An Investigation of the Relationship between Time of Testing and Test-Taking Effort[1]

Steven L. Wise, Lingling Ma, G. Gage Kingsbury, Carl Hauser

Northwest Evaluation Association

This study investigated the relationships between when a test is administered and the amount of test-taking effort exhibited by examinees. Three time-related variables were investigated: the time of year the test was administered, the day of the week the test event occurred, and the time of day that the test event occurred. Mean effort did not appear to vary across either time of year nor day of week. There was, however, a clear time of day effect, with mean effort decreasing throughout the day. In addition, mean effort was found to decreases across grade, and male examinees were found to exhibit lower mean effort than female examinees.

Whenever we administer an achievement test to a group of examinees, our goal is straightforward—to measure those examinees' levels of proficiency in a content area of interest. In practice, however, obtaining valid scores from a test administration is more complicated than simply presenting items to the examinees under standardized conditions. Obtaining valid scores also requires motivated examinees, who behave effortfully throughout a test event. It is not uncommon, though, for some examinees in a group to exhibit test-taking behavior during a test that is *non-effortful*. For these examinees, the resulting scores are likely to underestimate what the examinees actually know and can do.

Individual score validity (ISV; Hauser, Kingsbury, & Wise, 2008; Kingsbury & Hauser, 2007; Wise, Kingsbury, & Hauser, 2009) refers to the degree to which a test score for an examinee is a reasonable indicator of the examinee's proficiency on the construct of interest. The largest threat to ISV comes from construct-irrelevant variance (CIV; Haladyna & Downing, 2004), and a major source of CIV is examinee effort (Wise & Kong, 2005). This is particularly true in the absence of score-based consequences for examinees (such as graduation, licensure, or entrance into college). Whenever a test score is going to be used to make inferences about individual students, it is therefore important to identify threats to ISV. This improves the inference process, because it informs us about which scores are trustworthy to make inferences about, and which are not.

Two primary approaches to measuring examinee effort have been described in the literature. The first uses Likert scales to obtain a self-report from each examinee after the achievement test has been completed and asking the amount of effort he or she devoted to the test. The second approach uses unobtrusive measurement of the time an examinee takes to answer each test item as a basis for a measure of effort expended on the test.

---

[1] Paper presented at the 2010 annual meeting of the National Council on Measurement in Education, Denver, CO.

While the self-report approach is easy to implement across a variety of measurement settings, Wise and Kong (2005) indentified several disadvantages. First, it is difficult to ascertain how truthfully examinees will respond regarding their levels of effort. Those who did not give good effort may not respond honestly because they fear consequences resulting from an admission of low effort (even consequences as mild as having to re-take the test). Moreover, examinees who gave good effort but felt that they did not do well on the test might under-report their effort, because they are predisposed to attribute their perceived failures to lack of effort (Pintrich & Schunk, 2002). The general point is that self-reports of effort are vulnerable to biasing factors, and it is difficult to ascertain the validity of an individual's effort rating.

An additional disadvantage of self-report measures is that they provide only an overall measure of effort during the test. There is evidence, however, that examinee effort can vary dramatically during the test (Wise & Kong, 2005). Hence, an effort measure that could provide more detailed information (i.e., at the level of effort to individual items) would be useful in understanding the pattern of effort exhibited by an examinee during a test event.

The second approach to measuring examinee effort, and the one used in this paper, is based on item response time. Building on the work of Schnipke & Scrams (1997, 2002), Wise and Kong (2005) showed that examinees who were not giving effort to multiple-choice items will tend to respond very rapidly (i.e., much faster than it would take an examinee to read the item). Such responses are said to reflect *rapid-guessing behavior*, with the remaining responses reflecting *solution behavior*. Operationally, this requires the specification of a time threshold for each item; responses occurring faster than the item's threshold are classified as rapid guesses, with the rest being deemed solution behaviors. Based on this conceptualization, Wise and Kong developed and validated a measure of examinee test-taking effort, termed *response time effort* (RTE). RTE represents the proportion of an examinee's responses that were classified as solution behavior during a particular test event.

Distinguishing between solution behavior and rapid-guessing behavior has the advantages of being (a) based on examinee behavior, and therefore is not subject to the potential confounding factors that might affect a self-report measure and (b) able to provide effort information down to the level of the individual item response, which permits an identification of effort changes during a test event. Its primary disadvantage is that it requires response times to be captured as part of the test administration process. Currently, computer-based tests (CBT) provide the most accessible means of providing this capability.

Research on examinee motivation and test-taking effort has indicated that the degree of rapid-guessing behavior can have a material effect on test performance (Cronin, Bontempo, Kingsbury, Hauser, McCall, & Houser, 2005; Wise, Bhola, & Yang, 2006; Wise & DeMars, 2006, 2010), thus documenting the threat to ISV that is posed by low test-taking effort. This has led to research directed at better understanding the circumstances under which rapid-guessing behavior is most likely to occur.

Correlates of Examinee Effort

It is useful in the present discussion to consider the model of test-taking effort proposed by Wise and Smith (in press). In this model, a test can be considered a series of encounters, in which the *examinee* encounters an *item* in a particular *context*. The effort that the examinee expends on the item is influenced by all three elements. Examinees may differ on characteristics such as age, gender, or proficiency level. Items may differ in difficulty, the amount of reading required, the number of distractors, and a variety of other characteristics. In addition, the context in which an item is administered can have a strong effect on the effort expended: an examinee may be more likely to give effort to a high-stakes test than to a low-stakes test, an examinee might give greater effort to the 1st item of a 200-item test than to the 200th item, or a test item might receive more effort in a quiet administration setting than in a noisy one.

