

## **Taking the Time to Improve the Validity of Low-Stakes Tests: The Effort-Monitoring CBT<sup>1</sup>**

Steven L. Wise, Dennison S. Bhola, Sheng-Ta Yang

James Madison University

The attractiveness of computer-based tests (CBTs) is due largely to their capability to expand the ways we conduct testing. A relatively unexplored application, however, is using the computer to actively reduce construct-irrelevant variance while a test is being administered. This investigation introduces the effort-monitoring CBT, in which the computer monitors examinee effort (based on item response time) and displays warning messages to those exhibiting rapid-guessing behavior. The results of two experimental studies are presented, showing that an effort-monitoring CBT increased examinee effort, produced higher test performance, and yielded more valid test scores than a conventional CBT. This innovative testing procedure extends the capabilities of measurement practitioners to effectively manage the psychometric challenges posed by unmotivated examinees.

Over the past quarter century, the computer-based test (CBT) has become an increasingly familiar part of measurement practice. The attractiveness of CBTs is due largely to their capability to expand the types of items that can be administered, the manner by which test items can be administered, and the types of information that can be collected during test administration.

One type of CBT information that has long been of interest to researchers is item response time, which is defined as the amount of elapsed time between the display of an item and an examinee's response to that item. Several researchers have suggested methods for incorporating response time into proficiency estimation (Tatsuoka & Tatsuoka, 1980; Thissen, 1983; Wang & Hanson, 2005; Yamamoto, 1995). However, these methods have thus far had little impact on measurement practice.

Part of the challenge of using response time is that it has multiple influences—some of which are relevant to the construct being measured and some that are not. When an examinee encounters a multiple-choice test item, for example, the time it takes the examinee to select an answer is influenced by a number of factors. These factors can be generally classified as characteristics of the examinee, characteristics of the item, or the context in which the item is given. While examinees vary in their levels of proficiency, they can also vary on such variables as reading ability, cognitive processing speed, reaction time, fatigue, and motivation to perform well on the test. Items vary in

---

<sup>1</sup> Paper presented at the 2006 annual meeting of the National Council on Measurement in Education, San Francisco. Correspondence regarding this paper should be addressed to Steven L. Wise, Institute for Computer-Based Assessment, Center for Assessment and Research Studies, James Madison University, MSC 6806, Harrisonburg, VA 22807. E-mail: [wisesl@jmu.edu](mailto:wisesl@jmu.edu).

difficulty, but they also vary in the amount of reading required, and how mentally taxing the item is (e.g., how many steps it takes to complete a mathematics item). In addition, contextual factors such as item position, how free the test administration setting is from distractions or the consequences associated with test performance can influence item response time.

Collectively, these multiple influences make item response time an inherently complex variable for measurement practitioners to productively use, because of the difficulty of unequivocally interpreting the meaning of a given item response time value. There is, however, a region of the response time scale that is much less equivocal. For any multiple-choice test item, there is a time period so short that an examinee could not have had time to read, consider, and select the correct answer. Responses occurring within this time period, we believe, primarily occur because the examinee did not seriously attempt to answer the item.

Rapid responses to multiple-choice test questions have been investigated for over a decade, beginning with Bhola (1994) and Schnipke (1995). Each of these researchers studied the prevalence of short item response times at the end of speeded high-stakes CBTs and concluded that they signaled an examinee's switching from a response strategy of trying to answer the items (i.e., solution behavior) to rapidly submitting guesses to items as the testing time was expiring (i.e., rapid-guessing behavior). The correctness of responses given during rapid-guessing behavior has been shown to be at or near chance levels (Wise & Kong, 2005; Wise, in press) and this suggests that they are essentially random. Schnipke and Scrams further developed and refined this line of research in the context of high-stakes testing (Schnipke, 1996, 1999; Schnipke & Scrams, 1997, 2002).

More recently, Wise and Kong (2005) discovered that rapid-guessing behaviors are also present in the data from unspeeded low-stakes CBTs. Moreover, they found that these rapid-guessing behaviors can occur throughout a test and not just toward the end as Bhola (1994) and Schnipke (1995) had observed with speeded high-stakes tests. Wise and Kong demonstrated that rapid-guessing behaviors on low-stakes tests reflected a lack of examinee effort, and they developed an index, termed *response time effort (RTE)*, for measuring an examinee's overall test-taking effort. *RTE* scores are based on the conceptualization that a test session is comprised of a series of examinee-item encounters. In each encounter, the examinee makes a choice to engage in either solution or rapid-guessing behavior, reflected by the time he or she takes to respond to the item. Thus, for item  $i$ , there is a threshold,  $T_i$ , that represents the response time boundary between rapid-guessing behavior and solution behavior. Given an examinee  $j$ 's response time,  $RT_{ij}$ , to item  $i$ , a dichotomous index of item solution behavior,  $SB_{ij}$ , is computed as

$$SB_{ij} = \begin{cases} 1 & \text{if } RT_{ij} \geq T_i, \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

The index of overall response time effort for examinee  $j$  to the test is given by

$$RTE_j = \frac{\sum_{i=1}^k SB_{ij}}{k}, \quad (2)$$

where  $k$  = the number of items in the test.

