Journal of Computerized Adaptive Testing

Volume 2 Number 1

January 2014

The Utility of Adaptive Testing in Addressing the Problem of Unmotivated Examinees

Steven L. Wise

DOI 10.7333/1401-02010001

The *Journal of Computerized Adaptive Testing* is published by the International Association for Computerized Adaptive Testing
www.iacat.org/jcat

ISSN: 2165-6592

International Association for Computerized Adaptive Testing
IACAT
Advancing the Science and Practice of Human Assessment

Journal of Computerized Adaptive Testing

Volume 2, Number 1, January 2014

DOI 10.7333/1401-0201001

ISSN 2165-6592

The Utility of Adaptive Testing in Addressing the Problem of Unmotivated Examinees

Steven L. Wise

Northwest Evaluation Association

In low-stakes testing programs, test developers have the dual responsibilities of developing and administering a high-quality test that can yield valid scores, and motivating examinees to put forth their best effort to perform well on that test. However, unmotivated examinees present a major threat to validity of scores from these types of testing programs. This integrative review examines the motivational benefits of computerized adaptive tests (CATs), and demonstrates that they can have important advantages over conventional tests in both identifying instances when examinees are exhibiting low effort, and effectively addressing the validity threat posed by unmotivated examinees.

Keywords: *examinee motivation, effort, validity, adaptive testing, CAT, engagement*

Cognitive tests are used to provide valid information about what examinees know and can do. To that end, examinees are typically encouraged to demonstrate the highest level of performance of which they are capable (Cronbach, 1960). It has long been recognized, however, that "unless people are motivated to do their best on a test, their scores will not reflect their maximum performance capabilities" (Betz & Weiss, 1976a, p. 1). The impact of test-taking motivation on performance can be sizable. Wise and DeMars (2005) synthesized a set of research studies focused on the relationship between motivation and test performance, finding that more motivated examinees tended to outperform their less motivated peers by an average of 0.58 standard deviations. Thus, obtaining a valid score requires both a well-developed test and examinees who are willing to demonstrate their knowledge and abilities.

In a number of measurement settings, the issue of examinee motivation is usually not of great concern to test users. In these contexts, it is reasonable to assume that examinees are highly motivated to perform well because a high test score will help them attain something they desire, such as a course grade, graduation, scholarship, certification, or licensure—that is, examinees will perceive meaningful personal consequences associated with their test performance. Generally, if an unmotivated examinee chooses to perform with suboptimal effort in these contexts, it is usually not viewed as the responsibility of the test user, since the test user's primary concern is developing and administering a valid test (i.e., one that is capable of obtaining a valid score).

There are many assessment programs, however, in which tests are administered to examinees who might perceive few, if any, consequences associated with their performance. This type of *low-stakes* assessment (from the examinee's perspective) usually occurs in educational settings. For example, in the U.S. National Assessment of Educational Progress, students are asked to take an assessment that has no bearing on their school grades, and for which they do not receive a score. In the absence of personal consequences, test-taking motivation is driven by internal examinee factors such as competitiveness or academic citizenship (Wise & Smith, 2011), which makes it difficult to assess how motivated a particular examinee is during a low-stakes test administration. Hence, in low-stakes testing contexts, motivation is far less certain and becomes a serious threat to the validity of test scores. Because of this validity threat, ensuring that examinees give their best effort becomes an additional responsibility of the test user.

One of the unknowns with any low-stakes assessment program is the percentage of unmotivated examinees. Estimating this percentage is complicated, however, by the fact that there are varying degrees of motivation that an examinee might experience, and any classification scheme for identifying examinees whose test-taking behavior should be labeled "unmotivated" are necessarily based on arbitrary criteria. Nevertheless, if common criteria are applied across studies, some sense of the prevalence of unmotivated test-taking behavior might be gained.

A number of recent studies have used the *response time effort* (RTE) index introduced by Wise and Kong (2005) as a criterion for classifying a test event as unmotivated. Studies using RTE, which is based on the proportion of rapid guessing behavior exhibited by examinees, have found considerable variation in unmotivated behavior. In Kindergarten through Grade 12 (K–12) settings, Wise, Kingsbury, Thomason, and Kong (2004) found 1% of students in Grades 6–10 to be unmotivated when a .90 RTE value was used. In contrast, Wise, Ma, Kingsbury, and Hauser (2010), using a more stringent .85 RTE value, found the percentages of unmotivated examinees to be relatively low but increasing with grade—ranging from 1% (in Grade 3) to 7% (in Grade 9).

In higher education settings, unmotivated test taking on low-stakes tests has generally been more prevalent than that found in K–12 settings. Based on an RTE value of .90, Wise and Kong (2005) identified 7% of their examinees as unmotivated, Wise and DeMars (2010) identified 11%, and Swerdzewski, Harmes, and Finney (2011) identified 26%. Using an RTE value of .87 for a series of assessments over the course of a university semester, DeMars (2007) found that the percentages of unmotivated examinees ranged from 1% to 25%. The sizable variation in the percentages of unmotivated students across these studies reflects the diverse impact of test characteristics, examinee characteristics, and contextual factors on the likelihood that a particular examinee will respond with maximum effort during a particular test event.

