Providing Context for Identifying Effective Introductory Mechanics Courses

Jayson M. Nissen, Nissen Education Research and Design, Corvallis, OR
Ian Her Many Horses, University of Colorado Boulder, Boulder, CO
Ben Van Dusen, Iowa State University, Ames, IA
Manher Jariwala, Boston University, Boston, MA
Eleanor Close, Texas State University San Marcos, San Marcos, TX

Research-based assessments (RBAs) measure how well a course achieves discipline-specific outcomes. Educators can use outcomes from RBAs to guide instructional choices and to request resources to implement and sustain instructional transformations. One challenge for using RBAs, however, is a lack of comparative data, particularly given the skew in the research literature toward calculus-based courses at highly selective institutions.[1] In this article, we provide a large-scale dataset and several tools educators in introductory physics courses can use to inform how well their courses foster student conceptual understanding of Newtonian physics. The supplemental materials[2] include this dataset and these tools. Educators and administrators will often target courses with high drop, withdrawal, and failure rates for transformations to student-

centered instructional strategies. RBAs and the comparative tools presented herein allow educators to address critiques that the course transformations made the courses "easier" by showing that the transformed course supported physics learning compared to similar courses at other institutions. Educators can also use the tools to track course efficacy over time.

The supplemental material[2] includes the dataset and analysis code both as an Excel file and an R file for readers to create their own visualizations to understand and communicate one aspect of what is occurring in their courses. The dataset consists of courses from the online Learning About STEM Student Outcomes (LASSO) platform and from the literature that used either the Force Concept Inventory (FCI) or the Force and Motion Conceptual Evaluation (FMCE). The LASSO platform administers and scores a wide range of RBAs online, and we describe LASSO further in the "Methods" section. As we will show, the LASSO dataset is less skewed toward courses with high pretest scores than the literature, thereby it provides a more representative sample of course efficacy. The tools provided include a scatterplot of pretest and posttest scores for all courses in the dataset that can show how a course compares to other courses before and after instruction and a distribution of the effect sizes for the shifts on conceptual understanding for all courses. Educators can use the tools in the supplemental material to create visualizations comparing their target courses to the entire dataset.

Research-based assessments (RBAs) measure the effectiveness of a course in achieving discipline-specific, content-based learning outcomes or positive attitudinal shifts—for example, an instructor can administer the FCI prior to and after instruction to measure shifts in student understanding of Newtonian physics. For educators implementing new research-based instructional strategies, RBAs can provide evidence of their impact on student outcomes. Measuring this impact can show students the value of engaging in new pedagogies that may differ from their expectations, can help convince colleagues to adapt or improve their teaching practices systematically, and can persuade administrators to devote institutional resources to sustain or expand course transformations. On a larger scale, RBAs can aid in assessment of learning outcomes at the departmental or academic unit level and inform the development of instructional methods and materials.

One major benefit of RBAs over traditional assessments (e.g., a course final exam) is the ability to compare results to data from other educators and institutions. Comparisons with national datasets can provide a context for interpreting results from your own classroom. This context is important because these concept inventories can challenge students despite formal instruction "covering" the material. For example, raw gains from pretest to posttest on the FCI and FMCE are approximately 15% in lecture-based courses and 20% in transformed courses.[3] To make a case either for the need for additional support (if student learning is low) or for continued funding (e.g., for a learning assistant program facilitating interactive instruction), it is important to contextualize data to demonstrate need or effectiveness.

Many RBAs have normative data available in the literature, either in publications introducing the instrument or in literature reviews. While these publications provide useful comparisons, the student populations represented in them are typically from research-intensive, selective universities whose students are less diverse and better prepared mathematically than are average students[1] and therefore may not be generalizable to more diverse student populations and institutional contexts. Making the matter worse, many of the studies report their gains using normalized gain, a biased measure that favors high pretest score groups.[4] To address these limitations in the literature, we used data from both the literature and the online LASSO platform's multi-institution database. The LASSO data were almost exclusively collected online, whereas only two of the 17 studies from the literature reported some online data collection. Multiple studies indicate online administration and in-class administration of RBAs produces similar results.[5-7]

