Running Head: RESPONSE TIME EFFORT

Response Time Effort: A New Measure of Examinee Motivation in Computer-Based Tests

Steven L. Wise and Xiaojing Kong

James Madison University

(In Press, Applied Measurement in Education)

Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, April, 2005. Correspondence regarding this paper should be addressed to Steven L. Wise, Center for Assessment and Research Studies, James Madison University, MSC 6806, Harrisonburg, VA 22807. E-mail: wisesl@jmu.edu.

Response Time Effort: A New Measure of Examinee Motivation in Computer-Based Tests

When low-stakes assessments are administered to examinees, the degree to which examinees give their best effort is often unclear, complicating the validity and interpretation of the resulting test scores. This study introduces a new method for measuring examinee test-taking effort on computer-based test items based on item response time. This measure, termed response time effort (*RTE*), is based on the hypothesis that when administered an item, unmotivated examinees will answer too quickly (i.e., before they had time to read and fully consider the item). Psychometric characteristics of *RTE* scores were empirically investigated and supportive evidence for score reliability and validity was found. Potential applications of *RTE* scores and their implications are discussed.

When we measure an examinee's proficiency by administering an achievement test, we implicitly assume that the examinee will try to show us what he or she knows and can do. The validity of an inference made about the examinee on the basis of his or her test score, however, is dependent on the amount of effort that the examinee puts forth while taking the test. Without adequate effort, performance is likely to suffer, resulting in the examinee's test score under-representing his or her true level of proficiency. Low examinee effort thus represents a serious potential threat to test score validity.

In many measurement contexts, the issue of examinee effort is deservedly given little attention by test givers. Whenever a test is perceived as "high stakes" by examinees, it is reasonable to assume that they will devote good effort, particularly if there are important personal consequences associated with test performance. Examples of such high-stakes tests include high-school graduation exams, college entrance exams, and certification/licensure exams. In each instance, a strong test score is needed for an examinee to obtain something he or she desires (e.g., entrance to a prestigious university).

Even though one can conceive of some examinees not trying hard on a high-stakes test, few test givers worry about the validity of those examinees' scores, contending that it is the examinee's responsibility to give his or her best effort. In this view, not trying hard on a test is a choice made by an examinee to forgo personal benefits that could be gained through a strong test performance. Thus, the diminished test score validity for these examinees is typically neither the responsibility nor concern of those giving the test.

In contrast, there are a number of measurement situations in which examinee effort is viewed as a serious issue by test givers. Whenever a test is perceived as low stakes by examinees, it is reasonable to assume that some will not give their best effort, particularly if there are no important personal consequences associated with test performance. In these situations, examinees may little care if their test scores do not represent their true levels of proficiency, because (a) they will not receive sanctions for

poor test performance and (b) strong test performance will not help them obtain something they want.

Low-stakes testing can be characterized by the test giver potentially having a much stronger need for valid test performances than do examinees. This shifts the responsibility for obtaining valid test scores toward the test giver, and thus provides an incentive for test givers to understand and effectively manage the test-taking motivation of the examinees.

There are at least three low-stakes testing contexts in which examinee motivation is likely to be a concern for test givers. First, there are a number of low-stakes assessment programs that have serious potential consequences for institutions but little, if any, for individual examinees. In K-12 settings, there has been a growing emphasis on regular low-stakes assessment testing as a means to determine the quality of schools. In higher education, low-stakes assessment testing is increasingly being used to hold publicly funded institutions accountable for expenditures of taxpayer dollars. Second, there are times when high-stakes testing programs need to administer their tests (or test items) in low-stakes settings. For example, it is not uncommon for new testing programs to pilot test items in non-consequential settings, to obtain the data that are subsequently used in item calibration, test form construction, or linking/equating. Thus, even in high-stakes testing programs, low-stakes test administrations often serve a valued purpose. Finally, there is a great deal of measurement research that is conducted in low-stakes settings. For example, at colleges and universities, experimental research is frequently conducted using subject pools from which individuals serve as subjects to satisfy some type of participation requirement for a course.

It is important to note that many examinees appear to devote good effort to test taking in low-stakes settings, even though there are no personal consequences for test performance. The problem is that it is difficult to assess how many examinees try hard and how many do not, which makes it difficult to ascertain the degree to which low examinee effort has biased the test data. This issue has led to small body of research on how test-taking motivation influences test performance.

A number of studies have investigated the relationship between examinee effort and proficiency test performance. The results have been fairly consistent (and unsurprising): less motivated examinees perform less well than their more motivated counterparts. Wise and DeMars (in press) conducted a synthesis of 15 of these studies, finding an average effect size exceeding one half standard deviation between the two groups. Effect sizes of this magnitude would generally be viewed as meaningful by test givers.

In each of the studies cited in Wise and DeMars (in press) examinee test-taking effort was measured using post-test examinee self-reports. Although self-reports have value, there are several potential disadvantages to their use. First, it is unclear how many examinees respond truthfully when asked how hard they tried on a test they had just taken. Examinees can estimate how well they did on the test (accurately or not), and this

estimate might distort their responses to self-report items regarding how hard they tried. For example, motivational researchers have found that some individuals are predisposed to attribute failure on a task to lack of effort over lack of ability (see Pintrich & Schunk, 2002, for a good discussion of attribution theory). If such individuals estimate that they have not done well on a test, they might be biased toward reporting that they gave less effort than they actually did as a way of rationalizing their perceived failure. Our general point is self-report measures of effort are potentially vulnerable to bias through motivational processes, and it is difficult to ascertain the degree to which these factors have influenced a particular set of self-report data.

A related second disadvantage of self-report measures of test-taking effort is that their use requires the assumption that examinees who reported low effort truthfully answered the self-report instrument, but did not seriously try to do their best on the test they were taking minutes earlier. This inconsistency that the same examinees who *did not* take the test seriously *did* seriously complete the self-report items might call into question the credibility of self-report data. The assumption may be defended, however, on the grounds that answering self-report items required so little effort that examinees who were not motivated to answer the test items might reasonably have been quite willing to seriously complete the self-report items. Nevertheless, the assumption is a bit awkward.