Thus, examinee motivation represents a potentially serious threat to the validity of test scores. This problem is a distinctive aspect of educational measurement that complicates the science of measuring cognitive proficiency. Measurement in other fields, in contrast, rarely requires practitioners to be concerned about the motivation levels of their objects of measurement. As Cronbach (1960) noted over 50 years ago:

> In making a physical measurement—for instance, weighing a truckload of wheat—there is no problem of motivation. Even in weighing a person, when we put him on the scale we get a rather good measure no matter how he feels about the operation. But 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)

Effectively addressing the validity threat posed by examinee motivation involves pursuing two related goals. The first is to administer tests in a way that promotes examinee engagement and effort. The second is to detect situations in which examinees did not put forth their best effort to their tests. These two goals are part of a general approach to low-stakes measurement that recognizes that (1) examinee motivation can be enhanced through the choice of the testing methods that are adopted, and (2) despite all efforts, there will always be some unmotivated examinees and there needs to be some means of identifying them.

This paper presents an integrative review of the role that a computerized adaptive test (CAT) can play in addressing these two goals. First, the issue of whether a CAT is more motivating is explored based on a review of relevant research. Next, the identification of non-effortful test taking is examined and a unique advantage of a CAT is illustrated using data from a widely used CAT program administered to U.S. school children. Finally, the promise of future generations of adaptive tests to enhance examinee motivation in low-stakes settings is discussed, drawing on the relevant research literature.

Are CATs More Motivating Than Conventional Tests?

It has been frequently suggested that because CATs administer items that are well matched to an examinee's proficiency level, CATs can be more motivating than conventional tests (Linacre, 2000; Mead & Drasgow, 1993; Sands & Waters, 1997; Wainer, 1997; Weiss & Betz, 1973). These suggestions come in two forms. First, it is asserted that lower proficiency examinees taking a CAT will not become disengaged due to being frustrated or discouraged by items that are too difficult. Second, higher proficiency examinees taking a CAT should not become disengaged due to being bored by items that are too easy. Thus, the general claim is that CATs are more motivating than conventional tests because they are better at maintaining an examinee's engagement throughout the test event.

What is the evidence for this claim? The issue of whether a CAT is more motivating can be considered relative to two indicators. First, examinees might report that they perceived a CAT to be more motivating than a conventional test. Second, the relationship between motivation and test performance would suggest that if a CAT is more motivating, it should yield higher test performance relative to that yielded by a conventional test. While the second indicator is more stringent, if examinees reported higher motivation without accompanying evidence of improved test performance, the practical impact of higher reported motivation could be questioned.

There is indirect support in the research literature on student motivation for the claim of increased motivation with CATs. This body of research has consistently found that the most intrinsically motivating tasks are those that individuals find moderately challenging (Pintrich & Schunk, 2002, Chapter 6). Further evidence comes from research on a special type of computer-based test (CBT) called a *self-adapted test*. In a self-adapted test, before each item is administered, the examinee is allowed to choose its difficulty level from a set of discrete (usually 5–8) difficulty strata. Wise, Plake, Johnson, & Roos (1992) compared self-adapted tests and CATs, finding that examinees who were assigned to take a self-adapted test tended to choose difficulty levels that were similar to those they would have received on a CAT. That is, when they were given freedom to choose whichever difficulty levels they wished, most examinees chose levels with items that were about as difficult as those that would have been selected by a CAT algorithm. This finding suggests that the items administered by a CAT (i.e., those that are moderately challenging) are motivationally congruent with the difficulty levels that would be selfselected by examinees. In this way, the findings with self-adapted tests are consistent with research on intrinsic motivation.

Effects on Self-Reported Motivation

Relatively little research has been directed toward understanding the effects of CAT on self-reported motivation. Only three early research studies were found that studied this issue. Betz and Weiss (1976a) found that low-ability college-level examinees reported significantly higher levels of motivation on a stradaptive test (an early type of CAT) than on a paper-and-pencil test (PPT), while the reported motivation of high-ability examinees did not differ by test mode. The authors concluded that "the use of adaptive tests appears to result in comparable levels of motivation in examinees differing in ability level" (p. 31). Pine, Church, Gialluca, and Weiss (1979) administered CATs and PPTs to a group of high-school students, finding higher mean levels of motivation reported on the CAT. Moreover, a stronger motivation effect was found for African-American students, leading the authors to conclude "it may be possible to obtain more comparable motivational states across racial groups using computer-administered tests" (p. 34). These two studies suggest that a CAT might have the effect of creating a more homogeneous motivational environment across examinees, much as it yields proficiency estimates that are more homogeneous in precision. The only other early study found on self-reported motivation (Arvey, Strickland, Drauden, & Martin, 1990) compared CAT and PPT versions of the Armed Services Vocational Aptitude Battery (ASVAB) that were administered to a sample of military recruits. They found higher levels of reported motivation on the CAT.

A more recent study suggested a less positive effect of CAT on motivation. Ortner, Weisskopf, and Koch (2013) compared a nonverbal matrices CAT to a computerized conventional test version. They examined two aspects of examinee motivation: situational fear of failure, and perceived probability of success on a test, finding that examinees receiving a CAT reported higher fear of failure and lower perceived probability of success. It should be noted, however, that participants taking the CAT in the Ortner et al. study received no specific information about how a CAT works. In an earlier study comparing the same nonverbal matrices CAT and conventional test, Ortner and Caspers (2011) found that the test performance of high-anxiety examinees taking a CAT were negatively affected, but *only* when examinees were given no information regarding how CAT works. This finding suggests that the results of Ortner et al. (2013) should be interpreted cautiously, because it is common for operational testing programs using CAT to provide clear advance explanations to examinees about how the CAT experience differs from that of a conventional test.

Thus, while there are some indications that examinees report CAT to be more motivating, the evidence is not clear-cut. Moreover, in the research reported to date, most, if not all, of the examinees were experiencing a CAT for the first time. This leads to questions about the degree to which any observed higher motivation could represent a novelty effect, and the degree to which any motivational benefits would persist once examinees became more experienced with CAT remains unclear.

