

Measuring Learning Outcomes in Higher Education: Motivation Matters

Ou Lydia Liu¹, Brent Bridgeman¹, and Rachel M. Adler¹

With the pressing need for accountability in higher education, standardized outcomes assessments have been widely used to evaluate learning and inform policy. However, the critical question on how scores are influenced by students' motivation has been insufficiently addressed. Using random assignment, we administered a multiple-choice test and an essay across three motivational conditions. Students' self-report motivation was also collected. Motivation significantly predicted test scores. A substantial performance gap emerged between students in different motivational conditions (effect size as large as .68). Depending on the test format and condition, conclusions about college learning gain (i.e., value added) varied dramatically from substantial gain ($d = 0.72$) to negative gain ($d = -0.23$). The findings have significant implications for higher education stakeholders at many levels.

Keywords: accountability; assessment; higher education; motivation; outcomes assessment; regression analyses

Accountability and learning outcomes have received unprecedented attention in U.S. higher education over the past 5 years. Policymakers call for transparent demonstration of college learning (U.S. Department of Education, 2006). Accrediting associations have raised expectations for institutions to collect evidence of student learning outcomes and use such information for institutional improvement. For instance, the Council for Higher Education Accreditation (CHEA), the primary organization for voluntary accreditation and quality assurance to the U.S. Congress and Department of Education, has focused on the role of accreditation in student achievement by establishing the CHEA Award for Outstanding Institutional Practice in Student Learning Outcomes (http://www.chea.org/chea%20award/CA_2011.02-B.html). Various accountability initiatives press higher education institutions to provide data on academic learning and growth (Liu, 2011a; Voluntary System of Accountability, 2008). Facing mounting pressure, institutions turn to standardized outcomes assessment to fulfill accountability, accreditation, and strategic planning requirements. Outcomes assessment provides a direct measure of students' academic ability and is considered a powerful tool to evaluate institutional impact

on students (Kuh, Kinzie, Buckley, Bridges, & Hayek, 2006). Research on outcomes assessment has generated strong interest from institutional leaders, state officials, and policymakers. Based on outcomes assessment data, researchers are making conclusions about the current state of U.S. higher education and are offering policy recommendations (e.g., *Arum & Roksa*, 2011). However, a frequently discussed yet insufficiently researched topic is the role of students' performance motivation when taking low-stakes outcomes assessments. Although highly relevant to institutions, the test scores usually have no meaningful consequence for individual students. Students' lack of motivation to perform well on the tests could seriously threaten the validity of the test scores and bring decisions based on the scores into question. The current study is intended to contribute to the understanding of how motivation may affect outcomes assessment scores and, in particular, affect conclusions about U.S. higher education based on outcomes assessment results. The study also suggests practical ways to increase test takers' motivation on higher performance on low-stakes tests.

Outcomes Assessment in Higher Education

A systematic scrutiny of U.S. higher education was marked by the establishment of the Spellings Commission in 2005. The Commission lamented the remarkable lack of accountability mechanisms to ensure college success and the lack of transparent data that allow direct comparison of institutions (U.S. Department of Education, 2006). As a result, several accountability initiatives (e.g., Voluntary System of Accountability [VSA], Transparency by Design, Voluntary Framework of Accountability) were launched by leading educational organizations representing different segments of U.S. higher education (e.g., public institutions, for-profit institutions, community colleges). A core component of these accountability initiatives is the requirement that participating institutions provide evidence of student learning that is scalable and comparable. Take the VSA as an example: Among other requirements, it asks institutions to use one of three nationally normed measures (ETS[®] Proficiency Profile,¹ Collegiate Learning Assessment [CLA], or Collegiate Assessment of Academic Proficiency) to report college learning (VSA, 2008).

Both criticized and acclaimed, outcomes assessment has been gradually accepted by at least some in the higher education community. Since 2007, VSA alone has attracted participation from

¹Educational Testing Service, Princeton, NJ

361 institutions in 49 states. Over the past 5 years, more than one thousand higher education institutions have used at least one form of standardized outcomes assessment for purposes such as meeting accreditation requirements, fulfilling accountability demands, improving curricular offerings, and evaluating institutional effectiveness (Educational Testing Service [ETS], 2010; Kuh & Ikenberry, 2009; Liu, 2011a).

Accompanying the wide application of outcomes assessment is an emerging line of research focusing on the interpretation of college learning using outcomes assessment data (Liu, 2008), identifying proper statistical methods in estimating learning gain, or value-added (Liu, 2011b; Steedle, 2011), and comparing findings from outcomes assessments of different contents and formats (Klein et al., 2009).

Among recent research on outcomes assessment, a most noteworthy finding came from the book *Academically Adrift* (Arum & Roksa, 2011). The authors claimed that CLA data indicated that students gained very little academically from their college experience. By tracking the CLA performance of a group of freshmen to the end of their sophomore year, the authors found that on average, students made only a 7 percentile point gain (.18 in effect size) over the course of three college semesters. More than 45% of the students failed to make any progress as measured by the CLA. In addition, the performance gap tended to increase between racial/ethnic minority students and White students. The findings attracted wide attention from researchers and policy-makers and were frequently cited when U.S. students' minimal college learning was mentioned (Ochoa, 2011). However, this study was not accepted without criticism. Astin (2011) provided a substantial critique of this study, questioning its conclusion of limited college learning based on several major drawbacks: lack of basic data report, making conclusions about individual students without student-level score reliabilities, unsound statistical methods for determining improvement, and incorrect interpretation of Type I and Type II errors. What Astin didn't mention was the study's failure to consider the role of motivation when students took the CLA. Prior research found that the year-to-year consistency in institutional value-added scores was fairly low (0.18 and 0.55 between two statistical methods) when the CLA was used (Steedle, 2011). It seems likely that motivation may play a significant role in the large inconsistency in institutional rankings.

Research on Test-Taking Motivation

Students' motivation in taking low-stakes tests has long been a source of concern. In the context of outcomes assessment in higher education, institutions differ greatly in how they recruit students for taking the assessments. Some institutions set up specific assessment days and mandate students to take the test. Other institutions offer a range of incentives to students (e.g., cash rewards, gift certificates, and campus copy cards) in exchange for participation. However, because the test results have little impact on students' academic standing or graduation, students' lack of motivation to perform well on the tests could pose a serious threat to the validity of the test scores and the interpretation accuracy of the test results (Banta, 2008; Haladyna & Downing, 2004; Liu, 2011b; S. L. Wise & DeMars, 2005, 2010; V. L. Wise, Wise, & Bhola, 2006).

