Examinee Motivation in Low-Stakes Testing: Two Approaches to Identifying Data from Low-Motivated Students in an Applied Assessment Context

Peter Swerdzewski

Sara J. Finney

J. Christine Harmes

James Madison University

Abstract

Many universities rely on data gathered from tests that are low stakes for examinees but high stakes for the various programs being assessed. Given the lack of consequences associated with many collegiate assessments, the construct-irrelevant variance introduced by unmotivated students is a major potential threat to the validity of the inferences that institutions can make from their assessments. Two approaches to evaluating examinee motivation are discussed in this paper: one that employs a global paper-and-pencil self-report measure of students' motivation on all tests completed during the course of a testing session, and another computer-based method that non-intrusively measures the amount of time students spend on each item in a test. This study provides evidence that the two motivation filtering methods provide similar aggregate test scores; however, more data was removed using the global paper-and-pencil self-report technique.

Examinee Motivation in Low-Stakes Testing: Two Approaches to Identifying Data from Low-Motivated Students in an Applied Assessment Context

Introduction

With the increased prevalence of tests that are low-stakes for examinees come new issues that require attention by the assessment and measurement community. One such issue is the role that motivation plays in low-stakes contexts. With tests such as the SAT, ACT, and GRE, student test-taking motivation has traditionally not been an issue; however, as tests are used in situations such as program assessment and instrument development, students are increasingly being asked to complete tests for which there is little if any consequence for them when they do *not* put forth effort on the tests. We assume students are trying on the tests that we ask them to complete, but are they? If they are not exerting effort, what does that mean for the test results we use to inform policy decisions? The increasing prevalence of low-stakes tests requires us to re-evaluate our assumptions about testing and consider the critical role that motivation plays in the inferences we make from tests that have few consequences for the examinees who take them (Sundre & Kitsantas, 2004).

The role of student motivation in low-stakes testing contexts is a question of validity: if students are unmotivated to exert effort on a given test, the scores from those students will not accurately estimate the students' actual level on the construct of interest. By including data from unmotivated students in analyses, practitioners and researchers are contaminating results with construct-irrelevant variance (Eklöf, 2006). To ensure the efficacy of low-stakes testing programs, it is imperative that those who work with these programs take steps to minimize the error introduced by unmotivated students, thereby increasing the degree to which one can make valid inferences about scores from low-stakes tests (Erwin & Wise, 2002).

Purpose of the Current Study

Practitioners have explored two categories of techniques to minimize the influence that scores from unmotivated students have on aggregated low-stakes test results. One method is to provide a treatment prior to a low-stakes testing session with the hope that the treatment will increase the effort students exert on the low-stakes tests. These treatments include having speakers discuss the importance of the tests with students (Sundre & Moore, 2002), providing material incentives (i.e., prizes) to students (Cole, 2007), and publicly recognizing students for good performance (Pedulla et al., 2003).

The current study focuses on the second of the two categories of techniques often used to influence the role that contaminated data from unmotivated students has on aggregated score analyses: the use of motivation filtering to "clean" data after it has been collected. The process of motivation filtering requires the collection of some indicator of students' motivation during the collection of data on the construct of interest. This collection of motivation-related data often takes two forms: a global self-report measure provided at the conclusion of a testing session (cf. Sundre & Moore, 2002) or a non-intrusive technology-based solution that measures the amount of time a student spends on a given test or a given item (cf. S. L. Wise & DeMars, 2005; V. L. Wise, Wise, & Bhola, 2006). Specifically, in this study we examined the interrelationships between, and impacts of, these two motivation filtering techniques in an operational testing environment.

Motivation Issues in Low-Stakes Testing

As the use of low-stakes tests for accountability purposes such as NCLB and college accreditation has grown, so too has grown the research on the role of examinee motivation in low-stakes testing programs. Examinee motivation during low-stakes tests is an issue even among elementary-aged students, and is increasingly an issue among middle and high school students who are asked to take tests for school accountability purposes (Pedulla et al., 2003). Older K-12 students, particularly high school seniors, exert low effort on no-stakes tests such as the National Assessment of Educational Progress (NEAP, National Commission on 12th Grade Assessment and Reporting, 2004). Colleges and universities also must face the role of examinee motivation on low-stakes tests used for accountability purposes (Cole, 2007).

Examinee motivation is clearly a wide-spread phenomenon, but what *impact* does low motivation have on testing programs? Examinee motivation has been found to be a predictor of test performance among college students in situations in which the tests had low or no stakes for the examinees, yet had important consequences for the programs being assessed by those tests (Sundre, 1999; Sundre & Kitsantas, 2004; S. L. Wise & DeMars, 2005). Furthermore, K-12 teachers report that testing programs that rely upon low-stakes tests are less sustainable and produce results that have less utility for elementary, middle, and high school educators because of low student motivation (Pedulla et al., 2003). No matter the context (college, elementary school, high school), data from unmotivated examinees can tamper with the psychometric properties of a scale and ultimately the validity of the inferences made from the scores. (Eklöf, 2006; S. L. Wise & DeMars, 2006).

Recognizing the devastating impact of examinee motivation on low-stakes testing programs, Wise and DeMars (2005) looked to the literature on expectancy value theory to create a framework to parsimoniously explain examinee motivation (for an overview of motivation theory, see Eccles & Wigfield, 2002). In their framework, Wise and DeMars (2005) proposed that an examinee’s motivation is a function of (a) the examinee’s perceived success on a given test, (b) how much effort the examinee believes the test will consume, (c) the examinee's perceived importance of the test, and (d) the examinee's affective and emotional reactions to the individual test items. This framework not only has explanatory power, but also practical utility in understanding the methods that have been created by testing practitioners to stem the role of poor examinee motivation in testing programs.

