Performance and Participation Differences for In-Class and Online Administration of Low-Stakes Research-Based Assessments

Ben Van Dusen[1], Jayson Nissen[1], Manher Jariwala[2], & Eleanor Close[3]
[1]California State University Chico, Science Education
[2]Boston University, Physics
[3]Texas State San Marcos, Physics

Research-based assessments (RBAs), such as the Calculus Concept Inventory, have played central roles in many course transformations from traditional lecture-based instruction to research-based teaching methods. In order to support instructors in assessing their courses, the online Learning About STEM Student Outcomes (LASSO) platform simplifies administering, scoring, and interpreting RBAs. Reducing barriers to using RBAs will support more instructors assessing the efficacy of their courses and transforming their courses to improve student outcomes. The purpose of this study was to investigate the extent to which RBAs administered online and outside of class with the LASSO platform provided equivalent quantity and quality of data to traditional paper and pencil tests administered in class for both student performance and participation. We used an experimental design to investigate the differences between these two test modes. Results indicated that the LASSO platform can provide equivalent quantity and quality of data to paper and pencil tests.

Keywords: Assessment, quantitative methods, technology

Introduction

Research-Based Assessments (RBAs), such as the Calculus Concept Inventory (Epstein, 2007), are often used to both develop and disseminate research-based teaching methods that improve student outcomes. Subsequently, RBAs are the focus of many influential publications in physics education research, such as Hake's (1998) comparison of traditional and interactive-engagement courses. The large increase in the number of RBAs in physics education research coincided with a dramatic increase in the collaboration in the PER community (Sayre et al., 2017). Because of these successes, many educators are interested in using RBAs. Madsen et al. (2016), however, found that many instructors want support in choosing appropriate assessments, administering and scoring the assessments, and interpreting the results of their assessments. To address these needs the Learning Assistant Alliance developed the LASSO platform to host and administer RBAs online (LA Alliance, 2017). Hosting the RBAs online meets instructors' needs by allowing for the tests to be administered outside of class, to be promptly and automatically scored, and for instructors to be provided with a summary report to help interpret the results.

Extensive research has investigated the differences between computer based tests (CBTs) and pencil and paper tests (PPTs). Meta-analysis of the literature has revealed that there is no systematic difference in scores between these two modes of administering tests (Wang et al., 2007). However, the studies in these meta analyses were conducted using high-stakes standardized tests at the K-12 level, and most had the CBT being administered in class. Because the LASSO platform is designed to administer RBAs outside of class in order to free up class time, the results of this earlier work may not apply to the LASSO platform.

In a similar study to this one, Bonham (2008) conducted research in college astronomy courses and administered assessments online outside of class. Bonham and colleagues had students complete both a locally-made concept inventory and a research-based attitude survey.

The students were randomly assigned to two conditions with either the concept inventory done in class and the attitude survey done outside of class via an online system or the reverse. A matched sample was then drawn from the students who completed the surveys. They concluded that there was no significant difference between CBT and PPT data collection. In contrast to their findings, a close analysis of their results revealed that there was a small but meaningful difference in the data and that the study did not have a sufficient sample size to rule out any meaningful differences; their study was underpowered. Their results indicated that the online concept inventory scores were 6% higher than the in class scores, which was an effect size of approximately 0.30. While this is a small difference, lecture-based courses often have raw gains below 20% and a 6% difference would skew comparisons between data collected with CBT and PPT modes. Therefore, it is not clear from the prior literature that low-stakes tests provide similar data when collected in class with PPTs compared to outside of class with CBTs.

Research Questions

The purpose of the present study was to inform if data collected with LASSO is consistently different than data collected with paper tests. In pursuit of this purpose we asked:

(1) To what extent does the online administration of RBAs outside of class using the LASSO platform provide comparable data to the in-class administration of RBAs using PPTs? (2) How do instructor administration practices impact participation rates for low-stakes RBAs, if at all? (3) How are student course grades related to participation rates for low-stakes RBAs, if at all?

