Development and Validation of the Calculus Concept Inventory

Jerome Epstein,
Mathematics Department, Polytechnic University, Brooklyn, NY 11201 jepstein@poly.edu

Abstract

I discuss the development and validation of an instrument known as the *Calculus Concept Inventory* (CCI). Patterned closely on the highly successful *Force Concept Inventory* (FCI, physics), it tests conceptual understanding of the *most basic* principles of calculus. Results are discussed, and the future plan that it will serve as a tool for the evaluation of teaching methodologies, as the FCI has been in physics.

I. INTRODUCTION

The Calculus Concept Inventory (CCI) is a test of conceptual understanding (and only that — there is no computation) of the most basic principles of differential calculus. The idea of such a test follows the *Force Concept Inventory* (FCI) in physics (Halloun and Hestenes, 1985a, b, Hake et al 1998a, b), a test which has spawned a dramatic movement of reform in physics education and a large quantity of high quality research. The FCI showed immediately that a high fraction of students in basic physics emerged with little or no understanding of concepts that all faculty assumed their students knew at exit. More dramatic, the FCI in Hake's analysis showed a very dramatic correlation with teaching methodology, where the normalized gain (see below) in Interactive-Engagement (IE) sections exceeded that in Traditional Lecture (TL)-based sections by two standard deviations.

Mathematics education is mired in the "math wars" between "back-to-basics" advocates and "guided-discovery" believers. There is no possibility of any resolution to this contest between competing faiths without scientific evidence of what works and what doesn't. Such evidence requires widespread agreement on a set of basic concepts that students should be expected to master in, for example, first semester calculus. The CCI is a first element in such a development and is an attempt to define such a basic understanding.

The CCI has undergone extensive development and evaluation, funded by the National Science Foundation. It was developed by a panel of widely respected people in calculus education and a consultant who is nationally known for development and validation of standardized tests. The test shows good performance characteristics and exposes exactly what the physics test showed. In the Fall semester of 2006 the test was given at two new test sites with programs using alternative methodologies. Comparison of gain from two widely divergent methodologies then becomes possible and will be discussed in this paper. The paper will also discuss the development and validation process in some detail.. Statistical parameters (p-values, discrimination, and the Reliability number) are discussed for the test as well. Those interested in using the test should contact the author. Maintaining the security of the test is a crucial issue. The CCI is the first in a hoped-for series of "Concept Inventories" for various levels in mathematics (including high school and earlier) that can hopefully serve to provide a scientific basis for discussions about teaching methodology and curricula. We are seeking funding for an Algebra Concept Inventory.

II. CONCEPT INVENTORIES

The production of "concept inventories" has become a small cottage industry. These are tests of the *most basic* comprehension of foundations of a subject, not tests of computation. They are quite different from final exams and make no pretense of testing everything in a course. All of them trace their roots to the *Force Concept Inventory* (FCI) in physics (Halloun & Hestenes 1985a, b, Hestenes, Wells & Swackhammer 1992, Halloun, Hake et al 1995), and there is general agreement that physics education is well ahead of other disciplines in the use of concept tests as measures of teaching effectiveness. The FCI consists of multiple-choice items that test understanding of the basic foundations of Newtonian mechanics. The questions are carefully designed to test ability to use fundamental physical laws and principles in simple, qualitative, yet profound situations, where calculations are neither needed nor helpful. The FCI is designed to measure conceptual understanding that is considered to be absolutely fundamental for any useful understanding of physics. Halloun and Hestenes (1985a) say in their abstract:

“An instrument to assess the basic knowledge state of students taking a first course in physics has been designed and validated. Measurements with the instrument show that the student’s initial qualitative, common sense beliefs . . . have a large effect on performance in physics. But conventional instruction induces only a small change in those beliefs.”

Both the FCI and the CCI in calculus show that traditional instruction has remarkably little effect on basic conceptual understanding, and this has been the greatest shock to faculty. Research dating back at least 30 years has shown that most students emerge from standard introductory courses without a solid grasp of the basic concepts. This was documented in physics by Arnold Arons (1973, 1974). But prior to the development of the FCI, there was no generally accepted measure of how well students understood the basic concepts. It was thus difficult, if not impossible, to convince faculty of a need to consider changing the way they taught.

