CAE

Test Administration Procedures and Their Relationships with Effort and Performance on a College Outcomes Test

April 6, 2014

Jeffrey T. Steedle
Doris Zahner
Heather Kugelmass

Paper presented at the Annual Meeting of the American Educational Research Association
Philadelphia, Pennsylvania

AUTHOR NOTE

Currently, Jeffrey T. Steedle is a Research Scientist with Pearson. Doris Zahner is the Director of Test Development and a Measurement Scientist at the Council for Aid to Education. Heather Kugelmass is a Ph.D. candidate in the Department of Sociology at Princeton University.

Correspondence concerning this article should be addressed to Jeffrey T. Steedle, Pearson, 400 Center Ridge Drive, Austin, TX 78753. Contact: jeffrey.steedle@pearson.com

ABSTRACT

Poor motivation is commonly cited as a concern when interpreting results from low-stakes standardized tests administered to postsecondary students. This study investigates the associations between test administration procedures and students' self-reported effort and performance on the Collegiate Learning Assessment (CLA), an open-ended test of college students' critical-thinking and writing skills. Coefficient estimates from a series of hierarchical linear models revealed that paying students to take tests and offering performance-based incentives were positively associated with effort and performance. Mandatory testing, however, was negatively associated with effort and performance. Faculty involvement in recruiting, giving extra course credit, and offering prize raffle entries were not associated with effort or performance. Effort appeared to mediate the relationship between some test administration variables (e.g., payment and mandatory testing) and performance.

Keywords: higher education, institutional assessment, motivation

TEST ADMINISTRATION PROCEDURES AND THEIR RELATIONSHIPS WITH EFFORT AND PERFORMANCE ON A COLLEGE OUTCOMES TEST

Motivation to perform well on a test is a potentially problematic source of construct-irrelevant variance (Messick, 1995), especially when there are no stakes attached to performance. That is, interpretations of test scores as indicators of examinees' knowledge and skills may be compromised if those examinees do not put forth the effort necessary to demonstrate the full extent of their abilities. Low-stakes testing is commonplace in K-12 education in the United States, and concerns over suspect motivation are exacerbated when aggregate test results are used to measure teacher or school effectiveness (e.g., Guerriero, 2013). In postsecondary education, tests such as CAE's Collegiate Learning Assessment (CLA), ACT's Collegiate Assessment of Academic Proficiency, and the ETS Proficiency Profile are typically administered under low-stakes conditions to provide evidence of student learning for stakeholders, prospective students, and accreditors.

Poor motivation is commonly raised in critiques of postsecondary institutional assessment programs (Banta, 2008), and some colleges cite this concern as a reason for not using assessments such as the CLA. Statistical approaches to adjusting results for low motivation have been proposed (e.g., Sundre & Wise, 2003), but it would be better if no such adjustments were necessary. In addition to psychological variables, such as achievement goals and personality (Barry, Horst, Finney, Brown, & Kopp, 2010) and the format of the test (Sundre, 1999; Wise, 2006), test administration procedures potentially impact student motivation on tests. Indeed, inconsistencies in methods of recruiting and incentivizing examinees have been suggested as causes of variation in results across administrations (Hosch, 2010), and rigorous proctor training has been shown to increase self-reported effort (Lau, Swerdzewski, & Jones, 2009).

This study investigates test administration procedures and their possible associations with effort and performance on the CLA, a test of college students' critical-thinking and writing skills. In 2008, a post-administration survey was delivered to administrators at all participating institutions. This survey asked whether CLA testing was mandatory, whether faculty were involved in recruiting students, and how students were incentivized. In this study, a series of hierarchical linear models were employed to determine whether test administration procedures were significantly associated with test performance and self-reported effort after controlling for prior ability. Results could inform best practices for recruiting and incentivizing students to maximize motivation, thereby improving the validity of inferences from results of low-stakes college outcomes tests.