Finally, the practice of using a self-report measure of effort administered after a test precludes assessment of potential changes in effort that may have occurred during the course of the test. There is evidence, however, that the effort an examinee devotes to a test may dynamically vary throughout the test. For example, Wolf, Smith, and Birnbaum (1995) found that examinees' efforts varied as a function of how mentally taxing the test items were.

An alternative method that has been proposed for detecting lack of examinee effort is to compare an examinee's item response pattern to a theoretical measurement model. Person-fit statistics are designed to identify examinees with aberrant response patterns. One type of aberrant response behavior is random responding, which may be due to lack of motivation leading examinees to respond too quickly to test items (Meijer, 2003). Meijer and Sijtsma (2001) provide a good overview of person-fit research.

There are two drawbacks, however, to the practical use of person-fit statistics as measures of examinee effort. First, person-fit statistics are sensitive to a number of aberrant behaviors, some of which are inconsistent with lack of student motivation. In addition to random responding, they also have been used (a) to detect cheating, creative responding, careless responding, and lucky guessing (Meijer, 1996), (b) in cognitive diagnosis to help identify examinee misconceptions (Tatsuoka, 1996), and (c) to identify curricular differences among schools (Harnisch & Linn, 1981). Because a person-fit statistic can have multiple influences, it is difficult to conclude that examinee misfit is due to lack of effort. Second, person-fit statistics provide a global measure of response pattern aberrance. As with self-report measures, they are not sensitive to changes in effort that might occur during a test session.

The purpose of the present investigation was to develop a measure of examinee test-taking effort that is based more on direct records of examinee behavior than on self-reported judgments of behavior. In addition, we sought a measure that would provide information regarding effort on an item-by-item basis and would not be confounded with examinee proficiency. This led us to consideration of response time, which can be readily collected during a computer-based test (CBT).

Response time is the difference in seconds between when an item is presented and when it is answered by the examinee. There are several reasons why it provides a desirable basis for a measure of test-taking effort. First, the collection of response time data is unobtrusive and non-reactive; examinees taking a CBT will typically be unaware that response time data are being collected. Second, because it represents a direct observation of examinee behavior, it does not rely on examinee judgments, which may be biased by motivational processes. Third, response time data are available for each item, which would permit investigations of changes in examinee effort during a testing session.

Response time has long been attractive to measurement researchers as an outcome measure. Early research explored the use of response time information in obtaining more accurate proficiency level estimates (Rasch, 1960; Tatsuoka & Tatsuoka, 1980; Thissen, 1983). Response time has also been used to identify appropriate time limits (Bhola, Plake, & Roos, 1993; Gershon, Bergstrom, & Lunz, 1993; Halkitsis, Jones, & Pradhan, 1996; Reese, 1993), study speededness issues (Schnipke & Scrams, 1997; van der Linden, Scrams, & Schnipke, 1998), identify unusual test performance (Kingsbury, Zara, & Houser, 1993, 1994), and in filtering data to more accurately recover IRT item parameters (Bhola, 1994; Schnipke, 1995, 1996, 1999).

We found no prior research that studied the relationship between response time and examinee effort. A program of research that is particularly relevant to the present investigation, however, is that of Schnipke and Scrams (Schnipke, 1995, 1996, 1999; Schnipke & Scrams, 1997, 2002). A central concept in this research is the idea that two different response strategies may be used in a CBT that is speeded. In *solution behavior*, examinees actively seek to determine the correct answer to test items. When time is running out during the testing session, however, examinees may switch to a different response strategy, *rapid-guessing behavior*, in which they rapidly respond to remaining items before time expires. Rapid guessing behavior can be identified by responses occurring so rapidly that examinees did not have time to fully consider the item. Thus, answers given during rapid-guessing behavior are essentially random, and the correctness of these answers will be at or near chance levels.

Response Time Effort

We hypothesized that the two response strategies identified by Schnipke and Scrams for a high-stakes speeded test would also be operating during an unspeeded low-stakes test. Unlike Schnipke and Scrams, however, we believed that rapid-guessing behavior during a low-stakes test would indicate low effort rather than the hurrying-to-finish behavior identified by Schnipke and Scrams. That is, examinees who give good

effort to a test item will exhibit solution behavior. In contrast, examinees who do not try to do well on a test item will exhibit rapid-guessing behavior, because examinees who are not motivated will respond quickly. Instances of rapid-guessing behavior should therefore be identifiable by item response times so short that examinees could not have read and fully considered the item.

Occurrences of rapid-guessing behavior represent the basic component in our new measure of examinee effort, which we termed *response time effort (RTE).* Procedurally, a test is considered a series of items presented to an examinee, and the presence of solution behavior is evaluated for each item. 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

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

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

$$RTE_j = \frac{\sum SB_{ij}}{k},$$

where k = the number of items in the test. *RTE* scores range from zero to one, and represent the proportion of test items for which the examinees exhibited solution behavior. *RTE* values near one indicate strong examinee effort to the test, and the farther a value falls below one, the less effort the examinee expended.

An important issue in the calculation of *RTE* is the specification of the item thresholds used in determining rapid-guessing behavior. Conceptually, the threshold for a given item defines the range of response time that would be too short for an examinee to have a reasonable chance to read the item and identify the correct answer. We decided in this study to base the thresholds on two surface features of the items that have been shown to predict response time: item length (Bergstrom, Gershon, & Lunz, 1994; Halkitsis et al., 1996) and whether or not an item used a figure, an illustration, or some other reading material (Bergstrom et al., 1994).

Once *RTE* had been specified, our general goal became to explore its psychometric properties. To this end, we generated five research hypotheses that, if supported, would make a strong case for the validity of *RTE* as an index of examinee test-taking effort. These hypotheses were then empirically evaluated.

