Examinee Effort and Test Score Validity[1]

Steven L. Wise, James Madison University

Scores on achievement and aptitude tests are used to make inferences regarding what examinees know and can do on a construct of interest. The validity of these inferences is dependent, in part, on the qualities of the test. These qualities include the degree to which the test items adequately represent the construct and whether the number of items administered is enough to provide scores with sufficient reliability. In addition, the scores should be reasonably free from construct-irrelevant variance (Messick, 1984). This means that the observed scores predominantly reflect the construct of interest, and are not unduly influenced by factors irrelevant to that construct. Haladyna and Downing (2004) discussed a variety of sources of construct-irrelevant variance and encouraged measurement professionals to be mindful of these potential threats to inferences made on the basis of test scores.

One important potential threat to the validity of score-based inferences is the degree of effort devoted by examinees to the test. The test giver tacitly assumes that when an examinee takes a test, he or she will devote considerable effort to getting the items correct. There are times, however, when an examinee does not try to do his or her best; in these instances, the resultant test score is likely to underestimate what that examinee knows and can do. Thus, low effort typically leads to negatively biased estimates of examinee proficiency. Moreover, whenever test-taking effort varies across examinees, there will be a differential biasing effect, which will introduce construct-irrelevant variance into the test score data. To the degree to which low examinee effort was present in a test administration, inferences made on the basis of the test scores will be confounded. A low test score might be due to the examinee being of low actual proficiency, or it could be due to an examinee of much higher actual proficiency not trying very hard on the test.

To what extent should measurement professionals care about this potential problem? If the test scores have personal consequences for examinees (such as grades, diplomas, licensure, etc.), low effort is not generally considered a major validity threat. In these situations, test performance is considered to be the responsibility of the examinee, and if an examinee chooses to not give good effort to a test, it is generally not viewed as a meaningful threat to the validity of score-based inferences. In contrast, there are numerous measurement contexts in which the scores have important consequences for the test givers, but little consequences for examinees. Examples of this include the National Assessment of Education Progress (NAEP), the Third International Mathematics and Science Study (TIMMS), and many higher education assessment tests. In addition, testing programs often field test items in nonconsequential settings to obtain data that are subsequently used to calibrate items, construct test forms, and perform additional

[1] Paper presented at the October 2007 annual meeting of the Northeastern Educational Research Association, Rocky Hill, Connecticut.

linking/equating tasks. In these contexts, it is likely that some examinees will not give good effort, and the validity threat will increase with the number of these examinees.

Note that the contemporary high-stakes versus low-stakes distinction is not particularly useful here. The definition of a high-stakes test is that "it has serious consequences for students *or* [italics added] educators" (AERA, 2000, p. 24). From an effort perspective, it is the consequences for examinees that is the primary issue, regardless of the consequences for the test giver. Hence, many testing programs that are labeled high-stakes (e.g., NAEP) are vulnerable to the validity threat posed by low examinee effort. In this paper, test stakes will be viewed from the examinee's perspective; hence, a test that has little or no consequences for examinees will be considered low-stakes.

The purpose of this paper is to provide an overview of the current research base regarding test-taking motivation. First, the literature on examinee motivation will be reviewed, highlighting the degree to which examinee effort has been found to affect test performance. Second, a number of approaches that have recently been proposed for effectively managing this threat to score-based inferences will be discussed.

Examinee Motivation and Test Performance

A number of studies have investigated the relationship between examinee motivation and test performance. Many of these studies relied on either existing groups that were believed to differ in motivation or experimental conditions that were designed to manipulate the motivational levels of the examinees (Arvey, Strickland, Drauden, & Martin, 1990; Brown & Walberg, 1953; Jennings, 1953; Kim & McLean 1995; Kiplinger & Linn, 1995/1996; O'Neil, Sugrue, & Baker, 1995/1996; Rothe, 1947; Schiel, 1996; Sundre, 1999; Sundre & Kitsantas, 2004; Taylor & White, 1981; Wolf & Smith, 1995; Wolf, Smith, & DiPaulo, 1996). Wise and DeMars (2005) synthesized a number of these studies, and found that the more motivated examinee groups tended to show higher test performance than the less motivated groups by an average of 0.58 standard deviations. Other studies were nonexperimental, in which examinees provided self-reports regarding their test-taking motivation and effort. In these studies, motivation and test performance have been found to be positively related (Schiel, 1996; S. Wise & Kong, 2005).

