An Investigation of the Effects of Self-Adapted Testing on Examinee Effort and Performance in a Low-Stakes Achievement Test

Steven L. Wise, Kara M. Owens, Sheng-Ta Yang, Brandi Weiss,

Hilary L. Kissel, Xiaojing Kong and Sonia J. Horst

James Madison University

Paper presented at annual meeting of the National Council on Measurement in Education, Montreal, April, 2005.

An Investigation of the Effects of Self-Adapted Testing on Examinee Effort and Performance in a Low-Stakes Achievement Test

There are a variety of situations in which low-stakes achievement tests—which are defined as those having few or no consequences for examinee performance—are used in applied measurement. A problem inherent in such testing is that we often cannot assume that all examinees give their best effort to their test, which suggests that the test scores of some examinees will underestimate their actual levels of proficiency. This discrepancy between demonstrated and actual proficiency can be sizable in magnitude; previous research on the relationship between test-taking motivation and performance indicates that, on average, less motivated examinees tend to score greater than one half standard deviation below their more motivated peers (Wise & DeMars, 2005). In this sense, low test-taking effort introduces construct-irrelevant variance that has a systematic negative effect on test performance and constitutes a threat to test score validity (Haladyna & Downing, 2004).

Wise and Kong (in press) noted three commonly occurring situations in which examinee effort should be a concern to measurement professionals. First, there are many low-stakes assessment programs that have serious potential consequences for institutions but few personal consequences for individual examinees. There has been a growing emphasis on low-stakes assessment testing as a means to hold schools accountable for the quality of education they provide their students. The recent No Child Left Behind (NCLB) legislation has expanded the presence of such accountability testing in K-12 institutions. In higher education, low-stakes assessment testing has increasingly been used to hold publicly-funded institutions accountable for expenditures of taxpayer dollars. Second, high-stakes testing programs sometimes administer low-stakes tests, particularly in the early stages of the program. It is not uncommon, for example, 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. Finally, a great deal of measurement research is conducted in low-stakes settings. At colleges and universities, for example, measurement research frequently uses volunteers or students from subject pools, for whom there are typically minimal consequences associated with test performance. Because such behavior is difficult to reliably detect, there are typically no penalties imposed on subjects who show up to participate in a study, but do not try very hard to do their best when administered an achievement test.

It appears, however, that many (if not most) examinees tend to devote considerable effort when given achievement tests, even when there are no personal consequences associated with test performance. Although the reasons for this effort have not been well studied, there are probably several factors operating. First, for some examinees, trying hard on achievement tests is an acquired habit that has been reinforced through higher-stakes testing experiences (e.g., classroom tests). Second, there may be a tendency to give good effort out of a sense of competitiveness with other students, even if feedback on test performance will not be provided. Third, the effort of some examinees may reflect a desire to please more powerful others (such as teachers), or a general expression of academic citizenship (e.g., being mindful that trying hard on the test benefits the examinee’s institution). Finally, some examinees may find tests intrinsically challenging, because they enjoy engaging in demanding cognitive tasks.

Despite the various factors that motivate many examinees to give good effort, there will frequently be some examinees who do not try to perform their best. For these examinees, the resulting test data will likely provide a biased picture of what the examinees know and can do. Consequently, along with the traditional test development and administration responsibilities, measurement professionals who administer low-stakes tests have the additional challenge of identifying and adopting testing practices that elicit high effort from a maximum proportion of examinees.

This study is focused on the identification of testing methods that promote examinee effort. Specifically, its purpose is to investigate the effectiveness of several forms of computer-based tests (CBTs) on the effort and test performance of examinees administered a low-stakes test.

Theoretical Framework

The theoretical basis for this study is based on contemporary models of both achievement motivation and aptitude. One useful basis for representing examinee behavior during test taking is provided by *expectancy-value* models of achievement motivation, which specify that a student’s achievement behavior when approaching a task depends on two factors: expectancy, which represents the student’s beliefs or judgments that he or she can successfully complete a task, and value, which represents the beliefs held by the student regarding why he or she should complete the task. A good historical and comparative overview of expectancy-value models is provided by Pintrich and Schunk (2002).

