Abstract

Research investigating methods to influence examinee motivation during low-stakes assessment of student learning outcomes has involved manipulating test session instructions. The impact of instructions is often evaluated using a popular self-report measure of test-taking motivation. However, the impact of these manipulations on the psychometric properties of the test-taking motivation measure has yet to be investigated, resulting in questions regarding the comparability of motivation scores across instruction conditions and the scoring of the measure. To address these questions, the factor structure and reliability of test-taking motivation scores were examined across instruction conditions during a low-stakes assessment session designed to address higher education accountability mandates. Incoming first-year college students were randomly assigned to one of three instruction conditions where personal consequences associated with test results were incrementally increased. Confirmatory factor analyses indicated a two-factor structure of test-taking motivation was supported across conditions. Moreover, reliability of motivation scores was adequate even in the condition with greatest personal consequence, which was reassuring given low reliability has been found in high-stakes contexts. Thus, the findings support the use of this self-report measure for the valuable research that informs motivation instruction interventions for low-stakes testing initiatives common in higher education assessment.

AUTHORS

Sara J. Finney, Ph.D.
James Madison University

Catherine E. Mathers
James Madison University

Aaron J. Myers
James Madison University

CORRESPONDENCE

Email
finneysj@jmu.edu

Investigating the Dimensionality of Examinee Motivation Across Instruction Conditions in Low-Stakes Testing Contexts

Institutional accountability mandates prompt assessment of student learning (U.S. Department of Education, 2006). Although designed to accurately assess learning, many "accountability tests" are low stakes for students, meaning there are no personal consequences associated with performance for the examinee. Nonetheless, these tests are high stakes for universities in that scores are used to inform the evaluation and modification of programs, comparisons across institutions, accreditation, and resource allocation. With the prevalence of tests that are low stakes for examinees come issues that require attention from the assessment community. One such issue is the role that examinee motivation plays in low-stakes assessment contexts and its measurement.

The Need to Report Examinee Motivation

Examinee motivation is inherently linked to the validity of assessment interpretations. The more motivated examinees are to perform well, the better the test scores reflect ability (Wise & DeMars, 2005). Effortless test performance due to low motivation complicates inferences from test scores. Thus, low-stakes testing contexts, in particular, may result in test scores that are difficult to interpret for accreditation, strategic planning, and accountability purposes.

Consequently, score interpretations should be made in accordance with the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014). The *Standards* state, "In evaluation or accountability settings, test results should be used in conjunction with information from other sources when the use of the additional information contributes to the validity of the overall interpretation"

(p. 213). There is a specific call for information regarding "the degree of motivation of the test takers" in nonconsequential testing conditions as part of Standard 13.9.

Measuring Examinee Motivation via the Student Opinion Scale (SOS)

Reporting and interpreting examinee motivation requires its measurement. One particularly popular self-report measure of examinee motivation is the 10-item Student Opinion Survey (SOS; Thelk, Sundre, Horst, & Finney, 2009). The SOS has been implemented as a measure of examinee motivation in at least 9 countries, 33 universities, and 30 published studies (Sessoms & Finney, 2015). It has been used to examine the relationship between motivation and test performance (e.g., Abdelfattah, 2010; Swerdzewski, Harmes, & Finney, 2009; Wise & DeMars, 2005; Zilberberg, Finney, Marsh, & Anderson, 2014), to identify and filter out test scores from examinees with low motivation (e.g., Rios, Liu, & Bridgeman, 2014; Steedle, 2014; Swerdzewski, Harmes, & Finney, 2011), to examine personality characteristics that correlate with test-taking motivation (e.g., Barry, Horst, Finney, Brown, & Kopp, 2010; Barry & Finney, 2016; DeMars, Bashkov, & Socha, 2013; Kopp, Zinn, Finney, & Jurich, 2011), and to evaluate methods for increasing test-taking motivation (e.g., Finney, Sundre, Swain, & Williams, 2016; Hawthorne, Bol, Pribesh, & Suh, 2015; Liu, Bridgeman, & Adler, 2012; Steedle, 2010; Waskiewicz, 2011). The appropriateness of the use of the SOS for the latter is the focus of the current study.

The SOS was developed using expectancy-value (EV) theory (Wigfield & Eccles, 2000; Wolf & Smith, 1995). To determine the level of expended effort on a task, an individual considers (a) how well they *expect* to perform, and (b) the perceived *value* the task provides. EV theory applied to the context of test taking assumes an examinee's expended effort on the test is a function of their expected test performance and perceived value of the test. Assessing task value is essential in low-stakes testing contexts: examinees completing a test with no personal consequences for performance will likely put forth less effort because doing well has no attainment, intrinsic, or utility value. Thus, the resulting test scores may not be accurate representations of student ability (Wise & DeMars, 2005). This indirect effect of perceived test value on test performance (via test-taking effort) has been empirically supported in low-stakes contexts (Cole, Bergin, & Whittaker, 2008; Zilberberg et al., 2014).