Two examinee characteristics have been identified that are related to test-taking effort. The most consistent finding has been a gender effect; males have a tendency to exhibit lower mean levels of effort (Eklöf, 2007; Wise & DeMars, 2010; Wise, Kingsbury, Thomason, & Kong, 2004; Wise, Pastor, & Kong, 2009). In addition, one study found that lower-ability college students (as measured by SAT scores) were more likely than higher-ability students to exhibit non-effortful test-taking behaviors (Wise, Pastor, et al., 2009). This finding should be interpreted cautiously, however, as there have been several studies in the same measurement setting that did not find such a relationship (Kong, Wise, Harmes, & Yang, 2006; Wise et al., 2006; Wise & DeMars, 2006; Wise & Kong, 2005).

A number of item characteristics have been found to be related to effort. Items that require more reading by the examinee are more likely to receive rapid-guessing behavior (Wise, 2006; Wise, Pastor, et al., 2009). Wolf, Smith and Birnbaum (1995) found that effort on test items was influenced by their levels of mental taxation (i.e., how much mental effort was required by an examinee to reach a correct answer), with more taxing items receiving less effort. Wise, Pastor, et al. (2009) found that items presenting a greater number of response options received fewer rapid guesses, as did items containing a graphic.

Regarding the context in which an item is presented, the position at which an item occurs in a test has been related to the amount of effort it receives. Items occurring later in a test are more likely to receive rapid guessing (Setzer, Wise, & Allspach, 2008; Wise, 2006; Wise, Pastor, et al., 2009). Also, Wise, Owens, Yang, Weiss, Kissel, Kong, and Horst (2005) found significant differences among test session proctors in the average amount of effort exhibited by examinees in their sessions.

The current study investigated an additional type of context variable—the extent to which *when* a test is administered is related to test-taking effort. No previous research was found regarding how the time at which students are tested affects their levels of test-taking effort. There is a body of research, however, regarding the relationships between time of day and student achievement. Zagar and Bowers (1983) studied nonmedicated students who had

attention deficits, finding that they performed problem-solving tasks better in the morning. Barron, Henderson, and Spurgeon (1994) found that reading achievement showed more gains for those instructed in the morning than in the afternoon. Sjosten-Bell (2005) found that third graders administered math problems performed higher in the morning, followed by mid-morning, and then the afternoon. Other studies, however, have found conflicting results. Davis (1988) found that 8th graders who took English in the afternoon exhibited higher achievement than students who took English in the morning. A study of long-term memory retrieval found that the ability to retrieve stored information was stronger later in the day (Folkard, Monk, Bradbury, & Rosenthall, 1977).

The purpose of the current study was to assess several relationships between time of testing and the effort examinees give to achievement tests. Specifically, three time-related research questions were investigated:

1. To what extent does test-taking effort vary across the time of an academic year at which a test is administered?
2. To what extent does test-taking effort vary across the day of the week on which a test is administered?
3. To what extent does test-taking effort vary across the time of day at which a test is administered?

The study also investigated how grade and gender influenced effort as a function of the time of test administration.

Method

This research study was based on the data from an adaptive testing program in math and reading in grades three through nine. The tests were administered in a number of school districts in a single U.S. state, with each district deciding how often to test and when testing was to occur. The data analyzed in this study represent the aggregation of the adaptive test data collected in that state. The test events in each grade occurred at varying times of the academic year, on varying days of the week, and at varying times of the day. Testing occurred at three general time periods during the year (fall, winter and spring testing terms) and on five days of the week (Monday through Friday). For purposes of this study, the school day was divided into eight one-hour blocks (beginning from 7:00 a.m. to 2:00 p.m.). A given test event was assigned to the time of day variable according to the time block during which the student started his or her test.

Measures

Measures of Academic Progress (MAP). All of the tests administered were part of Northwest Evaluation Association's (NWEA's) MAP testing system. MAP tests are untimed interim computerized adaptive tests (CATs), with the tests in math being generally 50 items in length, while those in reading are generally 40 items in length[2]. MAP proficiency estimates are

[2] Occasionally, the test length was slightly longer for a given test, when a few additional items were administered to oversample the content goal categories that had higher standard errors of measurement.

expressed as scale (RIT) scores on a common scale that allows growth to be assessed as students are tested at different times. The standard errors of the RIT scores in math are about 3.0 points, while those in reading are about 3.2 points.

Response Time Effort (RTE). Each test event consisted of a series of examinee-item encounters. For each of these encounters, the testing software recorded the time that elapsed between the display of the item and when the examinee entered a response. These item response times provided the basis for evaluating examinee effort. Any response occurring within three seconds was classified as rapid-guessing behavior, with the remaining responses classified as solution behavior[3]. RTE scores (Wise & Kong, 2005), which are the proportion of the item responses that were classified as solution behaviors, were computed for each test event. RTE scores can range between 0.0 and 1.0, with higher scores indicating higher degrees of examinee effort.

Subjects

All of the MAP testing events in grades 3-9 from the fall, winter, and spring terms of the 2008-2009 academic year from a single state were retrieved from NWEA's *Growth Research Database*. This generated a total of 861,999 test events in math and 876,569 records in reading. A number of test events that fell out of acceptable ranges were excluded from this study. For instance, those test events occurring on weekend or starting earlier than 7 a.m. or later than 3 p.m. were excluded[4]. In total, there were 841,680 test events in math and 848,169 in reading included in this study. Because most students were tested two or three times during the year, however, the total number of examinees was lower (355,116 in math and 356,715 in reading).

