The Calculus Concept Inventory—Measurement of the Effect of Teaching Methodology in Mathematics

Jerome Epstein

Introduction

The Calculus Concept Inventory (CCI) is a test of conceptual understanding (and only that—there is essentially no computation) of the most basic principles of differential calculus. The idea of such a test follows the Mechanics Diagnostic Test (MDT, Halloun and Hestenes [11], [12]) and its successor the Force Concept Inventory (FCI) in physics (Hestenes, Wells, and Swackhammer [14]; Halloun et al. [13]; Hake [7], [8]), the last a test which has spawned a dramatic movement of reform in physics education and a large quantity of high quality research. The MDT and 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 students knew at exit and that a semester of instruction made remarkably little difference. More dramatic, the pre-to-post test average normalized gains g on the MDT and FCI in Hake's ([7], [8]) meta-analysis showed a strong correlation with teaching methodology: the average g for "interactive engagement" (IE) courses exceeded the average g for "traditional" (T) courses by about two standard deviations. No other variable, including the pretest score, had anywhere near this correlation with the gain.

Mathematics education is often mired in "wars" between "back-to-basics" advocates and "guided-discovery" believers. There seems to be no possibility of any resolution to this contest without hard, scientific evidence of what works. Such evidence requires widespread agreement on a set of very basic concepts that *all* sides agree students should—must—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.

Jerome Epstein was a faculty member of the Department of Mathematics at Polytechnic University. He continues to be affiliated with the Department of Mathematics at the same institution, now known as Polytechnic Institute of New York University. His email addresses are jerepst@att.net *and* jepstein@poly.edu.

DOI: http://dx.doi.org/10.1090/noti1033

The CCI has undergone extensive development and validation, funded by the National Science Foundation (Epstein and Yang [4]). It was developed by a panel of respected calculus educators and a consultant, nationally known for development and validation of standardized tests. The test shows good performance characteristics (see below) and exposes exactly what the FCI in physics showed. From the fall semester of 2006 through the fall semester of 2008, the test was given to approximately 5,000 students at test sites with many traditional programs as well as programs using alternative teaching methodologies. Test sites were in colleges and universities in the United States and about fifteen other countries. Comparison of gain from two widely divergent methodologies then becomes possible and is discussed in this paper. This paper also discusses the development and validation process in some detail. Those interested in using the test should contact this author (Epstein, email addresses in the footnote on the first page of this article). While this paper was being prepared, additional requests for the test have come in, typically one or two per week, but results received after 2006 are not included in this paper. All the more recent results seem to follow the pattern described below.

The CCI is the second in an anticipated series of Basic Conceptual Understanding instruments 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. The first is the Basic Skills Diagnostic Test (BSDT), which has already been used quite widely nationally. We are currently seeking funding for an Elementary Algebra Concept Inventory.

Concept Inventories

The production of "concept inventories" has become a small cottage industry. There does not seem to be a universally accepted definition of what constitutes a concept inventory, and the term is being used now quite widely. We use the following, taking the example of the FCI in physics: These are tests of the *most basic* conceptual comprehension of foundations of a subject and not of computation skill. 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 MDT and FCI in physics, and there is general agreement that physics education is ahead of other disciplines in the use of concept tests as measures of teaching effectiveness and in the development of programs that show much improved gain. The FCI consists of multiple-choice items that test understanding of the basic foundations of Newtonian mechanics. The questions are carefully designed to test one's 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 [11] 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 in physics 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 more than thirty years has shown that most students emerge from standard introductory courses without a solid grasp of the basic concepts. This was clearly documented in physics by Arnold Arons ([1], [2]). But, prior to the development of the FCI/MDT, there was no generally accepted measure of how well students understood the basic foundations. 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 twenty years. There are two main reasons why the FCI has been so effective in changing views, and these are instructive for mathematics also. First, faculty recognize 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 ([7], [8], [9]), and subsequently many others (shown in Hake [10]), 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, 100 classes, and 6,000 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*, <*g*>. The FCI is administered once at the start and once at the end of a first course in physics. 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* (not of individuals, hence the angle brackets) 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 more recent studies in physics have a normalized gain up to about 0.70. The question of calculating g for individual students is quite tricky; we do not consider it here.

