Comparing Student Performance on Low-Stakes and High-Stakes Evaluations of Conceptual Understanding

Rachel V Vitali
Department of Mechanical Engineering
University of Michigan
Ann Arbor, MI, USA
vitalir@umich.edu

Noel C Perkins
Department of Mechanical Engineering
University of Michigan
Ann Arbor, MI, USA
ncp@umich.edu

Cynthia J Finelli
Department of Electrical Engineering
University of Michigan
Ann Arbor, MI, USA
cfinelli@umich.edu

***Abstract* — This Research to Practice Full Paper investigates how student performance during a low-stakes evaluation of conceptual understanding relates to performance during a high stakes evaluation. Concept inventories, like the Dynamics Concept Inventory (DCI), are used to assess student understanding of a specific set of concepts, typically in a low-stakes setting. The DCI was offered for very modest course extra credit at the beginning and end of an introductory dynamics course at a large, research-intensive public university. This represents the low-stakes evaluation. By contrast, a midterm examination administered 12 weeks into the 15 week semester included: 1) a set of short-answer concept questions related to the concepts covered by the DCI and 2) two traditional long-answer problem-solving questions. This midterm examination represents the culmination of student understanding of rigid body dynamics, which is primarily what the DCI covers. It also represents the high-stakes evaluation as this examination strongly influences the students' final grade in the class. The synergy of the relationships between these three evaluations indicate the low-stakes DCI assessment is a valid measure in quantifying student conceptual understanding.**

Keywords—student assessment; course assessment; concept inventory; undergraduate.

I. INTRODUCTION

Diagnostic evaluation of student conceptual understanding takes a number of forms in education research [1, 2]. Some researchers advocate the use of interview sessions with students to probe their understanding of the subject matter [3, 4, 5], though this method of assessment is not scalable to a number of scenarios including large (100+ students) sections of introductory undergraduate courses. Others opt for a set of multiple choice questions populated with answers usually developed from students' past answers and typifying frequently misunderstood topics related to that concept (commonly known as distractors) [6, 7]. Concept inventories developed in recent years follow this latter trend. They undergo extensive evaluation to assess validity and reliability in assessing student conceptual understanding and ensure that similarly-groups items are testing the same underlying concept [8]. Concept inventories have been used for both high-stakes assessment (e.g., counted in the score on a final exam) and low-stakes assessment (e.g., offered outside of the normal course grading scheme to ascertain students' overall understanding). However, there is increasing concern that scores from low-stakes evaluations may not be able to produce unbiased representations of student knowledge [9].

The Dynamics Concept Inventory (DCI) is designed around important and/or commonly misunderstood concepts in engineering dynamics, identified via a modified Delphi process that included veteran course instructors for introductory dynamics [10, 11]. It is a 29-item instrument that assess 14 separate concepts. A recent study determined, "the DCI can function as a low-stakes instrument that educators can use to identify overall understanding of all concepts identified in the DCI" [8, pp. 479]. Similarly, Stites et al. show an abbreviated DCI is reliable and valid and correlates with student performance on traditional long-answer problem-solving exam questions that require problem-solving skills in a high-stakes setting [12]. These two studies ([8] and [12]) both provide support for the reliability of the DCI as an evaluative tool to measure student conceptual understanding.

This paper investigates how conceptual understanding in a high-stakes setting correlates with both conceptual understanding in a low-stakes setting and with traditional long-answer problem-solving exam questions. Here, we define a "high-stakes" situation as one in which there are significant consequences resulting from the students' performance (e.g., a midterm exam); while we define a "low-stakes" situation as one in which there are few consequences (e.g., an assignment where students receive extra credit for *completing* the assessment, independent of their performance).

This study was conducted in the context of a larger study evaluating the effects of an active learning intervention to a traditionally taught (lecture-only) course [12]. In the larger study, the results of the DCI are used to evaluate the hypothesis that the intervention leads to an increase in student conceptual

978-1-5386-1174-6/18/$31.00 ©2018 IEEE

understanding of dynamics. The results of the present study will inform the validity of using the low-stakes evaluation as a means of capturing changes in conceptual understanding by considering the relationships between the low-stakes and high-stakes evaluations.