Effects on Test Performance

Additional evidence of a CAT motivation effect would be provided by empirical studies showing that test performance on a CAT was higher than that yielded by a conventional test from the same measurement context (e.g., same measured construct, examinee population). The results from five meta-analyses comparing CBTs and PPTs, which are described below, are informative to this question. Note, however, that although these meta-analyses focused on score comparability, their results provide information about the magnitude of a CAT motivation effect (which would actually represent evidence of *non-comparability*).

Because the results of the five meta-analyses were not completely consistent, it is helpful in comparing them to note that each examined a standardized mean difference effect size (ES) computed as ($\text{Mean}_{\text{CBT}} - \text{Mean}_{\text{PPT}})/\text{SD}_{\text{Pooled}}$. Positive ESs were consistent with a motivation effect, as they indicate that the higher mean occurred with the CBT. Negative ESs indicated that the PPT mean was higher.

Bergstrom's (1992) meta-analysis compared CATs and PPTs, synthesizing 20 ESs from eight research reports. The examinees in 12 of the ESs were adults, while the remaining eight ESs were based on data from K–12 students. After deleting five ESs to attain a homogeneous set, Bergstrom reported a negligible mean ES of –.002.

Mead and Drasgow's (1993) meta-analysis examined 159 ESs comparing CBT and PPT performance on tests measuring the cognitive ability of young adults and adults. Of the 114 comparison studies they synthesized, 67 (59%) of the CBTs were CATs. They reported a mean ES for power tests of –0.03, which indicated that performance on the CBTs was slightly lower. Mead and Drasgow did not indicate whether adaptivity (i.e., whether the test was adaptive or non-adaptive) was a significant moderator of this mean difference ES. They did, however, report that adaptivity was not a significant moderator of the degree of correlation between CBTs and PPTs. Thus, the results from the first two meta-analyses provide little indication of a motivation benefit associated with a CAT.

The results from the other three meta-analyses suggest a more complicated story. Kim (1999) synthesized 226 ESs comparing CBTs and PPTs across a variety of samples and tests, reporting a mean ES of +.019. Kim found that the set of effect sizes was heterogeneous, and that adaptivity was a significant moderator variable (indicating that there were significant differences in the CBT-PPT ESs for adaptive and non-adaptive CBTs). Specifically, CATs showed an ES of –0.15 while the ES for non-adaptive CBTs was +0.10, suggesting that adaptivity had a negative effect on test performance. Wang, Jiao, Young, Brooks, and Olson (2007) noted, however, that K–12 students represented only 4% of the samples in the Kim (1999) study. They conducted a meta-analysis focused on comparisons between CBTs and PPTs for mathematics tests administered to K–12 students. This analysis showed that, for the 36 ESs selected for analysis, the overall mean ES was –0.06. Adaptivity was found to be a significant moderator, however, with the ESs for non-adaptive CBTs and CATs being –0.09 and 0.08, respectively. This indicated that adaptivity had a positive effect. The Wang et al. meta-analysis was repeated for K–12 tests in reading (Wang, Jiao, Young, Brooks, & Olson, 2008), with similar findings. The overall ES in reading was –0.004, adaptivity was again found to be a significant moderator, and the respective ESs for non-adaptive CBTs and CATs were –0.01 and 0.05. Thus, the Wang et al. meta-analyses both provided evidence (albeit weak) consistent with the presence of a CAT motivation effect.

The mixed findings from five meta-analyses make it difficult to draw strong conclusions about the presence of motivational benefits from CATs. The two analyses focused on K–12 students provide the most promising evidence. It should be noted, however, that PPTs and CATs differ in multiple ways that might distort test performance. For example, CATs generally do not permit examinees to review (and possibly change) their answers, which could negatively affect performance. Additionally, there is some evidence that CATs can make examinees more anxious (Betz & Weiss, 1974a), which could also negatively affect performance. This suggests that, relative to PPTs, CATs can affect examinees in multiple ways—some of which are performance enhancing and some of which are performance diminishing. Empirical comparisons of test performance between CATs and PPTs reflect the net effect of all of these influences, which makes it difficult to isolate the magnitude of a motivational effect. Nevertheless, because mean ESs found in the meta-analyses were consistently small in magnitude, it is probably safe to conclude that any motivational impact of a CAT on test performance is modest.

However, it is important to note that, whatever its magnitude, any potential motivational effect of a CAT on test performance is unlikely to be realized in operational use. This is because most CATs have been introduced in measurement contexts in which PPTs are already being used. Whenever both CAT and PPT versions of a test are administered, establishing the comparability of scores from the two versions is an important concern. For example, Guideline 22.1of the *International Guidelines on Computer-Based and Internet Delivered Testing* (International Test Commission, 2005) states that the scores from computerized and non-computerized versions should "produce comparable means and standard deviations or have been appropriately scaled to render comparable scores" (p. 11). In practice, this means that score comparability is often attained by equating the CAT scores to the scale established using the PPT. The implication of this is that, to the extent that a CAT increases motivation enough to improve test performance, *that improvement will be subsequently be eliminated by the equating process*. In essence, when CAT and PPT versions of a test are being used, the issue of whether the CAT possesses any motivational advantage is likely to be rendered moot by comparability concerns.

Conclusions

The claim that CATs are more motivating than conventional tests is frequently cited as an advantage of adaptive testing. The limited research on examinee self-reported motivation provides some support this claim. Moreover, the claim seems especially true for low-proficiency examinees, whose personal assessment histories tend to be characterized by test events in which the items were far too difficult for them.

