

Running Head: STUDENT OPINION SCALE

Examining Inferences about Test-Taking Motivation:
The Student Opinion Scale (SOS)

Amy D. Thelk, Donna L. Sundre, S. Jeanne Horst, and Sara J. Finney

James Madison University

Accepted for publication in the *Journal of General Education*.

For more information, please contact Donna L. Sundre at sundredl@jmu.edu or call: (540)568-3483. To review and download additional information from the Center for Assessment and Research Studies, go to: <http://www.jmu.edu/assessment> -- Look under the Assessment Resources section.

Keywords: Motivation, Assessment, Validity

Abstract

Institutions of higher education are encountering new and increasingly demanding expectations to address accountability and transparency of student learning and development efforts. Changes in federal administration have not dampened interest in the assessment and reporting of student learning (Jaschik, 2009), yet the centrality of general education to every educational mission will render it a problematic assessment domain. When assessment programs are implemented, how will we know that our students are engaged in completing the assigned tasks? The Student Opinion Scale (SOS; Sundre & Moore, 2002) was developed to assess examinee motivation during testing conditions. Based on the expectancy-value model of achievement motivation (Eccles, et al., 1983; Pintrich, 1989; Pintrich & DeGroot, 1990), the two subscores of the SOS are Effort and Importance. The 10-item questionnaire is administered at the close of testing and asks students to report the level of effort they invested and their perceived importance of the tasks. This article provides a thorough discussion and evaluation of validity evidence stemming from over 10 years of instrument use. We believe our review provides support for widespread and confident use of the instrument. Hence, SOS users should be able to draw more accurate inferences about examinee motivation using students' scores obtained via the SOS, instead of simply relying solely on test scores or anecdotal evidence. Moreover, faculty and staff will be able to address the difficult student engagement question with compelling evidence.

Motivation Matters: Using the Student Opinion Scale (SOS) to make Valid Inferences
about Student Performance

When college personnel need information about student achievement and growth, cognitive and/or noncognitive tests are commonly utilized. The ensuing analysis of test score results often leads to scrutiny of score interpretations. It is quite reasonable to speculate that in testing situations for which no personal consequences exist, low examinee motivation may result in test performances that underestimate true student ability. Thus, it becomes quite easy for stakeholders, generally relying on anecdotal evidence, to conclude that below-expected performance levels are due to lack of motivation, whether that be the true state of affairs or not. It is common for faculty members and committees when confronted with disappointing results to attribute low performance to a lack of motivation. This may or may not reflect reality.

Said another way, we gather assessment results to help us make sound decisions and interpretations about our academic programs. Every accrediting body, as well as our own desire for strong stewardship of our general education, academic major, and student affairs programs, advocates for assessment of student growth and development and use of this data. Without a method in place to actually measure and evaluate examinee test-taking motivation, the true meaning of resulting test scores remains dubious: Are the scores reflective of the true student achievement levels or simply an indication of how the students perform when not trying their hardest? In either case, the lack of information regarding examinee motivation thwarts attempts to adequately understand levels of student achievement. Overestimating examinee motivation, particularly when performance has no direct consequence for the examinee, represents a major threat to the

validity of test score interpretations. In other words, do the students' scores reflect what they know, or simply the level at which they were willing to perform at the time of testing?

The *Student Opinion Scale* (SOS; Sundre, 1997; Sundre & Moore, 2002) is a self-report tool that has been used in non-consequential testing contexts and provides a readily available and efficient means for estimating test-taker motivation. Over a decade of use in research and practice has garnered empirical support for the use of the SOS. In this article, we provide an organized evaluation of the properties of the SOS so faculty and staff can make informed decisions regarding its potential for their assessment and research related work. The subsequent articles in this special issue of the *Journal of General Education* provide additional perspectives and strategies for addressing complex issues associated with general education assessment. These challenging issues are all associated with validity.

Benson's Strong Program of Construct Validation

Construct validity refers to the extent to which instrumentation actually measures a theorized construct or trait. In order to feel confident that we are drawing accurate inferences about student motivation (the trait) from student responses to the SOS (the instrument), it is important that we build a body of validity evidence. Benson (1998) outlined a "strong program of construct validation" that included three stages: *substantive*, *internal* or *structural*, and *external*. Each stage builds upon the others and contributes to the strength of the overall inquiry into the quality of the measure. Information gained from an earlier stage may serve to inform a later stage, or information gained in a later stage may prompt a researcher to revisit an earlier stage of the validity

process. Thus, a strong program of construct validation entails an iterative process that involves the collection of evidence that is used to support or refine the theory or measure under study.

The first stage of Benson's (1998) strong program of construct validation, the substantive stage, involves theoretically and operationally defining the construct. That is, careful attention is paid to how we define a construct, from a theoretical standpoint. For example, the researcher must first articulate a theory of motivation. Once the theory has been carefully articulated, the researcher operationally defines the construct by writing items to cover the domain of all possible items. As individuals write items, care must be given to completely address the theory or construct (i.e., construct representation) and only items that are relevant to the construct should be included. Once a set of items has been written to operationally define the theoretically-defined construct, they may be evaluated through item analysis, reviewed through collaboration with other content experts, or given to focus groups of participants who respond to the items. This process of writing items to represent the trait being measured is not foreign to faculty; they engage in the process when writing tests to assess knowledge and skill in a particular domain

The second stage, the internal or structural stage, involves examination of the structure of item responses, that is, asking whether or not the item responses covary in ways that are predicted by the theoretical basis for the scale (Benson, 1998). As described in detail below, the SOS is comprised of two theoretically derived subscales, *Importance* and *Effort*; if we want to compute and ultimately make inferences from these subscale scores, it is important that there be empirical support for the computation of two theoretically distinct examinee motivation dimensions.

It is important to note that information gathered in this second, structural stage, involves only the item responses for a given instrument and not relationships with any other measures. Consequently, information gained from this stage addresses the instrument's structure and its theoretical alignment. However, information from this stage does not address whether the instrument represents the construct we purport to study. Hence, once support for the dimensionality of the scores is attained (i.e., we know how to score the measure), it is important to progress to the third stage.

