Measuring Teaching and Learning Performance: Interconnected Issues*

Richard Hake <rrhake@earthlink.net>, <http://www.physics.indiana.edu/~hake>
Indiana University, Emeritus

From Teaching to Learning

In their landmark wake-up call to American higher education “From Teaching to Learning: A New Paradigm for Undergraduate Education,” Barr & Tagg (1995) wrote: “A paradigm shift is occurring in American higher education. Under the traditional, dominant ‘Instruction Paradigm’ colleges are institutions that exist to provide instruction. Subtly but profoundly, however, a ‘Learning Paradigm’ is taking hold, whereby colleges are institutions that exist to produce learning. This shift is both needed and wanted, and it changes everything.”

Consistent with the broad theme of this international conference “Measurement and Evaluation in Education,” I think it might be appropriate to generalize Barr and Tagg’s rallying cry as follows:

“A paradigm shift is desirable in world-wide education. Under the traditional, dominant ‘Instruction Paradigm’ schools and colleges are institutions that exist to provide instruction. It is to be hoped that a ‘Learning Paradigm’ might eventually take hold, whereby educational institutions in all countries would exist to produce learning. This shift is both needed and wanted, and it would change everything.”

What to Measure and How to Measure

Investigation of the extent to which a paradigm shift from teaching to learning is taking place requires measurement of students’ learning in school and college classrooms. But Wilbert McKeachie (1987) has pointed out that the time-honored gauge of student learning – course exams and final grades – typically measures lower-level educational objectives such as memory of facts and definitions rather than higher-level outcomes such as critical thinking and problem solving. Furthermore, the same criticism (Hake, 2002a) as to assessing only lower-level learning applies to Student Evaluations of Teaching (SET’s), since, at least in American higher education, their primary justification as measures of student learning appears to lie in the modest correlation with overall ratings of course (+ 0.47) and instructor (+ 0.43) with “achievement” as measured by course exams or final grades (Cohen, 1981).

For general characterizations of higher-order learning see Anderson & Krathwohl (2001) and Shavelson & Huang (2003). The latter, in their Chart 1 - “Framework for Cognitive Outcomes,” display higher-level learning within knowledge domains, as might be measured and enhanced by disciplinary experts: “procedural - knowing how” – see, e.g., Anderson (2004); “schematic – knowing why”; and “strategic – knowing when certain knowledge applies, where it applies, and how it applies.” These contrast with the lower-order “declarative - knowing that.”


◊ Partially supported by NSF Grant DUE/MDR-9253965.
How then can we measure students’ higher-level learning? In this paper I shall concentrate on the measurement of learning in American higher education because: (a) my experience has been primarily in that area, (b) judging from the paper by Othman et al. (2006) “Accountability in Malaysian Higher Education” reported at this conference, the assessment problems in higher education in Malaysia are similar to those in the U.S., and (c) the measurement of learning in precollege courses presents problems similar to those in higher education. Much of the material to follow is drawn from “The Physics Education Reform Effort: A Possible Model for Higher Education” [Hake (2005a)].

For American higher education, several indirect (and therefore in my view problematic) gauges have been developed; e.g. Reformed Teaching Observation Protocol (RTOP), National Survey Of Student Engagement (NSSE), Student Assessment of Learning Gains (SALG), and Knowledge Surveys (KS’s) (Nuhfer & Knipp, 2003). [For a discussion and references for all but the last see Hake (2005b)].

On the other hand, direct measures of student learning have been developed by Hersh (2005) and Klein et al. (2005). Hersh codirects the Learning Assessment Project < http://www.cae.org/content/pro_collegiate.htm > that “evaluates students’ ability to articulate complex ideas, examine claims and evidence, support ideas with relevant reasons and examples, sustain a coherent discussion, and use standard written English.” But Shavelson & Huang (2003) warn that “learning and knowledge are highly domain-specific—as, indeed, is most reasoning. Consequently, the direct impact of college is most likely to be seen at the lower levels of Chart 1 – domain-specific knowledge and reasoning [my italics]. Klein et al. have devised tests that compare student learning across institutions in both domain-specific and broad-ability areas of the Shavelson & Huang 2003 “Framework of Cognitive Objectives” (SHFCO).

**Pre/Post Testing**

In sharp contrast to the above mentioned invalid (course exams, final grades, SET’s); indirect (RTOP, NSSE, SALG, KS’s); or general-ability [Hersh (2005), Klein et al. (2005)] measures discussed above, is the direct measure of students’ higher-level domain-specific learning through pre/post testing using (a) valid and consistently reliable tests devised by disciplinary experts, and (b) traditional courses as controls. It should be realized that domain specific learning is probably coupled to the broad-ability areas of the SHFCO, as suggested for physics by the recent research of Coletta & Phillips (2005).

