Listening. Learning. Leading.®

Research Report
ETS RR–15-19

Measuring Motivation in Low-Stakes Assessments

Bridgid Finn

December 2015

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Measuring Motivation in Low-Stakes Assessments

Bridgid Finn

Educational Testing Service, Princeton, NJ

There is a growing concern that when scores from low-stakes assessments are reported without considering student motivation as a construct of interest, biased conclusions about how much students know will result. Low motivation is a problem particularly relevant to low-stakes testing scenarios, which may be low stakes for the test taker but have considerable consequences for teachers, school districts, or educational and governmental institutions. The current review addresses the impact of motivation on assessment scores and research that have identified methods for minimizing the error introduced by unmotivated test takers. A comprehensive review of the test conditions that have been shown to influence motivation is provided. In addition, the review identifies a number of circumstances under which motivation can be enhanced or diminished. The benefits and limitations of the various measurement techniques that have been used to mitigate the negative impact of low-motivation test takers on score interpretation are discussed.

Keywords low-stakes assessment; motivation; filtering; self-report measures

doi:10.1002/ets2.12067

The current review addresses how research has approached the problem of low motivation in low-stakes assessment contexts. Assessments are considered to be low stakes if they are used to measure student achievement but have little or no consequences for the test taker. For example, performance on many low-stakes tests has no impact on the student's grade or class standing. A long history of research has shown that students' test-taking motivation is positively correlated with their performance (e.g., Pintrich & DeGroot, 1990). There is widespread evidence that students are not always highly motivated to perform well on low-stakes tests (Mislevy, 1995; Wolf & Smith, 1995). However, many low-stakes tests are used for accountability and comparison measures, underscoring the importance of considering motivation when interpreting low-stakes test performance. Low motivation poses a serious problem for determining the validity of the test, lowers the credibility of the assessment if it is seen as an invalid measure, and leaves institutions potentially drawing questionable conclusions about the status of their programs and the proficiency of their teaching faculty. A number of research reports have sought to investigate the problems of construct-irrelevant variance introduced when unmotivated students take low-stakes tests. Here, we review the impact of motivation on educational measurement and the ways that researchers have identified to minimize the error that is introduced by unmotivated test-takers.

This report is organized into several sections. The first section provides an overview establishing the wide-ranging effects that student motivation can have on score interpretation and the adverse consequences that can result when motivation is neglected as a construct of interest. We start the section by describing the impact of low motivation on construct validity. Next, the influence that low-stakes assessments have on teacher accountability outcomes is discussed. The second section provides a brief summary of expectancy value theory, a dominant model of achievement motivation that has been effective in relating student motivation to test performance. The third section provides a comprehensive review of test conditions that have been shown to influence motivation and identifies a number of circumstances under which motivation can be enhanced or diminished. The fourth section describes methodologies that researchers have used to identify low-motivation test takers. The benefits and limitations of the various measurement techniques that have been used to mitigate the negative impact of low-motivation test takers on score interpretation are discussed. The fifth section briefly touches on research relating individual differences to low motivation. We end the report with a summary of the main findings, limitations, and directions for future study.

Corresponding author: B. Finn, E-mail: BFinn@ets.org

Impact of Motivation on Test Performance, Validity, and Measures of Accountability

Impact of Low Motivation on Construct Validity

A major aim of educational measurement is to provide a valid index of what students know. However, currently no standard methodology of accounting exists for low effort and low motivation in students' scores on low-stakes tests. Eklöf (2010) noted that

> an achievement test can be viewed as a joint function of skill and will, of knowledge and motivation (also see Cronbach, 1960; Pintrich & DeGroot, 1990). However, when interpreting and using test scores, the will part is not always acknowledged and scores are mostly interpreted and used as pure measures of student knowledge. (p. 345)

Researchers have been giving increasing attention to the finding that unmotivated students can substantially impact low-stakes testing scores as well as the validity of interpretations that are based on these scores (Kane, 2006; O'Reilly & Sabatini, 2013; Wise, Wise, & Bhola, 2006). The basic problem is that if students do not give their full effort on an assessment, but motivation is not considered a construct of interest, the resulting scores may underestimate their levels of proficiency and yield a biased depiction of how they are performing.

In an assessment context, motivation typically refers to the amount of effort that students give to their test responses, with low-effort behavior characterized by guessing, omitting items, and rapid responding (Wise & Kong, 2005). Not surprisingly, low-effort behavior reduces performance and introduces construct-irrelevant variance to the resulting scores (Eklöf, 2006; Swerdzewski, Harmes, & Finney, 2011). The number of students who show low motivation in low-stakes testing contexts is not trivial. A study by Hoyt (2001) found that up to 22% of college students taking a low-stakes general education test reported giving little or no effort on the test. A similar study by Schiel (1996) found that in a sample of 20,000 university students asked to indicate the amount of effort that they had just given on a low-stakes university assessment, up to 28% of participants reported giving little or no effort.

Persistence is related to response time and is also seen as critical to performance on standardized tests (Duckworth, Quinn, Lynam, Loeber, & Stouthamer-Loeber, 2011). It has been measured in several ways: Persistence can reference time spent on a response attempt before moving on or may be a test-wide measure of how many items a student answered over the entire pool of test items. In a recent investigation of the latter measure of persistence, Boe, May, and Boruch (2002) examined the percentage of questions answered out of all questions asked on the Third International Mathematics and Science Study with students in Grades 3, 4, 7, and 8. The percentage was used as a measure of persistence and as a proxy for motivation. Their results showed that persistence, measured with respect to the amount of the test completed, accounted for a large percentage of the national differences in mean math and science scores.

Low motivation and low effort have a substantial impact on performance. Wise and DeMars (2005) reviewed 12 studies that experimentally manipulated test-taker motivation to examine the validity of test scores under low-motivation test scenarios. For example, students in the motivated condition might be told that they would receive extra course points based on their performance, whereas students in the unmotivated condition might be given standard test instructions. The researchers found that students who were motivated outperformed unmotivated students, with differences up to as large as .59 standard deviations. A meta-analysis of more than 50,000 test scores from 200 schools by Schiel (1996) presented students' reported effort (a one-item question) on the Collegiate Assessment of Academic Proficiency, a standardized assessment given in university settings. Schiel found that students who reported having put forth at least reasonable effort scored between .25 and 1.5 standard deviations higher than students who reported expending no effort.

These findings call attention to the dilemma of interpreting results that include scores from unmotivated students (DeMars, 2000; Eklöf, 2006; Wise, 2006a, 2006b; Wise & DeMars, 2005, 2006, 2010). A number of undesirable outcomes may follow from utilizing tests that include scores from low-performance test takers. For instance, if performance standards are based on values from data that underestimate what a student body actually knows, cut scores and difficulty parameters for higher stakes tests may end up being set too low (Wise, Wise, et al., 2006). Furthermore, low-motivation behaviors like rapid guessing distort the psychometric properties of test scores (Wise, Pastor, & Kong, 2009). For example, scores that include low-motivation scores show reduced convergent validity and erroneous increases in the internal consistency of the scores (Wise, 2006a, 2006b; Wise & DeMars, 2006; Wise et al., 2009). Researchers also worry about a contradictory problem — when administrators see low performance on the tests, they may begin to assume that the low

scores they are seeing are because of low motivation, when it may actually be the case that their students are not performing well (Thelk, Sundre, Horst, & Finney, 2009).

Impact of Low Motivation on Teacher Accountability Measures

Another hazard of interpretation and utilization of scores that include low-motivation test takers is that the scores are often tied to accountability measures for individuals or institutions. That is, a test that is low stakes for the examinee may be high stakes for the institution, teacher, or school district that is accountable for student achievement reflected by the scores. The federal education policy No Child Left Behind (NCLB) of 2001 and the federal education fund Race to the Top (RTTT) substantially increased emphasis on K-12 assessment for accountability in an effort to improve public education. For example, the provisions in NCLB require that states establish statewide standards of student achievement in elementary and secondary schools and give yearly assessments of academic proficiencies to determine which schools and districts are making appropriate progress (Linn, 2003). RTTT provides federal funding to states to implement a more rigorous approach to teacher accountability in which teachers are linked to student achievement and student growth data. Growth data refers to the amount of academic progress the student has shown over the course of the year as reflected in scores of assessments given at the beginning and at the end of the year.