Hypothesis 1: RTE scores should demonstrate adequate levels of reliability. Because SB_{ij}, the index of item solution behavior, is dichotomously scored, classical reliability methods can be used to estimate *RTE* reliability. In this study, coefficient alpha was used, with .80 judged to be the minimal criterion for adequate reliability.

Hypothesis 2: RTE scores should be correlated with other measures of examinee test-taking effort. This is traditional convergent validity evidence that two measures of the same construct should be correlated.

Hypothesis 3: RTE scores should not be correlated with measures of academic ability. If *RTE* scores are valid indicators of examinee effort, it is important that they are not simply surrogates of academic ability. This represents traditional discriminant validity evidence.

Hypothesis 4: Instances of rapid-guessing behavior should yield item scores that are correct at a rate consistent with chance. When the response time for an examinee-item encounter is less than the threshold for that item, examinees are assumed to be randomly choosing an answer. Demonstrating that rapid-guessing behavior yields item scores whose accuracy does not exceed chance would support the assertion that rapid-guessing behavior is distinct from solution behavior. In addition, support for Hypothesis 4 would address the concern that rapid responses are actually rapidly completed solution behaviors (i.e., that they do not reflect lack of effort).

Hypothesis 5: RTE scores should show motivation filtering effects similar to those found with other measures of examinee effort. In motivation filtering, the data from examinees exhibiting low test-taking effort on an achievement test are deleted, or filtered, from a data set. Previous research has shown that when motivation filtering based on self-reported examinee effort was used, (a) test performance improved, (b) test score reliability remained relatively constant, and (c) the correlation between test performance and an external variable showed a substantial increase (Sundre & Wise, 2003; Wise & DeMars, in press). Showing a similar motivation filtering effect with *RTE* scores would provide additional evidence of their validity as indicators of examinee test-taking effort.

Method

Examinees

The examinees used in this study were 506 freshmen at a medium-sized southeastern university who received computer-based assessment tests at the beginning of the fall, 2003 semester. All incoming freshmen (totaling over 3500) were required to participate in testing as part of the university's general education assessment. To assess each examinee's level of academic ability, Scholastic Assessment Test[1] (SAT) scores (both Verbal and Quantitative) were obtained from the student records database. One or more SAT scores were missing for 34 examinees; these examinees were consequently deleted from the analyses, leaving a final sample size of 472.

Measures

Achievement Test. The achievement test used in this study was an 80-item pilot form of the Information Literacy Test (ILT), which is a locally developed multiple-choice test used to assess student skills in finding, critically evaluating, and using information effectively. On many of the ILT items, examinees were provided ancillary online

materials, such as tables, figures, or web sites for them to read in order to answer the question posed in the item. The number of response options per item ranged from two to five. The computer-administered ILT was designed to match the Association of College and Research Libraries (ACRL) information literacy competency standards. ILT scores represent the percentage of items passed by the examinee. In the current study, the reliability of the ILT scores was acceptable, with coefficient alpha equal to .83.

Self-Reported Effort. To measure examinee self-reported effort on the ILT, the Reported Effort subscale of the Student Opinion Scale (SOS) was used (Sundre, 1999; Wolf & Smith, 1995). The SOS is a brief, ten-item motivation scale that yields a total score and two five-item subscale scores (Reported Effort, Perceived Importance of the Test). Each SOS item uses a five-point response scale ranging from strongly disagree to strongly agree. The scores of the Reported Effort subscale have a possible range of 5 to 25, with 5 indicating the lowest reportable effort. In the current study, the SOS was administered by computer.

Estimated reliability (using coefficient alpha) for each of the three SOS scores has consistently been in the mid to upper 80's in college samples, and favorable validity evidence has been found (Sundre, 1999; Sundre & Moore, 2002). Comparable score reliability was found in the current study; the Reported Effort subscale yielded a coefficient alpha equal to .83.

Person-Fit Statistic. The Modified Caution Index (MCI; Harnisch & Linn, 1981) was used as a measure of person fit. Karabatsos (2003) compared 36 person-fit indices, concluding that MCI was one of the five best indices for identifying aberrantly-responding examinees.

Procedure

Testing was conducted in several university computer laboratories. Although most of the testing (62%) was done on the Saturday before classes began, some (10%) was done a week earlier to accommodate examinee athletic commitments, with the remaining 28% occurring two to three weeks after the semester began, as makeup testing for those students who missed the Saturday assessment day.

The general assessment day procedures require each freshman to attend a 150-minute assessment testing session. A matrix-sampling format was used in which different subsets of students were randomly administered different combinations of assessment tests. Some students received paper-and-pencil tests, while others received computer-based tests. The examinees in the current study were administered the ILT first, followed by the SOS, and then one additional test. Although a time limit of 90 minutes was imposed during administration of the ILT, 99% of the examinees finished within 70 minutes, and all finished within 88 minutes. Thus, the ILT could be considered unspeeded. The SOS took approximately two additional minutes to complete.

The examinees were required to attend an assessment testing session during which the value of the assessment data to the university was explained to them by university proctors. Test performance carried no personal consequences for the examinees, however, and they were free to give less than their best effort if they chose. In addition, the examinees were provided no feedback on their test performance. Hence, from the perspective of the examinees, this was clearly a low-stakes testing session. As is typical in such testing, most examinees appeared to be cooperative and try hard. A small percentage, however, were observed by the assessment test proctors to exhibit a general lack of motivation during the assessment testing.

Results

An initial task was to identify the thresholds used to differentiate, for each item, rapid-guessing behavior from solution behavior. An examination of the response time distributions across items revealed similarity in the distribution shapes. Figure 1 shows the response time frequency distributions for two ILT items. For item 7 the distribution is unimodal and positively skewed, but with a small frequency spike occurring under 3 seconds. The distribution for item 35 has a similar shape except for a more pronounced short time spike that extended to about 5 seconds. These response time distributions with short time spikes are highly similar to those that tend to occur at the end of speeded tests (Schnipke, 1996, 1999; Schnipke & Scrams, 1995, 1997, 2002), yet they occurred for most, if not all, of the ILT items. This indicates that rapid-guessing behavior was occurring throughout the ILT testing session.