If the LASSO platform provided equivalent data to paper based administration, then the LASSO platform represents a much simpler entry point for instructors to begin assessing and transforming their own courses because it addresses many of the instructors' needs that Madsen et al. (2016) identified. A second major benefit of the widespread use of the LASSO system is that it automatically aggregates all of the data and makes this data available for research. The size and variety of this data allows for investigations that would have been underpowered if conducted at only a few institutions or lacking generalizability if only conducted in a few courses at a single institution.

Methods

The data was collected at a large regional Hispanic-serving university across two semesters in three different introductory physics courses: algebra-based mechanics, calculus-based mechanics, and calculus-based electricity and magnetism (E&M).

The study used a between-groups experimental design (Figure 1). Stratified random sampling created two groups within each section with similar representations across student gender, race, and honors status. One group completed a concept inventory (either the Force Concept Inventory [FCI] or Conceptual Survey of Electricity and Magnetism [CSEM]) online outside of class using the LASSO platform and an attitudinal survey (the Colorado Learning Attitudes about Science Survey [CLASS]) in class using paper and pencil. The other group completed the concept inventory in class and the attitude survey online outside of class. Both conditions were repeated at the beginning and end of the semester. Paper and pencil assessments were collected by the instructors, scanned using automated equipment, and uploaded to the LASSO platform. Student assessment data was downloaded from the LASSO platform and combined with student grades and demographic data provided by the university. The data analysis did not include students who joined the class late, dropped, or withdrew, leaving a total sample of 1,310 students in 25 course sections.

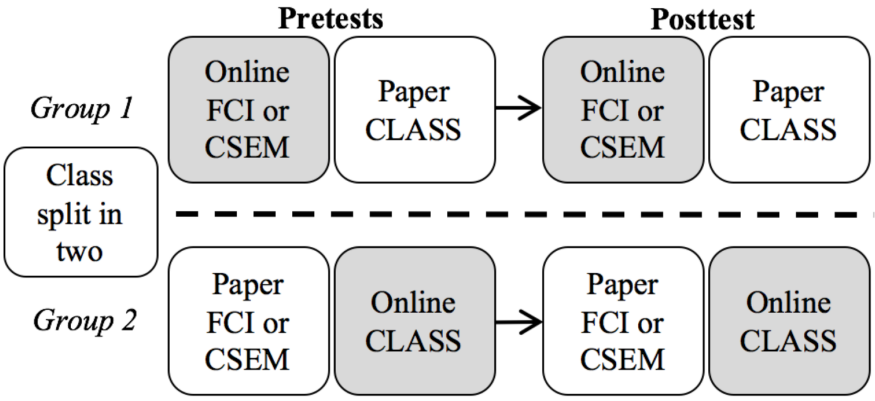

Figure 1. Design of the research conditions.

At the end of each semester of data collection participating faculty were interviewed to identify how the faculty motivated their students to complete the CBT. Four different practices were identified that we will refer to as *recommended practices*: 1) email reminders, 2) in class announcements, 3) participation credit for the pretest, and 4) participation credit for the posttest.

We used the HLM 7 software package to create multi-level models to analyze the performance and participation data. We analyzed the performance data for the concept inventories using 2-level Hierarchical Linear Models: test conditions (level 1) were nested within course types (level 2), no covariates were used. We analyzed the participation data using 3-level Hierarchical Generalized Linear Models: assessments (Level 1) were nested within students (level 2) nested within either course sections (level 3), the number of recommended practices and students grades in the courses were used as covariates.

The final models for performance and participation consisted of posttest score or participation as the outcome variables. The models were built in 3 or 4 steps: (1) no predictors, (2) add level 1 predictors, (3) add level 2 predictors, (4) add level 3 predictors (if applicable). This four-step process informed how much additional information was being explained by the addition of the new predictors in each step as indicated by a reduction in the variance for that variable.

