed. These methods look for the point at which accuracy exceeds what would be expected from random responding (Goldhammer, Martens, Christoph, & Lüdtke, 2016; Guo et al., 2016; Lee & Jia, 2014), which is typically defined for an item as the reciprocal of the number of response options. Each of the threshold methods combining response time and accuracy uses logic that results in minimal false positive classifications. The primary disadvantage of these methods, however, is that substantial response data per item is required to accurately detect the increase in accuracy needed to identify the threshold.

An example of the normative threshold method. To provide a practical illustration regarding how information about response time and accuracy can be combined in identifying thresholds, the normative threshold method was applied to two items from the Item Data Set. Although there were over 15,000 responses per item in this data set, relatively few responses with short response times were present, resulting in accuracy rates that were not very stable at these response times. To address this problem, moving average accuracy was calculated across overlapping 3-second intervals, which smoothed the trend across time. For example, the value at 5 seconds equaled the weighted average of the values at 4, 5, and 6 seconds.

The moving average trends for Items 1 and 2 are shown in Figure 2. For Item 1, accuracy remained fairly constant until 10 seconds, after which there was a consistent increase. This suggests that the threshold should be set at 10 seconds because, beyond that time, more accurate engaged responses had begun to appear. The graph for Item 1 also shows the normative threshold method value (9.40 seconds), which represents 10% of the mean response time for the item. For Item 2, the accuracy increase began sooner. After 7 seconds, accuracy showed a consistent increase, which was closely approximated by the 10% normative threshold value of 6.74 seconds.

Summary. Multiple threshold identification methods for identifying rapid guessing have been proposed. Although comparisons among methods have shown that they yield similar results (Kong, Wise, & Bhola, 2007), there are some clear differences among the methods regarding the relative occurrence of false positive and false negative classification errors.

Interpreting and Understanding Rapid-Guessing Behavior

Once we have identified rapid-guessing behavior, it is important that we can properly interpret its meaning. As described earlier, rapid guessing can result from two different test-taking scenarios—as strategic behavior from motivated test takers running out of time during high-stakes tests and as unmotivated test taking during low-stakes tests. There is, however, a third potential antecedent condition for rapid guessing. Whenever test takers can immediately recognize that they do not have the requisite knowledge, skills, or abilities to address the challenge posed by a particular item, they may also respond with a rapid guess. For example, if a trigonometry test was administered to a third grader who has had no instruction in trigonometry, after a few items the student would realize that they have no idea what cosines are. Any subsequent items that involve cosines might rationally elicit a rapid guess from the student, because they could quickly recognize they could not solve the problem.

This third scenario poses a problem for the interpretation of rapid-guessing behavior. We would like to interpret rapid guesses as uninformative regarding what the test taker knows and can do. But in the third scenario, a rapid guess *is* informative (i.e., the student does not understand cosines). Hence, the inference that rapid guesses are uninformative depends on the assumption that the third scenario has not occurred. This means that we assume the test takers under consideration are administered items for which they are capable of effortfully engaging.

If the above assumption is tenable, interpretation of rapid guessing appears straightforward: in high-stakes tests, rapid guesses represent strategic attempts to maximize one's score, whereas in low-stakes tests they represent unmotivated test taking. Although this is a generally accurate characterization, there are exceptions. For example, during a high-stakes test a test taker might decide to stop caring about the consequences of low performance and exhibit rapid guessing due to low motivation. Likewise, under low-stakes testing, an engaged test taker might begin to exhibit rapid guessing if the test event had a time limit that had nearly expired. Hence, while the context in which the test occurred suggests how rapid guessing should be interpreted, we may be less than totally confident in that interpretation. Despite the different testing contexts, however, it is important to note that the dynamics of rapid guessing appear to be largely the same in each. Regardless of *why* it occurs, in rapid guessing the test taker is not effortfully engaged with a test item. If our goal is simply to identify disengaged responses, it may not be necessary that we understand the reason for each occurrence.