Hake's findings are striking. They show that g is independent of the level μ_0 of the class at entrance (the correlation of g with μ_0 for the sixty-two courses surveyed was +0.02) and largely independent of instructor and text. It is, however, strongly dependent on the teaching methodology used. Classes that used a Traditional (T) approach, depending on "passive-student lectures, predefined laboratories and algorithmic homework and tests" (Hake's choice of words), 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 (standard deviation = 0.14), roughly two standard deviations above that of the T classes. The consistency and predictability, and the strong correlation with teaching methodology, make this difficult to ignore. The need for similar information in calculus, and math in general, is the prime motivation for this project. New data from calculus (see below) show exactly the same pattern. 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 attempt changes in the way they teach and to seek validation from the test. 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 new results on the CCI match those on the FCI (scores are even a bit lower, though the same pattern is seen); the gains are in the range 0.08—0.23.

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 appears to show otherwise. Studies by Mazur [15], Redish [17], Redish and Steinberg [18], and Saul [20] have found IE students solving standard problems are no worse than those in T courses. When he introduced Peer Instruction, Mazur expected—and looked for—a decline on standard "end-of-chapter" problems. In Peer Instruction the instructor spends much less time lecturing and working examples. Still, Mazur found no difference between T students and those using Peer Instruction. He did find the latter performed significantly better on tests of concept understanding. The mission of this project is to see if the same pattern holds for calculus instruction.

The studies in more basic mathematics (often back to elementary school) seem to show the same thing. Schoenfeld [21, p. 16] says:

> Now, more than a decade after the publication of the *[NCTM] 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 underrepresented minorities.

History of the CCI—What Defines IE?

In 2004, Epstein and Yang obtained NSF funding for the development and validation of a Calculus Concept Inventory (CCI). Howard Everson became consultant on that project in 2005 and has done validation studies on the test and the student data. More schools become involved on a regular basis (this author gets new requests for the test on a weekly basis), but we give the basics below. There has been no difficulty in getting T data from a good many schools. However, it has not been easy at all to find adequate numbers of clearly IE teaching sections in Calculus I to allow comparison with the T sections and thus get a handle on the effect of the different methodology, at least as measured by the CCI. In order to do this, one needs an *independent* definition of what IE means in practice. Clearly, simply lumping all of the high gain sections into IE will not be very convincing. We use as a definition essentially what was used by Richard Hake in his foundational paper (Hake, [8]).

> "Interactive Engagement" (IE) methods are those designed at least in part to promote conceptual understanding through interactive engagement of students in heads-on (always) and hands-on (usually) activities which yield immediate feedback through discussion with peers and/or instructors.

Our data in calculus indicate that the "immediate feedback" part of this is crucial, and the mere avoidance of lecture in favor of some alternative methodology does not *ipso facto* produce any change in <g>. The importance of immediate formative assessment has been emphasized by Black and Wiliam (1998) and Shavelson (2008). The feedback may be from an instructor or from other students or some combination of these. Feedback in this context means that students check the extent to which their answers are sensible and check for consistency with other concepts they already understand. The class structure must allow *time* for them to back up, revise strategy, and try again *immediately*. Programs based on computerized instruction have, so far, shown the same gain as the lecture sections. What is critical, and the great advantage conferred by an instrument like the CCI, is that one has a consistent measure of outcome, on which nearly all faculty will agree that students *should* succeed, to check what is making a difference and what is not.