This classification of item responses as either solution behaviors or rapid-guessing behaviors (i.e.,  $SB_{ij}$ s) has been used in two subsequent research applications. First, Wise (in press) used  $SB_{ij}$ s to develop an index of the amount of effort received by an item (termed *response time fidelity*), and found that it was highly correlated with both the amount of reading required by an item and the item's position in a test. Second,  $SB_{ij}$ s are a key element in the effort-moderated IRT model of Wise and DeMars (2006). They showed that when rapid-guessing behavior was present in low-stakes test data, the effort-moderated model showed better model fit and item parameter estimation, and yielded more valid scores than a traditional IRT model.

#### *A Proactive Approach to Managing Effort*

Despite their potential for innovative testing methods, contemporary CBTs have tended to follow a traditional pattern of administering an item, receiving an answer from the examinee, administering another item, and so on, until the test is completed and the examinee's score is calculated. That is, the computer has been a relatively passive agent in the examinee-item encounters, and generally only displays item content, records examinee responses, and calculates one or more test scores.

An alternative approach, which is the focus of this investigation, is to provide a more active role for the computer in the testing process. In an *effort-monitoring CBT*, the computer uses response time to classify examinee item responses as solution behaviors or rapid-guessing behaviors while the test is being administered, and takes actions designed to enhance examinee effort. In the current investigation, the actions taken consisted of displaying warning messages to examinees exhibiting rapid-guessing behavior.

In their thorough discussion of construct-irrelevant variance (CIV) associated with high-stakes testing, Haladyna and Downing (2004) noted that, "Compared with other sources of CIV, students potentially provide the most serious CIV threat to validity" (p. 23). In low-stakes testing situations, this threat comes primarily from lack of examinee effort, which degrades the integrity of test score data in two ways. First, to the degree to which an examinee does not give good effort, his or her test score is likely to be biased (i.e., demonstrated proficiency will underestimate actual proficiency). Second, because examinees typically vary in the effort they devote to low-stakes tests, there will be a differential biasing effect that introduces CIV into the test score data.

The rationale for an effort-monitoring CBT is straightforward. There are several approaches that test givers might use to motivate examinees to increase their effort on a low-stakes test. First, one might appeal to their sense of academic citizenship (e.g., you

should try hard on this test because your institution needs you to). Second, one might increase examinees' sense of accountability for their test performance (e.g., if you do not give good effort it will be noticed). Finally, one might impose consequences for lack of effort (e.g., if you do not give good effort something undesirable will happen). Assuming that it can accurately evaluate examinee effort, an effort-monitoring CBT could potentially automate any of these approaches.

### Study 1

The purpose of Study 1 was to conduct an initial experimental investigation of the effects of an effort-monitoring CBT on examinee effort and test performance. There were three research questions. First, would an effort-monitoring CBT increase examinee test-taking effort relative to that from a conventional CBT? Second, would this innovative CBT increase examinee test performance? Finally, would scores from an effort-monitoring CBT exhibit higher validity than those from a conventional CBT?

#### *Method*

When developing an effort-monitoring CBT, there are several decisions that must be made. First, how many messages will there be? Second, what are the criteria for displaying a particular message to an examinee? Third, what will the messages say? We decided to use two warning messages that were to be displayed to examinees exhibiting multiple consistent rapid-guessing behaviors. Specifically, if an examinee exhibited three consecutive rapid guesses, the first warning message would be displayed. For those examinees receiving the first message, if he or she again exhibited three consecutive rapid guesses, the second warning message would be displayed.

Table 1 shows the two warning messages, which were substantially different in tone. The first informed the examinee that his or her responses did not indicate effort, and reminded the examinee that the assessment data being collected were important to the university and the state. Essentially, the first warning message was designed to be persuasive by appealing to the examinee's sense of academic citizenship. In contrast, the second message was a bit more ominous; the examinee was identified by name and warned that continued lack of effort could have consequences for them (i.e., being required to attend a make-up assessment session). However, it is important to note that neither warning message indicated that the computer's effort evaluation was based on response time.