A useful theoretical basis for evaluating student test taking motivation is the expectancy-value model (Pintrich & Schunk, 2002). In this model, *expectancy* refers to students' beliefs that they can successfully complete a particular task and *value* refers to the belief that it is important to complete the task. Based on this theoretical model, researchers have developed self-report surveys to measure student motivation in taking low-stakes tests. For example, the Student Opinion Survey (SOS; Sundre, 1997, 1999; Sundre & Wise, 2003) is one of the widely used surveys that capture students' reported effort and their perception of the importance of the test. A general conclusion from studies investigating the relationship between student motivation and test performance is that highly motivated students tend to perform better than less motivated students (Cole & Osterlind, 2008; O'Neil, Sugrue, & Baker, 1995/1996; Sundre, 1999; S. L. Wise & DeMars, 2005; V. L. Wise et al., 2006). A meta-analysis of 12 studies consisting of 25 effect size statistics showed that the mean performance difference between motivated and unmotivated students could be as large as .59 standard deviations (S. L. Wise & DeMars, 2005). Besides relying on student self-report, researchers have also examined response time effort (RTE) for computer-based, unspeeded tests to determine student motivation (S. L. Wise & DeMars, 2006; S. L. Wise & Kong, 2005). Results show that RTE is significantly correlated with student self-reported motivation, but not with measures of student ability, and is also a significant predictor of their test performance.

To eliminate the impact of low performance motivation on test results, researchers have explored ways to filter responses from unmotivated students identified through either their self-report or response time effort (S. L. Wise & DeMars, 2005, 2006; S. L. Wise & Kong, 2005; V. L. Wise et al., 2006). The findings are consistent; after controlling for students' general ability (e.g., SAT scores), motivation filtering helps improve the validity of the inferences based on the test results (S. L. Wise & DeMars, 2005, 2010; V. L. Wise et al., 2006; Wolf & Smith, 1995).

Realizing the important impact of motivation on test results, researchers have explored ways to enhance student motivation to maximize their effort in taking low-stakes tests. Common practices include increasing the stakes of the tests by telling students that their scores contribute to their course grades (Sundre, 1999; Wolf & Smith, 1995), providing extra monetary compensation for higher performance (Baumert & Demmrich, 2001; Braun, Kirsch, & Yamamoto, 2011; Duckworth, Quinn, Lynam, Loeber, & Stouthamer-Loeber, 2011; O'Neil, Abedi, Miyoshi, & Mastergeorge, 2005; O'Neil et al., 1995/1996), and providing feedback after the test (Baumert & Demmrich, 2001; Wise, 2004). Increasing the stakes and providing extra payment for performance have been shown to be effective ways to motivate students (Duckworth et al., 2011; O'Neil et al., 1995/1996; Sundre, 1999). For instance, through a meta-analysis of random assignment experiments, the Duckworth et al. (2011) study found that monetary incentives increased test scores by an average of .64 standard deviations. Despite the intuitive appeal of providing feedback, it does not appear to have an impact on either student motivation or their test performance (Baumert & Demmrich, 2001; V. L. Wise, 2004).

Table 1
Descriptive Statistics by Institution

| | <i>N</i> | Female (%) | Test Scores ^a | | Part-time (%) | Language ^b (%) | White (%) | College GPA | |
|----|----------|------------|--------------------------|-----------|---------------|---------------------------|-----------|-------------|-----------|
| | | | <i>M</i> | <i>SD</i> | | | | <i>M</i> | <i>SD</i> |
| RI | 340 | 54 | 1,213 | 154 | 2 | 72 | 74 | 3.16 | .81 |
| MI | 299 | 63 | 1,263 | 145 | 1 | 73 | 81 | 3.33 | .52 |
| CC | 118 | 59 | 168 | 30 | 24 | 76 | 48 | 3.21 | .61 |

Note. RI = research university; MI = master's university; CC = community college.

^aThe numbers represent composite SAT scores or converted ACT scores for the research and master's institutions and composite placement test scores (reading and writing) for the community college.

^bEnglish as better language.

Rationale and Research Questions

Although motivation on low-stakes tests has been studied in higher education, there is a compelling need for such a study for widely used standardized outcomes assessment. Prior studies that experimentally manipulated motivational instructions examined locally developed assessments that were content-based tests in specific academic courses as opposed to large-scale standardized tests (Sundre, 1999; Sundre & Kitsantas, 2004; Wolf & Smith, 1995). It is unclear whether conclusions drawn from these course-based assessments can be extended to widely used standardized tests used for outcomes assessments. The distinction between these two types of examinations is critical because of the types of motivational instructions that are feasible differ by test type. In a course-based test, the instruction that the score will contribute to the course grade is believable. But for a general reasoning test of the type used for value-added assessments in higher education, an instruction indicating that the score would contribute to the grade in a specific course would not be plausible. In addition, most previous studies relied on data from a single program or single institution (Sundre & Kitsantas, 2004; S. L. Wise & Kong, 2005; V. L. Wise et al., 2006; Wolf & Smith, 1995), which may limit the generalizability of the findings. Furthermore, most previous studies either used self-report or item response time to determine examinees' motivation and use that information to investigate the relationship between motivation and performance. Very few studies created motivational manipulation to understand the magnitude of effect motivation may have on test scores.

By creating three motivational conditions that were plausible for a general reasoning test, we addressed three research questions in this study: What is the relationship between students' self-report motivation and test scores? Do motivational instructions affect student motivation and performance? Do conclusions drawn about college learning gain change with test format (i.e., multiple choice vs. essay) and motivational instruction?

Existing literature has addressed some discrete aspects of these questions, but no study has provided a complete answer to all of these questions for a large-scale standardized outcomes assessment. In sum, this study is unique on three aspects: (1) a focus on a large-scale general reasoning assessment, (2) the inclusion of multiple institutions in data collection, and (3) the creation of plausible motivational conditions with random assignment.

Methods

Participants

A total of 757 students were recruited from three higher education institutions (one research institution, one master's institution, and one community college) in three states. See Table 1 for participants' demographic information. The student profiles were similar between the research and master's institutions. The community college had a significantly larger percentage of part-time and non-White students than the two 4-year institutions. Participants were paid \$50 to complete the tests and the survey. We obtained information from each institution's registrar's office on the percentage of females, ethnic composition, and mean admission/placement test scores; the volunteer participants were representative of their home institutions in terms of gender, ethnicity, and admission/placement test scores.

Since first-year students may be more intimidated (and therefore more motivated) by taking even a low-stakes test, we recruited only students with at least 1 year of college experience at the 4-year institutions and students who had taken at least three courses at the community college.