The third stage, external validation, involves testing of theoretically based hypothesized relationships with other variables (Benson, 1998). A body of information gained from repeated administration of the measure in different settings, groups, and in conjunction with a variety of other measures is gathered. That is, expected relationships with other constructs are hypothesized and examined in multiple settings and with multiple groups of people. For example, in this stage we would hypothesize that scores from the SOS would relate to specific constructs (e.g., test performance) or discriminate between particular groups (e.g., students tested in different contexts) based on theory, and we would then test these hypotheses. Feedback from this or any stage may prompt the researcher to revisit the substantive stage or to reexamine the theory and operational definition of the construct as part of an ongoing process of construct validation. In sum, Benson's (1998) strong program of construct validation provides a useful framework for reporting and evaluating validity evidence that has been gathered for the SOS over multiple years and in multiple settings.

Methods

For this research, the authors reviewed dozens of studies in order to evaluate the SOS. Therefore, this methods section will be atypical, because we cannot provide specifics about each study's administrative conditions. Fortunately, at our institution, assessment testing, which is conducted twice a year, has changed little, allowing the data to be aggregated over administrations. Results emanating from other testing conditions will be briefly described to provide sufficient context for appropriate interpretation.

Administration

At James Madison University, students are required to engage in a half day of general education assessment testing at least twice during their undergraduate careers. A typical student will participate in their first Assessment Day as an entering first-year student. This Assessment Day is part of a required orientation process that takes place every fall just prior to the beginning of classes. The University also conducts an Assessment Day during the spring semester. On this day (typically in mid-February), classes are cancelled, and all students who have completed 45-70 collegiate credit hours are assigned to participate in Assessment Day. This group includes transfer students. We are interested in the performances of all of our students regardless of where those credits were earned. In fact, the impact of transfer, dual enrollment, International Baccalaureate and Advanced Placement credit hours on general education achievement are of growing institutional interest. Although only those students with 45-70 credit hours participate in this particular general education Assessment Day, many of our academic programs use this same day to assess their graduating seniors. Because all classes are cancelled, there are no time or room conflicts; all students and rooms are available. This design has

proven to be very cost effective, and we have been conducting these Assessment Days for over 20 years.

Students are required to participate in assessment testing; however, there are no personal consequences associated with their test performances. Therefore, this testing context is considered to be low-stakes or non-consequential for examinees. A typical assessment session is 2 to 3 hours long; students are randomly assigned (using student ID numbers) to computer labs and testing rooms, varying from 20 to 280 students per room. A set of tests is slated for each room and consists of both cognitive (e.g., quantitative and scientific reasoning) and non-cognitive (e.g. attitudinal) instruments. These instruments have been selected or, as in most cases, specifically crafted to assess the institution's stated general education student learning objectives

[\[http://www.jmu.edu/gened/gened_program.shtml\]](http://www.jmu.edu/gened/gened_program.shtml)]. Using this design, students are randomly assigned to different test batteries. In other words, no student takes all of the scheduled tests, but the institution will have large, representative, and random student data samples to analyze and interpret. Because students' IDs are used for room assignments, it is possible to retain the same test assignments over several years. This allows for both cross sectional analyses and true repeated measures of individual students. This is a very powerful design, and our faculty committees, as well as accreditors, find this information quite compelling.

The obtained scores are often incorporated into a profile of institutional results that have stakes for the university. Student performances are reported to regional accreditors, the state council, university administrators, and committees comprised of program faculty members. Therefore, the assessment results do have stakes for the

institution, and our faculty committees require more information to frame appropriate interpretations. In this setting, the SOS serves an important role. The SOS is the final assessment tool administered in every Assessment Day administration room, and these results are scrutinized by our faculty and researchers to learn more about examinee motivation.

SOS scores are reported in the aggregate and not as a means to make decisions about individual students. If the SOS works as intended, presentation of its subscale score distributions will greatly benefit interpretation of the general education test performances. For example, if low SOS scores are observed, interpretations of our general education test scores can be properly tempered with empirical evidence that students did not put forth their best effort and/or deemed the examinations as unimportant. Similarly, if SOS scores are high, we should be able to more confidently interpret the performances on our general education tests as a more accurate reflection of true student ability.

We have been administering the SOS as an integral part of our Assessment Day activities for over a decade. We have synthesized dozens of studies below in which the SOS was employed. Most of the studies referred to emanated from our Assessment Day testing context, a non-consequential testing condition. This is precisely the type of situation for which the SOS was developed.

Instrumentation

The Student Opinion Scale is based on an earlier unidimensional measure of examinee motivation, the Motivation Questionnaire, pioneered by Wolf and Smith (1993). Sundre (1997) consistently found that the original eight items were represented by two factors, and that these dimensions appeared to represent perception of importance

of the test (5 items), and amount of effort exerted on the test (3 items). Sundre (1999) was very interested in studying examinee motivation in low stakes testing conditions and thus reworded several items and added two effort items to further represent this important examinee motivation dimension, leading to the current 10-item SOS form.

Students respond to SOS items using a 1 to 5 scale (1=Strongly Disagree, 2=Disagree, 3=Neutral, 4=Agree and 5=Strongly Agree). Four items, two per subscale, are negatively worded and are therefore reverse-scored prior to summing the corresponding items to create the Effort and Importance subscale scores. A copy of the instrument and instructions for administration can be freely downloaded at <http://www.jmu.edu/assessment/resources/Overview.htm>. Table 1 presents the mapping of items to the theoretical dimensions of Effort and Importance.

Table 1
Test Blueprint for SOS

Subscale	Items
Importance Definition: How important doing well on the test is to the student (the consequence of the test for the student).	1. Doing well on these tests was important to me. 3. I am not curious about how I did on these tests relative to others. 4. I am not concerned about the scores I receive on these tests. 5. These were important tests to me. 8. I would like to know how well I did on these tests.
Effort Definition: The reported level of effort and persistence expended toward test completion.	2. I engaged in good effort throughout these tests. 6. I gave my best effort on these tests. 7. While taking these examinations, I could've worked harder on them. 9. I did not give these tests my full attention while completing them. 10. While taking these tests, I was able to persist to completion of the tasks.