The SOS was designed to operationalize the expended effort and test value components of test-taking motivation. Effort is defined as the level of effort expended toward test completion (e.g., "I engaged in good effort throughout this test"). Test value is defined as how important doing well is to the examinee (e.g., "Doing well on this test was important to me"). Again, theoretically, perceived importance influences expended effort. That is, importance and effort are considered theoretically distinct constructs. Empirical study of the dimensionality of the SOS scores has supported a two-factor structure over a one-factor structure of motivation in low-stakes testing contexts (e.g., Thelk et al., 2009). Invariance of the two-factor structure was found across age groups, gender, test modality (Thelk et al., 2009), and time (Sessoms & Finney, 2015) in low-stakes testing contexts.

Considering previous research examining the factor structure of noncognitive measures suggests dimensionality can differ across testing contexts (e.g., Barry & Finney, 2009; De Leeuw, Mellenbergh, & Hox, 1996), it is curious there have been no empirical studies assessing if the factor structure of the SOS is affected as the stakes or consequences of the test change. A difference in factor structure could impact the scoring of the SOS and, more important, could suggest test-taking motivation is conceptualized differently in different testing contexts. This issue becomes particularly important given the use of the SOS to evaluate the impact of increasing test consequences via test instructions. As called for in a recent issue of *Research & Practice in Assessment*, "Research on instruments that examine test-taker motivation on low-stakes tests is growing, but more is needed to fill the existing gap in the literature regarding examinee reactions to tests and the test conditions that affect performance and motivation" (Hawthorne et al., 2015, p. 36). We addressed this call by investigating the potential change in the psychometric properties of the SOS as the personal relevance and consequences for examinees were increased across test instruction conditions, as detailed below.

Evaluating Motivation Instruction Interventions Using the SOS

Given the relationship between examinee motivation and test scores in low-stakes testing contexts, assessment practitioners have investigated ways to increase motivation. One obvious solution is to increase the stakes for examinees (e.g., test scores impact grades). There are considerable complexities associated with a high-stakes testing program, which include the need to guard against and monitor cheating, the need for larger item pools for re-testing after remediation, the influence of test anxiety on test scores, and resistance from faculty (Wise & DeMars, 2005). Another option is to provide monetary compensation for performance (e.g., O'Neil, Sugrue, & Baker, 1995), which necessitates immense financial resources. Moreover, monetary incentives have not been proven consistently effective in improving test performance (O'Neil, Abedi, Miyoshi, & Mastergeorge, 2005).

Effortless test performance due to low motivation complicates inferences from test scores. Thus, low-stakes testing contexts, in particular, may result in test scores that are difficult to interpret for accreditation, strategic planning, and accountability purposes.

Another option presently receiving attention is motivation instruction interventions (e.g., Finney et al., 2016; Hawthorne et al., 2015; Kornhauser, Minahan, Siedlecki, & Steedle, 2014; Liu et al., 2012; Liu, Rios, & Borden, 2015; Waskiewicz, 2011). Motivation instruction studies involve evaluating the impact of test session instructions on examinee motivation and test performance. Of note, the test remains low stakes for examinees in that scores do not inform grades, graduation, or other academic outcomes. Instead, the instructions manipulate the message conveyed to examinees with the goal of making the test more personally relevant (see Appendix for a representative set of instructions). This active area of research may uncover an approach to influence examinee motivation in low-stakes contexts that requires no financial or human resources.

The effectiveness of motivation instructions is often evaluated using the SOS. That is, researchers examine if SOS scores differ, on average, across instruction conditions, with the goal of identifying instructions that increase motivation while preserving the low-stakes nature of the test. There is a considerable implicit assumption to this approach—one assumes that different instructions will potentially result in different average levels of motivation, yet other properties of the scores, such as the factor structure or reliability, will not be impacted. Of note, mean differences provide no insight into the stability of the factor structure and, hence, the scoring of the SOS; however, nonambiguous interpretation of mean differences necessitates no difference in factor structure across conditions. Surprisingly, there has been no empirical study evaluating if the factor structure of the SOS remains consistent across instruction conditions.

Is it reasonable to believe the factor structure of the SOS may change as instructions increase the personal relevance of the test for students? In a high-stakes context, the level of test importance should be high for examinees *and* they should put forth a great deal of effort. It is reasonable to presume that in a high-stakes testing environment, effort and importance may become indistinguishable (i.e., motivation becomes unidimensional). If this is the case, importance and effort items are interchangeable; an item from either subscale provides the same information regarding motivation. In turn, computing two subscales would no longer be appropriate. The factor structure of the SOS scores has not been examined in high-stakes contexts. Instead, the two-factor structure found in low-stakes contexts is simply assumed to generalize to high-stakes contexts, as reflected in the computation of the two subscales of effort and importance in high-stakes contexts. Importantly, there is evidence that the reliability of SOS scores differs across high- and low-stakes settings (Thelk et al., 2009). In high-stakes contexts, the SOS was sensitive to a ceiling effect, which decreased score variability, and in turn dramatically decreased estimates of internal consistency reliability. Given these results in high-stakes contexts, the reliability and dimensionality of SOS scores may differ across instruction conditions in low-stakes contexts.