LITERATURE REVIEW

Prior research has shown that attaching consequences to test performance can have large effects on motivation and test performance (Liu, Bridgeman, & Adler, 2012; Napoli & Raymond, 2004; Wise & DeMars, 2005; Wolf & Smith, 1995). Tests like the CLA, however, are almost exclusively administered in low-stakes contexts, so test administrators must rely on alternate means of fostering motivation. For example, it has been recommended to stress the importance and usefulness of a test because those feelings are correlated with effort, which partly mediates test performance (Cole, Bergin, & Whittaker, 2008). Alternatively, instilling a sense of competition with rival schools may have some effect on test performance (Bracey, 1996).

Uniformly high motivation has been reported in some international testing contexts (Baumert & Demmrich, 2001; Eklöf, 2007), with examinees reporting social responsibility, competitive spirit, interest, and personal or intrinsic motivators as the reasons for their motivation (Eklöf, 2008). Such findings sharply contrast with U.S. students taking low-stakes tests like the National Assessment of Educational Progress, where a large percentage of students report a low sense of importance and low effort (ETS, 1993).

Although monetary incentives have been effective for improving test scores in some content areas (Bettinger, 2010), performance-contingent financial rewards appear to be more effective

than a reward for simple test completion (Braun, Kirsch, & Yamamoto, 2011). In fact, bigger financial incentives appear to be more effective than smaller financial incentives on math and reading tests, but only if given at the time of testing (Levitt, List, Neckerman, & Sadoff, 2011). Even offering additional entries in a prize lottery for high math test scores has been effective in increasing self-reported effort and performance (Cole, 2007). The effectiveness of performance-contingent rewards, however, is not universal. In a series of studies, neither small nor large incentives improved 12th graders' math test performance (O'Neil, Abedi, Miyoshi, & Mastergeorge, 2005; O'Neil, Sugrue, & Baker, 1995/1996).

This study supplements prior research by examining the associations between monetary incentives (payment, performance-based incentives, and raffle entries), self-reported effort, and performance on a low-stakes college outcomes test. Additionally, this study included several variables not previously studied: the involvement of faculty in student recruitment, mandatory (versus voluntary) testing, and offering course extra credit. Separate analyses were conducted for freshmen and seniors due to different administration procedures for those two groups. Moreover, self-reported effort among older students may be different from younger students (Kiplinger & Linn, 1995/1996; Liu et al., 2012).

METHOD

Sample

The sample included 5,428 entering freshmen and 4,611 graduating seniors at 102 four-year colleges and universities (24 research universities, 47 master's colleges and universities, 30 baccalaureate colleges, and one seminary). Of those schools, 41 had a total enrollment greater than 10,000 students, 58 were public institutions, and four were classified as Historically Black Colleges and Universities. In terms of geographic distribution, 15 were located in the mid-Atlantic or New England regions, 15 in the Great Lakes or plains, 38 in the southeast, and 34 in the west or southwest. They had admissions rates ranging from 18% to 100% (median 66%), six-year graduation rates ranging from 19% to 92% (median 55%), and a percentage of White students ranging from 5% to 95% (median 68%). Because the statistical analyses controlled for students' prior academic abilities, only students with SAT or ACT scores on record were included in the data set.

Measures

CLA Performance Task (PT). Students had 90 minutes to analyze a set of documents representing a real-world problem and answer a series of essay questions asking them to analyze the provided information and then propose a solution. Students were randomly assigned one of many possible PTs. A group of trained scorers evaluated the responses using scales that described the quality of analysis, problem solving, and writing effectiveness and mechanics. Inter-rater correlations on PT total scores were typically around .85.

Self-reported effort. Following the PT, students completed a brief questionnaire, including an item that asked students how much effort they put into the test. The response options included "Made little effort," "Made some effort," "Mainly did my best," and "Tried my best."

Prior ability. As part of the regular CLA administration, SAT and ACT scores were collected from college registrars' offices. ACT total scores were converted to the SAT score scale using a concordance table (ACT, 2008).

Post-administration survey. Completed surveys were received from 102 (52%) of the 195 institutions participating in the 2007-2008 CLA administration. Relevant to this analysis, the survey asked whether students were required to take the CLA, whether faculty participated in CLA outreach to students, and how participants were incentivized. The incentives included in this analysis were payment (money, gift certificate, or university cash), performance-based incentives (e.g., the top five scorers receive $100), raffle entry, and course extra credit. Other incentives were excluded from analyses because they were employed by fewer than 10% of institutions (food, priority course registration, or none).