However, higher reported motivation has not conclusively been shown to be accompanied by higher test performance. There are several potential explanations for this. First, CATs might increase examinee motivation, but not by enough to significantly impact test performance. Second, any positive effects of motivation on test performance might be offset by other factors that have negative effects (such as not providing an opportunity to review or change answers). Finally, it is probably the case that a CAT motivationally benefits only some examinees. In most measurement contexts, the percentage of examinees that are found to be unmotivated is less than 10%. In these situations, a CAT could markedly improve the test performance of those examinees without meaningfully increasing overall mean performance. Additional research targeted specifically at unmotivated examinees would yield more conclusive results about the extent to which increased motivation translates into improved test performance.

Can Unmotivated Test Taking be More Readily Detected on a CAT?

Because it is an internal state, an examinee's test-taking motivation cannot be directly measured. However, the examinee's motivational state influences the *effort* directed toward the test, which is a behavior that can be measured. This suggests that the amount of effort expended during a test might be used as an indicator of an examinee's motivational state.

It should be noted that non-effortful test-taking behavior does not always indicate a lack of motivation. An examinee who is not feeling well might be highly motivated to do well on the test, while feeling too ill to give much effort. Alternatively, an otherwise motivated examinee might devote relatively little effort to a particular test item about a topic he or she has not yet had an opportunity to learn. Nevertheless, examinee effort generally provides a useful way to evaluate examinee motivation, and it can help identify instances when effort is so low that a score is not a trustworthy indicator of an examinee's level of proficiency.

Self-Report Measures

There are multiple ways to measure an examinee's test-taking effort. The most commonly used method is to ask the examinee, after the test has been completed, to report the level of motivation or effort during the test. Short, self-report instruments using a small number of Likert-style items, such as the Student Opinion Scale (Sundre & Moore, 2002) can produce reliable scores. The self-report method has several advantages. First, it is simple, economical, and can be administered in a short period of time (i.e., generally less than two minutes). Second, it can be used with both PPTs and CBTs. Because of its flexible ease of use, the self-report method has been the most widely used in research.

There are several disadvantages, however, to the self-report method. First, it is mildly intrusive, as examinees who have just completed their test are usually ready to focus on something other than the test they just took. Second, it is unclear how truthfully examinees will respond about their test-taking effort. Some examinees who did not put forth maximum effort might not want to admit it to test users they respect, or if they fear punishment for reporting lack of effort. Alternatively, some examinees are predisposed to attribute failure to lack of effort, and they might falsely report low effort if they believed they did not do very well on their test. Third, self-report measures provide a global assessment of effort, which makes it difficult to study any changes in effort that occur during a test event.

Measures Based on Test-Taking Behaviors

In recent years, there has been a growing interest in measuring effort based directly on behaviors exhibited by unmotivated examinees during a multiple-choice test event. There are several types of behaviors that might be observed. First, the examinee might simply omit items and not answer them. Second, he or she might give answers, but choose them randomly. Third, the examinee might answer very rapidly in an attempt to get the test event over with.

On a PPT, omitted items can be readily identified by the absence of responses. In addition, an examinee's test performance can be compared to that expected by random responding. Unfortunately, either omitting items or demonstrating test performance resembling that of random responding might also be exhibited by low proficiency examinees who would omit or guess at items because they did not know the correct answers. Thus, motivation and proficiency are potentially confounded when trying to interpret these types of behaviors with a PPT. Moreover, it is practically challenging with a PPT to identify instances in which an examinee answered an item very rapidly. Hence, when PPTs are used, the three test-taking behaviors that could indicate non-effortful test-taking either provide ambiguous information or cannot be measured.

As with a PPT, although a CBT can yield information about both omitted items and test performance, it is vulnerable to the same confounding between motivation level and proficiency level. This complicates interpretations of omitted items or low test performance as unambiguous indicators of low effort. Note that a key feature of a CBT is that, if it requires examinees to answer each item before moving on to the next, it can prevent omitted items. However, this feature is unlikely to improve examinee effort, because unmotivated examinees could readily respond randomly to items they otherwise would have omitted.

In contrast to a PPT, a CBT can unobtrusively measure numerous behaviors during a test event. The most important of these is response time, defined as the time elapsed between when an item is displayed and when the examinee enters a response. Item response time, which can readily be measured and recorded, can be used to identify rapid guessing, which has been shown to be a relatively unambiguous indicator of test-taking effort (Wise & Kong, 2005). A r*apid guess* is one that occurs much faster than it should take an examinee to read, understand, and enter a response. Such responses are called rapid guesses because their accuracy rates closely resemble those expected by chance from examinees who respond with random guesses to items.

A CAT, like all CBTs, can measure item response time, which can be used to identify rapid guessing, and permits assessment of effort at the individual item level. Omitted responses are typically not allowed on a CAT, because the item selection algorithm requires answers from all prior items in order to select subsequent items. A CAT, however, has an important advantage over a non-adaptive CBT. A distinctive feature of a CAT is that it can render low accuracy responses relatively unambiguous because it actively avoids administering items that are highly difficult for a given examinee. Assuming that the CAT (and its associated item bank) is able to administer items that are well targeted to an examinee's proficiency level, a set of responses with a very low accuracy rate can provide additional evidence of low effort.