In a test-taking context, a particularly suitable expectancy-value model was provided by the work of Eccles and Wigfield (e.g., Eccles, 1983; Wigfield & Eccles, 2000). In this model, students’ expectancy beliefs regarding success on a task are influenced by their perceptions of both the difficulty of the task and their competence levels. Value beliefs are influenced by several factors, including attainment value (e.g., the importance of doing well on the task), intrinsic value (e.g., the enjoyment gained from doing the task), utility value (e.g., how the task fits into one’s future plans), and perceived costs (e.g., what one has given up to do the task).

Another model that is particularly relevant to low-stakes assessment testing is provided by Snow’s concept of *aptitude* (Snow, 1989, 1992, 1994). This model specifies two parallel pathways that describe the psychological-level contributions to task performance: a performance and a commitment pathway. The performance pathway constitutes a process by which individuals draw on relevant cognitive resources in the service of accomplishing a particular task. The commitment pathway describes a separate process by which individuals draw on relevant conative and affective resources in guiding, energizing, and regulating behavior toward accomplishing the task (Lau & Roeser, 2002).

In the context of low-stakes assessment testing, both expectancy-value and aptitude models would predict that some examinees would not give good effort, because they will perceive minimal personal benefit from the assessment testing experience. For these examinees, the task of doing well on the test will have limited attainment, intrinsic, or utility value. In addition, these examinees are likely to be aware of the costs associated with the assessment test (i.e., being denied the opportunity to engage in activities that they value more highly). The Eccles-Wigfield expectancy-value model would therefore predict low effort (and consequently, diminished performance) from those examinees who hold weak value beliefs. Similarly, Snow's aptitude model would characterize examinees exhibiting low effort on low-stakes tests as having diminished commitment pathway resources.

Snow, Corno, and Jackson's (1996) *Provisional Taxonomy of Individual Difference Constructs* distinguished among the cognitive, affective, and conative functions of the mind. Ferrara, Duncan, Perle, Freed, McGivern, and Chilukuri (2003) described these three functions in terms of important factors potentially influencing examinee encounters with test items. Of particular interest to the current study are two implications that Ferrara et al. cite regarding the conative factors: "The challenge is to develop materials and activities that are interesting . . ." and "Building in a variety of tasks and stimulating situations . . . help to keep students' attention and enhance the positive aspects of motivation and volition on performance" (p. 19). Thus, Ferrara et al. underscored the importance of motivational factors when developing items and administering tests.

CBTs and Test-Taking Motivation

One of the advantages of CBTs over traditional paper-and-pencil tests is their potential for administering items and tests in innovative ways. The most prominent type of innovative CBT is the computerized adaptive test (CAT), which tailors the difficulty levels of the items administered to the proficiency level of the examinee. The primary reason for a CAT is improved testing efficiency over a fixed-item test (FIT). However, Wainer (1990) noted that a positive consequence of this tailoring of item difficulty to examinee proficiency is that the more proficient examinees would not be bored by receiving items that are too easy while the less proficient examinees would not be frustrated by items that are too difficult. This suggests that a CAT, by reducing some of the potentially negative aspects of the test-taking session (e.g., boredom, frustration), should have a positive effect on examinee motivation (as compared to a FIT). Given a CAT's psychometric origins as a means to administer tests more efficiently, however, any positive effect on motivation should be characterized as an unintended consequence.

The idea that a CAT should enhance examinee motivation is reinforced by research on achievement motivation. In fact, one of the most consistently observed findings in motivation research is that moderately challenging tasks are the most intrinsically motivating (Pintrich and Schunk, 2002). Thus, because a CAT explicitly tries to administer moderately challenging tasks (items) to examinees, it would be expected to be more intrinsically motivating (and possibly increase examinee effort) than a fixed-item test. It is interesting, however, that no empirical studies were found that have studied this expected relationship.