Analysis

A series of five HLMs were fit to the data. The unconditional model was first fit to estimate the variance in PT scores that is between and within schools.

$$\text{Level 1: } PT_{ij} = \beta_{0j} + r_{ij}$$

$$\text{Level 2: } \beta_{0j} = \gamma_{00} + u_{0j}$$

The model shown below treats students (level 1) as nested within institutions (level 2). At the student level (level 1), PTij is the PT score of student i at school j, and β_{0j} is the mean PT score at school j. At the school level, β_{0j} is modeled as the grand mean PT score plus the school-level residual, u_{0j}.

Next, a conditional model including only grand-mean centered SAT scores at level 1 was fit to see how much of the between-school variance was accounted for by SAT scores.

$$\text{Level 1: } PT_{ij} = \beta_{0j} + \beta_{1j}(SAT_{ij} - \overline{SAT}) + r_{ij}$$

$$\text{Level 2: } \beta_{0j} = \gamma_{00} + u_{0j}$$

$$\beta_{1j} = \gamma_{10}$$

SATij is the SAT score of student i at school j. Note that grand-mean centering has the effect of making β_{0j} equal to school j's PT mean adjusted for SAT scores.

Model 3 added test administration dummy variables at the school level (level 2).

$$\text{Level 1: } EFF_{ij} = \beta_{0j} + \beta_{1j}(SAT_{ij} - \overline{SAT}) + r_{ij}$$

Level 2:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}MAN_j + \gamma_{02}FAC_j + \gamma_{03}PAY_j + \gamma_{04}PER_j + \gamma_{05}RAF_j + \gamma_{06}CRE_j + u_{0j}$$

$$\beta_{1j} = \gamma_{10}$$

Here, adjusted mean PT scores (β_{0j}) are modeled as a function of six dummy variables indicating whether testing was mandatory (MAN_j), faculty were involved in recruitment (FAC_j), students were paid (PAY_j), performance-based incentives were employed (PER_j), raffle entries were offered (RAF_j), and course extra credit was offered (CRE_j) at school j.

Model 4 is identical to model 3 except that is adds self-reported effort EFF_{ij} at the student level.

$$\text{Level 1: } PT_{ij} = \beta_{0j} + \beta_{1j}(SAT_{ij} - \overline{SAT}) + \beta_{2j}(EFF_{ij} - \overline{EFF}) + r_{ij}$$

Level 2:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}MAN_j + \gamma_{02}FAC_j + \gamma_{03}PAY_j + \gamma_{04}PER_j + \gamma_{05}RAF_j + \gamma_{06}CRE_j + u_{0j}$$

$$\beta_{1j} = \gamma_{10}$$

$$\beta_{2j} = \gamma_{20}$$

Grand-mean centering on SAT scores and effort has the effect of making β_{0j} equal the mean PT score at school j adjusted for SAT scores and effort.

As demonstrated in previous research, effort may mediate the relationship between administrative procedures and CLA performance (Cole, Bergin, and Whittaker, 2010). When that is the case, one should expect coefficients for dummy variables to be significant in Model 3 but not when self-reported effort is included (Model 4). Model 5 used self-reported effort as the outcome to directly investigate the associations between administrative procedures and self-reported effort.

Level 1: $PT_{ij} = \beta_{0j} + \beta_{1j}(SAT_{ij} - \overline{SAT}) + r_{ij}$

Level 2:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}MAN_j + \gamma_{02}FAC_j + \gamma_{03}PAY_j + \gamma_{04}PER_j + \gamma_{05}RAF_j + \gamma_{06}CRE_j + u_{0j}$$

$$\beta_{1j} = \gamma_{10}$$

RESULTS

Table 1 shows the percentages of schools that employed the administration procedures examined in this study. Note that schools employed these procedures more frequently with seniors than with freshmen. This finding is consistent with common anecdotal reports from schools suggesting that seniors are more challenging to recruit than freshmen.