Test administration software was developed that could administer an achievement test either with or without the warning messages. This software provided the basis of the two experimental conditions used in the study. In the first, termed the Warning condition, warning messages would be displayed to examinees who deserved them (by satisfying the display criteria described above). In the second condition—the No Warning condition—warning messages were not displayed to those deserving them. For each item the test administration software recorded the examinee's response and the response time associated with that response. Response time was measured as the number of seconds elapsed between the display of the item and an examinee's submission of a response. During test administration, examinees could not omit items; that is, they were required to provide an answer to an item before they received the next one.

**Table 1.** *Warning Messages Displayed in the Effort-Monitoring CBT Used in Studies 1 and 2*

Warning Order	Warning Text
First	<p>Your responses to this test indicate that you are not giving your best effort.</p> <p>It is very important that you try to do your best on the tests you take on Assessment Day. These assessment data are used by the university to better understand what students learn at &lt;the university&gt;, and what improvements need to be made. In addition, &lt;the university's&gt; assessment data are reported to the state as evidence of what &lt;the university's&gt; students know and can do.</p>
Second	<p>&lt;Examinee Name&gt;, your responses continue to indicate that you are not trying to do your best on this test.</p> <p>These tests are very important, and you need to give them serious attention. Students who do not take Assessment Day activities seriously may be required to attend an Assessment Day make-up session.</p>

*Examinees.* The examinees were 318 mid-year sophomores from a medium-sized southeastern university who were administered computer-based assessment tests during the university's spring Assessment Day in February, 2005. Each Assessment Day, classes are cancelled and all sophomores are required to participate in a 2.5 hour session as part of the assessment of the university's general education program. These tests were considered low-stakes from the students' perspective, as there were no personal consequences based on test performance. The students were assigned to testing groups on the basis of the last two digits of their student identification numbers and these groups were then administered varying combinations of assessment tests. Thus, the students assigned the computer-based tests used in this study essentially constituted a 10% random sample from the sophomore class. From this random sample, students were randomly assigned to the treatment (Warning condition) and control (No Warning condition) groups.

*Measures.* Two instruments were administered to the examinees in Study 1. The first was a university-developed scientific reasoning test designed to assess the university's general education learning objectives in scientific reasoning. The test consisted of 42 multiple-choice items with 2 to 5 response options per item, and many of the items used graphs or short reading passages. The scientific reasoning test was administered via computer in both the Warning and No Warning formats.

The thresholds for distinguishing between rapid-guessing and solution behavior for an item can be identified either by inspecting the item's response time distribution (Wise, in

press) or through its surface features such as item length and whether or not a table, figure, or reading passage was included (Wise & Kong, 2005). Unfortunately, neither approach was practicable in the current study. Because the scientific reasoning test had not previously been administered as a CBT, no time data were available, which precluded threshold identification using response time distributions. In addition, many of the scientific reasoning items were quite mentally taxing to complete even though they did not require much reading. Therefore, it was judged that, for this test, the surface features associated with the amount of reading required by the items would be inadequate indicators of the proper item thresholds.

To address this problem, two of the authors timed themselves working through each scientific reasoning item, to estimate how long it should take an industrious, proficient examinee to complete each item. After independently timing themselves, and discussing their respective results, the authors agreed on a set of initial thresholds for the items.

The second instrument was a 114-item, multiple-choice, locally-developed fine arts test that measured the university's fine arts-related general education objectives. The test included a number of multimedia and text-based stimuli such as dramatic video, recorded music, artwork, and reading passages of literary and philosophical works. The fine arts test was administered only in a conventional (i.e., no warnings given) format.

An additional type of measure calculated for each of the CBTs was RTE, which for a given examinee represented the proportion of his or her item responses that were classified as solution behaviors (Wise & Kong, 2005). RTE scores could range between zero and one, with a value of one indicating highest effort (i.e., solution behavior exhibited on all items).

*Procedures.* The assessment tests were administered in three university computer labs containing between 30 and 102 computers. In the largest lab, the scientific reasoning test was administered first followed by the fine arts test. In the two smaller labs, the order of administration was reversed.