Analyses

Because there have been no previous investigations of examinee effort as it relates to time of testing, the nature of this study was essentially exploratory. Consequently, the analyses consisted of descriptive statistics and visual inspection of graphic representations of data trends.

Results

Descriptive statistics for the MAP scores by content area, grade, and gender are shown in Table 1. In each content area, the mean RIT scores increased across grades, which is to be expected because (a) all of the RIT values in a given content area are linked to the same measurement scale and (b) students learn more in math and reading as they progress through the grades. There is, however, one peculiar aspect to the RIT means. In both math and reading, the grade 9

[3] Because the MAP tests use item pools containing thousands of items, it was not convenient to establish a time threshold for each item in identifying rapid guessing. Instead, a common three-second threshold was used as an expedient. Investigations with items from the MAP item pools have indicated that this is a suitable, albeit conservative time threshold value to use with MAP items.

[4] These test events were interpreted as indications that that computer's clock was incorrect.

mean RIT is slightly lower than that found for grade 8. There are at least two factors that may cause this anomaly. First, the number of students taking MAP in grade 9 is substantially smaller than that for grade 8, which may indicate that select students are being tested in grade 9. One common practice is to continue to test students who are meaningfully different and less proficient (grade 9 students, here) while discontinuing testing of students who have demonstrated proficiency (grade 8 students, here). A second factor is that effort may be slightly lower in grade 9. Evidence of this factor is found by examining the mean RTE scores, which are also shown in Table 1. These means are generally very high, which is reassuring from a test giver's perspective. Beginning at grade 6, however, the means begin to drop, with a more pronounced drop at grade 9. The finding that the mean RIT anomaly occurs in conjunction with the largest decrease in mean RTE suggests that the decrease in examinee effort at grade 9 may have affected their RIT scores to the extent that their mean fell below that from grade 8.

Table 1. Descriptive Statistics for MAP in Math and Reading

			RIT Scores		Mean RTE		
Test	Grade	N	Mean	SD	All Students	Females	Males
Math	3	136,427	195.98	13.38	.997	.998	.996
	4	132,422	206.74	14.05	.997	.998	.996
	5	130,356	216.29	15.59	.997	.998	.996
	6	122,977	221.42	15.97	.994	.996	.991
	7	121,582	227.32	17.26	.992	.995	.988
	8	120,876	232.33	17.85	.991	.995	.987
	9	77,040	231.24	18.26	.985	.991	.979
Reading	3	138,016	192.94	15.91	.995	.996	.993
	4	134,535	201.82	15.39	.994	.996	.992
	5	131,818	208.61	14.98	.995	.997	.993
	6	122,963	212.39	15.39	.988	.993	.984
	7	121,273	216.01	15.60	.984	.991	.979
	8	120,375	219.24	15.52	.983	.990	.976
	9	79,189	218.60	16.20	.971	.982	.961

Table 1 also shows that males showed lower mean RTE scores than females at all grades. This finding is consistent with the gender differences in effort that have been found in previous research (e.g., Eklöf, 2007).

Figure 1 shows a histogram of the RTE scores for the MAP testing in math. It is clearly negatively skewed, most examinees exhibiting the desired value of 1.0. It is also apparent, however, that a few examinees had RTE values that were quite low; for these examinees, observed test performance underestimated their actual levels of proficiency. The corresponding graph for reading (not shown) was highly similar to the graph for math.

Figure 1. Histogram of RTE Scores for all MAP Test Events in Reading

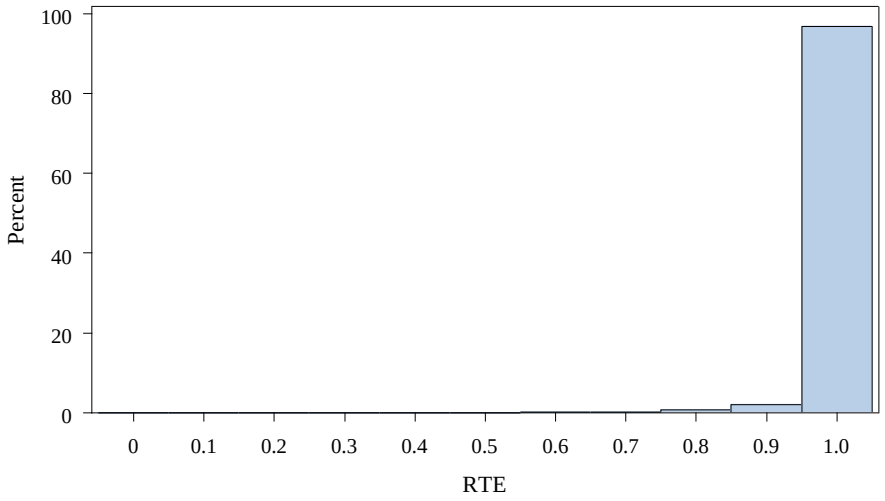


Table 2. Mean RTE, by Time Variables

		Math		Reading	
Time Variable	Value	Mean RTE	N	Mean RTE	N
Time of Year	Fall	.992	336,235	.987	337,041
	Winter	.994	194,667	.989	199,663
	Spring	.995	310,778	.990	311,465
Day of Week	Monday	.994	164,684	.988	185,268
	Tuesday	.994	182,925	.988	202,245
	Wednesday	.994	177,199	.988	180,526
	Thursday	.994	184,837	.988	165,568
	Friday	.994	132,035	.988	114,562
Time of Day	7:00 a.m.	.997	100,909	.994	99,651
	8:00 a.m.	.995	164,277	.990	169,133
	9:00 a.m.	.995	168,217	.990	171,513
	10:00 a.m.	.994	135,177	.989	135,631
	11:00 a.m.	.993	91,268	.987	89,118
	12:00 noon	.992	86,218	.985	86,099
	1:00 p.m.	.990	68,018	.982	68,992
	2:00 p.m.	.986	27,596	.974	28,032

Note: The time of day values refer to the hour-long block beginning at the labeled time.