Furthermore, this possibility of differing psychometric properties across instruction conditions is coupled with the perplexing practice by some researchers of scoring the SOS as a total motivation score (e.g., Kornhauser et al., 2014; Liu et al., 2012; Liu et al., 2015; Steedle, 2010). It is unclear if the authors of these studies uncovered a unidimensional solution when implementing motivation instructions and, hence, adapted the scoring of the SOS to align with this new conceptualization (i.e., a total SOS motivation score). If instruction

condition did impact the factor structure, this would imply a strong effect of instructions—the conceptualization of motivation differs depending on the instructions the examinees receive. Obviously, a difference in factor structure across instruction conditions makes comparisons of average motivation level across conditions obsolete.

Purpose of the Study

Using an operational low-stakes institutional accountability testing program, we examined the effects of gradually increasing test consequences on the psychometric properties of a popular measure of examinee motivation. Specifically, our purpose was to assess if the dimensionality and reliability of SOS scores differed across testing sessions that employed three different motivation instructions. Using confirmatory factor analysis (CFA) and SOS data from the three instruction conditions, we assessed the fit of the two-factor structure previously supported in low-stakes testing contexts and a one-factor structure implied by the creation of one total motivation score.

Assessing task value is essential in low-stakes testing contexts: examinees completing a test with no personal consequences for performance will likely put forth less effort because doing well has no attainment, intrinsic, or utility value.

Methods

Participants & Procedures

All students at a mid-sized university in the mid-Atlantic United States are required to participate in a three-hour large-scale testing session twice during their academic careers, once as incoming first-year students and again when they have accumulated 45-70 credit hours. Given students complete the same exams at both time points, this data collection scheme affords the computation of value-added scores associated with general education coursework. During the testing session, all students complete a battery of cognitive and noncognitive measures tied to general education and student affairs program objectives. Testing rooms differ in the exact measures administered and the size of the room (25 to 130 seats). Students are randomly assigned to testing room and, therefore, test configuration to ensure the desired sample size for each test. Although some students complete the tests via computer, the vast majority complete the tests via paper and pencil. All students selected for our study completed the tests via pencil and paper. Proctors in each room distribute and collect materials, read instructions, and encourage students to give their best effort. Test scores have no impact on students' academic record or graduation but do provide data for institutional accountability purposes.

Using data collected from this operational low-stakes testing program offered the unique and convenient opportunity to evaluate the impact of instructions on the psychometric properties of the SOS in an authentic testing environment. More specifically, the data analyzed in the current study were collected from incoming first-year students engaged in this testing program. To investigate the effects of test instructions on the factor structure of the SOS, students were randomly assigned to one of three conditions that incrementally increased the "dose" of the personal relevance of the test via the test instructions (see Appendix). In the first condition, students were instructed that their scores would be used in aggregate form at the institutional level (Institutional Condition). The Institutional instructions are the standard instructions all students have received over the past two decades of accountability testing at the institution. In the second condition, students were told that their scores would be used at the institutional level and their personal score would be available for individual feedback (Feedback Condition). In the final condition, students were told that their scores would be used at the institutional level, their personal score would be available to them for feedback, *and* their personal score would be released to faculty (Personal Condition). We purposefully selected these instructions as they are realistic in low-stakes testing contexts. Previous study of the SOS was conducted only in the Institutional condition; thus, it was unclear if the SOS would function adequately if institutions employed instructions similar to the Feedback and Personal condition instructions.

The assigned test instructions were read aloud to students and projected on the screen in front of the room. Proctors can positively affect effort (Lau, Swerdzewski, Jones, Anderson, & Markle, 2009); thus, all proctors received standardized training regarding administering

test instructions. Proctors were trained to draw students' attention to the test instructions to ensure the experimental conditions were understood. Furthermore, instructions for all conditions were presented with colored text (black for Institutional, blue for Feedback, & red for Personal) to draw attention to the conditions. Moreover, proctor gender, ethnicity, and age were held constant across instruction conditions to minimize any potential effect on motivation.

The study utilized one test configuration that was standardized across the instruction conditions. This configuration contained an arduous measure of scientific reasoning, which was administered first in the testing session, immediately followed by the SOS. Thus, student responses to the SOS represented students' perceived importance and expended effort for the scientific reasoning test just completed.

Of the 3,976 incoming first-year students engaged in the testing program, 1,287 were randomly assigned to one of the three test instruction conditions. A small proportion of students did not answer all SOS items, thus the effective sample size was reduced to N = 1,245. Of these students, 61.37% were female and the average age was 18.44 years. Students could self-identify in more than one ethnicity category, which resulted in 88.92% identifying as White; 5.06% as Black; 4.90% as Hispanic; 5.62% as Asian; 1.85% as American Indian; 0.96% as Pacific Islander; and 1.85% did not specify an ethnicity. These sample demographics align with the university demographics. At the university, 60% of students are female; 77.78% identify as White; 4.43% as Black; 5.75% as Hispanic; 4.35% as Asian; 0.18% as American Indian; 0.13% as Pacific Islander; and 3.48% unspecified. Of the 1,245 students, 385 received Institutional instructions, 385 received Feedback instructions, and 475 received Personal instructions. More examinees received the Personal instructions than the Institutional and Feedback instructions because this was the first administration of the Personal instructions, whereas Institutional and Feedback instructions had been administered in previous years.