One such category of methods is the global self-report measure, in which an examinee is asked to take a battery of tests and then, at the conclusion of the tests, completes a self-report instrument that measures the motivation an examinee believes he or she exerted during the testing session (Sundre & Moore, 2002; Wolf, Smith, & Birnbaum, 1995). Wolf and Smith (1995) devised a one-factor self-report instrument to measure examinee motivation that could be used to filter out data from unmotivated students. Building upon this instrument, Sundre and Moore (2002) created a two-factor instrument named the Student Opinion Scale (SOS) that aligns closely with expectancy-value theory (Pintrich & Schunk, 1996) and the four-pronged framework espoused by Wise and DeMars (2005). This new instrument is scored as two correlated factors that sum to a total test-taking motivation score: (1) the *effort* that an examinee reports putting forth on an instrument and (2) the *importance* an examinee places on the test(s). Eklöf (2006) built upon the work of Wolf and Smith (1996) and Sundre and Moore (2002) to create a third instrument that, like the other two, has utility in identifying unmotivated students and removing those students from a dataset prior to analysis.

These self-report measures of examinee motivation are resource-conscious techniques (e.g., they demand minimal testing and scoring time and only require adding a few items to a test administration) that practitioners can use in an attempt to increase the validity of the inferences one can make from test results. However, these methods have significant drawbacks: they require that students (a) know their level of motivation, (b) can use a scale to accurately report that level of motivation, and (c) *will* accurately report their level of motivation in lieu of providing a self-reported level of motivation that the student knows does not necessarily approach the truth.

Computer-based testing (CBT) technology allows for another option that does not involve indirect self-reports of motivation: Response Time Effort (RTE, S. L. Wise & Kong, 2005). RTE is a non-intrusive direct measure of examinee motivation that relies upon the assumption that students who are unmotivated during a low-stakes test will rapidly respond to items without taking the necessary time to read each item stem and thoughtfully consider each item's response options. During the course of a series of tests administered via CBT, an examinee views items one-at-a-time, and the CBT interface records the amount of time the examinee spends on each item before he or she can view the next item. The response time for Item 1 is recorded (two seconds for one examinee, seven seconds for another, etc.), and the students' response times are compared to a predetermined threshold representing the minimum possible time necessary to read and respond to the item (if the threshold for Item 1 is three seconds, the first examinee in the example above is likely to not have exerted effort on the item, whereas it is likely the second examinee *did* exert effort on the item). An examinee's overall index of RTE is calculated as the proportion of items on the test in which the examinee's response time exceeds the items' thresholds. In the current study, an examinee is said to have exerted effort on a test if his or her RTE index for the test exceeded 0.90 for the test. Said another way, a student exerted effort on a test if he or she spent enough time on 90% of the items to have read those items.

Recognizing the critical role that examinee motivation plays in low-stakes testing programs, and the necessity to develop motivation-filtering techniques to identify and remove suspect data from a dataset prior to analysis, we investigated three research questions in the current study:

1. To what degree does a global self-report measure of examinee motivation (the SOS) agree with a global as well as a test-specific non-intrusive direct measure of examinee motivation (RTE) in identifying students as either motivated or unmotivated?
2. If an economical global self-report measure of examinee motivation (the SOS) is used to filter out data from unmotivated examinees rather than a precise test-by-test direct non-intrusive method (RTE), how much data will be purged from the dataset? What is the nature of this purged data?
3. When applied to data from an operational battery of tests, how do the two motivation filtering techniques (global self-report versus direct non-intrusive) differ with respect to aggregated test scores?

Methods

Participants & Procedure

Data collected during the spring semester of 2006 for the assessment program at a mid-sized mid-Atlantic university were used to understand the characteristics of the two motivation-filtering techniques. This same data and similar methods were used to evaluated the phenomenon of student test avoidance of low stakes tests in a related research endeavor (Swerdzewski, Finney, & Harmes, 2007). All students who had between 45 and 70 credit hours (mostly second-year students) were required to participate in a series of cognitive assessments of their knowledge. The results from these assessments were used by the university's administrators to improve the quality of the institution's general education program. Students were also asked to complete a number of affective and developmental scales (i.e., non-cognitive scales), and the data from these scales were used for program improvement by the university's student affairs professionals.

The "Assessment Day", during which this data was collected, took place on the second Tuesday of February 2006. All classes at the university were cancelled to allow for the assessments. Each student who was required to complete assessments was randomly assigned to either a morning or afternoon session. All sessions were three hours in length and were managed by trained proctors. Due to the assessment design at the university, many students in the current study took some of the tests when they entered the institution as freshmen (i.e., 18 months earlier), thus the February 2006 administration served as a posttest for a number of the students in the sample.

A total of 2,965 students with between 45 and 70 earned credit hours were assigned to complete required assessments in either a traditional classroom through a paper-and-pencil test administration or in a computer lab using custom CBT software. Because of the scope of the current study's research questions, only data from those students who were randomly assigned to complete the assessments via CBT are included in the current study (N = 326).

The CBT sessions consisted of five non-cognitive (affective/developmental) measures. These included measures of (1) academic motivation and one's sense of belonging, (2) beliefs about learning related to the collegiate environment, (3) social self-efficacy, (4) worry, and (5) diversity. The two criterion-referenced measures of academic knowledge administered to the sample include measures of (1) quantitative/scientific reasoning and (2) the fine arts/humanities.

The final instrument administered to examinees was the global self-report measure of examinee motivation (the SOS; see Table 1 for a chart of measures).

Table 1

Short Description of Each Test

Order*	Test Name	Type**	Description
1.	Attitudes Toward Learning (ATL)	NC	Measures students' knowledge and beliefs related to learning. Scale consists of seven separate scales.
2.	Learning Environment Questionnaire (LEQ)	NC	Measures students' academic motivations and beliefs about learning under various environmental conditions in college.
3.	The Fine Arts	Cog	Measures students' understanding of the fine arts, humanities, and literature.
4.	Scale of Perceived Social Self-Efficacy (PSSE)	NC	Measures students' perceived abilities to engage in the social interactional tasks necessary to function in a college environment.
5.	Student Worry Questionnaire (SWQ)	NC	Measures both the process and content of worry as it is experienced by the traditional college student population.
6.	Interacting with Others (MGUDS)	NC	Measures students’ appreciation for, and comfort with diversity.
7.	Understanding our World (UOW).	Cog	Measures students' quantitative and scientific reasoning skills.
8.	Student Opinion Survey (SOS)	Mot	Measures the effort a student reports exerting on a test (or tests) and the importance he or she places on the test(s).

