Incentives, Motivation, and Performance on a Low-Stakes Test of College Learning

Jeffrey T. Steedle
Council for Aid to Education

Presented at the 2010 Annual Meeting of the American Educational Research Association
Denver, CO
April 30, 2010

Abstract

Many colleges face the challenge of recruiting examinees for institutional assessment programs. Unfortunately, students recognize these programs as low-stakes, and therefore may not have high performance motivation. This study examined the interrelationships between students' preferred incentives for taking the test, their performance motivation, and their observed performance on the Collegiate Learning Assessment (CLA). Results reveal that freshmen stated a preference for cash and prizes over recognition and feedback as incentives. However, preferred incentives bore no practically significant relationship with test performance. For individual students, motivation was a significant predictor of CLA scores even after controlling for entering academic ability, but this effect was not apparent when schools were the units of analysis. Thus, schools should be attentive to motivation when results are used for individual-level assessment, but motivation is less of a concern when interpreting average scores relative to other schools with students of similar entering academic ability.

Incentives, Motivation, and Performance on a Low-Stakes Test of College Learning

As part of institutional improvement programs and efforts to demonstrate learning gains, many post-secondary schools administer standardized tests of general college learning. Testing all students is logistically and financially infeasible, so schools administer such tests to samples of students, and this affords reasonably precise estimates of the performance of larger student bodies (Klein, Benjamin, Shavelson, & Bolus, 2007). Recruiting a small sample of students may seem like a trivial task, but many schools are challenged by this aspect of test administration. Furthermore, administrators and faculty members sometimes express concern about poor performance motivation due to the low-stakes nature of the tests.

The research presented here speaks to these concerns by addressing the questions “What are students’ preferred incentives for taking a low-stakes test of college learning?” and “How are preferred incentives, motivation, and performance interrelated on such tests?” Answers to these questions are derived from analyses of data from the fall 2005 Collegiate Learning Assessment (CLA) administration, which included a survey of preferred incentives for taking the test and a measure of student motivation. Results from the survey analysis inform future recruiting practices by identifying the incentives that are most likely to attract students. An examination of the correlations sheds light on whether students’ preferred incentives are related to motivation and test performance and also whether schools should be concerned about low motivation.

Background

The CLA is an open-ended test of critical thinking and writing skills that is administered at several hundred colleges and universities annually as part of institutional assessment programs. To obtain a snapshot of student performance, most schools attempt to recruit 100 freshmen and 100 seniors to take the test, but some are challenged by this aspect of testing, especially when recruiting seniors during the academic term leading up to graduation (Ekman & Pelletier, 2008; Reyes & Rincon, 2008).

To meet CLA sample size recommendations, schools employ a variety of tactics including emails, phone calls, letters to students, fliers, and announcements in classes, and nearly all schools provide some form of renumeration (e.g., money, food, gift certificates, or preferential course registration). A 2008 CLA post-administration survey of 109 institutions indicated that about 70% used multiple recruitment methods (Kugelmass, 2008). Survey respondents indicated that emails (the most common approach) were usually viewed as “moderately” or “very” effective. Peer outreach was seen as “moderately” effective, and announcements in classes (a common approach) were reported as “very” effective. Phone calls, letters to students, and especially fliers were reported as less effective. About 40% of schools mandate CLA participation for freshmen (27% for seniors), and, not surprisingly, mandatory participation was most often considered “extremely” effective.

Other ideas for improving recruiting and motivation were shared at the 2009 CIC/CLA Consortium Summer Session Roundtable on Senior Participation (Eskew, 2009). These include moving the senior test administration earlier in the academic calendar to reduce recruitment challenges and perceived motivation problems, ensuring that senior students understand the value of assessment results to the institution, working with faculty and students to foster a “campus culture” emphasizing assessment and improvement, encouraging competition by sharing previous results with test takers, and employing students to help with recruitment.

Prior research (described below) has focused exclusively on performance motivation rather than motivation to volunteer for an institutional assessment. An analysis of paired-comparison survey data collected for this study address this practical problem by identifying the incentives that students prefer most. Students responded to a series of items asking them to state their preference among pairs of possible incentives, and statistical scaling procedures were employed to estimate the latent “preferability” of 11 incentives. Subsequent analyses revealed whether students who prefer incentives such as cash and prizes have different motivation or performance than students who prefer incentives such as recognition and feedback.

Prior research has established that motivation is significantly related to test performance (Cole, Bergin, & Whittaker, 2008; Uguroglu & Walberg, 1979; Wise & DeMars, 2005), and some experimental evidence indicates that students are less motivated and perform worse in low-stakes testing conditions (Cole & Osterlind, 2008; Wolf & Smith, 1995). Thus, school administrators are justifiably concerned about performance motivation on low-stakes tests of college learning. It should be noted, however, that many students expressed high levels of motivation on low-stakes international comparative assessments such as PISA and TIMSS (Baumert & Demmrich, 2001; Eklöf, 2007). Such assessments are not altogether dissimilar from university institutional assessment programs because they both use the performance of a sample of students for benchmarking against other, similar entities (countries or schools).

For institutional assessment programs, the school (rather than the student) is the unit of analysis, so the school average CLA score is the outcome of primary interest. Even if individual students’ motivation levels are significantly related to their CLA scores, it is not necessarily true that average motivation will be significantly related to average CLA performance. Specifically, if average motivation (even low average motivation) is similar across schools, the relative standing of schools would be unaffected by motivation. Of course, the relative standing of schools would likely be affected by motivation if schools differed notably in average motivation. For instance, if some testing condition (e.g., mandatory testing late in the evening) caused uniformly low motivation at a school, it would not be surprising if that school's average CLA score was artificially low.