Completion rates for the PPT condition were 94% for the pretest and 74% for the posttest and for the CBT were 68% for the pretest and 54% for the posttest. For the performance analysis, missing concept inventory data (i.e. students who did not take either the pre or posttest) was replaced using Hierarchical Multiple Imputation (HMI) with the MICE package in R. HMI is a form of multiple imputation (MI) that takes into account the fact that students were nested in different courses and that their performance may have been related to the course they were in. MI addresses missing data by (1) imputing the missing data m times to create m complete data sets, (2) analyze each data set independently, and (3) combine the m results using standardized methods (Dong & Peng, 2013). Our MI produced m=10 complete data sets. Multiple imputation is preferable to list-wise deletion because it maximizes the statistical power of the study and has the same basic assumptions.

Findings

Performance

The model of student performance on concept inventories showed very little differences in either pretest or posttest performance across test conditions. The largest predicted effect of test

condition on student performance was on posttest for E&M students (Figure 2). This predicted effect bordered on being large enough to be meaningful because it indicated a 2.2 points higher posttest score for students doing the CBT and the overall predicted gain for the E&M students was only 11.6 points. However, the pre- and posttest across the three courses created six total measurements of the predicted effect for test condition; in three of those measurements the effect was nearly zero, in one it was positive, and in two it was negative. In addition to these inconsistencies in all six comparisons across condition there was large overlap in the 95% confidence intervals, indicating that the differences were not statistically reliable. Examination of the model variances showed that the inclusion of test conditions led to larger variances, indicating that conditions were not a reliable predictor of student performance.

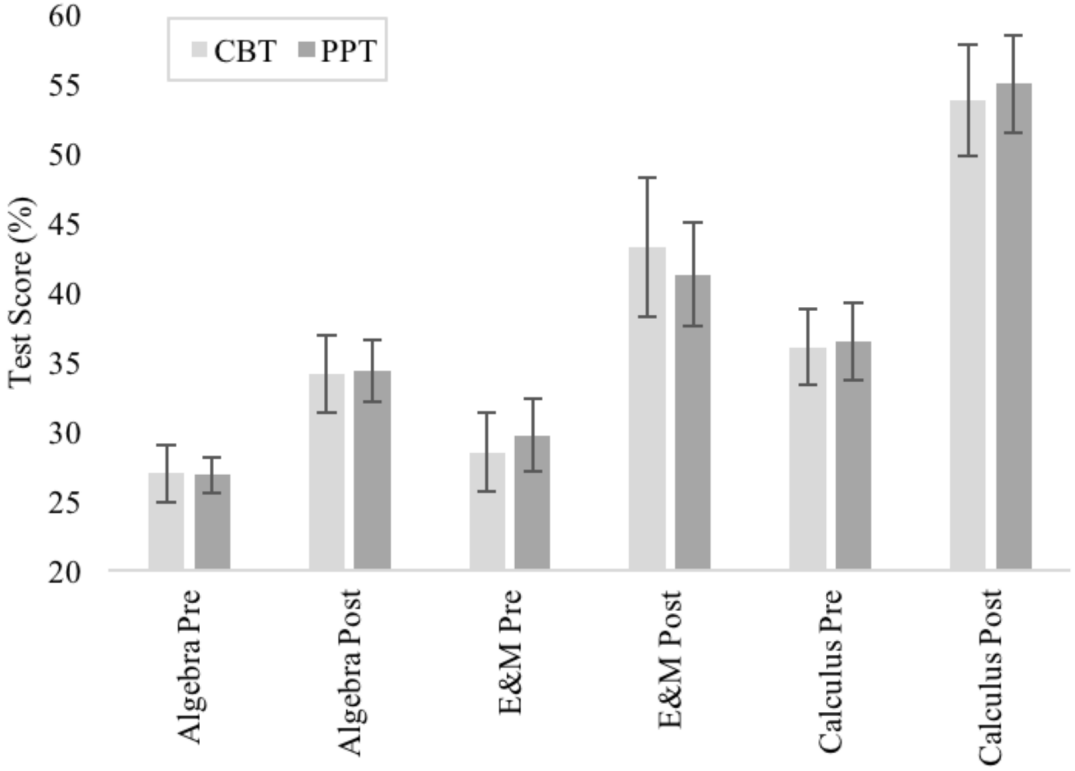

Figure 2. Predicted Mean Scores with 95% CIs.