There has recently been an interest in using item response time on computer-based tests (CBTs) as a means of measuring examinee effort. In this research, examinee item responses are classified as either *rapid-guessing behavior* (in which the examinee responds faster than the time needed to read and consider the item), or *solution behavior* (all other responses). Based on the hypothesis that rapid guesses are non-effortful, S. Wise and Kong (2005) developed a measure of test-taking effort, termed *response time effort (RTE)*. RTE, which represents the proportion of test items for which an examinee exhibited solution behavior, has been found to be positively correlated with test performance (Eklöf, 2007; S. Wise & Kong, 2005).

One of the more interesting findings of the response time-based studies is that most examinees in low-stakes testing situations have been found to exhibit considerable testing taking effort. Average RTE scores have ranged from .99 (S. Wise, Kingsbury,

Thomason, & Kong, 2004) down to .85 (DeMars, 2007). However, S. Wise and DeMars (2006) found that item parameter estimates could be noticeably biased with as few as 2.3% rapid guesses, and Kong (2007) found a similar impact on proficiency estimates.

Strategies for Managing Low Examinee Effort

Given the psychometric concerns posed by low examinee effort, it is important that measurement practitioners have strategies for effectively addressing this problem. There are several options available. First, one might appeal to the "citizenship" of the examinees, and encourage them to give good effort despite the absence of personal consequences. Second, one might provide feedback to examinees regarding test performance. Such feedback should not consist only of a test score, but be meaningfully interpretable by the examinee from either a norm-referenced or criterion-referenced perspective. Third, one might offer some type of incentive (e.g., monetary) to examinees for their performance.

It is difficult to assess, however, the effectiveness of these strategies in low-stakes testing situations. For example, O'Neil et al. (1995/1996) investigated the use of monetary incentives on NAEP testing, offering $1 per correct answer to both 8th and 12th grade samples. They found partial evidence that monetary incentives improved test performance for 8th graders, while the test performance of 12th graders was unaffected by the incentive. In another study, V. Wise (2004) investigated the effects of promising feedback to examinees taking a battery of low-stakes higher education assessment tests on their self-reported effort. She found that although the examinees in her study expressed a desire for feedback, neither self-reported effort nor test performance was improved by its being promised.

Recently, there have been four new methods proposed for improving the validity of test data in which low examinee effort is present. Three of these methods are statistically-based methods for culling out non-effortful data, while the fourth focuses on actively reducing the amount of non-effortful responses that occur.

Motivation Filtering

In many low-stakes testing situations, the goal of the testing program is to estimate the mean proficiency of a population of examinees. Examinees who do not give good effort will tend to receive test scores that underestimate their proficiency levels. This implies that the presence of test scores from these examinees will lower the mean proficiency of the group from what it would have been had all examinees given good effort (which is what the test givers want to assume). One solution to this problem would be to discard, or filter out, the data from examinees who did not give good effort. This procedure, termed *motivation filtering*, requires the assumption that effort is unrelated to actual proficiency. That is, it must not be the case that those who give little effort do so because they know little about the test content. Otherwise, by deleting the data from predominantly lower-proficiency examinees, the distribution of actual proficiency will be distorted, and the overall group mean would be positively biased.