Results from this study address these issues and supplement existing research by providing the correlation between motivation and performance on a low-stakes test of college learning. Additionally, linear regression was employed to study the value of motivation as a predictor of test performance after controlling for entering academic ability, a procedure sometimes used by institutional assessment programs to study performance relative to expected. These analyses were carried out twice: once treating students as the units of analysis and once treating schools as the units of analysis. Results from these two analyses tell somewhat different stories about the potential influence of motivation on test scores.

Methods

Subjects

In the fall of 2005, freshmen from 24 colleges and universities volunteered to take the CLA as well as surveys of testing motivation and preferred incentives as part of the first phase of a longitudinal study of student learning. The sample was 62% female and 65% White, with 12% reporting that English was not his or her primary language. The paired-comparison survey analysis used the sample of 2,242 students who completed the incentive survey, but correlation and regression results reflect the sub-sample of 1,826 students with complete data (CLA Total Score, Student Opinion Survey Total Score, and incentive survey—described below). All students had either SAT or ACT Total Scores on file.

Measures

Each participating student took the CLA, the Student Opinion Survey (SOS), and a paired-comparison survey of incentives. The CLA is an essay-based measure of college students' critical thinking and writing skills as applied to open-ended problems requiring students to analyze and evaluate information in order to make a decision, present and support a position, or find faults in an argument. Students completed 180 minutes of CLA testing that included a Performance Task, a Make-an-Argument prompt, and a Critique-an-Argument prompt (see www.cae.org/cla). A CLA Total Score was generated by weighting the subtests 50%, 25%, and 25%, respectively. The SOS is a motivation scale with 10 Likert-scale items ($\alpha = 0.83$) that measures a student's effort and belief that performing well is important (Sundre & Moore, 2002). The 22-item paired-comparison survey presented students with pairs of the incentives listed in Table 1 (some might be called "inducements" rather than incentives). Students responded to the prompt "Please select the choice that best describes which of the two options in a pair would be the stronger motivator to take this test" with answer options "Strongly prefer option A," "Prefer option A," "Options are equally attractive," "Prefer option B," and "Strongly prefer option B."

Table 1
Incentives included in the preferred incentives survey

	Incentive	Description
1.	adminask	A school administrator (President, Dean, Provost, etc.) asks me to (in a letter)
2.	cash	I am paid in money ($25)
3.	comparison	To see how I do compared to other students
4.	earlyreg	I am given early registration preference for the next semester
5.	facultyask	A faculty member (Professor, Advisor, etc.) asks me to (in a letter)
6.	helpschool	To help my school assess student learning
7.	inkind	I am paid in kind (discount at book store, gift certificate, etc.) (worth $25)
8.	printed	My school acknowledges my participation in printed materials (newsletter, annual report, etc.)
9.	prize	There is the chance to win a prize (1 in 10 chance of winning a $250 iPod)
10.	resume	To put my score on my resume
11.	strengths	To understand my strengths and weaknesses

Analyses

A multivariate generalization of the Bradley and Terry (1952) model for analyzing paired-comparison data akin to the many-faceted Rasch model (Linacre, 1989) was employed to estimate the locations of the 11 incentives along a unidimensional latent scale of "preferability." The following discussion explains the relevant notation and also describes how this model was built from components of other item response models.

Bradley and Terry (1952) proposed a logit model for paired comparisons of the form

$$\log[P(Y_{hi} = 1) / P(Y_{hi} = 0)] = \mu_h - \mu_i$$

or equivalently

$$P(Y_{hi} = 1) = \frac{e^{\mu_h - \mu_i}}{1 + e^{\mu_h - \mu_i}}$$

where μ_h and μ_i are the respective locations of objects *h* and *i* on a unidimensional latent scale of preferability. In this model, let $Y_{hi} = 1$ if object *h* is preferred over object *i*, and let $Y_{hi} = 0$ if object *i* is preferred over object *h*. When object *i* is more preferable than object *h* ($\mu_h - \mu_i < 0$), the probability of preferring object *h* is less than 0.50, and when object *h* is more preferable than object *i* ($\mu_h - \mu_i > 0$), the probability of preferring object *h* is greater than 0.50 (see Figure 1).

The Bradley-Terry model closely resembles the Rasch (1960) model for analyzing dichotomous assessment data, which takes the form

$$P(X_j = 1) = \frac{e^{\theta - \delta_j}}{1 + e^{\theta - \delta_j}}$$

where $P(X_j = 1)$ is the probability that a person with ability θ will correctly respond to item *j* with difficulty δ_j. In this model, θ and δ_j represent locations along the unidimensional latent trait scale of the construct measured by the assessment.