As examinees signed on to the scientific reasoning test, the administration software randomly assigned them to either the Warning or No Warning condition. For those assigned to the Warning condition, when the criteria for a warning had been met, the warning message popped up in a separate alert window in the middle of the computer screen, obscuring most of the previous item content. The warning box remained in view until the examinee closed it by clicking on its "OK" button. The time period during which a warning was displayed was not included in the response times of any of the items. That is, only after a warning box was closed was the content for the next item displayed, thereby initiating the new item's response time measurement.

The examinees were instructed to take the two assessment tests in their assigned order and that they would be allotted 45 minutes to take the scientific reasoning test. The response time data indicated that the examinees took, on average, 19.3 minutes to

complete the test, with a maximum of 37.8 minutes. Therefore, administration of the scientific reasoning test was considered unimpeded.

### *Results and Discussion*

For examinees in the No Warning condition, the respective reliabilities (coefficient alphas) of the scientific reasoning and fine arts scores were .80 and .72. For those in the Warning condition, the reliabilities were .70 and .73, respectively. First warnings on the scientific reasoning test were deserved by 48 of the No Warning condition examinees (34%) and 46 of the Warning condition examinees (26%). Reliabilities of the RTE scores for the conditions were .93 for the No Warning group and .87 for the Warning group.

Table 2 shows the effects of the experimental treatment on test performance and examinee effort, including significance test results (*t*-tests) and effect sizes (*d*). For all examinees, those in the Warning condition significantly outperformed their No Warning peers on both test performance (by about a quarter of a standard deviation) and RTE scores (by over a third of a standard deviation). Because the experimental treatment directly affected only about a fourth of the examinees (i.e., those deserving warnings), effect sizes of this magnitude suggest that the experimental manipulation had a sizable positive impact on examinee behavior.

Table 2 also shows that the results were more pronounced for those examinees who deserved a first warning. The mean difference between the groups on test performance, though not statistically significant, reflected a stronger effect size than the entire examinee group<sup>2</sup>. For RTE scores, the mean scores of those in the Warning condition were markedly higher than those in the No Warning condition ( $d = -0.92$ ). Moreover, only 35% of those in the Warning condition deserving a first warning went on to deserve a second warning, as opposed to 60% of those in the No Warning condition. Overall, the results clearly supported a conclusion that the effort-moderated CBT had a substantial impact on examinee effort and performance for examinees deserving a first warning.

While these results are encouraging, there was one finding that complicated interpretation of the above-mentioned results. Use of an effort-monitored CBT assumes that the test giver can effectively differentiate between rapid-guessing behavior and solution behavior. Wise and Kong (2005) found that rapid-guessing behaviors characteristically yielded responses whose accuracy rates closely resembled those expected by chance (i.e., through purely random responding). For the scientific reasoning test, the expected chance accuracy rate for random responding was .287. For the data in Study 1, however, responses classified as rapid guesses showed a much higher accuracy rate of .494. This indicated that the response time thresholds used tended to be too high, and that some of

---

<sup>2</sup> Because the experimental groups for the entire examinee sample showed significantly different test performance, the finding that the effect size for the examinees deserving a first warning exceeded that for the entire group was interpreted as supportive of the research hypothesis despite its lack of statistical significance.

**Table 2.** *Effects of the Experimental Conditions on Test Performance and Examinee Effort in Study 1*

Variable	Treatment Group				<i>t</i>	<i>df</i>	<i>p</i>	<i>d</i>
	No Warning (n = 142)		Warning (n = 176)					
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>				
	All Examinees							
Performance on Scientific Reasoning Test	25.80	6.12	27.11	5.04	-2.11	316	.036	-0.24
RTE on Scientific Reasoning Test	.82	.19	.88	.12	-3.42	316	.001	-0.39
	Examinees Deserving First Warning							
Performance on Scientific Reasoning Test	21.63	5.85	23.67	4.62	-1.88	92	.063	-0.39
RTE After Deserving First Warning	.51	.28	.76	.26	-4.35	89	<.001	-0.92
Proportion Deserving Second Warning	.60	.49	.35	.48	2.55	92	.013	0.52

*Note:* The numbers of examinees deserving a first warning in the No Warning and Warning conditions were 48 and 46, respectively.

the responses classified as rapid-guessing behaviors should have actually been classified as solution behaviors. Thus, it appeared likely that some examinees in the experiment received warnings that were unwarranted. Our solution to this problem was to establish more appropriate thresholds, with the benefit of the response time data that were now available, and replicate the experimental conditions from Study 1.

