

# Comparing student learning with multiple research-based conceptual surveys: CSEM and BEMA.

S. J. Pollock

*Department of Physics, University of Colorado, Boulder, CO 80309-0390*

**Abstract.** We present results demonstrating similar distributions of student scores, and statistically indistinguishable gains on two popular research-based assessment tools: the Brief Electricity and Magnetism Assessment (BEMA) and the Conceptual Survey of Electricity and Magnetism (CSEM). To deepen our understanding of student learning in our course environment and of these assessment tools as measures of student learning, we identify systematic trends and differences in results from these two instruments. We investigate correlations of both pre- and post- conceptual scores with other measures including traditional exam scores and course grades, student background (earlier grades), gender, a pretest of scientific reasoning, and tests of attitudes and beliefs about science and learning science. Overall, for practical purposes, we find the BEMA and CSEM are roughly equivalently useful instruments for measuring student learning in our course.

**Keywords:** course transformation, course assessment.

**PACS:** 01.40.-d, 01.40.Fk, 01.40.gb, 01.40.G-

## INTRODUCTION

Research-based conceptual instruments play a key role in course development, assessment, and even faculty awareness of student learning in introductory physics. In the domain of Electricity and Magnetism (E&M) several such instruments are widely used [1-3]. Two of the commonly used and cited evaluations are the Brief Electricity and Magnetism Assessment (BEMA [1]) and the Conceptual Survey of Electricity and Magnetism (CSEM [2]), both of which are broad surveys of the field of E&M. Each instrument has been evaluated for reliability and validity, although neither tries to span the entire domain of E&M. At the University of Colorado, we have used the BEMA for the past four years to assess our ongoing efforts at course reform [4-6], with well over 2500 students tested to date. Assessment data have been collected for a broad range of participants: ranging from introductory courses [4] to upper division students [7], learning assistants, graduate TAs, and even faculty.

The instruments are similar to one another in many general ways, but differ in the specifics of the majority of questions, with somewhat different content emphases. It is thus difficult to directly compare outcomes across institutions if they use different instruments. The purpose of this research study is to provide a first pass at comparing the two exams across one student population, to calibrate and compare the two instruments.

## COURSE SETTING

Our study measured students in the University of Colorado's (CU) calculus-based Physics II course in Fall 2007. The student population (N=425) is a mix of majors (70% engineering), is 76% male, and just over half are sophomores. The course was team taught, with the lead instructor (SJP) a member of the Physics Education Research group. We characterize this as a reformed, large-scale course, with ConcepTests and peer instruction [8] during three 50-minute lectures per week, online homework [9] and one 50-minute per week Tutorial using Washington Tutorials [10] with trained graduate TAs and undergraduate Learning Assistants [11]. There is a staffed help-room available for students. The introductory lab is decoupled from (although typically concurrent with) this course.

The average BEMA pretest at CU (averaged over 8 semesters) is 27%, with a typical standard deviation of 10%, and small (+/-1 to 2%) variations among terms [4]. Post-tests have larger variation - ranging from 50% to 61%. As detailed below, our BEMA results from Fall 2007 are on the high end at CU. The results from this study should thus not be taken as representing a broad spectrum of types of courses or teaching styles, at this point we are limited to the population and pedagogy currently in use at our institution. [4-6] The focus here is on *comparative* performance on the CSEM and BEMA within one population of students.

## DATA SOURCES

We collected data in Fall 2007 on several measures, including content assessments, grades, attitudes and beliefs (using the Colorado Learning Attitudes about Science Survey, CLASS [12]), and basic scientific reasoning (Lawson test [13]). The content (conceptual) surveys are issued in recitation sections; the rest of the survey instruments are given online for which students receive token participation credit. For the content surveys, we split the students based on their recitation times, giving half the class the pre- and post-BEMA, and the other half the CSEM. That is, all students in a single recitation received either the BEMA or CSEM survey.

The split was by recitation section. We ensured that the two groups equally represented the different graduate TAs, lecture times, recitation times, and rooms. After the fact, we have verified that there were no statistically significant differences between these two groups on measures available to us, including demographics (gender, major, or class), earlier grades from Physics I, concept evaluations in Physics I, CLASS scores, Lawson scores, or course grades, participation, or exam scores in this Physics II class.

Participation was typical for our institution, with 95% of the students taking the pre-concept instrument, and 78% taking matched, valid, pre *and* post concept instruments. In addition, 57% took the Lawson test (issued pre only), and 42% took the CLASS (matched, pre- and post). We have Physics I grades for 85%, and matched pre-post FMCE [14] for 65% of the class.