Results

Substantive stage

The items on the SOS were crafted to directly incorporate the theoretical underpinnings associated with value, specifically the way Eccles and her colleagues conceived it in the development of the expectancy-value model of achievement motivation theory (Eccles et al., 1983; Pintrich, 1989; Pintrich & DeGroot, 1990). *Value* is defined as how important it was to the student to do well on the test. The Importance scale is intended to assess the perceived value of the requested tasks to the student. Effort also corresponds to the theoretical feature of value. Wolf, Smith and Birnbaum (1995) defined effort as the amount of mental taxation the examinee is willing to put forth in responding to test items or tasks. When an individual perceives an assignment as having high subjective task value, he or she is more likely to engage in, or value, the current task.

Internal/Structural Stage

Assessing Dimensionality

A primary task of the structural stage of a strong program of construct validation is examination of whether the dimensionality of the scores is what one would expect based on theory (Benson, 1998). Recall that the SOS was developed based on the value component of the expectancy-value theory, with items written to represent importance and items written to represent effort. Therefore, we would expect two distinct factors of Importance and Effort to underlie SOS scores. Subscale score construction should align with the dimensionality that represents the data. For example, if a one-factor structure emerged for the SOS, it would be inappropriate to compute two subscale scores and

imply they represent distinct dimensions of Effort and Importance. By doing so, one would be assuming the two subscores carry differential meaning, when they do not. Instead, a total score of test-taking motivation would be computed and interpreted. Likewise, if a two-factor structure emerged, one should avoid creating a total scale score; such a procedure would assume a single dimension and would mask any differential relationships that the effort and importance factors have with external criteria. Thus, one can see the critical importance of assessing factor structure to inform the practical task of computing subscale scores.

Initial examination of two-factor vs. one-factor structure. One-factor and two-factor models of test-taking motivation were initially assessed using four samples of data: two samples of data from incoming freshmen completing Assessment Day (Freshmen 01, 02) and two samples of data from sophomores completing assessment activities (Sophomores 01, 02). Although it is tempting to draw conclusions about the factor structure based solely on one group of students, there is the risk that the findings are specific to that particular group of students and do not generalize to other groups of students (i.e., “capitalization on chance”). By evaluating and re-evaluating the factor structure of responses from several independent groups of students from several populations, a researcher can be more confident that the findings are “stable” (i.e., similar) across the groups of students and, consequently, be more confident in generalizing the findings to other similar groups of students (MacCallum, Roznowski, & Necowitz, 1992).

The one-factor and two-factor models did not represent the Freshmen 01 SOS scores (see Table 2). Although the two-factor model represented the data significantly

better than the one-factor model, suggesting that Effort and Importance are distinct, the two-factor model did not represent the data well in an absolute sense. It appeared that the relationship between items 1 and 5 from the Importance subscale and the relationship between items 7 and 9 from the Effort subscale were not well-reproduced. We believe this to be due to the fact that the item wording for items 1 and 5 is similar. That is, the items appear to be somewhat redundant, and, therefore share variance after controlling for the Importance factor. Similarly, we believe that items 7 and 9 share variance after controlling for the Effort factor due to the fact that these are the only reverse-coded items representing this factor. This issue of items sharing variance that is unrelated to the substantive factor (i.e. Effort and Importance) is not uncommon and is discussed often in the measurement literature (see Podsakoff, MacKenzie, Lee, & Podsakoff, 2003, for a review of this issue).

We thought it important to examine if the initial results replicated across an independent sample from the same population (Freshmen 02). As can be seen in Table 2, the results were extremely similar for the Freshmen 2002 sample (Freshmen 02). That is, the one-factor model did not represent the data and fit significantly worse than the two-factor model; however, there was some model misfit associated with the two-factor model. Importantly, the misfit was again associated with items 1 and 5 as well as items 7 and 9, indicating that the model misfit was stable across the two independent samples.

Because of the similarity in the wording of items 1 and 5 (redundant items written to represent the Importance subscale) and items 7 and 9 (negatively worded items written to represent the Effort subscale), we decided to model a relationship between them. In other words, we allowed items 1 and 5 to correlate above and beyond what was explained

by the Importance factor and allowed items 7 and 9 to correlate above and beyond what was explained by the Effort factor. Specifically, once controlling for the latent constructs in the model (i.e., Importance and Effort factors), we would typically assume that the remaining variation in item responses is random and not correlated. However, in this case, we believed that modeling a correlation between the unexplained variance (i.e., error terms) for items 1 and 5 and items 7 and 9 would account for the model misfit.

We included the relationship between these two item pairs in the model and assessed this modified model using Freshmen 01 and 02 samples (Table 2, *two-factor b* model). This model represented the SOS scores significantly better than the previous two-factor model. We believe that the addition of these error covariances is warranted given that these terms represent what we believe to be minor and common issues due to redundancy and negative item wording.

For both samples, Effort and Importance were moderately positively correlated (Freshmen 01 $r = .47$, Freshmen 02 $r = .46$), indicating that Effort and Importance are distinct constructs within the low-stakes test-taking context. In addition, the standardized relationships between the items and their respective factor (i.e. factor loadings) were quite similar and of an acceptable magnitude across the two samples ranging from .51 to .87 for the Effort items and from .51 to .76 for the Importance items.

