Impacts of Curricular Change: Implications from 8 Years of Data in Introductory Physics

Steven J. Pollock and Noah Finkelstein

Department of Physics, University of Colorado, Boulder, Colorado 80309 USA)

Abstract. Introductory calculus-based physics classes at the University of Colorado Boulder were significantly transformed beginning in 2004. They now regularly include: interactive engagement using clickers in large lecture settings, *Tutorials in Introductory Physics* with use of undergraduate Learning Assistants in recitation sections, and a staffed help-room setting where students work on personalized CAPA homework. We compile and summarize conceptual (FMCE and BEMA) pre- and post-data from over 9,000 unique students after 16 semesters of both Physics 1 and 2. Within a single institution with stable pre-test scores, we reproduce results of Hake's 1998 study that demonstrate the positive impacts of interactive engagement on student performance. We link the degree of faculty's use of interactive engagement techniques and their experience levels on student outcomes, and argue for the role of such systematic data collection in sustained course and institutional transformations.

Keywords: physics education research, course reform, assessment.
PACS: 01.30.Ib, 01.40.Di, 01.40.Fk, 01.40.G-, 01.40.gb

INTRODUCTION

Decades of education research show that transformed courses can improve student performance on conceptual measures.[1] At the University of Colorado (CU) Boulder, we restructured our calculus-based physics sequence in 2004 to include interactive engagement using clickers in the large-lecture setting, Tutorials [2] with trained undergraduate Learning Assistants [3] in recitations and a help-room where students work on personalized CAPA homework [4]. We have presented earlier results on investigations in this context of replication,[5] sustainability,[6] gender,[7] and more.[8] We continue to collect data in these classes, which we report on here.

Systematic and ongoing long-term data collection of this type [9] (i.e. without an explicit "design experiment" structure) supports intentional and sustained course transformation by informing the relative roles and impacts of curricular choices, institutional structures, student background, and faculty experience. It also suggests new questions regarding characterization of faculty choices, and their professional development. Our data confirm well-established positive impacts of interactive engagement in large-class settings, and suggest that faculty engaging in "deliberate practice" of IE [10] can show improved outcomes over time.

BACKGROUND

Introductory calculus-based physics at CU Boulder consists of 3 large classes/week (up to ~700 students split into two lectures). In all classes for which we have data, faculty have opted to use *Peer Instruction* [11] with clickers. Students participate in one hour of recitation each week, which use *Tutorials in Introductory Phys*ics [2]. Following Mazur [12], we characterize most of our classes as nominally IE-2 (high level interactive engagement), based solely on the classroom and recitation structures. Three of the earliest implementations of Physics 1 we label IE-1 (for partial interactive engagement) because the faculty chose to run more traditional recitations. We have no data for any completely traditional classes.

Data Collection and Demographics

Every semester since Sp '04, we have administered the FMCE[13] in the first week's recitation section, and again in the last or 2nd to last week of Physics 1. Altogether, we have collected over 7000 posttests representing ~70% of the students registered. We match valid pre- and posttest data for 90% of these students, constituting the data set represented in the figures to follow. In Physics 2, we began collecting BEMA data [14] in Fa '04, in much the same fashion as the FMCE. However, because of the stable (and low) pretest scores, many faculty (5 semesters) chose not to collect pretest data for Physics 2. We have collected almost 5000 valid posttests, representing 78% of the enrolled students. The average physics class grade of the students for whom we have posttests is ~2.75 in both classes, slightly higher than the overall average of 2.5, indicating that our sample is just slightly biased - we are missing some of the lower

performing students (who are more likely to miss the recitation in which the posttest is given).

Our cumulative dataset thus includes almost 9000 unique students spanning 17 semesters of Physics 1 and 16 semesters of Physics 2. Our classes are, on average, over 50% engineering majors, just over 25% female, and over 80% Caucasian. (All approximately the same for Physics 1 and 2.) This student population is well-prepared: we have SAT-Math data for over half of them, who average 650, (Math ACT average is 28.5), and the average high school GPA is 3.7/4.

Characterization of Faculty

At CU Boulder, two faculty are assigned to teach large-enrollment introductory courses. A lead faculty member lectures, with the second ("backup") working with grad Teaching Assistants and undergrad Learning Assistants (LAs) [3] preparing for Tutorials. Faculty cycle through these introductory courses over time.