### Study 2

The data from Study 1 provided response time distributions for the 42 scientific reasoning items. Inspection of these time distributions permitted an empirically-based identification of thresholds using the method described by Wise (in press), in which thresholds were chosen that corresponded to the end of the initial short time “spike” that was observed for most of the items. This process resulted in a new set of thresholds that were, relative to those used in Study 1, shorter for 33 items, unchanged for 8 items, and longer for one item. The new thresholds ranged from 2 to 30 seconds with a median of 5, whereas the Study 1 thresholds ranged from 3 to 30 seconds with a median of 10.

#### *Method*

Apart from the new set of item response time thresholds, the experimental design and procedures for Study 2 were identical to Study 1 with two exceptions. First, in Study 2, all examinees were administered the scientific reasoning test first, followed by the fine arts test. Second, all of the Study 2 test administrations were held in the largest (102 seat) computer lab.

*Examinees.* The examinees were 435 students who attended one of six makeup sessions that were conducted approximately three weeks after the 2005 spring Assessment Day. Some examinees had missed the original Assessment day for legitimate reasons (and had notified test administrators in advance). Most, however, had simply failed to show up for Assessment Day testing. All students who had missed Assessment Day testing were informed that they were required to attend a makeup session and, if they did not, they could not register for the next semester’s classes.

#### *Results and Discussion*

For 225 examinees randomly assigned to the No Warning condition, the respective reliabilities (coefficient alphas) of the scientific reasoning and fine arts scores were .79 and .77. For the 210 examinees in the in the Warning condition, the reliabilities were .75 and .69, respectively. Reliabilities of the RTE scores for the conditions were .97 for the No Warning group and .92 for the Warning group. Warning messages on the scientific reasoning test were deserved by 68 of the No Warning condition examinees (30%) and 48 of the Warning condition examinees (23%).

Using the new thresholds, there were 2044 responses classified as rapid guesses, 593 of which were correct, for an accuracy rate of .290. This closely matched the expected chance accuracy of .287, and indicated that the new thresholds performed much better than those used in Study 1.

Table 3 shows the effects of the experimental conditions on test performance and examinee response time effort. For all examinees, though the Warning group showed

significantly higher RTE scores, the test performance effect was nonsignificant. For those examinees who deserved a first warning, the effect of test condition on test performance also was nonsignificant. The effect size, however, was nearly a third of a standard deviation and was similar in direction and magnitude to that found in Study 1. As was found in Study 1, after deserving a first warning, those in the Warning condition exhibited significantly higher mean RTE scores. Moreover, the differences between the proportions of examinees deserving second warnings were again significant and similar in magnitude to that found in Study 1.

Figure 1 shows the box plots, by experimental condition, of the RTE scores. The distribution for the Warning condition is markedly closer to the maximum RTE value of 1.0. The corresponding box plots for the scientific reasoning test scores are shown in Figure 2. Though less pronounced than the plots for the RTE scores, there was a noticeably positive shift for the Warning condition distribution.

Two additional analyses in Study 2, the results of which are shown in Table 3, concerned the effects of the experimental conditions on examinee effort and performance when administered the fine arts test (which followed the scientific reasoning test). For those deserving the first warning on the scientific reasoning test, examinees in the Warning condition continued to show significantly higher RTE scores on the fine arts test. However, although the fine arts mean test scores did not significantly differ between the experimental groups, the effect size was similar to that found for the scientific reasoning scores.

Table 3 also shows the effects of the experimental conditions on examinees deserving a second warning. Of these analyses, the only statistically significant effect was found for the RTE scores on the scientific reasoning test. Although the other effects were nonsignificant, and had relatively low effect sizes, the analyses for those deserving a second warning were all in the hypothesized direction (higher performance and effort on both tests for examinees in the Warning condition). These findings suggest that the second warning had less of an impact on the examinees than the first warning. It is unclear, however, how to interpret this finding. It could have been because those receiving the second warning were “harder cases” who had already shown themselves to not be strongly affected by the first message. Alternatively, the results may have been due to a more friendly warning being more effective than a more ominously-worded one. Further research is needed to more fully understand this issue.