The results of the primary analyses that informed the three research questions of this study are found in Table 2. Regarding time of year, there was little variation in mean RTE scores, though it appeared that rapid-guessing behavior was slightly more likely to occur in the fall testing term. In addition, there was no indication of a day of the week effect in mean RTE. For time of day, however, there was a clear trend in both math and reading. As the school day progressed, mean RTE scores showed a gradual decline that appeared to accelerate in the afternoon. Thus, of the time-related variables, only the time of day that testing occurred appeared to be meaningfully related to examinee effort. In the remaining analyses, we will focus on exploring further the time of day effect.

Table 3 shows the gender differences in mean RTE, by time of day and content area. For both math and reading, as the day progressed, the mean difference between male and females increased. This pattern is illustrated in Figure 2 (for math) and Figure 3 (for reading). Comparison of these figures suggests that the gender effect was more pronounced, and increased more during the day, in reading.

Table 3. Mean RTE, by Time of Day and Gender

	Mean Math RTE				Mean Reading RTE			
Time of Day	Females	N	Males	N	Females	N	Males	N
7:00 a.m.	.998	50,032	.996	50,877	.996	48,992	.992	50,659
8:00 a.m.	.997	79,670	.993	84,607	.994	81,986	.987	87,147
9:00 a.m.	.997	82,144	.993	86,073	.994	83,477	.986	88,036
10:00 a.m.	.996	66,167	.992	69,010	.993	66,719	.985	68,912
11:00 a.m.	.996	45,084	.990	46,184	.993	43,137	.982	45,981
12:00 noon	.996	41,651	.989	44,567	.991	42,282	.980	43,817
1:00 p.m.	.994	33,342	.986	34,676	.988	33,719	.975	35,273
2:00 p.m.	.992	13,330	.981	14,266	.984	13,828	.964	14,204

Figure 2. Mean Math RTE, by Time of Day and Gender

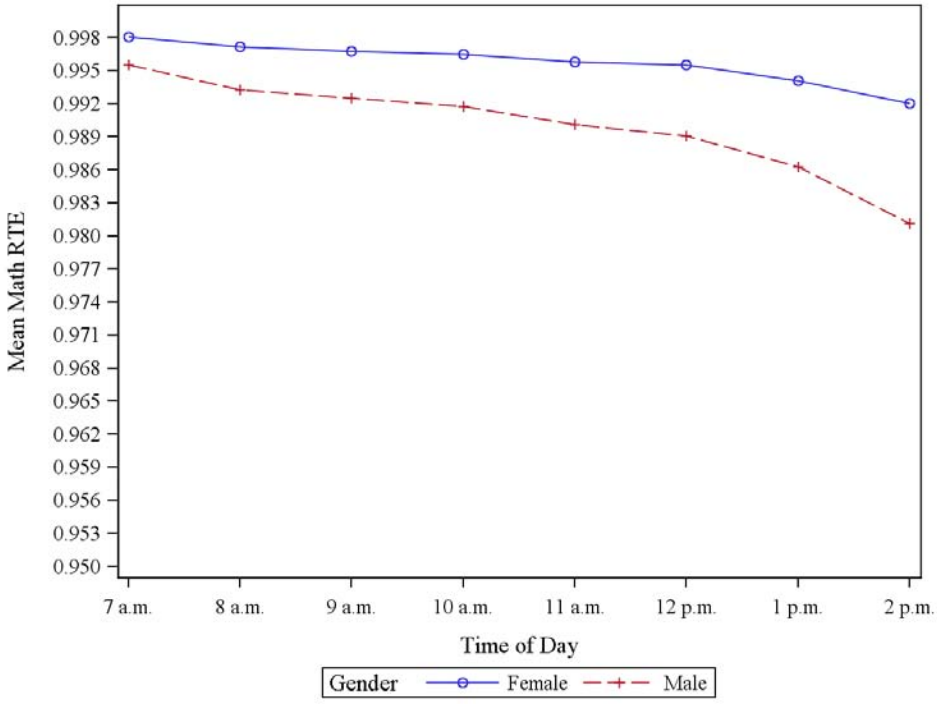
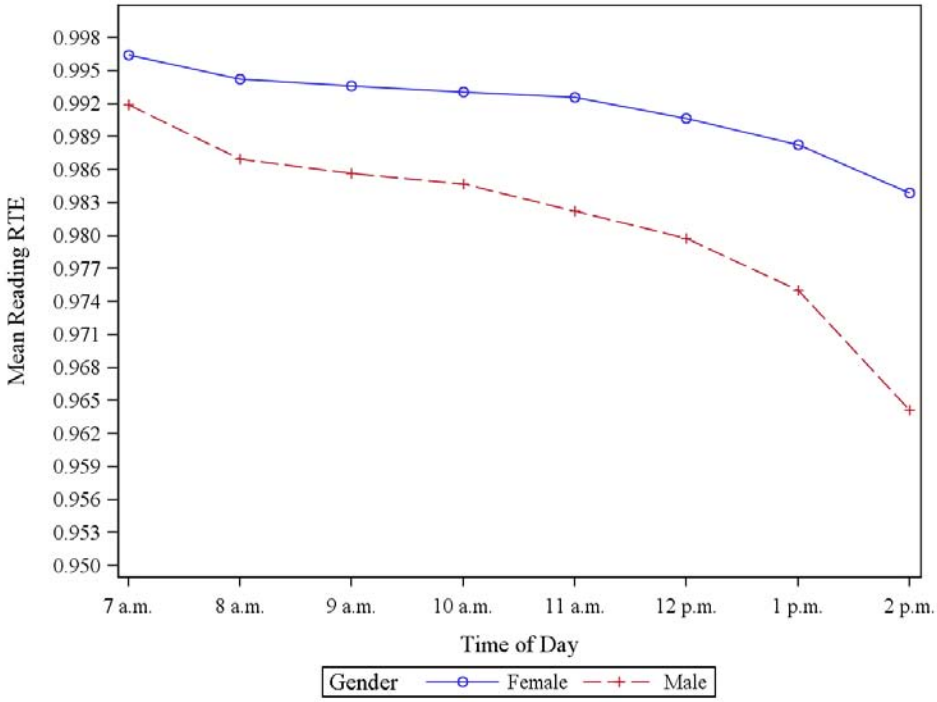