To illustrate this, consider a hypothetical examinee taking a CAT containing items that are calibrated using the Rasch model, and each multiple-choice item has four response categories. Assuming that the CAT is well targeted, a motivated examinee should have a relatively consistent probability around .50 of correctly answering items. An unmotivated examinee who guesses randomly, in contrast, would have a .25 probability of correctly answering items. The response of the examinee to a single item would not provide information sufficient to make a confident decision regarding whether the examinee was motivated or not. Over a set of items, however, confidence could be gained regarding whether the response pattern was more characteristic of a motivated examinee or an unmotivated one. For example, suppose that across a given set of 10 items during a test event, the examinee correctly answered two. Using the binomial theorem, the probability that a motivated examinee correctly answered exactly two items (assuming a .50 probability on each item) would be .04. The probability that an unmotivated examinee correctly answered exactly two items by random guessing would be .28. Hence, the outcome (two correct out of ten) is substantially more likely to have occurred if the examinee was unmotivated than if he or she were motivated. This is intuitively reasonable, because the observation that 20% of the items were correctly answered seems much more likely to have come from an examinee who was randomly guessing than one who one who had a 50% chance of correctly answering each item.

Thus, when a CAT is administered, the response times of individual items and response accuracy over sets of items can potentially be used to assess examinee effort. Table 1 summarizes the sources of effort-related information available under different types of tests. It is seen that CATs provide more information than CBTs, which in turn provide more information than PPTs, leading to the conclusion that non-effortful behavior can be most effectively identified when adaptive tests are used.

Effort Flagging of Test Events

How effort might be assessed on a CAT can be illustrated using the effort flagging criteria described by Wise, Ma, & Theaker (2012). They applied five criteria being studied for use by Northwest Evaluation Association with its *Measures of Academic Progress* (MAP) adaptive testing system. MAP is used to measure the academic growth of U.S. primary and secondary students in mathematics, reading, language arts, and science. MAP proficiency estimates are expressed as scale scores on a common scale that permits a student's growth to be assessed over time.

Table 1. Availability of Indicators of Effort Under Different Test Types

	Test Type		
Effort Indicator	PPT	Non-Adaptive CBT	CAT
Self-Report	Yes	Yes	Yes
Response Time	No	Yes	Yes
Response Accuracy	No	No	Yes

The five effort criteria (described in more detail below) are based on information from both item response time and response accuracy. If any of the effort flags are triggered for a test event, the score is classified as invalid due to low examinee effort. Two of the criteria are based on examinee behavior over the entire test event. Because examinees often exhibit non-effort during only a portion of a test event, however, three additional flagging criteria are based on subsets of the items during the test event.

Rapid guessing is identified based on the conceptualization that each item response can be classified as reflecting either *rapid-guessing behavior* or *solution behavior* (Schnipke & Scrams, 1997, 2002). This classification is done using pre-established time thresholds for each item based on the normative threshold method (Wise & Ma, 2012) set at 10%. This means that the threshold for an item is set at 10% of the average time examinees have historically taken to answer the item. RTE for a test event equals the proportion of the examinee's responses that are solution behaviors (Wise & Kong, 2005).

The first two effort flagging criteria are based on RTE. The first uses the overall RTE for the test event:

Flag A: *If the examinee gave rapid guesses to at least 15% of the items (overall RTE $\leq .85$).*

The second flag—designed to detect non-effort during only a portion of the test event—is based on rolling subsets of items. For example, for subsets of size 10, items 1–10 would be considered, then 2–11, then 3–12, and so on, until the end of the test. Based on rolling subsets of size 10, an additional RTE-based flag was developed for evaluating low effort on a more local level:

Flag B: *If the examinee exhibited low RTE (local RTE $\leq .70$) on at least 20% of the rolling subsets.*

The next two flagging criteria are based on response accuracy. Low-accuracy responses should be evaluated carefully to make sure that they were not due to the examinee receiving items that were much too difficult. It is important that the CAT item bank and selection algorithm be capable of administering items that are well targeted to an examinee's proficiency level throughout the test event. To accomplish this, a bank adequacy requirement is imposed specifying that low response accuracy will only be considered for CAT test events in which, at least 60% of the time, the examinee receives an item with difficulty no more than three scale score points away from the examinee's momentary proficiency estimate[1]. This led to the development of two additional flags related to response accuracy:

Flag C: *If the examinee correctly answered fewer than 30% of the items (overall accuracy ≤ .30) and at least 60% of all of the administered items were within three scale score points of the examinee's momentary proficiency estimate.*

Flag D: *If the examinee exhibited low accuracy (local accuracy ≤ .20) on at least 20% of the rolling subsets and at least 60% of all of the administered items were within three scale score points of the examinee's momentary proficiency estimate.*

The final effort flag is based on the joint occurrence of rapid guessing and low accuracy on any of the rolling subsets of items:

Flag E: *If the examinee correctly answered no more than two items (local accuracy ≤ .20) and gave three or more rapid guesses (local RTE ≤ .70) on any 10*-item subset, and *at least 60% of all of the administered items were within three scale score points of the examinee's momentary proficiency estimate.*

The five flagging criteria were applied to a set of MAP test events. Testing records in reading from the fall and spring testing terms of the 2010–2011 academic year in a single U.S. state were retrieved from the Northwest Evaluation Association's *Growth Research Database*. The test records were limited to 287,690 students in Grades 3–9 who were tested in both testing terms as part of their district sponsored testing programs.

The numbers of effort flags triggered during the fall test administrations are shown in Table 2. Over 31,000 (11%) of the test events were classified as invalid due to low effort. Nearly twice as many test events triggered rapid guessing flags as accuracy flags or joint rapid guessing and accuracy flags. The total number of flags triggered exceeded 47,000, indicating that it was common for multiple flags to be triggered by a test event.