Development of the CCI

A panel of expert faculty[1] with decades of experience was assembled and produced the first form for pilot testing in the spring semester of 2005. This was preceded by extensive discussion to agree on a small set of basic constructs (behaviors we believed the test should measure) and content domains to be covered (limited to Calculus I, differential only). From these, the writing of items began and a pilot test was given first in February 2005 as a pre-test. There were about 250 students at six institutions. It showed results mostly at the random guess level, even though substantial numbers of students had taken calculus previously. This was actually not a surprise. At the end of the semester we found that there was *no gain anywhere*, and some classes scored lower than at pre-test (though not by a significant amount). We had shown, at least as measured by that pilot test, the same thing that was shown by Halloun and Hestenes: a semester of instruction had essentially no effect at all. We rather expected that, but "no gain anywhere" was clearly a shock, even to us.

A reappraisal was in order. We were coming to the conclusion that, as basic as the items were, they were too hard for the purposes for which the test was intended. One department head said the questions needed to be at a level of "point to your foot" and there will still be students who will get it wrong. We came to the conclusion that the items needed to be at a level where most faculty would believe the items were utterly trivial, and we would be about right. A revised field test was developed for administration in the fall of 2006. Except for two items (see below), this is the form in which the CCI now exists. It was given in the fall of 2006 to 1,100 students at fifteen institutions in the U.S. and one in Finland (no difference between Finland and the U.S.). Pre-test scores were now significantly above random guess levels, and at semester's end there was some gain everywhere (not much at most places). We seemed to be, at least on the surface, where the FCI had arrived after some years of development.

Dr. Everson came on board the team in the summer of 2005 and began with his graduate student to do evaluations of the results with the intent of ascertaining the validity of the instrument from the psychometric point of view. We discuss this below. A set of Cognitive Laboratories was done (also see below) to give us a better window into the student mental process when faced with the test items, and then a second field test was given in the fall of 2007. We have attempted to keep the project going after the ending of the grant, in order to obtain adequate data from IE sections to appropriately test the fundamental hypothesis of a connection between gain and teaching methodology. Some quite dramatic results were obtained in the fall 2008 semester.

Results on the CCI through 2007

Early on, 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 generally just not a satisfactory measure (however, see below on the results from the University of Michigan), and it was clear to us that all sections were predominantly lecture. During the field tests of the fall semesters of 2006 and 2007, of most interest were obviously any sections that could be viewed as clearly alternative teaching methods. This is a critical issue. We got data from Uri Treisman at the University of Texas, from a strongly IE-based instructor at Oregon State University (through Tevian Dray, along with several traditional sections), and from two sections at St. Mary's College of Maryland, of clearly IE methodology (David Kung). The absence of lectures is not sufficient. Assigning conceptual homework problems and going over them in class the next day is not sufficient. Real IE sections are not so easy to find.

All of these IE sections showed g between 0.30 and 0.37. While this gain does not seem very high, especially when compared with results from physics, it is well outside the range of what we

[1] *D. Yang (Polytechnic), C. Shah (Polytechnic (deceased)), D. Flath (Macalester), M. Terrell (Cornell), K. Vincent (Washington State) and M. Robinson (Seattle).*

were seeing for T sections. This was enticing and optimistic but simply not enough data, we felt, to make any broad conclusions. But very recent data have improved matters dramatically. In the fall semester of 2008, there was more data from Oregon State and once again the same instructor had dramatically better gain than the lecture-based sections. But most important we obtained a large amount of data from the University of Michigan. We discuss this next. These results were reported at the Joint National Meetings in Washington (Rhea [19]).

University of Michigan Calculus Program

In the fall of 2008, the CCI was administered to all sections of Calculus I at the University of Michigan (U-M). Introductory mathematics courses at U-M are taught in sections of thirty-two (or fewer) students (though this is by no means unique to U-M). In the fall of 2008, there were fifty-one sections of Calculus I, with 1,342 students. Homework assignments and all exams (two midterms and a final) are uniform, but instructors have full responsibility for teaching their individual section(s)—delivery of material, preparing and grading in-class quizzes, and grading collected homework. All sections of the course used the Harvard Project Calculus text (Hughes-Hallett et al., 1998).