Instruments

We administered the ETS Proficiency Profile, including the optional essay, to the 757 college students. The Proficiency Profile measures college-level skills in critical thinking, reading, writing, and mathematics and has been used by over 500 institutions as an outcomes assessment for the past 5 years. The reliabilities for the subscales are over .78 for student-level data and over .90 for institution-level data (Klein et al., 2009). Abundant research has been conducted examining the test's construct validity, content validity, predictive validity, and external validity (Belcheir, 2002; Hendel, 1991; Klein et al., 2009; Lakin, Elliott, & Liu, in press; Liu, 2008; Livingston & Antal, 2010; Marr, 1995). Students with higher Proficiency Profile scores tend to have gained more course credits (Lakin et al., in press; Marr, 1995). Students' Proficiency Profile performance is consistent with the skill requirements of their major fields of study, with humanities majors scoring higher than other students on critical thinking and writing and mathematics and engineering students scoring higher on mathematics (Marr, 1995). Proficiency Profile scores are also highly correlated with scores from tests that measure similar constructs (Hendel, 1991; Klein et al., 2009). In addition, the Proficiency Profile is

able to detect performance differences between freshmen and seniors after controlling for college admission scores (e.g., SAT) (Liu, 2011b). Although researchers have examined various aspects of validity for the Proficiency Profile, one less explored aspect is how the test scores predict post-college performance in various academic, workforce, and community settings. Such evidence is also scarce for other types of outcomes assessment. The only study that we are aware of is the follow-up study to Arum and Roksa's (2011) study, which we discuss at the end of the article under "A Cautionary Note."

There are two versions of the Proficiency Profile, a 108-item test intended to yield valid scores at the individual student level and a 36-item short form intended primarily for group-level score reporting (ETS, 2010). Because of the limited amount of testing time, we used the short form, which can be completed in 40 minutes.

An essay, which measures college-level writing ability, is an optional part of the Proficiency Profile. The essay prompt asks students to demonstrate their writing ability by arguing for or against a point of view. For example, the prompt may provide one point of view and solicit students' opinions about it. Students are asked to support their position with justifications and specific reasons from their own experiences and observations. It took the students about 30 minutes to complete the essay. In each testing session, students took the online version of the Proficiency Profile and the essay with a proctor monitoring the testing room.

After completing the tests, students filled out the SOS by hand (Sundre, 1997, 1999; Sundre & Wise, 2003). The SOS is a 10-item survey that measures students' motivation in test taking. The survey has been widely used in contexts of outcomes assessment similar to this study.

Following the test administration, undergraduate admission test scores were obtained for the students at the research and master's institutions, and placement test scores were obtained for the students from the community college. All test scores were obtained from the registrars' offices.

Experimental Conditions

To address the three research questions described in the introduction, we designed an experiment with three motivational conditions, represented by three different consent forms. Within each testing session, students were randomly assigned to conditions before they took the tests. The consent forms were identical for the three conditions, except that the following instructions were altered based on the different motivational conditions:

Control condition: Your answers on the tests and the survey will be used only for research purposes and will not be disclosed to anyone except the research team.

Personal condition: Your answers on the tests and the survey will be used only for research purposes and will not be disclosed to anyone except the research team. However, your test scores may be released to faculty in your college or to potential employers to evaluate your academic ability.

Institutional condition: Your answers on the tests and the survey will be used only for research purposes and will not be disclosed to anyone except the research team. However, your test scores will be averaged with all other students taking the test at your college.

Only this average will be reported to your college. This average may be used by employers and others to evaluate the quality of instruction at your college. This may affect how your institution is viewed and therefore affect the value of your diploma.

The three instructions were highlighted in bold red letters so students would likely notice them before giving their consent. After the data collection was completed, students in the treatment conditions were debriefed that their test scores would not be shared with anyone outside of the research team. Among the three conditions, we expected the personal condition to have the strongest effect on students' motivation and performance as it is associated with the highest stakes for individual students. We also expected the institutional condition to have some impact on students' motivation and performance as maintaining their institution's reputation could be a motivator for students to take the test more seriously than usual. The conditions were approved by the Institutional Review Board at both the researcher's institution and the three institutions where the data collection took place. The students in the institutional and personal conditions were debriefed after the data collection was completed and were assured that their scores would not actually be reported to faculty or potential employers.

Because students were randomly assigned to the conditions within a testing room, before the testing they were instructed to raise their hand if they had a question instead of asking that question in front of the class; thus, no student could realize that other students in their room had different instructions.

Analyses

Multiple linear regression analyses were used to investigate the relationship between self-reported motivation and test scores. The predictors were SOS scores and admission (or placement) test scores, and the outcome variables were the Proficiency Profile and essay scores, respectively. For students from the two 4-year institutions, the admission scores were the composite SAT critical reading and mathematics scores (or converted ACT scores based on the concordance table provided by ACT and the College Board at <http://www.act.org/aap/concordance/>). For students from the community college, the placement scores were the composite reading and writing scores from the eCompass, an adaptive college placement test. The regression analysis was conducted separately for each institution and each dependent variable. The admission (or placement test) scores were entered into the equation first, followed by mean SOS. The change in R^2 was examined to determine the usefulness of the predictors. Pearson correlations were also calculated among test scores, admission scores, and SOS scores.

An ANOVA was conducted to investigate the impact of the motivational conditions on self-reported motivation and on test scores. The Bonferroni correction was used for post hoc comparisons between conditions to adjust the Type I error rate for multiple comparisons. Standardized mean differences were computed between the three motivational conditions on the SOS, the Proficiency Profile, and essay scores. A separate analysis was conducted for each measure and each institution. Two-way ANOVAs were also conducted to investigate any interaction between the three institutions and the motivational instructions.

Table 2
Pearson Correlations Among Test Scores and Predictors

| | Test Score ^a | SAT ^b | Self-Report Motivation |
|------------------------|-------------------------|------------------|------------------------|
| RI | | | |
| Test score | — | 0.71** | 0.29** |
| SAT | 0.34** | — | 0.18* |
| Self-report motivation | 0.25** | 0.18* | — |
| MI | | | |
| Test score | — | 0.61** | 0.39** |
| SAT | 0.27** | — | 0.16* |
| Self-report motivation | 0.32** | 0.16* | — |
| CC | | | |
| Test score | — | 0.31** | 0.24** |
| Placement | 0.51** | — | 0.07 |
| Self-report motivation | 0.27** | 0.07 | — |

Note. RI = research university; MI = master's university; CC = community college.

^aUpper diagonal values are the Proficiency Profile total scores and lower diagonal values are the essay scores.

^bFor the community college this is the placement test scores.

* $p < .05$. ** $p < .01$.

A general linear model (GLM) analysis was used to address the research question on college learning gain in SPSS. In the GLM, the Proficiency Profile and essay scores were used as separate outcomes variables, with motivational condition and class status being fixed factors, and SAT scores as a covariate. In the case of this study, the GLM analysis is equivalent to a two-way analysis of covariance. A homoscedasticity test was conducted to evaluate the homogeneity assumption for the GLM. Note that only students from the two 4-year institutions were included for this analysis since the learning gain was indicated by the performance between sophomores and seniors. The class status was classified based on number of credits completed: sophomore (30-60 credits), junior (60-90 credits), and senior (more than 90 credits). The analyses were done separately for the Proficiency Profile and the essay.