Our data span courses taught by 27 faculty in lead or secondary teaching roles. Seven are female (two of whom were in the lead role). We have not systematically observed the classroom dynamics across our dataset, although we have such information for several semesters.[8] The syllabi, curricula, homework materials, and exam styles are similar across most terms, with faculty often sharing resources like exam and clicker questions. For the purposes of this paper, we bin faculty into 3 broad groups: 1) PER faculty, all of whom are highly experienced in this teaching environment 2) experienced faculty, which we define as having already taught at least once in this class using the described IE methods, and 3) first-time faculty in this particular environment. A few faculty members we label "PER-neutral": such faculty members adopt many but not all of the reform materials and classroom structures, and do not participate in faculty discussions about the use or intent behind interactive engagement, and/or express private skepticism of the value of such methods.

RESULTS

The average CU FMCE pretest score is 34%. Our average FMCE posttest is 65%, for an overall average normalized gain of 47%. The typical standard deviation in any given semester is ~20% on the pretest, and ~27% on posttest, but when looking across semesters, the standard deviation of *class averages* is a mere 2.5% for the pretest, and 7% for the posttest. Thus, the average pretest scores are extremely stable from semester to semester. There is a small, persistent (~3.5%) fall-spring difference, which we attribute to a different population in the "off-sequence" semester.

For the BEMA, the average CU pretest score is 27%. The average BEMA posttest is 57%. Comparing to national results, [15] our pretest is indistinguishable from peer institutions. Our posttest is well above the 40-45% seen in traditionally taught courses, [15] and similar to results from *Matter and Interactions* courses.[16] Typical standard deviations on pre and post-tests are 10% and 15% respectively. As with the FMCE, the standard deviation across terms of the *class average scores* is much smaller, 1.4% for pretest, 4% for posttest, indicating remarkably stable pretest results (with no fall-spring semester effect evident), and a surprisingly small spread in posttests, at well.

Declared physics majors (just under 10% of the population) score slightly higher (4%) than the class average, both pre and post, on both the FMCE and BEMA, but their final course grades are comparable.

We observe a persistent, statistically significant gender gap [7] in scores on the conceptual surveys. Female students' FMCE averages are ~13% below the males, both pre and post. In Physics 2, the gender gap on the BEMA is smaller, ~2% pre, but 6% post. There is only a small (but still statistically significant) gender gap (in the same direction) in course grades, roughly 0.1 on a 4-point scale in Physics 1, 0.2 in Physics 2.

IE and Student Performance

A summary of many of the individual results described above can be shown with a histogram of class average normalized gains.[17] On this plot (see Fig 1), we include Hake's original FCI results (dashed outlines) with the caveat that normalized gains on FMCE may not be perfectly comparable to those for the FCI. [18] The results at CU are separated into the three "IE1" courses (light fill) which did not use Tutorials in recitation, and the remaining 14 IE2 semesters (solid black).

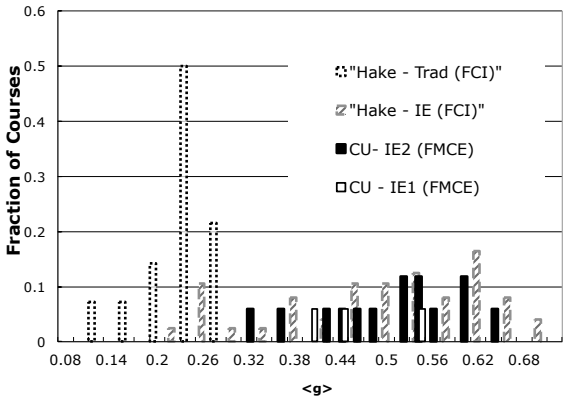

FIGURE 1. CU results shown in solid (light fill showing IE1, and solid black IE2) Hake's FCI data for traditional and interactive engagement (IE) classes (dash) shown (dashed) in the background (reproduced with author's permission).

These results provide confirmation of Hake's result for interactive engagement classes – we see the same spread, and scale, of normalized gains across our 17 terms, which now represents a local data set comparable in size to Hake's broader set. Two of the three IE1 classes are near the bottom end of our distribution. The highest of the three IE1 classes is an interesting case – although Tutorials were not used, students in those recitations worked in small groups on workbook activities from their text [19], not the completely traditional recitation that the other two IE1 courses offered. The average score for that term lies in roughly the middle of the CU distribution.

The Role of Faculty

To investigate the origins of the observed variance in normalized gains, we organize student performance data by faculty experience level in Fig 2. (Note that this is not a histogram like Fig 1, it shows normalized gain as a function of faculty member teaching.) PER faculty are consistently in the upper end of the range of gains, and experienced non-PER faculty are distributed similarly.