Eighteen of the fifty-one sections (35 percent) were taught by instructors new to the course. For many years, U-M has trained its instructors to teach in an IE style. All new instructors attend a week-long presemester training workshop and weekly coursewide meetings to support this instruction. Experienced instructors are also encouraged to attend weekly meetings, and most do. The new instructors were new postdocs and graduate students, either new to teaching or new to the course. In order to preserve the integrity of the test, the CCI was administered online in a proctored lab. No instructors had access to the test. Students took the pretest during the first week of class and the post-test during the last week of class. Students were allowed thirty minutes and one attempt on each test. As incentives, the pre-test counted as credit for an in-class quiz grade. The post-test counted for 5 percent of the final exam grade. Of the 1,342 students who completed the course, 1,284 students (96 percent) took both the pre- and post-tests. This alone is *quite extraordinary* and, we think, an important independent result. Generally, drop-out rates from pre-test to post-test have been much higher in other schools. The g values at all schools have been calculated only with students who completed both tests, but we have checked that the drop-outs are randomly distributed among the pre-test scores.

We were interested in the students' perception of whether the class was taught in an interactive classroom style, so students were asked two questions on a preface page to the CCI post-test. Instructors were asked corresponding questions on an anonymous survey. These questions were:

- If an interactive classroom is one in which students actively work on underlying concepts and problems during the class and receive feedback from the instructor or other students on their work in class, how would you describe your class this semester: Very Interactive; Interactive; Somewhat Interactive; Not Interactive?
- On average, about what percent of your time in class would you say was spent with you working on problems and receiving feedback from your instructor and/or your classmates: 76–100 percent; 51–75 percent; 26–50 percent; 1–25 percent; 0 percent?

Tabulation of Results at U-M:

The results of the CCI at U-M include the following:

- The average gain over all fifty-one sections was 0.35.
- Ten sections had a gain of 0.40 to 0.44.
- The range of the gain scores was 0.21 to 0.44.
- Assignments to "interactivity" responses were 4 = very interactive; 3 = interactive; 2 = somewhat interactive; and 1 = not interactive.
- Coursewide, the average student perception of "interactivity" was 2.7; range 1.8 to 3.7.
- For the ten sections with a gain over 0.40, the average interactivity score was 3.02.
- Percent of time in class spent on "interactive engaged" activities was defined by taking the midpoint of the percent range chosen (e.g., 51–75 percent was scored as 63 percent, etc.). Student perception of the percent of time spent on these activities showed the following:
- Average for the overall course: 48 percent; range 30 percent to 74 percent.
- Average percent time in the ten high gain sections: 55.1 percent.

We note that the *lowest* gain at U-M was 0.21, which was about the *highest* gain among the T sections. The lowest gain section at U-M contained twelve students, all known to be at risk for failure. That this section had the same gain as the *highest* T section seems to be quite dramatic. The next lowest section at U-M had a g of 0.27, significantly above *all* the T sections at *all* schools. There are still questions of interpretation, and we claim no final conclusions. We discuss at the end some of the possible alternate explanations.

Follow-up and Analysis: Students at U-M and the other institutions who have contributed to the study were all given the same test. The analysis of gain, <*g*>, was determined using the same measure at all institutions, and this definition models the measure used in physics. Previous calculus sections showing gain at or above 0.30 have been from sections taught in a clearly IE style (based on self-descriptions of how the class is run). The fact that over one-third of the sections at U-M were taught by instructors new to the course demonstrates that this style of teaching can be fostered and developed.

This follows the history of the FCI quite well. However, the results thus far indicate that, as a community, this is surely worthy of pursuit. There are numerous ways that the FCI has changed the face of physics instruction. (Most recently, see the *NY Times* article regarding physics instruction at MIT: `http://nytimes.com/2009/01/13/us/13physics.html?emc=eta`.) Physics has led the way, but we now have a means to pursue this analysis in mathematics. We are happy to see more schools use the CCI and contribute to the data and the dialog. Please inquire with this author.