Although the ILT items consistently exhibited short time spikes in their response time distributions, the "width" of the spike varied substantially across items, extending from only a few seconds up to 10 seconds. Our inspection of the items and their characteristics suggested that the variance in spike width was strongly related to the amount of reading required by each item. Consequently, we measured the total length of each item's stem and options (in characters) and established three initial item response time thresholds according to the following rules. If an item was shorter than 200 characters, a 3-second threshold was used. If an item was longer than 1000 characters, or if the item provided some particular ancillary reading for the first time, a 10-second threshold was used. For the remaining items, a 5-second threshold was used.

The three thresholds appeared to adequately represent the short time frequency spikes for all but four of the ILT items. For these items, each of which had an initial threshold of 5 seconds and no ancillary reading, the short time spike spanned roughly 3 seconds. The four items were very similar in content; in each the examinee was provided an entry from a reference list and asked to identify its type. We believe that to correctly answer these items, an examinee did not have to read the entire reference; rather, the examinee needed only to scan for key elements of the reference (such as a volume number which would indicate a journal article). This implies that completion of solution behavior could reasonably require less than 5 seconds. Consequently, the thresholds for these items were changed to 3 seconds. Across the 80 ILT items, the final number of 3-second, 5-second, and 10-second thresholds was 27, 46, and 7, respectively.

The item thresholds were then used to differentiate rapid-guessing behavior from solution behavior in calculating *RTE* scores for each examinee. Figure 2 shows a histogram of the *RTE* scores for the study sample. It is highly negatively skewed, indicating that most examinees exhibited response times that consistently exceeded the item thresholds (i.e., *RTE* scores near 1.0). There were 299 examinees with RTE equal to 1.0. A number of examinees, however, exhibited *RTE* scores markedly below 1.0, indicating substantial rapid-guessing behavior. The lowest observed *RTE* score was .20; this examinee exhibited solution behavior for only about one fifth of the items.

Research Hypotheses

Hypothesis 1: RTE scores should demonstrate adequate levels of reliability. Because the *RTE* scores were based on a summation of the dichotomous indices of item solution behavior (SB_{ij}), coefficient alpha could be used to estimate *RTE* reliability. The observed value of alpha was .97, indicating a very high degree of internal consistency. Thus, the *RTE* scores could be considered reliable, and Hypothesis 1 was supported.

Hypothesis 2: RTE scores should be correlated with other measures of examinee test-taking effort. Table 1 shows the correlations among a number of the measures used in the current study. *RTE* scores showed significant correlations with both self-reported effort and person fit. The magnitude of the correlations, however, suggests that the three effort measures may not be measuring exactly the same construct. A scatter plot of the relationship between *RTE* and self-reported effort is shown in Figure 3. Hypothesis 2 was also considered supported.

Test performance showed significant correlations with self-reported effort, *RTE* scores, person fit, and total test time, as well as with the two SAT subscale scores. Interestingly, *RTE* scores showed the highest correlation, with values exceeding even those of the SAT subscale scores.

Hypothesis 3: RTE scores should not be correlated with measures of academic ability. Table 1 shows that the correlations between *RTE* scores and the SAT subscale scores were near zero, providing support for Hypothesis 3. Hence, the *RTE* scores showed evidence of discriminant validity.

Total test time was included in Table 1 to provide a basis of comparison. The different patterns of correlations for *RTE* scores and total test time suggest that the two time-based measures are distinct. *RTE* scores were more strongly correlated with those measures with which they were expected to be related (i.e., test performance, self-reported effort, person fit), and less strongly correlated with measures with which they were not expected to be related (i.e., SAT subscales).

The relationships between *RTE* scores and several of the other study variables were further illustrated by dividing the examinee sample in three groups: those with *RTE* scores less than .80 (those giving the lowest effort to the ILT), those with *RTE* scores between .80 and .90, and those with *RTE* scores exceeding .90 (those giving the highest

effort). Table 2 shows descriptive statistics for the three groups along with analyses of variance and effect sizes. The groups showed significant differences in test performance, with the lowest and highest *RTE* groups showing performance differences that exceeded two standard deviations. Similarly, consistent with Hypothesis 2, the *RTE* groups differed in self-reported effort and person fit in a manner that was consistent with their being measures of test-taking effort. Finally, the *RTE* groups showed minor, nonsignificant differences in SAT scores, which was consistent with Hypothesis 3.

Hypothesis 4: Instances of rapid-guessing behavior should yield item scores that are correct at a rate consistent with chance. If an examinee's response time did not exceed the threshold for that item, his or her response was inferred to have reflected rapid-guessing behavior. If this was true, then the accuracy of responses under rapid-guessing behavior should have not exceeded what would have been expected by chance under random responding. Table 3 shows the accuracy of rapid-guessing responses and solution responses for a subset of 38 ILT items[2]. The accuracy of the rapid-guessing responses, which was far lower than that of the solution responses, did not significantly exceed chance level. Therefore, Hypothesis 4 was supported.

Hypothesis 5: RTE scores should show motivation filtering effects similar to those found with other measures of examinee effort. Previous studies of motivation filtering (Sundre & Wise, 2003; Wise & DeMars, in press) found that by deleting from a sample of test data examinees who report low effort, test score validity can be markedly improved. Table 4 shows the results of motivation filtering based both on *RTE* scores and on self-reported effort. The results are highly similar for both filtering variables. Successively filtering greater numbers of examinees yielded ILT scores with increased means and decreased standard deviations. Coefficient alpha decreased, but this was largely due to the decreased observed score standard deviations, as evidenced by the relatively constant standard errors of measurement (SEM). At the same time, filtering did not appear to delete predominantly lower ability examinees; the SAT-Verbal scores remained virtually unchanged. Most important, the filtering had the effect of enhancing the convergent validity correlation between ILT scores and an outside criterion (in this case, SAT-Verbal scores). Thus, because the motivation filtering results for *RTE* scores and self-reported effort were similar, Hypothesis 5 was supported.