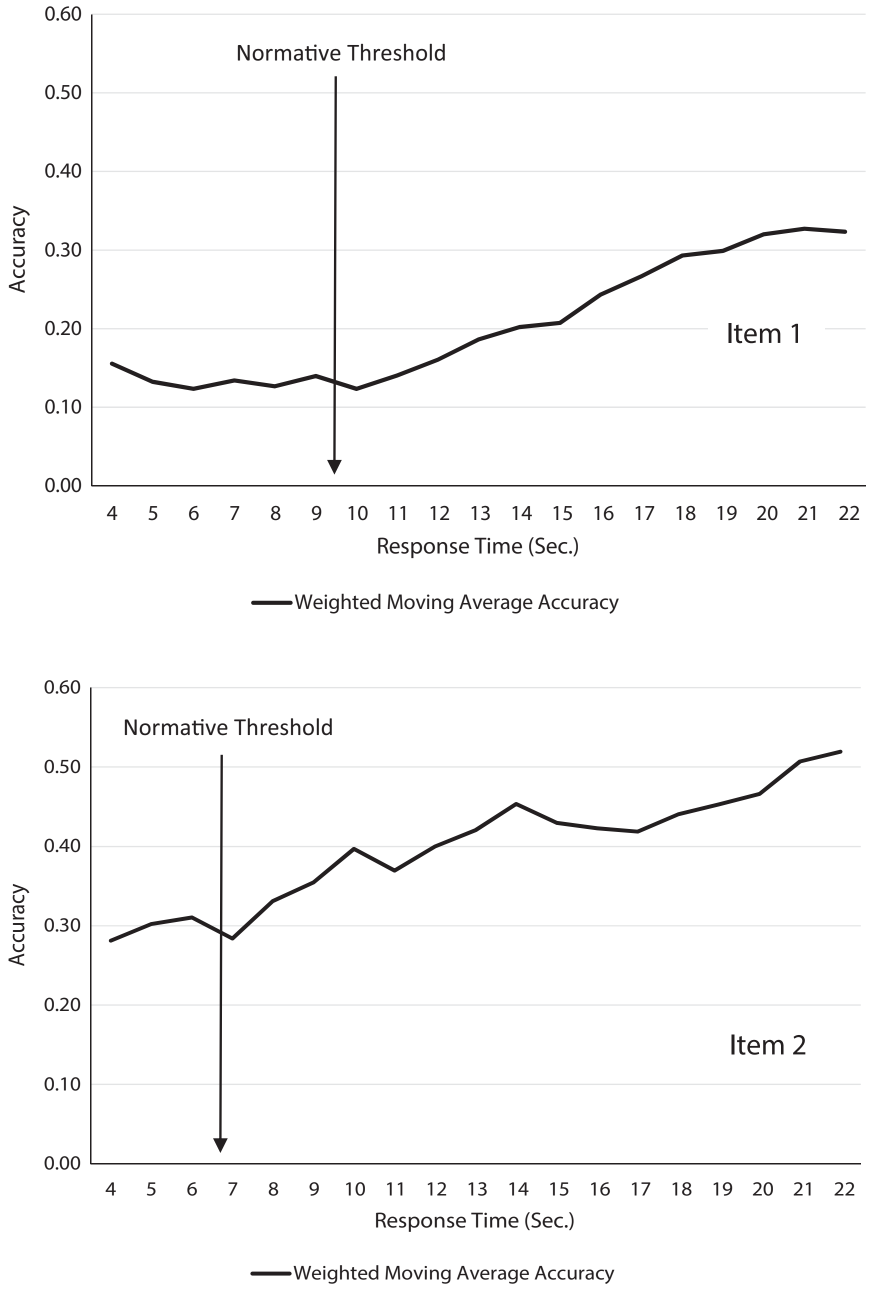

FIGURE 2. Weighted moving averages of response accuracy across response times for two MAP Growth Reading items, along with their normative threshold values.

A Theory of Rapid-Guessing Behavior

The term "rapid guessing" is a convenient shorthand term for the test-taking behavior addressed in this article. But it does not explain what happens during the behavior, and its implications for educational measurement. In this section, I present a theory of what I believe the test taker does during rapid-guessing behavior. Then, in the following section, I will present evidence consistent with the theory.

It is useful to consider a multiple-choice test event as a series of discrete encounters between a test taker and a test item such that, at any point in the test, a presented item must receive a response before the test taker can proceed to the next item. Shortly after an item is presented, the test taker makes a choice between solution behavior or rapid guessing. If solution behavior is chosen, the test taker will then apply his or her knowledge, skills, and abilities to attempt to identify the correct response option for the item. In contrast, if rapid guessing is chosen, the test taker will then rapidly make a second choice regarding which response option to select. This second choice will not reflect an effortful attempt by the test taker to identify the correct option, nor will it be a random response. After the selected option has been submitted by the test taker, the next item is presented and the decision process repeats.