Results from research using the FCI have caused a dramatic transformation in a modest, but rapidly increasing, number of physics programs in the last ten years. There are two main reasons why the FCI has been so effective in changing views, and these are instructive for mathematics also. First, one recognizes in the FCI questions that arise in any practical use of basic principles, *including* those requiring standard computations. All acknowledge that the concepts measured are absolutely necessary (but not sufficient) for any useful understanding. Second, Hake (1998, 2001) has shown that the FCI provides a *reproducible* and *objective* measure of how a course improves *comprehension* of principles, not merely how bright or prepared the students are, nor what they have memorized. In a study of some 20 institutions, 60 classes, and 6000 students Hake compared FCI scores at entry with scores at exit. Patterns found in the data led to a performance measure that Hake calls the *normalized gain.* The FCI is administered once at the start and once at the end. The class performance is measured by the *normalized gain*, defined to be

$$g = \frac{\mu_f - \mu_0}{100 - \mu_0},$$

where μ_0 is the mean score of the class at the start and μ_f is the mean score at the end (in percent correct). This measures the gain in the class’s performance on the FCI as a fraction of the maximum possible gain. Few of the groups studied had a normalized gain much less than 0.15. On the other hand, the best performing classes in recent studies in physics have a normalized gain of about 0.70.

Hake’s findings are striking. They show that *g* is independent of the level μ_0 of the students at entrance, and largely independent of instructor and text. It is, however, strongly dependent on the teaching methodology used. Classes that used a *Traditional-Lecture* (TL) approach had an average normalized gain of 0.23 (standard deviation 0.04). In contrast, classes that used an *Interactive Engagement* (IE) approach had an average normalized gain of 0.48 (SD = 0.14), roughly two standard deviations above that of the TL classes.

A decision on what is an IE class is not trivial. Hake uses the rubric that in an IE class, students are actively engaged at all times, in developing concepts, developing strategies to solve problems of a non-routine kind, testing solutions for sensibility as well as correctness, and then Hake adds the critical component that student work in class must receive *immediate feedback*, whether from an instructor or from other students or some combination of these, feedback that requires sense-making and checks for consistency with other concepts already understood and allows the student to revise his conceptions accordingly.

The consistency and predictability of results in Hake's 1998 study, and the strong correlation with teaching methodology, make this difficult to ignore. They provide strong evidence that IE methods are more effective than TL methods. An increasing number of departments use FCI results to measure the effectiveness of physics courses, and this movement, while still small, is growing rapidly. The data and analysis have provided *objective* evidence, which convinced many to change the way they teach. The growth in this movement in physics has been impressive, and there are now concept tests in more advanced parts of physics, and new concept inventories in biology, astronomy, mathematics (the CCI), chemistry and others. The results on the CCI (nearly all data so far from TL classes) match those from the FCI, the gains all cluster around 0.15 – 0.23, with one exception (see later).

Many, particularly in mathematics, are skeptical, believing that students taught with IE are less able to do standard computational problems. There is, however, much physics research that shows otherwise. Studies by Mazur (1997), Redish (1999), Redish & Steinberg (1999), and Saul (1998) have found IE students in standard problems are no worse than those in TL courses. When he introduced *Peer Instruction*, Mazur expected — and looked for — a decline on standard problems. In *Peer Instruction* the instructor spends much less time lecturing and working examples. Still, Mazur found no difference between TL students and those using *Peer Instruction* on standard "end-of-chapter" problems. He *did* find the latter did significantly better on tests of conceptual understanding. *Workshop Physics*, developed by Priscilla Laws (1991) at Dickinson College, has no lectures at all and has produced similar outcomes (and highest *g* values of any). The studies in more basic mathematics seem to show the same thing. Schoenfeld (2002) says (page 16):

"Now, more than a decade after the publication of the Standards, hard data on large-scale implementations of these curricula are beginning to come in. To briefly summarize
1. On tests of basic skills, there are no significant performance differences between students who learn from traditional or reform curricula.
2. On tests of conceptual understanding and problem solving, students who learn from reform curricula consistently outperform students who learn from traditional curricula by a wide margin.
3. There is some encouraging evidence that reform curricula can narrow the performance gap between whites and under-represented minorities."