In an effort to obtain funding, many states require that students in K-12 take low-stakes assessments that will be used as part of teacher evaluations. Some states link teacher pay to teacher effectiveness measured, in part, by students' performance on low-stakes tests. Arne Duncan, secretary of education, cites performance pay for teachers as the department's "highest priority" (Toch, 2009, para. 1). State legislators in Rhode Island, Delaware, Louisiana, and most recently, Tennessee have passed measures that base licensure renewal on teacher evaluations that depend in part on improved scores on low-stakes tests (Banchero, 2013). Some states do tie student performance on the statewide exams to student outcomes, such as whether the student will pass to the next grade or graduate from high school (Heubert & Hauser, 1999). But K-12 standardized tests are often primarily used as accountability and performance measures at the grade, school, or district level. Schools that do not meet test score targets may become ineligible for state or federal funding and may even be at risk for school closure or takeovers (Goertz & Duffy, 2003; Nichols, Glass, & Berliner, 2012).

During the school year, students may also take the National Assessment on Educational Progress (NAEP) assessments. NAEP, also known as the "Nation's Report Card," tests students in Grades 4, 8, and 12. These tests are used as a measure of U.S. student proficiency in subjects such as math and reading but have no direct consequences for the student. Indeed, the student is never even given feedback about his or her performance on the test. High school seniors in particular have shown low motivation on NAEP and other low-stakes tests (Braun, Kirsch, Yamamoto, Park, & Eagan, 2011; Swerdzewski et al., 2011), which is symptomatic of the trend for students to show lower effort on the test as they move through school. There are also low-stakes international assessments such as the Programme for International Student Assessment, the International Reading Literacy Study, and the Trends in International Mathematics and Science Study, which are used to make international comparisons of student proficiency in a number of educational domains. For example, in 2008, the Organization of Economic Cooperation and Development began international contrasts of learning outcomes using the Assessment of Higher Education Learning Outcomes (AHELO; Liu, Rios, & Borden, 2014).

Higher education has also been the focus of attention of policy makers who want accountability evidence (Liu et al., 2014; U.S. Department of Education, 2006). At the university or college level, assessments are used to provide data on an institution's impact on student achievement. Other assessments, such as the AHELO, may use scores to contrast performance of participating countries (Tremblay, Lalancette, & Roseveare, 2012). Many universities also use assessments for program evaluation to improve the quality of the educational programs at the university. Student scores may be reported to regional accreditors, state councils, university administrators, and faculty committees (Thelk et al., 2009) and can influence policy and funding recommendations. For example, state universities often are required to demonstrate student proficiencies to funding sources such as state legislatures and accrediting agencies that want to see evidence of student learning in college (Cole, Bergin, & Whittaker, 2008; Flowers, Osterlind, Pascarella, & Pierson, 2001; Palomba & Banta, 1999; Stone & Friedman, 2002). A number of accountability initiates have been launched by educational organizations to demonstrate evidence of student learning (Liu et al., 2014; U.S. Department of Education, 2006) and are used by universities and government funding agencies to make decisions about programs and funding for the institution (Sundre, 1999; Sundre & Kitsantas, 2004; Wise & DeMars, 2005). The initiatives appeal to institutions to assess student learning in ways that allow comparisons across institutions. For example, the Voluntary Framework of Accountability requires

participating institutions to use one of the three standardized tests: the Educational Testing Service (ETS) Proficiency Profile, the Collegiate Learning Assessment owned by the Council for Aid to Education, or the Collegiate Assessment of Academic Proficiency, which is owned by American College Testing (ACT). Scores are used to report student learning outcomes and to examine value added by the institutions to the educational experience (Hosch, 2012; Liu et al., 2014; Voluntary System of Accountability, 2008). In sum, low motivation creates a number of problems for interpreting test scores, and there are increasing calls for methods of analysis that account for low motivation to ensure score validity.

The vast majority of the research investigating the impact of motivation on assessment has been done in a university assessment context and is aimed at establishing best practices for managing test scores that include uninterpretable, low-effort scores. To minimize the influence of unmotivated students on test scores, researchers and practitioners have introduced techniques before the examinee takes the test to influence the effort that the examinee gives on the subsequent test (Swerdzewski, Harmes, & Finney, 2009). Another approach has been to examine the impact of filtering the assessment data by removing unmotivated test-takers. In the following, we first review expectancy value theory, which underpins much of the research investigating the conditions of the test that influence motivation. We then turn to the methods of motivational filtering that researchers have identified to improve test validity.

Expectancy Value Theory

Expectancy value theory has been influential in explaining the relationship between test-taking motivation and performance on low-stakes tests. According to the expectancy value framework of achievement motivation (Atkinson, 1957; Eccles & Wigfield, 2002; Feather, 1982; Pintrich, 1989; Pintrich & Schunk, 2002), motivation depends on both the student's expectation for success and the value that the student places on the task, that is, how important or useful the student believes the task to be (Eccles & Wigfield, 2002). The expectation component of expectancy value theory predicts that when there is a discrepancy between individual ability and test difficulty (i.e., the test is harder or easier than the student expected it to be), the student's motivation will be impacted (Eccles & Wigfield, 2002). Wolf, Smith, and Birnbaum (1995) elaborated on the expectancy component of the model as also including an estimation of how much mental effort the task will require, or in other words, how mentally taxing the task will be. Effort here was considered as the amount of mental work the test taker will give when responding to items on the test (Wolf et al., 1995). Other considerations related to expectancy value theory have been brought to bear on test motivation. For example, Pintrich and collaborators (Pintrich, 1989; Pintrich & DeGroot, 1990) have also discussed how the student's affective reactions to the test influence achievement motivation.

In line with expectancy value theory, Cole et al. (2008) explored whether rated interest, usefulness, and importance significantly predicted test-taking effort and whether effort predicted performance. Students took the College Base, a 3.5-hour, standardized university achievement exam that a number of universities require students take prior to graduation but that does not impact their record in any way (Osterlind, Robinson, & Nickens, 1997). Results showed that students' perceived usefulness and importance of taking the test were important predictors of test-taking effort — which predicted test performance. The researchers explored this question separately for the English, Mathematics, Science, and Social Studies subtests. Effort was found to be a strong predictor of test performance in all subjects.

Test Conditions Shown to Influence Motivation

Consequences

A number of studies have investigated the importance of consequences on test-taker motivation. One argument is that when low-stakes tests, such as NAEP or standardized tests used for accountability, have no consequences for the student taking the test, the student has little motivation to give effort. According to Wolf and Smith (1995), tests that obtain norms "under conditions that are not consequential to students may underestimate actual achievement in students" (p. 239; but see National Center for Education Statistics, 2007). The researchers argued that this finding could create serious measurement problems if schools use the norms from low-stakes tests as a comparison against which to measure their own students' performance on tests that did have consequences. In essence, the schools could potentially overstate their abilities relative to the U.S. norms (Wolf & Smith, 1995).

To examine the impact of consequences of test-taker motivation and performance, Wolf and Smith (1995) grouped students taking a college-standardized test into a consequence group and a no consequence group. The consequence group

was told that performance would count toward course grades, and the no consequence group was told that performance would not count toward grades. The consequence group reported higher test-taking motivation and significantly outperformed the no consequence group. Note that the effect size was larger for the motivation ratings than for the actual test performance (1.45 vs. 0.26, respectively). Napoli and Raymond (2004) similarly found that whether the exam would be graded or not graded influenced test performance.

This pattern has been shown in a number of other investigations (Sundre, 1999; Sundre & Kitsantas, 2004; Terry, Mills, & Sollosy, 2008; Wolf & Smith, 1995; Wolf, Smith, & DiPaolo, 1996). For example, other studies have shown that taking a test that has consequences for course placement, promotion, graduation, or admission influences test performance via motivation (Cole & Osterlind, 2008; DeMars, 2000; Liu et al., 2014; Rothe, 1947). Cole et al. (2008) found that college students taking the CollegeBASE exam, an exam that helps in making high-stakes decisions, scored significantly higher than students who took the test as a low-stakes exam. Wolf et al. (1995) showed that 10th graders who were told that the test would be used to determine whether they would be placed into an 11th-grade remedial math class (an undesirable situation) were more motivated than 11th graders who did not have any consequences attached to their performance. Sundre and Kitsantas (2004) found that for tests that had no consequences, motivation was a significant predictor of performance; however, the relationship was absent in the exam with consequences, presumably because most students in the consequences group were highly motivated to perform well.

Given that consequences have a large influence on test-taking effort, one way to encourage students to do their best may be to make sure that the test has consequences for the student. One approach, discussed by Wise and DeMars (2005) as practiced at their institution (James Madison University), is to require students to take and pass the standard university assessment before they can register for the following semester's courses. Another approach that has been proffered is to require students to attain a minimum score to graduate from high school or grade in class (Sundre, 1999).