What determines the first choice—whether to engage in solution behavior or rapid guessing? Wise and Smith (2011) introduced their demands-capacity model of test-taking effort. In the Wise-Smith model, items vary in their *resource demands*—how much reading they require, how mentally taxing they are to complete, and how difficult they are.

Likewise, test takers vary in their *effort capacity*, which represents the amount of effort a test taker is currently willing, or able, to devote to solution behavior. Effort capacity is affected by myriad factors such as test stakes, time pressure, fatigue from answering earlier items, how interesting earlier items were, or a desire to please teachers or parents. Under the demands-capacity model, whenever an item is presented the test taker compares his or her effort capacity to the resource demands of the item. If effort capacity is high enough, the test taker will engage in solution behavior; otherwise, rapid guessing will occur. Because both the item resource demands and test taker effort capacity may vary unsystematically across items, sequences of rapid-guessing behavior and solution behavior during a test event can appear idiosyncratic.

The idea that a test taker chooses between two fundamentally diverse ways of responding to an item is also grounded in dual-process theory, which views cognitive processes as being divided into two systems (Kahneman, 2011). Stanovich and West (2000) originally proposed the terms *System 1* and *System 2* and described the distinctions between the two. System 1 can be described as relatively fast, low-effort, intuitive, and relatively undemanding of cognitive capacity. System 2, in contrast, is relatively slow, effortful, involves analytic processing, and is demanding of cognitive capacity. In a test-taking context, rapid guessing appears to be a manifestation of System 1, while solution behavior reflects System 2. Hence, the Wise-Smith model can be viewed as a test taker's deciding whether or not to override the default System 1 by engaging in System 2, with the decision driven by the test taker's current effort capacity as compared to the resource demands of the item.

If a test taker chooses to give a rapid guess to an item (i.e., uses System 1), what determines the second choice regarding which response option to select? System 1 provides us with impressions and associations that we can use to respond intuitively to our environment. System 1 has learned associations between ideas, and mental activities can become fast and automatic through practice. When faced with an unfamiliar multiple-choice item, however, System 1 can do little more than make a choice among the response options based on the test taker's rapid assessment of the option most likely to be correct. In essence, rather than use System 2 to effortfully identify the correct answer to an item, the intuitive System 1 provides a rapid guess based on its *impression* of what appears to be the correct answer.

Empirical Evidence for the Test-Taking Theory

There is a variety of evidence that is consistent with the theory described above. Some of the evidence (described below) comes from prior research, and some comprises new results from the Test Event Data Set and the Item Data Set.

Rapid guessing is idiosyncratic. Several researchers have conceptualized disengaged test taking as following a state model, in which all test takers begin a test in an engaged state, with some shifting into a disengaged state at some point and remaining there for the remainder of the test event (Bolt, Cohen, & Wollack, 2002; Cao & Stokes, 2008; Jin & Wang, 2014; Yamamoto & Everson, 1995). This would suggest that rapid guessing ought to conform to a pattern of all solution behavior up to a point, and then all rapid guessing thereafter. This conceptualization, however, appears simplistic. Wise and Kingsbury (2016) found evidence that the occurrence of rapid guessing is more idiosyncratic. To illustrate this with the Test Event Data Set, I focused on test takers who had exhibited at least four rapid guesses. After the fourth rapid guess, only 40% of the subsequent item responses were rapid guesses, which is far less than the 100% that would be expected under a state model.

An idiosyncratic occurrence of rapid guessing, however, is not inconsistent with the Wise and Smith (2011) capacity-demands model. Because resource demands vary across items, and effort capacity varies both within and across test takers, the finding that a test taker might show an irregular pattern of rapid guessing on some items, but not others, is unsurprising.

Two distinct response behaviors. There are three types of evidence for the claim that there are two distinct response behaviors. First, the amount of time test takers spend on each behavior is markedly different. Second, solution behaviors and rapid guesses typically have very different accuracy rates (Schnipke, 1995; Wise, 2015). Third, the response behaviors exhibit clear differences in the amount of psychometric information they provide.