In my opinion, the physics-education reform model – measurement and improvement of cognitive gains by faculty disciplinary experts in their own courses – can provide a crucial complement to the top-down approaches of Hersh (2005) and Klein et al. (2005). Such pre/post testing, pioneered by economists [Paden & Moyer (1969)] and physicists [Halloun & Hestenes (1985a,b)], is rarely employed in higher education, in part because of the tired old canonical objections recently lodged by Suskie (2004) and countered by Hake (2004a) and Scriven (2004). Despite the nay-sayers, pre/post testing is gradually gaining a foothold in introductory astronomy, economics, biology, chemistry, computer science, economics, engineering, and physics courses [see Hake (2004b) for references].
It should be emphasized that such low-stakes formative pre/post testing is the polar opposite of the high-stakes summative testing mandated by the U.S. Department of Education’s No Child Left Behind Act (NCLB) for pre-college grades (USDE 2005a). The pre-to-post-test gain data for a course is only for the use of the teacher of that course, so as to enable the teacher to gauge the degree of student learning in the course and thereby the effectiveness of her or his teaching. The teacher’s salary and prestige are not affected the gain data averages. Likewise a student’s course grade or prestige are not affected by her or his pre-to-post-test gain. Such formative assessment acts to improve student learning while avoiding the consequences of “Campbell’s Law” [Campbell (1975), Nichols & Berliner (2005)]:

“The more any quantitative social indicator is used for social decision making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor.”

Judging from the paper by Hussan et al. (2006) “Integrity in Public Examinations: A Malaysian Experience” reported at this conference, the Malaysian Ministry of Education is engaged in a continual heroic battle to avoid the consequences of Campbell’s Law and thus preserve the integrity of its summative pre-college testing that occurs at the ends of primary education, lower secondary education, upper secondary education, and the pre-tertiary program.

What Physics Has Learned

Physics education researchers (PER’s) have employed formative pre/post testing to show that traditional (T) introductory physics courses promote very little change in students’ understanding of basic physics concepts; regardless of the experience, enthusiasm, talents, and motivation of their professors. This has driven some physicists to develop novel “interactive engagement” (IE) methods, among them: Microcomputer-based Labs, Concept Tests, Modeling, Active Learning Problem Sets, Overview Case Studies, and Socratic Dialogue Inducing Labs (for references see Hake 2002b). That such IE methods are relatively effective in promoting student higher-level learning has been demonstrated by the nearly two-standard deviation [cf. Bloom’s (1984) “two sigma problem”) superiority in normalized average learning gains <g> of IE courses over T (traditional) courses [Hake (1998a,b; 2002b,c) and corroborative references therein]. Notable examples are large enrollment courses at Harvard [Crouch & Mazur (2001)], North Carolina State University [Beichner & Saul (2004)], MIT [Dori & Belcher (2004)], the University of Colorado at Boulder [Pollock (2004)], and California Polytechnic State University at San Luis Obispo [Hoellwarth et al. (2005)].

Some definitions are in order. In the above paragraph (a) the average normalized gain <g> is the actual gain [%post% - %pre%] divided by the maximum possible gain [%100% - %pre%], where the angle brackets indicate the class averages; (b) “traditional” (T) courses are operationally defined courses as those reported by instructors to make little or no use of “interactive engagement” (IE) methods, relying primarily on passive-student lectures, recipe labs, and algorithmic problem exams; (c) IE courses are operationally defined as 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.
With regard to the Malaysian Ministry of Education’s exemplary program to enhance school-based formative assessment [Hassan (2006)], it should be emphasized that IE-methods with their immediate feedback to students also entail immediate feedback to teachers on the knowledge state of the students, thereby facilitating continual formative assessment. Two examples are the continual Socratic questioning employed in “Socratic Dialogue Inducing Labs” [Hake (1992, 2002d)], and Mazur’s (1997) continual Concept Testing by means of an electronic classroom communication system [see e.g. Burnstein & Lederman (2003)].

For links to over 50 U.S. PER groups, over 200 PER papers published in the American Journal of Physics since 1972, and tests of cognitive and affective conditions see, respectively, Meltzer (2005a), Meltzer (2005b), and NCSU (2005). The very active PER discussion list PhysLrnR <http://listserv.boisestate.edu/archives/physlrnr.html> logged over 750 posts in 2005. As far as I know, no other discipline is so actively researching undergraduate student learning. For reviews see McDermott & Redish (1999), Redish (1999), Thacker (2003), Heron & Meltzer (2005), and Wieman & Perkins (2005).