Measures

To evaluate the dimensionality of the SOS across the three testing conditions, examinees completed a cognitive test of scientific reasoning and then immediately indicated their motivation with respect to that scientific reasoning test.

Scientific Reasoning Test. Scientific reasoning was assessed using the Natural World Test, Version 9 (SR; Sundre & Thelk, 2010; Sundre, Thelk, & Wigtil, 2008), a 66-item cognitive test designed to measure students' scientific reasoning skills. This test has been in use since its creation in 1996. It was designed to assess the scientific reasoning student learning objectives upon which a 10-12 credit hour curriculum has been designed. Faculty who teach this curriculum wrote every test item. Thus, the learning objectives and curriculum have been aligned. This cognitively demanding test typically takes an hour to complete. This test was the first test completed in the testing session, followed immediately by the SOS.

The [Student Opinion Scale] SOS was designed to operationalize the expended effort and test value components of test-taking motivation.

SOS. The Student Opinion Scale (Thelk et al., 2009) is a 10-item, self-report measure of test-taking motivation consisting of five effort items and five importance items (see Table 1). The SOS instructions referred to the scientific reasoning test and the SOS was completed directly after the scientific reasoning test. The Effort subscale consists of five items that measure the degree to which examinees put forth effort on a given test (e.g., "I gave my best effort on this test"). The Importance subscale consists of five items that measure the degree to which examinees view a given test as important (e.g., "Doing well on this test was important to me"). Examinees responded to the items using a scale of 1 (Strongly Disagree) to 5 (Strongly Agree).

Results

Data Screening

Prior to formally testing the fit of the one-factor and two-factor models to the SOS data from the three instruction conditions, the item-level data were examined (see Table 1). Inter-item correlations foreshadowed the dimensionality. In general, correlations among effort items

We addressed this call by investigating the potential change in the psychometric properties of the SOS as the personal relevance and consequences for examinees were increased across test instruction conditions...

Table 1
Correlations and Descriptive Statistics for the SOS by Test Instruction Condition

Institutional Condition (n = 385)										
	Item									
Item	1	3	4	5	8	2	6	7	9	10
1. Doing well on this test was important to me. [I]	1									
3. I am not curious about how I did on this test relative to others. [I]*	.252	1								
4. I am not concerned about the score I receive on this test. [I]*	.469	.495	1							
5. This was an important test to me. [I]	.585	.260	.483	1						
8. I would like to know how well I did on this test. [I]	.466	.530	.484	.393	1					
2. I engaged in good effort throughout this test. [E]	.423	.205	.234	.280	.440	1				
6. I gave my best effort on this test. [E]	.451	.180	.221	.341	.408	.601	1			
7. While taking this test, I could have worked harder on it. [E]*	.262	.167	.229	.273	.229	.436	.586	1		
9. I did not give this test my full attention while completing it. [E]*	.384	.179	.257	.281	.422	.519	.606	.590	1	
10. While taking this test I was able to persist to completion of the task. [E]	.265	.066	.113	.197	.327	.490	.387	.310	.421	1
Mean	3.633	3.321	3.272	2.974	3.672	4.121	4.015	3.214	3.861	3.964
SD	0.896	1.155	1.073	0.927	1.032	0.759	0.855	1.103	0.941	0.823
Skew	-0.438	-0.407	-0.384	0.071	-0.631	-1.131	-0.924	-0.211	-0.955	-1.082
Kurtosis	0.321	-0.637	-0.480	-0.183	0.010	2.582	1.246	-0.800	0.823	2.196
Feedback Condition (n = 385)										
1. Doing well on this test was important to me. [I]	1									
3. I am not curious about how I did on this test relative to others. [I]*	.317	1								
4. I am not concerned about the score I receive on this test. [I]*	.356	.329	1							
5. This was an important test to me. [I]	.497	.243	.484	1						
8. I would like to know how well I did on this test. [I]	.390	.484	.332	.313	1					
2. I engaged in good effort throughout this test. [E]	.500	.263	.329	.372	.338	1				
6. I gave my best effort on this test. [E]	.486	.304	.229	.253	.350	.626	1			
7. While taking this test, I could have worked harder on it. [E]*	.386	.193	.252	.277	.196	.540	.576	1		
9. I did not give this test my full attention while completing it. [E]*	.373	.198	.257	.331	.264	.613	.564	.654	1	
10. While taking this test I was able to persist to completion of the task. [E]	.284	.231	.089	.126	.252	.448	.439	.355	.421	1
Mean	3.722	3.479	3.405	3.003	3.868	4.129	4.018	3.163	3.835	3.987
SD	0.781	1.025	0.958	0.880	0.855	0.736	0.865	1.116	1.002	0.818
Skew	-0.451	-0.544	-0.496	0.155	-0.818	-0.755	-0.855	-0.134	-0.813	-0.722
Kurtosis	0.465	-0.260	-0.051	0.022	1.109	0.975	0.885	-0.837	0.190	0.574
Personal Condition (n = 475)										
1. Doing well on this test was important to me. [I]	1									
3. I am not curious about how I did on this test relative to others. [I]*	.403	1								
4. I am not concerned about the score I receive on this test. [I]*	.491	.490	1							
5. This was an important test to me. [I]	.643	.315	.482	1						
8. I would like to know how well I did on this test. [I]	.436	.533	.533	.363	1					
2. I engaged in good effort throughout this test. [E]	.532	.292	.356	.405	.374	1				
6. I gave my best effort on this test. [E]	.509	.296	.348	.408	.397	.697	1			
7. While taking this test, I could have worked harder on it. [E]*	.347	.227	.202	.286	.327	.483	.586	1		
9. I did not give this test my full attention while completing it. [E]*	.386	.312	.308	.338	.403	.563	.636	.617	1	
10. While taking this test I was able to persist to completion of the task. [E]	.368	.225	.169	.276	.300	.462	.494	.369	.382	1
Mean	3.676	3.381	3.444	2.994	3.734	4.118	4.023	3.184	3.826	3.983
SD	0.887	1.063	1.051	0.953	0.965	0.771	0.837	1.103	0.982	0.831
Skew	-0.514	-0.372	-0.481	0.099	-0.707	-0.970	-0.723	-0.164	-0.886	-0.870
Kurtosis	0.224	-0.455	-0.319	-0.258	0.401	1.889	0.348	-0.729	0.602	1.098