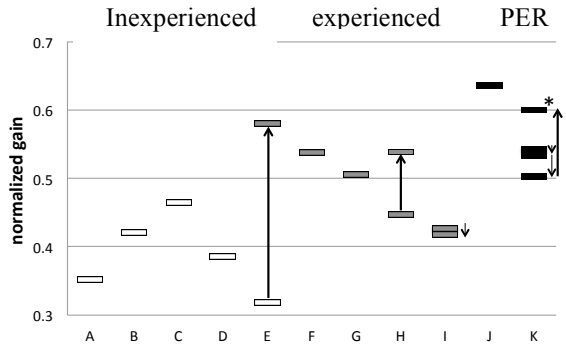

FIGURE 2. CU data for normalized gain on FMCE (now the vertical axis, note the suppressed zero.) Letters on horizontal axis indicate distinct faculty. Faculty A-E were all relatively inexperienced teachers. (The second time E teaches is categorized as 'experienced'.) Arrows indicate shifts for faculty teaching multiple times. The three shortest arrows are not statistically significant shifts. The asterisk indicates a semester when SmartPhysics [20] was used. Faculty I is characterized as "PER neutral".

There are three instances where a faculty member taught the lead role twice. In two cases, there was a statistically significant improvement in class performance the second time. (Standard error of the mean for normalized gains within a given semester is consistently ~±0.02) The longest arrow in Fig. 2 shows the improvement of a junior faculty member teaching a large course for the first time in his career. The first time, his backup faculty member was also another inexperienced teacher, while the second time, his backup was PER faculty. The medium-length arrow shows the shift for a senior faculty member, H, who has taught many classes, including large ones. He was backed up with an inexperienced and "PER neutral" partner the first time, and with an enthusiastic but novice partner the second round. The short black downwards arrow in "I" is for our only "PER neutral" lead instructor (characterized in the previous section.) The negative shift is not statistically significant. One PER faculty member taught the class four times. His normalized gains, chronologically, were .54, .53, .50, and .60. The lowest of those was the only term he was backed up by a novice partner. The highest is statistically significantly different from the others, and that term he was backed up by a PER faculty member and used SmartPhysics [20] before lectures.

These results suggest that the spread in posttest results may be attributed in part to faculty experience, with some role (which we cannot separate out) played by the experience of the backup faculty, as well as the degree to which the lead has incorporated the principles and curricula of interactive engagement in their teaching. Further support for these ideas are found in the BEMA results from Physics 2. (See Fig 3)

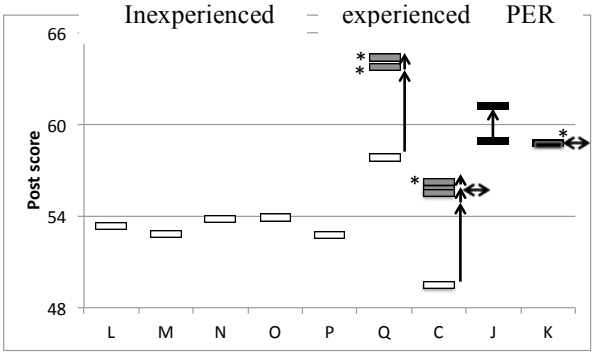

FIGURE 3. BEMA post-test scores. (note the highly suppressed zero on the vertical scale) Faculty L-Q were inexperienced teachers. Prof. Q had served as backup for the (higher scoring) class labeled H in Fig 2. Faculty C, J, and K correspond to the same lettered faculty in Fig 2. Arrows indicate shifts of individual faculty teaching multiple times. Only the three longest arrows are statistically significant. Asterisks indicate semesters using SmartPhysics [20].

The results in Physics 2 are qualitatively similar to Physics 1, although the distribution of scores is narrower than it was in Physics 1. The collection of curricular materials used in these Physics 2 classes is *also* less varied than it is in Physics 1: all of these faculty borrowed freely from the original set of concept tests developed by the lead PER faculty members. Again, we see that experienced faculty and PER faculty largely represent the top half of the distribution. And, again, we see that inexperienced faculty teaching a second time show statistically significant improvements. The largest improvement occurred for a first-time faculty teaching by himself,

the second time he was backed up by a PER faculty member. The next largest improvement is from a junior faculty member whose backup changed from experienced to less experienced, with the addition of SmartPhysics used before lectures.[20]

DISCUSSION & SUMMARY

In these data, which span 33 semesters, 9000 unique students and 17 different lead instructors, we reproduce the well-established finding that IE courses result in comparatively high post-test results. These findings hold across faculty and semesters.