Students take the DCI at a time and place of their choosing during the week the survey is available at the beginning and end of the semester. They receive 0.5% extra credit for filling out any portion of the instrument independent of their performance. Given the leniency of these testing conditions, our hypothesis is:

> **Student performance on the low-stakes evaluation (DCI *post*) will correlate with the high-stakes evaluations (*short*-answer concept and *long*-answer problem-solving scores), but the correlations will be low.**

Researchers have previously shown that conceptual understanding can be correlated with problem-solving skills [12, 13]. Therefore, a secondary hypothesis is:

> **There will be a statistically significant correlation between the student scores on the two high-stakes performance assessments (the *short*-answer concept and *long*-answer problem-solving scores).**

As a result of these correlations, our third hypothesis is:

> **Student performance on the two high-stakes evaluations will be predictive of their performance on the low-stakes evaluation.**

II. Methods

A. Course and Participants

This study was conducted in an undergraduate introductory mechanical engineering dynamics course covering concepts in three-dimensional particle motion, planar rigid body motion, and basic vibration. This course is required for multiple engineering disciplines at a large, research-intensive public university. In a typical semester, two to three sections are offered with enrollment typically ranging from 60-120 students each. Course grades for the 15-week semester are based on student performance on 13 homework and experiment-based assignments (15%), 2 midterms (50%), and a final (35%).

B. Low-Stakes Evaluation

The DCI was offered twice during the term (during weeks 2 and 15), both times in a low-stakes situation. Both offerings were administered online without a time limitation, and students who completed any portion of the DCI received 0.5% extra credit each time for a maximum of 1.0% total extra credit. For each implementation, the overall DCI scores were calculated as the percentage of 29 questions answered correctly. For this study, we investigate student performance on the end-of-semester DCI. Fifty-seven of the 70 students enrolled in this section of the course completed the DCI at that administration.

C. High-Stakes Evaluation

A midterm examination was administered 12 weeks into the 15-week long semester. That midterm included: 1) a set of four *short*-answer concept questions related to the course concepts and 2) two traditional *long*-answer problem-solving questions. The four short-answer questions are formulated around six of the 14 concepts on the DCI.

As an example of one of the *short*-answer concept questions, students were asked to identify the relationship between angular velocities measured at two points on the same rigid body. One of the *long*-answer problem-solving questions asked students to complete kinematic analyses (e.g. velocity of a specific point) of a wheel rolling without slipping with a constant angular speed. Each *long*-answer problem-solving question was graded by the same person (either the instructor or the teaching assistant) for consistency across all students in the section.

All six questions on the midterm exam focused on concepts for the unit on rigid body dynamics, and performance on this examination contributed to 25% of the students' final course grade. This examination was offered at the conclusion of the unit on rigid body dynamics; thus, it likely corresponded to the peak in student understanding of the material.

D. Statistical Analysis

Due to the nonnormalities in the data, the Spearman correlation coefficients were calculated to evaluate the relationship between student performance on the low-stakes evaluations (DCI offered for extra credit) and the high-stakes evaluations (midterm exam). Additionally, stepwise linear regressions (backwards elimination) were conducted to determine a final model with the best fit. The predicted (outcome) variable is overall DCI score at the end of the semester (DCI *post*), and the regressors are two corresponding scores on the high-stakes midterm evaluation (*short*-answer concept scores and *long*-answer problem-solving scores). Normality and heteroscedasticity of the residuals are confirmed for the final model.

III. Results

A. Low-Stakes Evaluation

Table I documents student performance on the DCI and summarizes the low-stakes evaluation of student conceptual understanding. Reported are the mean ± standard deviation of the overall DCI scores taken at the start (*pre*) and end (*post*) of the term. The *pre* and *post* values are comparable to those documented previously in the literature [8, 10, 11], and although not included in the analysis to follow, the *pre* scores are offered for reference.

TABLE I. Low-Stakes Evaluation of Student Performance.

	Pre	***Post***
Score	40.5% ± 16.1%	48.0% ± 19.7%

B. High-Stakes Evaluation

Table II documents student performance on the midterm examination, which represents the high-stakes evaluation of student conceptual understanding. Reported are the mean ± standard deviation of the four *short*-answer concept questions and the two *long*-answer problem-solving questions used to evaluate student conceptual understanding.

TABLE II. HIGH-STAKES EVALUATION OF STUDENT PERFORMANCE.

	Short	***Long***
Score	75.3% ± 19.1%	84.5% ± 13.2%

C. Correlations

To evaluate the pairwise relationships between all three scores (*post*, *short*, and *long*), the Spearman correlation coefficients are calculated for each pair and are reported in Table III.

TABLE III. SPEARMAN CORRELATION COEFFICIENTS BETWEEN PAIRS OF LOW-STAKES AND HIGH-STAKES EVALUATION SCORES.

		High-stakes	
		Short	***Long***
High-Stakes	***Long***	0.68***	
Low-Stakes	***Post***	0.42**	0.43***

[a] *Significant at α = *0.05, **0.01, ***0.001*

D. Statistical Regression

To evaluate the relationships between the low-stakes evaluation (*post*) and the high-stakes evaluation (*long* and *short*), stepwise linear regression with backwards elimination was conducted. The final model showed the low-stakes evaluation is related most strongly with just the *long*-answer high-stakes evaluation ($F(1,56)=8.35$, $p<0.01$, $R^2=0.13$) such that scores on the *long*-answer problem-solving questions positively predicts scores on the low-stakes evaluation ($\beta=0.38$, $p<0.01$). Removing the *short*-answer concept question scores as a predictor variable did not significantly change the fit of the model ($F(1,54)=1.84$, $p=0.18$).