Note; The order column refers to the order in which all students encountered the eight tests. Not all students were asked to take the UOW. NC = Non-cognitive test; Cog = Cognitive test; Mot = Global self-report measure of test-taking motivation. More detailed information about the scales is reported in Appendix A.

Data Cleaning and Management

Of the 326 participants who were randomly assigned to complete the required assessments via CBT, 17 were removed from the study due to incomplete student information system data (i.e., the data used to verify the student’s identity at the institution) and an additional six were removed due to incomplete or unusual test data (the CBT software prohibited missing data, so the data removed at this step was due to technical issues, not due to student-related issues). This resulted in a final dataset of 303 participants. Because of the specific testing plan used by the institution, not all students were asked to complete the UOW; therefore, the effective sample size for analyses concerning the UOW is 137. If a student was not asked to complete the UOW, this was taken into account in analyses used to answer the four research questions for this study.

Instrumentation

Students completed a total of seven or eight tests of varying length and content. Two of the tests, referred to in this study as the cognitive tests, evaluate the effectiveness of the fine arts and humanities component (the Fine Arts test) and the quantitative and scientific reasoning component (the UOW test) of the university's general education program. The remaining five tests, referred to in this study as the non-cognitive or developmental tests, assessed students in terms of their psychological and educational development. The remaining test was the SOS (Sundre & Moore, 2002), a *global* self-report measure of examinee motivation across *all* seven tests previously completed by examinees during the testing session. Information about the scales used in this analysis is reported in Appendix A and a short description of each test, as well as the order in which each was given, is in Table 1.

Identifying Motivated and Unmotivated Students

Students were identified as either motivated or unmotivated using both motivation-filtering techniques: the SOS, a global self-report measure of examinee motivation; and RTE, a direct non-intrusive measure of motivation that relies upon the assumption that students must spend a sufficient amount of time on a given item to answer the item.

Using the SOS (the global self-report method of motivation filtering), a student is identified as “unmotivated” across all tests. This idea of creating two groups of students (unmotivated and motivated) is similar to the RTE method: RTE data can be used to identify unmotivated students across *all* tests by creating a cut-point (i.e., RTE index exceeded 0.90 across all tests) or for each test (i.e., RTE index exceeded 0.90 for the specific test). Using the SOS data, a student was deemed unmotivated if he or she had an average score on the scale that signifies a disagreement with items such as “I engaged in good effort on these tests” and “I gave my best effort on these tests” (i.e., had an average score less than 3.0 on the instrument). Because the SOS is given after all other tests have been administered to examinees, the examinees are asked to respond to the SOS items in terms of *all* tests they completed during the three-hour session; therefore, per the SOS an examinee is either “motivated” or “unmotivated” across *all* tests. All data from “unmotivated” examinees across all tests is identified and removed from a dataset prior to analysis according to this method.

Using RTE (the non-intrusive direct measure of examinee motivation), a student is identified as “unmotivated” for a given test if he or she did not spend enough time on at least 90% of a test’s items to read those items. Because RTE is collected non-intrusively (students are not aware that the amount of time they spend on each item is being recorded), the RTE approach to motivation-filtering can be at the test- or item-level. For the current study, we chose to identify instances in which a student is unmotivated on a given test, not a given item. All data from a given *test* is removed per this method if the examinee’s RTE index is below 0.90.

For this study, each student therefore has up to eight dichotomous (“motivated” / “unmotivated”) measures of motivation: one self-report measure of motivation across all tests and seven test-specific measures of motivation (collected via RTE).

Results

Agreement Between the Two Motivation-Filtering Techniques

Because the SOS provides a single indicator of examinee motivation across all tests during an administration and RTE provides test-level indicators of examinee motivation, we shall explore agreement between these two methods using two approaches. First, we will look at global agreement between the two approaches, and, second, we will look at agreement between the global SOS motivation approach and test-level motivation scores as determined by RTE.

Agreement when RTE is at a global level. To determine the degree to which the two measures agree on a global level (i.e., across *all* tests administered during a given administration), we first converted test-level motivation scores determined by RTE to a global indicator of motivation for each student. Recognizing the importance of quality data, we decided that a student would be termed “unmotivated” via RTE methodology if he or she was unmotivated on at least one of the seven tests as measured by RTE. For example, if a student did not exert effort on the Fine Arts test (per RTE), he or she would be identified as an “unmotivated student” and all data from the examinee across the examinee’s seven tests would be filtered out of the dataset. Despite the severe nature of this approach, the decision to have a single “unmotivated” attempt at a test (per RTE) effectively “throw out” all the data for a given student across all tests seemed most analogous to the decisions students make when responding to a self-report measure (such as the SOS). The results from the analysis are in Table 2.

Table 2

Percent Agreement between SOS and RTE for Global Test Motivation

Agreement	Percent Agreement	*N*

No agreement between SOS and RTE	36.63%	111
Agreement between SOS and RTE	63.37%	192
Grand Total	100.00%	303

Note: 107 students were deemed “unmotivated” using the SOS method, whereas 144 students were deemed unmotivated using the global RTE method.

When RTE is treated similarly to SOS (i.e., test-level information is not considered; decisions are made on a global or administration-wide level only), the two methods agreed for 63.37% of the sample. In other words, the two methods, when used on a global level, similarly identified 192 students of 303 students as either motivated or unmotivated. The remaining 111 students were identified as motivated using one method and unmotivated using another method. Of these 111 students, 37 were identified as unmotivated by the RTE method but motivated by the SOS method, whereas 74 students were identified as unmotivated per the SOS method but motivated by the RTE method. In cases in which the two methods differentially classify a student, it is most often because the student *indicated* that he or she was unmotivated via the SOS, not because the non-intrusive direct measure of timing the student’s responses identified the student as unmotivated (even though we used the extreme cutoff of being unmotivated on only one test).