Results

Reliabilities

The Cronbach's alpha for the abbreviated Proficiency Profile was .83 for the research institution, .86 for the master's institution, and .85 for the community college. The Cronbach's alpha for the SOS motivation scale was .84 for the research institution, .85 for the master's institution, and .84 for the community college.

Relationship Between Self-Report Motivation and Test Performance

Pearson correlations among SAT (or placement) scores, Proficiency Profile test scores (multiple choice and essay), and SOS scores, separately for each institution, are in Table 2. Multiple choice test scores are above the diagonal and essay scores below. All correlations were significant ($p < .05$) except for the correlation between SOS and placement scores at the community college.

After controlling for SAT or placement scores, self-report motivation was a significant predictor of both the Proficiency

Profile and essay scores, and the finding was consistent across the three institutions (see Table 3). The standardized coefficients ranged from .17 to .26 across institutions. After the variable mean SOS was added to the equation, the change in R^2 was significant across institutions and tests. The R^2 values were consistently higher for the multiple-choice Proficiency Profile questions than for the essay.

The Impact of the Motivational Instructions

Motivational instructions had a significant impact on SOS scores (Table 4). At all three institutions, students in the personal condition reported significantly higher levels of motivation than students in the control group, and the average difference was .31 *SD* between the control and institutional conditions and .43 *SD* between the control and the personal conditions. The largest difference was .57 *SD* between the control and personal conditions for students at the community college. No statistically significant differences were observed between the institutional and personal conditions across the three institutions.

Motivational condition also had a significant impact on the Proficiency Profile scores. Students in the personal group performed significantly and consistently better than those in the control group at all three institutions and the largest difference was .68 *SD*. The average performance difference was .26 *SD* between the control and institutional conditions and .41 *SD* between the control and the personal conditions. No statistically significant differences were observed between the institutional and personal conditions across the three institutions.

Similarly, students in the personal condition had consistently higher essay scores than students in the control condition across all three institutions. The largest effect size was .59 *SD*. Again, no statistically significant differences were observed between the institutional and personal conditions across the three institutions.

Results from the two-way ANOVAs showed that the interaction between institutions and motivational conditions was not statistically significant ($F = .51$, $df = 4$, $p = .73$ on mean SOS scores; $F = .86$, $df = 4$, $p = .49$ on Proficiency Profile scores; and $F = .83$, $df = 4$, $p = .51$ on essay scores). Given that the institutions did not interact with the conditions, we combined all students for additional analyses and included the results in Table 4. When all the students were included, the performance difference was .23 *SD* between the control and institutional conditions and .41 *SD* between the control and personal conditions.

Sophomore to Senior Learning Gain

A homoscedasticity test was provided to examine the homogeneity assumption of general linear regression. The Levene's test of equality of error variances was not significant ($F = 1.25$, $df_1 = 8$, $df_2 = 557$, $p = .27$ for the Proficiency Profile; and $F = 1.18$, $df_1 = 8$, $df_2 = 557$, $p = .31$ for the essay), which suggests that the data were suitable for this analysis. Table 5 presents the results from the GLM analyses. After controlling for SAT, motivation condition was a significant predictor for both tests ($p = .001$ for both). Class status was a significant predictor of the Proficiency Profile scores, but not significant for the essay. The interaction between motivation condition and class status was not significant for either test.

Table 3
Standardized Regression Coefficients With Self-reported Motivation and Standardized Test Scores Predicting Proficiency Profile and Essay Scores

| | Proficiency Profile | | | Essay | | |
|--------------------------------------|---------------------|----------|--------|----------|----------|--------|
| | RI | MI | CC | RI | MI | CC |
| Self-report motivation | .17*** | .26*** | .22** | .20*** | .26*** | .17* |
| SAT (or placement test) ^a | .68*** | .54*** | .50*** | .31*** | .32*** | .29** |
| ^b ΔR^2 | .03 | .06 | .05 | .04 | .04 | .04 |
| F(ΔR^2) | 15.87*** | 24.81*** | 6.36** | 13.57*** | 12.13*** | 6.05** |
| R ² | .53 | .42 | .31 | .16 | .13 | .11 |

Note. RI = research university; MI = master's university; CC = community college.

^aThe regression analysis was conducted separately for each institution by test. For both the research and master's institutions, composite SAT scores or converted ACT scores were used as a covariate. For the community college, composite placement test scores were used as a covariate.

^b ΔR^2 is the change in R² after the variable mean Student Opinion Survey was added to the regression equation.

* $p < .05$. ** $p < .01$. *** $p < .001$.

Table 4
Comparison by Motivational Condition and by Institution

| Self-Report Motivation Score | | | | | | | | | | | | | | |
|------------------------------|---------|------|------|-------------|------|------|----------|------|------|----------|----------|----------|-------|-------|
| | Control | | | Institution | | | Personal | | | d_{CI} | d_{CP} | d_{IP} | F | p |
| | n | M | SD | n | M | SD | n | M | SD | | | | | |
| RI | 111 | 3.65 | .59 | 116 | 3.80 | .59 | 113 | 3.88 | .64 | .25 | .37* | .13 | 4.43 | .010 |
| MI | 99 | 3.59 | .60 | 99 | 3.76 | .60 | 98 | 3.88 | .61 | .28 | .48** | .20 | 5.81 | .003 |
| CC | 40 | 3.57 | .69 | 42 | 3.93 | .65 | 36 | 3.95 | .65 | .54* | .57* | .03 | 4.06 | .02 |
| Total | 250 | 3.61 | .63 | 257 | 3.81 | .60 | 247 | 3.89 | .63 | .31** | .43*** | .14 | 13.68 | <.001 |

| Proficiency Profile Score | | | | | | | | | | | | | | |
|---------------------------|---------|-----|-------|-------------|-----|-------|----------|-----|-------|----------|----------|----------|-------|-------|
| | Control | | | Institution | | | Personal | | | d_{CI} | d_{CP} | d_{IP} | F | p |
| | n | M | SD | n | M | SD | n | M | SD | | | | | |
| RI | 111 | 453 | 18.13 | 116 | 460 | 20.66 | 113 | 461 | 21.79 | .37* | .40** | .04 | 5.37 | .005 |
| MI | 99 | 460 | 20.19 | 99 | 462 | 19.27 | 98 | 467 | 19.64 | .13 | .37* | .25 | 3.5 | .032 |
| CC | 40 | 435 | 20.74 | 42 | 443 | 18.48 | 36 | 450 | 21.08 | .37 | .68** | .35 | 4.79 | .010 |
| Total | 250 | 453 | 21.11 | 257 | 458 | 20.84 | 247 | 462 | 21.62 | .26* | .41*** | .16 | 11.19 | <.001 |