However, as was shown in Table III above, the *short*-answer concept question scores are in fact strongly correlated with the *long*-answer problem-solving question scores. We therefore re-conducted the regression and removed the *long*-answer scores as a predictor variable to verify the relationship between the *short*-answer concept questions and the low-stakes evaluation. The final model showed the low-stakes evaluation is also strongly related with the *short*-answer high stakes evaluation ($F(1,56)=7.81$, $p<0.01$, $R^2=0.12$) such that the scores on the *short*-answer concept questions positively predicts scores on the low-stakes evaluation ($\beta=0.37$, $p<0.01$). As expected, this relationship is nominally weaker than that between the low-stakes evaluation and the *long*-answer problem-solving question scores.

IV. DISCUSSION

A. Correlations

Our first hypothesis is supported by the moderate correlation between the *long*-answer problem-solving questions and the DCI *post* scores ($r_S=0.43$, $p<0.001$). The correlation between performance on the midterm *short*-answer concept questions and DCI *post* scores ($r_S=0.42$, $p<0.01$) indicates that conceptual understanding as evaluated by the high-stakes midterm exam is moderately correlated with conceptual understanding as evaluated by the low-stakes DCI.

The correlation between the midterm *short*-answer concept questions and the traditional *long*-answer problem-solving questions ($r_S=0.68$, $p<0.001$) confirms our second hypothesis that student performance on the concept questions is related to problem-solving skills. These findings confirm research demonstrating significant relationships between conceptual understanding and problem-solving [14, 15].

B. Statistical Regression

Students' performance on the low-stakes evaluation (*post*) was predicted by their performance on the *long*-answer problem-solving questions on the midterm evaluation. Furthermore, their low-stakes evaluation (*post*) was also predicted by their performance on the *short*-answer concept questions on the midterm evaluation. Coupled with the moderate Spearman correlation coefficient, this confirms our third hypothesis that students' performance on the low-stakes evaluation is predicted by the conceptual understanding demonstrated on the high-stakes evaluation.

C. Theoretical Frameworks

As our study is an exploratory one, we did not set out to prove an underlying theoretical framework, but there are several which shed light on our findings. The Theory of Planned Behavior is one possible theoretical framework that could explain why students' performance on the low-stakes evaluation is indicative of their performance - and therefore conceptual understanding - on a high-stakes evaluation [16]. The Theory of Planned Behavior links an individual's behaviors and beliefs, and it includes the key concept of perceived behavioral control. This concept suggests that the sense of control an individual has over a specific behavior influences their decision to engage (or not engage) in that behavior, an ideal closely related to Bandura's notion of self-efficacy [17]. Another key concept is behavioral intention, which describes one's willingness to perform a given behavior. In general, an individual is more likely to engage in behaviors that are expected to result in a positive outcome. In the context of this study, we posit that students know they will successfully earn the small extra credit award independent of their actual performance, and that makes them more willing to complete the DCI survey. Thus, they have a high degree of perceived behavioral control, and that results in an increase in behavioral intention. This supposition is supported by the fact that, for our larger, long-term study involving the administration of the DCI in 4 separate semesters, our overall response rate has been very high (88%).

Another theory that could help explain why students put forth effort on this low-stakes assessment is Eccles' Expectancy Value Theory [18]. This theory notes that students' motivation to complete a task depends on their expectation for succeeding at the task and the value they place on the task [19]. Students in this study have a very high expectation for success on the DCI since they earn the full extra credit amount by simply completing the instrument. As for value, past research has shown students' test-taking effort and performance on a 3.5 hour test that had not impact on their academic record was related to how useful and important students perceived the test to be [20]. In the context of this study, the usefulness and importance of completing the DCI could be linked with their understanding that they are contributing data for a larger research study. This hypothesis is supported by the implication from the results presented by Cole, Bergin, & Whittaker [21] that students who perceive high usefulness and importance of a low-stakes evaluation will put forth more effort if that usefulness and importance is communicated to them by the administrators of the evaluation. Before they fill out the DCI, and both times an active learning intervention was implemented in the class as part of the larger study, the students were reminded of the fact that their performance on the DCI will inform the results of the research project.

V. Conclusion

The synergy between the low-stakes DCI evaluation and the high-stakes midterm examination evaluation indicates the DCI is a reliable measure for quantifying student conceptual understanding, even under low-stakes conditions. These moderate correlations suggest students are exhibiting comparable levels of effort and value between the two forms of evaluation, which supports the hypothesis that the low-stakes DCI evaluation is a reliable portrayal of student conceptual understanding.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Award ID No. 1609204 through NSF's Division of Undergraduate Education (DUE) as a part of the Improving Undergraduate STEM Education (IUSE) program. We thank the instructor and students for their participation.