Recent Results

In recent years, the CCI has been given outside the United States in three provinces of Canada and about a dozen other countries. The results will surprise many. While there are small variations between countries, and between those countries and the United States, the results are generally very much the same, with a few striking exceptions, and there is no dramatic difference between these countries and the U.S. overall. In spite of the common mythology, the problem of basic conceptual understanding is extremely widespread. This result has been a source of some distress not only in the U.S., but also in almost all the other countries—except one.

The results in China (Shanghai) looked very much like the results from Michigan. They were a reasonably normal distribution, but two standard deviations above the general U.S. results, except for the clearly Interactive-Engagement sections in Michigan, the sections with Uri Treisman (Texas), David Kung (St. Mary's College, MD), and I think two other small sections. The size of the difference is enormous—two Gaussians with no overlap. Clearly we know about the Interactive-Engagement methods in the U.S., but some explanation is needed for the results in China. We have no conclusive explanation for this and hope to do more investigation at some point.

There is a common (mis-)conception that Chinese students do well only because they have been drilled hard and have become very good at memorized, low-level tasks. The data from the CCI clearly indicate that this view is insupportable, and the Chinese are doing something much more fundamentally right. It looks like the conclusion will be that the rest of the world has a great deal to learn from China in mathematics education, and almost certainly, one needs to go back to much earlier stages than calculus. We hope that we, or others, can pursue this further in the future.

Validation of the CCI

Cognitive Laboratories (Garavaglia 2001, Ericcson and Simon 1993) are of great help in knowing what test items are really measuring in the student mind, 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 in discovering this phenomenon. They are a highly structured interview technique where individual students (only a modest number of course) are asked to think out loud as they work on a problem. Probing questions are then used to access the student's 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. Cognitive Labs are helpful 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 wrong reasons, or getting wrong answers due to wording of the item.

A set of Cognitive Labs was done with students from the fall semester of 2006 early into the following 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 panel stripped out this item and one other, where we were not satisfied that we had found an item to measure the construct of "the derivative as a measure of the sensitivity of the output to change in the input"—that is, essentially the tangent line approximation. This left a final test of twenty-two items. Dr. Everson presented to us a 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 interest from the psychometric point of view was the reliability coefficient, which came in at 0.7—considered modestly 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 our results were respectable. There is much more data analysis possible with the data we already have, and more comes in on a regular basis. We expect a further report on validation and on the deeper meaning in future years.

Interpretation and Conclusions—the Future

On the surface the results do look good, but caution is always in order. Are there alternative explanations for the results seen? The data sets for T and IE sections are widely spread apart, essentially two Gaussians with no overlap, so an explanation on the basis of random statistical fluctuation seems quite unlikely. The effect size is about 2.0, highly significant. The question remains of whether there might be an unanticipated *systematic* error, some variable which might be skewing the results. We have made an attempt to think of any variables that might be hidden in this data which could distort the results in apparent favor of the IE sections. We make no claim that these results are final. We have every expectation that many institutions will be using the test in the coming years and their results will independently either verify or call into question these results.

We have considered the following possible confounding variables:

- Class size,
- Instructor experience,
- Assignment to IE or T categories,
- Textbook,
- Time spent in calculus class,
- Student preparation at entrance.

Class size: We have little data from large IE sections. It is no small feat to do Interactive Engagement in a large class. The only approach to doing this that we know of is (for ordinary faculty) the use of Personal Response Systems (so-called "Clickers"). This is also a development that comes from physics. We are not aware of any implementation of this methodology in large calculus classes but would be happy to know of results if any such exist. We do, however, have lots of data from small calculus classes. The class sizes at U-M where most of the IE data come from are all less than or equal to thirty-two. But that is also the case for the Polytechnic data, Cornell, and most of the other T sections. The small classes do not appear to have made any difference in those cases. There is one case of a relatively large IE class. Uri Treisman at the University of Texas, Austin, had a class of eighty-five and did not expect good results. His gain of 0.30 falls into the IE range and corresponds with his well-known commitment to interactive teaching.