| Essay Score | | | | | | | | | | | | | | |
|-------------|---------|------|------|-------------|------|------|----------|------|------|----------|----------|----------|-------|-------|
| | Control | | | Institution | | | Personal | | | d_{CI} | d_{CP} | d_{IP} | F | p |
| | n | M | SD | n | M | SD | n | M | SD | | | | | |
| RI | 111 | 4.20 | .84 | 116 | 4.46 | .82 | 113 | 4.60 | .93 | .31 | .45* | .16 | 6.24 | .002 |
| MI | 99 | 4.19 | .88 | 99 | 4.30 | .93 | 98 | 4.53 | .83 | .12 | .39* | .26 | 3.73 | .025 |
| CC | 40 | 3.30 | 1.18 | 42 | 3.81 | .99 | 36 | 3.97 | 1.08 | .47 | .59* | .15 | 4.04 | .020 |
| Total | 250 | 4.07 | .96 | 257 | 4.29 | .93 | 247 | 4.46 | .95 | .23* | .41*** | .18 | 12.93 | <.001 |

Note. RI = research university; MI = master's university; CC = community college. d_{CI} = standardized mean difference (d) between the control and institutional conditions. d_{CP} = standardized mean difference (d) between the control and personal conditions. d_{IP} = standardized mean difference (d) between the Institutional and Personal conditions.

* $p < .05$. ** $p < .01$. *** $p < .001$.

Figures 1a and 1b illustrate the estimated Proficiency Profile and essay scores by motivational condition and class status (sophomores, juniors, seniors), after controlling for SAT scores. Within each class status group, students in the personal condition scored

highest on the Proficiency Profile and on the essay, followed by students in the institutional condition, with the control group showing the lowest performance. The only exception was the seniors in the institutional and control groups, who had equal

Table 5
Results From the General Linear Models

| Proficiency Profile | | | | | | |
|------------------------|-------------------------|-----|--------------|---------|-------|---------------------|
| Source | Type III Sum of Squares | df | Mean Square | F | p | Partial Eta Squared |
| Corrected model | 110,882.23 | 9 | 12,320.25 | 59.34 | <.001 | .49 |
| Intercept | 1,041,497.58 | 1 | 1,041,497.58 | 5016.10 | <.001 | .90 |
| SAT | 99,110.37 | 1 | 99,110.37 | 477.34 | <.001 | .46 |
| Condition ^a | 3,232.73 | 2 | 1,616.36 | 7.78 | <.001 | .03 |
| Class | 4,088.74 | 2 | 2,044.37 | 9.85 | <.001 | .03 |
| Condition × Class | 399.67 | 4 | 99.92 | .48 | .750 | .00 |
| Error | 115,442.80 | 556 | 207.63 | | | |
| Total | 121,140,988 | 566 | | | | |
| Corrected total | 226,325.04 | 565 | | | | |

| Essay | | | | | | |
|-------------------|-------------------------|-----|-------------|-------|-------|---------------------|
| Source | Type III Sum of Squares | df | Mean Square | F | p | Partial Eta Squared |
| Corrected model | 48.50 | 9 | 5.39 | 8.74 | <.001 | .12 |
| Intercept | 51.46 | 1 | 51.46 | 83.43 | <.001 | .13 |
| SAT | 32.40 | 1 | 32.40 | 52.54 | <.001 | .09 |
| Condition | 8.67 | 2 | 4.34 | 7.03 | <.001 | .02 |
| Class | 3.32 | 2 | 1.66 | 2.69 | .069 | .01 |
| Condition × Class | 2.88 | 4 | .72 | 1.17 | .324 | .01 |
| Error | 341.09 | 553 | .62 | | | |
| Total | 11,562.00 | 563 | | | | |
| Corrected total | 389.60 | 562 | | | | |

Note. R^2 was .49 for the Proficiency Profile and .13 for the essay.
^aIs the motivation condition.

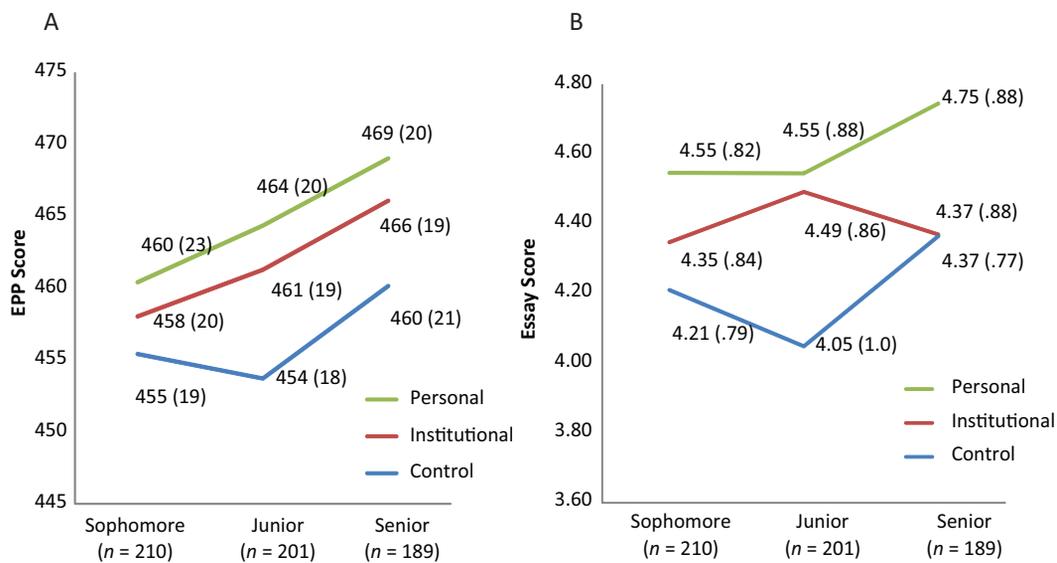


FIGURE 1. Proficiency Profile (EPP) and essay scores (and standard deviations) by condition and by class status, adjusted by college admission SAT/ACT scores.

essay scores. Although the interaction between class status and motivation condition was not statistically significant, there was a larger score difference between the personal and control groups for juniors and seniors than for sophomores on the Proficiency Profile (Figure 1a). On the essay (Figure 1b), the personal condition demonstrated a substantial impact across all classes as compared to the control group: .41 *SD* for sophomores, .53 *SD* for juniors, and .45 *SD* for seniors.