## RESULTS

The broad purpose of this study is to help characterize, calibrate, and better understand these two commonly used assessment instruments. We are interested in overall difficulty, whether learning gains are similar for each survey, and whether pre- and post-test scores are similarly correlated with other pre-factors, and/or with outcome measures. Looking at common items on the two instruments also allows for a direct comparison. These results are all presented below. We find statistically comparable outcomes, so the choice of instrument may ultimately be determined more by individual match to local course goals.

### Main result: Difficulty and gains

Our average BEMA posttest has been 56%, with a typical standard deviation of 16%. Data for BEMA and CSEM for Fall 2007 are shown in Table 1. (Note that the semester of this study had slightly higher BEMA post-scores than our historical average at CU.)

We show pre and postscores, gain ( $\langle \text{Final} \rangle - \langle \text{Initial} \rangle$ ), and normalized gain ( $\langle \text{Final} \rangle - \langle \text{Init.} \rangle / (100 - \langle \text{Init.} \rangle)$ ). Table 1 shows the gain of the averages, which in both cases was close to the average of individual gains.

**TABLE 1.** Comparison of BEMA and CSEM average results for Fa07. Only matched, valid scores are included. Numbers shown in parentheses are standard deviations.

Test	Pre (SD)	Post (SD)	gain	Norm. gain
BEMA (N=162)	26% (9%)	<b>61%</b> (15%)	35%	0.47
CSEM (N=168)	32% (10%)	<b>66%</b> (16%)	34%	0.50

Average CSEM scores, pre-and post, are higher than BEMA scores by 5-6%, a statistically significant ( $p < 0.05$ , 2-tailed t-test) difference, with moderate effect size (difference/standard deviation=0.5 pre, 0.4 post). Both the absolute and normalized gains are statistically indistinguishable for the two tools. It appears that (for this population) the CSEM is slightly easier, but both exams are comparable with respect to evaluating student learning. [15]

## Correlations

Assessment tools serve many functions besides measures and comparisons of student learning. Individual items inform teaching, and pretest scores can help serve as part of at-risk indicators [16,17]. We are also interested in how performance on post-tests matches with performance in the class as evaluated in other ways. Table 2 shows correlation coefficients between several measures of interest, for each half of the class taking the BEMA and CSEM respectively.

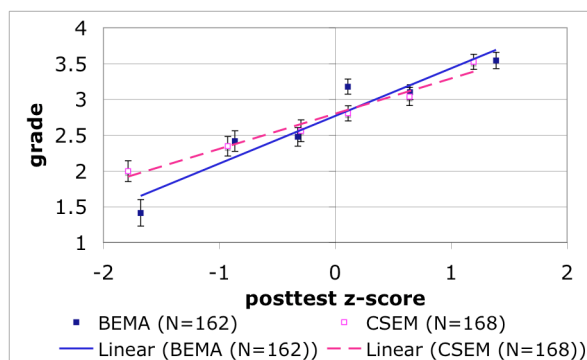
**TABLE 2.** Pearson correlation coefficients for populations taking (matched pre-post) BEMA or CSEM. Correlations ( $r$ ) for students taking the BEMA/CSEM are shown separated in each table entry by a slash, "/". ( $r$  for CSEM is given in italics) E.g., the first entry in the upper left square says  $r(\text{pre with post})$  for students who took the BEMA=0.4, but  $r(\text{pre with post})$  for students taking the CSEM=0.5. All entries except  $r(\text{BEMA pre to Lawson})=0.1$  are statistically significantly different from 0 ( $p < .05$ )

	Post	Grade	Lawson	Prev. grade
Pre	0.4/0.5	0.2/0.3	0.1/0.5	0.2/0.4
Post	-	0.7/0.6	0.3/0.6	0.5/0.5
Course grade	-	-	0.5/0.5	0.7/0.7
Lawson	-	-	-	0.4/0.4

The first two rows of Table 2 show correlations of pre and post-test scores with several other measures (final grade in course, score on Lawson test of scientific reasoning taken at the start of the term, and

most recent previous grade from their Physics I course). The correlation of pre-score to post-score is just slightly higher for the CSEM than for the BEMA. The same holds for course grade correlated to pretests, with the CSEM pre-score correlating slightly more with final grade in the class.

The post-test shows the opposite trend, with the BEMA post correlating slightly *more* with course grade. Neither difference is large. Because both tests have rather low (and narrowly distributed) pretest scores, normalized gains closely track post scores (correlation of post-test to normalized gain is 0.9 for both BEMA and CSEM). Fig 1 shows these results graphically, plotting course grades as a function of conceptual post-test scores (binned into 6 roughly equal sextiles). Fig 1 demonstrates that conceptual post-tests reflect student performance in the class, with considerable fluctuation largely hidden in the binning.