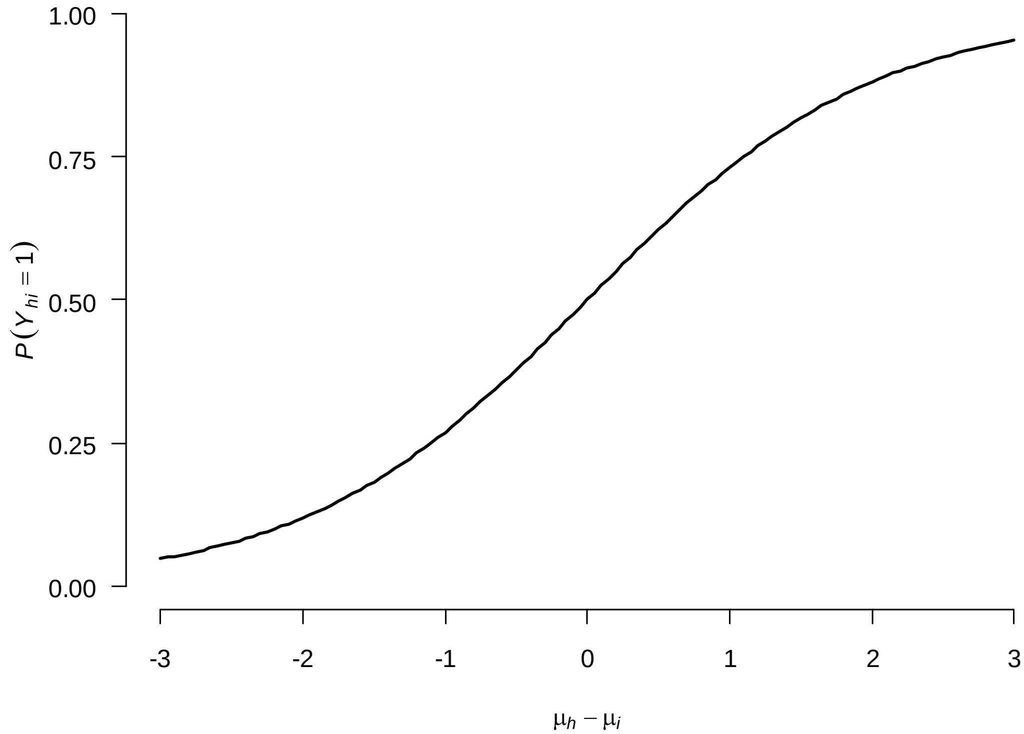

Figure 1. Response curve for the Bradley-Terry model.

The models described above are only appropriate for dichotomous data (i.e., scored 0 or 1), but data from tests and surveys are often polytomous. For instance, in a 5-choice paired-comparison survey like the one analyzed here, subjects select from response options reflecting strong preference for *i*, moderate preference for *i*, no preference, moderate preference for *h*, and strong preference for *h* (i.e., $h << i$, $h < i$, $h = i$, $h > i$, and $h >> i$).

There are numerous extensions of the Rasch model that allow for the analysis of polytomous data. Andrich's (1978) Rating Scale Model (RSM) is one such extensions that is well suited to analyzing data from surveys in which all items have the same set of response options reflecting ordered categories. The category response function of the RSM is

$$P(X_j = x) = \frac{\exp\left(\sum_{s=0}^{x}[\theta - (\delta_j + \tau_s)]\right)}{\sum_{k=0}^{M}\exp\left(\sum_{s=0}^{k}[\theta - (\delta_j + \tau_s)]\right)}$$

also defining

$$\sum_{s=0}^{0}[\theta - (\delta + \tau_s)] \equiv 0 \text{ and } \sum_{j=1}^{N}\sum_{s=0}^{M}(\delta_j + \tau_s) \equiv 0$$

where X_j takes on the possible values 0,1, …, M and $\tau = (0,\tau_2,\ldots,\tau_M)$ is a vector of M "step" parameters used for all N items. The term "step" reflects the notion that it is generally more "difficult" to reach higher scores (i.e., requires more of whatever trait is being measured). Consider that there are M steps between the M+1 scale points. When $X_j = 0$ (the lowest possible score), no steps have been passed. When X_j = M (the highest possible score), all M steps have been passed.

The model that follows reflects a modification of the RSM for the analysis of paired comparison data (roughly following the notation of Agresti, 1992). Let I be the number of objects to compare, and let J be the number of response categories. When $J = 5$, the response categories would be $h << i$, $h < i$, $h = i$, $h > i$, and $h >> i$ for j = 1, 2, 3, 4, and 5, respectively. In this model, let $P(Y_{hi} = j)$ be the probability that response j will result from the comparison of objects h and i, where $j = 1$ strongly favors object i, and $j = J$ strongly favors object h. Let $\tau = (0,\tau_2,\ldots,\tau_J)$ be a vector of object location adjustments for response categories $1,\ldots,J$. When comparing object h to object i, the probability that response $Y_{hi} = j$ (for $j = 1,\ldots,J$) given the locations of objects h and i (μ_h and μ_i) and location adjustment vector τ is

$$P(Y_{hi} = j \mid \mu_h, \mu_i, \tau) = \frac{\exp\left(\sum_{s=1}^{j}[\mu_h - (\mu_i + \tau_s)]\right)}{\sum_{y=1}^{J}\exp\left(\sum_{s=1}^{y}[\mu_h - (\mu_i + \tau_s)]\right)}$$

also defining

$$\sum_{s=1}^{1}[\mu_h - (\mu_i + \tau_s)] \equiv 0 \text{ and } \sum_{i=1}^{I}\mu_i \equiv 0$$

to anchor the scales.

For a graphical explanation, consider Figure 2, which provides generic category response curves based on the model. Notice that, when objects *h* and *i* are equally preferable (i.e., $\mu_h = \mu_i$ or $\mu_h - \mu_i = 0$), the most likely response is “Objects equal” (orange dotted line in Figure 2). When object *i* is preferable to object *h* ($\mu_h - \mu_i < 0$), students tend to choose “Prefer *i*” or “Strongly prefer *i*” (the red and black lines). The “Prefer *h*” and “Strongly prefer *h*” options (the green and blue lines) become most likely when object *h* is preferable to object *i* ($\mu_h - \mu_i < 0$).