It should be noted that motivation filtering using *RTE* scores yielded modestly more favorable results than did self-reported effort. That is, across successively stringent filtering using *RTE* scores, both mean test performance and convergent validity correlation showed greater increases, even though fewer examinees were deleted. This suggests that, at least in the current study, *RTE* scores may have been more potent measures of examinee effort than was self-reported effort.

Consistency of Effort During the Testing Session

One of the purported advantages of a response time based measure of examinee effort was that it provided information regarding examinee effort down to the level of individual items. This would permit investigations, for example, of whether effort

systematically changes during a testing session. Figure 4 graphs the response times across the 80 ILT items for an examinee who appeared to change response strategies during the test. His *RTE* score was .53, his self-reported effort score was 11, and his MCI was .29. During the first 50 items, the examinee's response times were relatively long, indicative of primarily solution behavior. Around item 50, however, the response times abruptly changed to consistent rapid-guessing levels (with response times of one to two seconds for nearly all of the remaining items). This pattern suggests that the examinee gave up and stopped trying about two thirds of the way through the test. This type of examination of item response times provides more information about this examinee's testing session than would the global self-report effort and MCI measures.

Discussion

When an examinee receives a test item, he or she chooses whether or not to engage in solution behavior. When the test is high-stakes—with personal consequences for the examinee—this choice is almost always in the affirmative. When the test is low-stakes, however, many examinees may not choose solution behavior. One option available to them would be to simply not respond to the test items. This behavior, however, is easy to spot in the test data, and these examinees could be readily discarded from the sample. Another option for examinees is to engage in rapid-guessing behavior, perhaps to finish the test sooner and move on to more valued activities. Such behavior is inconsistent with the measurement models currently used in practice (i.e., classical test theory or item response theory). Also, it is more difficult to differentiate rapid-guessing behavior from solution behavior solely on the basis of the item responses. The current study has shown however, that such differentiation is possible when response time information is considered.

The purpose of this study was to develop a measure of examinee test-taking effort that was (a) based directly on examinee behavior rather than examinee self-reports and (b) able to provide information regarding examinee effort down to the level of individual items. Using response time as a basis for an effort measure allowed us to meet each of these goals. Moreover, *RTE* scores are unobtrusive and non-reactive, which are highly desirable characteristics.

The five research hypotheses in this study were all supported. *RTE* scores were found reliable, and they showed evidence of both convergent and discriminant validity. Clear evidence was found of two different examinee behaviors, with occurrences of rapid-guessing behavior yielding item responses whose accuracy did not exceed chance levels. In addition, both *RTE* scores and self-reported effort showed very similar motivation filtering effects. Collectively, the results of this study are consistent with the assertion that *RTE* scores measure examinee test-taking effort.

Potential Applications of RTE Scores in Measurement Practice

Assuming, then, that *RTE* scores are measuring examinee effort, what value could they have for measurement practitioners? First, they could be generally valuable as

indicators of how much effort examinees gave during a test. In many instances, test givers recognize the potential for low student effort to influence achievement test data, but have little empirical evidence concerning the degree to which low effort was actually present. *RTE* scores might make more concrete what were previously speculations regarding examinee motivation and effort.

Second, *RTE* scores could provide valuable information to researchers actively trying to understand the dynamics of examinee test-taking motivation. Having a measure that can provide information at the item level should encourage additional research regarding the testing practices that optimally facilitate examinee effort. One might study, for example, how effort varies across item types, the functional relationship between effort and test length, the impact of item feedback on examinee effort, or how adaptive testing affects effort. There is much we do not know about how examinees behave when taking tests, and *RTE* scores may be valuable in increasing our understanding of the dynamics of test taking.

A third application of *RTE* scores is in motivation filtering. In K-12 settings, low-stakes tests are often used to assess how many students in a group (school, district, state, etc.) have met a proficiency standard. To the extent to which students who are truly proficient do not give good effort—and thereby fail to demonstrate proficiency—the observed proportion of students whose scores exceeded the proficiency standard will underestimate the true proportion of students who are proficient. This could have serious implications if, say, a school faced sanctions if it did not bring its students up to a targeted level of proficiency, and it was misclassified as failing to meet its target. If students with low *RTE* scores were filtered out of the sample, the proportion of remaining students attaining proficient scores may provide a more accurate estimate of the true proportion of proficient students in the group. The logic of motivation filtering, however, requires the assumption that examinee effort is unrelated to ability.

Motivation filtering might also be useful to high-stakes testing programs that are administering items in a low-stakes testing setting, to collect data to be used in item calibration or test equating. If examinees with low *RTE* scores were filtered out, item calibrations based on the remaining data should yield parameter estimates that better represent the characteristics of the items in the operational high-stakes setting. For example, in a simulation study based on a speeded, high-stakes test, Schnipke (1996) found that the deletion of rapid-guessing responses yielded more accurate recovery of item parameters.

Additional Research

This study introduced the rationale underlying response time effort and reported the results of the initial investigation of its properties. There are, however, a number of research questions that remain.

One question concerns the item thresholds used to distinguish between rapid-guessing and solution behavior. We used a relatively simple method of establishing three

thresholds based on item length and whether a particular ancillary reading was provided for the first time. There was some indication that these thresholds may not have been appropriate for all of the items (see Figure 4). Because of the importance of these thresholds to the *RTE* score, more research should be directed toward improving threshold identification. An important issue is whether to establish thresholds empirically, in some data-independent manner (such as using surface features of the items), or through some combination of the two types of information.