Note. *Denotes items that are reversed prior to scoring. Respondents rate their agreement with the 10 items on a 5-point Likert scale from Strongly Disagree (1) to Strongly Agree (5) with higher scores indicating higher levels of reported effort and importance.

[I] Denotes items from importance subscale.

[E] Denotes items from effort subscale.

were stronger than correlations among effort and importance items. Likewise, correlations among importance items were generally stronger than correlations among importance and effort items. This pattern suggests better fit for a two-factor than a one-factor model across all three instruction conditions.

Moreover, data were screened to assess univariate and multivariate normality as this impacts the choice of estimation method when formally estimating the models. Across conditions, all items were univariately normal with skew values less than |1.2| and kurtosis values less than |2.2|. Mardia's multivariate kurtosis coefficients ranged from 150.01 to 160.05 across the three conditions. Given multivariate non-normality, we chose an estimation method that accounts for the multivariate kurtosis of the data (Finney & DiStefano, 2013). That is, the CFA models were estimated using maximum likelihood estimation and the Satorra-Bentler adjustment was used to adjust the fit indices and the standard errors (Satorra & Bentler, 1994).

The factor structure of the SOS scores has not been examined in high-stakes contexts. Instead, the two-factor structure found in low-stakes contexts is simply assumed to generalize to high-stakes contexts, as reflected in the computation of the two subscales of effort and importance in high-stakes contexts.

Model-Data Fit

Two models were fit to the data in each condition: a two-factor model that aligns with the development of the SOS and a one-factor model that aligns with the (questionable) use of a total score. The robust root mean square error of approximation (RMSEA), robust comparative fit index (CFI), and standardized root mean square residual (SRMR) were used to assess global model-data fit. When using the Satorra-Bentler correction for multivariate non-normality, the following cutoffs have been suggested as indicators of good model fit: robust RMSEA ≤ .05, robust CFI ≥ .95, and SRMR ≤ .07 (Yu & Muthén, 2002). However, because the cutoffs are based on one study and are considered overly sensitive (i.e., result in rejecting adequate models), suggested cutoffs should be used as guidelines rather than strict criteria (Marsh, Hau, & Wen, 2004). Moreover, global fit indices simply summarize the overall model-data fit, whereas correlation residuals indicate local misfit of a model (under- or overestimated relationships between items). Correlation residuals greater than |.15| were flagged for inspection.

Table 2

Model Fit Indices for One-Factor and Two-Factor Models of the Student Opinion Scale

Model	*df*	χ^2_{SB}	$RMSEA_{SB}$	CFI_{SB}	SRMR	Correlation Residuals > |.15|
Institutional						
One-Factor	35	309.97*	0.16	0.78	0.110	7
Two-Factor	34	142.23*	0.09	0.93	0.067	0
Feedback						
One-Factor	35	187.84*	0.12	0.89	0.089	4
Two-Factor	34	97.00*	0.07	0.96	0.057	1
Personal						
One-Factor	35	366.92*	0.16	0.84	0.091	6
Two-Factor	34	166.36*	0.09	0.95	0.054	0

Note. χ^2_{SB} = Satorra-Bentler chi-square; $RMSEA_{SB}$ = robust root mean square error of approximation; CFI_{SB} = robust comparative fit index; SRMR = standardized root mean square residual. Chi-square difference tests were not conducted to compare the ill-fitting one-factor model to the two-factor model. Nested model difference tests are useful when evaluating competing models that represent the data well; thus, a nested model comparison is unnecessary here when one of two models grossly misrepresents the data (Bandalos & Finney, 2010).
*$p < .001$.

Using confirmatory factor analysis (CFA) and SOS data from the three instruction conditions, we assessed the fit of the two-factor structure previously supported in low-stakes testing contexts and a one-factor structure implied by the creation of one total motivation score.

Global fit indices and the number of correlation residuals greater than |.15| are located in Table 2. Not surprisingly, the χ^2_{SB} values were significant for both models; χ^2 tests are influenced by sample size, thus slight model misfit will be statistically significant with large samples. All global fit indices for the one-factor model were unsatisfactory within each condition. The large number of correlation residuals (ranging from |.16| to |.30| across conditions) reiterates the lack of fit of the one-factor model. In short, the one-factor model does not represent the data well.