This model can also accommodate the inclusion of a parameter characterizing the response behavior of each person. When a rater facet is included, the model becomes

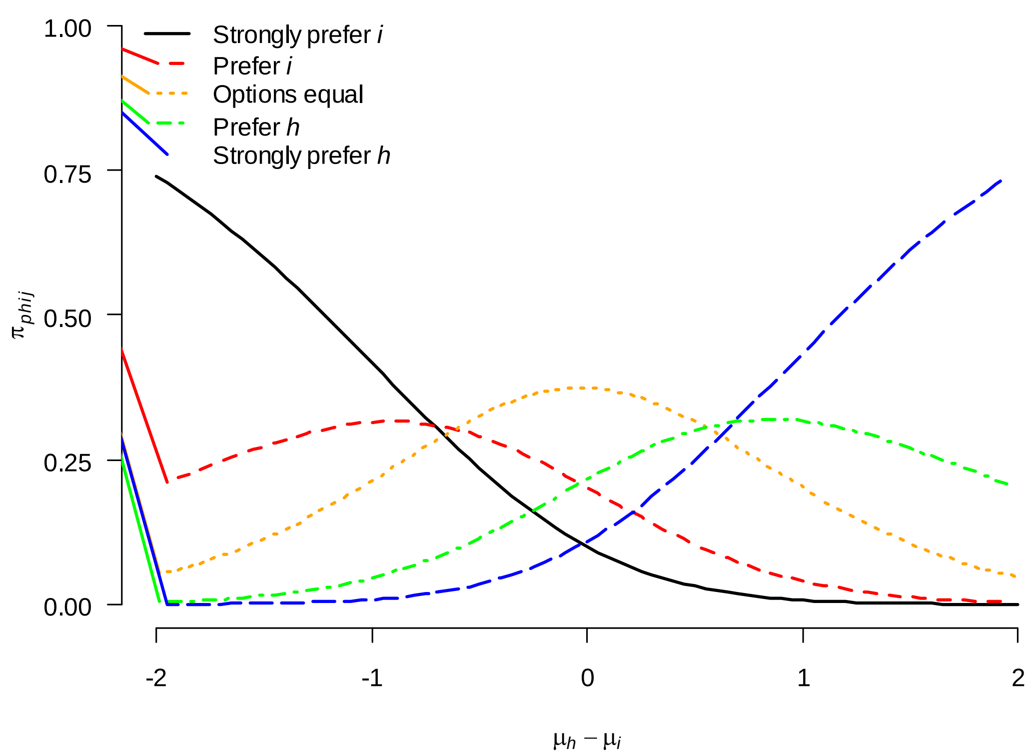

Figure 2. Generic category response curves based on the polytomous paired-comparison model.

$$P(Y_{phi} = j \mid \mu_h, \mu_i, \tau, \lambda_p) = \pi_{phij} = \frac{\exp\left(\sum_{s=1}^{j}[\lambda_p + \mu_h - (\mu_i + \tau_s)]\right)}{\sum_{y=1}^{J}\exp\left(\sum_{s=1}^{y}[\lambda_p + \mu_h - (\mu_i + \tau_s)]\right)}$$

also defining

$$\sum_{s=1}^{1}[\lambda_p + \mu_h - (\mu_i + \tau_s)] \equiv 0$$

and

$$\sum_{i=1}^{I}\mu_i \equiv 0 \text{ and } \sum_{p=1}^{N_p}\lambda_p \equiv 0$$

to anchor the scales. Here, λ_p is the parameter that characterizes person *p*, and π_{hipj} is the probability that person *p* selects response *j* when comparing objects *h* and *i*. Note that $\lambda_p \sim N(0,1)$. Markov Chain Monte Carlo estimation was carried out using WinBUGS (Spiegelhalter, Thomas, Best, & Lunn, 2003). The Appendix provides the WinBUGS code used to fit this model.

OUTFIT statistics (outlier-sensitive fit statistics, Wright & Masters, 1982) were computed for the person parameters in order to evaluate person fit, which in this case, indicates closeness to the consensus view. The OUTFIT statistic for person *p* is the sum of his or her squared residuals reflecting the difference between observed and expected responses and standardized using the square root of the variance of the squared residuals. The formula for an OUTFIT statistic for person *p* is

$$u_p = \sum_{n=1}^{N} z_{np}^2 / N$$

where

$$z_{np} = \frac{Y_{np} - E_{np}}{\sqrt{\sum_{j=1}^{J}(j - E_{np})^2 \pi_{npj}}}$$

is the standardized residual for person *p* on item *n* (comparing some combination of objects *h* and *i*). Here, *Y* is the observed response, and *E* is the expected response

$$E_{np} = \sum_{j=1}^{J} j\pi_{npj} .$$

Generally, an OUTFIT statistic near 1.0 reflects good person fit. In this context, good person fit indicates that a student's preferred incentives were close to the consensus view.

Correlations between CLA scores, SOS scores, and OUTFIT statistics were computed. In addition, linear regression analyses were employed to evaluate the contribution of motivation to explaining variation in CLA scores after controlling for entering academic ability (as measured by the SAT or ACT). These analyses were carried out using individual-level data first and then using school average scores.

Results

Preferability of Incentives

Indicators of MCMC convergence were favorable and overall model data fit was good (Gelman, Carlin, Stern, & Rubin, 2004). Figure 3 shows the posterior density distributions of the latent locations of the incentives (μ parameters) along a unidimensional scale of "preferability," and Table 2 provides the mean values of those distributions. Results indicate that cash was the most preferable incentive, followed by early course registration, non-cash compensation, and a chance to win a prize. These were followed by learning about one's strengths and weaknesses, being asked by faculty, and being asked by administrators (all similarly preferable). The least preferable incentives were helping one's school, comparing oneself to other students, having something to put on one's resume, and getting recognition in printed materials.