Agreement when RTE is on a test-level. As noted earlier, although self report measures such as the SOS are generally collected at the end of a battery of tests or at the conclusion of a testing session (Sundre & Moore, 2002), an advantage of RTE is that the CBT interface collects motivation-related information for every item (i.e., the amount of time an individual spends viewing a given item), which allows for an evaluation of motivation on a test-by-test level, rather than across all tests. When a student's motivation level as measured by RTE ("motivated" or "unmotivated") for a given test was compared to the student's reported level of motivation as collected via the SOS, the two methods agreed between 62.71% and 68.61% of the time (see Table 3). Interestingly, the percent agreement between the two methods is rather similar across the tests.

For those students for whom the global SOS does not indicate the same level of motivation as the test-level RTE method, in the majority of instances, approximately 80% of students reported they were unmotivated, whereas the non-intrusive RTE method found that they *were* motivated. The exception was the PSSE (Perceived Scale of Social Self-Efficacy), in which the split was nearly 50/50.

Table 3

Percent Agreement between the Global SOS Measure of Motivation and a Non-Intrusive Test-by-Test Measure of Motivation (RTE)

	Percent Agreement	*N*	*N* (%) unmotivated per RTE	*N* (%) unmotivated per SOS
ATL				
Agreement	67.33%	204	-	-
Disagreement*	32.67%	99	14 (14.14%)	85 (85.86%)
LEQ				
Agree	64.69%	196	-	-
Disagree	35.31%	107	24 (22.43%)	83 (77.57%)
Fine Arts				
Agree	67.66%	205	-	-
Disagree	32.34%	98	19 (19.39%)	79 (80.61%)
MGUDS				
Agree	67.66%	205	-	-
Disagree	32.34%	98	17 (17.34%)	81 (82.65%)
PSSE				

Agree	62.71%	190	-	-
Disagree	37.29%	113	52 (46.02%)	61 (53.98%)
SWQ				
Agree	66.67%	202	-	-
Disagree	33.33%	101	20 (19.80%)	81 (80.20%)
UOW*				
Agree	68.61%	94	-	-
Disagree	31.39%	43	9 (20.93%)	34 (79.07%)

Note: Only 137 students were asked to take the UOW

* *Disagreement* refers to the disagreement between the global SOS motivation filtering method and the test-specific RTE method in identifying examinees as either motivated or unmotivated. For example, 204 of the original 303 students who completed the ATL were identically classified using the to methods. Conversely, 99 examinees (32.67% of the original sample of 303) who took the ATL were classified differently using the two methods; 14 of the 99 (14.14%) were classified as unmotivated using the test-specific RTE method, and 85 of the 99 (85.86%) were classified as unmotivated using the global SOS method.

Amount of Data Purged by Using a Global Self-Report Filtering Technique Rather Than a Test-Specific Non-Intrusive Filtering Technique

As noted earlier, a drawback to the global self-report approach to motivation filtering is that a student is asked to consider his or her motivation across all tests during a given session when responding to the single set of ten self-report items. One would naturally posit that a student may be motivated on some tests and not on others, yet the global self-report method does not collect test-specific information about examinee motivation. The self-report method *is*, however, a far easier and more economical method of filtering, so in this research question we explore how much data is removed using the global SOS-filtering method that would *not* have been removed had a non-intrusive direct measure such as RTE been employed for *each* test.

Each student in the initial dataset (*N* = 303) was identified as either motivated or unmotivated per the SOS filtering method. Test data for all seven tests from those students flagged as unmotivated normally would be removed from the dataset using the SOS method. This data was analyzed to determine the number of instances in which a student who self-reported as being unmotivated across *all* tests actually exerted effort on *a given test* per the test-level RTE method (see Table 5).

Table 5

Amount of Test Data Removed From the Dataset Using the SOS Filtering Method That Would Not Have Been Removed Using the Test-Level RTE Method

Test	Initial *N*	*N* (%) Retained Via SOS	*N* (%) Retained Via RTE	*N* (% of initial sample of 303) Removed Per SOS but not per RTE
ATL	303	196	267	85 (28.05%)
LEQ	303	196	255	83 (27.39%)
Fine Arts	303	196		79 (26.07%)
MGUDS	303	196		81 (26.73%)
PSSE	303	196		61 (20.13%)
SWQ	303	196		81 (26.73%)
UOW	137	196		34 (24.82%)

Using the more economical SOS filtering method caused the removal of an additional 20.13% (PSSE) to 28.05% (ATL) of the initial number of test scores above those removed via the RTE method. These test scores would have been retained had the non-intrusive direct RTE method been employed. Interestingly, one might imagine that the global SOS method would apply more accurately to tests later in the testing session than earlier, as students complete the self-report instrument at the very end of the session and the experience of taking more recent test (e.g., the UOW and the SWQ) are more prescient than tests completed nearly three hours earlier (e.g., the ATL and the LEQ). In other words, we expected that tests administered later in the testing session would have greater agreement (i.e., motivated/motivate or unmotivated / unmotivated) between the two filtering methods. Our analyses do not support this scenario, as the removal of test data using the SOS versus the RTE is rather evenly spread across all tests. The frequency with which we observed each pattern of effort across tests further demonstrates that students' test-taking motivation (per the direct RTE method) is spread across the seven tests and is not concentrated on a single test (see Table 6).