Methods

Data collection

LASSO is an online platform hosted on the Learning Assistant Alliance website (https://learningassistantalliance.org). LASSO provides course instructors with the ability to collect

DOI: 10.1119/5.0023763

student assessment data online through a process designed to reduce the instructor and class time to give an assessment and to use the collected data. LASSO collects two levels of data: course-level data and student data. Instructors provide course-level data when setting up the assessment by describing the course context and teaching method, and providing a student roster with emails. Student-level data include demographic data, assessment metadata (such as time taken on the assessment), and responses to the individual assessment questions. Students complete the assessment via a website link sent to their email. LASSO processes the student responses to provide instructors with a report that includes summative statistics and a histogram of student performance. The reports are being updated to include the visualizations presented in this publication. For students who agreed to share their data for research, the platform anonymizes the data and makes it available to researchers.

To expand the dataset beyond the LASSO database, we collected descriptive statistics from studies using either the FMCE or FCI. We searched journals that typically publish physics education research (PER): *Physical Review Physics Education Research*, the *PER Conference Proceedings*, the *American Journal of Physics*, and *The Physics Teacher* for relevant articles. We included all 17 articles that reported pretest and posttest mean scores. Whenever a study reported data for individual courses, we included the data for each course. Several studies only reported statistics that combined courses, which is why the literature had more students but fewer courses than LASSO. The data from the literature do not overlap with the LASSO data. We excluded studies that only reported normalized gain because our purpose was to look at the relationships between pretest and posttest means. The supplementary material includes an Excel document with all the data used in the study.[2] The dataset includes 309 courses: 202 courses from the LASSO platform for 12,879 students and 107 courses from the literature for 23,882 students.

Multiple imputation

We handled missing data and the uncertainty it introduces using multiple imputation (MI).[8] In statistics, imputation replaces missing values with probable values. In PER studies using pretests and posttests, researchers often handle missing data by removing every student that does not have both a pretest and a posttest, a procedure known as complete case analysis. Statisticians do not recommend this practice because it often biases the results and instead refer to MI as the gold standard for handling missing data.[8] MI allows analyses to use all the data by imputing the missing values multiple times to create multiple complete datasets and then combining the results from all of those analyses to produce unbiased estimates with accurate uncertainties. Nissen, Donatello, and Van Dusen[9] showed that MI produces more accurate results than complete case analysis with RBA data. Schafer[8] discusses MI in detail.

For the LASSO data, we filtered the data to remove students that did not take the test seriously by only including tests where students took more than five minutes and completed more than 80% of the test. We then used MI to impute missing data for individual student scores and calculated a single dataset of course averages from the multiple imputed datasets. We then combined the LASSO data with the data from the literature. Several of the studies in the literature were missing either pretest or posttest means or standard deviations. We used MI to impute these missing data points and averaged the multiple imputed datasets into a single dataset. Averaging multiple imputed datasets is not the best practice as it creates artificially small error bars; however, it served our purpose of being able to share data and resources that any instructor could use to interpret the results in their courses.

Multiple linear regression

To understand what variables were important for looking at the relationships between pretest scores and posttest scores, we built multiple linear regressions. Our regression models predicted class mean posttest scores (*posttest mean*) using *pretest mean, pretest mean squared, test, course type*, and *instruction type*. *Pretest mean* is the course's mean pretest score. Pretest mean acted as a proxy for the courses' student population's prior physics knowledge and opportunities. We included pretest mean because prior performance is consistently the best predictor of future performance. Including it helped to account for systemic differences between course types and institution types. We included *pretest mean squared* to account for any non-linearity in the model that could occur from courses with either very low (floor effects) or very high (ceiling effects) means. For example, as pretest scores get closer to 100% gains will decrease; *pretest mean squared* allowed the overall relationship between *posttest mean* and *pretest mean* to curve and account for both ceiling and floor effects. This curved relationship occurs in Fig. 2. *Test* informed if the relationships differed for the FCI and FMCE. *Course type* differentiated between algebra- and calculus-based mechanics courses. Instruction type differentiated between lecture-based and interactive engagement (IE) instruction. The multiple linear regressions produced models using every combination of the predictor variables to predict posttest means. We used Akaike information criterion corrected (AICc)[10] scores to identify which model provided the most information and to remove any redundant or uninformative variables.