#### The Argument for Enhanced Validity

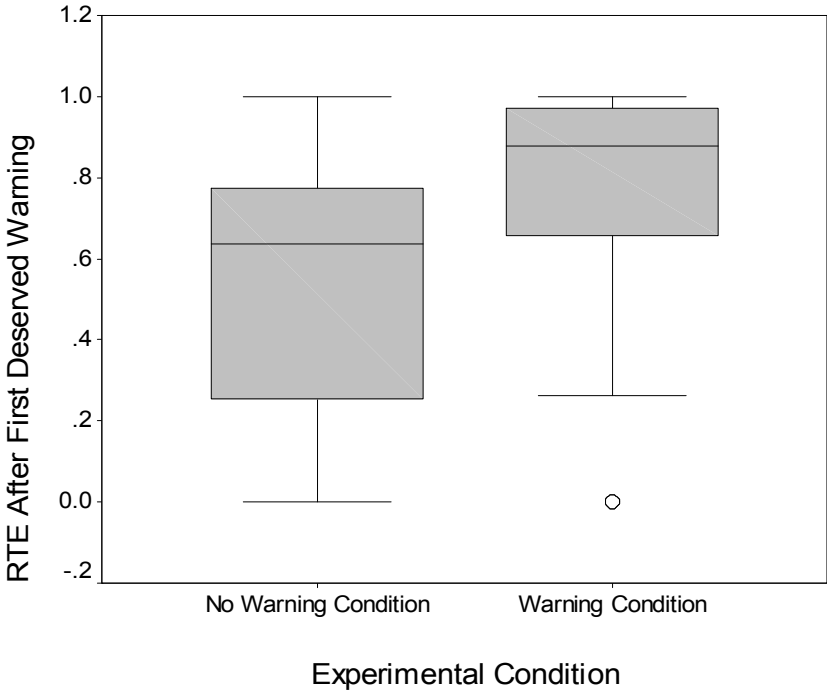
The central research hypothesis of this investigation was that an effort-monitoring CBT will reduce effort-related CIV and, as a result, will yield scores that have higher validity. Although the argument for enhanced validity is limited by the lack of an identifiable criterion for scientific reasoning proficiency, the results from Studies 1 and 2 clearly provided three types of supportive evidence.

First, the findings from both studies that the experimental manipulation resulted in increases in both RTE scores and test performance imply that the stronger test performance by examinees in the Warning condition was attributable to higher effort.

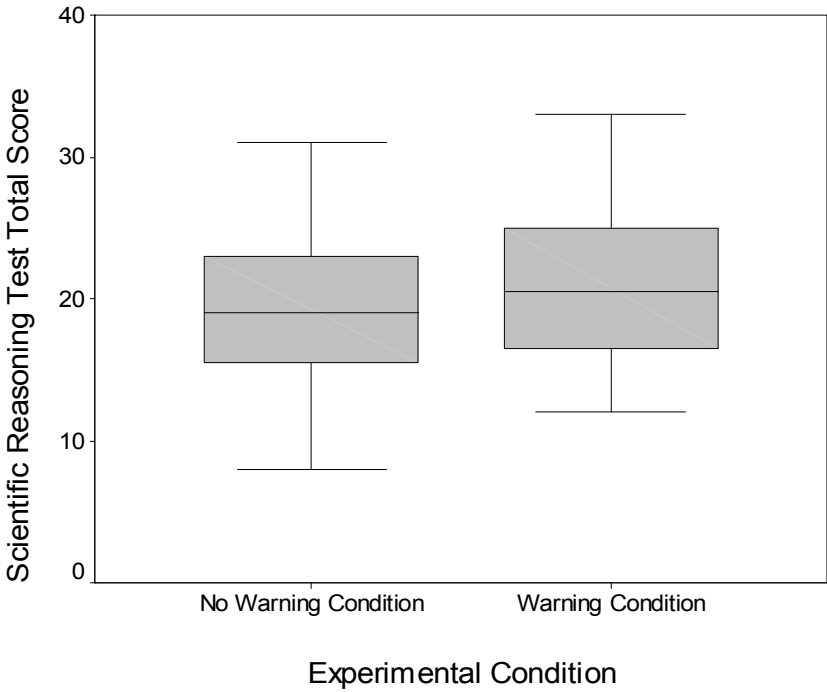
**Table 3.** *Effects of the Experimental Conditions on Test Performance and Examinee Effort for the Two Assessment Tests in Study 2*