Some worry that there may be limitations to attaching consequences to all student assessments. For example, a number of researchers acknowledge that increasing consequences for the student may serve to increase test anxiety, which has also been shown to impact negatively on performance and threaten test validity (Cassady & Johnson, 2002). For example, Smith and Smith (2002) found that the best scenario for test taking was a combination of high motivation and low anxiety. In contrast, high motivation – high anxiety students showed the same level of performance as low motivation – low anxiety test takers. Thus, even when students are highly motivated, anxiety can have a cost on performance. A second limitation of making all tests have high-stakes consequences is that assessments that help in making high-stakes decisions are comparatively costly in terms of administration, test development, and analysis required to carry them out (A. R. Brown & Finney, 2011; Wise & DeMars, 2003).

Relevance

Self-Relevance

One problem that has been identified with taking low-stakes tests is that students do not often feel like the test has particular relevance for them. Karmos and Karmos (1984) found that for students in Grades 6 – 9 performance on a standardized achievement test was related to perception of effort, importance of the test, and how the test results would be used. Seegers and Boekaerts (1993) found that effort and performance on a math learning task were related to how personally relevant the 11- and 12-year-old students judged the task to be. Curran and Harich (1993) showed a similar result with undergraduate business students.

One approach that has been offered for increasing the self-relevance of the test is to have the institution sponsoring the test (i.e., the university) provide a certificate of student performance that could potentially be used in job applications (Liu et al., 2014). Similarly, students could list the test on their transcript as having fulfilled a university requirement (Wise & DeMars, 2005). One limitation of these approaches is that the interviewer or admissions officer looking at the transcript or certificate might not know anything about what the test measured.

Another potential method of increasing the perceived self-relevance of the test is to provide students with feedback about their performance. Providing feedback on a low-stakes test may make the test seem less personally meaningless (Peterson & Irving, 2008; Zilberberg, Anderson, Finney, & Marsh, 2013) because, at the very least, the student would get an indication of his or her proficiency in a given area. Though providing students with feedback on their performance seems like an obvious way to make the test more self-relevant, research suggests that it does not have an immediate

impact on motivation. For example, Baumert and Demmrich (2001) found that promising feedback did not improve performance or motivation ratings on the test. It is not clear from these studies, however, whether providing feedback on one test would impact motivation on subsequent tests. Another issue is the format in which the feedback is presented. Providing students with their test scores may not necessarily be useful, as the students might not understand what the particular metric indicates (Wise & DeMars, 2003). Instead, it may be more valuable to provide students with explanatory information about what those scores indicate about their proficiencies as measured by the test.

Motivational Framing

Another approach, closely related to increasing self-relevance and aimed at increasing student motivation by increasing the relevance and importance of the test, has been to frame the test as important for the test taker and/or the institution (Liu et al., 2014; Sundre & Moore, 2002). This method is also closely related to methods that increase test consequences. For example, Sundre and Moore (2002) found that talking with students about the importance of the test increased motivation. More recently, Liu et al. (2014) examined the impact of motivation on the ETS Proficiency Profile, a test that measures college level skills in critical thinking, reading, writing, and mathematics and is widely used by institutions of higher education to report accountability information. The tests used to assess learning gains from the first to second year of college often showed minimal learning (Arum & Roska, 2011). But, as Liu et al. noted, the tests that have been used in these analyses were low stakes for the students. They argued that the small learning gains that have been shown may be more attributable to students' lack of motivation than to their lack of learning.

Liu et al. (2014) were interested in whether motivational instructions would affect student motivation and performance. To investigate, the researchers contrasted three motivational conditions in which students were either told that (a) the test would only be used for research purposes (control condition), (b) their scores may be released to faculty or potential employers (personal condition), or (c) their scores on average may be used to evaluate the quality of instruction at the college (institutional condition). Self-reported motivation was also collected. A total of 757 students took a 36-item short-form version of the ETS Proficiency Profile. The students were also given an essay-writing task, which is an optional part of the ETS Proficiency Profile that measures college level writing ability. After the test, the students completed the student opinion scale (SOS), a self-report questionnaire of test-taking motivation.

As was consistent with other research, results showed that self-reported motivation showed a significant relationship with both multiple choice and essay scores. This finding was true even when controlling for college admission scores and placement test scores. More relevant, the instructions were shown to influence student motivation and, as a result, scores on both the multiple choice and essay tests. Students in both the personal and the institutional consequences conditions outperformed the control condition. Notably, there was no difference in performance for the personal and institutional consequences groups. Thus institutional framing was equally motivating to students as the framing condition in which the personal consequences were highlighted but may give rise to more test anxiety.

Cole et al. (2008) suggested that staff and test administrators communicate the importance and usefulness of the test to the student body. However, one limitation of the instructional framing approach is that framing may not be successful in getting students to sign up to take a voluntary exam. Hosch (2012) found that efforts to recruit seniors at a regional university in Connecticut to participate in the accountability assessments were not improved by emphasizing the benefits of seeing their performance compared with the performance of seniors nationwide or to the benefit of the institution gathering useful information. We discuss noncompliance in further detail later in the review.

Incentives

In general, incentives have been used in two ways with respect to low-stakes assessment: to get students to sign up to take the test or to get students to give effort while taking it. To encourage voluntary participation in assessments, some universities have used incentives such as movie passes, gift cards, free food, or a discount or full waiver on graduation fees and graduation regalia (Hosch, 2012; Palomba & Banta, 1999). Research has also investigated the impact of performance-based incentives, such as prizes, public recognition (Pedulla et al., 2003), and compensation for higher performance (Baumert & Demmrich, 2001; Braun et al., 2011; Duckworth et al., 2011; O'Neil, Abedi, Miyoshi, & Mastergeorge, 2005; O'Neil, Sugrue, & Baker, 1996), on test motivation. For example, O'Neil et al. (1996) showed that financial incentives increased eighth-grade NAEP test scores. In their study, 8th and 12th graders were randomly assigned to either a financial incentive

group, in which students received $1 for every correct item; an ego instruction group, in which students were told that the goal of the test was to compare students' math abilities; a task instruction group, in which students were told that the goal of the test was to provide opportunity for personal accomplishment; or a control group, which was not told any additional information about the test. With the eighth-grade students, the financial incentives condition showed greater reported effort, and these students scored higher on the exam compared with the other groups, which were not significantly different from one another. However, 12th-grade students showed no benefit of financial incentives on either effort or performance. These results were largely replicated by Braun et al. (2011). Duckworth et al. (2011) and Sundre (1999) have shown that performance-based compensation motivates students. However, Baumert and Demmrich (2001) reported that financial rewards do not improve performance or reported effort. Together, these studies paint an unclear picture of if and when incentives enhance test-taking effort. Moreover, incentives can be expensive, and some institutions may find it unethical to pay students to complete a test. Critics of incentive approaches have described these kinds of inducements as fundamentally artificial (e.g., Wolf & Smith, 1995).

Test Difficulty

Wolf et al. (1995) found that performance differences between lower and higher motivation students were primarily found on those items that were the most challenging. That is, more moderate items showed similar levels of performance for low- and high-motivation students. Performance diverged when the item required more effort. Likewise, Sundre (1999) and DeMars (2000) found that performance on high-effort items, such as essay questions, suffers more when motivation is low (Wise & DeMars, 2005). Asseburg and Frey (2013) found that, in general, test-taking effort was greater and boredom was reduced when the items were more moderate.

Wise and DeMars (2005) noted that the challenge is then to create items that measure higher order thinking but are not mentally taxing. They argue that items can become mentally taxing when they require that the student completes many lower-order steps. Wise, Wise, et al. (2006) suggested that assessments should be designed to be more intrinsically motivating by being moderately challenging—but not too difficult. This idea is in line with motivation research showing that students are most motivated by moderately challenging tasks (Pintrich & Schunk, 2002).

Wise et al. (2009) found that items that have more text, that have more response options, and that depend on information that occurs later in the text are related to increases in rapid guessing, a typical form of low-motivation test-taking behavior. Items that included a graphic showed decreased rapid guessing and mitigated the impact end of test effects on rapid guessing behavior. A study by DeMars (2007) found that over a series of assessments (university students were required to take eight tests over 4 weeks), later tests had more rapid guessing and worse performance. The challenge with these approaches is making modifications without oversimplifying or changing the construct coverage and, hence, diminishing construct validity. To the extent possible, assessment tasks should correspond to the full range of stimulus and task demands one might find outside the test context. This may include facing challenging and complex task environments and items.