Effect size

We analyzed the course-level gains from pretest to posttest using Cohen's *d* with Hedges' correction as a measure of effect sizes, given in Eq. (1). Cohen's *d* calculates how many standard deviations the mean score for a course shifted and is a very common metric for measuring the efficacy of a course or intervention in the education literature.[11] In addition to its prevalence, we used Cohen's *d* because it handles both ceiling and floor effects by using standard deviation in the denominator. When course means fall near the floor or ceiling of a measurement, the individual scores will cluster more tightly and have a smaller standard deviation. Hedges' correction ad-

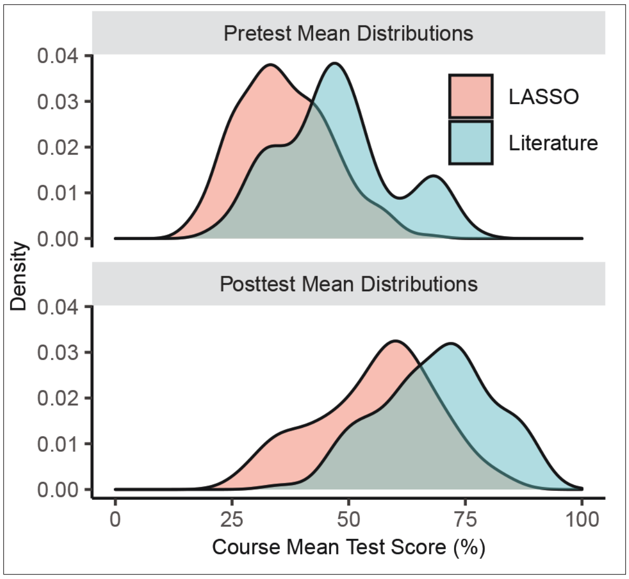

Fig. 1. Distribution of mean pretest and posttest scores in LASSO and the literature, showing that LASSO represents a different distribution of courses than those in the published literature.

dresses bias toward larger effect sizes when analyzing smaller sample sizes (N), which in this case was the number of students in a course. The supplemental materials[2] include a tool to conduct these calculations.

$$d = \frac{\bar{x}_{\text{post}} - \bar{x}_{\text{pre}}}{\left(sd_{\text{post}} + sd_{\text{pre}}\right)/2}\left(\frac{N-3}{N-1.25}\right)\sqrt{\frac{N-2}{N}}. \qquad (1)$$

Findings

Comparing LASSO and the literature

The distributions of scores indicated that the 207 courses in the LASSO data included more courses with lower pretest scores and fewer courses with very high pretest scores than the 102 courses from the literature. We illustrate these differences in Fig. 1. Combining the LASSO and literature data provides a larger and broader sample of courses for comparison.

Statistical relationships between pretest and posttest scores

Our final linear regression model, shown in Eq. (2), indicated that all the variables other than test type (*test*) were important to include. This model indicates that gains in the average course with a pretest near either the bottom or the top of the data distribution (20 or 65 percentage points, respectively) will be around 13 percentage points, while courses with a pretest in the middle of the distribution (near 45 percentage points) will have gains around 18 percentage points (see Fig. 2). Algebra-based courses tended to have larger gains by about 2.4 percentage points, and courses that used IE instruction tended to have 4.1 percentage point larger gains, which represents a 22% to 31% increase in learning in those courses and aligns with prior research on IE instruction.[12,13]

$$x_{post} = 0.65 + 1.81^{*}\, x_{pre} - 0.00947^{*}(\, x_{pre})^{2} + 4.05 * IE + 2.41 * Algebra. \qquad (2)$$