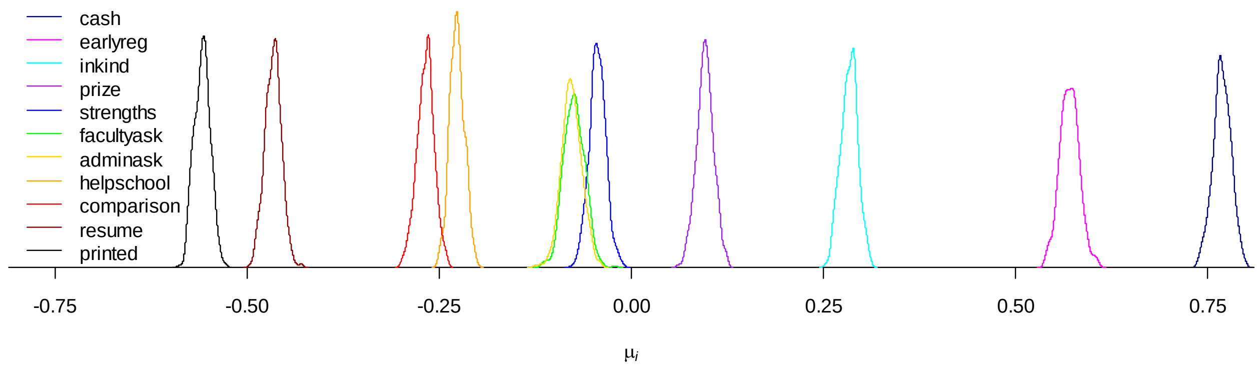

Figure 3. Posterior distributions of latent preferability.

Table 2

Mean of posterior distributions of latent preferability

Incentive	Mean
cash	0.77
earlyreg	0.57
inkind	0.28
prize	0.10
strengths	-0.04
facultyask	-0.07
adminask	-0.08
helpschool	-0.23
comparison	-0.27
resume	-0.47
printed	-0.56

Student-level Analysis

OUTFIT statistics were used as an index of students preferred incentives. Students with good person fit tended to express the consensus view (heavily favoring cash and prizes), and students with poor person fit tended to prefer other incentives such as helping one's school, comparing oneself to other students, and learning about one's strengths and weaknesses. The vast majority of students had OUTFIT statistics near 1.0 (91.5% below 2.0), indicating general closeness to the consensus view.

OUTFIT statistics (labeled u_p) were correlated with SOS Total Scores and with CLA Total Scores (Figure 4). Although both correlations were statistically significant ($p < .05$), neither was notably different from zero (0.08 and -0.05, respectively). Thus, preferred incentives, as indicated by OUTFIT statistics, bore no practically significant relationship with motivation or CLA performance. Note that these low correlations may reflect restriction of range (i.e., little variation) in students' preferred incentives. Had there been a larger number of students with "non-consensus" views, relationships among the variables might have been apparent.

The correlation between motivation and CLA performance was 0.23 ($p < .001$), which means that motivation accounts for 5% of the variation in CLA scores (Figure 5). Some standardized tests of college learning report scores that control for entering ability (e.g., value-

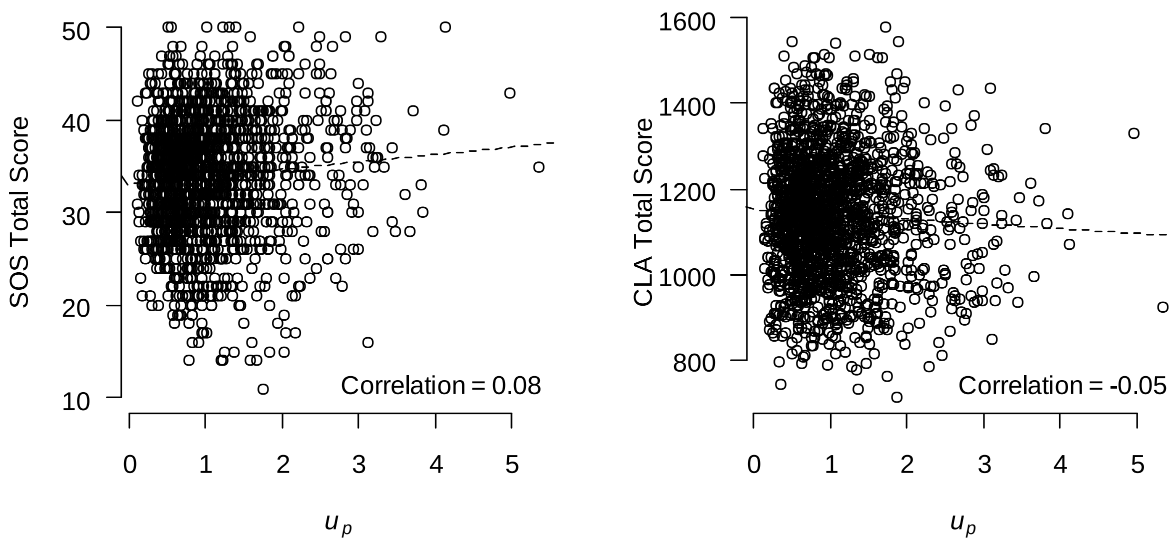

Figure 4. Scatterplots of SOS Total Score versus OUTFIT statistics and of CLA Total Score versus OUTFIT statistics.

added scores), so motivation was also correlated with CLA scores after removing the component linearly associated with SAT (or converted ACT) scores. This correlation ($r = 0.29$, $p < .001$) indicates that motivation remains a significant predictor of CLA scores even after controlling for entering ability (Figure 5). In fact, SAT scores were uncorrelated with motivation ($r = -0.005$, non-significant), which means that the 5% of CLA score variability accounted for by motivation does not overlap with the 34% accounted for by SAT scores. Multiple regression results indicate