Figure 3. Mean Reading RTE, by Time of Day and Gender



To provide some perspective regarding when this study's test events actually occurred, Table 4 shows, for each grade, the numbers of test events in math that occurred at each hour of the day. Table 5 shows the corresponding numbers for reading. For each content area, most of the test events occurred during the early to mid-morning, with a marked decrease occurring in the afternoon (particularly in grades 3-5).

Table 4. Number of Test Events in Math, by Time of Day and Grade

	Grade						
Time of Day	3	4	5	6	7	8	9
7:00 a.m.	21,851	18,857	19,636	13,536	11,254	9,325	6,450
8:00 a.m.	29,276	25,386	27,119	24,493	21,264	20,519	16,220
9:00 a.m.	30,108	31,585	28,733	22,855	21,516	20,774	12,646
10:00 a.m.	22,030	24,377	24,956	16,949	17,301	18,899	10,665
11:00 a.m.	12,704	12,330	13,165	13,765	13,705	16,022	9,577
12:00 noon	14,672	13,357	10,815	14,046	12,500	12,760	8,068
1:00 p.m.	5,513	6,076	5,251	12,602	15,593	14,615	8,368
2:00 p.m.	273	454	681	4,731	8,449	7,962	5,046

Table 5. Number of Test Events in Reading, by Time of Day and Grade

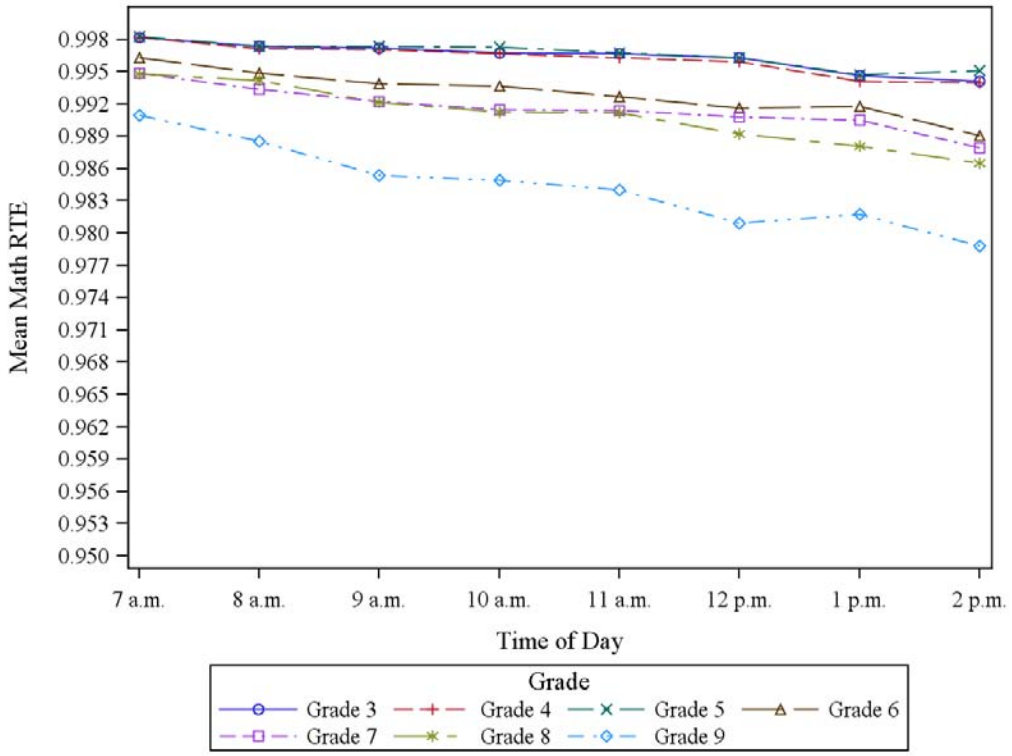
	Grade						
Time of Day	3	4	5	6	7	8	9
7:00 a.m.	22,563	18,953	19,887	12,922	10,814	9,050	5,462
8:00 a.m.	30,454	27,912	28,785	23,615	21,074	20,423	16,870
9:00 a.m.	31,329	31,497	29,135	23,730	21,584	20,535	13,703
10:00 a.m.	21,817	24,470	24,532	17,885	17,306	18,649	10,972
11:00 a.m.	12,138	12,280	12,720	13,467	13,173	16,220	9,120
12:00 noon	13,979	13,253	10,788	14,095	12,808	12,245	8,931
1:00 p.m.	5,493	5,669	5,509	12,608	15,388	15,247	9,078
2:00 p.m.	243	501	462	4,641	9,126	8,006	5,053

Table 6 shows the mean RTE scores in math, broken down by time of day and grade. Figure 4 illustrates the time of day trends for each grade. The results show that for grades 3-5, mean RTE remained high throughout the day, with a very slight decrease. For grades 6-8, mean RTE started a bit lower and decreased a bit more during the day. In grade 9, the mean RTE started even lower and diminished more rapidly throughout the day, dropping below .98 at the 2:00 hour.