A second type of CBT that holds potential for influencing examinee motivation is a self-adapted test (S-AT), which was developed by Rocklin and O'Donnell (1987). In a S-AT, an item pool is divided into several (typically 5-8) ordered difficulty levels, or strata, based on the items' difficulty parameters. An examinee begins a S-AT by choosing the difficulty level of his or her first item, which causes an item from the chosen level to be randomly selected and administered. After the item is answered, feedback is typically given regarding the correctness of the answer, after which the examinee chooses the difficulty level of the next item. This process continues until a stopping criterion is reached (either a predetermined number of items or a target precision of proficiency estimation is attained). After the test administration is completed, the examinee's test performance is typically calculated using an IRT-based proficiency estimation method.

Rocklin and O'Donnell (1987) who developed the S-AT procedure noted that, "instead of being tailored to the examinee's estimated ability level, a self-adapted test is tailored to the examinee's self-perceived ability as well to his or her motivational and affective characteristics" (p. 315). Thus, in contrast to a CAT, the S-AT procedure was explicitly intended to have an effect on examinees. This effect, however, is not intended to influence the examinee's standing on the proficiency construct of interest, but rather the extraneous factors that have been found to have a systematic debilitative effect on test performance. For example, examinees taking S-ATs have reported lower posttest state anxiety and higher test performance than those receiving CATs or FITs [see Pitkin & Vispoel (2001) or Rocklin, O'Donnell, & Holst (1995) for an overview of S-AT research findings].

Although there have been numerous studies of the effects of a S-AT on examinee anxiety, there have been virtually none concerning its effects on test-taking motivation. Vispoel and Coffman (1994) found that examinees reported a significant preference for a S-AT over a CAT. We hypothesized that, because examinees are continually making difficulty level choices, those taking a S-AT will be more engaged in the test-taking session, which should increase effort. A primary goal of the present study was to experimentally explore this hypothesis.

We additionally explored the effects of a modified S-AT that provided a more "game-like" experience for examinees taking low-stakes tests. In an examinee-informed, stratum-scored S-AT (EISS S-AT), an examinee is provided a set of point awards (for correct answers) and penalties (for incorrect answers) associated with each difficulty stratum. The examinee is told that after answering each item, his or her total test score will—depending on the correctness of the answer—increase or decrease by an amount dictated by the points associated with the chosen difficulty stratum (termed a *stratum score*). To illustrate this, Table 1 contains point values associated with a six-stratum EISS S-AT. For example, if an examinee chooses stratum 4, then his or her score increases by four points if the answer to the presented question is correct, but decreases by 3 points if the answer is incorrect. In effect, because the examinee is provided the stratum scores before he or she makes a difficulty stratum choice, stratum scores can be viewed as wagers that the examinee makes via his or her difficulty choices. It was hypothesized that the use of stratum scores—which examinees can dynamically see change through out the test—will introduce a "game-like" testing environment that would engage and sustain student effort more effectively than a FIT.

Table 1

Stratum Scores for Six Difficulty Strata

	Stratum					
	1	2	3	4	5	6
Item Score if correct	+1	+2	+3	+4	+5	+6
Item Score if incorrect	-6	-5	-4	-3	-2	-1

Wise (1999a, 1999b) showed that stratum scores (i.e., the summed item stratum scores) can provide a good approximation to maximum likelihood scoring, which over an accumulated sequence of administered items (as in a CAT) can be largely conceptualized by two basic principles:

1. Whenever an item is passed, an examinee's proficiency estimate is increased by an amount that is largely dependent on the difficulty level of the item. The more difficult the passed item, the larger the score increase.

2. Whenever an item is failed, an examinee's proficiency estimate is decreased by an amount that is largely dependent on the difficulty level of the item. The less difficult the failed item, the larger the score decrease.

The stratum scores in Table 1 are consistent with these two principles. Wise (1999) found that stratum scores correlated very highly with maximum-likelihood proficiency estimates (*r* = .98-.99).

The objective of the current study was to conduct an experiment investigating the motivational impact of four types of CBT: FIT, CAT, S-AT, and EISS S-AT. These CBTs were compared in terms of examinee test performance and effort.