Table 1. The Impact of Motivation Filtering on Convergent Validity Correlations

Study	Test Content	External Measure	Correlation Between Test Score and External Measure		Effort Measure
			Before Filtering	After Filtering	
Sundre & S. Wise (2003)	Natural World	SAT-Total	.45	.64	SR
	Quantitative Reasoning	SAT-Total	.45	.62	SR
	Correct Reasoning	SAT-Total	.41	.52	SR
S. Wise, Kingsbury, Thomason, & Kong (2004)	Mathematics	GPA	.55	.55	SR
	Mathematics	GPA	.51	.50	RTE
S. Wise & DeMars (2005)	American Experience	SAT-Total	.34	.53	SR
S. Wise & Kong (2005)	Information Literacy	SAT-Verbal	.50	.54	SR
	Information Literacy	SAT-Math	.50	.58	RTE
V. Wise, S. Wise, & Bhola (2006)	Information Literacy	SAT-Verbal	.36	.46	SR
	Information Literacy	SAT-Math	.12	.24	SR
	Fine Arts	SAT-Verbal	.54	.55	SR
	Fine Arts	SAT-Math	.22	.27	SR
	Scientific Reasoning	SAT-Verbal	.46	.54	SR
	Scientific Reasoning	SAT-Math	.38	.43	SR
	American Experience	SAT-Verbal	.56	.59	SR
	American Experience	SAT-Math	.24	.26	SR
	Sociocultural Dimension Assessment	SAT-Verbal	.34	.42	SR
	Sociocultural Dimension Assessment	SAT-Math	.22	.35	SR
Kong, S. Wise, & Bhola (2007)	Information Literacy	SAT-Verbal	.36	.47	RTE

Note. For the effort measure, SR = self-report and RTE = response time effort.

There have been several empirical investigations of motivation filtering. Sundre and S. Wise (2003) looked at several university assessment tests in which the data from examinees self reporting low effort were filtered out. Their basic findings were that (a) after filtering, overall mean scores increased by .27 to .40 standard deviations, (b) the convergent validity correlations between test scores and an external variable (SAT) increased markedly, and (c) the correlation between effort and SAT scores was approximately zero. Sundre and S. Wise concluded that motivation filtering was

successful at removing untrustworthy non-effortful data while not distorting the distribution of actual examinee proficiency. Similar results were found in subsequent studies of motivation filtering, all of which were conducted in higher education settings (Kong, S. Wise, & Bhola, 2007; S. Wise & DeMars, 2005; S. Wise & Kong, 2005; V. Wise, S. Wise, & Bhola, 2006). Table 1 summarizes the changes in convergent validity correlations for these studies.

One study of motivation filtering, conducted on data from a statewide achievement testing program yielded anomalous results (S. Wise et al., 2004). In their study, (a) mean scores did not increase after filtering, (b) convergent validity correlations did not increase (see Table 1), and (c) self-reported motivation showed a weak, though statistically significant, correlation with student grade point average (GPA). However, because the study differed from the others in a number of ways (younger examinees from grades 6-10; only 1% of the responses were rapid guesses; GPA was used as an external measure), it is difficult to draw conclusions concerning the anomaly. The findings of S. Wise et al. (2004) point to the need for additional studies regarding the generalizability of motivation filtering results found across higher education settings.

Rapid-Response Filtering

In motivation filtering, the entire set of an examinee's item responses are filtered out. S. Wise & Kong (2005), however, found evidence that some examinees may exhibit solution behavior during early items in a test and then at some point switch to rapid-guessing behavior for the remaining items. This suggests that a portion of an examinee's response record may provide useful information in estimating proficiency. S. Wise (2006) studied rapid-response filtering, in which rapid guesses are deleted from an examinee's set of responses, with the examinee's test score consisting of the proportion of items passed under solution behavior. He found that both the item difficulties and item-total correlations of easier items were most affected by this method. In addition, he found that rapid-response filtering yielded scores with higher convergent validity, as shown in Table 2. Kong (2007) also studied rapid-response filtering, finding that one of her two convergent validity correlations increased, while the other was unchanged (see Table 2).