Participation

The results of our HGLM model of the student data, indicate that the more recommended practices instructors used, the higher the participation rates were for their CBT assessments. Student course grades were also a statistically reliable predictor of student participation.

Figure 3 illustrates the predicted student participation rate based on student course grades and the number of recommended practices that instructors used. In terms of data collection, the posttests represented the limiting case as predicted participation rates on the posttests for both the PPT and CBT were lower than on the pretests. With the exception of the PPT pretest there was a large difference in predicted participation based on course grades. The number of recommended practices that instructors used dramatically increased predicted participation rates such that when instructors implemented all four recommended practices the participation rates of the CBT and PPT posttest were very similar. The impact of recommended instructor practices on predicted

participation rates occurred for all students, but was largest for high achieving students. Relationships between student participation, grades, and instructor practices on the CBT pretest were similar to those on the CBT posttest. These results indicated that similar participation rates to those on PPT can be achieved via CBT when instructors use all four recommended practices.

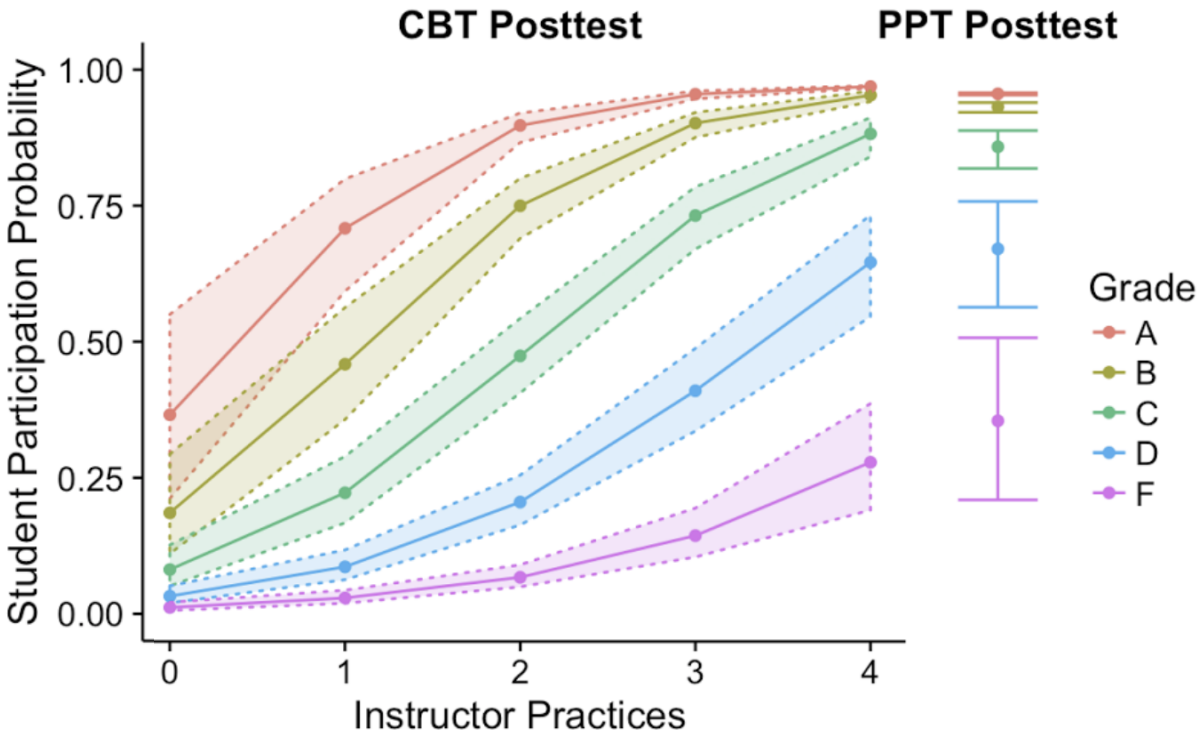

Figure 3. Predicted student participation rates with 95% CIs. Only posttest predictions are shown as it is the test with the lower participation rates and is the primary limiter for data collection.