Table 1
Percentages of schools using administration procedures

	Freshmen	Seniors
Mandatory testing (MAN)	18%	33%
Faculty involvement (FAC)	60%	68%
Payment (PAY)	50%	70%
Performance-based incentives (PER)	15%	19%
Raffle (RAF)	24%	30%
Course extra credit (CRE)	21%	21%

Table 2 provides descriptive statistics on the outcome measures. As would be expected due to learning in college, the seniors scored higher on the CLA, but some of this difference can be accounted for by differences in ability, which were apparent from the mean difference in SAT scores. On average, freshmen and seniors reported similar levels of effort, with more than 70% of students reporting that they mainly did their best or tried their best. The correlation between self-reported effort and CLA performance was .24 ($p < .001$) in both samples, and this result is fairly typical (Steedle, 2014).

Table 2
Sample demographics and descriptive statistics

	Freshmen	Seniors
Female	60%	60%
White	72%	74%
English spoken at home	91%	91%
Mean SAT (400-1600 scale)	1086	1117
Mean self-reported effort (1-4 scale)	2.9	3.0
Mean CLA PT	1080	1194
Made little effort	4%	4%
Made some effort	25%	23%
Mainly did my best	44%	44%
Tried my best	27%	29%

Results from the analysis of freshmen are shown in Table 3. Models 1 and 2 indicated that 25% of the variance in CLA scores was between schools, but, after controlling for SAT scores, 7% was between schools. In Model 3, only the coefficient for payment had a significant positive coefficient ($p < .05$), but the negative coefficient for mandatory testing was nearly significant ($p = .07$). When self-reported effort was added in Model 4, only mandatory testing had a nearly significant coefficient ($p = .09$). Significant coefficients for payment in Models 3 and 5 but not Model 4 suggested effort as a mediator between paying students and test performance. Model 5 also had a significant positive coefficient for performance-based incentives ($p < .01$).

In the senior analysis, 20% of the CLA score variance was between schools, and 5% was between schools after controlling for SAT scores (Table 4). In Model 3, there was a significant negative coefficient for mandatory testing ($p < .05$) and a significant positive coefficient for performance-based incentives ($p < .001$). Only performance-based incentives had a significant coefficient ($p < .001$) in Model 4. Model 5 had significant coefficients for mandatory testing ($p < .01$), payment ($p < .01$), and performance-based incentives ($p < .05$). The comparison of Models 3, 4, and 5 reveals the possible mediating effect of effort between mandatory testing and performance.

Table 3
Coefficients from models fit to entering freshman data

		Model 1	Model 2		Model 3		Model 4		Model 5 (Outcome variable: EFF)	
		Coef.	Coef.	Std. err.	Coef.	Std. err.	Coef.	Std. err.	Coef.	Std. err.
Student level										
SAT	γ_{10}		0.57***	0.01	0.57***	0.01	0.57***	0.01	0.00009	0.00007
EFF	γ_{20}						42.82***	2.59		
School level										
Intercept	γ_{00}		456.5***	16.07	1069.8***	10.79	948.3***	12.29	2.83***	0.06
MAN	γ_{01}				-23.01	12.40	-19.44	11.36	-0.08	0.06
FAC	γ_{02}				-5.27	10.32	-3.31	9.43	-0.04	0.05
PAY	γ_{03}				20.89*	10.44	11.89	9.56	0.22***	0.05
PER	γ_{04}				18.18	13.00	10.10	11.87	0.19**	0.07
RAF	γ_{05}				12.88	11.27	12.60	10.33	0.00	0.06
CRE	γ_{06}				-17.06	11.86	-14.24	10.87	-0.07	0.06
Variance										
Student level	σ^2	29328	23800		23803		22714		0.62403	
School level	τ_{00}	9859	1854		1462		1168		0.03912	