Table 2. The Impact of Rapid-Response Filtering on Convergent Validity Correlations

			Correlation Between Test Score and External Measure	
Study	Test Content	External Measure	Before Filtering	After Filtering
S. Wise (2006)	Information Literacy	SAT-Verbal	.36	.39
	Information Literacy	SAT-Math	.12	.15
	Information Literacy	GPA	.24	.25
Kong (2007)	Scientific Reasoning	SAT-Verbal	.49	.49
	Scientific Reasoning	SAT-Math	.42	.45

Comparisons of Tables 1 and 2 suggest that rapid-response filtering tends to have a weaker impact on convergent validity. In addition, the credibility of scores based on rapid-response filtering is unclear, because test scores are based on different sets of items across examinees and classical test theory-based scoring is used. Therefore, its use is not recommended in practice.

The Effort-Moderated IRT Model

S. Wise and DeMars (2006) extended the logic of rapid-response filtering into an IRT-based framework. They reasoned that an item's response function under solution behavior would follow a familiar *s*-shaped item characteristic curve (ICC), whereas for rapid-guessing behavior the ICC would be flat—indicating that the probability of an examinee passing the item under rapid guessing would be near chance level regardless of examinee proficiency. S. Wise and DeMars proposed an *effort-moderated IRT model* under which an examinee's response was modeled either using the solution behavior or rapid guessing ICC, depending on how the item response was classified (using response time). They found that when rapid guessing was present in test data, an effort-moderated model showed better model fit, more accurately estimated item parameters, and yielded scores with higher convergent validity than did a standard IRT model. Similar results were reported by Kong (2007). DeMars (2007) found that an effort-moderated model had a weaker effect on convergent validity, though those results may have been due to her use of course grade as the external measure. Table 3 shows the convergent validity results for these three studies.

Table 3. The Effect of Using an Effort-Moderated IRT Model on Convergent Validity Correlations

			Correlation Between Test Score and External Measure	
Study	Test Content	External Measure	Standard 3PL Model	Effort-Moderated 3PL Model
S. Wise & DeMars (2006)	Information Literacy	SAT-Verbal	.33	.40
	Information Literacy	SAT-Math	.13	.19
	Information Literacy	GPA	.24	.27
Kong (2007)	Scientific Reasoning	SAT-Verbal	.49	.52
	Scientific Reasoning	SAT-Math	.41	.47
DeMars (2007)	Unspecified	Course Grade	.52	.52
	Unspecified	Course Grade	.50	.50
	Unspecified	Course Grade	.43	.46
	Unspecified	Course Grade	.27	.27
	Unspecified	Course Grade	.21	.20
	Unspecified	Course Grade	.08	.14

One of the less obvious effects of the presence of rapid guessing is that it tends to increase the internal consistency of the test score data. This unexpected finding that rapid guessing decreases convergent validity while increasing internal consistency was observed by S. Wise (2006). S. Wise and DeMars (2006) found that, for real data, test information was higher for the standard IRT model than for the effort-moderated model. However, using simulated data, they showed that the test information was spuriously high for the standard model, and that the effort-moderated model better reproduced the true test information functions. S. Wise and DeMars (in press) provide an additional discussion of the relationship between rapid guessing and internal consistency.

The Effort-Monitoring CBT

Motivation filtering, rapid-response filtering, and effort-moderated IRT models represent strategies that measurement professionals might use to manage test data that contain non-effortful examinee responses. A more proactive approach is to reduce the number of non-effortful behaviors that occur. With this goal in mind, S. Wise, Bhola, and Yang (2006) developed *the effort-monitoring CBT*, in which the computer monitors effort during an examinee's test administration, and displays warning messages to those beginning to consistently exhibit rapid-guessing behavior. Wise et al. conducted an experiment in which an effort-monitoring CBT was compared to a traditional CBT for examinees taking a higher education assessment test. The effort-monitoring CBT yielded higher RTE scores. In addition, for examinees deserving at least one warning, the proportion of examinees deserving an additional warning decreased and convergent validity correlation increased. In a follow-up study, Kong, Wise, Harmes, and Yang (2006) replicated the findings from the Wise et al. study, with the additional finding that examinees deserving at least one warning showed higher test performance when assigned to the effort-monitoring CBT condition. Table 4 contains the convergent validity correlations from these two studies. Generally, the increases in convergent validity appear to be higher for the effort-monitoring CBT than for the filtering methods or the effort-moderated IRT model. This may indicate that the effort-monitoring CBT has a more potent ameliorative effect on test score validity.