All three types of evidence can be demonstrated using the two MAP Growth data sets. Table 1 shows the mean response times for solution behavior and rapid guesses, across different levels of response time effort (RTE; Wise & Kong, 2005). RTE equals the proportion of administered items that were solution behaviors during a test event, and provides a useful index of a test taker's overall engagement. The mean response times for rapid guesses were under 5 seconds, while the values for solution behaviors averaged over a minute. However, it might be expected that the most disengaged test takers (i.e., RTE < .40) would also exhibit shorter solution behavior response times that more closely resembled those from rapid guessing. In fact, the reverse was found. Table 1 shows that the mean times for these test takers were the longest for

Table 1. Mean Response Time and Accuracy for Solution Behavior and Rapid Guessing From Test Events Exhibiting Various Intervals of Response Time Effort (RTE)

		Mean Response Time (Seconds)		Mean Accuracy[a]	
Group	*N*	Solution Behavior	Rapid Guessing	Solution Behavior	Rapid Guessing
RTE = 1.0	316,319	82.88	–	.50	–
.80 ≤ RTE < 1.0	90,998	67.17	4.82	.50	.25
.60 ≤ RTE < .80	11,248	65.67	3.02	.53	.26
.40 ≤ RTE < .60	2,769	75.83	2.10	.56	.27
RTE < .40	534	110.88	1.40	.55	.27

[a] Accuracy is defined as the proportion of items passed.

solution behaviors; the mean time spent during solution behavior was 79 times that spent during rapid guessing. These results indicate that test takers spend far more time on solution behaviors, even when a high proportion of rapid guesses are present.

The difference between the two response behaviors are further seen in their accuracy rates (also found in Table 1). The accuracy of solution behaviors remained near the .50 value expected based on the CAT item selection algorithm.[2] The mean accuracy of rapid guesses, in contrast, remained near the .25 value that would be expected under random guessing to four-choice multiple-choice items.

The most important difference between rapid guessing and solution behavior, however, lies in the relative amounts of psychometric information they provide. All psychometric models currently used in achievement and ability testing share the foundational assumption that the likelihood an item is passed is positively related to the test taker's standing on the latent construct the item measures. On a reading test, for example, it is assumed that the stronger a test taker is in reading, the more likely they are to pass a given reading item. It is because of this relationship that the test taker's observed score on the item (incorrect, correct) provides information about his or her standing on the measured construct. Such information, aggregated across a set of items, provides the basis for educational measurement.

Conversely, if the likelihood an item is passed is unrelated to the construct being measured, responses to the item do not provide psychometric information, and therefore do not contribute to measurement. Figure 3 shows the relationship between response accuracy and reading achievement, by response behavior, for the six items in the Item Data Set. For each item, the distribution of MAP Growth Reading scores of test takers who received the item was divided into five equivalent-sized strata (i.e., quintiles), and the accuracy rates of rapid guesses and solution behaviors were computed for each quintile. Each of the items shows a similar pattern. Solution behaviors exhibited a monotonic increase in accuracy across quintiles, consistent with the assumption that accuracy on the reading items was positively related to reading achievement. The patterns for rapid guesses, in

