CAE

The Effect of Attaching Stakes to a Performance-Base Assessment of Critical Thinking and Written Communication

April 10, 2017

Jonathan Lehrfeld, Ph.D.
Doris Zahner, Ph.D.

Introduction

Low motivation to perform well on a test is a potentially problematic source of construct-irrelevant variance (Messick, 1980), especially when there are low or no stakes attached to performance. That is, interpretations of test scores as indicators of students' knowledge and skills may be compromised if those students do not put forth the effort necessary to demonstrate the full extent of their abilities.

High motivation has been reported in some international testing contexts (Baumert & Demmrich, 2001; Eklöf, 2007), owing to factors including social responsibility, competitive spirit, interest, and personal or intrinsic motivators (Eklöf, 2008). Such findings sharply contrast with U.S. students taking low-stakes tests like NAEP where a large percentage of students report a low sense of importance and low effort (ETS, 1993).

Prior research has shown, however, that attaching stakes or 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). Therefore, the ability to raise the stakes of an assessment might lead to better scores, potentially via the causal effect of stimulating motivation to perform well. This would be a powerful finding, particularly as international comparative testing becomes more common.

This study investigates two cohorts of students and the possible associations with effort and performance on CLA+, an international assessment of college students' critical-thinking and writing skills. The first cohort was administered the assessment in a low-stakes condition where student performance was not associated with any outcomes. The second cohort was administered the assessment in a high-stakes condition where the students' CLA+ results would be used in conjunction with other criteria (e.g., GPA) to decide on their placement into a post-university program.

Method

Participants

The data used in this study come from 4,748 students in 2014 and 4,616 students in 2016, all of whom were part of a domestic military program that is offered at multiple higher education institutions. The students from 2014 were part of the low-stakes condition since their CLA+ scores were not used for placement decisions into their post-university program. The students from 2016 were part of the high-stakes condition because their CLA+ scores were used in conjunction with other variables such as GPA in placement into their post-university program. All students attended four-year institutions in the United States or its minor outlying islands. A matched data set was created for the students from the two administrations. After matching, the data set contained 5,666 students total, 2,833 students from each of the two administrations.

Instrument

The CLA+ is an internationally administered standardized assessment that measures critical thinking and written communication that consists of two sections. First, for the Performance Task (PT), students have 60 minutes to read a set of documents and respond to a prompt that asks them to analyze the information from a document library and write a solution to a real-world problem. Trained scorers evaluate the responses using scales that describe the quality of analysis and problem solving and writing effectiveness and mechanics. Following the PT, students have an additional 30 minutes to answer 25 document-based selected-response questions (SRQs) that are aligned to the same construct as the PT. Subscores for each section and a total scale score are awarded to each participant.

Propensity Score Matching

Propensity scores were generated using a logistic regression model. The outcome variable for the data set was having tested in 2016 (the high-stakes condition). In the propensity score

model, two types of predictor variables were used: student level and institution level. The student-level predictors were age, gender, race/ethnicity, parental education, field of study, and English language status. The institution-level predictors were Carnegie classification, Barron's selectivity index, public vs. private institution, HBCU status, and institution size.

After propensity scores were generated, one-to-one matching without replacement was implemented using the Matching package (Sekhon, 2011) in R (R Core Team, 2016). We followed Austin's (2011) recommendation for choosing caliper widths (i.e., the caliper width should be one-fifth the standard deviation of the sample distribution of logit-transformed propensity scores). Balance diagnostics were examined, and it was determined that adequate balance was achieved using this methodology.

Effect of Stakes

For graphical illustration, we provide a bar chart of mean differences between the low- and high-stakes groups on their CLA+ scores and on their self-reported motivation on the assessment. To examine the effect of stakes and other predictors on subsequent CLA+ scores, we use a series of linear regression models[1]. The models provide estimates for the average increase (or decrease) in score points associated with each predictor, including being in the high-stakes condition. Since all predictors were entered into the same model, the effect of stakes can be interpreted net of the other predictors.