Based on the estimated means produced from the GLM analyses, sophomore to senior year score gain was calculated. The standardized mean differences were used as the effect size (Figures 2a and 2b). Within the same motivational condition (Figure 2a), the control group showed comparable learning gains on the Proficiency Profile and the essay (.25 vs. .23 in *SD*). However, the difference was striking for the institutional condition: While no learning gain (.02 *SD*) was observed on the essay, the gain was

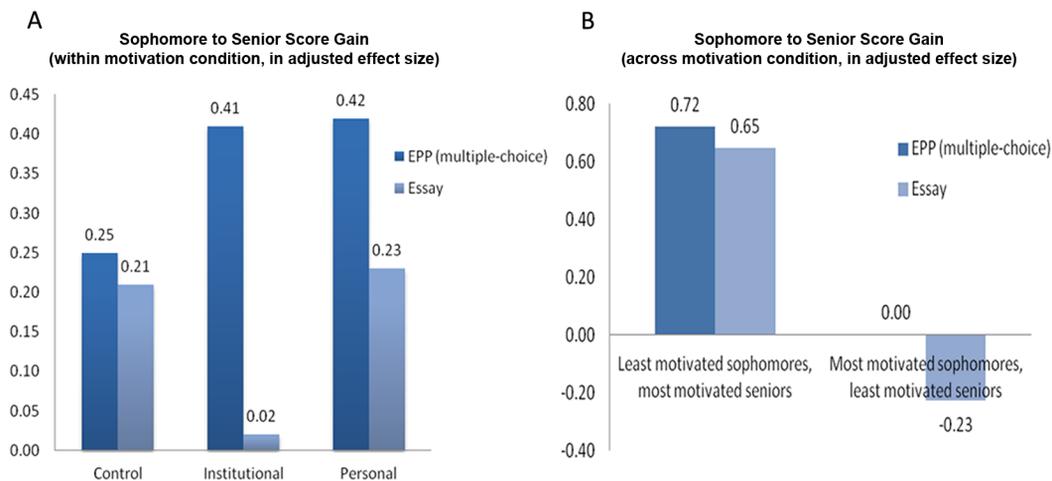


FIGURE 2. *Sophomore to senior score gain (value-added) in effect size adjusted for SAT scores, within and across motivation conditions. EPP = Proficiency Profile.*

substantial using the Proficiency Profile (.41 *SD*). The personal condition also showed a considerable difference in value-added learning between the multiple-choice and the essay tests: .23 *SD* on the essay and .42 *SD* on the Proficiency Profile.

In most value-added calculations, it is assumed that the levels of motivation remain somewhat equal between the benchmark class (e.g., freshmen or sophomores) and the comparison class (e.g., juniors or seniors). However, students in lower classes may be more motivated than their upper-class peers for multiple reasons, such as still being intimidated by tests or being less busy. Here we illustrated two extreme cases where least motivated sophomores and most motivated seniors were compared, and vice versa. Substantial gains on both the Proficiency Profile (.72 *SD*) and the essay (.65 *SD*) were observed when groups of least motivated sophomores and most motivated seniors were tested (Figure 2b). However, little or even negative gain (-.23 *SD*) was observed when groups of most motivated sophomores and least motivated seniors were considered.

Conclusions

We draw three conclusions from this random assignment experiment. First, self-report motivation has a significant and consistent relationship with test scores, for both multiple-choice and essay tests, even after controlling for college admission scores or placement test scores. Second, manipulation of motivation could significantly enhance student motivation in taking low-stakes outcomes assessments and in turn increase their test scores on both multiple-choice and essay tests. The results also confirmed researchers' concern (e.g., Banta, 2008; Liu, 2011a) that students do not exert their best effort in taking low-stakes outcomes assessments. Students in the two treatment conditions performed significantly better than students in the control condition. Between the two treatment conditions, there was no statistically significant performance difference, but students in the personal condition showed a small advantage as compared to the students in the institutional condition ($d = .16$ for the Proficiency Profile and $d = .18$ for the essay). Last, when using outcomes assessment scores to determine institutional value-added gains, one has to take into

consideration students' levels of motivation in taking the assessment and the format of the assessment instrument (i.e., multiple choice or constructed response). As shown in this study, conclusions about value-added learning changed dramatically depending on the test of choice and the motivation levels. These findings are fairly consistent with findings from previous studies using course-based assessments (e.g., Sundre, 1999; Sundre & Kitsantas, 2004; Wolf & Smith, 1995). To summarize, motivation plays a significant role in low-stakes outcomes assessment. Ignoring the effect of motivation could seriously threaten the validity of the test scores and make any decisions based on the test scores questionable.

Although previous studies (e.g., Duckworth et al., 2011) have demonstrated the value of monetary incentives, such incentives are not a practical alternative for most institutional testing programs given the fiscal challenges institutions currently face. This study demonstrated that once institutions recruit students to take the test, they can use motivational strategies that do not involve extra financial costs to produce significant effects on student performance.

One potential limitation of this study is that the administration of the multiple-choice and essay tests was not counterbalanced due to logistic complications with the random assignment within a testing session. All students took the multiple-choice test first, which may have impacted their overall motivation in taking the following essay test. However, our results showed that students' self-report motivation predicted both tests to about the same degree (Tables 2 and 3), and the effect of the motivational instructions was comparable on the two tests (Table 4), which suggests that the impact of the order of the test administration was probably minimal. A potential explanation is that both the multiple-choice and the essay test were pretty short (40 and 30 minutes) and therefore students were not exhausted by the end of the first test.

Implications

Implications for Researchers, Administrators, and Policymakers. Findings from this study have significant implications for

higher education stakeholders at many levels. For educational researchers, the limited college learning reported from prior research is likely an underestimate of true student learning due to students' lack of motivation in taking low-stakes tests. The book *Academically Adrift* (Arum & Roksa, 2011) surprised the nation by reporting that overall, students demonstrated only minimal learning on college campuses (.18 *SD*), and at least 45% of the students did not make any statistically significant gains. They concluded that "in terms of general analytical competencies assessed, large numbers of U.S. college students can be accurately described as academically adrift" (p. 121). The Arum and Roksa study analyzed the performance of a group of students when entering their freshman year and at the end of their sophomore year using the CLA, a constructed-response test.

We want to bring it to the readers' attention that the limited learning gain reported in the Arum and Roksa (2011) study (.18 *SD*) is very similar to the small learning gain (.23 *SD*, Figure 2a) observed in this study for students in the control group on the essay. However, we've shown in this study that with higher levels of motivation, students can significantly improve their test performance and demonstrate a much larger learning gain (Figure 2a). In addition, conclusions about college learning can also change with the test of choice. Findings from this study show that more learning gain was consistently observed on the multiple-choice test than on the essay test (Figures 2a and 2b). The reason could be that it takes more effort and motivation for students to construct an essay than to select from provided choices. Figure 1b shows that the institutional condition was not able to motivate the seniors on the essay test. It may take a stronger reason than caring for one's institutional reputation for seniors to be serious about writing an essay.