*** p < .001, ** p < .01, * p < .05

Table 4
Coefficients from models fit to graduating senior data

		Model 1	Model 2		Model 3		Model 4		Model 5 (Outcome variable: EFF)	
		Coef.	Coef.	Std. err.	Coef.	Std. err.	Coef.	Std. err.	Coef.	Std. err.
Student level										
SAT	γ_{10}		0.56***	0.02	0.55***	0.02	0.54***	0.02	0.00035***	0.00008
EFF	γ_{20}						43.06***	3.03		
School level										
Intercept	γ_{00}		561.9***	18.28	1183.2***	12.21	1057.6***	14.35	2.91***	0.07
MAN	γ_{01}				-20.06*	9.87	-11.90	9.18	-0.18**	0.06
FAC	γ_{02}				-11.05	9.82	-9.21	9.09	-0.04	0.06
PAY	γ_{03}				10.11	10.45	3.64	9.71	0.16**	0.06
PER	γ_{04}				41.03***	10.63	33.69***	9.84	0.16*	0.06
RAF	γ_{05}				6.84	9.69	8.12	9.00	-0.03	0.06
CRE	γ_{06}				-3.47	11.49	-3.95	10.69	0.01	0.07
Variance										
Student level	σ^2	32371	26880		26907		25842		0.60117	
School level	τ_{00}	8236	1600		1036		824		0.0453	

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

DISCUSSION

This study investigated associations between test administration procedures, self-reported effort, and performance for a low-stakes test of college outcomes. The analysis of senior data revealed a significant negative association between requiring students to take the CLA and performance after controlling for other variables (nearly significant for freshmen). That is, results are consistent with the notion that mandating testing, rather than soliciting volunteers, can negatively impact student effort and performance.

Some have suggested that institutional assessment programs would benefit from faculty buy-in and involvement (Steedle, 2010). In this study, faculty involvement in recruitment was not significantly associated with effort or performance, after controlling for other variables. However, the extent of faculty involvement was unknown.

Of the incentives, payment for task completion (money, gift certificate, or university cash) and performance-based incentives (e.g., the top five scorers receive $100) showed positive associations with effort and performance, after controlling for other variables. Raffle entries and course extra credit did not. Performance-based incentives for freshmen and payment for seniors were correlated with self-reported effort but not test performance. In some cases, there was evidence that effort acted as the mediator between administration procedures and performance. For example, results are consistent with the notion that mandatory testing of senior students reduced effort, thereby reducing performance. In a similar fashion, higher performance among freshmen who were paid can be accounted for by the association between payment and effort.

The practices of paying students to participate and offering performance-based incentives appear to support testing effort and performance, thereby strengthening the validity of test-score interpretations on low-stakes tests of college outcomes. In contrast, the practice of mandating testing (e.g., in randomly selected freshman seminars or senior capstone courses) appears to depress testing effort and performance, which could negatively impact validity. This conclusion could be subjected to experimentation in future studies. Until then, it is recommended that students be solicited to volunteer for testing and offered financial incentives, while making every effort to ensure that the tested sample is representative of the larger student body.

REFERENCES

ACT. (2008). ACT-SAT Concordance. Retrieved July 21, 2011, from http://www.act.org/aap/concordance/

Banta, Trudy W. (2008). Editor's Notes: *Trying to Clothe the Emperor*. Assessment Update, 20(2), 3-4, 15-16.

Barry, Carol L., Horst, S. Jeanne, Finney, Sara J., Brown, Allison R., & Kopp, Jason P. (2010). *Do Examinees Having Simlar 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(3), 441-462.

Bettinger, Eric. (2010). *Paying to Learn: The Effect of Financial Incentives on Elementary School Test Scores*. NBER Working Paper 16333. Cambridge, MA: The National Bureau of Economic Research.

Bracey, Gerald W. (1996). *Altering the motivation in testing*. Phi Delta Kappan, 78(3), 251-252.

Braun, Henry, Kirsch, Irwin, & Yamamoto, Kentaro. (2011). *An Experimental Study of the Effects of Monetary Incentives on Performance on the 12th-Grade NAEP Reading Assessment*. Teachers College Record, 113(11), 2309-2344.

Cole, James S. (2007). *Motivation to do well on low-stakes tests*. University of Missouri-Columbia, Columbia, MO.