The next step of the analysis concerned whether the results were generalizable to a sophomore population. As sophomores have more experience with test-taking and most of these students had been through this particular low-stakes testing program one year earlier, it could be possible that the measure functions differently for sophomore students. As shown in Table 2, the results from two sophomore samples were identical to the two

freshmen samples. That is, the one-factor model did not represent the SOS scores and was significantly worse than the two-factor model. In addition, items 1 and 5 from the Importance subscale and items 7 and 9 from the Effort subscale shared variance after controlling for their respective factors. Thus, these item-wording issues generalized across samples. As with the freshmen samples, the modified two-factor model represented the SOS scores adequately and Effort and Importance were moderately positively correlated (Sophomore 01 $r = .50$, Sophomore 02 $r = .49$). Similar to the findings from the freshmen samples, the standardized relationships between the items and their respective factor (i.e., factor loadings) were quite similar and of an acceptable magnitude across the two sophomore samples ranging from .49 to .92 for the Effort items and from .50 to .78 for the Importance items.

Cross-validation of modified two-factor model. It was important to assess the fit of the modified two-factor model on independent samples that were not used to inform the modification to the model (Maccallum et al., 1992). The results in Table 3 indicate that the modified two-factor model adequately represented SOS scores from two independent samples of freshmen and sophomores. As found with the previous four samples, Effort and Importance were positively related but distinct (Freshmen 03 $r = .51$, Sophomore 03 $r = .41$) and factor-item relationships ranged from .51 to .90 for Effort and .51 and .80 for Importance. Given the stability of the two-dimensional structure of the SOS scores across six independent samples, it appears appropriate to score the SOS as two separate subscale scores of Importance and Effort.

Table 2

Fit of the One-factor and Two-factor Models of Test-Taking Motivation

Model	χ^2	df	CFI	SRMR
Fresh01 one-factor	2981.17	35	.68	.12
Fresh01 two-factor	1358.41	34	.86	.07
Fresh02 one-factor	3169.35	35	.68	.12
Fresh02 two-factor	1552.38	34	.84	.08
Fresh01 two-factor b	737.35	32	.92	.07
Fresh02 two-factor b	810.95	32	.92	.08
Soph01 one-factor	2927.87	35	.68	.13
Soph01 two-factor	1324.06	34	.86	.08
Soph02 one-factor	3621.83	35	.68	.13
Soph02 two-factor	1574.24	34	.86	.07
Soph01 two-factor b	594.45	32	.94	.08
Soph02 two-factor b	746.04	32	.94	.07

Note. Model two-factor b included error covariance terms between items 1 and 5, and items 7 and 9.

χ^2 = Maximum Likelihood χ^2 ; CFI = comparative fit index; SRMR = standardized root mean residual. Given the conflicting recommendations concerning fit index cutoff criteria (e.g., Marsh, Hau, & Grayson, 2005; Marsh, Hau, & Wen, 2004; Sharma, Mukherjee, Kumar, & Dillon, 2005) we followed a strategy suggested by Vandenberg and Lance (2000). Specifically, when examining the CFI, we used a value of .90 as a lower bound of good model-data fit, with values of .95 or above indicating a well-fitting model. When examining the SRMR, we used a value of .10 as a lower-bound of good model-data fit, with values of .08 or less indicating a well-fitting model.

$N_{Freshman\ 01} = 2573$; $N_{Freshman\ 02} = 2623$; $N_{Sophomore\ 01} = 2000$; $N_{Sophomore\ 02} = 2382$.

Table 3

Fit for the One-factor and Modified Two-factor Models using Independent Samples

Model	χ^2	df	CFI	SRMR
Fresh03 one-factor	3089.90	35	.70	.11
Fresh03 two-factor b	832.12	32	.92	.07

Sopn03 one-factor	3378.78	35	.64	.15
Soph03 two-factor b	637.06	32	.94	.08

Note. Two-factor model b included error covariance terms between items 1 and 5, and items 7 and 9.

$N_{Freshman\ 03} = 2674$; $N_{Sophomore\ 03} = 2035$.

Generalizability of factor structure across gender and testing medium. Given that faculty will often want to compare SOS scores across groups (e.g., computer-based test takers vs. paper-and-pencil test takers, male vs. female), and that faculty may want to aggregate SOS data composed of different groups, we felt it necessary to investigate if the SOS was functioning equivalently across these groups. For example, as we reported above, we found the two-factor structure fit the data from six independent samples of students. However, all of these students completed the SOS via paper and pencil. Given the ease and efficiency of computer-based data collection, we foresee an increase in collection of SOS scores via that medium. One cannot ignore the vast differences in interface between computer-based assessment and paper-and-pencil assessment. The functioning of the SOS could be very different across these two mediums and must be investigated. Additionally, in practice, responses from male and female students are often combined into one overall dataset in order to assess test-taking motivation. It is important to evaluate if the measure is being used in the same manner by males and females. Both of these research questions focus on assessing differential item function (DIF) across groups. When using a structural equation modeling framework, DIF is assessed in three

steps as will be demonstrated below: *configural invariance*, *metric invariance*, and *scalar invariance*.

SOS functioning across computer-based and paper-and-pencil testing medium.

All eligible sophomores completed the SOS as part of Assessment Day (Spring 06; this sample was independent from the samples used in the previous CFA analyses). Given that more students completed the paper-and-pencil SOS than the computer-based SOS, a random sample of paper-and-pencil SOS responses was selected for the analysis in order to control for sample size.

Step 1, configural invariance, the baseline model, examines whether both groups are conceptualizing the latent construct in a similar fashion (Cheung & Rensvold, 2002; Vandenberg & Lance, 2000). That is, Step 1 addresses whether students in one group interpret the items as representing the same underlying constructs as students in the other group. If configural invariance is met, a two-factor model should be supported for both groups with items corresponding to the same factors across both groups. There was support for configural invariance across testing medium (see Table 4).

In Step 2, metric invariance was assessed. If metric invariance is supported, it suggests that the strength of the item-factor relationships is similar across groups. That is, if metric invariance is supported, not only is the number and form of the factors the same across groups (two factors of Effort and Importance found across mediums; configural invariance), but the extent to which items represent the constructs is similar across groups. In the current study, metric invariance was supported.¹ That is, the factor loadings are essentially equivalent across testing mediums.