Wise, Bhola, and Yang (2006) looked at the impact of effort monitoring on low-motivation behaviors such as rapid responding. While a test was being administered, students were given warning messages if the computer had identified rapid guessing behavior. Tests conducted with such warnings in place yielded higher performance scores and had higher validity than a traditional computer-based test (Kong, Wise, Harmes, & Yang, 2006; Wise, Bhola, et al. 2006).

Students tend to respond more rapidly at the end of the test (DeMars, 2007). Rapid responding could be occurring because students are running out of time at the end of the test or because they are fatigued and are no longer giving effort to their responses. One suggestion for reducing the problem caused by rapid responding is to make the test have fewer items. DeMars (2007) has also suggested that when students are expected to take many low-stakes tests, matrix sampling can reduce the total number of tests that the student has to take.

Motivation Filtering

As an alternative to introducing techniques and inducements to improve motivation, researchers have tried to identify indicators of low motivation and low effort on the test so that they can filter out total examinee scores from data analyses. Two widespread indicators of test-taking motivation are self-report measures of motivation and test-timing data (Sundre & Moore, 2002; Wise & DeMars, 2005; Wise, Wise, et al., 2006). Self-report measures typically determine the students'

self-perceived levels of motivation for the test specifically and academic motivation more generally. A second approach to motivation filtering has been to use test-timing data, such as the response time to complete a question, to determine whether students are investing effort in the exam. Self-report and response times are used as indirect indicators to make inferences about the complex construct of motivation.

Motivation filtering, developed by Sundre and Wise (2003) and Wise and DeMars (2005), is a method of identifying and removing low-motivation students so that only meaningful scores will be included in the analyses, with the goal of improving the validity of the inferences that can be drawn from the scores. The idea behind filtering is that students who invest low effort will have artificially low test scores. Critically, filtering also assumes that motivation is related to test performance and that motivation is not related to ability. Nevertheless, one concern with using filtering is that those students who tend to show low motivation and low effort may just be less proficient students. If the variables used to determine motivation filters were biased toward primarily low-ability student scores, the scores remaining after the filter had been applied would be an unrepresentative sample of the test group. If this were the case, then filtering out these students' scores would mean that a subset of the true sample was being excluded and scores would be inflated. Wise and collaborators have investigated this issue and have not found a relationship between effort and academic ability (Sundre & Wise, 2003; Wise & DeMars, 2005; Wise & Kong, 2005; Wise, Wise, et al., 2006).

Global Self-Report Measures

Global self-report measures are ratings by the test taker of the motivation and effort that the test taker thinks he or she exerted over the whole test (Sundre & Moore, 2002; Wolf et al., 1995). Wolf and Smith (1995) found that students' ratings on a global motivation questionnaire taken at the end of a standardized testing session were correlated with their performance on the test. Researchers have since used global self-report of motivation to apply motivation filtering to their test data. A number of self-report measures have been used to investigate test-taking motivation, such as the Motivated Strategies for Learning Questionnaire (Pintrich, Smith, Garcia, & McKeachie, 1993) and the Test-Taking Motivation Questionnaire (Eklöf, 2006). We focus on the SOS, which is the self-report measure that has been most extensively utilized in the context of motivational filtering.

The SOS was developed by Wolf and Smith (1995) and revised and extended by Sundre (1999). The SOS is a short, Likert-style, 10-item self-report measure of motivation that is usually used at the end of an assessment. It measures effort (five items) and personal importance (five items) and either subscale can be used separately without reduction in reliability (Sundre & Moore, 2002). A number of studies have demonstrated evidence of validity for the measure (Sundre, 1999; Sundre & Moore, 2002; Sundre & Wise, 2003).[1]

Limitations of Self-Report Methodologies

Swerdzewski et al. (2011) warned that while global self-report methods are easy to implement and score, they suffer from the pitfalls of all self-report measures, namely, that they require that the student accurately be able to report his or her level of motivation. Self-report measures have been shown to have problems with validity because people can demonstrate inaccurate self-knowledge (Kruger & Dunning, 1999; Ziegler, MacCann, & Roberts, 2011). In addition, as with all measures of self-report, it is difficult to ascertain whether test takers are being truthful when reporting their test effort. They may show response biases; for example, they may say that they gave high effort to avoid punishment or they may not be sensitive to changes in effort during tests (Wise & Kong, 2005; Wise & Ma, 2012). A final caveat: if students give low effort on the test itself, how can researchers and test administrators ensure that they are not also giving low effort on the motivation measure?

Variables That Influence Ratings on the Student Opinion Scale

Wise and DeMars (2003) administered the SOS after students took a low-stakes general education assessment. As is described in greater detail later, the motivation scores were rank ordered from low to high, and students with the lowest level of reported motivation were removed from the data. Results show that as the filter became more stringent (i.e., a higher motivation score was needed to be included in the data analyses), overall measures of performance significantly increased (Sundre & Wise, 2003; Wise & DeMars, 2005; Wise, Wise, et al., 2006). When students take the SOS after taking

a test that helps in making high-stakes decisions, very little variability in motivation ratings is shown in comparison with that seen when the measure is administered after a low-stakes assessment (Wise & DeMars, 2003). With a test that helps in making high-stakes decisions, most students report high levels of motivation, effort, and importance. Higher levels of self-reported motivation are reported on the SOS in settings where motivation-enhancing strategies are used, such as a welcoming speaker or talking to students about the assessment (Sundre & Moore, 2002). Lower motivation is reported in settings when problems occur with administering the test, for example, if incorrect answer sheets are handed out (Wise & DeMars, 2003). Lower motivation is also reported when the test includes difficult tasks, such as essay writing (Wise & DeMars, 2003).

Further research by Thelk et al. (2009) evaluated the relationship between effort reported on the SOS and a test of quantitative and scientific reasoning with community college students and found a positive correlation between effort and performance. In addition, the effort subscale was shown to be positively correlated with response time, another measure of motivation. When the test taken was high stakes, effort scores on SOS increased. Thelk et al. concluded that higher reported effort for tests that help in making high-stakes decisions indicates that SOS scores provide an accurate measure of effort.

Response Time

Researchers have also focused on response time data as another index of test-taking motivation. One argument in favor of response time as a measure of motivation as opposed to self-report measures of motivation is that response times can be measured without student awareness and so are a more objective marker of motivation on the test.

The initial research on rapid responding centered on high-stakes, timed tests and showed that as the time limit of the test approached, students often switched from trying to work out answers to rapidly responding (Schnipke & Scrams, 1997, 2002). This trend has been termed *rapid guessing behavior*. Performance on items answered by rapid guessing was generally at chance levels. Rapid guessing, at least on tests that help in making high-stakes decisions, was thought to be a strategy to attempt to get some items correct that otherwise would have been wrong if left blank (Schnipke & Scrams, 1997). Subsequent studies on response time found that rapid guessing behavior occurs during low-stakes, untimed tests and also results in chance-level performance (Wise & Kong, 2005). In the case of low-stakes tests, the student's strategy seems to be one of ending the exam as quickly as possible rather than attempting to maximize performance. As might be expected, overall test time also shows a relationship to performance. Recently, Hosch (2012) found that the total amount of time needed for taking a low-stakes college assessment was related to test performance, with those who scored higher taking longer on the test. In addition, students who spent more time on the test outperformed their expected scores, which were based on their combined scores on the *SAT*® test.

A number of researchers have examined response time as an index of effort (e.g., Schnipke & Scrams, 1997; Wise & Kong, 2005) and focus on response times at the item level. When tests are administered on the computer, response times can be calculated for each question and can be used as an index of student effort. As students take a test, research has shown that they will primarily engage in solution behavior or rapid guessing behavior, in which they scroll through items and randomly click on response alternatives (Thelk et al., 2009). Thus, if a student is unmotivated to take the test, he or she will move through many of the items quickly and fail to demonstrate solution behavior, in which the student reads each item and considers response alternatives thoughtfully. Motivation filtering can be used to exclude meaningless data from students who have been identified as disengaged via response times that are too fast to represent a meaningful response.