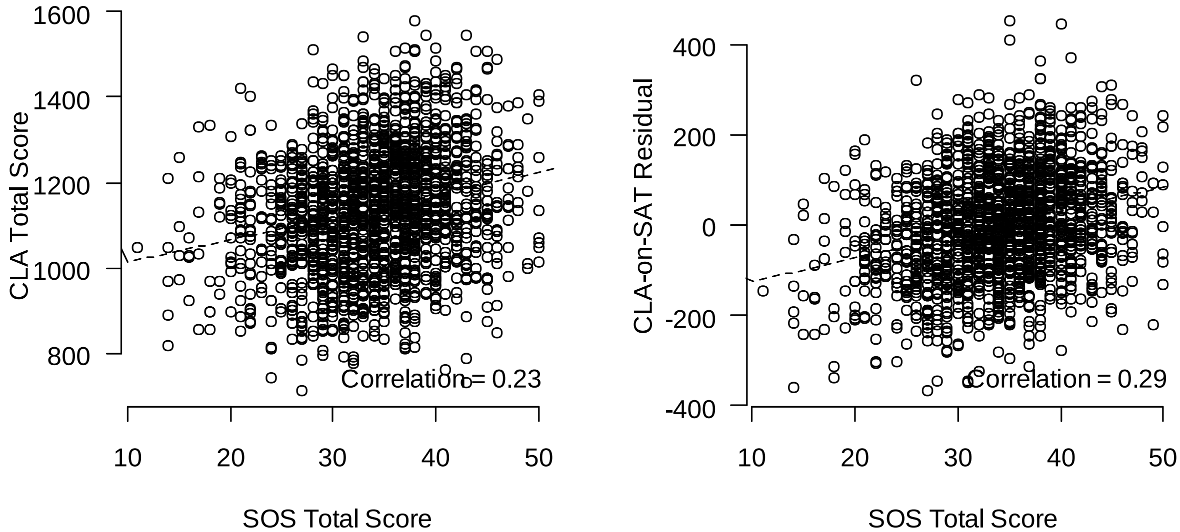

Figure 5. Scatterplots of CLA Total Score versus SOS Total Score and of CLA Total Score after controlling for ability (CLA-on-SAT Residual) versus SOS Total Score.

Table 3
Regression coefficients and standardized regression coefficients for models predicting CLA scale total scores using student's SOS and SAT scores

Model	Coefficient	Estimate	Standardized estimate	R^2
1	SOS	5.232	0.232	0.054
2	SAT	0.458	0.583	0.340
3	SOS	5.297	0.235	0.395
	SAT	0.459	0.584	

Note: All coefficients were significant at $p < .001$.

that, together, entering ability and motivation accounted for 39.5% of the variation in CLA scores (Table 3).

School-level Analysis

The student-level analysis was replicated using data from 19 schools with at least 25 students having CLA, SOS, and SAT scores. Note that one school with exceptionally low average motivation was excluded to prevent it from having undue influence on the correlations and regression results. The plots provided in this section have the same x- and y-axes as those in Figures 4 and 5 to allow for comparisons and to show that school average scores have much less variance that individual scores.

The scatterplots and correlations between mean OUTFIT statistics and mean SOS scores and between mean OUTFIT statistics and mean CLA scores are provided in Figure 6. Despite there being very little variation in mean OUTFIT statistics and mean SOS total scores, there is a significant positive relationship between the two ($r = 0.53$, $p < .05$). This suggests that schools

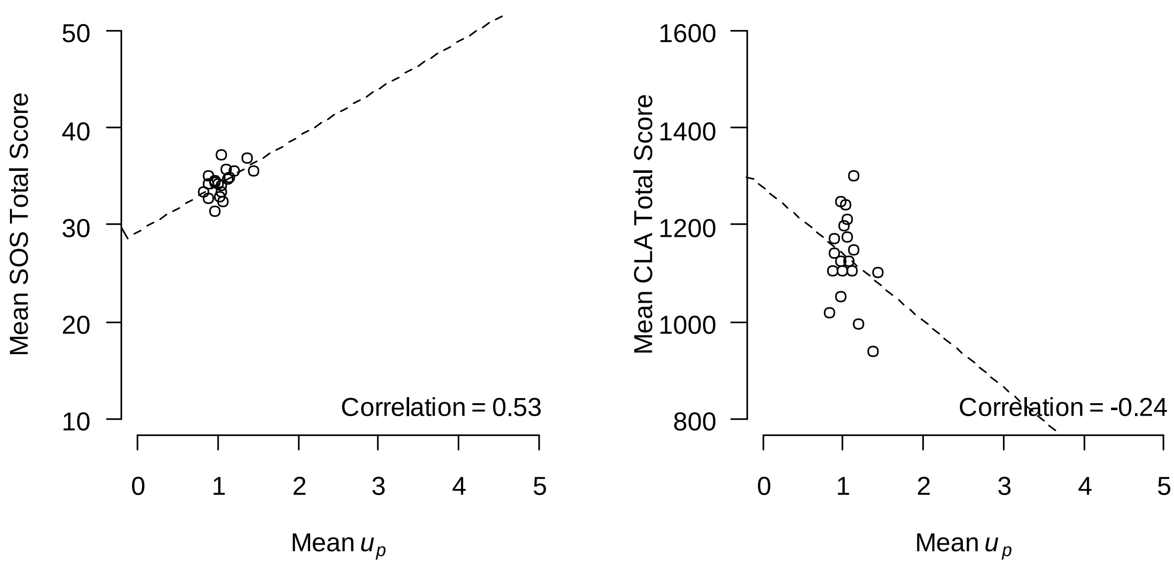

Figure 6. Scatterplots of SOS Total Score versus preferred incentives (u_p) and of CLA Total Score versus preferred incentives (u_p).

with more students who are motivated by things other than cash and prizes tend to have higher average motivation. However, the higher average motivation of the students in those schools did not translate into higher average CLA scores, as indicated by the non-significant correlation between mean OUTFIT statistics and mean CLA scores (*r* = -0.24, *p* = 0.32)

This was corroborated by the non-significant correlations between mean motivation and mean CLA scores (*r* = -0.38, *p* = .11) and between mean motivation and mean CLA scores after controlling for mean SAT scores (*r* = 0.11, *p* = .64) shown by the scatterplots in Figure 7. Corresponding regression results are provided in Table 4. Average motivation was not a significant predictor of average CLA scores regardless of whether controls were present for average SAT. It was noted earlier that this finding would be expected if all schools had a similar average motivation level, and this seems to be the case. The average of the average SOS scores was 34.4 with a standard deviation of 1.5 on a scale between 10 and 50.