Instructor experience: It seems very clear that this variable has no effect. A large fraction of the teachers at Michigan were brand new to the course or brand new to teaching entirely, whereas all at Polytechnic had many years of experience. Yet the g values at Polytechnic were much lower.

Assignment to IE or T categories: We have done the best we could to make this assignment based on independent information on how the class runs, and not on the gain score after the fact. The largest body of IE data is from Michigan, and they have had a strong program of training and supervising IE teaching for more than ten years (described above). The other cases of IE are also clearly within the definition, although this is based on self-reporting. One interesting case is a school with a long-time program in calculus using Mathematica. This course produced the same range of gain scores as the T sections. The lack of lectures alone seems to make no difference.

Textbook: Both Michigan and Polytechnic (and we suspect others as well) used the CCH book from Harvard and had done so for many years. The results are dramatically different. We made no systematic attempt to investigate this, and more data on this possible variable will appear in the future.

Time spent in calculus class: As far as we know, all of the sections at all schools were in the typical range of 4±1 hours per week in class (including some recitation). We saw no pattern indicating a significant effect. We have seen anecdotal reports that schools who immerse students for much longer hours (as many as ten per week, largely working with student peer tutors) have shown clear improvements in pass rates. We have no further information on this, and we note that multiple hours of one-on-one or one-on-few teaching might qualify as IE methods anyway.

Student preparation at entrance: We can examine any correlation of the normalized gain with the pre-test score. Such a correlation, if significant, would indicate that the normalized gain is in part predicted by the level of the student at entrance. The information we have indicates that, just as in physics, there is no significant correlation between the entry level score and the normalized gain. For example, at the University of Michigan, this correlation turns out to be +0.04. A similar result (no correlation) is also found by Hake for the FCI in physics. The correlation is quite small. One would expect that a reasonable measure of preparation, especially for this test, would be the pre-test score. We also asked on the test whether the student had previously taken calculus, either in high school or college. In general, no effect was seen on g from the student taking calculus previously. The data from Michigan seem to show actually that those who have taken calculus previously have somewhat smaller gain. At this point, we do not attribute any significance to this, though it might turn out later to be interesting.

We expect that, with the continuing accumulation of more data, the possible effect of any confounding variables will become clearer, and again we strongly encourage others to join the project and make use of the test.

How is the future of the CCI likely to evolve? Can we learn something of this from the history of 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 thirteen years between Halloun and Hestenes's publication of the original test and Hake's analysis. We expect something quite similar to happen with the CCI.

Richard Hake has communicated the following to me:

> I agree with your view of the dissemination history as outlined in your proposal. I think the papers by Halloun and Hestenes [11], [12] were virtually unknown outside the tiny and nearly invisible Physics Education Research community until Mazur, Redish, and Hake gave them wider publicity.

The real explosion began following Hake's work of 1998. This showed rather conclusively that the effects of IE programs were not due to enrollment of better students, nor to smaller classes, nor to better instructors who would have fared equally well with T programs. It was the *teaching methodology* that made a two standard-deviation difference. The experience at Michigan strongly supports this conclusion, as we see from the uniformly improved results across a large group of instructors, many with no prior experience in the methodology or even in teaching at all.