In sum, for both multiple-choice and constructed-response tests, students' performance motivation could dramatically change the conclusions we make about college learning. The limited college learning as reported in the Arum and Roksa (2011) study, as well as that found in this study for the students in the control condition, is likely an underestimation of students' true college learning. It is dangerous to make conclusions about the quality of U.S. higher education based on learning outcomes assessment data without considering the role of motivation.

For institutions, this study provides credible evidence that motivation has a significant impact on test scores. Without motivational manipulation, the performance difference between sophomores and seniors was 5 points (Figure 1a, control condition). With motivational manipulation, sophomores were able to gain 5 points in the personal condition, which suggests that the motivational effect for sophomores was as large as 2 years of college education. When administering outcomes tests, institutions should employ effective strategies to enhance student motivation so that students' abilities will not be underestimated by the low-stakes tests. Although we paid students \$50 to take the test in the study, the motivational instructions used to boost student performance did not involve any additional payment. Institutions can use other incentives (e.g., offering extra credits) to recruit students to take the tests and use practical strategies to motivate them, such as stressing the importance of the test results to the institution and emphasizing potential consequences of the results

to individual students. This way, students' scores are likely to be improved at no extra financial cost to the institutions.

An important message to policymakers is that institutions that employ different motivational strategies in testing the students should be compared with great caution, especially when the comparison is for accountability purposes. Accountability initiatives involving outcomes assessment should also take into account the effect of motivation when making decisions about an institution's instructional effectiveness. Institutions doing a good job of motivating students could achieve significantly higher rankings than institutions doing a poor job of motivating students, even though their students may have comparable academic abilities. Figure 2b illustrates how significant the effect of motivation could be: If we compare the most motivated (personal condition) sophomores to the least motivated (control condition) seniors on the Proficiency Profile, we would come to the conclusion that students did not learn anything during the 2 years time. However, if we compare the least motivated sophomores with the most motivated seniors also on the Proficiency Profile, we would come to a radically different conclusion, that students gained substantial knowledge (0.72 *SD*). The difference is starker on the essay. A comparison of the most motivated sophomores with the least motivated seniors leads to the conclusion that not only did students not make any progress, but that they were even set back by a college education as indicated by the negative gain score (-0.23 *SD*).

The importance of the findings extends well beyond the United States as outcomes assessment is being used in international studies assessing college learning across multiple countries. For example, the Assessment of Higher Education Learning Outcomes (AHELO) project sponsored by the Organization of Economic and Cooperation Development (OECD) tests what college graduates know and can do in general skills such as critical thinking, writing, and problem solving and has attracted participation from 17 countries. Although AHELO does not endorse ranking, the higher education systems of the participating countries will likely be compared once the data are available. Differential motivation across countries is likely to significantly impact how U.S. students stand relative to their international peers (Barry, Horst, Finney, Brown, & Kopp, 2010; S. L. Wise & DeMars, 2010). As S. L. Wise and DeMars (2010) noted, results from international comparative studies such as PISA may be questionable as the level of mean student motivation may vary across countries. In fact, differential motivation between freshmen and sophomores, in addition to the low motivation in general, was likely the key factor responsible for the limited learning reported in the Arum and Roksa study (2011).

A Cautionary Note. We wanted to make a cautionary note that college learning outcomes are much broader than what's captured by learning outcomes assessments. College learning covers learning in disciplinary subjects, interdisciplinary domains, general skills, and in many other aspects. Although students' scores on outcomes assessments are in general valid predictors of their course work preparation (Hendel, 1991; Lakin et al., in press; Marr, 1995), they only reflect a fraction of what students know and can do. Generalizing outcomes scores to college learning or even to the quality of higher education is questionable. In

addition, sampling issues could further thwart the validity of the conclusion about an institution's instructional quality using outcomes assessment (Liu, 2011a).

In addition, although research has been conducted concerning other aspects of validity for outcomes assessment, little is known about its consequential validity (Messick, 1995), in this case, whether outcomes assessment can assist administrators better prepare students for performance in the workforce. The follow-up study to Arum and Roksa's (2011) study found that graduates scoring in the bottom quintile are more likely to be unemployed, living at home, and having amassed credit card debt (Arum, Cho, Kim, & Roksa, 2012). However, graduates in the top quintile were only making \$97 more than those in the bottom quintile (\$35,097 vs. \$35,000), and graduates in the middle three quintiles were making even less than the bottom quintile cohort (\$34,741). The consequential validity of learning outcomes assessments awaits further confirmation.

Next Steps

In future research, efforts should be made to identify effective and robust strategies that institutions can adopt to boost student motivation in taking low-stakes tests. We are particularly interested in further exploring the function of the institutional condition used in this study. Although not producing effects as large as the personal condition, in general this condition was effective in motivating students. In addition, as what is said about the personal condition (that students' scores will be used by potential employers to evaluate their academic ability) may not be true, what is described for the institutional condition is often true given many institutions do rely on outcomes learning data for improvement and accountability purposes. This strategy can be easily customized or even enhanced by individual institutions. For instance, instead of including it in the consent form, institutions can train proctors to motivate students with a short speech emphasizing the importance of the test scores to their institution and the relevance of the test results to students.

The reason underlying the effect of the personal condition lies in the relevance of the test scores to students. A possible solution along the same line is for the test sponsors to provide a certificate to students attesting to their performance. Students then can choose to present the certificate to potential employers in evaluating their academic ability. With a certificate, results from learning outcomes assessment are not only important for institutions, but are meaningful for students as well.

In this study, although we are able to observe consistent motivation effects across the participating institutions, only three institutions were included. It is important to see whether the findings from this study can be replicated with more institutions. Knowledge about effective and practical strategies that institutions can use to enhance student motivation will greatly help improve the validity of outcomes assessment and largely contribute to the evidence-based, data-driven, and criterion-referenced evaluation system that U.S. higher education is currently developing.

NOTE

¹Formerly known as the Measure of Academic Proficiency and Profile (MAPP).