Step 3 involved testing for scalar invariance, which, when present, implies that group differences in the means of the SOS items are due to group differences in the means of the latent constructs (Effort and Importance). In other words, establishing scalar invariance permits us to interpret observed mean differences as reflections of true differences between the groups on the underlying constructs of Effort and Importance. Scalar invariance across paper-pencil and computer-based testing was supported in the current study (see Table 5).² Finding support for configural, metric, and scalar invariance provides: 1) important empirical evidence for the construct validity of the SOS; 2) allows for more confidence in the interpretation of SOS scores collected via paper-and-pencil and computer-based formats; and 3) allows for the aggregation of responses collected via paper-and-pencil and computer-based mediums.

Table 4
Tests of Invariance of SOS across Testing Medium (Sophomores S06)

Model	χ^2	df	$\Delta\chi^2$	Δdf	CFI	ΔCFI
Step 1: Configural Invariance	236.11	64	_____	_____	.9677	_____
Step 2: Metric Invariance	260.22	72	24.11*	8	.9647	<.01
Step 3: Scalar Invariance	309.43	80	49.21*	8	.9570	<.01

Note. χ^2 = chi-square; $\Delta\chi^2$ = chi-square difference; Δdf = degrees of freedom difference; CFI = comparative fit index; ΔCFI = CFI difference. $\Delta\chi^2$ and ΔCFI tests were conducted between each step and the previous step.

* $p < .05$

$N_{paper} = 323$; $N_{computer} = 323$

SOS functioning across gender. Incoming freshman completed the SOS as part of Assessment Day (Fall 07). This sample was independent from those used to assess structure. Given the greater number of female than male students at the university, a random sample of female responses was used in the analysis to control for sample size.

As described earlier, the same invariance steps were followed. Similar to testing medium, full measurement invariance of the SOS across gender was found (Table 5). Again, these results provide: 1) additional empirical evidence for the construct validity of the SOS; 2) more confidence in the interpretation of SOS scores across males and females; and 3) support for the aggregation of male and female SOS responses.

Table 5
Tests of Invariance of SOS across Gender (Freshman F06)

Model	χ^2	df	$\Delta\chi^2$	Δdf	CFI	ΔCFI
Step 1: Configural Invariance	537.66	64	_____	_____	.9463	_____
Step 2: Metric Invariance	561.79	72	24.13*	8	.9445	<.01
Step 3: Scalar Invariance	590.34	80	28.55*	8	.9442	<.01

Note. χ^2 = chi-square; $\Delta\chi^2$ = chi-square difference; Δdf = degrees of freedom difference; CFI = comparative fit index; ΔCFI = CFI difference. $\Delta\chi^2$ and ΔCFI tests were conducted between each step and the previous step.

* $p < .05$

$N_{Females} = 862$. $N_{Males} = 856$.

Internal Consistency (Reliability)

After establishing the dimensionality of the SOS, internal consistency of the scores for each subscale can be completed. In order to draw valid inferences from these scores, it is important for instruments to be functioning consistently and for scores to reflect actual ability. Cronbach's coefficient alpha, a measure of internal consistency reliability, provides an index of the strength of the relationship between items. With an upper bound of 1 and a lower bound of 0, reliability values above .80 are desirable (Nunally, 1978, as cited in Lance, Butts, & Michels, 2006). Table 6 contains the reliability estimates observed over several administration years and from our own and a variety of partner institution settings. Reliability estimates for the Importance subscale scores range from .80-.89, whereas reliability estimates for the Effort subscale scores ranged from .83-.87. These are quite acceptable reliability values.

Table 6
Cronbach's alpha, raw scores, and percent scores for SOS subscales from low-stakes testing administrations in a variety of settings

Setting	Subscale	Reliability (α)	Raw Score Average	Raw Score SD
General Education; mid-Atlantic; 4-year; public; liberal-arts				
First-year students ($N = 3111$)	Importance	.80	14.74	4.06
	Effort	.84	17.20	4.04
First-year students ($N = 3343$)	Importance	.80	14.94	3.98
	Effort	.83	17.62	3.96
Sophomores ($N = 1965$)	Importance	.83	13.84	4.36
	Effort	.85	16.97	3.97
Sophomores ($N = 2210$)	Importance	.82	13.37	4.28
	Effort	.86	17.08	4.06

General Education; mid-Western; 4-year; public; liberal-arts school				
Seniors (<i>N</i> = 1002)	Importance	.84	17.44	4.13
	Effort	.85	17.91	3.88
Exit exams; mid-Atlantic; 2-year; public; community college				
Graduating students (<i>N</i> =332)	Importance	.89	17.29	4.41
	Effort	.87	18.29	3.93
Exit exams; mid-Atlantic; 2-year; public; community college (23 institutions in state)				
Graduating students (<i>N</i> =2045; includes above sample)	Importance	.86	17.43	4.36
	Effort	.86	17.99	4.02
Course embedded; mid-Western; 4-year; public state Research I institution				
Various grade levels (<i>N</i> =1029)	Importance	.80	15.51	3.76
	Effort	.84	16.41	3.66

Note: For each setting, data were collected in a single session. Sessions were held fall 2003 to spring 2008. Possible scores for each scale range from 5 to 25. Higher scores reflect higher amounts of effort or importance.

External stage

A “crucial” (Benson, 1998, p. 14) feature of a strong program of construct validation is the ongoing collection of external evidence, which includes examination of theoretically-based hypothesized relationships with other constructs. For example, scores on the SOS would be predicted to be positively correlated with other measures of importance and effort. Additionally, students who report higher levels of motivation, as measured by the SOS, would be expected to perform on cognitive tests at a higher level than those who report lower levels of motivation.

In the sections that follow, relationships between SOS subscale scores and external measures of effort and perception of importance are presented. Typically, most research on test-taking motivation has dealt with effort. Wise (this issue) discusses the

weighty implications of low examinee effort on validity of score inferences. On the other hand, student perceptions of importance neither stimulate policy decisions nor lead researchers to conceive of novel ways to impact students. Importance, as measured on the SOS, seems to be a function of perceived context. The scores we have observed are in line with what one expects theoretically: as testing stakes increase, scores on the Importance subscale increase. In other words, students basically understand the nature of the testing situation in which they are asked to perform.