Conclusion and Implications

Our study shows that CBT and PPT administrations of low-stakes assessments can lead to similar student performance and participation. This similarity indicates that when our recommended practices are implemented instructors and researchers can use online systems, such as the LASSO platform, to collect valuable information about the impacts of their courses that is comparable to prior research that was collected with paper and pencil tests. Collecting data with the LASSO system can also greatly reduce the barriers to instructor's use of RBAs since instructors do not need to dedicate class time to collect the data or their own time to sort, scan, and analyze the data. It is important to note, however, that instructors do need to make some effort to motivate their students to complete the online assessments. We have found that by making announcements in class, sending out email reminders, and giving credit to students who complete the RBAs instructors can achieve similar participation rates on CBT assessments as on PPT assessments. Our hope is that reducing the barriers to using RBAs use will lead more instructors to assess the efficacy of their courses and, subsequently, to adopt research-based teaching practices that support student success.

In addition to promoting the use of RBA's developed by the DBDR community, the LASSO platform anonymizes, aggregates, and makes its database available to researchers with

appropriate IRB protocols. The LASSO database has already provided multi-level large-scale data to examine questions of equity in student outcomes (Van Dusen & Nissen, in press), effects of near-peer mentors on student outcomes (White et al., 2016), and effects of instructor experience on their effectiveness (Caravez, in press). As the LASSO dataset grows, it will allow the DBER community the ability to quickly access a dataset designed to support the investigation of student outcomes from across the country.

While these findings are generally encouraging, there are several unexamined factors that could strengthen the conclusions and generalizability of the work. Useful areas for future research includes: (1) examining the associations between student demographics and student participation and performance in CBT and PPT conditions, (2) comparing student performance at the item-level (rather than total score) on CBT and PPT conditions, and (3) replicating the study in diverse institutional settings.

References

Bonham, S. (2008). Reliability, compliance, and security in web-based course assessments. *Physical Review Special Topics-Physics Education Research*, *4*(1).

Caravez, D., De La Torre, A., Nissen, J., & Van Dusen, B. (In Press) Longitudinal Associations between Learning Assistants and Instructor Effectiveness, *Proc. 2017 Physics Education Research Conference*.

Dong, Y., & Peng, C. Y. J. (2013). Principled missing data methods for researchers. *SpringerPlus*, *2*(1), 222.

Epstein, J. (2007, September). Development and validation of the Calculus Concept Inventory. In *Proceedings of the ninth international conference on mathematics education in a global community* (Vol. 9, pp. 165-170). Charlotte, NC.

Hake, R. R. (1998). Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses, *American Journal of Physics*, 64–74.

LA Alliance (2017). https://www.learningassistantalliance.org/

Madsen, A., Mckagan, S. B., Martinuk, M. S., Bell, A., & Sayre, E. C. (2016). Research-based assessment affordances and constraints: Perceptions of physics faculty. *Physical Review Physics Education Research*, *12*, 1–16.

Van Dusen, B. and Nissen, J. (In Press) Systemic Inequities in Introductory Physics Courses: the Impacts of Learning Assistants, *Proc. 2017 Physics Education Research Conference*.

Wang, S., Jiao, H., Young, M. J., Brooks, T., & Olson, J. (2007). A Meta-Analysis of Testing Mode Effects in Grade K-12 Mathematics Tests. *Educational and Psychological Measurement*, *67*(2), 219–238. https://doi.org/10.1177/0013164406288166

White, J.S.S., Van Dusen, B., & Roualdes, E. (2016, December). The Impacts of Learning Assistants on Student Learning of Physics. In D. L. Jones, L. Ding, & A. Traxler (Eds.), *Physics Education Research Conference Proceedings* (pp. 384–387). http://doi.org/10.1119/perc.2016.pr.091