Table 6

Pattern of Test-Taking Motivation Using the Direct Non-Intrusive RTE Method

N	ATL	LEQ	Fine Arts	MGUDS	PSSE	SWQ	UOW	Description
97	Yes	Yes	Yes	Yes	Yes	Yes		Motivated across all tests; not required to complete UOW
62	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Motivated across all tests; required to complete UOW
18	Yes	Yes	Yes	Yes	No	Yes		Not motivated on PSSE
18	Yes	Yes	Yes	Yes	No	Yes	Yes	Not motivated on PSSE
8	Yes	Yes	Yes	Yes	Yes	Yes	No	Not motivated on UOW
6	No	No	No	No	No	No		Not motivated on any test
5	Yes	Yes	No	Yes	Yes	Yes	Yes	Not motivated on the Fine Arts test
4	Yes	No	Yes	Yes	No	Yes	Yes	Not motivated on LEQ and PSSE
4	Yes	No	Yes	Yes	No	Yes		Not motivated on LEQ and PSSE
4	Yes	Yes	No	Yes	Yes	Yes	No	Not motivated on Fine Arts and UOW tests
4	No	No	No	No	No	No	No	Not motivated on any test
4	Yes	Yes	No	Yes	Yes	Yes		Not motivated on Fine Arts test
4	Yes	Yes	Yes	No	Yes	Yes		Not motivated on MGUDS

4	Yes	Yes	Yes	Yes	Yes	No	Not motivated on SWQ

Note: Only patterns of motivation observed for four or more students are reported in this table. A total of 59 different patterns were observed among the 303 students in the sample.

"Yes" = the student was motivated on the test per the RTE method; "No" = the student was not motivated on the test using the RTE method. Note that a blank cell under the UOW column indicates the students who contributed that the motivation pattern described in the row were not requested to take the UOW.

How do aggregated test scores differ when each of the two methods is used?

The full, unfiltered dataset of 303 participants was filtered using each method: the global SOS method and the test-level method using RTE, and resulting aggregated scores from the two samples were compared to one another and to the unfiltered dataset (see Table 4). When the original dataset (*N* = 303) of examinees was filtered using the global SOS method, a dataset of 196 students resulted. Again, due to the administration-wide (i.e., not test-specific) nature of the SOS filtering method, a student was entirely removed from the dataset if he or she indicated not putting forth an acceptable degree of effort on the entire set of exams. Conversely, when the original dataset (*N* = 303) of examinees was filtered using the test-specific RTE filtering method, a student's test data was only removed for a particular test if the RTE index indicated the student did not exert effort on the test. Therefore, using the RTE method a student may be included in the aggregated test scores for the ATL (and its associated subscales) but may not be included in the aggregated test scores for the Fine Arts test. The new test-level sample sizes that resulted from the RTE filtering method range from 101 (for the UOW, only 137 students were initially asked to complete the UOW) to 267 (the ATL). Cohen's *d*s were calculated for each comparison as a measure of the practical significance of each difference.

Comparison of the SOS and RTE Test-Level Filtering Methods. Although the two filtering methods yielded substantially different sample sizes, the aggregated test scores for the resulting filtered datasets were surprisingly similar. Only two scores differed by a Cohen's *d* larger than 0.10: The Enjoyment subscale of the LEQ (*d* = 0.12) and the Perceived Social Self-Efficacy score (*d* = 0.19).

The Enjoyment subscale of the LEQ includes items such as "I really enjoyed the courses I've taken" and "I'd recommend the courses that I've taken to others". Higher scores indicate higher levels of enjoyment. Although the difference on the Enjoyment subscale of the LEQ was only marginal, it is interesting to note that the aggregated Enjoyment score was *higher* for the self-report filtering method rather than the non-intrusive RTE filtering method. If one assumes that the RTE method is a more objective (thus more valid) filtering method (a reasonable but not as yet fully-validated assumption), then it can be said that filtering using RTE removes scores from students who were not motivated *and*, on average, happened to select higher response options for the Enjoyment items (despite not having actually read the items).

The second of two scores that differed by a Cohen's *d* of greater than 0.10, the Scale of Perceived Social Self-Efficacy (PSSE), elicits students' perceived abilities to engage in social tasks. Sample indicators include, "Put yourself in a new and different social situation" and "Go to a party or social function where you won't know anyone". Higher scores indicate higher levels of social confidence. The aggregated averaged PSSE score using the two filtering methods differed by *d* = 0.19, a respectable magnitude of practical significance. Filtering using the global self-report (SOS) method yields a *higher* social self-efficacy score than filtering using the test-level RTE method. In sum, the aggregated filtered scores were very similar across the two methods.

Table 4

Comparison of Filtered Scores Across Methods

	No Filtering			SOS Filtered			Test-Level (RTE) Filtered			SOS vs. RTE Filtering		No Filtering vs. SOS Filtering		No Filtering vs. RTE Filtering	
	Mean	SD	N	Mean	SD	N	Mean	SD	N	*d*	|*d*|	*d*	|*d*|	*d*	|*d*|
ATL Subscale Score:															
Mastery-Approach	5.80	0.95	303	5.80	0.87	196	5.83	0.93	267	-0.04	0.04	0.00	0.00	-0.04	0.04
Mastery-Avoidance	4.57	1.01	303	4.51	0.95	196	4.53	0.99	267	-0.03	0.03	0.06	0.06	0.04	0.04
Performance-Approach	5.36	1.38	303	5.40	1.31	196	5.37	1.36	267	0.02	0.02	-0.03	0.03	-0.01	0.01
Performance-Avoidance	4.74	1.42	303	4.68	1.36	196	4.72	1.41	267	-0.03	0.03	0.04	0.04	0.01	0.01
Work-Avoidance	2.92	1.24	303	2.77	1.19	196	2.82	1.19	267	-0.04	0.04	0.12	0.12	0.08	0.08
Theories of Intelligence	3.46	0.85	303	3.46	0.83	196	3.51	0.81	267	-0.07	0.07	0.01	0.01	-0.06	0.06
Metacognitive Awareness-Reg	3.66	0.46	303	3.68	0.48	196	3.67	0.43	267	0.03	0.03	-0.05	0.05	-0.02	0.02
LEQ Subscale Scores:															
Autonomy	5.23	0.87	303	5.28	0.85	196	5.20	0.81	255	0.09	0.09	-0.05	0.05	0.03	0.03
Competence	5.87	0.87	303	5.89	0.85	196	5.87	0.80	255	0.03	0.03	-0.02	0.02	0.00	0.00
Interest	2.61	0.60	303	2.59	0.57	196	2.56	0.51	255	0.06	0.06	0.04	0.04	0.09	0.09
Enjoyment	3.74	0.71	303	3.76	0.70	196	3.68	0.66	255	0.12	0.12	-0.03	0.03	0.08	0.08
Intrinsic Motivation to Know (IMTK)	5.33	1.13	303	5.33	1.08	196	5.29	1.11	255	0.03	0.03	0.01	0.01	0.04	0.04
Intrinsic Motivation Toward Accomplishment (IMTA)	4.74	1.41	303	4.75	1.38	196	4.67	1.35	255	0.05	0.05	0.00	0.00	0.05	0.05
Intrinsic Motivation to Experience Stimulation (IMES)	3.79	1.52	303	3.74	1.49	196	3.67	1.43	255	0.04	0.04	0.04	0.04	0.08	0.08
Extrinsic Motivation Identified (EMID)	6.02	0.89	303	6.02	0.84	196	6.05	0.79	255	-0.04	0.04	0.00	0.00	-0.04	0.04
Extrinsic Motivation Introjected (EMIN)	5.10	1.40	303	5.10	1.40	196	5.07	1.36	255	0.02	0.02	0.00	0.00	0.02	0.02