COGNITIVE LABORATORIES

Cognitive Laboratories (Garavaglia 2001, Ericcson & Simon 1993) are of great help in knowing what test items are really measuring, and they were used in the validation of the CCI. Scores on items and on tests can tell a lot when properly analyzed, but it is surely true that students get right answers for wrong reasons and can get wrong answers that are at least in part the fault of the item. Cognitive labs (sometimes called "analytic interviews") are a marvelous technique to discover this. They are a highly structured interview technique where individual students are asked to think out loud as they work on a problem. Probing questions are then used to access the student mental process (*not* to tutor the student!). These probing questions for each item are contained in a carefully designed protocol. It is subtle to design this protocol. We utilized consultant services to do this for the CCI. One wants to use a Lab on an item with poor discrimination (good students got it wrong and/or poor students got it right), but also on a few items that perform well, to be sure that students are not getting right answers for the wrong reasons or getting wrong answers due to a problem in wording the item.

DEVELOPMENT OF THE CCI

Some more history on the CCI is needed and relevant.

It is rare that developers of tests in mathematics submit their product to scientific validation, with the exception of national tests such as the SAT. There is a huge literature and many trained professionals who validate tests. It is a subject known as “psychometrics”. College faculty almost never consult such people. We think this is a mistake and have incorporated what is known about the validation of tests from the outset. A highly recommended resource is the volume from the National Research Council (NRC 2001). The process begins with the identification of the *fundamental constructs* that the test is designed to measure. One then puts together a panel of item writers with expertise in the subject matter and, hopefully, some knowledge of the principles of good assessment.

An outline of steps for developing and validating the CCI (or any other scientifically validated test) includes:

> *Test specifications, Item specifications, Item development*: *Item review*, (Distracters are develop *and analysis*: An item must display good ‘performance characteristics’ – statistical measures including the *p*-value, measuring the difficulty, and the discrimination, measuring how a student’s score for the item correlates with the score on the full test, and thus whether the item really does distinguish students who understand a concept from those who do not – *Field testing andanalysis, Post–examination analysis*:.

We gave the first pilot test of the CCI to about 250 students at 6 schools in the Spring of 2005. There was *no gain anywhere*, and scores were near the random guess level of 20% (even at post-test). This shocked even us. Extensive discussion among the panel led to a significant modification of the test, and to making it considerably easier. The conclusion was that if most faculty believe the test is trivial, we are probably about right. The panel developed the first field test of the CCI for the Fall of 2005 for about 1100 students in 12 American universities and one in Finland. Performance was much improved. The average scores (36%) were now well above random guess, and there was some gain everywhere. The normalized gain *g* clustered very strongly between 0.15 and 0.23, essentially the same as for the FCI in TL courses. There was one small section with *g* = 0.41. We could discover no reason for this section to be different after discussion with the instructor. We suspected teaching to the test, or it was just a statistical outlier.

We made an attempt to survey instructors in a self-administered survey on the degree of “interactive engagement” in teaching. This showed – not surprisingly – no correlation with gain score. Instructors’ own views of their interactivity are just not a satisfactory measure, and it was clear to us that all sections were predominantly lecture in any case. A set of Cognitive Labs was done with students from the Fall semester early in the Spring semester. These confirmed that all of the test items except one were indeed hitting the misconceptions they were designed to hit. Students were not being tripped up by confusing wording, or on some other unanticipated issue. The problem item was causing an issue of whether a function is increasing if it is negative and getting more negative (i.e. confusion with the magnitude). When the panel met, it was decided that the offending item could not be saved, and one other item was stripped out.