In short, effort and importance items were well represented by two correlated factors of effort and importance, not one over-arching motivation factor, when students were told their scores were available to them personally and when students were told their personal scores could be viewed by faculty.

The SRMR and most robust CFI values were satisfactory for the two-factor model across all conditions. The robust CFI within the Institutional condition was only slightly below the cutoff. RMSEA values did not meet the suggested cutoff; however, the correlation residuals indicated acceptable fit of the two-factor model, aligning with the use of the fit index cutoff as an imprecise guideline (Marsh et al., 2004). Only one residual was greater than |.15| in the Feedback condition. The correlation residual between items 3 and 8 was .18, indicating the relationship between these items was underestimated by the two-factor model. Although both items represent the Importance subscale, the relatively larger observed correlation between these items compared to the other importance items may be due to a wording effect. Given satisfactory fit of the two-factor model across conditions, the latent factor correlation and reliability estimates were examined to further investigate the effects of test instructions.

Factor Correlation and Reliability

The correlation between the Effort and Importance factors was .61, .68, and .69 for the Institutional, Feedback, and Personal conditions, respectively. Notice the correlation increased negligibly as test consequences increased.[1] Moreover, the highest factor correlation indicated the two factors were related but not redundant. Importantly, internal consistency reliability of the Effort subscale scores (α = .83, .84, .84) and Importance subscale scores (α = .79, .74, .81) were adequate across the Institutional, Feedback, and Personal conditions, thus supporting their use. The magnitude and similarity of the reliability estimates were expected given the values of the factor loadings across conditions (see Table 3). Notice the relationship between each item and the corresponding factor differ negligibly across instruction condition.

Discussion and Implications

Given previous test-taking motivation research, support for the two-factor structure of test importance and expended effort in the Institutional condition was not surprising. The SOS has consistently been shown to be comprised of two moderately correlated yet distinct factors when examinees are told test scores are used solely for institutional accountability purposes. Our results reinforce the idea that when test scores have no personal relevance to students, test-taking motivation is *not* unidimensional in structure, and should not be scored as one total score. We found identical results when we increased the personal relevance or consequence for students. In short, effort and importance items were well represented by two correlated factors of effort and importance, not one over-arching motivation factor, when students were told their scores were available to them personally and when students were told their personal scores could be viewed by faculty.

Moreover, for the Institutional condition, 37.21% of the variance was shared between the effort and importance factors. When additional consequences were added in the Feedback and Personal conditions, 45.69% and 47.19% of the variance was shared, respectively. These results suggest as consequences are increased, effort and importance become only slightly less distinct. Despite the slight convergence of the two factors, most of the variance associated with

1 Formal measurement invariance tests were conducted to assess not only the equivalence of the factor structure across conditions (the focus of the current study), but also the equivalence of the factor pattern coefficients (i.e., equal unstandardized factor loadings, which is typically referred to as metric invariance), the covariance between Effort and Importance factors, and the correlation between Effort and Importance factors. All models (configural invariance, metric invariance, factor covariance invariance, and factor correlation invariance) fit well in an absolute sense (i.e., adequate values of fit indices) and, each model did not fit worse than the baseline configural model. Hence, in addition to the SOS having the same two-factor structure across motivation instruction conditions, the SOS items also had equal saliency to the factors across conditions (i.e., metric invariance) and an equivalent relationship between Effort and Importance factors across conditions.

Table 3
Standardized Factor Pattern Coefficients Across Motivation Instruction Conditions

	Institutional		Feedback		Personal	
Item	Effort	Importance	Effort	Importance	Effort	Importance
1. Doing well on this test was important to me.		.71		.72		.78
3. I am not curious about how I did on this test relative to others.		.54		.51		.59
4. I am not concerned about the score I receive on this test.		.69		.58		.69
5. This was an important test to me.		.66		.64		.70
8. I would like to know how well I did on this test.		.72		.58		.66
2. I engaged in good effort throughout this test.	.72		.80		.79	
6. I gave my best effort on this test.	.81		.77		.87	
7. While taking this test, I could have worked harder on it.	.69		.74		.68	
9. I did not give this test my full attention while completing it.	.77		.78		.75	
10. While taking this test I was able to persist to completion of the task.	.53		.54		.56	

Note. Given each item represents only one factor, the values above can be interpreted as correlations and squared to indicate the amount of variance explained in the item by the factor.

Fortunately, the current study supports the use of the SOS for the continued evaluation of motivation instruction interventions.

effort and importance was not shared, further supporting the distinction between perceived test importance and expended effort in low-stakes contexts.

Given these results, in low-stakes testing contexts the SOS may be perceived as having increased utility for two purposes: (a) reporting and interpreting examinee motivation to align with the *Standards for Educational and Psychological Testing* (2014); and (b) researching the effectiveness of motivation instructions. Regarding the first purpose, as noted earlier, when gathering data for accountability purposes in low-stakes testing contexts, assessment practitioners should collect and interpret examinee motivation information to inform inferences from test scores. We realize that the instructions communicated to university students in these low-stakes contexts differ across institutions, and those differences are tied to the personal relevance of the scores for students. We have provided evidence that the SOS importance and effort scores are appropriate to report and interpret in low-stakes contexts that differ in the message conveyed to students.