The data also suggest that the degree of experience faculty have with these interactive approaches matters. Faculty teaching in these ways for the first time consistently have students performing on the lower end of the data set. Faculty teaching for a second or third time show consistent improvement in student performance measures. The only cases where we do not see shifts in student performance are by faculty who are either classified as "PER" (who are familiar with these IE approaches and underpinning philosophy and structure, and consistently score at the highest end of our range), or as "PER-neutral" (skeptical and unengaged in discussions around these curricula). These results suggest that faculty can learn, and that deliberative, or intentional, practice has a positive impact on the educational experiences for students.

There are additional tantalizing implications that are more tentative. It appears that the experience of the backup faculty member (the individual leading the TA/LA meetings for the Tutorials) may have an impact. Those lead faculty who have been supported by more experienced secondary faculty generally show improvement from terms when the backups have been more junior. Similarly, there are several instances when it appears the SmartPhysics preflights [20] might have positive impact. Instructor K, a seasoned veteran of the environment with very stable scores makes his most dramatic shift in Physics 1 when supported by a PER faculty member and employs the SmartPhysics preflights. Similarly Instructor Q makes the most dramatic gains when issuing preflights. That said, Instructor K makes no shift in student performance when employing SmartPhysics in another term (Fig 2). These findings need further investigation, and larger datasets to distinguish among these (and other) factors.

Collecting and analyzing these data is good not only for individual course assessments, but also for studying and supporting systematic transformation. We can use such data to move beyond assessments of a single instructor and a single course to observe the factors that support the widespread adoption and effective implementation of educational practices. For instance, at CU, the data serve as a mechanism for change. Collecting and reporting these data has become a part of departmental practice. Faculty are privately informed of their performance each semester, and given anonymized versions of these plots to contextualize their performance. While far from perfect, it allows faculty and department the option to use these results to move beyond the standard end-of-term student evaluation as the sole metric of quality. We are beginning to couple teaching with learning.

Examination of these data over time can inform the department on steps for improvement. Results suggest that faculty development and buy-in are key. We demonstrate these are possible for a department to achieve. Lastly, if we were to unpack why and how these transformations take root at CU, a framework for studying STEM change [21] proves useful, and suggests a mix of prescriptive (Tutorials and FMCE) and emergent (faculty practice) approaches, and a mix of top down (institutional), along with bottom-up (faculty decision making and engagement) support.

ACKNOWLEDGMENTS

Thanks to PhysTEC (APS/AIP/AAPT), NSF CCLI (0410744, 0737118), NSF LA-TEST (0554616), the CU Science Education Initiative, the CU PER group, and many faculty and students who have contributed.

REFERENCES

1. D. Meltzer, R. Thornton, *Am. J. Phys* **80** (478) 2012.
2. L. McDermott and P. Schaffer, *Tutorials in Introductory Physics* (Prentice-Hall, Upper Saddle River, NJ, 2002).
3. V. Otero et al., *Am. J. Phys.**78** (1218), 2010.*
4. CAPA, http://www.lon-capa.org/
5. N. Finkelstein, S. Pollock, *Phys Rev STPER*. **1,** 010101 (2005).
6. S. Pollock, N. Finkelstein, *PR STPER*. **4,** 010110 (2008).
7. L. Kost-Smith et al., *PR STPER* **6(2),** 020112 (2010), and *PR STPER* **5(1),** 010101 (2009).
8. C. Turpen & N. Finkelstein, *PR STPER*, **5,** 020101 (2009)
9. C. Crouch, E. Mazur, *Am. J. Phys.* **69,** 970 (2001)
10. K. Ericsson et al., *Psychological Rev.* **100(3)**, 363 (1993)
11. E. Mazur, *Peer Instruction: a users manual* (Prentice Hall: Upper Saddle River NJ, 1997).
12. M. Lorenzo et al., *Am. J. Phys.* **74**, 118 (2006).
13. R. Thornton, D. Sokoloff, *Am. J. Phys*. **66,** 228 (1998).
14. L. Ding, et al., *PR STPER* **2,** 010105 (2006).
15. M. Kohlmyer et al., *PR STPER* 5, 020105 (2009).
16. R. Chabay, B. Sherwood, *Am. J. Phys* **74,** 329 (2006).
17. R.R. Hake, *Am. J. Phys.* **66,** 64 (1998).
18. R. Thornton, *PR STPER* **5** 010105 (2009).
19. R. Knight, *Physics for Scientists and Engineers,* Addison Wesley, Boston, MA, 2003).
20. www.smartphysics.com, T. Steltzer et al, *Am. J. Phys.* **77(2),** 184 (2009)
21. C. Henderson et al, *JRST*, 48 (8), 952-984, (2011).