In an effort to develop a filter metric of test effort based on response time, Wise and Kong (2005) developed response time effort (RTE). RTE is a score that represents the proportion of the items for which the student showed solution behavior as opposed to rapid guessing behavior. The scores range from 0 at the minimal end of effort to 1, which represents maximum effort. To determine the RTE score, a response time threshold designating the minimum amount of time required to read and respond is computed for each item in the test and is compared to the amount of time a student actually spends on each item. RTE is a measurement of the proportion of items on the test for which students' response times exceed the threshold, that is, the proportion of items that they spent more than the minimum amount of time needed to answer the question. Students' data can be filtered out if their RTE is lower than the cutoff that the researchers or practitioners establish.

Setting the Threshold

In determining the response time threshold, the goal is to identify as many noneffortful items as possible while avoiding marking effortful responses as noneffortful. A number of approaches have been taken toward threshold setting. Wise and Kong (2005) presented an extensive discussion of RTE threshold setting. Briefly, Wise and Kong based thresholds on surface features of the items such as length. Wise (2006a) looked at response time frequency distribution and selected time thresholds that corresponded to the end of the initial frequency spike. Schnipke and Scrams (1997) argued that rapid guessing behavior and effortful answering show different response time distributions and used a two-state mixture model to identify the time threshold of each item. A comparison of Schnipke and Scrams (1997) and Wise (2006a) found that they produced similar thresholds (Kong, Wise, & Bhola, 2007). See also Wise and Ma (2012) who developed the normative threshold method to determine the time threshold for each test item. The normative threshold model determines the threshold by the percentage of elapsed time between the onset of the item and the mean of the response time distribution for that item up to a maximum of 10 seconds. They discussed the benefits of the normative model to the more commonly used threshold to identify low-effort test taking.

Studies examining the efficacy of RTE have typically found a bimodal distribution with a small distribution of rapid responders at the lowest end of the response time continuum and a large distribution comprising the rest of the response time continuum (Swerdzewski et al., 2011). For example, Wise, Kingsbury, Thomason, and Kong (2004) found that of more than 2,300 tested middle and high school students, only 27 students had RTE scores that had reached the threshold. Interestingly, 23 of the 27 were male. (We discuss gender difference in low-stakes testing motivation in more detail later in the review.) RTE has been shown to be significantly correlated with self-reported motivation and test performance but is not related to student ability. Wise and Kong (2005) showed that when total scores on a low-stakes test were filtered out of the data set based on RTE, there were no differences in the average SAT scores (presumably a high-effort test for students) before the filter was applied and after it was applied. Though removing scores based on an RTE threshold should result in higher mean SAT test scores for the remaining students in the data set (if those same students showed similar low motivation on the SAT), the Wise and Kong findings suggest that low effort is equally likely to occur for students anywhere in the ability distribution, providing convergent evidence that low motivation is not only a phenomenon that occurs among low-ability students. Converging research evaluating the psychometric properties of the RTE scores acquired from a computer-based university assessment found high internal consistency reliability ($\alpha = .97$) and convergent validity, as demonstrated by significant positive correlations with self-report and person-fit statistics, as well as near-zero correlations with SAT scores and motivational filtering effects similar to those found with other filtering methods.

Reaction Times Versus Self-Report Measures

Does one method do a better job of improving scores while filtering the fewest students? Wise and Kong (2005) compared the efficacy of filters that were based on reaction times or self-reported motivation on university assessments. In both cases, removing data led to improved performance on the assessments. However, their findings favored the response time filter because fewer students ended up being filtered from the data than when the self-report measure was used. Swerdzewski et al. (2011) found that filters based on self-report and filters based on reaction times were fairly consistent in identifying low-motivation students. Similar to the findings of Wise and Kong (2005), when there were differences between the measures, the self-report measure removed more student data than the response item measure, but even so, there were no meaningful differences in mean scores by filter type. Converging findings come from Rios, Liu, and Bridgeman (2014) who similarly found that response time measures were more accurate in identifying unmotivated examinees than were self-report measures.

Pros and cons are associated with each approach. Response time is an objective measure, has high internal consistency, allows analysis of effort at the item level, and enables evaluations of how effort is changing over the course of the test (Wise & Ma, 2012). An important limitation of the response time measure is that response times can only be collected if the test is computer administered. In contrast, self-report measures can be collected with pencil-and-paper tests or computer-administered tests. However, researchers urge practitioners to consider how even the measure of test-taking motivation could be impacted by low motivation or demand characteristics. Despite the advantages and disadvantages afforded by each measure, proponents of filtering agree that using either measures of RTE or self-report questionnaires in

motivation filtering improves validity and enhances inferences based on test scores (Wise & DeMars, 2005, 2010; Wolf & Smith, 1995).

Other Filtering Methods

Several other methods of filtering are based on motivational variables. Lee and Chen (2011) provided a comprehensive review of various filtering methods that are based on reaction time data. Other methods include filtering based on performance relative to that which was expected as compared to self-reported effort (Steedle, 2014). Steedle evaluated this filtering method with the Collegiate Learning Assessment (CLA), an open-ended test of general college education. Results showed that both types of filters improved CLA scores but that the relative to expected method did so by filtering fewer students. Wise (2006a) developed response time fidelity, which filters by effort at the item level.

Statistical Modeling to Identify Low Motivation

A number of researchers have suggested that student motivation be incorporated into measurement models (Frary, Tideman, & Watts, 1997; Karabatsos, 2003; Wise & DeMars, 2005, 2006; for a review, see Zerpa, Hachey, van Barneveld, & Simon, 2011). For example, might it be possible to expand item response theory (IRT) to include effort? Wise and DeMars (2005) worried that using effort in IRT modeling could only work for low-stakes tests because students could learn to report lower levels of motivation so that the score would be adjusted up to account for it. Another IRT modification that has been offered is for IRT to specify item characteristic curves for solution behavior and for rapid guessing behavior (Wise & DeMars, 2006). This type of model frequently fits the response data better than a standard IRT model. Another modeling approach uses the *lz* index statistic, which was developed by Drasgow, Levine, and Williams (1985). The *lz* index is a standardized statistical estimate used to detect the percentage of low-motivation test-taking behaviors like guessing and omissions in tests that help in making high-stakes and low-stakes decisions. The index only uses test scores and does not account for effort. Other model adjustments include using person-fit measures that identify when a test taker has an aberrant response pattern. Others have suggested using individual differences, such as achievement goals, personality, and ability to model low effort (Barry, Horst, Finney, Brown, & Kopp, 2010). See Lee and Chen (2011) for a recent review of statistical models that account for low motivation via response times in tests that help in making high-stakes and low-stakes decisions.

Individual Difference Predictors and Indicators of Low Motivation

Noncompliance

Another threat to the validity of low-stakes assessment is student noncompliance. For example, only 55% of 12th-grade students selected actually participated in taking the NAEP tests (National Commission on NAEP 12th Grade Assessment and Reporting, 2004). Higher education accountability results are also threatened by low compliance (A. R. Brown & Finney, 2011; Ewell, 2008; Wall-Smith, 2006). If a large number of students does not show up for the test, the scores will not represent the knowledge and abilities of the true student population, leading to problems in assessing the validity of test scores (A. R. Brown & Finney, 2011; Swerdzewski et al., 2009).

A. R. Brown and Finney (2011) recently explored the impact of noncompliance on test scores at their home university, where all first-year students are required to take a 3-hour assessment that measures general education skills. They are also required to take the same tests in their second or third year so that growth can be determined. Despite the fact that a hold will be placed on their records if they do not take the test, many students skip the scheduled assessment day and attend a makeup session. A. R. Brown and Finney deemed these students as noncompliant. Noncompliance is thought to indicate very little motivation to take the test, and indeed, noncompliant students who take the makeup test have lower scores than do compliant students (A. R. Brown & Finney, 2011). The researchers tested the hypothesis that students who do not attend the original testing session are more reactant than compliant examinees. Reactance refers to a reaction to threats to perceived loss of freedoms (Brehm & Brehm, 1981). Noncompliance is a common response to feelings of reactance (Brehm & Brehm, 1981). Reactance was measured with the Hong Psychological Reactance Scale (Hong & Page, 1989). DeMars (2000) found that male students were less likely to attend the regular session, with 30% versus 22% of male and

female students, respectively, failing to attend the initial testing session. Some research suggests that communicating the importance of the exams (S. M. Brown & Walberg, 1993; Liu et al., 2014) may work to increase compliance.