Therefore, the reader should note that the external scores reported below are most often Effort scores, while actual changes in administrative procedures leading to increased stakes are more relevant to the review of Importance scores.

Relationships with External Measures

Response time effort. Response time effort (RTE; Wise & Kong, 2005) is an example of an alternate measure of motivation that is theoretically related to the SOS effort scores. The RTE is based on the concept that students taking a computer-based test predominantly engage in either *solution behavior* or *rapid-guessing behavior* (simply scrolling through items and rapidly and randomly clicking response alternatives on a computer-based test). Ranging from 0 to 1, the RTE represents the proportion of items on a test on which the student engaged in solution behavior. Consequently, students who are not motivated and simply click through the questions on a computer-based test will score low on RTE, while students who spend time on each item will score high. One of the appeals of the RTE measure is that, unlike the SOS, it is *not* a self-report measure, and is computed without student awareness. Wise and Kong hypothesized that students reporting that they put forth more effort (as measured by the SOS Effort subscale) would

have a higher RTE score. As hypothesized, SOS Effort scores were positively correlated with RTE ($r = .54$).

Wise and Kong (2005) also created groups of students based on RTE. Specifically, they computed SOS Effort scores for three groups of students: 1) low RTE (i.e., less than 80% solution behavior on a computer-based test and greater than 20% rapid-guessing behavior); 2) mid-range RTE (between 80%-90% solution behavior); and 3) high RTE (i.e., greater than 90% solution behavior and less than 10% rapid-guessing behavior). The three groups significantly differed in their SOS Effort scores, with students with lower RTE reporting lower SOS Effort scores, and those with higher RTE having higher SOS Effort scores.

Performance. Effort positively correlates with achievement scores. Schiel (1996) looked at the relationship between reported effort (as measured by a single multiple-choice item) and Collegiate Assessment of Academic Proficiency (CAAP) scores. His analysis of over 50,000 test scores from nearly 200 different schools revealed that students who expended “reasonable” effort earned scores 0.25 to 1.5 standard deviations higher than students expending no effort. Another study investigating performance and effort was conducted by Thelk (2006). She administered the SOS along with a general education measure of quantitative and scientific reasoning to a group of graduating community college students. The correlation between Effort and test score was $r = .30$. Wise and Kong (2005) correlated the SOS with a test of information literacy. Effort was found to be moderately correlated with test score ($r = .34$). Of particular note, Effort was not significantly related to SAT Verbal or SAT quantitative scores ($r = .14$ and $r = .01$,

respectively). This finding is theoretically sound, given the observation that general ability (for which SAT scores can be used as a proxy) has no relationship with Effort.

Swerdzewski, Harmes and Finney (this issue) compared performance and motivation of students who attended a regular assessment session with those who failed to attend but participated in a required make-up testing session. The students who avoided the regular session and did not report high SOS Effort on the actual tests performed lower than those students in make-up testing who did exert effort. For the students in make-up testing who reported giving effort, their scores looked comparable to the students who took the tests at the regular assessment session. These results provide additional support for confident interpretation of SOS Effort scores.

Motivation filtering. Scores from the SOS Effort subscale have also been used as a motivation filter (Sundre & Wise, 2003; Wise & Kong, 2005), providing further evidence for the scores' use in drawing inferences about student motivation. Sundre and Wise (2003) reported no relationship between student SAT and SOS scores. They also reported incremental increases in scientific reasoning test performance and external validity coefficients with incremental removal of low motivated examinees. A more recent study conducted by Wise and Kong (2005) revealed that when information literacy test scores were incrementally removed from a data set based on the students' SOS Effort scores, the average information literacy test scores increased incrementally. In other words, removing students with low reported Effort scores resulted in higher cognitive test scores; this result continued through incremental motivation filtering.

Swerdzewski, Finney, and Harmes (2007) evaluated the consistency between the SOS and RTE methods of motivation filtering. They found that using the SOS to identify

unmotivated examinees resulted in the expulsion of more data than the RTE method afforded. However, the mean scores for each group were essentially the same, suggesting that testing professionals could use either method for data filtering. However, using the SOS may be preferable due to its compatibility with either computer-based or pencil-and-paper testing formats. Wise (this issue) provides a thorough discussion of motivation filtering as one method of actually dealing with and understanding the impact of data from unmotivated examinees.

Change in stakes. As anticipated based on expectancy-value theory, we have observed that the SOS Importance and Effort scores reflect changes in testing stakes. One study of SOS scores in a high-stakes setting at James Madison University involved graduating social work majors who completed the SOS following their compulsory comprehensive examination, a very high-stakes assessment. Social Work seniors must complete and pass both the written and oral comprehensive examinations to receive their baccalaureate Social Work (BSW) degree. Table 7 presents SOS data from this condition over six administrations. With all cohorts but one, means exceed 21, indicating that after recoding negatively-worded items, most students responded with at least a 4 (“Agree”) on the five-point scale across all five subscale items. The observed scores for high-stakes testing conditions are consistently much higher than those observed in low-stakes, non-consequential testing contexts, providing additional evidence that SOS scores provide an accurate reflection of examinee effort and importance.

Table 7
Cronbach’s alpha and descriptive statistics for SOS subscale scores from high-stakes administration (Social Work comprehensive exam) at four-year institution

Cohort (N)	Score	Reliability	Raw Score	Raw Score SD
------------	-------	-------------	-----------	--------------

		(α)	Average	
Fall 2008 (N=11)	Importance	.96	22.45	4.70
	Effort	.95	19.91	5.13
Spring 2008 (N=25)	Importance	*	22.80	1.22
	Effort	.62	21.28	2.37
Fall 2007 (N=14)	Importance	.83	21.07	3.17
	Effort	.84	21.36	3.30
Spring 2007 (N=40)	Importance	.73	22.20	2.71
	Effort	.66	21.80	2.39
Fall 2006 (N=15)	Importance	.65	21.93	2.22
	Effort	.81	21.47	3.02
Spring 2006 (N=25)	Importance	.62	22.32	2.63
	Effort	.71	22.04	3.05

*The value is not interpretable (negative) due to lack of variance of the subscale for this cohort. The low standard deviation reflects this phenomenon.