References

[1] D.F. Treagust, "Development and use of diagnostic tests to evaluate students' misconceptions in science," *Intl. J of. Sci. Educ.*, 10, pp. 159–169, 1988.

[2] G. W. Fulmer, H.-E. Chu, D. F. Treagust, K. Neumann, "Is it harder to know or to reason? Analyzing two-tier science assessment items using the Rasch measurement model," *Asia Pac. Sci. Educ.,* 1:1, 2015.

[3] R. J. Osborne and J. K. Gilbert, "A technique for exploring the students' view of the world," *Phys. Educ.*, 50, pp. 376-379, 1980.

[4] M. Watts, "Exploring pupils' alternative frameworks using the interview-about-instances method," in *Proc. of the Intl. Workshop on Problems Concerning Students' Representation of Phys. and Chem. Knowledge*, Ludwigsburg, West Germany, pp. 365-386, 1981.

[5] Nelson M. A., "Oral Assessments: Improving Retention, Grades, and Understanding," in *PRIMUS*, 21(1), pp. 47-61, 2011.

[6] P. Tamir, "An alternative approach to the construction of multiple choice test items," *J. of Bio. Educ*., 5, pp. 305-307, 1971.

[7] Kirbulut Z. D. and Geban O., "Using three-tier diagnostic test to assess students' misconceptions of states of matter," in *Eurasia J. Math. Sci. Technol. Ed.*, 10(5), pp. 509-521, 2014.

[8] N. Jorion, B. D. Gane, K. James, L. Schroeder, L. V. DiBello, J. Pellegrino, "An analytic framework for evaluating the validity of concept inventory claims," *J. of Eng. Educ*., 104(4), pp. 454-496, 2015.

[9] B. Finn, "Measuring motivation in low-stakes assessments," Research Report (RR-15-19), Educational Testing Service, Princeton, NJ, 2015.

[10] G. L. Gray, F. Costanzo, D. Evans, P. Cornwell, B. Self, J. L. Lane, "The dynamics concept inventory assessment test: a progress report and some results," in *Proc. of American Society of Engineering Education, Annual Conference*, Portland, OR, pp. 4819-4833, 2005.

[11] N. Jorion, B. Self, K. James, L. Schroeder, L. V. DiBello, J. Pellegrino, "Classical test theory analysis of the dynamics concept inventory," in *Proc. of the 2013 American Society of Engineering Education, Annual Conference,* Riverside, CA, 2013.

[12] N. Stites, D. A. Evenhouse, M. Tafur, C. M. Krousgrill, C. Zywicki, A. N. Zissimopoulos, D. B. Nelson, J. DeBoer, J. F. Rhoads, E. J. Berger, "Analyzing an abbreviated dynamics concept inventory and its role as an instrument for assessing emergent learning pedagogies," in *Proc. of the American Society of Engineering Education, Annual Conference*, New Orleans, LA, 2016.

[13] R. V. Vitali, N. C., Perkins, C. J. Finelli, "Introduction and Assessment of i-Newton for the Engaged Learning of Engineering Dynamics," in *Proc. Of the American Society of Engineering Education, Annual Conference,* Salt Lake City, UT, 2018.

[14] S. Ates and E. Catologlu, "The effects of students' cognitive styles on conceptual understandings and problem-solving skills in introductory mechancis," *Res. Sci. Technol. Educ*., 25(2), pp. 167-178, 2007.

[15] K. L. Malone, "Correlations among knowledge structures, force concept inventory, and problem-solving behaviors," *Phys. Rev. Spec. Top. Educ. Res.*, 4(2), 020107, 2008.

[16] I. Ajzen, "The theory of planned behavior," *Organ. Behav. Hum. Decis. Process*, 50(2), pp. 179-211, 1991.

[17] A. Bandura, "Self-efficacy mechanism in human agency," *Am. Psychol*., 37(2), pp. 122-147, 1982.

[18] A. Wigfield and J. S. Eccles, "Expectancy-Value Theory of Achievement Motivation," *Contemporary Educational Psychology*, 25, pp. 68-81, 2000.

[19] J. S. Eccles and A. Wigfield, "Motivational beliefs, values, and goals," *Annu. Rev. of Psychol*., 53, pp. 109-132, 2002.

[20] S. J. Osterlind, R. D. Robinson, and N. M. Nickens, "Relationship between collegians' perceived knowledge and congeneric tested achievement in general education," *J. of Coll. Stud. Dev*., 38(3), pp. 255–265, 1997.

[21] J. S. Cole, D. A. Bergin, and T. A. Whittaker, "Predicting student achievement for low stakes tests with effort and task value," Contemporary Educational Psychology, 33(4), pp. 609-624, 2008.