	No Filtering			SOS Filtered			Test-Level (RTE) Filtered			SOS vs. RTE Filtering		No Filtering vs. SOS Filtering		No Filtering vs. RTE Filtering	
	Mean	SD	N	Mean	SD	N	Mean	SD	N	*d*	|*d*|	*d*	|*d*|	*d*	|*d*|
Extrinsic Motivation															
External Regulation	5.50	1.17	303	5.45	1.21	196	5.49	1.13	255	-0.04	0.04	0.04	0.04	0.00	0.00
Amotivation	1.44	0.92	303	1.33	0.68	196	1.34	0.65	255	-0.01	0.01	0.13	0.13	0.13	0.13
Fine Arts % Correct	0.57	0.09	303	0.58	0.09	196	0.59	0.08	256	-0.09	0.09	-0.11	0.11	-0.20	0.20
MGUDS	3.81	0.51	303	3.76	0.47	196	3.78	0.45	260	-0.04	0.04	0.10	0.10	0.07	0.07
PSSE	1.63	0.29	303	1.63	0.29	196	1.57	0.27	205	0.19	0.19	0.02	0.02	0.21	0.21
SWQ Subscale Scores															
Worrisome Thinking	3.01	1.02	303	2.97	1.05	196	2.97	0.99	257	0.00	0.00	0.04	0.04	0.04	0.04
Significant Others' Well-Being Concern	2.94	1.09	303	2.90	1.09	196	2.92	1.09	257	-0.02	0.02	0.04	0.04	0.02	0.02
Social Adequacy Concern	2.96	0.94	303	2.95	0.93	196	2.93	0.92	257	0.02	0.02	0.00	0.00	0.02	0.02
Financial-Related Concern	2.63	1.05	303	2.57	1.01	196	2.55	1.00	257	0.01	0.01	0.06	0.06	0.07	0.07
Academic Concern	3.60	0.87	303	3.58	0.86	196	3.64	0.83	257	-0.07	0.07	0.03	0.03	-0.04	0.04
Generalized Anxiety Disorder Symptoms	3.11	0.94	303	3.10	0.95	196	3.09	0.89	257	0.01	0.01	0.02	0.02	0.03	0.03
UOW % Correct Score	0.58	0.15	303	0.64	0.13	196	0.64	0.11	101	-0.03	0.03	-0.37	0.37	-0.40	0.40
Average Cohen's *d*											0.05		0.05		0.07

Comparison of the Unfiltered Dataset and the SOS-Filtered Dataset. When aggregated average scores computed for the unfiltered dataset (*N* = 303) were compared to aggregated scores after motivation filtered using the SOS method, the differences for most subscales were minimal; only four scales were found to have a practically-significant differences exceeding a Cohen's *d* of 0.10. These scales were the ATL's Work Avoidance subscale (*d* = 0.12; the unfiltered data had the higher average score), the LEQ's Amovitation subscale (*d* = 0.13; the unfiltered data had the higher average score), the Fine Arts test (*d* = 0.11; the unfiltered data had the lower average score), and the UOW test (*d* = 0.37; the unfiltered data had the lower average score). Assuming that responses from unmotivated students to cognitive tests are lower (as they are likely closer to chance guessing), the SOS filtering technique appears to provide an advantage to practitioners and researchers because the method essentially removes misleading low-score data for cognitive instruments such as the Fine Arts and UOW tests. Interestingly, but not surprisingly, when the global SOS filtering technique was applied to the dataset the average aggregated Work Avoidance and Amotivation scores dropped (compared to the unfiltered dataset). This suggests that work avoidance and amotivated students also self-reported they had low test-taking motivation, causing them to "removed themselves" from the sample due to SOS filtering. This finding of lower amotivation and work avoidance suggests that these "unmotivated" students *may* be attending to scales, but, again, due to their overall ratings of test-taking motivation, they are removed from the sample.

Comparison of the Unfiltered Dataset and the RTE-Filtered Dataset. When aggregated average scores for each scale in the unfiltered dataset are compared to aggregated average score for each scale in the RTE-filtered dataset, four scales differ in their average score by more than a Cohen's *d* of 0.10. Three of these four differing scores were for the same scales that differed in the comparison of scored between the unfiltered dataset and the SOS-filtered dataset, suggesting again that the two motivation-filtering methods produce similar results: the LEQ’s Amotivation scale (*d* = 0.13 between the RTE-filtered and unfiltered datasets; *d* = 0.13 between the SOS-filtered and unfiltered dataset as well), the Fine Arts aggregated average test score (*d* = 0.20 between the RTE-filtered and unfiltered datasets; *d* = 0.11 between the SOS-filtered and unfiltered datasets), and the UOW aggregates average test score (*d* = 0.40 between the RTE-filtered and unfiltered datasets; *d* = 0.37 between the SOS-filtered and unfiltered datasets). Interestingly, there is a sizable difference in the aggregated average score for the PSSE depending on the type of filtering employed. When the SOS filtering method is employed, there is little difference between the PSSE scores (*d* = 0.02); however, when the RTE-based filtering method is employed, the difference between the unfiltered and filtered datasets on the PSSE is *d* = 0.21.