Results

As is shown in Figure 1, there were significant differences between low-stakes and high-stakes conditions in all three CLA+ test scores (total, PT, and SRQ). Students who tested under high stakes had mean scores approximately 60 to 80 points higher than students who tested under low stakes. In addition to being statistically significant, the effect sizes indicated that the results are indicative of large and meaningful differences in the population.

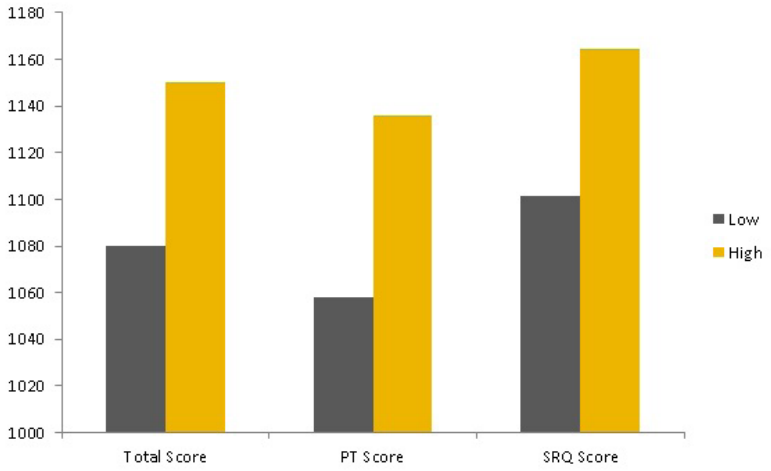

Figure 1. Effect of stakes on all CLA+ scores.

Table 1 shows the effect of motivation on CLA+ performance, net of the predictors in the propensity score models. The effect of stakes is always positive and statistically significant, meaning that students in the high-stakes condition performed significantly better than their counterparts in the low-stakes condition, even after accounting for all the other ways students could differ from each other before being made aware of the stakes of their assessment. As is shown in the rest of the table, these pre-existing differences include student-level variables such as age, gender, race/ethnicity, parental education, area of study, and native English speaking status, as well as institution-level characteristics such as Carnegie classification, Barron's selectivity index, institutional control, HBCU status, and institutional size. Specifically, having high stakes was associated with an average increase of 71 points in CLA+ total scores, 78 points in CLA+ PT scores, and 64 points in CLA+ SRQ scores, even after accounting for all the other predictors.

[1] We use regression instead of a series of independent-samples t-tests due to the recommendation of Ho, Imai, King, and Stuart (2007), who argue that any parametric analyses conducted after creating the matched data sets should control for the effects of the predictor variables from the propensity score model.

Table 1. Effect of Stakes and Other Predictors on CLA+ Scores

	Average Effect on Total Score	Average Effect on PT Score	Average Effect on SRQ Score
Having higher stakes	**+71**	**+78**	**+64**
Being…			
Female	-2	+2	-6
Asian	-11	+9	-31
Black	-80	-74	-86
Hispanic	-23	-16	-30
Science & Engineering Major	+36	+20	+52
Social Sciences Major	+30	+25	+36
Humanities Major	+11	+6	+15
Business Major	+13	+8	+17
Native English Speaker	+46	+42	+50
One More Year of…			
Age	-4	-3	-5
Parental Education	+8	+7	+8
Attending a School That…			
Grants Bachelors and Masters Degrees	+3	+2	+4
Grants Bachelors, Masters, and Doctoral Degrees	+8	+4	+12
Is Competitive	+9	+9	+9
Is Very Competitive	+21	+20	+22
Is Highly Competitive	+40	+35	+45
Is Most Competitive	+80	+73	+87
Is Public	-20	-22	-18
Is an HBCU	-51	-46	-56
Has 5,000-9,999 Students	+1	-1	+3
Has 10,000-19,999 Students	+6	+5	+8
Has 20,000+ Students	+16	+14	+17

Note. All effects are net of the effects of the other predictors in the table. Shaded cells represent statistically significant effects ($p < .05$). Levels of competitiveness refer to Barron's Selectivity Index. HBCU = Historically Black College or University. Counts of students refer to full-time equivalent undergraduate students.