The need for a CCI or something similar was already very clear in 2004 when this project began. As of the date this is written (June 2012), further requests for the test come in on a regular basis (one or two each week). In the coming years, we will attempt to assemble the large mass of data, item by item, that should be out there. It should be a very worthwhile repository of valuable information on mathematics education in general. There had already been two decades of papers and articles making clear that some sort of evaluation of reform programs was critical. Presentations at conferences have consistently led to requests for the test and in particular postings on the Web at sites frequented by people in mathematics education research have spread wide interest and greatly increased demand for the test. Those who know about it are now doing presentations themselves at meetings on calculus instruction, and that is how the department leader at the University of Michigan (Rhea) heard about it (we have made a consistent attempt to have everyone requesting the test agree to a set of nondisclosure requirements). The consciousness that some sort of scientific means of "deciding what works" is needed in all of mathematics. The Web gives an immediate means of propagating information that such an instrument exists and has already produced many requests for the test and one request to use it in Ph.D. thesis research. It is our hope that the CCI and other instruments to follow will provide a research tool that can begin to provide answers to issues that have been around for a very long time.

Acknowledgments

I am pleased to acknowledge with gratitude the support of the National Science Foundation program in Course, Curriculum, and Laboratory Improvement (CCLI). This research would not have been possible without it. I also gratefully acknowledge extensive comments and suggested revisions from Richard Hake. Also I gratefully acknowledge help and suggestions from Howard Everson of Fordham University and Karen Rhea of the University of Michigan.

References

[1] A. Arons, Toward a wider public understanding of science, *Amer. J. Physics* **41**(6), 1973, 769–782.

[2] ______, Addendum to Toward a Wider Public Understanding of Science, *Amer. J. Physics* **42**(2), 1974, 157–158.

[3] J. Epstein, What is the Real Level of Our Students? or What Do Diagnostic Tests Really Measure?, American Association of Physics Teachers conference, College Park, MD, August 1996.

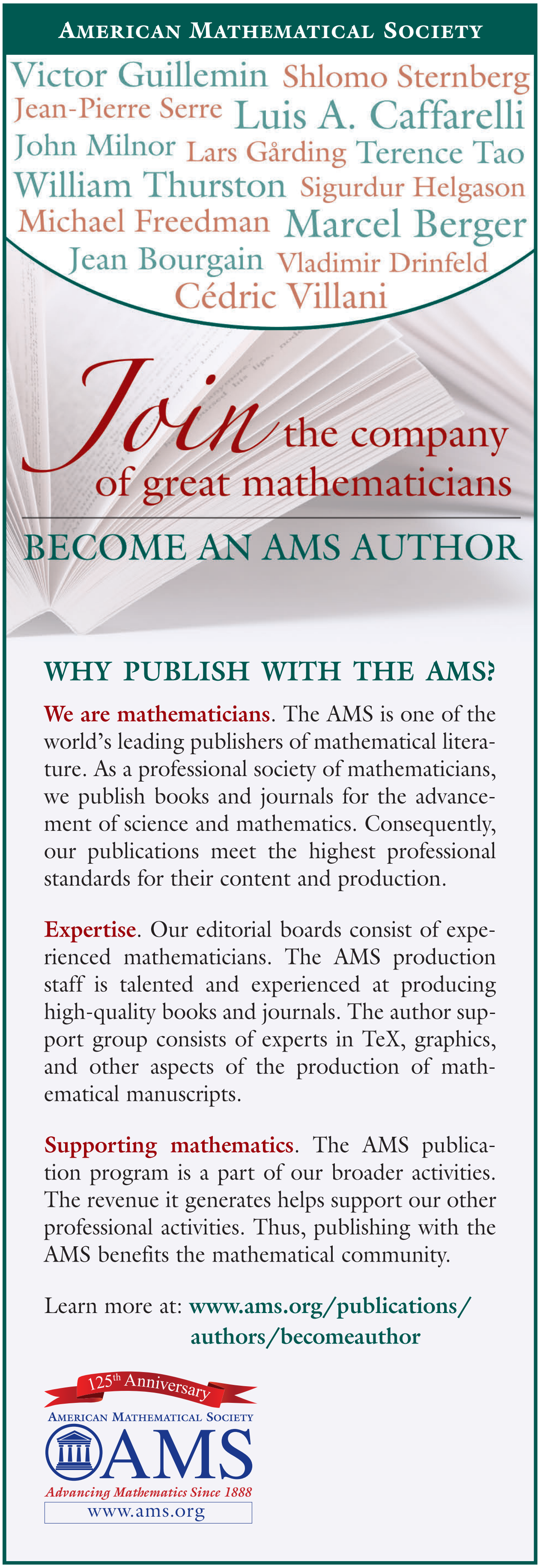
[4] J. Epstein and D. Yang, Development and Validation of the Calculus Concept Inventory, National Science Foundation, CCLI program, Grant # 04-04818, 2004.