Table 6. Mean Math RTE, by Time of Day and Grade

	Grade						
Time of Day	3	4	5	6	7	8	9
7:00 a.m.	.998	.998	.998	.996	.995	.995	.991
8:00 a.m.	.997	.997	.997	.995	.993	.994	.989
9:00 a.m.	.997	.997	.997	.994	.992	.992	.985
10:00 a.m.	.997	.997	.997	.994	.991	.991	.985
11:00 a.m.	.997	.996	.997	.993	.991	.991	.984
12:00 noon	.996	.996	.996	.992	.991	.989	.981
1:00 p.m.	.995	.994	.995	.992	.991	.988	.982
2:00 p.m.	.994	.994	.995	.989	.988	.987	.979

Figure 4. Mean Math RTE, by Time of Day and Grade

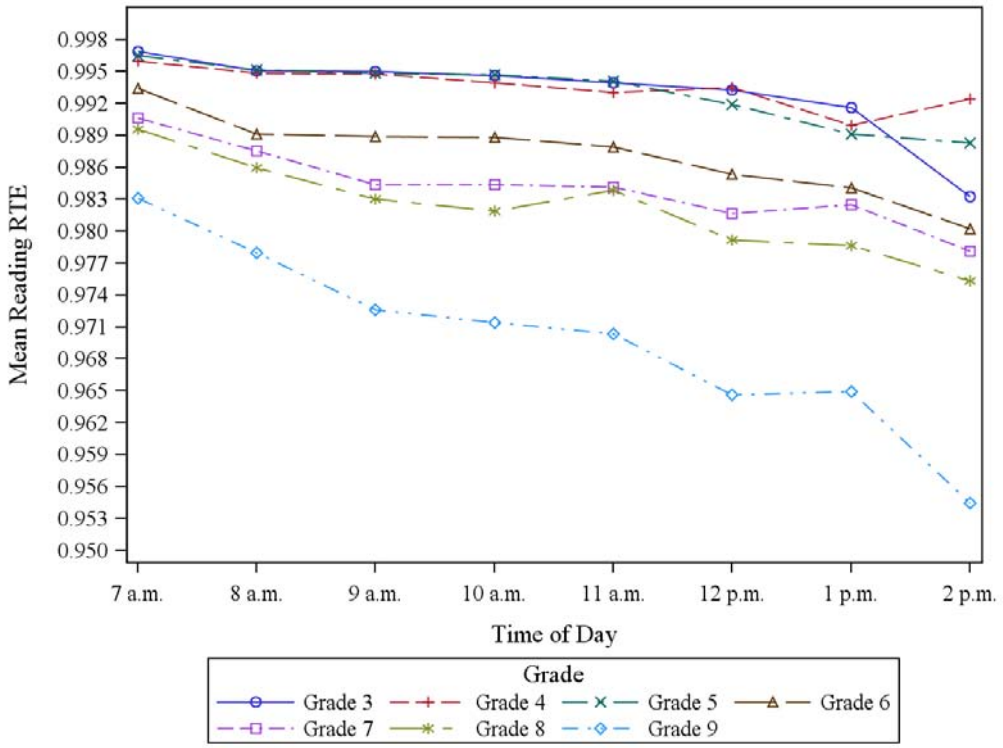


The corresponding results for reading are found in Table 7 and Figure 5, respectively. The pattern of results for reading resembles that found for math, except that it was more pronounced, both in starting values and decreases during the day. In particular, the grade 9 mean RTE score dropped precipitously, falling below .96 at the 2:00 hour. The general finding that the results for reading indicated lower mean effort than for math was similar to that found for the gender differences from Table 3.

Table 7. Mean Reading RTE, by Time of Day and Grade

	Grade						
Time of Day	3	4	5	6	7	8	9
7:00 a.m.	.997	.996	.997	.993	.991	.990	.983
8:00 a.m.	.995	.995	.995	.989	.988	.986	.978
9:00 a.m.	.995	.995	.995	.989	.984	.983	.973
10:00 a.m.	.995	.994	.995	.989	.984	.982	.971
11:00 a.m.	.994	.993	.994	.988	.984	.984	.970
12:00 noon	.993	.994	.992	.985	.982	.979	.965
1:00 p.m.	.992	.990	.989	.984	.983	.979	.965
2:00 p.m.	.983	.992	.988	.980	.978	.975	.954

Figure 5. Mean Reading RIT, by Time of Day and Grade



Although assessment of the mean RTE scores is useful in describing the general relationship between time of day and effort, it does not fully convey the extent to which ISV is sufficiently threatened to render a given test score untrustworthy. Wise, Kingsbury, et al. (2009) developed a set of indicator flags for low examinee effort—one of which was based on the overall RTE score for a test event. They examined an RTE flagging criterion of .85, which means that any test score associated with an RTE value less than .85 would be flagged as invalid due to low effort. Table 8 shows the percentages of test events in the current study that triggered this effort flag, broken down by grade, time of day, and test content. Three basic patterns are apparent from inspection of this table. First, at every grade, there was a clear increase throughout the day in the percentages of flagged test events. Second, at each time of day, there was a substantial increase in the percentages of flagged test events across grades. Third, the percentages were much higher in reading than in math.