Cole, James S., Bergin, David A., & Whittaker, Tiffany A. (2008). *Predicting student achievement for low stakes tests with effort and task value*. Contemporary Educational Psychology, 33, 609-624.

Eklöf, Hanna. (2007). *Test-Taking Motivation and Mathematics Performance in TIMSS 2003*. International Journal of Testing, 7(3), 311-326.

Eklöf, Hanna. (2008). *Test-taking motivation on low-stakes tests: A Swedish TIMSS 2003 example*. IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments. IEA-ETS Research Institute. Hamburg, Germany and Princeton, NJ.

ETS. (1993). *Data compendium for the NAEP 1992 mathematics assessment of the nation and the states.* (Report No. 23-ST-04). Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Guerriero, Michael. (2013). *Seattle's Low-Stakes Testing Trap*. The New Yorker. http://www.newyorker.com/online/blogs/newsdesk/2013/03/seattles-low-stakes-testing-trap.html

Hosch, Braden J. (2010). *Time on Test, Student Motivation, and Performance on the Collegiate Learning Assessment: Implications for Institutional Accountability*. Paper presented at the Association for Institutional Research Annual Forum, Chicago, IL.

Kiplinger, Vonda L., & Linn, Robert L. (1995/1996). *Raising the Stakes of Test Administration: The Impact on Student Performance on the National Assessment of Educational Progression*. Educational Assessment, 3(2), 111-133.

Lau, Abigail R., Swerdzewski, Peter, & Jones, Andrew T. (2009). *Proctors Matter Strategies for Increasing Examinee Effort on General Education Program Assessments*. General Education, 58(3), 196-217.

Levitt, Steven D., List, John A., Neckerman, Susanne, & Sadoff, Sally. (2011). *The Impact of Short-term Incentives on Student Performance*. http://bfi.uchicago.edu/events/20111028_experiments/papers/Levitt_List_Neckermann_Sadoff_Short-Term_Incentives_September2011.pdf

Liu, Ou Lydia, Bridgeman, Brent, & Adler, Rachel M. (2012). *Measuring learning outcomes in higher education: Motivation matters*. Educational Researcher, 41(9), 352-362.

Messick, Samuel. (1995). *Validity of Psychological Assessment: Validation of Inferences From Persons' Responses and Performances as Scientific Inquiry Into Score Meaning*. American Psychologist, 50(9), 741-749.

Napoli, Anthony R., & Raymond, Lanette A. (2004). *How Reliable Are Our Assessment Data?: A Comparison of the Reliability of Data Produced in Graded and Un-Graded Conditions*. Research in Higher Education, 48(8), 921-929.

O'Neil, Harold F., Abedi, Jamal, Miyoshi, Judy, & Mastergeorge, Ann. (2005). *Monetary Incentives for Low-Stakes Tests*. Educational Assessment, 10(3), 185-208.

O'Neil, Harold F., Sugrue, Brenda, & Baker, Eva L. (1995/1996). *Effects of Motivational Interventions on the National Assessment of Educational Progress Mathematics Performance*. Educational Assessment, 3(2), 135-157.

Steedle, Jeffrey T. (2014). *Motivation Filtering on a Multi-Institution Assessment of General College Outcomes*. Applied Measurement in Education, 27(1), 58-76.

Sundre, Donna L. (1999). *Does Examinee Motivation Moderate the Relationship between Test Consequences and Test Performance?* Paper presented at the Annual Meeting of the American Educational Research Association, Montreal, Canada.

Sundre, Donna L., & Wise, Steven L. (2003). '*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.

Wise, Steven L. (2006). *An Investigation of the Differential Effort Received by Items on a Low-Stakes Computer-Based Test*. Applied Measurement in Education, 19(2), 95-114.

Wise, Steven L., & DeMars, Christine E. (2005). *Low Examinee Effort in Low-Stakes Assessment: Problems and Potential Solutions.* Educational Assessment, 10(1), 1-17.

Wolf, Lisa F., & Smith, Jeffrey K. (1995). *The Consequence of Consequence: Motivation, Anxiety, and Test Performance.* Applied Measurement in Education, 8(3), 227-242.