Boredom

A number of studies have looked at how boredom impacts effort and motivation (Acee et al., 2010; Asseburg & Frey, 2013; Kim & Pekrun, 2014). Boredom, typically measured through self-report, is a negative affective state characterized by feelings that the task is unchallenging and as though one's actions have little value in the current situation (Pekrun, Goetz, Daniels, Stupnisky, & Perry, 2010; Van Tilburg & Igou, 2011). Pekrun (2006) argued that experiencing a mismatch between task difficulty and individual ability can give rise to boredom and highlights the relevance of boredom in achievement motivation. In general, effort and boredom show a negative relationship (Acee et al., 2010). Asseburg and Frey (2013) explored the relationship between effort and self-reported boredom during a low-stakes assessment with more than 9,500 German ninth-grade students. The researchers found that test-taking effort and boredom were negatively correlated ($r = -.44$). They also found a negative correlation between boredom and math performance.

Gender Differences

A variety of gender differences related to motivation have been shown on low-stakes tests. DeMars, Bashkov, and Socha (2013) analyzed data from students taking a low-stakes university assessment. The researchers found that there were more men in the band of lowest reported effort on reported motivation in self-report measures using the SOS. Other studies have shown that female students reported having higher effort, more cooperativeness, and greater test-taking motivation (Butler & Adams, 2007; Cole et al., 2008; Eklöf, 2007; Karmos & Karmos, 1984; O'Neil et al., 2005; Wise et al., 2009). DeMars et al. (2013) found that male students had more rapid response behavior than female students, an indicator of low motivation. Pattern making with responses (ABCABCABC) is another indication of rapid guessing behavior and has been shown to be much more prevalent among male than female students on NAEP (Freund & Rock, 1992). Note, however, that Wise et al. (2009) did not find gender differences in response times on a low-stakes test.

Conscientiousness and Agreeableness

Though there is not extensive research on the link between personality measures and test-taking effort, a few studies have demonstrated that personality dimensions are related to test-taking effort (Ackerman & Kanfer, 2009; Barry et al., 2010; DeMars et al., 2013; Liem, Lau, & Nie, 2008; Yeo & Neal, 2008). For example, as might be expected, conscientiousness and agreeableness have a positive relationship with test-taking effort. Conscientiousness has also been shown to be related to lower levels of fatigue throughout extended low-stakes testing sessions (Ackerman & Kanfer, 2009). Women score higher on both conscientiousness and agreeableness (DeMars et al., 2013). The findings suggest that personality traits may be a fruitful area of further research to explain the differences in test-taking motivation between male and female students. Gender, in particular, is a stable characteristic that may prove resistant to interventions designed to increase motivation in low-stakes testing scenarios (Barry et al., 2010; Zilberberg, Brown, Harmes, & Anderson, 2009).

Discussion

The current review focused on research investigating the interactions between the test and the test taker and the interventions and analysis solutions that researchers have identified to mitigate construct-irrelevant variance introduced by unmotivated test takers. A number of interventions, some more effective than others, have been used before the start of the test in attempts to increase motivation at the front end. For example, many institutions have offered students performance incentives that have been shown to have varying success in increasing motivation. Motivational framing may offer more promise. Liu et al. (2014) found that instructions emphasizing the importance of the scores for the reputation of test takers' institutions were effective in increasing test takers' motivation and, accordingly, their test scores. Instructional framing highlighting the consequences for the institution was equally as effective as instructions that highlighted the consequences of the test for the individual. Additional studies should be conducted to further investigate the impact of framing on test motivation.

Motivational filtering, in which motivation scores are obtained via students' self-reported motivation or their test response times, is a technique of managing the error introduced by unmotivated test takers once the test is complete. Filtering methods based on self-report or response times each have particular advantages. For example, self-report is well suited to pencil-and-paper tests, whereas response times can be measured objectively and unobtrusively as the student takes the test. There is not yet consensus on whether one method provides a stronger approach to filtering. The research does demonstrate that both methods of filtering yield more valid test scores than those obtained without filtering.

There is growing concern that when scores are reported without consideration of motivation as a construct of interest, biased conclusions about student knowledge will result. Low motivation is a problem particularly relevant to low-stakes testing scenarios, which may be low stakes for the test taker but have considerable consequences for teachers, school districts, or institutions. Nearly 15 years ago, in its *Guidelines on Test Use*, the International Testing Commission called for test users to consider "qualities which may have artificially lowered or raised results when interpreting scores" (International Test Commission, 2013, p. 21, Guideline 2.7.7). Barry et al. (2010) argued that test-taking effort is such a quality and should be examined and reported. They are not alone in this sentiment. The Standards for Educational and Psychological Testing, which have been developed by the American Educational Research Association, the American Psychological Association, and the National Council for Measurement in Education, recommend collecting and reporting measures of effort and using those measures in interpreting test scores. Increasingly, researchers have been arguing that measuring student motivation is critical to determining whether a test has measured knowledge or motivation (Eklöf, 2010; O'Reilly & Sabatini, 2013). The current review demonstrates that motivation should be an essential consideration in evaluating and interpreting low-stakes testing scores and highlights areas of promise, such as motivational framing, that may have a positive impact on motivation with few costs.

Note

1 The SOS and instructions for administration can be found at http://www.jmu.edu/assessment/resources/Overview.htm.

References

Acee, T. W., Kim, H., Kim, H. J., Kim, J. I., Chu, H. N. R., Kim, M., & Wicker, F. W. (2010). Academic boredom in under- and over-challenging situations. *Contemporary Educational Psychology*, *35*(1), 17 – 27. doi:10.1016/j.cedpsych.2009.08.002

Ackerman, P. L., & Kanfer, R. (2009). Test length and cognitive fatigue: An empirical examination of effects on performance and test-taker reactions. *Journal of Experimental Psychology: Applied*, *15*(2), 163 – 181.

Arum, R., & Roska, J. (2001). *Academically adrift: Limited learning on college campuses*. Chicago, IL: University of Chicago Press. doi: 10.7208/chicago/9780226028576.001.0001

Asseburg, R., & Frey, A. (2013). Too hard, too easy, or just right? The relationship between effort or boredom and ability-difficulty fit. *Psychological Test and Assessment Modeling*, *55*, 92 – 104.

Atkinson, J. W. (1957). Motivational determinants of risk-taking behavior. *Psychological Review*, *64*, 359 – 372. doi:10.1037/h0043445

Banchero, S. (2013, August 13). Teachers face license loss: Tennessee to join three other states in yanking privileges based on student scores. *Wall Street Journal*. Retrieved from http://online.wsj.com/news/articles/SB10001424127887323455104579014764151835816

Barry, C. L., Horst, J., Finney, S. J., Brown, A. R., & Kopp, J. P. (2010). Do examinees have similar test-taking effort? A high-stakes question for low-stakes testing. *International Journal of Testing*, *20*, 342 – 363. doi:10.1080/15305058.2010.508569

Baumert, J., & Demmrich, A. (2001). Test motivation in the assessment of student skills: The effects of incentives on motivation and performance. *European Journal of Psychology of Education*, *16*(3), 441 – 462. doi:10.1007/BF03173192

Boe, E. E., May, H., & Boruch, R. F. (2002). *Student task persistence in the Third International Mathematics and Science Study: A major source of achievement differences at the national, classroom, and student levels*. Philadelphia, PA: University of Pennsylvania, Center for Research and Evaluation in Social Policy.

Braun, H., Kirsch, I., Yamamoto, K., Park, J., & Eagan, M. K. (2011). An experimental study of the effects of monetary incentives on performance on the 12th-grade NAEP reading assessment. *Teachers College Record*, *113*(11), 2309 – 2344.

Brehm, S. S., & Brehm, J. W. (1981). *Psychological reactance: A theory of freedom and control*. New York: Academic Press.

Brown, A. R., & Finney, S. J. (2011). Low-stakes testing and psychological reactance: Using the Hong Psychological Reactance Scale to better understand compliant and non-compliant examinees. *International Journal of Testing*, *11*, 348 – 270. doi: 10.1080/15305058.2011.570884

Brown, S. M., & Walberg, H. J. (1993). Motivational effects on test scores of elementary students. *Journal of Educational Research*, *86*(3), 133–136. doi:10.1080/00220671.1993.9941151

Butler, J., & Adams, R. J. (2007). The impact of differential investment of student effort on the outcomes of international studies. *Journal of Applied Measurement*, *8*, 279–304.

Cassady, J. C., & Johnson, R. E. (2002). Cognitive test anxiety and academic performance. *Contemporary Educational Psychology*, *27*(2), 270–295. doi:10.1006/ceps.2001.1094

Cole, J. S., Bergin, D. A., & Whittaker, T. A. (2008). Predicting student achievement for low stakes tests with effort and task value. *Contemporary Educational Psychology*, *33*, 609–624. doi:10.1016/j.cedpsych.2007.10.002

Cole, J. S., & Osterlind, S. J. (2008). Investigating differences between low-and high-stakes test performance on a general education exam. *The Journal of General Education*, *57*(2), 119–130.

