

# Multiple Roles of Assessment In Upper-Division Physics Course Reforms

Steven Pollock, Rachel Pepper, Stephanie Chasteen, Katherine Perkins

*Science Education Initiative and Department of Physics, University of Colorado, Boulder, Colorado 80309 USA*

**Abstract.** The University of Colorado at Boulder has been involved in a systematic program of upper-division undergraduate course transformations. The role of assessment has been critical at multiple, interconnected scales: (1) formative evaluation focused on the *course* itself in the design phase; (2) formative assessment focused on *students* in the instructional phase and (3) summative assessment to determine student performance *and* the success of course design. We summarize the role and nature of assessments at each of these levels. At the design scale, investigative measures include observations and surveys of students and student work. In the classroom, assessments to determine and address student difficulties include clicker questions and tutorials. At the summative scale, assessments include faculty interviews and course and tutorial-scale posttests. We discuss examples, affordances, outcomes, and challenges associated with these different layers of assessments at the upper-division level.

**Keywords:** physics education research, course reform, assessment, upper-division.

**PACS:** 01.30.Ib, 01.40.Di, 01.40.Fk, 01.40.G-, 01.40.gb

## INTRODUCTION

At the University of Colorado at Boulder, we are investigating student learning in select upper-division physics courses and making corresponding targeted changes in pedagogy and curriculum<sup>1,2,3,4,5,6,7,8</sup>. We have introduced techniques previously demonstrated to improve student learning in introductory physics<sup>9</sup>: developing consensus learning goals, applying interactive techniques such as concept tests and small-group tutorials, focusing on known student difficulties with the material, and aligning course materials with explicit expectations for student achievement. This paper focuses on the role and nature of *assessment*<sup>10</sup> in multiple stages of course transformation: development, implementation, and evaluation. We examine productive classes of assessments, borrowing from lower-division efforts, focussing on new challenges and constraints associated with the upper-division.

## BACKGROUND

Our department graduates ~50 majors/year, with core upper-division classes of roughly 25-75 students taught each term. We have ~50 faculty who cycle through teaching assignments across the curriculum, so any given course rarely sees the same instructor for more than a term or two, and different instructors may teach sequential courses. Our introductory sequence has been transformed for over a decade following PER-based principles and curricula (e.g. Peer Instruction<sup>11</sup>, U. Washington Tutorials<sup>12</sup>, undergraduate Learning Assistants<sup>13</sup>, online

homeworks, and a student-centered helproom.) Faculty cycle through these introductory courses as well, so a majority of our faculty have personal experience with interactive engagement practices. We have organized brown-bag lunches to discuss issues in undergraduate education for several years. These meetings have been well-attended, with 31 CU Physics faculty having participated in at least one<sup>14</sup>. Our department is characterized by faculty who are largely supportive of, and engaged with, PER efforts across the curriculum<sup>15</sup>.

## LAYERS OF ASSESSMENT

To inform our course transformations in a scholarly way, and to help individual students and instructors, we require feedback and assessment at multiple levels. While creating materials for a transformed course, feedback on, and assessment of, our central design questions are required, a process we refer to as "formative evaluation", which we define and describe below. Once a class has been modified, a key pedagogical element is formative assessment of students throughout the course. At the end of each semester, we employ summative assessment(s) to evaluate individual students' learning, inform and motivate faculty and inform research and ongoing development of the courses. The distinctions between these layers are, of course, fluid - "formative assessment" (e.g. clicker questions in a classroom) can also provide summative data to evaluate and compare classes. Similarly, summative assessment (e.g. post-tests) can provide a formative role too, informing ongoing course development, as we discuss below.

## Design Stage: Formative Evaluation

"Formative evaluation" refers to frequent, low stakes, bi-directional (between faculty and developers) reflections and investigations that inform and alter design processes. Paraphrasing Redish<sup>9</sup>, we refer to "probes of an individual student's learning as *assessment* and to probes of our instruction as a whole as *evaluation*". Formative evaluation thus refers to elements of input and feedback at the course level.

**Faculty brown-bags.** Formative evaluation played a central role in our course transformation efforts during regular faculty brown bag meetings, which focused on developing consensus course learning goals and related assessments. Discussions extended well beyond simple lists of topical coverage, to include discussion of higher-level goals such as problem-solving skills, math-physics connections, checking one's solution, communication skills, and more<sup>1,5</sup>. In this way, faculty buy-in was driven from the bottom up - that is, by creating transformations that value and align with faculty goals, rather than trying to "sell" transformations after the fact<sup>16</sup>.