Table 2. Numbers of Test Events With Different Types of Effort Flags Triggered

Effort Flag Type	*N*	% of Total Sample
Zero Flags	255,956	89.0
At Least One Flag	31,734	11.0
Rapid Guessing Flag (Overall or Local)	22,398	7.8
Accuracy Flag (Overall or Local)	11,730	4.1
Joint Rapid Guessing and Accuracy Flag	13,330	4.6

Table 3 shows that when only a single type of flag was triggered, the type of flag varied. Rapid guessing flags occurred most often, followed by accuracy flags, with joint flags occurring relatively rarely. Particularly noteworthy is the finding that nearly a quarter of the test events triggering flags (7,622 out of 31,734) were identified only through accuracy flags. This under-

[1] Standard errors of examinee scores in reading are typically about 3.2 scale score points.

scores the unique advantage of CATs in identifying non-effortful behavior that is characterized by low accuracy, even when rapid guessing does not occur.

Table 3. Numbers of Test Events in Which Only A Single Type of Flag was Triggered

Effort Flag Type	*N*	% of Total Sample
Rapid Guessing Flag (Overall or Local)	10,584	3.7
Accuracy Flag (Overall or Local)	7,622	2.6
Joint Rapid Guessing and Accuracy Flag	1,361	0.5

Evidence for the validity of the flags can be found in the correlation of test scores with other variables if the scores from flagged test events are removed from the sample. The correlation between the entire sample's fall and spring MAP scores was .82. When examinees with flagged test events in either testing session were removed from the data, the correlation increased to .86 even though the variances of both the fall and spring scores were reduced. These results are consistent with the idea that deleting the data from examinees exhibiting non-effortful behavior has the effect of improving the validity of a set of test scores by removing construct-irrelevant variance (Haladyna & Downing, 2004) from the data.

How Can CATs Be Modified to Better Address Examinee Motivation?

It is useful for test users to have methods of detecting when examinees have exhibited low examinee effort, because they help identify scores with low individual score validity (Wise, Kingsbury, Hauser, & Ma, 2012). These methods, however, are designed to be applied *after* the test event has occurred. Consideration might also be given to what could be done during a test event to promote or maintain motivation—that is, to what extent can non-effortful behavior be reduced by the way the test event is conducted? The possibilities of what might be done range from relatively minor changes to the basic CAT algorithm and features to more fundamental changes in how adaptive testing is operationalized. Such changes require psychometricians to think beyond the traditional CAT algorithm and to consider testing methods that can have motivational benefits even in the absence of psychometric benefits.

One potential change in the CAT algorithm that has been investigated is the difficulty of the items that are administered to an examinee. Item selection in a CAT is typically driven largely by a maximum information criterion based in item response theory, which is intended to maximize the efficiency of test events and yield scores with maximal precision. This results in test events in which examinees correctly answer their items about 50% of the time. Some researchers have argued that this success rate is too low and threatens the motivation of examinees who are used to correctly answering a much higher percentage of items when they take tests.

Studies of CATs targeted at higher success rates (usually 60%, 70%, and 80%) have found that examinees do report higher levels of motivation compared to that from a CAT using a 50% success rate, but without significant motivational benefits in terms of better test performance (Bergstrom, Lunz, & Gershon, 1992; Häusler & Sommer, 2008; Lunz & Bergstrom, 1994; Tonidandel, Quiñones, & Adams, 2002). In addition, these studies have demonstrated that CATs using higher success rates yield proficiency estimates with higher standard errors than those from traditional 50% CATs. Thus, there is a tradeoff between motivation and precision; examinees tend to find higher success rates somewhat more motivating, but at the cost of slightly less precise proficiency estimates. Because higher motivation has not been found to result in higher performance, however, precision has been a priority for test developers and few, if any, operational CATs use a higher success rate. Nevertheless, despite arguments that measurement efficiency is the primary reason for a CAT (Wainer, 1993), success rate should be considered a factor that could be potentially be manipulated to enhance examinee motivation.

An additional feature that could potentially be provided to examinees taking a CBT is item feedback regarding the correctness of their responses. The rationale for providing this feature is that, by providing item feedback, examinees' engagement will be maintained because they will be motivated to see if they answered items correctly or not. On a high-stakes fixed-length test, providing item feedback would naturally raise questions about the validity threat posed by exposing item content. But with a CAT (using large item banks) in a low-stakes testing context (in which test-taking motivation is apt to be of greatest concern), providing item feedback would likely raise fewer item exposure concerns.

The research findings on the relationship between item feedback and motivation are mixed. Betz and Weiss (1976a) found that low-ability examinees reported lower motivation when item feedback was provided, while high-ability examinees reported higher motivation. Similarly, Betz and Weiss (1976b) found that item feedback resulted in higher test performance for the high-ability group, but lower performance for the low-ability group. Pine et al. (1979) found that African-American examinees reported more negative attitudes toward item feedback and indicated that item feedback made them nervous and interfered with their concentration. These results suggest that the impact of item feedback varies across examinees such that some might find it motivating while others might find it anxiety producing or distracting. The mixed impact of item feedback on motivation is similar to the mixed impact of item feedback on test performance [see Vispoel (1998) for a review of this research]. Thus, test users trying to improve motivation should use such feedback cautiously, because it might induce other types of construct-irrelevant factors such as anxiety.

Expanded Types of Adaptive Tests

Standardization is one of the foundational concepts of modern measurement. It embodies the "principle that valid interpretation of test scores relies on the expectation that every test administration has been conducted under the same, standardized conditions of measurement" (McCallin, 2006, p. 634). The primary goal of standardization is to reduce the impact of group-level construct-irrelevant variance on test scores. For example, use of a standard time limit with a test ensures that differential time limits do not introduce construct-irrelevant variance that can compromise interpretations of test scores.