Curran, M. T., & Harich, K. R. (1993). Performance attributions: Effects of mood and involvement. *Journal of Educational Psychology*, *85*, 605–609. doi:10.1037//0022–0663.85.4.605

DeMars, C. E. (2000). Test stakes and item format interactions. *Applied Measurement in Education*, *13*(1), 55–77. doi:10.1207/s15324818ame1301_3

DeMars, C. E. (2007). Changes in rapid-guessing behavior over a series of assessments. *Educational Assessment*, *12*(1), 23–45. doi:10.1207/s15326977ea1201_2

DeMars, C. E., Bashkov, B. M., & Socha, A. B. (2013). The role of gender in test-taking motivation under low-stakes conditions. *Research and Practice in Assessment*, *8*, 69–82.

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*, *38*(1), 67–86. doi:10.1111/j.2044-8317.1985.tb00817.x

Duckworth, A. L., Quinn, P. D., Lynam, D. R., Loeber, R., & Stouthamer-Loeber, M. (2011). Role of test motivation in intelligence testing. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(19), 7716–7720. doi:10.1073/pnas.1018601108

Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology*, *53*(1), 109–132.

Eklöf, H. (2006). Development and validation of scores from an instrument measuring student test-taking motivation. *Educational and Psychological Measurement*, *66*, 643–656. doi:10.1177/0013164405278574

Eklöf, H. (2007). Test-taking motivation and mathematics performance in TIMSS 2003. *International Journal of Testing*, *7*, 311–326. doi:10.1080/15305050701438074.

Eklöf, H. (2010). Skill and will: Test-taking motivation and assessment quality. *Assessment in Education: Principles, Policy, and Practice*, *17*(4), 345–356. doi:10.1080/0969594X.2010.516569

Ewell, P. T. (2008). Assessment and accountability in America today: Background and context. *New Directions for Institutional Research*, *2008*(S1), 7–17. doi:10.1002/ir.258

Feather, N. T. (1982). Human values and the prediction of action: An expectancy-value analysis. In N. T. Feather (Ed.), *Expectations and actions: Expectancy-value models in psychology* (pp. 263–289). Hillsdale, NJ: Lawrence Erlbaum.

Flowers, L., Osterlind, S. J., Pascarella, E. T., & Pierson, C. T. (2001). How much do students learn in college? Cross-sectional estimates using the College BASE. *Journal of Higher Education*, *72*, 565–583. doi:10.2307/2672881

Frary, R., Tideman, T., & Watts, T. (1997). Indices of cheating on multiple-choice tests. *Journal of Educational Statistics*, *6*, 152–165. doi:10.2307/1164808

Freund, D. S., & Rock, D. A. (1992). *A preliminary investigation of pattern-marking in 1990 NAEP data*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA. Retrieved from ERIC Document Reproduction Service. (ED 347189)

Goertz, M., & Duffy, M. (2003). Mapping the landscape of high-stakes testing and accountability programs. *Theory Into Practice*, *42*, 4–11. doi:10.1207/s15430421tip4201_2

Heubert, J. P., & Hauser, R. M. (Eds.) (1999). *High stakes: Testing for tracking, promotion, and graduation*. Washington, DC: National Academy Press.

Hong, S. M., & Page, S. (1989). A psychological reactance scale: Development, factor structure and reliability. *Psychological Reports*, *64*(3c), 1323–1326. doi:10.2466/pr0.1989.64.3c.1323

Hosch, B. J. (2012). Time on test, student motivation, and performance on the Collegiate Learning Assessment: Implications for institutional accountability. *Journal of Assessment and Institutional Effectiveness*, *2*, 55–76.

Hoyt, J. E. (2001). Performance funding in higher education: The effects of student motivation on the use of outcomes tests to measure institutional effectiveness. *Research in Higher Education*, *42*(1), 71–85.

International Test Commission. (2013). *International guidelines for test use*. Retrieved from http://www.intestcom.org/Guidelines/Test+Use.php

Kane, M. (2006). Content-related validity evidence in test development. In T. M. Haladyna & S. M. Downing (Eds.), *Handbook of Test Development* (Vol. 1, pp. 131–153). Mahwah, NJ: Lawrence Erlbaum.

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty six person-fit statistics. *Applied Measurement in Education*, *16*, 277 – 298. doi:10.1207/S15324818AME1604_2

Karmos, A. H., & Karmos, J. S. (1984). Attitudes toward standardized achievement tests and their relation to achievement test performance. *Measurement and Evaluation in Counseling and Development*, *17*, 56 – 66.

Kim, C., & Pekrun, R. (2014). Emotions and motivation in learning and performance. In *Handbook of research on educational communications and technology* (pp. 65 – 75). New York: Springer.

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, CA.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, *77*(6), 1121 – 1134. doi:10.1037//0022 – 3514.77.6.1121

Lee, Y. H., & Chen, H. (2011). A review of recent response-time analyses in educational testing. *Psychological Test and Assessment Modeling*, *53*(3), 359 – 379.

Liem, A. D., Lau, S., & Nie, Y. (2008). The role of self-efficacy, task value, and achievement goals in predicting learning strategies, task disengagement, peer relationship, and achievement outcome. *Contemporary Educational Psychology*, *33*(4), 486 – 512.

Linn, R. L. (2003). Accountability: Responsibility and reasonable expectations. *Educational Researcher*, *32*(7), 3 – 13. doi:10.3102/0013189X032007003

Liu, O. L., Rios, J. A., & Borden, V. (2014, April). *The effect of motivational instruction on college students' performance on low-stakes assessment*. Paper presented at the American Educational Research annual meeting, Philadelphia, PA.

Mislevy, R. J. (1995). Test theory and language-learning assessment. *Language Testing*, *12*(3), 341 – 369. doi:10.1177/026553229501200305

Napoli, A. R., & Raymond, L. A. (2004). How reliable are our assessment data? A comparison of the reliability of data produced in graded and un-graded conditions. *Research in Higher Education*, *45*(8), 921 – 929. doi:10.1007/s11162-004-5954-y

National Center for Education Statistics. (2007). *Mapping 2005 State Proficiency Standards onto the NAEP scales* (Report No. NCES 2007 – 482). Washington, DC: Author.

National Commission on NAEP 12th Grade Assessment and Reporting. (2004). *12th grade student achievement in America*: *A new vision for NAEP*. *A report to the National Assessment Governing Board*. Washington, DC: Author.

Nichols, S., Glass, G., & Berliner, D. (2012). High-stakes testing and student achievement: Updated analyses with NAEP data. *Education Policy Analysis Archives*, *20*, 1 – 31.

O'Neil, H. F., Abedi, J., Miyoshi, J., & Mastergeorge, A. (2005). Monetary incentives for low-stakes tests. *Educational Assessment*, *10*(3), 185 – 208. doi:10.1207/s15326977ea1003_3

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

O'Reilly, T., & Sabatini, J. (2013). *Reading for understanding: How performance moderators and scenarios impact assessment design* (Research Report No. RR-13-31). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2013.tb02338.x

Osterlind, S. J., Robinson, R. D., & Nickens, N. M. (1997). Relationship between collegians' perceived knowledge and congeneric tested achievement in general education. *Journal of College Student Development*, *38*(3), 255 – 265.

Palomba, C. A., & Banta, T. W. (1999). *Assessment essentials: Planning, implementing, and improving assessment in higher education*. San Francisco, CA: Jossey-Bass.

Pedulla, J. J., Abrams, L. M., Madaus, G. F., Russell, M. K., Ramos, M. A., & Miao, J. (2003). *Perceived effects of state-mandated testing programs on teaching and learning: Findings from a national survey of teachers*. Boston, MA: National Board on Educational Testing and Public Policy, Boston College.

Pekrun, R. (2006). The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. *Educational Psychology Review*, *18*(4), 315 – 341.

Pekrun, R., Goetz, T., Daniels, L. M., Stupnisky, R. H., & Perry, R. P. (2010). Boredom in achievement settings: Exploring control – value antecedents and performance outcomes of a neglected emotion. *Journal of Educational Psychology*, *102*(3), 531 – 549. doi:10.1037/a0019243

Peterson, E. R., & Irving, S. E. (2008). Secondary school students' conceptions of assessment and feedback. *Learning and Instruction*, *18*(3), 238 – 250. doi:10.1016/j.learninstruc.2007.05.001

Pintrich, P. R. (1989). The dynamic interplay of student motivation and cognition in the college classroom. *Advances in Motivation and Achievement*, *6*, 117 – 160.