The SOS is sensitive to stakes in that a ceiling effect may be noticed in high-stakes administrations. As personal consequences increase, Effort and Importance scores go up, leading to a decrease in group variance. Reliability estimates will decrease as a direct result. This phenomenon can be observed in Table 7, in which the Social Work majors report highly homogenous responses to the SOS items. This result reinforces the notion that motivation will vary most in low-stakes situations. In other words, when less is at stake, students display greater variation in the amount of effort they choose to expend and the importance they assign to the assessment activities. In contrast to the data in the rest of the table, the reliability estimates for the Fall 2008 students are notably

higher. Upon further investigation, the Social Work department chair conveyed that one student had inadvertently reversed the item response scale. This phenomenon relates how a single student can spuriously increase variability and, as a result, reliability estimates.

It is very interesting to note the differences in student self reports of Importance and Effort across testing conditions that vary in whether or not there are personal consequences. Again reviewing the Social Work graduating seniors' SOS scores from a very high stakes testing situation, and comparing them with a sample of Assessment Day sophomores, the differences are quite stunning. Table 8 contains the effect size calculations of the differences in subscales across these two distinct testing conditions. The effect sizes are of statistical and practical significance. The Importance subscale had an effect size of over two standard deviations. This was a full standard deviation greater than that observed for the Effort subscale, which was still quite large at 1.30. However, while the perceived importance of the assessment tasks clearly reflects the reality of the situation, it is consoling to note that Effort was not as heavily impacted. It is the Effort subscale that has proven to be the most useful for our studies of examinee motivation. We can inform students of the importance of their performances to the institution, but this may not relate to any changes in student perceptions or behavior. We are most interested in influencing the amount of effort with which they engage in our assessment tasks. The Lau, Swerdzewski, Jones, Anderson, & Markle paper (this issue) provides a very interesting discussion of their studies and strategies to impact student effort in testing conditions.

Table 8

Effect sizes (Cohen's d) for Differences between High Stakes (Social Work seniors) and Low Stakes (Spring 2006 assessment day examinees) Administrations for SOS Subscales

Importance	Effort
2.30	1.30

Note: Cohen's d can be interpreted as the difference between the two means, in terms of standard deviations. A d value above 0.50 can be considered to represent a large effect size (Kline, 1998, p. 149).

Sundre and Kitsantas (2003) used the SOS to report differences in student motivation and performances across consequential conditions (counted toward grade; did not count toward grade) and test modalities (multiple choice; essay). These studies were conducted using course embedded tests in which students took parallel forms of tests that either counted or did not count. Their study provided clear evidence that Importance and Effort were impacted in theoretically expected ways. More specifically, SOS scores and performances declined dramatically when the test 'did not count' for their grade. More importantly, their study also revealed that arduous tasks such as essays, really take a toll when there are no consequences associated with them—even when students were prepared to complete them. These results have led to more studies associated with effort and how to improve it in testing conditions (see Lau et al in this issue).

Conclusion

The use of the SOS has provided us with many opportunities to directly address the concerns of many faculty members and their committees in discussing and accepting general education assessment results. Our faculty members have expressed greater confidence in interpreting data derived from low-stakes Assessment Day activities as a result of using SOS data in tandem with score dissemination. A few of our general

education steering committees have specifically requested analyses that incorporate SOS scores in order to determine how students within different majors or schools react to testing and perform outside the classroom. Some of our general education steering committees are now requesting that motivation filtering techniques be employed to remove examinees with low motivation from our analyses and reporting structures.

Our ability to monitor examinee motivation across assessment tasks has shaped not only the way we conduct our general education test administrations but it has also impacted the crafting of assessment instruments. For example, we know from the results of our focus groups conducted with students, that their stated expectancy for success on test items is increased by not labeling the tests as being related to a specific science course. We have titled our test of scientific and quantitative reasoning “The Natural World Test,” because this is less threatening. All of our items have been developed to assess scientific reasoning and thinking, not specific scientific content. This is a reflection of our evolving understanding of what general education is. During a focus group, one of the students indicated relief at the nature of the items and said that she felt that “she had a chance to perform well” with this type of item. Further, our students have indicated to us that if we create interesting items, they will spend more time trying to solve them. They actually like items that incorporate tables, graphs, and figures, rather than extended reading passages. Both of these modifications will positively impact student self efficacy. We also know that expectancy and value directly influence student behavior during assessment testing.

Low examinee motivation is a potential form of test bias. It is a form of systematic error that negatively influences student test performance. It is evident that

students can show up and elect not to perform to capacity during a testing session. This is systematic error that diminishes test scores. We now have evidence that the SOS can provide solid estimates of the presence and magnitude of this error. We have found the SOS Importance and Effort subscores to be highly useful in framing appropriate interpretations of scores.

We have found the study of examinee motivation to be a very fruitful area of applied and theory based research. In an effort to increase the importance of assessment tests, we are currently working on new methods to provide feedback to our students on their test scores using our eCampus information system. We have designed an experiment to assess whether or not access to scores and norm and criterion referenced interpretive information will impact their performances and motivation. The SOS will help us to monitor these experimental manipulations. We invite you to use the SOS and to join us in this important conversation. There is much more to be learned and shared about student motivation in general education learning and assessment.