Table 8. Percentages of Test Events that Triggered an Effort Flag (RTE < .85)

		Grade						
Test	Time of Day	3	4	5	6	7	8	9
Math	7:00 a.m.	0.3	0.3	0.2	0.7	1.1	1.1	1.7
	8:00 a.m.	0.5	0.5	0.5	1.1	1.4	1.1	2.4
	9:00 a.m.	0.4	0.5	0.5	1.2	1.6	1.6	3.3
	10:00 a.m.	0.6	0.7	0.5	1.3	1.8	1.8	3.3
	11:00 a.m.	0.5	0.6	0.5	1.6	1.8	1.8	3.6
	12:00 noon	0.6	0.8	0.6	1.8	2.0	2.3	4.2
	1:00 p.m.	0.8	1.2	1.0	1.8	2.0	2.6	4.1
	2:00 p.m.	1.1	1.1	0.9	2.4	2.7	3.0	4.6
Reading	7:00 a.m.	0.5	0.9	0.8	1.5	2.2	2.4	4.1
	8:00 a.m.	1.1	1.1	1.0	2.5	3.0	3.3	5.4
	9:00 a.m.	1.1	1.1	1.1	2.7	3.9	4.2	6.8
	10:00 a.m.	1.1	1.4	1.1	2.6	3.8	4.5	7.2
	11:00 a.m.	1.3	1.5	1.2	3.0	3.9	3.9	7.6
	12:00 noon	1.5	1.4	1.8	3.5	4.5	5.2	8.5
	1:00 p.m.	1.9	2.5	2.7	3.9	4.1	5.3	9.1
	2:00 p.m.	3.7	1.2	3.2	4.8	5.6	5.9	11.6

Test performance information for each combination of time of day and grade are found in Table 9. At each time of day, the mean RIT scores increase across grade, with some time periods exhibiting the same grade 9 score reversal that was noted in Table 1. In addition, at each grade there was a general decrease in mean RIT scores throughout the day.

Table 9. Mean Math and Reading RIT Scores, by Time of Day and Grade

		Grade						
Test	Time of Day	3	4	5	6	7	8	9
Math	7:00 a.m.	198	208	218	223	230	234	234
	8:00 a.m.	196	206	216	221	227	233	231
	9:00 a.m.	196	207	216	221	226	232	231
	10:00 a.m.	196	207	216	222	227	233	230
	11:00 a.m.	196	206	216	222	228	232	231
	12:00 noon	196	206	215	221	227	232	231
	1:00 p.m.	193	205	214	222	227	232	231
	2:00 p.m.	192	201	217	221	226	232	231
Reading	7:00 a.m.	195	203	210	213	217	220	220
	8:00 a.m.	193	202	209	212	216	219	219
	9:00 a.m.	193	202	209	213	216	218	219
	10:00 a.m.	192	202	208	213	216	219	219
	11:00 a.m.	192	201	209	213	216	221	218
	12:00 noon	192	201	207	212	216	219	219
	1:00 p.m.	188	200	206	212	216	219	218
	2:00 p.m.	189	198	208	212	216	219	217

It is clear that rapid-guessing behavior has the effect of negatively biasing test scores. But to what extent are RIT scores affected by a given amount of rapid guessing? To better understand this, we simulated the impact of varying amounts of rapid guessing in a CAT.

All grade 5 math test records (n = 8682) that were free of rapid guesses were selected for the simulation. The item responses in the second half of each test record were systematically replaced with a "response" simulated as a random guess. For each test record, response replacement began at the last item in the record (item 50) and receded back one item at a time to item 26. After each step, all tests were rescored using the maximum likelihood estimation method normally used to compute RIT scores.

The results of this simulation, which are shown in Table 10, indicates that an examinee who rapid guessed to 8 items (the smallest number that would have resulted in the test event being effort flagged) would be expected to receive a RIT score 2.11 points lower than they would have under full effort. This represents a RIT score underestimate of roughly two-thirds of a standard error of proficiency estimation. Additional rapid guessing would lead to greater

proficiency underestimation; for example, an examinee who had 20 rapid guesses would be expected to have observed a RIT score more than five points lower.

Table 10. Expected Impact of Rapid Guesses on Math RIT Scores

Number of Rapid Guesses	RTE	Expected Decrease in RIT
1	.98	-0.26
2	.96	-0.52
3	.94	-0.77
4	.92	-1.05
5	.90	-1.32
6	.88	-1.58
7	.86	-1.84
8	.84	-2.11
9	.82	-2.36
10	.80	-2.64
11	.78	-2.90
12	.76	-3.18
13	.74	-3.43
14	.72	-3.67
15	.70	-3.95
16	.68	-4.25
17	.66	-4.49
18	.64	-4.83
19	.62	-5.13
20	.60	-5.36
21	.58	-5.62
22	.56	-5.87
23	.54	-6.26
24	.52	-6.52
25	.50	-6.78

Discussion

The validity of an achievement test score depends upon two key elements—having both a well-constructed test and an examinee who is willing to expend the effort needed to show what he or she knows and can do. The test giver can exert a great deal control over the first element through careful test development procedures. The second element, in contrast, lies largely outside the control of the test giver whenever the test administered has few personal consequences that are important to examinees. In those instances, the test giver implicitly depends on test-taking motivation coming from internal examinee factors, such as examinee's perceived level of proficiency and amount of test preparation, competitiveness, expectations regarding how demanding the test will be, the degree to which the examinee wishes to please

the test giver, and the examinee's general sense of assessment citizenship (Wise & Smith, in press). Because examinees can vary in the degree to which they possess these factors, test-taking effort can also vary.