Method

Participants

Prior to conducting this study, a power analysis was performed to determine how many students in each of the four CBT conditions would be needed to ensure that the study would be sensitive enough to detect differences in the conditions. Given a minimum meaningful effect size of *d* = .50, a level of power at .80, and four experimental groups, a minimum of 80 examinees per CBT condition was needed. The study was designed based on these needs. A total of 711 participants were recruited using the subject pool for introductory psychology courses at a mid-sized Southern university. Of the sample of student participants, 23% were male and 77% were female.

Measures

CBTs. For the purposes of this study, four CBTs were created: FIT, CAT, S-AT, and EISS S-AT. These tests used an item bank of retired American College Testing program (ACT) mathematics items obtained from four released 60-item forms of the ACT mathematics test. Item parameters were obtained using Bilog-MG (Zimowski, Muraki, Mislevy, & Bock, 1996). A three-parameter logistic item response model was used to calculate the parameters. Item parameters were computed for all but the last six items of each form, because it was judged that examinee responses to the last ten percent of each form were likely to be influenced by the test's time limit. In addition, one item was dropped because it appeared to have been answered correctly by all examinees. The final item pool consisted of 214 items.

The mathematics items were calibrated using samples of scored test data received from ACT. These data files contained item responses for 2748 to 2921 examinees for each of the four test forms received. It was assumed that the data for each of the forms represented randomly equivalent groups, thus eliminating the need for a linking step.

Once the item parameters were computed, the researchers determined each item's difficulty stratum, for use in the S-AT and EISS S-AT conditions. First, items were ranked in ascending order according to their difficulty parameter estimate. Next, the ordered set was divided into six equal-sized parts to form the difficulty strata. When examinees selected a difficulty level during the test, they received an item drawn at random, without replacement, from that stratum.

Students in each testing condition were administered one of the CBTs. The FIT was comprised of a fixed set of 40 items selected from one of the ACT forms that matched the content specifications for the entire form. For this test, the average *a* parameter was 1.10 (range: 0.45 to 2.16), the average *b* parameter was 0.01 (range: -1.52 to 1.54) and the average *c* parameter was .19 (range: .07 to .44).

The CAT began by administering 4 items randomly selected from the 20 most informative items at theta = 0.00, which provided the initial provisional proficiency estimate. Beginning with the 5th item, each item was chosen using maximum information item selection, the item was administered, and the provisional proficiency estimate was updated. No content balancing or item exposure constraints were imposed on the item selection. The CAT ended when 40 items had been administered.

In the S-AT, an examinee chose a difficulty stratum prior to each item, and then was administered an item drawn randomly from that stratum. After the item was answered, the examinees were given feedback regarding whether the answer given was correct or not, told which stratum they had last chosen, and asked to choose the difficulty level of the next item. This process repeated until 40 items had been administered.

The EISS S-AT was similar to the S-AT, except that examinees were given point values associated with passing and failing an item from each stratum, according to the values shown in Table 1. Examinees were told they began the test with 100 points (which was used to avoid negative scores during the test), and were encouraged to try to attain the highest score they could. After an item was answered, examinees were provided information about (a) whether or not they had passed the item, (b) the number of points they gained or lost on the item, (c) their updated score, and (d) the last stratum they had chosen. They were then asked to choose the difficulty level of their next item. The process repeated until 40 items had been administered. To permit comparability between the EISS S-AT condition and the other experimental conditions, maximum-likelihood proficiency estimates (bounded by -4.0 and +4.0) were computed and used in the data analyses rather than total stratum scores.

Self-Reported Effort. The Student Opinion Survey (SOS; Sundre & Moore, 2002) is a 10-item paper-and-pencil survey designed to measure examinee motivation. The SOS provides three scores: an effort score (5 items), an importance score (5 items) and a total motivation score. Sundre and Moore reported internal consistency reliabilities ranging from .80-.89 for the three scales. In addition, Sundre and Moore found that all three motivation scores are positively correlated with performance scores. Moreover, they found that both motivation and performance scores differed significantly in high- and low-stakes testing conditions. For the current study, only the five items pertaining to examinee effort (SOS-Effort) were involved in the analysis because the effort that examinees reported exerting was judged to be more important than the importance they placed on the test.