Pintrich, P. R., & DeGroot, E. V. (1990). Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology*, *82*(1), 33 – 40. doi:10.1037//0022 – 0663.82.1.33

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

Pintrich, P. R., Smith, D. A. F., Garcia, T., & McKeachie, W. J. (1993). Reliability and predictive validity of the Motivated Strategies for Learning Questionnaire (MSLQ). *Educational and Psychological Measurement*, *53*, 801–813. doi:10.1177/0013164493053003024

Rios, J. A., Liu, O. L., & Bridgeman, B. (2014, April). *Identifying unmotivated examinees on student learning outcomes assessment: A comparison of two approaches*. Paper presented at the annual conference of the National Council on Measurement in Education, Philadelphia, PA.

Rothe, H. F. (1947). Distributions of test scores in industrial employees and applicants. *Journal of Applied Psychology*, *31*, 480–483. doi:10.1037/h0060979

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

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*(3), 213–232. doi:10.1111/j.1745-3984.1997.tb00516.x

Schnipke, D. L., & Scrams, D. J. (2002). Exploring issues of examinee behavior: Insights gained from response-time analyses. In C. N. Mills, M. T. Potenza, J. J. Fremer, & W. C. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments* (pp. 237–266). New York: Laurence Erlbaum.

Seegers, G., & Boekaerts, M. (1993). Task motivation and mathematics achievement in actual task situations. *Learning and Instruction*, *3*(2), 133–150. doi:10.1016/0959-4752(93)90012-O

Smith, L. F., & Smith, J. K. (2002). Relation of test-specific motivation and anxiety to test performance. *Psychological Reports*, *91*(3), 1011–1021.

Steedle, J. T. (2014). Motivation filtering on a multi-institution assessment of general college outcomes. *Applied Measurement in Education*, *27*(1), 58–76. doi:10.1080/08957347.2013.853072

Stone, J., & Friedman, S. (2002). A case study in the integration of assessment and general education: Lessons learned from a complex process. *Assessment and Evaluation in Higher Education*, *27*(2), 199–211. doi:10.1080/02602930220128760

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, Canada.

Sundre, D. L., & Kitsantas, A. L. (2004). An exploration of the psychology of the examinee: Can examine self-regulation and test-taking motivation predict consequential and non-consequential test performance? *Contemporary Educational Psychology*, *29*(1), 6–26. doi:10.1016/S0361-476X(02)00063-2

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

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

Swerdzewski, P. J., Harmes, J. C., & Finney, S. J. (2009). Skipping the test: Using empirical evidence to inform policy related to students who avoid taking low-stakes assessments in college. *Journal of General Education*, *58*, 167–195.

Swerdzewski, P. J., Harmes, J. C., & Finney, S. J. (2011). Two approaches for identifying low-motivated students in a low-stakes assessment context. *Applied Measurement in Education*, *24*, 162–188. doi:10.1080/08957347.2011.555217

Terry, N., Mills, L., & Sollosy, M. (2008). Student grade motivation as a determinant of performance on the business Major Field ETS exam. *Journal of College Teaching and Learning*, *5*(6), 20–25.

Thelk, A. D., Sundre, D. L., Horst, S. J., & Finney, S. J. (2009). Motivation matters: Using the Student Opinion Scale to make valid inferences about student performance. *Journal of General Education*, *58*(3), 129–151.

Toch, T. (2009, October 11). Five myths about merit pay for teachers. *Washington Post*. Retrieved from http://www.washingtonpost.com/wp-dyn/content/article/2009/10/09/AR2009100902571.html

Tremblay, K., Lalancette, D., & Roseveare, D. (2012) *Assessment of higher education learning outcomes feasibility study report: Vol. 1. Design and implementation*. Paris: OECD.

U.S. Department of Education. (2006). *A test of leadership: Charting the future of American higher education*. Washington, DC: Author.

Van Tilburg, W. A. P., & Igou, E. R. (2011). On boredom: Lack of challenge and meaning as distinct boredom experiences. *Motivation and Emotion*, *36*, 181–194.

Voluntary System of Accountability. (2008). *Information on learning outcomes measures*. Retrieved from: http://www.voluntarysystem.org/docs/cp/LearningOutcomesInfo.pdf

Wall-Smith, S. B. (2006). Assessment measures. *Assessment Update: Progress, Trends, and Practices in Higher Education*, *18*, 14–15.

Wise, S. L. (2006a). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement*, *43*(1), 19–38. doi:10.1111/j.1745-3984.2006.00002.x

Wise, S. L. (2006b). An investigation of the differential effort received by items on a low-stakes computer-based test. *Applied Measurement in Education*, *19*(2), 95–114. doi:10.1207/s15324818ame1902_2

Wise, S. L., Bhola, D. S., & Yang, S. T. (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. (2003, June). *Low examinee effort in low-stakes assessment*: *Problems and potential solutions*. Paper presented at the annual meeting of the American Association of Higher Education Assessment Conference, Seattle, WA.

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment*, *10*(1), 1 – 17. doi:10.1207/s15326977ea1001_1

Wise, S. L., & DeMars, C. E. (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement*, *43*(1), 19 – 38. doi:10.1111/j.1745 – 3984.2006.00002.x

Wise, S. L., & DeMars, C. E. (2010). Examinee noneffort and the validity of program assessment results. *Educational Assessment*, *15*, 27 – 41. doi:10.1080/10627191003673216

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*(2), 163 – 183. doi:10.1207/s15324818ame1802_2

Wise, S. L., & Ma, L. (2012). *Setting response time thresholds for a CAT item pool*: *The normative threshold method*. Paper presented at the annual meeting of the National Council on Measurement in Education, Vancouver, Canada.

Wise, S., Pastor, D. A., & Kong, X. (2009). Correlates of rapid-guessing behavior in low stakes testing: Implications for test development and measurement practice. *Applied Measurement in Education*, *22*(2), 185 – 205.

Wise, V. L., Wise, S. L., & Bhola, D. S. (2006). The generalizability of motivation filtering in improving test score validity. *Educational Assessment*, *11*(1), 65 – 83. doi:10.1207/s15326977ea1101_3

Wolf, L. F., & Smith, J. K. (1995). The consequence of consequence: Motivation, anxiety, and test performance. *Applied Measurement in Education*, *8*(3), 227 – 242.

Wolf, L. F., Smith, J. K., & Birnbaum, M. E. (1995). Consequence of performance, test, motivation, and mentally taxing items. *Applied Measurement in Education*, *8*(4), 341 – 351. doi:10.1207/s15324818ame0804_4

Wolf, L. F., Smith, J. K., & DiPaolo, T. (1996). *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.

Yeo, G., & Neal, A. (2008). Subjective cognitive effort: A model of states, traits, and time. *Journal of Applied Psychology*, *93*(3), 617 – 631.

Zerpa, C., Hachey, K., van Barneveld, C., & Simon, M. (2011). Modeling student motivation and students' ability estimates from a large-scale assessment of mathematics. *SAGE Open*, *1*(2), 1 – 9. doi:10.1177/2158244011421803

Ziegler, M., MacCann, C., & Roberts, R. (Eds.) (2011). *New perspectives on faking in personality assessment*. Oxford, England: Oxford University Press.

Zilberberg, A., Anderson, R. D., Finney, S. J., & Marsh, K. R. (2013). American college students' attitudes toward institutional accountability testing: Developing measures. *Educational Assessment*, *18*(3), 208 – 234. doi:10.1080/10627197.2013.817153

Zilberberg, A., Brown, A. R., Harmes, J. C., & Anderson, R. D. (2009). How can we increase student motivation during low-stakes testing? Understanding the student perspective. In D. M. McInerney, G. T. L. Brown, & G. A. D. Liem (Eds.), *Research on sociocultural influences on motivation and learning: Vol. 9. Student perspectives on assessment: What students can tell us about improving school outcomes?* (pp. 255 – 277). Greenwich, CT: Information Age.

Suggested citation:

Finn, B. (2015). *Measuring motivation in low-stakes assessments* (Research Report No. RR-15-19). Princeton, NJ: Educational Testing Service. doi:10.1002/ets2.12067

Action Editor: John Sabatini

Reviewers: Tenaha O'Reilly and Ou Liu