Some construct-irrelevant factors, however, do not affect all of the examinees in a group. Person-specific construct-irrelevant variance is a systematic error that can affect an individual examinee's score (Haladyna & Downing, 2004). Test-taking motivation is an example of person-specific construct-irrelevant variance that might negatively affect individual examinee scores, and thus can threaten individual score validity.

Person-specific construct-irrelevant variance due to internal factors such as motivation will generally not be reduced through test standardization. Instead, it might be effectively managed through an *individualization* of the test administration. Specifically, a test might adapt to the presence of a construct-irrelevant factor during a particular examinee's test event by altering the test administration in a way that might mitigate the effect of the validity threat. Cronbach (1960) considered the management of construct-irrelevant factors a particular type of standardization:

> We may better understand the problem of framing directions and arousing motivation if we realize that the psychometric tester tries to standardize the behavior of the subject, as well as the test stimuli. Even though he is measuring individual differences, his procedures are designed to *eliminate* individual differences—to eliminate, that is, variation in every characteristic save the one that his test is supposed to measure. (p. 59–60)

A good example of individualization relevant to motivation is the *effort-monitoring CBT* introduced by Wise, Bhola, and Yang (2006). In this type of test, the computer algorithm monitors item response times. If it detects that an examinee has begun to exhibit rapid-guessing behavior, a message is displayed to the examinee noting that a decrease in effort has been detected and encouraging the examinee to increase his or her effort. Wise et al. found that those examinees receiving messages exhibited increased levels of motivation as indicated by longer response times and higher success rates on subsequent items. A replication of the Wise et al. study by Kong, Wise, Harmes, & Yang (2006) found that examinees receiving effort messages performed significantly better on subsequent items than examinees who deserved, but did not receive messages.

In an effort-monitoring CBT, messages are displayed only to those whose test-taking behavior warrants them. The remaining examinees do not receive messages. In this way, the effort-monitoring CBT is less standardized than a conventional PPT or CBT because its intervention is conditional on the behavior of the examinee[2]. It focuses on maximizing the validity of individual test scores by identifying and reducing construct-irrelevant factors that affect particular examinees rather than maintaining a standard test administration for all examinees. Such a strategy is designed to maximize the collective individual score validity of the scores from a group of examinees.

In what other ways might a CAT adapt to unmotivated examinees other than through manipulation of item difficulty? One answer is that future CATs might select or present items based on a pre-knowledge of an examinee's interests (which would presumably be more motivating). This would require items that could be framed within a variety of contexts. If it were known in advance that a particular examinee liked fishing, for example, a particular mathematics item might be framed in the context of a fishing example or scenario. If a different examinee liked sewing, the same item might be presented in a sewing context. The challenge to the test user would be to develop items that could be plausibly re-framed according to the examinee's interests (the motivational need) while maintaining reasonably consistent difficulty levels (the psychometric need). This idea is speculative, however, and research is needed to understand its feasibility.

Conclusions

Wainer (1993) cautioned that "if we are to improve the practice of testing, we must allow later generations of tests to be better than earlier ones" (p. 19). CATs represent a major advancement in how efficiently information can be gathered about examinee proficiency. While testing efficiency has been a primary reason for implementing CAT programs in education, there is evidence that the targeting of item difficulty to examinee proficiency brings with it modest motivational benefits that can improve individual score validity. It is important, however, that these motivational benefits not be equated away when CATs are used in conjunction with conventional

[2] A traditional CAT can also be considered individualized relative to a conventional test in that the difficulty levels of the items administered depend on the correctness of an examinee's responses to earlier items. The result is that each examinee receives a test with item difficulty levels that are adapted to his or her level of proficiency.

tests (i.e., PPTs or non-adaptive CBTs). Understanding that CATs can improve score validity by reducing construct-irrelevant variance in ways that conventional tests cannot might lead to recognition that test performance differences between testing modes might be viewed as a positive result rather than a threat to comparability for which corrections must be made.

Unmotivated examinees represent an individual score validity threat that is likely to be present whenever low-stakes tests are used in education. Without personal consequences associated with test performance, not all examinees will be motivated to perform with maximum effort. It is, therefore, important that effective methods for identifying instances of low examinee effort are available, because they will indicate test events with scores that have low individual score validity. It has been shown in this review that CATs have clear advantages over conventional tests in identifying such instances—a finding that enhances the value of CATs to those administering low-stakes tests.

A CAT exemplifies the psychometric advantages that can be obtained when a test is less standardized than its conventional counterpart. Additional validity-related advantages might be found in a new generation of CATs that can adapt to the motivational or affective states of examinees. Developments in this direction will require a broader perspective on what it means for a test to adapt. It will also require a consideration of the potential validity gains that might be obtained by adopting individualized test practices that are directed toward reducing the impact of construct-irrelevant variance on test scores.

Low-stakes testing programs should provide high-quality tests that are capable of yielding valid scores and motivating examinees to respond to the test with maximum effort. This review has shown that, by its unique nature, a CAT encourages examinees to maintain engagement during a test while allowing psychometricians to better detect instances when disengagement has occurred.

References

Arvey, R. D., Strickland, W., Drauden, G., & Martin, C. (1990). Motivational components of test taking. *Personnel Psychology, 43*, 695–716. CrossRef

Bergstrom, B. (1992, April). *Ability measure equivalence of computer adaptive and paper and pencil tests: A research synthesis.* Paper presented at the annual meeting of the American Education Research Association, San Francisco.