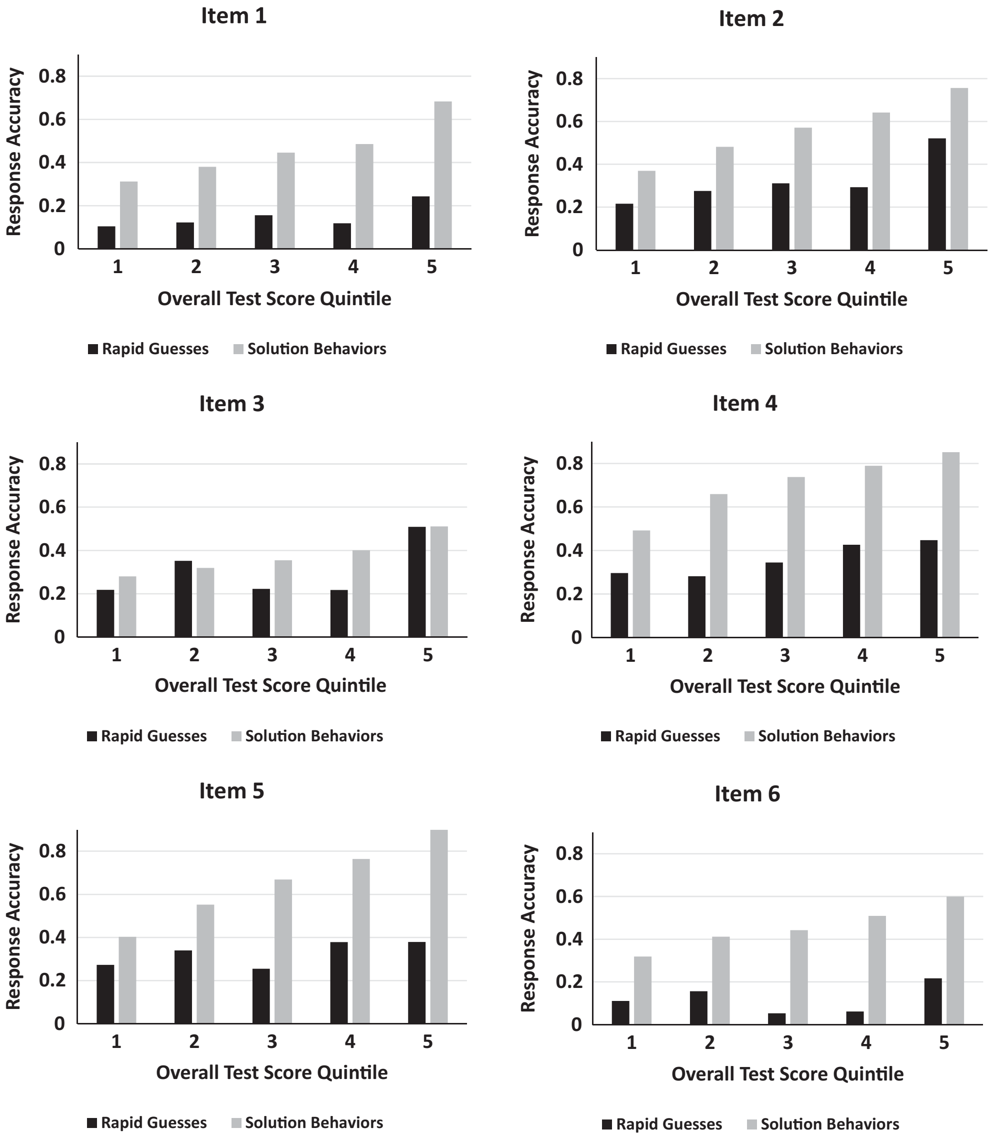

FIGURE 3. Accuracy rates of rapid guesses and solution behaviors, by overall test score quintile, for six MAP Growth Reading items.

Table 2. Distributions of Responses Under Solution and Rapid-Guessing Behavior to Six MAP Growth Reading Items

Item	Time Threshold (Seconds)	Percentage of Responses That Were Rapid Guesses	Percentage of Responses to Each Response Option							
			Solution Behavior				Rapid-Guessing Behavior			
			A	B	C	D	A	B	C	D
1	9.40	4	15	26	13	**46**	28	35	23	**14**
2	6.74	3	**56**	18	11	15	**29**	32	25	14
3	7.06	3	28	19	**37**	16	29	33	**26**	12
4	5.03	1	14	**70**	13	3	28	**34**	25	13
5	5.09	2	**64**	19	10	7	**31**	29	29	10
6	4.66	1	28	16	11	**45**	28	37	23	**12**

Note: The numbers of student responses for each item ranged from 15,092 to 19,878, and the correct option for each item is shaded. Time thresholds were calculated from the item pool data using the normative threshold method (10%).

contrast, showed little to no evidence of a positive relationship. Thus, while the accuracy of item responses under solution behaviors showed clear evidence of psychometric information, those under rapid guessing showed virtually no evidence.

Collectively, the three types of evidence from the MAP Growth Reading data support the claim that solution behaviors and rapid guesses represent two fundamentally different response behaviors. Responses under solution behavior took a sizable amount of time to complete, were correct at a rate expected for a CAT, and were psychometrically informative. Responses under rapid-guessing behavior took little time to complete, were correct at a much lower rate that resembled that from random guessing, and were psychometrically uninformative.