**Interviews, observations and surveys.** Formative evaluation included interviews (by a postdoc) with faculty who had recently taught the course in question. These interviews took place before and/or after the faculty brown-bags, to further reflect on and inform design. Additionally, for 1-2 semesters before and during initial implementations of transformed materials, we observed classes and interviewed undergraduates in order to investigate upper-division student difficulties. We also surveyed alumni to evaluate student attitudes and beliefs about the nature of learning in these courses, and to inform what we emphasize (whether we aim to prepare students for graduate work or engineering careers, for example, or focus on math and formalism vs. concrete examples).

These formative evaluations guided the creation of the elements of the transformed course - including the concept tests, activities, modified homework, and tutorials. The materials were focused on our learning goals, including sense-making, visualization, computation, and estimation. These elements both support and extend beyond the more traditional analytic calculational focus.

## Formative Assessment in the Classroom

Formative assessment can inform and alter instruction in real time. At the lower-division, formative assessments (in the form of clicker/concept questions, small group activities and UW Tutorials,

interactive lecture demonstrations, etc.) form the core of research-based transformations<sup>9</sup>. We modeled our upper-division course transformations on several categories of formative assessment, outlined below. In all cases, we found significant affordances, while recognizing that there are issues that we have not yet resolved. It may be that modeling upper-division transformations on successes from lower-division research is too limited, given the highly self-selected, more experienced learners populating our advanced undergraduate physics-major courses.

**Conceptual clicker questions in class.** Clicker questions and peer instruction build on a base of educational research<sup>9</sup>. At the upper-division, we focus clicker questions on conceptual issues, applications and extensions of high-level ideas, checking, sense-making, math-physics connections, and explicit use of multiple representations. In an ideal implementation, we use clicker questions to support articulation of reasoning, which allows us to hear multiple student voices and to scaffold increasingly sophisticated argumentation skills. Clicker questions relate to our consensus learning goals, such as communication skills, problem-solving strategies, and "intellectual maturity", which in this context is operationalized as student awareness of what they do not understand. This can be evidenced by students asking questions and taking actions to move beyond their difficulties. Further discussion of upper-division clicker questions can be found in Ref 5.

This method of formative assessment, while sharing many of the affordances seen in the lower division, introduces new issues and difficulties at the upper-division. Both the type of question and the facilitation of peer instruction must be adapted for the more sophisticated upper-division population. Questions are limited by their multiple choice nature, which tends to constrain the span of discussion and does not match as well with the increasingly complex procedures and ideas we are teaching. The short time typically allotted to them further disconnects them from the greater focus on calculational and procedural skills in the upper-division.

**Tutorials in and out of class.** Building on the U. Washington model<sup>12</sup>, we use tutorials in all of our transformed upper-division classes. They may be held outside of class in optional 1-credit "co-seminar" sessions. In other cases, we have adapted research-based materials for use during lecture periods<sup>17</sup>.

Tutorials build on conceptual underpinnings and model sense-making and checking (both internal, and via feedback from instructors). By promoting student discussion of the physics, tutorials provide faculty with valuable insights into student thinking and

provide PER researchers further understanding of student difficulties. They also give students directed feedback on their own understanding both from instructors and from peers. Tutorials provide a mechanism by which formative assessment can address broader goals than simply content - they target visualization and use of multiple-representations, expecting and checking solutions, organization of knowledge, and building on earlier material.

When tutorials at CU are out-of-class, they are only modestly well-attended (about 30-50% of the class)<sup>7</sup>. They add logistical complexity, additional time demands for faculty (in courses which do not traditionally have a recitation), and (in our case) require trained, advanced undergraduate Learning Assistant support<sup>13</sup>. Some faculty do not value the conceptual focus implicit in tutorials, because of the increasing emphasis on development of calculational skills at this level. The greatest weakness of these materials may be the dearth of research on student difficulties in the upper-division - continuing research and development of tutorials are sorely needed.