Bergstrom, B. A., Lunz, M. E., & Gershon, R. C. (1992). Altering the difficulty in computer adaptive testing. *Applied Measurement in Education, 5*, 137–149. CrossRef

Betz, N. E., & Weiss, D. J. (1976a). *Effects of immediate knowledge of results and adaptive ability testing on ability test performance.* (Research Report 76–3). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program. Retrieved from http://iacat.org/biblio

Betz, N. E., & Weiss, D. J. (1976b). *Psychological effects of immediate knowledge of results and adaptive ability testing* (Research Report 76–4). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program. Retrieved from http://iacat.org/biblio

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

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

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing.

Educational Measurement: Issues and Practice, 23(1), 17–27. CrossRef

Häusler, J., & Sommer, M. (2008). The effect of success probability on test economy and self-confidence in computerized adaptive tests. *Psychology Science Quarterly, 50*, 75–87.

International Test Commission (ITC) (2005). *International guidelines on computer-based and internet delivered testing* (Version 2005). Available from http://www.intestcom.org

Kim, J. (1999, October). *Meta-analysis of equivalence of computerized and P&P tests on ability measures.* Paper presented at the annual meeting of the Midwestern Educational Research Association, Chicago.

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.

Linacre, J. M. (2000). *Computer-adaptive testing: A methodology whose time has come.* (MESA Memorandum No. 69). University of Chicago: MESA Psychometric Laboratory.

Lunz, M. E., & Bergstrom, B. A. (1994). An empirical study of computerized adaptive testing conditions. *Journal of Educational Measurement, 31*, 251–263. CrossRef

Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin*, *114*, 449–458.

McCallin, R. C. (2006). Test administration. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development*. Mahwah, NJ: Lawrence Erlbaum Associates.

Ortner T. M., & Caspers, J. (2011). Consequences of test anxiety on adaptive versus fixed item testing. *European Journal of Psychological Assessment*, *27*, 157–163. CrossRef

Ortner, T. M., Weisskopf, E., & Koch, T. (2013). I will probably fail: Higher ability students' motivational experiences during adaptive achievement testing. *European Journal of Psychological Assessment.* CrossRef

Pine, S. M., Church, A. T., Gialluca, K. A., & Weiss, D. J. (1979). *Effects of computerized adaptive testing on black and white students.* (Research Report 79-2). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program. Retrieved from http://iacat.org/biblio

Pintrich, P. R., & Schunk, D. H. (2002). *Motivation in education: Theory, research, and applications* (2nd ed.). Upper Saddle River, NJ: Merrill Prentice-Hall.

Sands, W. A., & Waters, B. K. (1997). Introduction to ASVAB and CAT. In W.A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation*. Washington, D.C.: American Psychological Association. CrossRef

Schnipke, D. L., & Scrams, D. J. (1997). Modeling item response times with a two-state mixture model: A new method of measuring speededness. *Journal of Educational Measurement, 34*(3), 213–232. CrossRef

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 Associates.

Sundre, D. L., & Moore, D. L. (2002). The Student Opinion Scale: A measure of student motivation. *Assessment Update*, *14*(1), 8–9.

Swerdzewwski, P. J., Harmes, J. C., & Finney, S. J. (2011). Two approaches for identifying low-motivated students in a low-stakes assessment context. *Applied Measurement in Education, 24*, 162-188. CrossRef

Tonidandel, S., Quiñones, M. A., & Adams, A. A. (2002). Computer-adaptive testing: The impact of test characteristics on perceived performance and test takers' reactions. *Journal of Applied Psychology, 87*, 320–332. *CrossRef*

Vispoel, W. P. (1998). Psychometric characteristics of computer-adaptive and self-adaptive vocabulary tests: The role of answer feedback and test anxiety. *Journal of Educational Measurement, 35*, 155–167. *CrossRef*

Wainer, H. (1993). Some practical considerations when converting a linearly administered test to an adaptive format. *Educational and Psychological Measurement, 12*, 15–20.

Wainer, H. (1997). *Computerized adaptive testing: A primer*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Wang, S., Jiao, H., Young, M. J., Brooks, T. E., & Olson, J. (2007). A meta-analysis of testing mode effects in Grade K–12 mathematics tests. *Educational and Psychological Measurement, 67*, 219–238. *CrossRef*

Wang, S., Jiao, H., Young, M. J., Brooks, T. E., & Olson, J. (2008). Comparability of computer-based and paper-and-pencil testing in K–12 reading assessments: A meta-analysis of testing mode effects. *Educational and Psychological Measurement, 68*, 5–24.

Weiss, D. J., & Betz, N. E. (1973). *Ability measurement: Conventional or adaptive?* (Research Report 73-1). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program. Retrieved from http://iacat.org/biblio

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. *CrossRef*

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

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

Wise, S. L., Kingsbury, G. G., Hauser, C., & Ma, L. (2012). *How do I know that this score is valid? The case for assessing individual score validity*. Manuscript submitted for publication.

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.

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. *CrossRef*

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., Ma, L., & Theaker, R. A. (2012, May). *Identifying non-effortful student behavior on adaptive tests: Implications for test fraud detection*. Paper presented at the Conference on the Statistical Detection of Potential Test Fraud, Lawrence, Kansas.

Wise, S. L., Plake, B. S., Johnson, P. L., & Roos, L. L. (1992). A comparison of self-adapted and computerized adaptive tests. *Journal of Educational Measurement*, 29, 329–339. *CrossRef*

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. *CrossRef*

Acknowledgments

The author thanks Lingling Ma for her assistance in both the data extraction and data analyses.

Author Address

Steven L. Wise, Northwest Evaluation Association, 121 NW Everett Street, Portland, Oregon 97209, USA. Email: steve.wise@nwea.org.