Variable	Treatment Group				<i>t</i>	<i>df</i>	<i>p</i>	<i>d</i>
	No Warning (n = 225)		Warning (n = 210)					
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>				
All Examinees								
Performance on Scientific Reasoning Test	25.09	6.15	25.85	5.51	-1.35	433	.177	-0.13
RTE on Scientific Reasoning Test	.86	.23	.92	.13	-3.40	433	.006	-0.32
Examinees Deserving First Warning								
Performance on Scientific Reasoning Test	19.03	5.58	20.81	5.34	-1.72	113	.088	-0.32
RTE After Deserving First Warning	.55	.33	.78	.24	-4.15	111	<.001	-0.78
Proportion Deserving Second Warning	.63	.49	.42	.50	2.26	113	.026	0.42
Performance on Second Test (Fine Arts)	48.40	9.12	51.06	8.39	-1.58	110	.118	-0.30
RTE on Fine Arts Test	.74	.24	.83	.21	-2.00	110	.048	-0.38
Examinees Deserving Second Warning								
Performance on Scientific Reasoning Test	16.90	5.07	18.00	5.27	-0.78	60	.435	-0.21
RTE After Deserving Second Warning	.42	.33	.69	.31	-3.04	58	.004	-0.83
Performance on Following Test (Fine Arts)	46.56	9.39	47.60	7.10	-0.44	59	.664	-0.12
RTE on Fine Arts Test	.69	.25	.76	.25	-1.08	59	.286	-0.28

*Note:* The numbers of examinees deserving a first warning in the No Warning and Warning conditions were 67 and 48, respectively. For those deserving a second warning, the respective numbers from the No Warning and Warning conditions were 42 and 20.



**Figure 1.** Boxplot, by experimental condition, of response time effort (RTE) scores on the scientific reasoning test for examinees who deserved first warnings (Study 2).



**Figure 2.** Boxplot, by experimental condition, of performance on the scientific reasoning test for examinees who deserved first warnings (Study 2).

Given the assumption that the warnings did not somehow cause examinees to know more about scientific reasoning (i.e., attain higher actual proficiency), a reasonable conclusion is that the higher performance was due to the suppression of some systematic, construct-irrelevant factor that would have otherwise caused demonstrated proficiency to be lower than actual proficiency. The largely consistent findings for RTE scores and test performance suggest that examinee effort is the source of CIV that was influenced in Studies 1 and 2.

The second type of evidence comes from the principle that the reduction of CIV should reduce the variance of observed test scores. For all examinees in Study 1, the variance of the scientific reasoning scores (shown in Table 2) was 32% lower for examinees in the Warning condition [ $F(143,175) = 1.47, p = .008$ ]. For all examinees in Study 2, the corresponding variance reduction was 20% [ $F(224,209) = 1.25, p = .051$ ; see Table 3]. Thus, both studies found variance differences consistent with the conclusion that the effort-monitoring CBT reduced CIV.

The third type of evidence concerns the convergent validity of the scores from the two experimental conditions. For the examinees in Studies 1 and 2, information on college grade point average (GPA) and SAT scores (both Verbal and Math subtests) were obtained from university records. These variables, as they are each indicators of general academic ability, would be expected to exhibit positive correlations with scientific reasoning scores. Table 4 shows the convergent validity correlations for the examinees deserving first warnings in each experimental condition. In all six instances, the correlations were higher for the scores from the Warning condition. The tests of the difference between the correlations were statistically significant only for the correlations involving scientific reasoning and SAT-Math; however, the power of the significance tests was constrained by the limited number of examinees deserving a first message. It is interesting to note that the correlations were higher for the Warning condition, despite the reduced variance in the scientific reasoning scores (which, because of restriction of range, should have been expected to yield *lower* correlations). Given both the consistent direction of the differences and the finding that the differences in shared variance were sizable (ranging from 5 to 17%), the conclusion that the effort-monitoring CBT yielded scores with higher convergent validity appears warranted.

#### Conclusions, Implications, and Future Directions

The main finding of this investigation is that an effort-monitoring CBT—by taking an active role in the testing process—can be used to influence examinee effort and thereby enhance the validity of interpretations made on the basis of test scores. This represents an attractive new reason why measurement practitioners might choose a CBT format, particularly to the extent that they are concerned that lack of examinee effort will undermine score validity.

There are a variety of important testing programs that have few, if any, consequences for examinees and consequently are particularly vulnerable to effort-related CIV. These include the various K-12 statewide accountability testing programs, the National Assessment of Educational Progress (NAEP), the Trends in International Mathematics

and Science Study (TIMMS), and the Programme for International Student Assessment (PISA). In addition, it is not uncommon for high-stakes testing programs (e.g., certification and licensure exams; high school graduation exams) to administer test items in low-stakes settings, particularly in the early stages of the program. For example, a state might administer pilot test items for a new high school exam in a nonconsequential setting to obtain the initial psychometric data that are subsequently used to calibrate items and construct test forms. In each of these situations, an effort-monitoring CBT might yield test data that more accurately reflect what students know and can do.