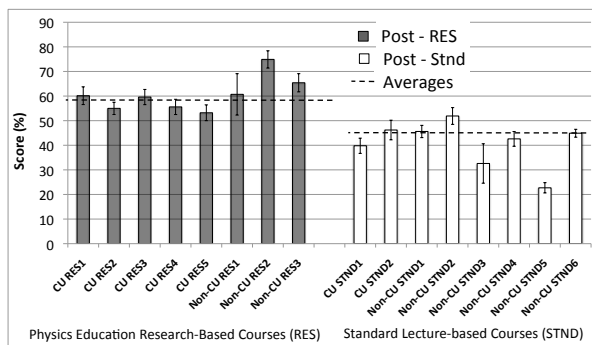
**Preflights.** A third form of assessment is frequent online feedback from students in the form of weekly "preflights" modeled on either the U. Washington Tutorial pretests<sup>12</sup> or JITT-style preflights developed at the US Air Force Academy<sup>18</sup>. Preflights ask short conceptual questions based on recent classes or upcoming reading. These can elicit student ideas, and focus both the instructor and the students on the daily learning goals. Students receive feedback on preflights in real time, or by the next class. Although fairly demanding of both students' and instructors' time, they provide micro-level data useful to refine and assess the impact and effectiveness of our tutorials and homeworks, potentially resulting in modified instruction on very short time scales<sup>18</sup>.

## Summative Assessments

Summative assessment refers to the use of performance measures - typically snapshots near the end of instruction - to provide information about individual students' learning and the overall success of classroom approaches. Conceptual post-tests do not count towards student grades, but are useful for students as a study-guide. (They can thus *also* serve a formative role for students, while still allowing us to summatively compare classes within and across institutions.) Student work on summative measures informs us as PER researchers of persistent student difficulties, providing yet another (formative!) tool for feedback as we modify curriculum and classroom methods.

**Conceptual Posttests.** We have developed three end-of-term in-class instruments: the CUE (Colorado Upper-Division Electrostatics evaluation), QMAT (Quantum Mechanics Assessment Tool), and CCMI (Colorado Classical Mechanics and Math Methods instrument, under construction)<sup>1,3,4,6,14</sup>. Significant time and effort was invested in evaluating the validity and reliability of these instruments. They target learning goals that traditional final exams can miss, including upper-level concepts, problem-solving approaches, and skills such as sketching, interpretation of formalism, and explanatory abilities.

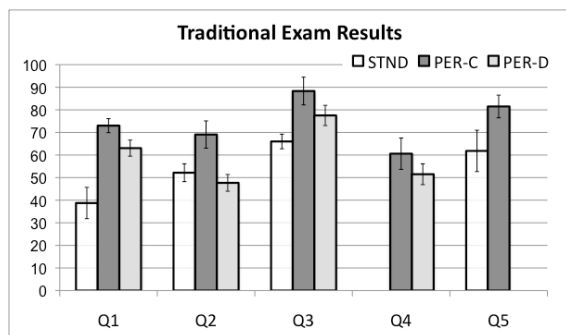
Fig. 1 shows CUE scores from 16 classes. Seven are from CU: two were traditional lecture-based courses taught by popular, experienced faculty. The remaining five used our suite of research-based materials and formative assessments. Data outside of CU reflect a mix of institutions and class sizes, with three courses using our materials, and six following traditional methods. The trend is clear - the average CUE score in research-based course implementations is ~15 points higher than in the standard lecture-based courses, an effect size of ~1 standard deviation. These data have proven influential at CU and beyond in helping faculty see the value of interactive engagement methods at this level.



**FIGURE 1.** Class average CUE scores from 16 courses at CU and elsewhere. See Ref 1. On the left are courses using Colorado materials (clicker questions, tutorials, and homeworks). On the right are traditionally taught courses.

There remain serious issues with conceptually-focused post-test summative assessment in the upper-division. Short answer (or even more challenging, multiple choice) questions are poorly matched to the high-level problem solving skills targeted in these courses, and can provide only a narrow view of student learning. The more such exams target higher-level learning goals, the greater the grading (and corresponding reliability) challenges. Furthermore, faculty must value and demand the conceptual aspects being measured by the particular instrument - if their goals center on calculational proficiency, the outcomes of such post-tests are less useful as feedback to faculty and less likely to merit class time.