This left a ‘final’ test of 22 items. Dr. Howard Everson presented detailed psychometric analysis, which looked pretty good. Discrimination numbers were all acceptable. There seemed to be two ‘dimensions’ to the exam, which correlate well internally, but not as well with each other. These were roughly, (a) ‘Functions’ and (b) ‘Derivatives’, and a smaller third dimension on limits, ratios and the continuum. Of most interest from the psychometric point of view was the reliability coefficient, which came in at 0.7 – considered respectable, given the wide variety of testing circumstances. Professional test developers like to see 0.8, and the SAT consistently comes in around 0.85. But Dr. Everson assured us that we were quite respectable.

The second field test has just finished, and at the time of this writing we are just getting the data analyzed. Of most interest were obviously any sections that could be viewed as clearly alternative teaching (IE) methods. We got data from Uri Treisman’s group at the University of Texas, and from the *Calculus with Mathematica* group at the University of Illinois (Davis et al, 1994). The group from Illinois came in at the same *g* as the lecture based groups (quite low). We suppose changing to *Mathematica* by itself is insufficient to see any change without some thought about how the computer exercises work with the student mind — but actually we do not yet have any firm handle on this. The most optimistic results were from Texas. Uri Treisman has received a MacArthur award for his extraordinary history of success in mathematics instruction. He did not expect much, he said, because he was stuck with a large class of some 85 students. Nevertheless, he came in with $g = 0.30$ which is well outside the range of all the standard lecture based sections (0.15 to 0.23), though significantly lower than what was seen in physics. Obviously the amount of data from good alternative instruction is far too small for any final conclusions, and the foundational question of whether teaching methodology strongly affects gain (on the CCI) as it does for physics (on the FCI) will have to await further data. Significant new data will come from several IE programs this Spring semester (2007). In the next few months, we will publish the results and make the test available to qualified faculty via the web. Others will use it and accumulate meaningful data far faster than we could. This is what happened with the FCI.

The explosion of physics education reform arose *after* the publication of the FCI, and use of the test did in fact feed back into improved education. The dramatically improved gain scores (up to 0.70) arose over a period of 13 years between Halloun and Hestenes' publication of the original test and Hake's analysis. We expect something quite similar to happen with the CCI.

V. PRINCIPLES OF EVALUATION

In 1998 the National Research Council (NRC), with support from NSF, convened a Committee on the Foundations of Assessment, which produced *Knowing What Students Know* (NRC, 2001). This volume assembled into useable form for both researchers and educators what was known about the science of assessment in education. The field has exploded in the past 20 years, and resources of knowledge vastly exceed what was known in 1985 when the FCI was developed. I cite central points.

The NRC study stresses throughout that assessment rests on three pillars: *cognition*, *observation*, and *interpretation*. Very briefly: 'Cognition' is a *model* of how the student represents knowledge and develops deeper understanding of concepts. 'Observation' means a method of taking data (not necessarily just pencil and paper tests). 'Interpretation' means how the researcher or educator evaluates data from observation to learn what is actually happening in the student's mind. We add a fourth leg, that quality assessment must be aligned with curriculum and teaching to have any effect.

1. Evaluation must make visible the processes happening in the student mind when faced with a task. This was done explicitly in the design of the CCI.
2. Assessment should distinguish between short-term recall and "expert" use of concepts stored in long-term memory. Assessments claimed to test for higher level functioning should set tasks that require higher level cognition, not just recall of facts or formulas.
3. Design of assessments should consider the types of mental processes that the student is expected to use in order to demonstrate competence. NRC says (page 62)

   > *"This information is difficult to capture in traditional tests, which typically focus on how many items examinees answer correctly or incorrectly, with no information being provided about how they derive those answers or how well they understood the underlying concepts. Assessment of cognitive structures and reasoning requires more complex tasks revealing information about thinking patterns, reasoning strategies, and growth in understanding over time."*

4. Assessment should provide information that can be used to improve instruction (our fourth leg).

   > *" . . . most current large-scale tests provide very limited information that teachers and educational administrators can use to identify why students do not perform well. . . . Such tests do not reveal whether students are using misguided strategies to solve problems or fail to understand key concepts. . . Indeed, it is entirely possible that a student could answer certain types of test questions correctly and still lack the most basic understanding."* NRC, page 27