Other interesting effects were those of gender, age, race/ethnicity, parental education, and field of study. Notably, there were no significant gender effects. Age was always significantly associated with lower CLA+ performance; this effect was always significant, but it should be noted that each extra year of age was only associated with a decrease in scores of between three and five points, depending on the type of score. However, this means that two students, identical on all other covariates except for the fact that one is four years older than the other, would be expected to differ in CLA+ scores by between 12 and 20 points, with an even greater difference as this gap increases[2]. The older student is expected to perform worse.

The effect for Asian students, like that of female students, was reversed depending on which outcome variable was examined: Asian students performed better on the PTs but worse on the SRQs. However, this effect was only significant for SRQ scores. Much larger effects were seen for Black and Hispanic students. Hispanic students scored, on average, 16 to 30 points lower, and Black students scored, on average, 74 to 86 points lower. It is notable that the effect of being Black was similar in magnitude to the effect of having high vs. low stakes on the assessment. These race/ethnicity effects are concerning and need to be addressed in future item/test development. They may also reflect underlying differences in educational opportunities or experiences among minority students in the normal course of their education.

[2] Note that this is the effect of age, not the effect of class level. All students in these models were juniors, so the effect of age must be interpreted as the effect of being an older student in the same class—for instance, non-traditional students returning to college (or beginning college) after an atypical break between high school and college.

Parental education also had a significant, albeit weak, relationship to assessment scores. For each extra year of parental education, there is a seven to eight point increase in CLA+ performance. Thus, the average difference between a student whose parents completed high school and a student whose parents completed college (and who are otherwise identical on all other covariates in the model), would be 28 to 32 points.

Finally, there are certain types of majors associated with better CLA+ performance than others. The reference category for field of study was helping/services majors (education; health care; home economics; law enforcement; parks, recreation, leisure studies, and sports management; physical education; public administration; and social work). Relative to the reference class, all other fields of study were associated with higher scores, but the estimates were modest for business (accounting, business administration, marketing, management, etc.) and humanities majors (art history, communications, English and literature studies, liberal studies, general studies, philosophy, religion, and visual or performing arts). On the other hand, the estimates associated with social sciences (anthropology; economics; ethical, cultural, or area studies; geography; history; interdisciplinary studies; political science; psychology; and sociology) or with sciences and engineering (agriculture, architecture, astronomy, biology, botany, chemistry, computer and information sciences, earth sciences, engineering and technology, mathematics, physics, and zoology) were much larger (i.e., 20 to 52 points better).

Finally, self-reported effort and engagement were higher for the high-stakes test takers (Figure 2) than for the low-stakes test takers. Effort and engagement together form a composite indicator of self-reported motivation on the assessment. Thus, the high-stakes test takers were more motivated than the low-stakes test takers. Again, the effect sizes indicate large and meaningful differences in the population.

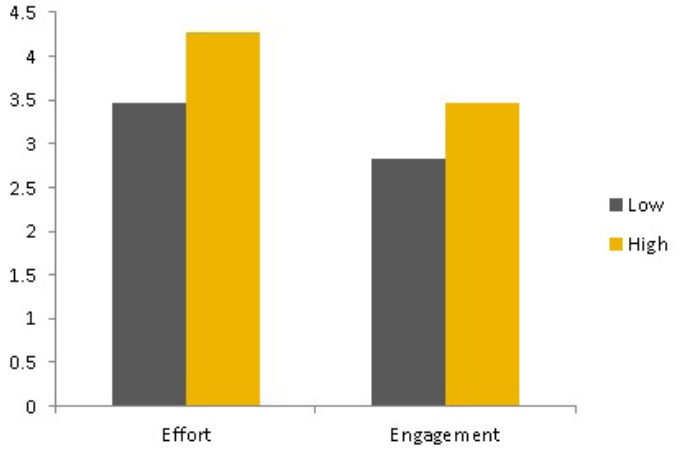

Figure 2: Effect of stakes on self-reported effort and engagement levels.