References

- Benson, J. (1998). Developing a strong program of construct validation: A test anxiety example. *Educational Measurement: Issues and Practice, 17*, 10-17.
- Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*, 233-255.
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity of psychological tests. *Psychological Bulletin, 52*, 281-302.
- Eccles, J. F., Adler, T. F., Futterman, R., Goff, S. B., Kaczala, C. M., Meece, J. L., & Midgley, C. (1983). Expectancies, values, and academic behaviors. In J. T. Spence (Ed.), *Achievement and Achievement Motives* (pp. 75-146). San Francisco, CA: W.H. Freeman.
- Hancock, G. R. (1997). Structural equation modeling methods of hypothesis testing of latent variable means. *Measurement and Evaluation in Counseling and Development, 30*, 91-105.
- Corno, L. & Kanfer, R. (1993). The role of volition in learning and performance. *Review of Research in Education, 19*, (1), 301-341.
- Jaschik, S. (2009). Assessing assessment. *Inside Higher Ed*, retrieved January 23, 2009 from <http://insidehighered.com/news/2009/01/23/assess>.
- Kline, R. B. (1998). *Principles and Practice of Structural Equation Modeling*. New York: Guilford.
- Lance, C. E., Butts, M. M., & Michels, L. C. (2006). The sources of four commonly reported cutoff criteria. *Organizational Research Methods, 9*(2), 202-220.

- Lau, A. R., Swerdzewski, P. J., Jones, A. T., Anderson, R. A., & Markle R. E. (this issue). Proctors matter: Strategies for increasing examinee effort on general education program assessments. *Journal of General Education*.
- MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin, 111*, 490-504.
- Marsh, H. W., Hau, K-T, & Grayson, D. (2005). Goodness of fit in structural equation models. In A. Maydeu-Olivares & J. McArdle (Eds.), *Contemporary Psychometrics. A festschrift for Roderick P. McDonald* (pp. 275–340). Mahwah, NJ: Lawrence Erlbaum.
- Marsh, H. W., Hau, K-T, & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers to overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling, 11*, 320–341.
- Pintrich, P. R. (1989). The dynamic interplay of student motivation and cognition in the college classroom. *Advances in Motivation and Achievement: Motivation Enhancing Environments, 6*, 117-160.
- Pintrich, P. R., & DeGroot, E. V. (1990). Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology, 82*(1), 33-40.
- Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, M. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology, 88*, 879-903.

- Quintana, S. M. & Maxwell, S. E. (1999). Implications of recent developments in structural equation modeling for counseling psychology. *The Counseling Psychologist, 27*, 485-527.
- Schiel, J. (1996). *Student effort and performance on a measure of postsecondary educational development*. (ACT Research Report No. 96-9). Iowa City, IA: ACT.
- Sharma, S., Mukherjee, S., Kumar, A., & Dillon, W. R. (2005). A simulation study to investigate the use of cutoff values for assessing model fit in covariance structure models. *Journal of Business Research, 58*, 935–943.
- Steenkamp, J., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research, 25*, 78–90.
- Sundre, D. L. (1997). *Differential examinee motivation and validity: A dangerous combination*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.
- Sundre, D. L. (1999). *Does examinee motivation moderate the relationship between test consequences and test performance?* (Report No. TM029964). Harrisonburg, Virginia: James Madison University. (ERIC Documentation Reproduction Service No. ED432588).
- Sundre, D. L., & Kitsantas, A. (2003). An exploration of the psychology of the examinee: Can examinee self-regulation and test-taking motivation predict consequential and non-consequential test performance? *Contemporary Educational Psychology, 29*, 6-26.

- Sundre, D. L. & Moore, D. L. (2002). The Student Opinion Scale: A measure of examinee motivation. *Assessment Update*, 14 (1), 8-9.
- Sundre, D. L., & Wise, S. L. (2003, April). 'Motivation Filtering:' *An exploration of the impact of low examinee motivation on the psychometric quality of tests*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.
- Swerdzewski, P. J., Finney, S. J., & Harmes, J. C. (2007, October). *Examinee motivation in low-stakes testing: Two approaches to identifying data from low-motivated students in an applied assessment context*. Paper presented at the annual meeting of the Northeastern Educational Research Association, Rocky Hill, CT.
- Swerdzewski, P. J., Harmes, J. C. & Finney, S. J. (this issue). Skipping the test: Using empirical evidence to inform policy related to students who avoid taking low-stakes assessments in college. *Journal of General Education*.
- Thelk, A. D. (2006). *Examinee awareness of performance expectation and its effects on motivation and test scores*. Unpublished doctoral dissertation, James Madison University.
- Vandenberg, R. J. & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, 3, 4-69.
- Wise, S. L. (this issue). Strategies for managing the problem of unmotivated examinees in low-stakes testing programs. *Journal of General Education*.

- Wise, S. L. & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education, 18*(2), 162-183.
- Wolf, L. F. & Smith, J. K. (1993, April). *The effects of motivation and anxiety on test performance*. Paper presented at the annual meeting of the American Educational Research Association, Atlanta.
- Wolf, L. F., Smith, J. K., & Birnbaum, M. E. (1995). Consequences of performance, test motivation, and mentally taxing items. *Applied Measurement in Education, 8*(4), 341-351.
- Yin, P. & Fan, X. (2003). Assessing the factor structure invariance of self-concept measurement across ethnic and gender groups: Findings from a national sample. *Educational and Psychological Measurement, 63*, 296-318.

Footnotes

¹ Because the metric model was nested within the configural model, both the chi-square difference ($\Delta\chi^2$) and CFI difference (ΔCFI) were computed to compare these nested invariance models. Both difference indexes were used because the $\Delta\chi^2$ is an exact test: “It should be noted that, like the χ^2 statistic, the $\chi^2_{\text{comparison}}$ statistic is an exact test, which may be overly restrictive” (Quintana & Maxwell, 1999, p. 506). Steenkamp and Baumgartner (1998) recommended evaluating change in fit indices in addition to the $\Delta\chi^2$ test to evaluate measurement invariance (see also Yin & Fan, 2003). More specifically, Cheung and Rensvold (2002) recommended reporting ΔCFI values and suggested that values at or less than .01 indicate a change in fit that is not practically significant.

² Although the values of the descriptive fit indexes were slightly worse for the scalar model than for the configural or metric models, the values were still acceptable. In addition, the ΔCFI indicated that the overall fit of the scalar model was not practically worse than the fit of the metric model (see Cheung & Rensvold, 2002).