The nonrandom second choice. When an item is presented, the first choice a test taker makes is whether to respond using solution behavior or rapid guessing. If rapid guessing is chosen, then the test taker must then quickly select a response option. But how is this selection made? One might expect that a disengaged test taker repeatedly chooses some readily identifiable option (e.g., option A), but that rarely occurs. To illustrate this, using the Test Event Data Set, I selected test takers who had exhibited between 8 and 12 rapid guesses. Across those rapid guesses, I calculated the numbers of different response option positions (A–D) selected during a test event. Most of the time (56%), all four option positions were selected, at least three were selected 91% of the time, and a single position was selected only 1% of the time. The diversity of option selections suggests that test takers tend to make a deliberate choice among the different options when they rapid guess, rather than simply reflexively choosing a single option.

The strength of this tendency can be seen with the most disengaged test takers. If we focus on the test events for which RTE was less than .40, of the rapid guesses that occurred during the last 10 items of the test event at least three option positions were selected 76% of the time. What is most remarkable about this finding is how quickly the choice was made; Table 1 shows that the mean response times of rapid guesses were only 1.40 seconds! Response times that short suggest that instead of fully reading and understanding the challenge posed by the item before they select a response, rapid guessers quickly scan only the response options and make an intuitive (and usually wrong) System 1–based judgment about their correctness.

The accuracy of rapid guesses. As noted above, the overall accuracy rates of rapid guesses, across items, is similar to that expected from random guessing. This does not mean, however, that this accuracy rate is consistently seen for individual items. In fact, the accuracy rates can vary substantially. Table 2 shows, for the Item Data Set, the percentages of responses to each response option for each type of response behavior. Under solution behavior, the correct option was consistently the most frequently chosen for each item, though the percentages to the other options varied. Under rapid guessing, however, an interesting pattern emerged. Option B was almost always the most frequently chosen option, option D was the least frequently chosen, with the frequencies for options A and C somewhere in between. Thus, the relative attractiveness of the four options was highly consistent across items, regardless of the location of the correct answer. As a result, the accuracy rate of rapid guesses to a given item was primarily due to the position of the correct answer. Item 4 (correct option B) was passed 34% of the time, well above what would be expected by random responding, while items 1 and 6 (correct option D) were passed only 14% and 12% of the time, respectively.

Table 2 illustrates that the accuracy rates of rapid guessing can vary considerably for individual items. This has an important implication for threshold identification methods that combine response time and accuracy. Several of these methods (Goldhammer et al., 2016; Guo et al., 2016; Lee & Jia, 2014) identify an item's time threshold based on the accuracy of responses relative to that expected from random responding (operationally defined as the reciprocal of the number of response options). The results from Table 2 suggest that the actual rapid guessing accuracy for a particular item can deviate considerably from this value, which limits the utility of threshold methods that use random accuracy as a reference point.

Closing Comments

Over fifty years ago, Cronbach (1960) noted that "in a psychological test the subject must place himself on the scale, and unless he cares about the result he cannot be measured" (p. 52). Cronbach's point, in the context of cognitive tests of ability and achievement, can be summarized simply: measurement requires engagement. But just as test-taking engagement can vary across items during a test event, so too can the validity of information provided by item responses.

This article has shown evidence of two different types of response behaviors. In the first, the test taker approaches an item in a deliberate, effortful fashion. The resulting response provides us information about what the test taker knows and can do, and is the engaged response behavior assumed by our measurement models. The second type of response behavior, in contrast, is quickly executed and noneffortful, with the resulting responses being uninformative about the construct we are trying to measure. This has an important implication: *once a test taker chooses to give a rapid guess to an item, measurement is momentarily suspended*. An item has been administered, and a response will be given, but because that response does not reflect what that test taker knows and can do, measurement will not occur. It may then resume on the next item, or not, depending on the choice made about that item. In this sense, a rapid guess to an item represents a choice by the test taker to opt out of being measured. It does not matter whether it is an unmotivated test taker during a low-stakes test or a highly motivated test taker who is running out of time during his high-stakes test.

Because rapid guesses do not constitute measurement, it makes little sense to include them in scoring. If their only effect is to add unsystematic noise in the form of construct-irrelevant variance to test scores, we might choose to tolerate their presence. But because accuracy rates under rapid guesses tend to differ (and typically be lower) from those of solution behavior, rapid guessing systematically biases scores. The degree of bias increases with the number of rapid guesses, and the amount of distortion can get very large. It has been found, for example, that in extreme cases MAP Growth scores can be distorted by four logits or more (Wise, 2015). This bias can be reduced considerably by using scoring methods that exclude rapid guesses (Rios et al., 2016; Wise & DeMars, 2006).