Response Time Effort. Wise and Kong (2005) showed that item response time can also be used to measure examinee effort on a CBT. Response time effort (RTE) is based on the idea that if an examinee answers an item quickly (i.e., before he or she had time to read the item, comprehend the task presented, and identify the correct answer) then the response represents *rapid-guessing behavior*, which is indicative of lack of effort. Otherwise, the response represents *solution behavior*. An examinee's RTE score is the proportion of items for which he or she exhibited solution behavior, and an examinee exhibiting solution behavior on all items would attain the maximum RTE score of 1.0. To differentiate rapid-guessing from solution behavior, a response time threshold of 10 seconds was used for all items.

Procedures

Sixteen separate data collection sessions, comprising four sessions for each of the four CBT conditions, were set up using a psychology department subject pool. In each session, all examinees received the same type of CBT. Students enrolled in specific psychology courses within the university were required to participate in research projects or complete alternative written assignments for course credit. A maximum of 50 participants for each session were recruited. One proctor and one research assistant led each of the data collection sessions, with a total of four different proctors and two different research assistants used across the 16 sessions. While 16 data collection sessions were planned, one session was not completed. The missing session, which would have administered a S-AT, was not included in the final sample because a problem with the computer network server caused an interruption of 20 minutes during the testing session. As a result, the researchers felt that this interference might influence student responses and motivation. When a second attempt was made to collect data for this S-AT condition on a different day, a fire alarm occurred, which disrupted the session. At this point, the lack of available participants resulted in the researchers abandoning efforts to collect data for the 16th session. Consequently, the total number of participants in the S-AT condition was lower than those in the other CBT conditions.

To control for the possibility of experimenter/proctor effects, sessions were counterbalanced across the different CBT conditions. To illustrate, each proctor led a total of four data collection sessions, one session for each of the CBT conditions (FIT, CAT, S-AT, EISS S-AT). To ensure the standardization of the data collection procedures, a script was written and followed by each proctor. Research assistants helped only with the distribution and collection of materials and therefore did not require a script.

The CBT conditions were unspeeded and all students completed the 40-item test within 60 minutes. Examinees were read instructions about their particular test, after which they progressed through the CBT at their own pace. Once an examinee completed the CBT, he or she was administered the SOS, after which the examinee was free to leave the testing session.

Data Analyses

There were three dependent variables investigated in this study: estimated proficiency, self-reported effort (SOS-Effort) and response time effort (RTE). There was one independent variable (test type) and one classification variable (examinee gender) used in the analysis. However, because participants signed up for test sessions and sessions were randomly assigned to a particular CBT, in this experimental design participants were nested within session and session was nested within test type. This indicated that hierarchical ANOVAs were appropriate for testing treatment effects for the three dependent variables. For all analyses of treatment effects, a .05 level of significance was used.

Results

Table 2 shows descriptive statistics for estimated proficiency, broken down by test type and examinee gender. All three of the adaptive tests yielded similar mean proficiency, while the FIT showed the highest mean. The hierarchical ANOVA, however, revealed nonsignificant effects for test type [$F(3,17.02) = 1.04$, $p = .399$], gender [*F*(1,677) = 0.03, *p* = .858], and the test type by gender interaction [*F*(3,677) = 0.77, *p* = .509]. Thus no treatment effects were observed for test type.

Table 2

Descriptive Statistics for Estimated Proficiency, by Test Type and Examinee Gender

	Test Type				
Gender	FIT	CAT	S-AT	EISS S-AT	All Examinees
Males					
Mean	0.41	0.26	0.08	0.06	0.24
SD	0.56	1.54	1.25	1.02	1.12
N	54	45	30	34	163
Females					
Mean	0.28	0.15	0.16	0.21	0.20
SD	0.65	1.39	0.64	0.87	0.95
N	136	140	108	149	533
All Examinees					
Mean	0.32	0.17	0.14	0.18	0.21
SD	0.63	1.42	0.80	0.90	0.99
N	190	185	136	183	696