Discussion

In the current study we sought to explore two popular methods used to address motivation issues in low-stakes testing. The first method, the use of a global self-report instrument administered at the completion of a set of tests, was compared to the second method, the use of a direct non-intrusive technique in which data is collected on the amount of time an examinee spends on a given item. The methods were applied to an operational dataset to determine (a) the degree to which the two methods agree with one another, (b) the amount of data purged from the dataset using the more economical global self-report measure that would not have been removed using the test-specific direct non-intrusive method, and (c) the difference in aggregated test scores across the two methods.

In general, the two methods were surprisingly consistent in the degree to which they similarly identified students who were unmotivated or motivated during the low-stakes assessments. Interestingly, when the two methods differed from one another, the self-report method was actually more *liberal* in the removal of suspect data than the RTE method, suggesting that practitioners and researchers interested in the integrity of their datasets, but who are not concerned with decreased sample sizes, may wish to use the self-report method over the test-specific direct non-intrusive method.

Furthermore, when the two methods are applied to the same unfiltered dataset and the filtered scores are compared to one another, the differences between the scores were minimal or nonexistent. This finding provides evidence that practitioners and researchers interested only in aggregated means and standard deviations will find equal utility in using either method.

This study was specifically designed to address questions that practitioners and researchers face in low-stakes operational testing programs, thus there are numerous limitations to the current study. First, if stakes are attached to test results (e.g., grades, material rewards), one would expect that the self-report measure of motivation would be severely confounded with the degree to which students decide to tell the truth. In the current study, with no stakes attached to students' responses, there was no reason for a student to *not* tell the truth.

The use of cut scores to identify motivated and unmotivated students is an additional limitation for the current study. Cut scores are inherently objective, but nonetheless necessary for operational testing programs, and it is important to note that the findings from the current study could have been wildly different had the cut scores been set using different methods.

A final, yet critical limitation to the current study is that this is the first such study that we could find in our review of the literature. This study must be replicated across different samples and under different conditions if one is to make operational decisions about the filtering methods employed in actual testing programs.

Conclusion

Students who do not put forth effort on low-stakes tests introduce crippling construct-irrelevant variance into data. These unmotivated examinees tend to yield test scores that are not indicative of the examinees' true levels on the construct of interest. Practitioners and researchers must continually evolve new methods to increase test-taking motivation among examinees, but if this isn't possible or methods don't work, practitioners may want to remove data from unmotivated students after the data has been collected in order to make more valid inferences. This is an issue of validity: an issue that must be addressed if the testing community is to maintain the integrity of the assessments we administer to examinees in low-stakes contexts.

Appendix A

Descriptions of Scales Used in Analysis

Attitudes Toward Learning (ATL): Measures students' knowledge and beliefs related to learning. 65 items. Scale consists of seven separate scales:

Subscale Name	Sample Item	Scale Range
(1) Mastery-approach goal orientation	"I want to learn as much as possible this semester"	1 - 7 ("Not at all true of me" to "Very true of me")
(2) Mastery-avoidance goal orientation	"I'm afraid that I may not understand the content of my classes as thoroughly as I'd like"	1 - 7 ("Not at all true of me" to "Very true of me")
(3) Performance-approach goal orientation	"My goal this semester is to get better grades than most of the other students"	1 - 7 ("Not at all true of me" to "Very true of me")
(4) Performance-avoidance goal orientation	"I just want to avoid doing poorly compared to other students this semester"	1 - 7 ("Not at all true of me" to "Very true of me")
(5) Work-avoidance goal orientation	"I want to do as little work as possible this semester"	1 - 7 ("Not at all true of me" to "Very true of me")
(6) Dweck's Theories of Intelligence	"You have a certain amount of intelligence, and you can't really do much to change it"	1 - 6 ("Strongly Agree" to "Strongly Disagree")
(7) Metacognitive Awareness Inventory-Regulation of Cognition Subscale	"I find myself pausing regularly to check my comprehension"	1 - 5 ("Always False" to "Always True")

Learning Environment Questionnaire (LEQ): Measures students' academic motivations and beliefs about learning under various environmental conditions in college. 66 items.

Subscale Name	Sample Item	Scale Range
(1) Perceived autonomy	"I feel that my instructors provide me choices and options"	1 - 7 ("Almost never" to "Almost always")
(2) Perceived competence	"I feel confident in my ability to learn the material in my courses"	1 - 7 ("Not at all true of me" to "Very true of me")
(3) Interest in academics	"I think what I've learned in my courses is important"	1 - 7 ("Almost never" to "Almost always")
(4) Enjoyment of academics	"I really enjoy the courses I've taken"	1 - 7 ("Almost never" to "Almost always")

The next seven subscales compose an instrument known as the "Academic Motivation Scale", which measures students' reasons for attending college. All begin with the stem "Why are you going to college?"