Another research question concerns how to explain the modest magnitude of the correlation found in this study between *RTE* scores and self-reported effort. There are several potential explanations. First, the two might be measuring slightly different constructs. Second, the correlation might be influenced by examinee attribution biases that affect only self-reported effort. Third, the explanation might be statistical. The negatively skewed distribution of *RTE* scores could represent a type of ceiling effect in which examinees exhibiting a range of higher level of effort all receive a similar *RTE* score of one. In this case the modest correlation between *RTE* scores and self-reported effort could be attributed to a restriction of range effect. Fourth, *RTE* scores may identify only a portion of the unmotivated examinees. There may be, for instance, some examinees who respond slowly but without effort. More research is needed to better understand the relationship between *RTE* scores and self-reported effort.

A final research question concerns the use of response time information in the proficiency estimation process, as has been explored previously (Rasch, 1960; Tatsuoka & Tatsuoka, 1980; Thissen, 1983). If response time information can accurately differentiate between rapid-guessing behavior and solution behavior in low-stakes testing, then we might be able to more effectively model examinee test taking and thereby more validly assess examinee proficiency.

References

Bergstrom, B. A., Gershon, R. C., & Lunz, M.E. (1994, April). *Computer adaptive testing: Exploring examinee response time using hierarchical linear modeling.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

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.

Bhola, D. S., Plake, B.S., & Roos, L. L. (1993, October). *Setting an optimum time limit for a computer-administered test.* Paper presented at the annual meeting of the Midwestern Education Research Association, Chicago, IL.

Gershon, R. C., Bergstrom, B. A., & Lunz, M. (1993, April). *Computer adaptive testing: Exploring examinee response time.* Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta, GA.

Halkitis, P. N., & Jones, J. P. (1996). *Estimating testing time: The effects of item characteristics on response latency.* Paper presented at the annual meeting of the American Educational Research Association, New York, NY.

Harnisch, D. L., & Linn, R. L. (1981). Analysis of item response patterns: Questionable test data and dissimilar curriculum practices. *Journal of Educational Measurement, 18,* 133-146.

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education, 16*, 277-298.

Kingsbury, G.G., Zara, A. R., & Houser, R. L. (1993, April). *Procedures for using response latencies to identify unusual test performance in computerized adaptive tests.* Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta, GA.

Kingsbury, G.G., Zara, A. R., & Houser, R. L. (1994, April). *Modeling item response latencies in computerized adaptive tests.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Meijer, R. R. (1996). Person-fit research: An introduction. *Applied Measurement in Education, 9*, 3-8.

Meijer, R. R. (2003). Diagnosing item score patterns on a test using item response theory-based person-fit statistics. *Psychological Methods, 8*, 72-87.

Meijer, R. R., & Sijtsma, J. (2001). Methodology review: Evaluating person fit. *Applied Measurement in Education, 25*, 107-135.

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

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danmarks Paedogogishe Institut.

Reese, C. M. (1993, April). *Establishing time limits for the GRE computer adaptive tests.* Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta, GA.

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., Fremer, J. J., & Ward, W. C. (Eds.). *Computer-based testing: Building the foundation for future assessments.* Mahwah, NJ: Lawrence Erlbaum Associates.

Sundre, D. L. (1999, April). *Does examinee motivation moderate the relationship between test consequences and test performance?* Paper presented at the annual meeting of the American Educational Research Association, Montreal. (ERIC Document Reproduction Service No. ED432588)

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

Sundre, D. L., & Wise, S. L. (2003, April). *'Motivation filtering': An exploration of the impact of low examinee motivation on the psychometric quality of tests.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Tatsuoka, K. K. (1996). Use of generalized person-fit indices, zetas for statistical pattern classification. *Applied Measurement in Education, 9*, 65-76.

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.

van der Linden, W. J., Scrams, D. J., & Schnipke, D. L. (1998). *Using response-time constraints in item selection to control for differential speededness in computerized apaptive testing* (Report No. 98-06). Enschede, Netherlands: Twente University.

Wise, S. L., & DeMars, C. E. (in press). Examinee motivation in low-stakes assessment: Problems and potential solutions. *Educational Assessment.*

Wolf, L. F., & Smith, J. K. (1995). The consequence of consequence: Motivation, anxiety, and test performance. *Applied Measurement in Education, 8*, 227-242.

Wolf, L. F., Smith, J. K., & Birnbaum, M. E. (1995). Consequence of performance, test motivation, and mentally taxing items. *Applied Measurement in Education, 8*, 341-351.

Footnote

[1]This test is also known as the Scholastic Aptitude Test or the Scholastic Achievement Test.

[2]It was discovered after the data were collected that middle response options (i.e., choices *b* or *c*) were correct roughly two thirds of the time—a rate that was much higher than would have been expected if the position of the correct options for the items had been designated randomly. It was also found that during rapid-guessing responding examinees tended to choose middle categories more often than would have been expected under random responding. This potential confounding was removed by identifying (using stratified random selection) a subset of 38 ILT items that was balanced regarding the position of the correct option.

Table 1

Correlations Among Test Performance, Effort Measures, Person Fit, Test Time, and SAT Scores

Measure	1	2	3	4	5	6	7
1. Test Performance	--						
2. Self-Reported Effort	.34**	--					
3. Response Time Effort	.54**	.25**	--				
4. Person Fit	–.22**	–.17**	–.42**	--			
5. Total Test Time	.26**	.22**	.44**	–.15*	--		
6. SAT-Verbal	.50**	.14*	.06	–.08	–.18**	--	
7. SAT-Quantitative	.34**	.01	–.02	–.01	–.11	.44**	--

Note. $N = 472$.

*$p < .01$ **$p < .001$

Table 2

Means, Standard Deviations, and One-Way Analyses of Variance for Three Response Time Effort Groups on Four Variables

	Response Time Effort Group						ANOVA		
	RTE less than .80		*RTE* between .80 - .90		*RTE* greater than .90				
Measure	*M*	*SD*	*M*	*SD*	*M*	*SD*	*F*(2,469)	*p*	ω^2
ILT Test Performance	37.87	9.26	47.33	7.82	63.26	9.48	87.22	<.001	.26
Self-Reported Effort	12.50	2.95	13.33	3.42	16.38	3.55	16.41	<.001	.06
SAT-Verbal	554.00	67.85	574.00	61.50	575.77	64.76	1.08	.36	.00
SAT-Quantitative	589.50	77.22	582.67	64.20	582.72	65.35	0.10	.66	.00

Note. The sample sizes for the three *RTE* groups were 20, 15, and 437, respectively.