**Traditional Exam questions.** Beyond the validated, research-based questions described above, faculty still put great stock in traditional exam questions that focus on computational skills. In Fig 2, we see several exam questions, given in multiple classes<sup>1,7</sup>. Here the general trend is that on computational problems (Q2-5 in the figure), we find no evidence that students do worse in the research-based transformed courses, despite the nominal shift in class time and emphasis. [See Ref 14 for more examples from other classes.]



**FIGURE 2.** A comparison of student performance on common midterm and final exam questions given in standard (STND) and transformed (PER C and D) courses. Q1 was more conceptual than the rest, which were largely computational. (See Refs 1 and 7.)

Here again there are serious issues with this type of summative assessment, when looking beyond individual student evaluation. Such questions are typically not research-validated, nor have they been constructed to target consensus learning goals. Furthermore, exam questions of this type are sensitive to idiosyncracies in course and instructor emphasis - performance on a given exam problem may be heavily influenced by the particular homework and related example problems solved or emphasized in class. These questions often deemphasize conceptual issues (and other high-level consensus goals), limiting the space of student learning we wish to assess.

**Surveys.** We briefly mention a third summative assessment tool we make use of - end of term surveys which focus on students' developing attitudes and beliefs about learning. Survey data provides feedback on student concerns and satisfaction. Sharing results with faculty has been a productive mechanism for communication and systematic improvement in classroom methods and pedagogical approaches. See Ref 14 for more examples and discussion.

## DISCUSSION

The development of well-articulated faculty consensus learning goals, which span both content and higher-level learning goals, have led us to include

multiple layers and types of assessment at the upper-division: before, during, and after both course development *and* instruction. These assessments inform our curricula, provide feedback to students, faculty and the PER community, and they provide a variety of measures of outcomes useful to different audiences. Upper division courses have different, novel issues<sup>19</sup> from those which have been well-investigated at the introductory level. These include different student populations, different background and skill levels, and different expectations from both faculty and students - in short, a different culture. We have evidence that adapting methods and assessments successful in Physics 101 are beneficial at the upper-division. But, it is possible that we could do better still by developing novel approaches, using richer and more appropriate teaching methods and assessments which requires ongoing research and development.

## 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 the many faculty and students who have significantly contributed to this work.

## REFERENCES

- S. Chasteen et al., *J. Coll. Sci. Teach* **40**(4), ('11) p. 70
- S. Chasteen, Pollock, *PERC proc.*, *AIP 1064* ('08) p. 911
- S. Chasteen, Pollock, *PERC proc.*, *AIP 1179* ('09) p. 7
- S. Chasteen, Pollock, *PERC proc. AIP 1179* ('09) p. 109
- S. Pollock et al., *PERC proc. AIP 1289*, ('10), p. 261
- S. Goldhaber et al, *PERC Proc. AIP 1179* ('09), p. 145.
- S. Chasteen et al., submitted to PERC 2011.
- For a full set of course materials and learning goals, see [www.colorado.edu/sci/departments/physics.htm](http://www.colorado.edu/sci/departments/physics.htm)
- E.F. Redish *Teaching Physics with the Physics Suite*, Maryland, John Wiley & Sons, 2003, and refs therein
- D. Wiliam, in *Second Handbook of Research on Mathematics Teaching and Learning*, ed. F. Lester (Information Age Publishing, Greenwich, CT, 2007), p.1053.
- E. Mazur, *Peer Instruction*, Prentice Hall 1997
- L. McDermott et al., *Tutorials in Introductory Physics*, Prentice Hall 2002
- V. Otero et al., *Am. J. Phys* **78** (11) ('10) p. 1218
- R. Pepper et al, submitted to PERC 2011
- K. Perkins, Turpen, *PERC proc.*, *AIP 1179* ('09) p.225
- C. Henderson and M. Dancy, *Am. J. Phys* **76** (1) ('08) p.79
- B. Ambrose, *Am. J. Phys.* **72** ('04) 453, see also Intermediate Mechanics Tutorials at [perlnet.umaine.edu/imt/](http://perlnet.umaine.edu/imt/)
- G. Novak et al., *JITT*, Benjamin Cummings (1999), and G. Novak and S. Novotny, private communications
- See e.g. C. Manogue, et. al, *Am. J. Phys.* **74** ('06) p. 344; J. Bilak and C. Singh, *PERC Proc. AIP 951* ('07) p. 49; T. Bing and E. Redish, *PERC Proc AIP 883*, ('06), p. 26