(5) Intrinsic motivation to know (IMTK)	"Because I experience pleasure and satisfaction while learning new thing."	1 - 7 ("Does not correspond at all [to me]" to "Corresponds exactly [to me]")
(6) Intrinsic motivation toward accomplishment (IMTA)	"For the pleasure I experience while surpassing myself in studies"	1 - 7 ("Does not correspond at all [to me]" to "Corresponds exactly [to me]")
(7) Intrinsic motivation to experience stimulation (IMES)	"For the intense feelings I experience when I am communicating my own ideas to others"	1 - 7 ("Does not correspond at all [to me]" to "Corresponds exactly [to me]")
(8) Extrinsic motivation identified (EMID)	"Because eventually it will enable me to enter the job market in a field that I like"	1 - 7 ("Does not correspond at all [to me]" to "Corresponds exactly [to me]")
(9) Extrinsic motivation introjected (EMIN)	"To prove to myself that I am capable of completing my college degree"	1 - 7 ("Does not correspond at all [to me]" to "Corresponds exactly [to me]")
(10) Extrinsic motivation external regulation (EMER)	"In order to obtain a more prestigious job later on"	1 - 7 ("Does not correspond at all [to me]" to "Corresponds exactly [to me]")
(11) Amotivation (AM)	"I once had good reasons for going to college; however, now I wonder whether I should"	1 - 7 ("Does not correspond at all [to me]" to "Corresponds exactly [to me]")

Fine Arts: Measures students' understanding of the fine arts, humanities, and literature. 118 items.

Subscale Name	Sample Item	Scale Range
(The Fine Arts Test is a single scale and does not include subscales)	"In most ancient civilizations, most monumental architecture, ceremonies, and art primarily served to A. provide the focus of a complex economy depending on the skills of many artisans. B. enhance the buildings in which humans resided and worshipped. C. celebrate the power and place of god(s) in shaping human life and society. D. reveal the creative and innovative spirit of the human intellect in facing a frightening world."	All items are dichotomously scores as correct or incorrect. Total scores can range from 0 points to 108 points.

Scale of Perceived Social Self-Efficacy (PSSE): Measures students' perceived abilities to engage in the social interactional tasks necessary to function in a college environment.

Subscale Name	Sample Item	Scale Range

(The PSSE is a single scale and does not include subscales)	Students are asked to indicate using a Likert scale how much confidence they have as a student to engage in various activities, such as: "Start a conversation with someone you don’t know very well."	1 - 5 ("No confidence at all" to "Complete confidence")

Student Worry Questionnaire (SWQ): Measures both the process and content of worry as it is experienced by the traditional college student population. 35 items.

Subscale Name	Sample Item	Scale Range
(1) Worrisome thinking	"I worry a lot about many daily life events and situations."	1 - 5 ("Almost never characteristic of me" to "Almost always characteristic of me")
(2) Significant others' well being	"I worry that a close family member might become seriously ill or injured."	2 - 5 ("Almost never characteristic of me" to "Almost always characteristic of me")
(3) Social-adequacy concerns	"I worry about saying the right things when expressing my opinion in discussions with other people."	3 - 5 ("Almost never characteristic of me" to "Almost always characteristic of me")
(4) Financial-related concerns	"I worry about not having enough money for the basic necessities of life (for example, clothing, food, rent)."	4 - 5 ("Almost never characteristic of me" to "Almost always characteristic of me")
(5) Academic concerns	"I worry about keeping up with or handling my academic workload."	5 - 5 ("Almost never characteristic of me" to "Almost always characteristic of me")
(6) General anxiety symptoms	"I feel physically tired and exhausted when I worry about things."	6 - 5 ("Almost never characteristic of me" to "Almost always characteristic of me")

Interacting with Others (MGUDS): Measures students’ appreciation for, and comfort with diversity. 45 items.

Subscale Name	Sample Item	Scale Range
(1) Relativistic appreciation	"It is very important that a friend agrees with me on most issues."	1 - 6 ("Strongly disagree" to "Strongly agree")
(2) Diversity of contact	"I would like to join an organization that emphasizes getting to know people from different countries."	1 - 6 ("Strongly disagree" to "Strongly agree")
(3) Comfort with differences	"I am only at ease with people of my own race."	1 - 6 ("Strongly disagree" to "Strongly agree")

Understanding our World (UOW): Measures students' quantitative and scientific reasoning skills. 60 items.

Subscale Name	Sample Item	Scale Range

(Although one can score the UOW using subscales, this analysis only uses the total score)	"Which of the following is not an essential feature of experimental design? a. manipulation of the independent variable b. at least one control group or comparison c. use of modern technology d. measurement of the dependent variable"	All items are dichotomously scores as correct or incorrect. Total scores can range from 0 points to 42 points.
Student Opinion Survey (SOS): Measures the effort a student exerts on a test (or tests) and the importance he or she places on the test(s). 10 items.		
Subscale Name	Sample Item	Scale Range
(1) Effort	"I gave my best effort on this test."	1 - 5 ("Strongly disagree" to "Strongly agree")
(2) Importance	"This was an important test to me."	1 - 5 ("Strongly disagree" to "Strongly agree")

References

Cole, J. E. (2007). *Motivation to do well on low-stakes tests.* Unpublished Doctor of Philosophy dissertation, University of Missouri-Columbia.

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

Eklöf, H. (2006). Development and validation of scores from an instrument measuring student test-taking motivation. *Educational and Psychological Measurement, 66*(4), 643-656.

Erwin, T. D., & Wise, S. L. (2002). A scholar-practitioner model for assessment. In T. W. Banta (Ed.)*, Building a scholarship of assessment. the jossey-bass higher and adult education series* (pp. 67-81). San Francisco: Jossey-Bass.

National Commission on 12th Grade Assessment and Reporting. (2004). *12th grade student achievement in America: A new vision for NAEP*. Washington, D.C.: National Assessment Governing Board. Retrieved March 6, 2007, from http:// www.nagb.org/ release/ 12_gr_commission_rpt.pdf

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. Research report published by the National Board on Educational Testing and Policy: Chestnut Hill: MA.

Pintrich, P. R., & Schunk, D. H. (1996). *Motivation in education : Theory, research, and applications*. Englewood Cliffs, N.J: Merrill.

Sundre, D. L. (1999). *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, Quebec, Canada.

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*(1), 6-26.

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

Swerdzewski, P. J., Finney, S. J., & Harmes, J. C. (2007 April). *Skipping the test: Using evidence to inform policy related to those students who avoid taking low-stakes assessments in college*. Poster Presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL.

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

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.

Wise, S. L., & Kong, X. J. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education, 18*(2), 163-183.

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.

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.