**Table 4.** *Correlations Between Scientific Reasoning Test Score and Several External Variables, By Study and Treatment Group, for the Examinees Deserving First Warning*

Variable	Treatment Group				Difference in Shared Variance <sup>1</sup>	Significance <sup>2</sup>
	No Warning		Warning			
	<i>N</i>	<i>r</i>	<i>N</i>	<i>r</i>		
Study 1						
SAT-Verbal	48	.22	46	.31	+.05	.325
SAT-Math	48	-.01	46	.33	+.11	.049
GPA	48	.32	46	.46	+.11	.219
Study 2						
SAT-Verbal	67	.22	46	.33	+.06	.273
SAT-Math	67	.11	46	.43	+.17	.038
GPA	67	.22	46	.35	+.07	.236

<sup>1</sup>  $r_{Warning}^2 - r_{NoWarning}^2$

<sup>2</sup> Probability associated with *t*-test of the difference between independent correlations

This initial investigation has shown that the computer can play an active role in promoting test score validity. More research will be needed, however, to fully understand and effectively develop effort-monitoring CBTs. One key question concerns both the message contents and the criteria that must be satisfied for a given message to be displayed. There are many potential strategies that test givers might take in interacting with examinees. In this investigation, we intended to deter rapid-guessing behavior through warning messages. Wise and Kong (2005), however, noted that some examinees will exhibit solution behavior through part of the test and then switch over to rapid-guessing behavior for the remaining items. For these examinees, it might be more

effective to provide encouraging messages while they are engaging in solution behavior, and thereby perhaps preempt a shift to rapid-guessing behavior. Alternatively, one might display warning messages to examinees exhibiting rapid-guessing behavior and then display encouraging messages if they then return to solution behaviors.

There are other potential directions for future research. First, how generalizable are the results of this investigation? Can effort-monitoring CBTs be effective in K-12 settings or with NAEP? Second, can the effort-monitoring approach be expanded beyond multiple-choice items to include other item formats as well?

Previous research has shown that item response time can be useful in measuring examinee effort (Wise & Kong, 2005). Moreover, an effort-moderated IRT model (Wise & DeMars, 2006) uses item response time to more effectively model examinee behavior and consequently can better estimate both proficiency and item parameters. Both of these research contributions, however, are decidedly *post hoc* in nature. The unique contribution of the effort-monitoring CBT is its potential to suppress many rapid-guessing behaviors before they occur. This procedure extends the capabilities of measurement practitioners to effectively manage the psychometric challenges posed by unmotivated examinees, and thereby enhance the validity of low-stakes test scores.

## References

- Bhola, D. S. (1994). *An investigation to determine whether an algorithm based on response latencies and number of words can be used in a prescribed manner to reduce measurement error*. Unpublished doctoral dissertation, University of Nebraska-Lincoln.
- Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice*, 23(1), 17-27.
- Schnipke, D. L. (1995, April). *Assessing speededness in computer-based tests using item response times*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.
- 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. (ERIC Document Reproduction Service No. ED400276)
- Schnipke, D. L. (1999). *The influence of speededness on item-parameter estimation* (Computerized Testing Report No. 96-07). Princeton, NJ: Law School Admission Council. (ERIC Document Reproduction Service No. ED467809)
- 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 Mills, C. N., Potenza, M.T., Fremmer, J. J., & Ward, W. C. (Eds.), *Computer-based testing: Building the foundation for future assessments* (pp. 237-266). Mahwah, NJ: Lawrence Erlbaum Associates.
- Tatsuoka, K. K., & Tatsuoka, M. M. (1980). A model for including response time data in scoring achievement tests. In D. J. Weiss (Ed.), *Proceedings of the 1979 computerized adaptive testing conference* (pp. 236-256). Minneapolis, MN: University of Minnesota, Department of Psychology, Psychometric Methods Program.
- Thissen, D. (1983). Timed testing: An approach using item response testing. In D.J. Weiss (Ed.), *New horizons in testing: Latent trait theory and computerized adaptive testing* (pp. 179-203). New York: Academic Press.
- Wang, T. & Hanson, B. A. (2005). Development and calibration of an item response model that incorporates response time. *Applied Psychological Measurement*, 29, 323-339.
- Wise, S. L. (in press). An investigation of the differential effort received by items on a low-stakes, computer-based test. *Applied Measurement in Education*.
- 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., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-cased tests. *Applied Measurement in Education*, 16, 163-183.
- Yamamoto, K. (1995). *Estimating the effects of test length and test time on parameter estimation using the HYDRID model* (ETS Research Report RR-95-2). Princeton, New Jersey: Educational Testing Service. (ERIC Document Reproduction Service No. ED 395035).