Similar results were found for the two dependent variables measuring effort. Table 3 shows the descriptive statistics for SOS-Effort. There was only minor variation in means among test type and gender groups. This analysis showed nonsignificant effects for test type [*F*(3,16.88) = 0.21, *p* = .885], gender [*F*(1,631) = 0.23, *p* = .631], and the test type by gender interaction [*F*(3,677) = 1.04, *p* = .372]. The descriptive statistics for RTE are given in Table 4. The different groups exhibited very similar means and, again, nonsignificant effects were found for test type [*F*(3,14.62) = 0.92, *p* = .454], gender [*F*(1,677) = 1.26, *p* = .261], and the test type by gender interaction [*F*(3,677) = 0.89, *p* = .446].

Additional Analyses

The ANOVAs for the three dependent variables consistently showed that test type did not affect either the test performance or effort levels of the examinees. We did, however, discover an unexpected influence on the dependent variables. Although not originally part of our research questions, the hierarchical experimental design used in this study also permitted an analysis of proctor effects.

Table 3

Descriptive Statistics for SOS-Effort, by Test Type and Examinee Gender

	Test Type				
Gender	FIT	CAT	S-AT	EISS S-AT	All Examinees
Males					
Mean	16.41	17.09	16.89	17.23	16.86
SD	3.56	3.86	3.83	3.53	3.67
N	51	44	28	31	154
Females					
Mean	17.64	17.40	16.66	16.90	17.17
SD	3.51	3.90	4.07	3.80	3.82
N	124	134	100	138	496
All Examinees					
Mean	17.28	17.32	16.71	16.96	17.10
SD	3.56	3.88	4.01	3.74	3.78
N	175	178	128	169	650

Table 4

Descriptive Statistics for RTE, by Test Type and Examinee Gender

	Test Type				
Gender	FIT	CAT	S-AT	EISS S-AT	All Examinees
Males					
Mean	.96	.95	.94	.95	.95
SD	.05	.08	.09	.08	.07
N	54	45	30	34	163
Females					
Mean	.96	.97	.95	.95	.96
SD	.05	.05	.07	.08	.06
N	136	140	108	128	533
All Examinees					
Mean	.96	.96	.95	.95	.96
SD	.05	.06	.07	.08	.06
N	190	185	138	183	696

Table 5 shows the descriptive statistics for estimated proficiency broken down by proctor and examinee gender. Much larger differences among the groups are present, relative to those found for test type. A hierarchical ANOVA was performed with proctor (nested within session) and gender as the factors; the results are found in Table 6. There were significant effects for both proctor and the proctor by gender interaction. The proctor effect showed a large effect size (η^2). To better understand the interaction, tests of simple effects were performed, the results of which are also shown in Table 6, revealed that the proctor effect was found for male examinees but not for females.

Table 5

Descriptive Statistics for Estimated proficiency, by Proctor and Examinee Gender

	Proctor				
Gender	A	B	C	D	All Proctors
Males					
Mean	0.47	0.36	0.55	-0.16	0.24
SD	0.90	1.08	0.87	1.30	1.12
N	42	30	32	59	163
Females					
Mean	0.18	0.12	0.39	0.18	0.20
SD	0.99	0.85	1.07	0.93	0.95
N	145	158	102	128	533
All Examinees					
Mean	0.24	0.15	0.43	0.07	0.21
SD	0.97	0.89	1.03	1.07	0.99
N	187	188	134	187	696

The analyses of proctor effects for the effort variables were mixed. For SOS-effort, the corresponding hierarchical ANOVA found no significant effects. For RTE, however, the results were similar to those found with estimated proficiency. Table 7 shows the descriptive statistics for RTE. The proctor/gender combination RTE means ranged from .93 to .97. In addition, the overall mean of .96 indicates that 4% of the examinee responses represented rapid-guessing behavior. The ANOVA for RTE is shown in Table 8. There was a significant effect for proctor, as well as for session within proctor, but no proctor by gender interaction. The effect size for proctor was similar to that found for estimated proficiency. It is useful to note that the RTE scores were substantially skewed, which calls in question the appropriateness of the parametric ANOVA analysis reported for these scores. Although there is not a nonparametric counterpart to the hierarchical ANOVA model analyzed here, we conducted a Kruskal-Wallis test (i.e., analogous to a one-factor ANOVA) on RTE and again found a significant proctor effect.