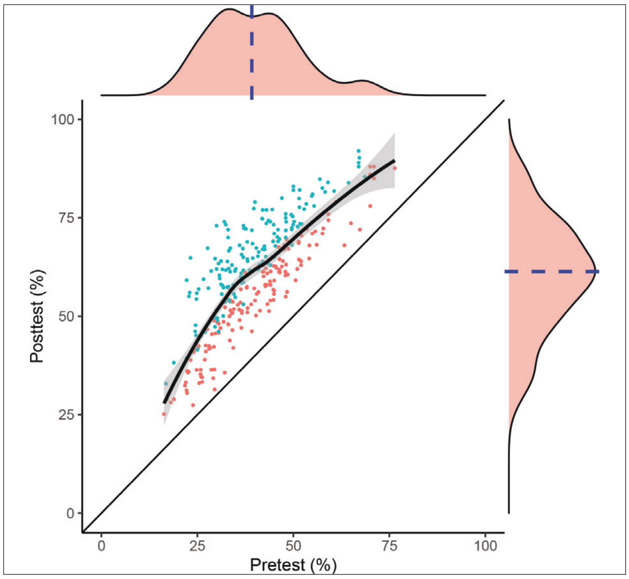

Fig. 2. Scatterplot of mean pretest and posttest scores for all 309 courses. The scatterplot includes a LOESS line of best fit with 95% confidence intervals shown in gray. The confidence intervals describe the certainty of the LOESS fit line. The colors differentiate between courses with effect sizes above and below the median value. The density plots represent the distribution of mean pretest and posttest scores with the median indicated by the blue dashed line.

Plot of data for educators

Figure 2 plots pretest and posttest scores for all 309 courses. Instructors can use this plot to inform how their courses compare to the courses in our dataset. The supplemental files[2] include an Excel file and R code to generate the plots with one's own courses overlaid on the plots. The black line on the scatterplot is the LOESS line of best fit, and it shows the average posttest value for any given pretest. The further a course falls from the black fit line on the scatterplot, the more exceptional the course. The shaded region around the black line represents the uncertainty in the estimate with 95% confidence intervals and shows how uncertainty increases due to the smaller number of courses near both ends of the fit line. The density plots show the spread of the data for pretest and posttest scores. The blue dashed lines on the density plots represent the median values. Educators using the plot should keep in mind that it may take several data points to get a consistent picture of course outcomes. If an instructor's or a department's courses tend to show the same picture (above, below, or average), then they can act on that information. Instructors seeking to improve their instruction can find resources at the PhysPort and Learning Assistant Alliance websites.

Effect size as an additional metric for looking at instruction

We also used effect size to investigate the effectiveness of courses and provide a second method for contextualizing course outcomes. Figure 3 shows the distribution of effect sizes for all 309 courses with the median value of 1.03 standard deviations. In addition to plotting their course pretest and posttest scores, educators can interpret the effectiveness

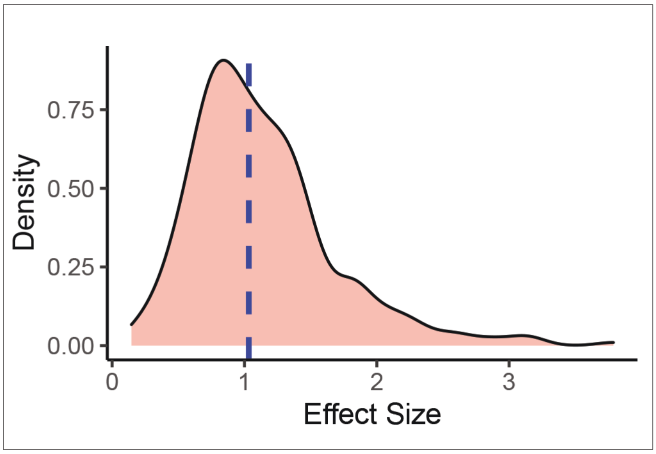

Fig. 3. A density plot of the effect sizes with a dashed blue median line showing that a typical effect size is very near 1. This density plot shows the probability of different effect sizes; the area under the curve integrates to 1.

of their instruction using a rule of thumb of an effect size of 1.0 as being average on the FCI and FMCE. Figure 2 shows consistency between effect size and the line of best fit for the pretest and posttest scores as courses with effect sizes greater than the median tended to have test scores above the line of best fit.