**FIGURE 1.** Final course grade (0-4) as a function of (binned) post-test scores on the BEMA (solid) and CSEM (dash/open points) (z-score is (score-average)/standard dev)

The correlation of BEMA/CSEM pre and post-tests with Lawson's test shows slightly different results for the two instruments. It appears that the Lawson measure of scientific reasoning is better correlated with pre and post scores on the CSEM than with the BEMA. A similar (but weaker) trend is seen from students' previous physics grade (from Physics I), which also correlates slightly more with CSEM pre than BEMA pre (but no difference for post-tests). These correlations might be associated with the slightly easier nature of the CSEM, and are thus perhaps reflective of test-taking skills. We investigated correlations of CLASS (pre) with BEMA and CSEM, and again found only small differences - the correlation of CLASS (pre) to conceptual post-scores was 0.5 (for the BEMA) and 0.4 (for the CSEM), comparable to the predictive power of CLASS (pre) with course grade. [6,12]

In summary, it appears that both instruments are similar regarding correlations of pre and post with

each other, with respect to other measures of student preparation, and with student final grades. The CSEM pre-score is more strongly correlated with measures of previous student test-taking success, including grade in Physics 1, Lawson test, and FMCE scores from a previous term ( $r=0.5$ ). The BEMA post is somewhat more strongly correlated with performance in *this* class. We find no compelling evidence here for or against either instrument as a better measure of student learning, nor as an at-risk indicator, although the BEMA appears perhaps slightly more coupled to our own learning goals as measured by final grades.

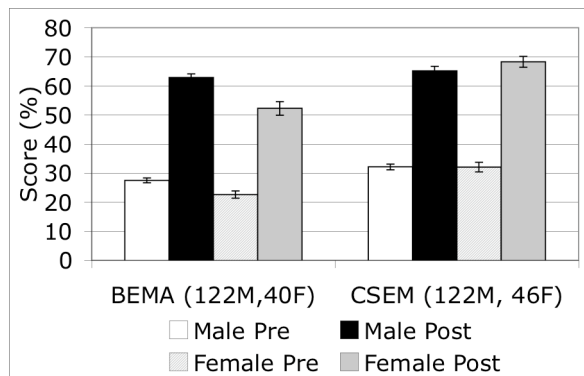
## Comparing Items

Individual items provide useful information for understanding and improving instruction, and in this case can also provide additional calibration between the two instruments [3]. There are six questions (out of 31 total on the BEMA, 32 on the CSEM) which are identical or nearly identical (although the CSEM has only 5 answer choices, while the BEMA has up to 10 possible answers on some questions, and the question order is different), in addition to three common ECCE questions [3] which we added to the end of all surveys.

Identical individual questions on the instruments did not generate identical average scores. One or two specific common questions generated differences as large as 15%, although more typically less than  $\pm 5\%$  different, perhaps reflecting effects such as question placement as well as the different number of distractors. However, the average difference (post-test) for all 9 common questions is 3% (higher for the CSEM population). If several of the common questions are scored as recommended on the BEMA[1] (looking for internal consistency rather than correctness) the difference between these 9 common questions is 0% on the post-test. This subset of questions is 5% higher on the CSEM *pre*-test, possibly reflecting the smaller number of distractors, but not statistically significantly different from zero. In summary, we find that on average, students perform the same overall on the overlapping set of common questions. This also gives additional confirmation that our two groups are equivalent.

## Gender Breakdowns

Building on prior work on the gender gap [16,18] we examine differences by gender between performance on the BEMA and CSEM instruments. Our population contains relatively small samples, ( $N(\text{female, CSEM})=46$  and  $N(\text{female, BEMA})=40$ ) so one must be quite cautious about interpreting statistical significance. The main outcomes are shown in Fig. 2.



**FIGURE 2.** Pre and Post BEMA and CSEM scores (as %) for male and female students (matched, valid only). Error bars show standard error of means.

Pretest scores are statistically significantly different by gender on the BEMA (male pre=27.5±/-.8%, female pre=22.6±/-.1.3%,  $p < .05$ ), but *not* significantly different on the CSEM (male pre = 32±/-.1%, female pre=32±/-.1.7%) Post-test scores are again statistically significantly different on the BEMA (male post=62.9±/-.1.2%, female post=52.2±/-.2.3%,  $p < .05$ ) but *not* significantly different on the CSEM (male post = 65.1±/-.1.5%, female post=68.3±/-. 1.8%) The shift is thus 6 points greater for males than females on the BEMA ( $p < 0.05$ ) but is 3 points *lower* for males than females on the CSEM (not statistically significant).