We generally observe good effort from the majority of examinees even when there are few, if any, personal consequences for test performance. Unfortunately, this is not always the case and the effort from some examinees will be low. In these cases it is desirable to be able to identify these instances of non-effortful test-taking, particularly when score-based inferences are to be made about individual examinees. Being able to recognize when *not* to trust a score is an important aspect of the validity of the inferential process.

The current investigation was undertaken to better understand the factors associated with non-effortful test-taking behavior. Three time-related variables were studied, and although effort differences were not observed related to the time of year or day of week, there was a clear indication of a time of day effect. At all grades, and for both content areas, rapid-guessing behavior occurred more often as the day progressed. The reasons for this are unclear. It should be noted that because random assignment to testing times was not used in this study, there are unknown confounding variables that might be operating. Hence, we should be cautious in interpreting the meaning of the time of day effect, and more research is needed to better understand it. Nevertheless, test givers should be aware that testing students in the morning appears to yield more valid scores. In fact, the test givers in the state investigated in this study may have been aware of this, because Tables 4 and 5 showed that the vast majority of student testing occurred in the morning.

This study also yielded additional results, some of which are consistent with what has been found in other studies, while others support what many believe. There was clear evidence that males gave less effort across all grades and time periods. This is consistent with previous research (Eklöf, 2007; Wise & DeMars, 2010; Wise et al., 2004; Wise, Pastor, et al. 2009). Reading tests received less effort than math tests, which is consistent with previous findings that items requiring more reading are more likely to receive rapid guesses (Wise, 2006; Wise, Pastor, et al, 2009). The finding that the occurrence of rapid guessing increased across grade, and showed a marked jump in grade 9, supports the common belief that adolescents are less compliant than younger students.

Note that the results of this study indicate the presence of influences on effort associated with all three elements of the Wise-Smith model: examinee (gender, grade), item (test content), and context (time of day). In addition, the findings illustrate the compounding effects of these influences. For example, a reading test given to a ninth grader at 2:00 yielded lower mean RTE scores than were found for all reading tests, all ninth graders, or all 2:00 tests.

One of the most important outcomes of this study, however, is that provides a practical perspective on the magnitude of the ISV threat posed by low student effort. The percentages of students triggering effort flags were not high, which is good news, as high percentages would call into doubt the validity and credibility of the testing program. Furthermore, the patterns in

the data provide clear indications where ISV is most likely to be threatened (i.e., male examinees, later grades, reading tests administered later in the school day). This information should help test givers strategize how and where to focus their efforts to improve ISV, and to monitor the impact of these strategies.

References

Barron, B. G., Henderson, M. V., & Spurgeon, R. (1994). Effects of time of day instruction on reading achievement of below grade readers. *Reading Improvement, 31*, 59-60.

Cronin, J., Bontempo, B., Kingsbury, G. G., Hauser, C., McCall, M., & Houser, R. (2005, April). *Using item response time and accuracy on a computer adaptive test to predict deflated estimates of performance*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.

Davis, Z. T. (1988). The effect of time of day of instruction on eighth grade students' English and mathematics achievement. *High School Journal, 71(2),* 78-80.

Eklöf, H. (2007). Test-taking motivation and mathematics performance in TIMSS 2003. *International Journal of Testing, 7,* 311-326.

Folkard, S., Monk, T. H., Bradbury, R., & Rosenthall, J. (1977). Time of day effects in schoolchildren's immediate and delayed recall of meaningful material. *British Journal of Psychology, 68*, 45-50.

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice, 23(1),* 17-27.

Hauser, C., Kingsbury, G, G., & Wise, S. L. (2008, March). *Individual validity: Adding a missing link.* Paper presented at the annual meeting of the American Educational Research Association, New York, NY.

Kingsbury, G. G., & Hauser, C. (2007, April). *Individual validity in the context of an adaptive test.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

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

NWEA (2009). *Technical Manual for Measures of Academic Progress and Measures of Academic Progress for Primary Grades.* Portland, OR: Northwest Evaluation Association.

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

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.

Setzer, C., Wise, S. L., & Allspach, J. R. (2008, April). *An investigation of examinee test-taking effort on the Major Field Test in business.* Paper presented at the annual meeting of the American Educational Research Association, New York.

Sjosten-Bell, W. (2005). *Influence of time-of-day on student performance on mathematical algorithms.* Unpublished masters thesis, Dominican University of California, San Rafael, CA.

Wise, S. L. (2006). An investigation of the differential effort received by items on a low-stakes, computer-based test. *Applied Measurement in Education, 19,* 95-114.

Wise, S. L., Bhola, D., & Yang, S. (2006). Taking the time to improve the validity of low-stakes tests: The effort-monitoring CBT. *Educational Measurement: Issues and Practice 25(2),* 21-30.

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

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

Wise, S. L., Kingsbury, G. G., & Hauser, C. (2009, April). *How do I know that this score is valid? The case for assessing individual score validity.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego..

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

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

Wise, S. L., Owens, K. M., Yang, S., Weiss, B., Kissel, H. L., Kong, X., & Horst, S. J. (2005, April). *An investigation of the effects of self-adapted testing on examinee effort and performance in a low-stakes achievement test*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal.

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

Wise, S. L., & Smith, L. F. (in press). A model of examinee test-taking effort. In J. Bovaird (Ed.) *Contemporary issues in high stakes testing.* Washington, DC: American Psychological Association.

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.

Zagar, R., & Bowers, N. D. (1983). The effect of time of day on problem solving and classroom behavior. *Psychology in the Schools, 20*, 337-345.