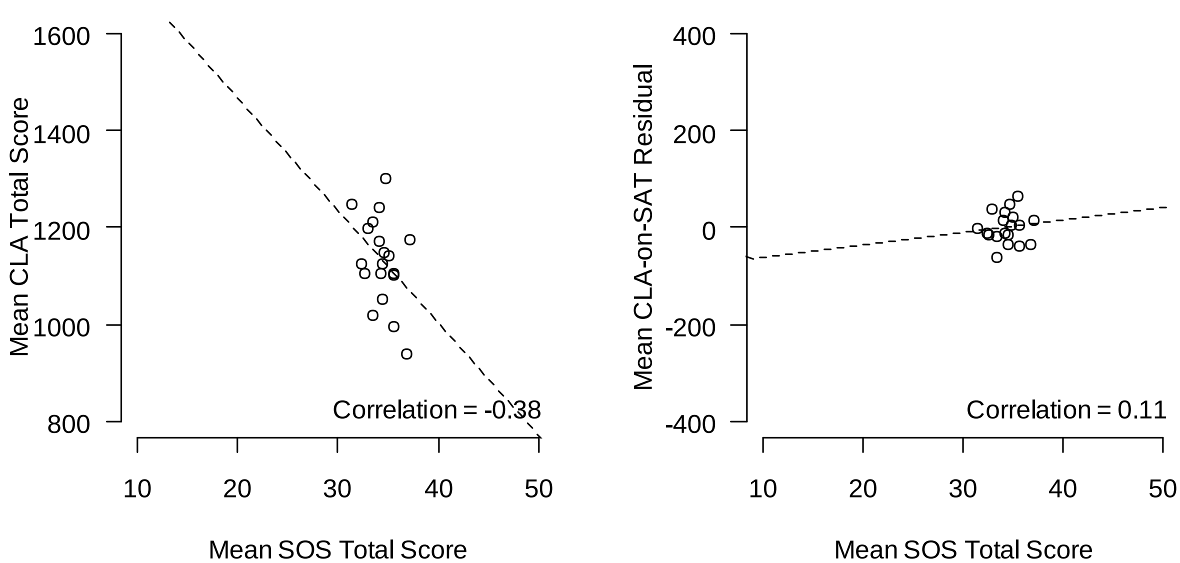

Figure 7. Scatterplots of CLA Total Score versus SOS Total Score and of CLA Total Score after controlling for ability (CLA-on-SAT Residual) versus SOS Total Score.

Table 4

Regression coefficients and standardized regression coefficients for models predicting school mean CLA scale total scores using school mean SOS and school mean SAT

Model	Coefficient	Estimate	Standardized estimate	R^2
1	Mean SOS	-23.23 (ns.)	-0.38 (ns.)	0.141
2	Mean SAT	0.60	0.93	0.873
3	Mean SOS	3.13 (ns.)	0.05 (ns.)	0.875
	Mean SAT	0.62	0.96	

Note: All coefficients were significant at $p < .001$ unless labeled nonsignificant (ns.).

Discussion and Conclusions

The paired-comparison survey results presented here indicate that, when it comes to recruiting college freshmen, cash and prizes are preferred to non-renumeration alternatives. It may be the case that freshmen do not yet see the long-term benefits of learning about one's academic strengths and weaknesses or building one's resume. The fact that being asked by faculty (or administrators) was moderately preferable suggests that greater faculty involvement could improve recruiting yields. Of course, none of these strategies are likely to be as effective as requiring students to take a test.

One major limitation of this study is that it was not known how students were recruited to take the CLA. If many of them were recruited using cash and prizes, it would not be surprising if these students reported cash and prizes as the most preferable incentives. According to 2008 post-administration survey results (Kugelmass, 2008), only 10% of schools had voluntary participation with no incentives. Nearly all schools (even many that mandated participation) gave students something for their participation (money, gift certificates, priority course registration, food, extra credit). Different sorts of students might be recruited if schools emphasized the personal benefits of participating in institutional assessment and eschewed cash and prizes. However, the results presented here suggest that this strategy might not draw a sufficient number of students and would not influence test results.

This analysis revealed no practically significant relationship between preferred incentives and test performance (though the relationship between preferred incentives and *motivation* was significant at the school level). Thus, the incentives that schools employ to recruit students should not impact institutional assessment results. Even though being asked by faculty to take the CLA was only moderately preferable, improved faculty involvement may still have a positive impact on motivation (and possibly subsequent test performance) if faculty stress the importance of the test. As Baumert and Demmrich (2001) noted, "…accentuating the societal utility value of the test, and thus inducing situational interest, is in itself a sufficient condition for the generation of test motivation" (p.458). However, this would likely require extensive professional development because many faculty members express skepticism about the utility of standardized tests of college learning and see such testing as an unwelcome intrusion into academic affairs. Anecdotally, students have reported being told by professors that the "test doesn't matter." Note that the effects of mandatory testing on motivation and performance are still unknown.

Even after controlling CLA scores for entering academic ability, motivation accounted for 5% of CLA score variation for individual students (a notable incremental improvement over the 34% accounted for by entering ability). Thus, it is sensible to have concerns about motivation on low-stakes tests of college learning when results are used to compare students or identify their individual strengths and weaknesses. Specifically, students with low performance motivation (and subsequent poor performance) may not receive scores that can be validly interpreted.

In the school-level analysis, average motivation was not a significant predictor of average CLA scores. It accounted for 14% (non-significant) of average CLA score variation, but after controlling for average SAT scores, average motivation provided no incremental improvement in average CLA score prediction. That is, accounting for average motivation did not impact the relative standing of schools, in particular when controls for student ability are employed. This result validates the use of school average scores for norm-referenced comparisons between schools (and highlights the value of using controls for entering ability), but the same cannot necessarily be said for criterion-referenced interpretations of results. If a school's average score is to serve as an indicator of maximum performance relative to some standard of proficiency (not relative to other schools), valid interpretations could be jeopardized in the presence of low motivation (CLA results are commonly framed as indicators of typical rather than maximum performance).