The gender gap is thus (statistically significantly) increased on the BEMA, and (insignificantly) decreased on the CSEM [19]. We do not have a mechanistic explanation for these results on gender differences, but find them intriguing and worth further investigation.

## CONCLUSIONS

With a broad availability of validated, research-based conceptual assessment tools in introductory physics, the particular choice of instrument is often based on a sense of connection to course goals, or the collective published base with which to compare results. To assist in this choice, we have administered BEMA and CSEM pre- and post-tests to a large group of introductory students. Overall we find close similarities in measurement of learning using both CSEM and BEMA. The differences between the two instruments in one class (of order 5% on overall difficulty) are of the same scale as differences we have seen between different semesters on the BEMA over time. The BEMA appears slightly harder, with a slightly smaller correlation of pretest results to either prior measures or learning gains in the course, and a slightly stronger correlation of post-test to course

grades, but with most differences quite small. We observe there to be no gender gap on the CSEM but a nonzero gap on the BEMA. We encourage other faculty to engage in similar studies, to allow and support comparison of courses across a broader spectrum of institutions and pedagogies.

## ACKNOWLEDGMENTS

Thanks to PhysTEC (APS/AIP/AAPT), NSF CCLI (DUE0410744) and NSF LA-TEST (DRL0554616). Thanks also to the University of Colorado, the CU Physics Department, the PER at Colorado group, Prof. V. Gurarie who team-taught this course, and to the students of Physics 1120, Fa07.

## REFERENCES

1. L. Ding et al, *Phys. Rev. STPER*. **2** (2006) 010105. See also [www.ncsu.edu/per/TestInfo.html](http://www.ncsu.edu/per/TestInfo.html)
2. D. Maloney et al, *Am. J. Phys.* **69** (2001) S12
3. We supplement both exams with three questions from the ECCE instrument of Thornton and Sokoloff, see [physics.dickinson.edu](http://physics.dickinson.edu). For a list of other instruments, see <http://www.ncsu.edu/per/TestInfo.html>
4. S. Pollock and N. Finkelstein, *Phys. Rev. ST Phys. Educ. Res.* **4**, 010110 (2008)
5. N. Finkelstein, S. Pollock, *Phys. Rev. ST Phys. Educ. Res.* **1**, 010101. (2005).
6. S. Pollock, *2004 PERC Proc* 790, p.137, S. Pollock, *2005 PERC Proc.* 818, p.141, S. Pollock and N. Finkelstein, *2006 PERC Proc.* 883, p.109,
7. S. Pollock, *2007 PERC Proc.* 951, p.172
8. E. Mazur. (1997). *Peer Instruction*, Prentice Hall
9. CAPA, <http://www.lon-capa.org/>
10. L. McDermott and P. Schaffer, *Tutorials in Introductory Physics* (Prentice-Hall, Upper Saddle River, NJ, 2002) See also L. McDermott, E. Redish, *Am. J. Phys* **67** (1999) p.755 for many other references.
11. V. Otero, *et. al.*, *Science* **28** July 2006, p. 445.
12. Adams, et al., *Phys Rev ST: PER*, **2**, 010101 (2006) and reference list at <http://class.colorado.edu>
13. A. Lawson, *J. Res. Sci. Teach.* **15**(1), 11 (1978). See also V. Coletta and J. Phillips, *Am. J. Phys.* **73**, 1172 (2005)
14. R. Thornton, D. Sokoloff, *Am. J. Phys.* **66**, (1998) p. 338
15. The CSEM post-test has nearly equal mean, median, and mode, whereas the BEMA post-test has a slightly higher median (3% higher than the average) arising from a slight "tail" of lower scores. .
16. S. Pollock et al., *Phys. Rev. ST Phys. Educ. Res.* **3**, 010107 (2007). See also L. E. Kost, et al., *PERC Proceedings* 2007, 951, p. 136
17. R. Thornton, *Proc. E. Fermi Summer School*, CLVI. E. Redish, M. Vicentini eds. p 591 (Italian Phys Soc, 2004).
18. M. Lorenzo, *et. al.*, *Am. J. Phys* **74**(2), 118 (2006).
19. Average final grades for males and females in this course, and also across the BEMA and CSEM-taking subpopulations, are statistically indistinguishable.