VI. CONCLUSION

Even if later results show clear superiority of IE approaches in calculus as they did in physics, the hardest part of making progress will be to reach traditional mathematics faculty who usually do not go to meetings, or sessions at meetings, involving mathematics education. There is no easy fix for this, since, at national meetings and most regional meetings, it is abundantly clear which are the education sessions. They are clearly delineated in conference programs, and are often actually under different sponsorship (M.A.A. instead of A.M.S.). There are many who will avoid like the plague the “education” type, though this is improving. There is now one advantage that some sort of improvements will actually take hold, an advantage which did not exist a generation ago. There is a large research base on what "learning” actually means and how students actually learn. From that research base flows the current consensus among most of the relevant national organizations on the broad shape of the path that should be pursued. Thus it seems that the national organizations themselves will likely lead in trying to acquaint the greater mass of college faculty and pre-college school system personnel at least that there is a serious field of real research into education in mathematics and science. Many do not believe it (or, at the pre-college level, are frightened of it), and it is a great impediment to progress. It is our hope and expectation that a validated CCI will be a tool to aid the national organizations in this effort.

REFERENCES:

Arons, A. (1973) Toward a Wider Public Understanding of Science, *Am. J. Physics.* 41(6), 769-782

Arons, A. (1974) Addendum to Toward a Wider Public Understanding of Science, *Am. J. Physics*, 42(2), 157-158

Davis, B., H. Porta, and J. Uhl, (1994) Calculus and Mathematica, Addison–Wesley.

Ericcson, K.A. & Simon, H. (1993), Protocol Analysis, MIT Press

Garavaglia, D. R. (2001). The Relationship Between Item Format and Cognitive Processes in Wide-Scale Assessment of Mathematics. Unpublished doctoral dissertation.

Hake, R. R. (1998a) Interactive Engagement Methods in Introductory Mechanics Courses, <http://www.physics.indiana.edu/~sdi>.

____ (1998b) Interactive Engagement Versus Traditional Methods: a Six–thousand Student Survey of Mechanics Test Data for Physics Courses, *Am. J. Physics* 66, 64. <http://www.physics.indiana.edu/~sdi>.

____(2001) Lessons from the Physics Education Reform Effort, <http://www.physics.indiana.edu/~hake>

Halloun, I. and D. Hestenes, (1985a) Common Sense Concepts About Motion, *American Journal of Physics* 53, 1056–1065.

___(1985b) The Initial Knowledge State of College Physics Students, *American Journal of Physics* 53, 1043–1055.

Halloun, I., R. R. Hake, E. P. Mosca, and D. Hestenes, (1995) Force Concept Inventory (Revised), <http://modeling.la.asu.edu/modeling.html>.

Hestenes, D., M. Wells, and G. Swackhammer, (1992) Force Concept Inventory, *Physics Teacher* 30, 141–158.

Laws, P.(1991) *Workshop Physics: Learning Introductory Physics by Doing It*, Change 23, 20–27.

Mazur, E. (1997) Peer Instruction: A User’s Manual, Prentice–Hall, 1997.

National Research Council (2001), Knowing What Students Know: The Science and Design of Educational Assessment, Pellegrino, J., Chudnowsky, N., and Glaser, R., National Academy Press. Washington, DC

Redish, E. F. (1999): *Millikan Lecture:* Building a science of Teaching Physics, *Am. J. Physics* 67, 562–573.

Redish, E.F. and R. N. Steinberg, (1999) Teaching Physics: Figuring Out What Works, *Physics Today* 52, 24–30, <http://www.physics.umd.edu/rgroups/ripe/perg/cpt.html>.

Saul, J. M. (1998) Beyond Problem Solving: Evaluating Introductory Physics Courses Through the Hidden Curriculum, Ph.D. thesis, University of Maryland.

Schoenfeld, A.H. (2002): Making Mathematics Work for All Children, *Educational Researcher* 31 #1, 13-25