Table 4. The Effect of an Effort-Monitoring CBT on Convergent Validity Correlations for Examinees Deserving Warning Messages

			Correlation Between Test Score and External Measure	
Study	Test Content	External Measure	Traditional CBT	Effort-Monitoring CBT
S. Wise, Bhola, & Yang (2006)	Scientific Reasoning	SAT-Verbal	.22	.31
	Scientific Reasoning	SAT-Math	-.01	.33
	Scientific Reasoning	GPA	.32	.46
Kong, S. Wise, Harmes, & Yang (2006)	Scientific Reasoning	SAT-Verbal	.08	.31
	Scientific Reasoning	SAT-Math	.15	.30

One of the attractive features of the effort-monitoring CBT is that it does not require the assumption that examinee effort is unrelated to actual proficiency. It does, however, require that a CBT is used and that the testing software can administer the test in an effort-monitoring format. Nevertheless, the effort-monitoring CBT appears to be a promising way to address examinee motivation challenges.

Conclusions and Recommendations

There is a growing body of research on the relationship between examinee motivation and test performance. This research has clearly demonstrated that unmotivated examinees give less than full effort on test items, and to the extent to which examinees exhibit non-effortful behaviors, the psychometric properties of test data are degraded. At the level of the individual examinee, test scores from unmotivated examinees are likely to underestimate their actual proficiency levels. Furthermore, if the unit of analysis is a group of examinees, or if the test data are used to calibrate a set of test items or to estimate relationships between test performance and other variables, low effort from even a small percentage of examinees can meaningfully distort the data analyses. It is therefore important that measurement professionals be mindful of the potential for problems to be caused by low examinee effort and have effective strategies for dealing these problems.

There are a variety of methods that might be employed by test givers as they strive to elicit maximal examinee effort in low-stakes testing situations. Appeals to examinees' sense of citizenship may be effective, as well as providing them with interpretable feedback regarding their test performance. Monetary incentives might also be useful, though the financial costs to the testing program may render their use prohibitive, and the specific conditions under which monetary incentives will be effective are not yet understood.

Three methods were discussed for dealing with test data from unmotivated examinees. First, motivation filtering deletes all item data from examinees who have been identified as giving low effort to the test. Whenever it is reasonable to assume that effort is unrelated to actual proficiency, motivation filtering can effectively cull out a sizeable proportion of non-effortful item responses and improve score validity. A key advantage of motivation filtering is that it can be used with both paper-and-pencil tests (using self-report measures of effort) and CBTs (using RTE scores). Second, rapid-response filtering was designed to filter data at the item response level (and thereby preserve data from examinees who give good effort to part of their test). Its use in practice is not recommended, due to questions concerning the interpretability of its proportion-correct scores and its relatively weak improvement in convergent validity. The third method is the effort-moderated IRT model. As with rapid-response filtering, individual item responses are classified as solution behaviors or rapid guesses, but more rigorous IRT methods are used in test scoring. Effort-moderated models may prove to be quite useful in practice. Note, also, that more than one of the three methods can be used in concert. DeMars (2007) investigated the use of motivation filtering to delete the examinees exhibiting particularly low effort, and used an effort-moderated model with the remaining examinees.

The effort-monitoring CBT represents a new type of CBT. Traditionally, measurement professionals have focused on trying to develop CBTs that are equivalent to paper-and-pencil tests. That is, the computer has played a relatively passive role in test administration. The primary exception is the computerized adaptive test (CAT); however, CATs were developed to increase testing efficiency and not to increase test score validity[2]. In contrast, an effort-monitoring CBT plays an active role in test administration by identifying rapid guessing and intervening as needed to reduce subsequent rapid guessing. The evidence that effort-monitoring CBTs yield increased examinee effort and increased score validity is encouraging, though more research is needed to better understand how to most effectively implement these types of tests.