Table 6

ANOVA Results for Estimated Proficiency, by Proctor and Examinee Gender

Source	*SS*	*df*	*MS*	*F*	*F-prob*	η^2
Proctor	22.88	3	5.09	4.80	.011	.407
Proctor at Males	16.63	3	5.54	5.73	.001	.025
Proctor at Females	5.07	3	1.69	1.75	.156	.008
Gender	0.74	1	0.74	0.76	.382	.001
Proctor x Gender	10.90	3	3.63	3.75	.011	.016
Session within Proctor	12.12	11	1.10	1.14	.3271	.018
Error	667.14	688	0.97			
Total	687.19	695				

Table 7

Descriptive Statistics for RTE, by Proctor and Examinee Gender

	Proctor				
Gender	A	B	C	D	All Proctors
Males					
Mean	.96	.96	.97	.93	.95
SD	.06	.06	.04	.09	.07
N	42	30	32	59	163
Females					
Mean	.97	.96	.95	.94	.96
SD	.05	.06	.06	.07	.06
N	145	158	102	128	533
All Examinees					
Mean	.97	.96	.95	.94	.96
SD	.05	.06	.06	.08	.06
N	187	188	134	187	696

Table 8

ANOVA Results for RTE, by Proctor and Examinee Gender

Source	*SS*	*df*	*MS*	*F*	*F-prob*	η^2
Proctor	0.075	3	0.025	3.739	.032	.406
Gender	0.002	1	0.002	0.448	.504	.001
Proctor x Gender	0.017	3	0.006	1.391	.244	.006
Session within Proctor	0.086	11	0.008	1.947	.031	.031
Error	2.719	677	0.004			

Discussion

Being able to effectively address the problem of low examinee effort is important for measurement practitioners who are administering tests in low-stakes contexts. The goal of the present study was to investigate whether a variety of adaptive tests could induce greater levels of examinee effort and test performance than those found with a fixed item CBT. The results of this study clearly showed that test type influenced neither effort nor performance.

It is likely that the two self-adapted tests (S-AT and EISS S-AT) increased examinee engagement. However, they also required more effort for the examinees to complete. Examinees were asked to make item difficulty level choices as well as answer the test items, which additionally increased the time required to complete the test. It is possible that any increased examinee engagement was offset by the additional effort required by the test. Wise (2004) found that the effort an item received was positively related to its position in the test. This effect might be re-interpreted as indicating that the longer a test takes, the more likely that that effort will diminish. Thus, the self-adapted tests may have had dual contrasting effects, one of which tended to facilitate test-taking effort, and the other having a debilitative effect. This might explain the low net effect of the self-adapted tests.

It was also found that a CAT did not yield effort and performance exceeding that from a FIT. From a psychometric standpoint, this makes perfect sense, as it merely indicates that the invariance principle underlying IRT is maintained in this study’s data. Operationally, in cases where there are CAT and FIT (either CBT or paper-and-pencil) versions of the same test, we strive for equivalence between the two test types. From a motivation theory standpoint, however, the results were unexpected. Because a CAT explicitly seeks to administer moderately challenging items to examinees, the intrinsic motivation of a low-stakes test would be expected to increase, leading to increased examinee effort and—ultimately—improved test performance. It may be that in the context of low-stakes testing, providing moderately challenging items may have limited practical effect. More research is needed on this issue.

The discovery of a significant proctor effect was surprising and stands in sharp contrast to the absence of a test type effect. The proctors and research assistants were all female graduate students who read identical sets of test instructions in the same testing lab during the same hours of the testing days. Apparently, it was the male examinees who were most affected by who administered the test. The results of this study indicate that proctor effects represent a source of construct-irrelevant variance that may have a meaningful effect on test score validity. Because little is known about the dynamics of proctor effects, additional research should be devoted toward better understanding their effects.