Discussion and Conclusion

The results of this study yielded a number of interesting findings. Stakes have an effect on student performance on CLA+, an internationally administered assessment of critical thinking and written communication. The students testing under high stakes performed significantly better than their matched counterparts from the low-stakes administration. One possible explanation is that knowing that their CLA+ performance would influence their future in a post-university program increased student motivation. This led to an increase in the amount of effort they put forth in the task, which yielded significantly higher test scores that the low-stakes cohort. In comparison, when CLA+ performance was not thought to affect placement in their post-university program, students from the low-stakes cohort were less motivated to perform well since their CLA+ results would not be used in any decision-making for placement. This explanation is supported by the self-reported effort and engagement data, which together form an indicator of student motivation on the assessment.

Despite the contributions this study offers to the literature on the effect of stakes on motivation and performance, there are a number of limitations worth noting. First, this study used a specific sample of college students entering one type of post-college program, so the results may not generalize to all college students. Second, it is very difficult to implement a

high-stakes testing situation at the college level, so even if these results did generalize to a larger population of college students, it might be difficult to enact a program that actually made use of these effects. Finally, the current study only assessed motivation through self-reports of students' levels of effort and engagement on the assessment. If these results could be replicated using other measures of motivation, that would add more weight to the findings. For instance, the Student Opinion Scale (Sundre, 2007) is a more direct measure of motivation and could add further insight into the causal mechanism underlying the relationship between stakes and performance.

This study has both theoretical and practical applications. From a theoretical standpoint, the study adds information to the literature on the relationship between stakes, motivation, and performance on an assessment. From a practical standpoint, information from this current study may help guide future attempts at increasing student motivation on assessments, as it illustrates that attaching higher stakes does in fact increase student performance on an assessment.

References

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, *46*, 399-424. doi:10.1080/00273171.2011.568786

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. doi:10.1007/BF03173192

Eklöf, H. (2007). Test-taking motivation and mathematics performance in TIMSS 2003. *International Journal of Testing*, *7*, 311-326. doi:10.1080/15305050701438074

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

ETS. (1993). Data compendium for the NAEP 1992 mathematics assessment of the nation and the states. Research Report No. 23-ST-04. Princeton, NJ: ETS.

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, *15*, 199-236. doi:10.1093/pan/mpl013

Liu, O. L., Bridgeman, B., & Adler, R. M. (2012). Measuring learning outcomes in higher education: Motivation matters. *Educational Researcher*, *41*, 352-362. doi:10.3102/0013189X12459679

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist*, *35*, 1012-1027. doi:10.1037/0003-066X.35.11.1012

Napoli, A. R., & Raymond, L. A. (2004). How reliable are our assessment data? A comparison of the reliability of data produced in graded and ungraded conditions. *Research in Higher Education*, *48*, 921-929. doi:10.1007/s11162-004-5954-y

R Core Team. (2017). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Com puting. https://www.R-project.org/

Sekhon, J. S. (2011). Multivariate and propensity score matching software with automated balance optimization: The Matching package for R. *Journal of Statistical Software*, *42*(7), 1-52. doi:10.18637/jss.v042.i07

Sundre, D. L. (2007). The Student Opinion Scale (SOS): A measure of examinee motivation: Test manual. Retrieved from https://www.researchgate.net/profile/Donna_Sundre/publication/251666711_The_Student_Opinion_Scale_A_Measure_of_Examinee_Motivation/links/53ede95e0cf23733e80b16f3.pdf

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potention solutions. *Educational Assessment*, *10*, 1-17. doi:10.1207/s15326977ea1001_1

Wolf, L. F., & Smith, J. K. (1995). The consequence of consequence: Motivation, anxiety, and test performance. *Applied Measurement in Education*, *8*, 227-242. doi:10.1207/s15324818ame0804_4