REFERENCES

- Arum, R., Cho, E., Kim, J., & Roksa, J. (2012). *Documenting uncertain times: Post-graduate transitions of the academically adrift cohort*. Brooklyn, NY: Social Science Research Council.
- Arum, R., & Roksa, J. (2011). *Academically adrift: Limited learning on college campuses*. Chicago, IL: University of Chicago Press.
- Astin, A. W. (2011, February 14). In "Academically Adrift," data don't back up sweeping claim. *The Chronicle of Higher Education*. Retrieved from <http://chronicle.com/article/Academically-Adrift-a/126371>
- Banta, T. (2008). Trying to clothe the emperor. *Assessment Update*, 20, 3-4, 16-17.
- Barry, C. L., Horst, S. J., Finney, S. J., Brown, A. R., & Kopp, J. (2010). Do examinees have similar test-taking effort? A high-stakes question for low-stakes testing. *International Journal of Testing*, 10(4), 342-363.
- Baumert, J., & Demmrich, A. (2001). Test motivation in the assessment of student skills: The effects of incentives on motivation and performance. *European Journal of Psychology of Education*, 16, 441-462.
- Belcheir, M. J. (2002). *Academic profile results for selected nursing students* (Report No. 2002-05). Boise, ID: Boise State University.
- Braun, H., Kirsch, I., & Yamamoto, K. (2011). An experimental study of the effects of monetary incentives on performance on the 12th grade NAEP reading assessment. *Teachers College Record*, 113, 2309-2344.
- Cole, J. S., & Osterlind, S. J. (2008). Investigating differences between low- and high-stakes test performance on a general education exam. *The Journal of General Education*, 57, 119-130.
- Duckworth, A. L., Quinn, P. D., Lynam, D. R., Loeber, R., & Stouthamer-Loeber, M. (2011). Role of test motivation in intelligence testing. *Proceedings of the National Academy of Sciences*, 108, 7716-7720.
- Educational Testing Service. (2010). *Market research of institutions that use outcomes assessment*. Princeton, NJ: Author.
- Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice*, 23, 17-27.
- Hendel, D. D. (1991). Evidence of convergent and discriminant validity in three measures of college outcomes. *Educational and Psychological Measurement*, 51, 351-358.
- Klein, S., Liu, O. L., Sconing, J., Bolus, R., Bridgeman, B., Kugelmass, . . . Steedle, J. (2009). *Test validity study report*. Retrieved from http://www.voluntarysystem.org/docs/reports/TVSReport_Final.pdf
- Kuh, G. D., & Ikenberry, S. O. (2009). *More than you think, less than we need: Learning outcomes assessment in American higher education*. Urbana, IL: University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment.
- Kuh, G. D., Kinzie, J., Buckley, J. A., Bridges, B. K., & Hayek, J. C. (2006). *What matters to student success: A review of the literature* (Report commissioned for the National Symposium on Postsecondary Student Success: Spearheading a Dialog on Student Success). Washington, DC: National Postsecondary Education Cooperative.
- Lakin, J., Elliott, D., & Liu, O. L. (in press). Investigating the impact of ELL status on higher education outcomes assessment. *Educational and Psychological Measurement*.
- Liu, O. L. (2008). *Measuring learning outcomes in higher education using the Measure of Academic Proficiency and Progress (MAPPTM)* (ETS Research Report Series RR-08-047). Princeton, NJ: Educational Testing Service.
- Liu, O. L. (2011a). An overview of outcomes assessment in higher education. *Educational Measurement: Issues and Practice*, 30, 2-9.
- Liu, O. L. (2011b). Value-added assessment in higher education: A comparison of two methods. *Higher Education*, 61, 445-461.

- Livingston, S. A., & Antal, J. (2010). A case of inconsistent equatings: How the man with four watches decides what time it is. *Applied Measurement in Education*, 23(1), 49-62.
- Marr, D. (1995). *Validity of the academic profile*. Princeton, NJ: Educational Testing Service.
- Messick, S. (1995). Validity of psychological assessment: Validation of references from persons' responses and performances on scientific inquiry into score meaning. *American Psychologist*, 50, 741-749.
- Ochoa, E. M. (2011, March). Higher education and accreditation: The view from the Obama administration. *Career Education Review*. Retrieved from http://www.careereducationreview.net/featured-articles/docs/2011/CareerEducationReview_Ochoa0311.pdf
- O'Neil, H. F., Abedi, J., Miyoshi, J., & Mastergeorge, A. (2005). Monetary incentives for low-stakes tests. *Educational Assessment*, 10, 185-208.
- O'Neil, H. F., Sugrue, B., & Baker, E. L. (1995/1996). Effects of motivational interventions on the National Assessment of Educational Progress mathematics performance. *Educational Assessment*, 3, 135-157.
- Pintrich, P. R., & Schunk, D. H. (2002). *Motivation in education: Theory, research, and applications* (2nd ed.). Upper Saddle River, NJ: Prentice Hall.
- Steedle, J. (2011). Selecting value-added models for postsecondary institutional assessment. *Assessment and Evaluation in Higher Education*, 1-16.
- Sundre, D. L. (1997, April). *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, April). *Does examinee motivation moderate the relationship between test consequences and test performance?* Paper presented at the annual meeting of the American Educational Research Association, Montreal.
- Sundre, D. L., & Kitsantas, A. L. (2004). An exploration of the psychology of the examinee: Can examinee self-regulation and test-taking motivation predict consequential and non-consequential test performance? *Contemporary Educational Psychology*, 29(1), 6-26.
- Sundre, D. L., & 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.
- U.S. Department of Education. (2006). *A test of leadership: Charting the future of American higher education* (Report of the commission appointed by Secretary of Education Margaret Spellings). Washington, DC: Author.
- Voluntary System of Accountability. (2008). *Information on learning outcomes measures*. Author.
- Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment*, 10(1), 1-17.
- Wise, S. L., & DeMars, C. E. (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement*, 43(1), 19-38.
- Wise, S. L., & DeMars, C. E. (2010). Examinee noneffort and the validity of program assessment results. *Educational Assessment*, 15, 27-41.
- Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, 18(2), 163-183.
- Wise, V. L. (2004). *The effects of the promise of test feedback on examinee performance and motivation under low-stakes testing conditions* (Unpublished doctoral dissertation). University of Nebraska-Lincoln, Lincoln, NE.
- Wise, V. L., Wise, S. L., & Bhola, D. S. (2006). The generalizability of motivation filtering in improving test score validity. *Educational Assessment*, 11(1), 65-83.
- Wolf, L. F., & Smith, J. K. (1995). The consequence of consequence: Motivation, anxiety, and test performance. *Applied Measurement in Education*, 8, 227-242.

AUTHORS

OU LYDIA LIU is a senior research scientist at ETS, 660 Rosedale Road, Princeton, NJ 08540; lliu@ets.org. Her research focuses on learning outcomes assessment in higher education and innovative science assessment.

BRENT BRIDGEMAN is a distinguished presidential appointee at Educational Testing Service, 660 Rosedale Rd., Princeton, NJ 08540; bbridgeman@ets.org. His research focuses on validity research, in particular threats to score interpretations from construct irrelevant variance.

RACHEL M. ADLER is a research assistant at ETS, 660 Rosedale Road, Mailstop 9R, Princeton, NJ 08541; radler@ets.org. Her research focuses on validity issues related to assessments for higher education and English Language Learners.

Manuscript received April 12, 2012

Revisions received June 1, 2012, and July 23, 2012

Accepted July 24, 2012