Given this potential impact on validity, one might expect the identification of rapid-guessing behavior to have become a common practice in educational measurement. This is not yet the case, however, for at least two reasons. Detecting rapid guessing requires the use of CBTs which, though they are growing in popularity, are currently used in only a relatively small proportion of testing programs. In addition, rapid guessing has thus far been used to detect disengaged test taking on multiple-choice items that require an answer from the test taker. This constraint, coupled with the trend toward an increased use of constructed response items suggests that the utility of rapid guessing may be limited. This limitation may be diminishing, however, as recent research has developed expanded response time–based methods for detecting disengagement on both omitted and constructed responses (Wise & Gao, in press).

Beyond the practical limitations, in discussions of rapid guessing with colleagues and other researchers, several reservations are commonly expressed. First, there is concern that fast, effortful responses might be misclassified as noneffortful rapid guesses, which would unfairly penalize fast-thinking test takers. Second, some have noted that there may be slower noneffortful responses that may go undetected by rapid-guessing criteria. Third, it has been suggested that if it were to become known that rapid guesses would be identified and deleted from scoring, then test takers might use this to their advantage by intentionally giving rapid guesses when they otherwise would not. These types of reservations indicate a different type of reluctance to implementing rapid-guessing detection, as they suggest that such detection methods may have unintended consequences that threaten score validity.

The first two reservations are actually two sides of the same issue. If the response time distributions overlap in a manner similar to Figure 1, there must be misclassifications of some sort. Avoiding false positives will increase the number of false negatives, and vice versa. Methods like the normative threshold method are designed to be conservative—choosing threshold values meant to avoid classifying effortful responses as rapid guesses (i.e., minimize false positives). This, in turn, implies that there will be some (hopefully few) noneffortful responses that will mistakenly be deemed solution behaviors. I believe this is sensible, because the point of rapid-guessing detection is to identify disengaged item responses that are distortive and psychometrically uninformative. The fact that the responses classified as rapid guesses may not comprise *all* of the noneffortful responses in a test event does not detract from the improved validity that can be realized by identifying those we can.

The third reservation suggests that if test takers know in advance that rapid guesses will be identified and removed from scoring, they might try a strategy to improve their scores by intentionally giving rapid guesses to items they can quickly identify that they do not know. This presumes, however, that a test taker can be skillful in quickly recognizing that they cannot solve the item. Given the assumption discussed earlier that the test be comprised of items that a test taker should be capable of attempting, I doubt that the strategy would be feasible. In my investigations of rapid guessing, I have observed that time thresholds are typically much shorter than the time a test taker would need to read and comprehend the problem posed by an item. That is, I do not believe that test takers could reliably differentiate “an item I can answer” from “an item I cannot answer” within the item’s time threshold. This assertion is, of course, conjecture informed by experience, and is worthy of more rigorous investigation.

Educational measurement professionals have a responsibility to identify and adopt practices most likely to yield scores with strong validity. Validity requires engaged test takers, and the research on rapid-guessing behavior has shown that we can now identify disengaged item responses that distort measurement. The detection of rapid guessing is a unique capability of CBTs, and as their use becomes more common the identification of disengaged item responses should become a best practice that supports our ability to provide valid scores to educators.

Acknowledgments

The views expressed in this article are solely those of the author and they do not necessarily reflect the positions of NWEA. Special thanks to Gage Kingsbury and James Soland for their helpful comments and suggestions.

Notes

[1] Throughout this article, it is assumed that the items on low-stakes CBTs require an answer to be entered before a test taker can proceed to the next item. That is, omitted answers are not allowed on presented items.

[2] Solution behavior accuracy showed a small increase as RTE decreased. Wise and Kingsbury (2016) showed that rapid guessing in a CAT leads to a mis-targeting of item difficulty. Consequently, test takers exhibiting rapid guessing tend to receive easier items than they otherwise would have, resulting in accuracy rates above the .50 level expected during the CAT.

References

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (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.

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

Bowe, B., Wise, S. L., & Kingsbury, G. G. (2011, April). *The utility of the effort-moderated IRT model in reducing negative score bias associated with unmotivated examinees*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Cao, J., & Stokes, S. L. (2008). Bayesian IRT guessing models for partial guessing behaviors. *Psychometrika*, *73*, 209–230.