[5] ______, Development and validation of the Calculus Concept Inventory, in *Proceedings of the Ninth International Conference on Mathematics Education in a Global Community*, September 7–12, 2007, edited by Pugalee, Rogerson, and Schinck; online at http://bit.ly/bqKSWJ. (48 kB).

[6] ______, The Calculus Concept Inventory—New Data. Correlation with Teaching Methodology, Joint Mathematics Meeting, January 5–10, 2009; abstract online at http://bit.ly/18WehnP. (32 kB).

[7] R. R. Hake, (1998a), Interactive engagement methods in introductory mechanics courses, 1998, http://bit.ly/aH2JQN.

[8] ______, (1998b), Interactive engagement versus traditional methods: A six-thousand student survey of mechanics test data for physics courses, *Amer. J. Physics* **66**(1), 1998, 64–74, http://bit.ly/9484DG.



[9] ______, Lessons from the Physics Education Reform Effort, *Ecology and Society* **2**: 28, 2002, `http://bit.ly/aL87VT`.

[10] ______, (2008), *Design-based research in physics education research: A review*, in *Handbook of Design Research Methods in: Education: Innovations in Science, Technology, Engineering, and Mathematics Learning and Teaching* [Kelly, Lesh, and Baek (2008)]—publisher's information at `http://bit.ly/dkLabI`; a pre-publication version of Hake's chapter is online as a 1.1 MB pdf at `http://bit.ly/9kORMZ`.

[11] I. HALLOUN and D. HESTENES, (1985a), Common sense concepts about motion, *Amer. J. Physics* **53**, 1985, 1056–1065.

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

[13] I. HALLOUN, D. HESTENES, R. HAKE, E. P. MOSCA, and D. HESTENES, Force Concept Inventory (revised), 1995, `http://bit.ly/b1488v` (scroll down).

[14] D. HESTENES, M. WELLS, and G. SWACKHAMMER, Force concept inventory, *Physics Teacher* **30**, 1992, 141–158.

[15] E. MAZUR, *Peer Instruction: A User's Manual*, Prentice-Hall, 1997.

[16] C. A. PAULSON, An exploratory study of cognitive laboratories for development and construct validation of reading and mathematics achievement test items, dissertation, Univ. of Pennsylvania, 1999; abstract online at `http://bit.ly/1OUNXoU`.

[17] E. F. REDISH, Millikan Lecture: Building a science of teaching physics, *Amer. J. Physics* **67**, 1999, 562–573.

[18] E. F. REDISH and R. N. STEINBERG, Teaching physics: Figuring out what works, *Physics Today* **52**, 1999, 24–30, `http://bit.ly/13IYrvv`.

[19] KAREN RHEA, Using the CCI at a Large Research University, presentation at the 2009 Joint National Meetings of AMS/MAA, Washington DC, January 2009.

[20] J. M. SAUL, Beyond problem solving: Evaluating introductory physics courses through the hidden curriculum, Ph.D. thesis, University of Maryland, 1998.

[21] A. H. SCHOENFELD, Making mathematics work for all children, *Educational Researcher* **31** (1), 2002, 13–25.

[22] R. J. SHAVELSON, 2008, Formative assessment, guest editor's introduction, special issue of *Applied Measurement in Education*; online at `http://stanford.edu/dept/SUSE/SEAL/`.