Table 3

Test Performance Relative to Random Responding for Rapid-Guessing Responses and Solution Responses on a Subset of 38 ILT Items

Response Behavior	*M*	*SD*	*df*	*t*	*p value*	*Effect Size*
Rapid-Guessing Responses	28.74	21.33	37	1.16	.25	.19
Solution Response	61.66	27.66	37	8.23	<.001	1.33

Note. Test performance scores had an expected value under random responding of 24.73. The effect size index was given by $(M_{\text{obtained}} - M_{\text{chance}})/SD_{\text{obtained}}$.

Table 4

The Effects of Motivation Filtering for the ILT Data Using Response Time Effort and Self-Reported Effort as Filters

Data Analyzed	*n*	*M*	*SD*	Coefficient Alpha	SEM	Correlation Between Test Score and SAT-V	Mean SAT-V Score
			Response Time Effort (*RTE*)				
All Examinees	472	61.68	11.02	.83	4.54	.50	574.79
Examinees with *RTE* ≥ .60	462	62.32	10.17	.80	4.56	.54	574.98
Examinees with *RTE* ≥ .70	454	62.66	9.89	.79	4.55	.55	575.31
Examinees with *RTE* ≥ .80	452	62.73	9.85	.79	4.54	.55	575.71
Examinees with *RTE* ≥ .90	437	63.26	9.48	.77	4.53	.58	575.77
			Self-Reported Effort (SRE)				
All Examinees	472	61.68	11.02	.83	4.54	.50	574.79
Examinees with SRE ≥ 10	460	61.82	10.90	.82	4.57	.53	573.93
Examinees with SRE ≥ 11	446	62.02	10.80	.82	4.56	.54	574.13
Examinees with SRE ≥ 12	423	62.53	10.49	.81	4.55	.53	575.46
Examinees with SRE ≥ 13	397	62.67	10.36	.81	4.54	.54	576.50

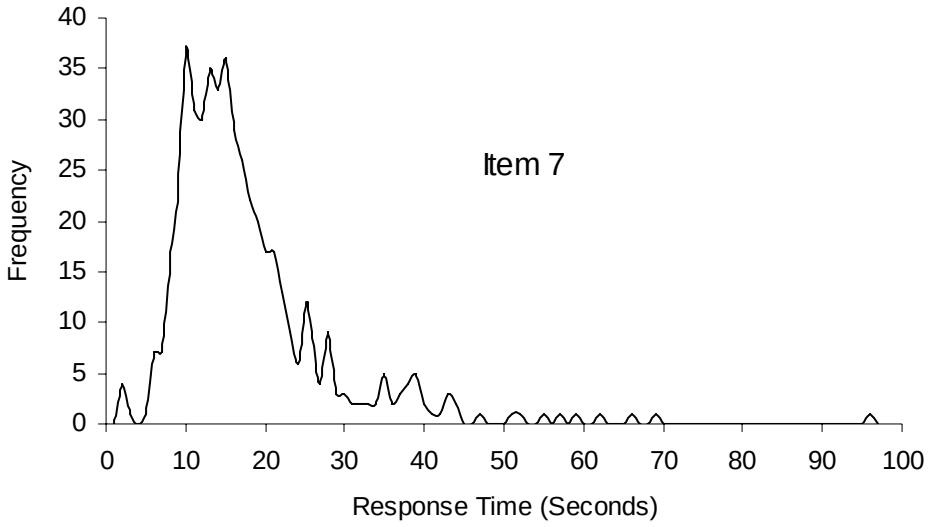
Figure Captions

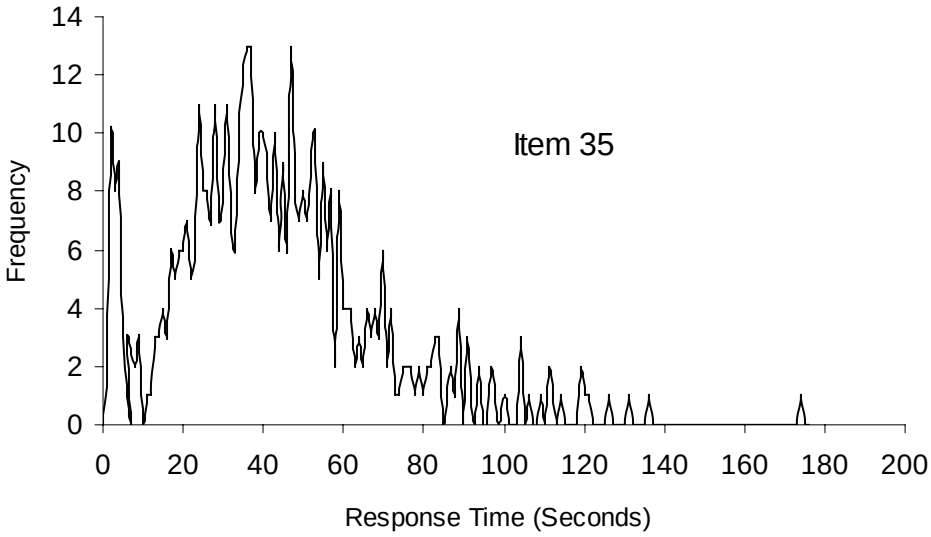
Figure 1. Response time frequency polygons for two ILT items.