Regarding the second purpose, additional study of the effectiveness of motivation instructions is needed given previous research employs small samples (e.g., Hawthorne et al., 2015; Kornhauser et al., 2014), relies on volunteers who may not represent the university population (e.g., Liu et al., 2012; Liu et al., 2015), utilizes institutions with a fairly homogenous demographic composition (e.g., Finney et al., 2016), and confounds instruction

interventions with financial incentives (e.g., Liu et al., 2012; Liu et al., 2015). Fortunately, the current study supports the use of the SOS for the continued evaluation of motivation instruction interventions.

Although obvious, we feel it is important to reiterate that the SOS is a self-report measure. Assessment practitioners must rely on examinees providing responses to the SOS that represent true levels of perceived test importance and expended effort. Measures of effort such as response time effort (RTE) do not rely on accurate self-reporting but rather actual behavior as indexed by time (Wise & Kong, 2005). If self-report measures are necessary given lack of access to computerized testing to gauge RTE (as was the case in the current study), there is evidence of the alignment between RTE and self-report SOS scores (e.g., Rios et al., 2014; Swerdzewski et al., 2011). Nevertheless, we encourage gathering multiple measures of motivation when possible to provide additional insight into the effectiveness of motivation interventions and further validity evidence for self-report measures. Moreover, this study was based on a large, representative sample of first-year students, thus results should not be generalized to other student populations. We encourage researchers to conduct additional study of the properties of the SOS in motivation instruction conditions using other student populations.

In conclusion, the expectation for institutions to collect outcomes assessment data is not expected to decline, thus low-stakes testing will likely remain prevalent in higher education contexts. Consequently, the need to report and interpret examinee motivation will remain critical, as will the need to uncover a feasible intervention to increase motivation in these contexts. Fortunately, the SOS can be utilized for both purposes, allowing assessment practitioners to focus on possible solutions to the vexing problem of examinee motivation rather than its measurement.

Appendix

Institutional Condition Test Instructions

Please make sure you have correctly filled in your name and ID number on the scan form. After you have done this please write NW-9 in the top right corner of the scan form.

The Natural World test is designed to assess your quantitative and scientific reasoning. At this university, we define these as thinking processes for obtaining and evaluating knowledge of the natural world. This instrument was developed by faculty who teach in the university's award winning General Education program. The results are used to inform and improve our General Education program.

You will have 60 minutes to complete the 66 multiple-choice items on this test. You will have a piece of scrap paper to help you during this test. If you need more scrap paper, raise your hand. Make sure to read all test directions carefully, and answer the items to the best of your ability.

Thank you in advance for your effort and concentration on this important test. You may begin.

Feedback Condition Test Instructions

Please make sure you have correctly filled in your name and ID number on the scan form. After you have done this please write NW-9 in the top right corner of the scan form.

The Natural World test is designed to assess your quantitative and scientific reasoning. At this university, we define these as thinking processes for obtaining and evaluating knowledge of the natural world. This instrument was developed by faculty who teach in the university's award winning General Education program. The results are used to inform and improve our General Education program.

You will have 60 minutes to complete the 66 multiple-choice items on this test. You will have a piece of scrap paper to help you during this test. If you need more scrap paper, raise your hand. Make sure to read all test directions carefully, and answer the items to the best of your ability.

We are pleased to let you know that you will be able to find out how you scored on the quantitative and scientific reasoning measures and what your scores tell you about these reasoning skills. Later in the semester, you will receive an e-mail providing you with instructions for accessing your scores and the interpretive information.

When you become eligible for assessment again as a sophomore or junior, we will make every effort to make sure that you are assigned to take this instrument again so you can compare those scores with the ones you earn today. It is the hope of the faculty that we will see a lot of growth and development. We are pleased to offer this feedback to you.

Thank you in advance for your effort and concentration on this important test. You may begin.

Personal Condition Test Instructions

Please make sure you have correctly filled in your name and ID number on the scan form. After you have done this please write NW-9 in the top right corner of the scan form.

The Natural World test is designed to assess your quantitative and scientific reasoning. At this university, we define these as thinking processes for obtaining and evaluating knowledge of the natural world. This instrument was developed by faculty who teach in the university's award winning General Education program. The results are used to inform and improve our General Education program.

You will have 60 minutes to complete the 66 multiple-choice items on this test. You will have a piece of scrap paper to help you during this test. If you need more scrap paper, raise your hand. Make sure to read all test directions carefully, and answer the items to the best of your ability.

We are pleased to let you know that you will be able to find out how you scored on the quantitative and scientific reasoning measures and what your scores tell you about these reasoning skills. Later in the semester, you will receive an e-mail providing you with instructions for accessing your scores and the interpretive information.

When you become eligible for assessment again as a sophomore or junior, we will make every effort to make sure that you are assigned to take this instrument again so you can compare those scores with the ones you earn today. It is the hope of the faculty that we will see a lot of growth and development. We are pleased to offer this feedback to you.

Later in the semester, your personal test scores will be released to your faculty.