A limitation of the correlation and regression results is that they may not generalize beyond the 180-minute version of the CLA (students usually take either one Performance Task or the combination of Make-an-Argument and Critique-an-Argument). In addition, there were only 19 schools available for the school-level analyses, which limited the statistical precision of the school-level correlation and regression results. Conclusions about the relationships between school average motivation and test performance might have been different with a larger, more diverse sample of schools. Previous research has indicated that average self-reported effort accounts for an additional 3% to 7% of average CLA score variance beyond the 70% accounted for by average entering ability (Klein et al., 2007), but increasing from 70% to 75% is a relatively small incremental improvement compared to the increase from 34% to 39% observed in the student-level analysis.

To sum up, for schools that do not require their students to sit for low-stakes tests of college learning, cash and prizes seem to be the most preferable recruiting incentives for freshmen. With regard to motivation, results from this study suggest that schools may be justifiably concerned in some testing scenarios: when results are used to evaluate individual students and when results are intended to facilitate criterion-referenced inferences about school performance. In these situations, schools should seek to optimize motivation in order to ensure the interpretability of results. Most research (this study included) treats recruiting and motivation as separate issues, but future research could examine testing conditions that simultaneously incentivize students to participate and motivate them to put forth effort.

References

Agresti, A. (1992). Analysis of ordinal paired comparison data. *Applied Statistics, 41*(2), 287-297.

Andrich, D. A. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*(4), 561-573.

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.

Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs: The method of paired comparisons. *Biometrika, 39*, 324-345.

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

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*(2), 119-130.

Eklöf, H. (2007). Test-taking motivation and mathematics performance in TIMSS 2003. *International Journal of Testing, 7*(3), 311-326.

Ekman, R., & Pelletier, S. (2008). Assessing student learning: A work in progress. *Change, 40*(4), 14-19.

Eskew, R. (2009). Notes from roundtable on senior participation [personal email].

Gelman, A., Carlin, B. P., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis*. New York, NY: Chapman & Hall/CRC.

Klein, S., Benjamin, R., Shavelson, R., & Bolus, R. (2007). The collegiate learning assessment: Facts and fantasies. *Evaluation Review, 31*(5), 415-439.

Kugelmass, H. (2008). *2008 cla post-administration survey report*. New York, NY: Council for Aid to Education.

Linacre, J. M. (1989). *Many-faceted Rasch measurement*. Chicago, IL: MESA.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.

Reyes, P., & Rincon, R. (2008). The Texas experience with accountability and student learning assessment. In V. M. H. Borden & G. R. Pike (Eds.), *Assessing and accounting for student learning: Beyond the Spellings commission: New directions for institutional research, assessment supplement 2007* (Vol. 2008, pp. 49-58). San Francisco, CA: Jossey-Bass.

Spiegelhalter, D., Thomas, A., Best, N., & Lunn, D. (2003). *WinBUGS user manual*. Cambridge, UK: MRC Biostatistics Unit, Institute of Public Health.

Sundre, D. L., & Moore, D. L. (2002). The student opinion scale: A measure of examinee motivation. *Assessment Update*, *14*(1), 8-9.

Uguroglu, M. E., & Walberg, H. J. (1979). Motivational achievement: A quantitative synthesis. *American Educational Research Journal*, *16*(4), 375-390.

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

Wolf, L. F., & Smith, J. K. (1995). The consequence of consequence: Motivation, anxiety, and test performance. *Applied Measurement in Education*, *8*(3), 227-242.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis: Rasch measurement*. Chicago, IL: MESA.

Appendix: WinBUGS code (with comments shown in green)

```
# WinBUGS code for analyzing pairwise comparison data
# I = number of objects, J = number of response categories
# Np = number of persons, Nc = number of comparisons (items)
# Hc = vector of is (first object in comparisons), Ic = vector of js (second object)
# Y = Np x Nc matrix of survey responses

model {
  # Prior probabilities for object locations
  for(i in 1:(I-1)) {
    mu[i] ~ dnorm(0,1)
  }
  mu[I] <- -sum(mu[1:(I-1)]) # Set sum of mus = zero for anchoring

  # Prior probabilities for location adjustment (step) parameters
  tau[1] <- 0
  for(j in 2:J) {
    tau[j] ~ dnorm(0,1)
  }

  # Prior probabilities for rater facets
  for(p in 1:(Np-1)) {
    lambda[p] ~ dnorm(0,1)
  }
  lambda[Np] <- -sum(lambda[1:(Np-1)]) # Set sum of lambdas = zero for anchoring

  # Compute probabilities
  for(p in 1:Np) {
    for(c in 1:Nc) {
      diffterm[p,c,1] <- 0
      for(j in 2:J) {
        diffterm[p,c,j] <- lambda[p] + mu[Hc[c]] - (mu[Ic[c]] + tau[j])
      }
    }
  }

  for(p in 1:Np) {
    for(c in 1:Nc) {
      for(j in 1:J) {
        numer[p,c,j] <- exp(sum(diffterm[p,c,1:j]))
      }
    }
  }

  for(p in 1:Np) {
    for(c in 1:Nc) {
      denom[p,c] <- sum(numer[p,c,1:J])
      for(j in 1:J) {
        pi[p,c,j] <- numer[p,c,j]/denom[p,c]
      }
    }
  }

  # Model response of person p comparing objects i and j
  for(p in 1:Np) {
    for(c in 1:Nc) {
      Y[p,c] ~ dcat(pi[p,c,1:J])
    }
  }
}
```