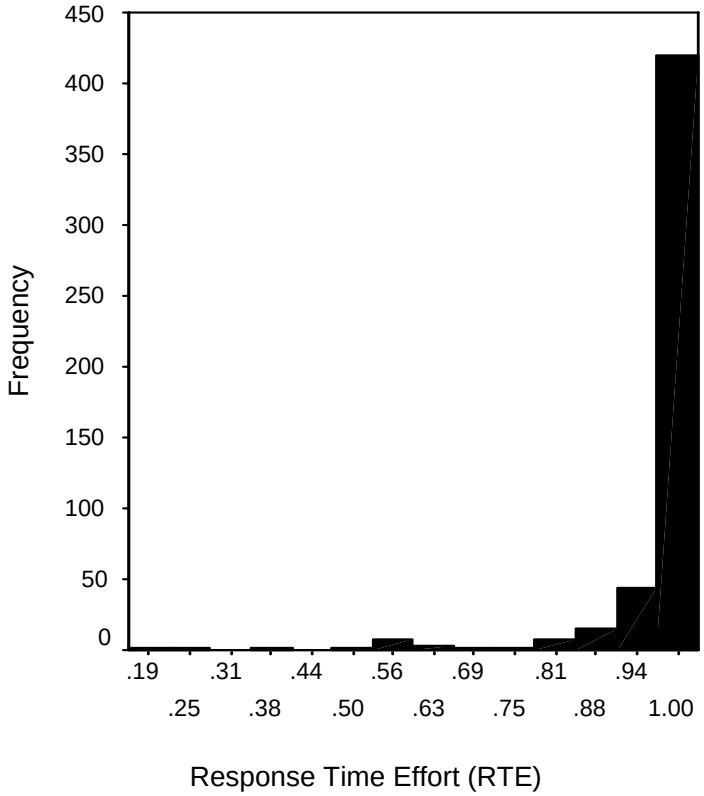
Figure 2. Histogram of Response Time Effort (*RTE*) scores.

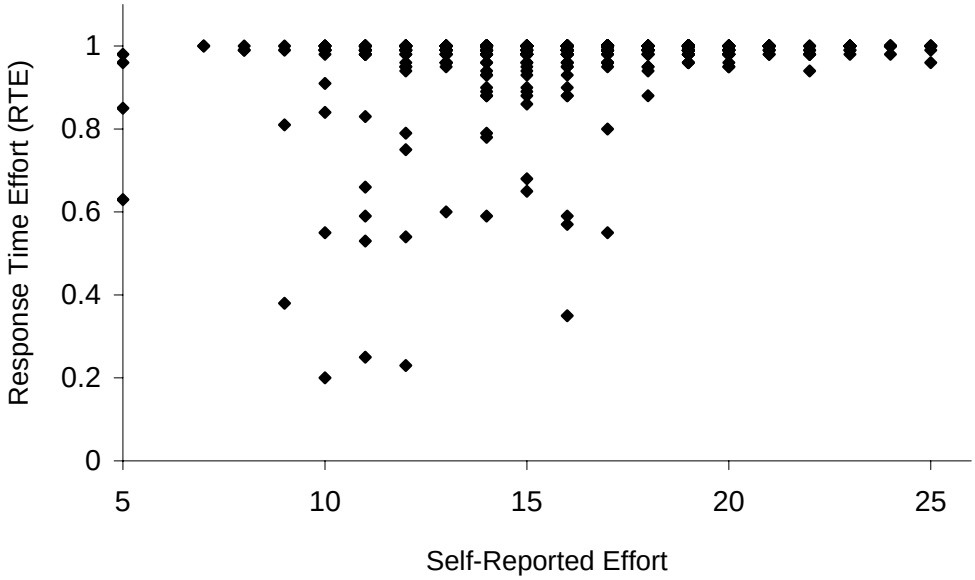
Figure 3. Scatter plot of *RTE* scores by self-reported effort.

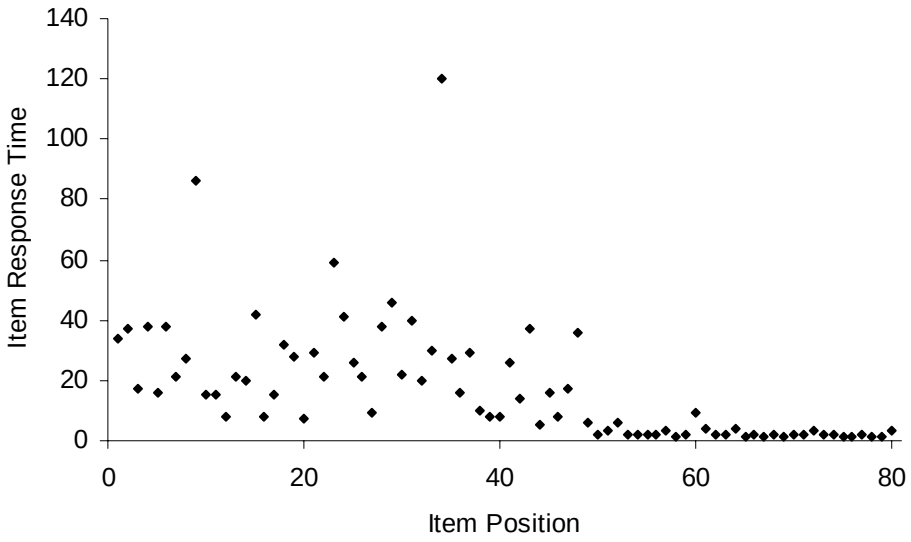
Figure 4. Distribution of item response times by item position for an examinee who appeared to switch response strategies in the middle of the test.


Item 7
Frequency
Response Time (Seconds)
0
5
10
15
20
25
30
35
40
0
10
20
30
40
50
60
70
80
90
100


Item 35
Frequency
Response Time (Seconds)
0
2
4
6
8
10
12
14
0
20
40
60
80
100
120
140
160
180
200


450
400
350
300
250
200
150
100
50
0
Frequency
.19
.25
.31
.38
.44
.50
.56
.63
.69
.75
.81
.88
.94
1.00
Response Time Effort (RTE)


1
0.8
0.6
0.4
0.2
0
Response Time Effort (RTE)
5
10
15
20
25
Self-Reported Effort


140
120
100
80
60
40
20
0
Item Response Time
0
20
40
60
80
Item Position