Discussion

Instructors interested in interpreting their course's gains on RBAs have lacked a large dataset for comparison. Normative measures in the literature skew toward courses with high pretest scores at research-intensive institutions and are limited by the common reporting of normalized gain rather than descriptive statistics and effect size measures with statistical foundations. We overcame these issues through a combination of published studies and data collected online through the LASSO platform.

Using data from the literature and the LASSO platform, we built a dataset and visualizations of FCI and FMCE scores instructors can use to inform how their courses compare to normative data across a broad spectrum of student preparation levels. These tools allow instructors to compare their courses' pretest scores, posttest scores, gains, and effect sizes to the larger dataset. For educators interested in a simpler metric than a scatterplot, we recommend using an effect size measure. We found that the median effect size was approximately 1. While *d* equals 1 is a good rule of thumb for FCI and FMCE scores, educators and researchers should not apply it to other instruments without further investigations.

For educators who wish to measure and improve their instruction, online resources have made it easier than ever to examine student outcomes (e.g., using LASSO) and find pedagogical practices to improve them (e.g., using PhysPort or the Learning Assistant Alliance). The LASSO platform on the website for the Learning Assistant Alliance hosts a Shiny App that can generate these figures.

References:

1. Stephen Kanim and Ximena C. Cid, "The demographics of physics education research," *Phys. Rev. Phys. Educ. Res.* **16**, 1–17 (2020).
2. Readers may view these materials at *TPT Online*, http://10.1119/5.0023763, under the Supplemental tab.
3. Xochith Herrera, Jayson M. Nissen, and Benjamin Van Dusen, "Student outcomes across collaborative learning environments," *Proc. 2018 Phys. Educ. Res. Conf.* 1-4 (2018).
4. Jayson M. Nissen, Robert M. Talbot, Amreen Nasim Thompson, and Ben Van Dusen, "Comparison of normalized gain and Cohen's d for analyzing gains on concept inventories," *Phys. Rev. Phys. Educ. Res.* **14**, 1-12 (2018).
5. Bethany R. Wilcox and Steven J. Pollock, "Investigating students' behavior and performance in online conceptual assessment," *Phys. Rev. Phys. Educ. Res.* **15**, 1-10 (2019).
6. Jayson M. Nissen, Manher Jariwala, Eleanor W. Close, and Ben Van Dusen, "Participation and performance on paper-and computer-based low-stakes assessments," *Int. J. STEM Educ.* **5**, 1-17 (2018).
7. Scott Bonham, "Reliability, compliance, and security in web-based course assessments," *Phys. Rev. ST Phys. Educ. Res.* **4**, 1-8 (2008).
8. Joseph L. Schafer, "Multiple imputation: A primer," *Stat. Methods Med. Res.* **8**, 3-15 (1999).
9. Jayson M. Nissen, Robin Donatello, and Ben Van Dusen, "Missing data and bias in physics education research: A case for using multiple imputation," *Phys. Rev. Phys. Educ. Res.* **15**, 1-15 (2019).
10. Y. Sakamoto, M. Ishiguro, and G. Kitagawa, *Akaike Information Criterion Statistics* (D. Reidel, Dordrecht, The Netherlands, 1986).
11. Matthew A. Kraft, "Interpreting effect sizes of education interventions," *Educ. Res.* **49** (4), 241–253 (2020).
12. Scott Freeman, Sarah L. Eddy, Miles McDonough, Michelle K. Smith, Nnadozie Okoroafor, Hannah Jordt, and Mary Pat Wenderoth, "Active learning increases student performance in science, engineering, and mathematics," *PNAS* **111** (23), 8410–8415 (2014).
13. Richard R. Hake, "Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses," *Am. J. Phys.* **66**, 64–74 (Jan. 1998).

Nissen Education Research and Design, Corvallis, OR; jayson.nissen@gmail.com