Cronbach, L. J. (1960). *Essentials of psychological testing* (2nd ed.). New York, NY: Harper & Row.

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

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

Goldhammer, F., Martens, T., Christoph, G., & Lüdtke, O. (2016). *Test-taking engagement in PIAAC* (OECD Education Working Papers, No. 133). Paris, France: OECD Publishing.

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*, 173–183.

Jin, K., & Wang, W. (2014). Item response theory models for performance decline during testing. *Journal of Educational Measurement*, *51*, 178–200.

Kahneman, D. (2011). *Thinking, fast and slow*. New York, NY: Farrar, Straus, and Giroux.

Kong, X., 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*, 606–619.

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

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*(8), 1–24.

Meyer, J. P. (2010). A mixture Rasch model with item response time components. *Applied Psychological Measurement*, *34*, 521–538.

Pastor, D. A., Strickman, S. N., & Ong, T. Q. (2015, April). *Patterns of solution behavior across items in low-stakes assessment*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

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*, 1–31.

Schnipke, D. L. (1995). *Assessing speededness in computer-based tests using item response times* (Unpublished doctoral dissertation). Johns Hopkins University, Baltimore, MD.

Schnipke, D. L. (1996, April). *How contaminated by guessing are item-parameter estimates and what can be done about it*? Paper presented at the annual meeting of the National Council on Measurement in Education, New York, NY.

Schnipke, D. L. (1999). *The influence of speededness on item-parameter estimation*. Princeton, NJ: Law School Admission Council.

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*, 213–232.

Schnipke, D. L. & Scrams, D. J. (2002). Exploring issues of examinee behavior: Insights gained from response-time analyses. In C. N. Mills, M. T. Potenza, J. J. Fremer, & W. C. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments*. Mahwah, NJ: Lawrence Erlbaum.

Setzer, J. C., & Allspach, J. R. (2007, October). *Studying the effects of rapid guessing on a low-stakes test: An application of the effort-moderated IRT model*. Paper presented at the annual meeting of the Northeastern Educational Research Association, Rocky Hill, CT.

Setzer, J. C., Wise, S. L., van den Heuvel, J. R., & Ling, G. (2013). An investigation of test-taking effort on a large-scale assessment. *Applied Measurement in Education*, *26*, 34–49.

Silm, G., Must, O., & Täht, K. (2013). Test-taking effort as a predictor of performance in low-stakes tests. *Trames*, *17*, 433–448.

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences*, *23*, 645–726.

Wise, S. L. (2006). An investigation of the differential effort received by items on a low-stakes, computer-based test. *Applied Measurement in Education*, *19*, 93–112.

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

Wise, S. L., Bhola, D., & Yang, S. (2006). 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, S. L., & Cotten, M. R. (2009). Test-taking effort and score validity: The influence of student conceptions of assessment. In D. M. McInerney, G. T. L. Brown, & G. A. D. Liem (Eds.), *Student perspectives on assessment: What students can tell us about assessment for learning* (pp. 187–206). Greenwich, CT: Information Age.

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

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

Wise, S. L., & Gao, L. (in press). A general approach to measuring test-taking effort on computer-based tests. *Applied Measurement in Education*.

Wise, S. L., & Kingsbury, G. G. (2016). Modeling student test-taking motivation in the context of an adaptive achievement test. *Journal of Educational Measurement*, *53*, 86–105.

Wise, S. L., Kingsbury, G. G., Thomason, J., & Kong, X. (2004, April). *An investigation of motivation filtering in a statewide achievement testing program*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, *18*, 163–183.

Wise, S. L., & Ma, L. (2012, April). *Setting response time thresholds for a CAT item pool: The normative threshold method*. Paper presented at the annual meeting of the National Council on Measurement in Education, Vancouver, Canada.

Wise, S. L., Ma, L., Kingsbury, G. G., & Hauser, C. (2010, May). *An investigation of the relationship between time of testing and test-taking effort*. Paper presented at the annual meeting of the National Council on Measurement in Education, Denver, CO.

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

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

Yamamoto, K., & Everson, H. T. (1995). *Modeling the mixture of IRT and pattern responses by a modified HYBRID model* (ETS Research Report RR-95-16). Princeton, NJ: Educational Testing Service (ERIC Document Reproduction Service No. ED 395036).