Within the context of low-stakes testing, this study's results should be a matter of concern to measurement professionals. Test type—which is very much in control of the test giver—had very little effect on examinees, whereas assigned proctor—which apparently can have a meaningful impact—is far less controllable by test givers. That is, although we readily can designate who serves as a proctor and who does not, unless we understand which types of proctors have which type of effects on examinee effort, proctor effects may be difficult to manage. From a practical standpoint, in low-stakes settings it is often challenging to find individuals willing to serve as proctors, and to have to select only some of the available individuals effectively diminishes their supply.

Low-stakes tests pose more challenges to obtaining valid scores than their high-stakes counterparts. The effort expended by an examinee taking low-stakes tests is vulnerable to a number of construct-irrelevant influences, such as how much reading an item requires or how late in the test the item appears (Wise, 2004), how mentally taxing an item is (Wolf, Smith, & Birnbaum, 1995), and who proctors the test. Because of the strong personal consequences associated with test performance, an examinee taking a high-stakes test is far less vulnerable to these threats to test score validity. Therefore, because validity appears to be a more fragile characteristic in low-stakes testing contexts, it is important that we better understand the most serious threats to validity and identify strategies for effectively dealing with them.

References

Eccles, J. (1983). Expectancies, values, and academic behaviors. In J. T. Spence (Ed.), *Achievement and achievement motives* (pp. 75-146). San Francisco: Freeman.

Ferrara, S., Duncan, T., Perle, M., Freed, R., McGivern, J., & Chilukuri, R. (2003, April). *Item construct validity: Early results from a study of the relationship between intended and actual cognitive demands in a middle school science assessment.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

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

Lau, S. & Roeser, R. W. (2002). Cognitive abilities and motivational processes in high school students' situational engagement and achievement in science. *Educational Assessment, 8,* 139-162.

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

Pitkin, A. K., & Vispoel, W. P. (2001). Differences between self-adapted and computerized adaptive tests: A meta-analysis. *Journal of Educational Measurement*, *38*, 235-247.

Rocklin, T. R., & O'Donnell, A. M. (1995). Self-adapted testing : A performance-improving variant of computerized adaptive testing. *Journal of Educational Psychology, 79,* 315-319.

Rocklin, T. R., O'Donnell, A. M., & Holst, P. M. (1995). Effects and underlying mechanisms of self-adapted testing. *Journal of Educational Psychology, 87,* 103-116.

Snow, R. E. (1989). Cognitive-conative aptitude interactions in learning. In R. Kanfer, P. L. Ackerman, & R. Cudeck (Eds.), *Abilities, motivation, and methodology* (pp. 435-474). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Snow, R. E. (1992). Aptitude theory: Yesterday, today, and tomorrow. *Educational Psychologist, 27,* 5-32.

Snow, R. E. (1994). Abilities in academic tasks. In R. J. Sternberg & R. K. Wagner (Eds.) *Mind in context: Interactionist perspectives on human intelligence* (pp. 3-37). Cambridge: Cambridge University Press.

Snow, R. E., Corno, L., & Jackson, D. (1996). Individual differences in affective and conative functions. In D. C. Berliner & R. C. Calfee (Eds.), *Handbook of educational psychology* (pp. 243-310). New York: Macmillan.

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

Vispoel, W. P., & Coffman, D. D. (1994). Computerized-adaptive and self-adapted music-listening tests: Psychometric features and motivational benefits. *Applied Measurement in Education, 7,* 25-51.

Wainer, H. (1990). Introduction and history. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (pp. 1-21). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc..

Wigfield, A., & Eccles, J. (2000). Expectancy-value theory of achievement motivation. *Contemporary Educational Psychology, 25,* 68-81.

Wise, S. L. (1999a, April). *The rationale and principles of stratum scoring*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal.

Wise, S. L. (1999b, April). *Comparison of stratum scored and maximum likelihood-scored CATs*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal.

Wise, S. L. (2004). An investigation of the differential effort received by items on low-stakes, computer-based tests. Manuscript under review.

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

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.

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.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). BILOG-MG [Computer software]. Chicago: Scientific Software International.