Thank you in advance for your effort and concentration on this important test. You may begin.

References

Abdelfattah, F. (2010). The relationship between motivation and achievement in low-stakes examinations. *Social Behavior and Personality*, 38, 159-168.

AERA, APA, & NCME (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Bandalos, D. L., & Finney, S. J. (2010). Exploratory and confirmatory factor analysis. In G.R. Hancock & R.O. Mueller, (Eds.), *Quantitative methods in the social and behavioral sciences: A guide for researchers and reviewers*. (pp. 125-155). Florence, NY: Routledge.

Barry, C. L., & Finney, S. J. (2009). Does it matter how data are collected? A comparison of testing conditions and the implications for validity. *Research & Practice in Assessment, 3*, 1–15.

Barry, C. L. ,& Finney, S. J. (2016). Modeling change in effort across a low-stakes testing session: A latent growth curve modeling approach. *Applied Measurement in Education, 29*, 46-64.

Barry, C. L., Horst, S. 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, 10*, 342-363.

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.

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.

De Leeuw, E. D., Mellenbergh, G. J., & Hox, J. J. (1996). The influence of data collection method on structural models: A comparison of a mail, a telephone, and a face-to-face survey. *Sociological Methods & Research, 24*, 443–472.

Finney, S. J., & DiStefano, C. (2013). Nonnormal and categorical data in structural equation modeling. In G. R. Hancock & R. O. Mueller, (Eds.), *Quantitative methods in education and the behavioral sciences: Issues, research, and teaching*. (pp. 439-492), Charlotte, NC: Information Age Publishing.

Finney, S. J., Sundre, D. L., Swain, M., & Williams, L. (2016). The validity of value-added estimates from low-stakes testing contexts: The impact of change in test-taking motivation and test consequences. *Educational Assessment, 21*, 60-87.

Hawthorne, K. A., Bol, L., Pribesh, S., & Suh, Y. (2015). Effects of motivational prompts on motivation, effort, and performance on a low-stakes standardized test. *Research & Practice in Assessment, 10*, 30-39.

Kopp, J. P., Zinn, T. E., Finney, S. J., & Jurich, D. P. (2011). The development and evaluation of the academic entitlement questionnaire. *Measurement and Evaluation in Counseling and Development, 44*, 105-129.

Kornhauser, Z. G. C., Minahan, J., Siedlecki, K. L., & Steedle, J. T. (2014, April). *A strategy for increasing student motivation on low-stakes assessments*. Paper presented at the annual meeting of the American Educational Research Association, Philadelphia, PA.

Lau, A. R., Swerdzewski, P. J., Jones, A. T., Anderson, R. D., & Markle, R. E. (2009). Proctors matter: Strategies for increasing examinee effort on general education program assessments. *The Journal of General Education, 58*, 196-217.

Liu, O. L., Bridgeman, B., & Adler, R. M. (2012). Measuring learning outcomes in higher education: Motivation matters. *Educational Researcher, 4*, 352-362.

Liu, O. L., Rios, J. A., & Borden, V. (2015). The effects of motivational instruction on college students' performance on low-stakes assessment. *Educational Assessment, 20*, 79-94.

Marsh, H. W., Hau, K., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling: A Multidisciplinary Journal, 11*, 320-341.

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

O'Neil, H. F., Jr., Abedi, J., Miyoshi, J., & Mastergeorge, A. (2005). Monetary incentives for low-stakes tests. *Educational Assessment, 10*, 185-208.

Rios, J. A., Liu, O. L., & Bridgeman, B. (2014). Identifying low-effort examinees on student learning outcomes assessment: A comparison of two approaches. *New Directions for Institutional Research, 161*, 69-82.

Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye & C. Clogg (Eds.), Latent variables analysis: *Applications for developmental research* (pp. 399-419). Thousand Oaks, CA: Sage.

Sessoms, J. C., & Finney, S. J. (2015). Measuring and modeling change in examinee effort on low-stakes tests across testing occasions. *International Journal of Testing, 15*, 356-388.

Steedle, J. T. (2010, April). *Incentives, motivation, and performance on a low-stakes test of college learning*. Paper presented at the annual meeting of the American Educational Research Association, Denver, CO.

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

Sundre, D. L., & Thelk, A. D. (2010). Advancing assessment of quantitative and scientific reasoning. *Numeracy, 3*, 2.

Sundre, D. L., Thelk, A. D, & Wigtil, C. (2008). *The Natural World Test, Version 9: A measure of quantitative and scientific reasoning, test manual.*

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. *The 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.

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. *The Journal of General Education, 58*, 129-151.

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

Waskiewicz, R. A. (2011). Pharmacy students' test-taking motivation-effort on a low-stakes standardized test. *American Journal of Pharmaceutical Education, 75*(3), 1-8.

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

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

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

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

Yu, C., & Muthén, B. (2002, April). *Evaluation of model fit indices for latent variable models with categorical and continuous outcomes*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Zilberberg, A., Finney, S. J., Marsh, K. R., & Anderson, R. D. (2014). The role of students' attitudes and test-taking motivation on the validity of college institutional accountability tests: A path analytic model. *International Journal of Testing, 14*, 360-384.