In low-stakes testing programs, the primary goal of the measurement professional is to obtain valid test scores. When there are little or no personal consequences for examinees, the responsibilities of the measurement professional extend beyond the development of a sound assessment test to the establishment of testing procedures and analytic techniques that yield data with sufficient quality to support the intended score-based inferences. To that end, low-stakes testing programs should routinely conduct "effort analyses" to assess the degree to which examinee motivation is creating psychometric problems, and take steps to address these problems.

[2] Proponents of CATs have claimed increased examinee motivation as a feature of a CAT, though this has not been clearly demonstrated.

References

American Educational Research Association (2000). Position statement of the American Educational Research Association concerning high-stakes testing in PreK-12 education. *Educational Researcher, 29(8),* 24-25.

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

Brown, S. M., & Walberg, H. J. (1993). Motivational effects of test scores of elementary students. *Journal of Educational Research*, *86*, 133-136.

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

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

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

Jennings, E. E. (1953). The motivation factor in testing supervisors. *Journal of Applied Psychology*, *37*, 168-169.

Kim, J. G., & McLean, J. E. (1995, April). *The influence of examinee test-taking motivation in computerized adaptive testing*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Kiplinger, V. L., & Linn, R. L. (1995/1996). Raising the stakes of test administration: The impact on student performance on the National Assessment of Educational Progress. *Educational Assessment, 3,* 111-133.

Kong, X. J. (2007). *Using response time and the effort-moderated model to investigate the effects of rapid guessing on estimation of item and person parameters.* Unpublished doctoral dissertation, James Madison University.

Kong, X. J., Wise, S. L., & Bhola, D. S. (2007). Setting the response time threshold parameter to differentiate solution behavior from rapid-guessing behavior, *Educational and Psychological Measurement, 67*, 606-619.

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.

Messick, S. (1984). The psychology of educational measurement. *Journal of Educational Measurement, 21,* 215-237.

O'Neil, Jr. H. F., Sugrue, B., & Baker, E. L. (1995/1996). Effects of motivational interventions on the National Assessment of Educational Progress mathematics performance. *Educational Assessment, 3,* 135-157.

Rothe, H. F. (1947). Distribution of test scores in industrial employees and applicants. *Journal of Applied Psychology*, *31*, 480-483.

Schiel, J. (1996). *Student effort and performance on a measure of postsecondary educational development* (ACT Rep. No. 96-9). Iowa City, IA: American College Testing Program.

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., & Kitsantas, A. (2004). An exploration of the psychology of the examinee: Can examinee self-regulation and test-taking motivation predict consequential and non-consequential test performance? *Contemporary Educational Psychology, 29*, 6-26.

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.

Taylor, C., & White, K. R. (1981, April). *Effects of reinforcement and training on Title I students' group standardized test performance*. Paper presented at the annual meeting of the American Educational Research Association, Los Angeles. (ERIC Document Reproduction Service No. ED206655)

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. (2005). Examinee motivation in low-stakes assessment: Problems and potential solutions. *Educational Assessment, 10,* 1-18.

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. (in press). A clarification of the effects of rapid guessing on coefficient α: A note on Attali's "Reliability of Speeded Number-Right Multiple-Choice Tests." *Applied Psychological Measurement*.

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

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

Wise, V.L. (2004). *The effects of the promise of test feedback on examinee performance and motivation under low-stakes testing conditions.* Unpublished doctoral dissertation, University of Nebraska-Lincoln.

Wise, V. L., Wise, S. L., & Bhola, D. S. (2006). The generalizability of motivation filtering in improving test score validity. *Educational Assessment, 11*, 65-83.

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., & DiPaolo, T. (1996, April). *The effects of test specific motivation and anxiety on test performance*. Paper presented at the annual meeting of the National Council on Measurement in Education, New York.