**Synapse Stimulation**

The fact that IE methods are far more effective in promoting conceptual understanding than traditional passive-student methods is probably related to the “enhanced synapse addition and modification” induced by those methods. Bransford et al. (2000) wrote: “. . . synapse addition and modification are lifelong processes, driven by experience. In essence, the quality of information to which one is exposed and the amount of information one acquires is reflected throughout life in the structure of the brain. This process is probably not the only way that information is stored in the brain, but it is a very important way that provides insight into how people learn.” Leamnson (1999, 2000) has also stressed the relationship of biological brain change to student learning. In his Chapter 5 “Teaching and Pedagogy,” Leamnson (1999) wrote, “Teaching must involve telling, but learning will only start when something persuades students to engage their minds and do what it takes to learn.” Another reminder that the affective and the cognitive are inextricably linked, as recently emphasized by Nuhfer (2005).

**The Challenge**

I see no reason that student learning gains far larger than those in traditional courses could not eventually be achieved and documented in other disciplines from arts through philosophy to zoology IF their practitioners would: (a) reach a consensus on the crucial concepts that all beginning students should be brought to understand, (b) undertake the lengthy qualitative and quantitative research [see, e.g., Halloun & Hestenes (1985a,b)], required to develop multiple-choice tests (MCT’s) of higher-level learning of those concepts, and (c) develop Interactive Engagement methods suitable to their disciplines. Why MCT’s? So that the tests can be given to thousands of students in hundreds of courses under varying conditions in such a manner that meta-analyses can be performed, thus establishing general causal relationships in a convincing manner.
But can multiple-choice tests measure higher-order learning? Wilson & Bertenthal (2005) think so, writing (p. 94): “Performance assessment is an approach that offers great potential for assessing complex thinking and learning abilities, but multiple choice items also have their strengths. For example, although many people recognize that multiple-choice items are an efficient and effective way of determining how well students have acquired basic content knowledge, many do not recognize that they can also be used to measure complex cognitive processes. For example, the Force Concept Inventory . . . [Hestenes et al. 1992 – available in Malaysian!] . . . is an assessment that uses multiple-choice items to tap into higher-level cognitive processes.”

Lessons Learned
Can nearly all university disciplines develop synapse-stimulating interactive engagement methods, and also valid and reliable multiple-choice tests of affective and cognitive conditions to measure their effectiveness? I would bet “Yes,” provided they care enough about student learning to mount the necessary research and development effort.

Aside from the advantages of pre/post testing, perhaps physics education researchers’ most important lessons (Hake 2002b) for higher education are Lessons #1, 3, and 4:

L1: The use of Interactive Engagement strategies can increase the effectiveness of conceptually difficult courses well beyond that obtained with traditional methods.

L3: High-quality standardized tests of the cognitive and affective impact of courses are essential for gauging the relative effectiveness of non-traditional and traditional educational methods. For examples of such physics tests see the listing at NCSU (2005).

L4: Education Research and Development by disciplinary experts (DEs), and of the same quality and nature as traditional science/engineering R&D, is needed to develop potentially effective educational methods within each discipline. But the DEs should take advantage of the insights of DEs engaged in education R&D in other disciplines, cognitive scientists, faculty and graduates of education schools, and classroom teachers.

In the U.S., calls for the accountability of higher education in promoting student learning are becoming more forceful, both from inside the university, e.g., Duderstadt (2000), Weber & Duderstadt (2004), Hersh (2005), Hersh & Merrow (2005), Bok (2005a,b,c); and outside the university, e.g., by the U.S. Dept. of Education’s new “Commission on the Future of Higher Education” (USDE 2005b). For reports on the Commission’s first two meetings and commissioner’s comments on the possibility of NCLB-like testing in higher education, and on the declining literacy of college graduates (NAAL 2005), see Lederman (2005a, b).

As Hersh (2005) observes: “. . . in an era when the importance of a college diploma is increasing while public support for universities is diminishing, [assessment of student learning] is desperately needed. The real question is who will control it. Legislators are prepared to force the issue: Congress raised the question of quality during its recent hearings on the reauthorization of the Higher Education Act; all regional accrediting agencies and more than forty states now require evidence of student learning from their colleges and universities; and pressure is rising to extend a No Child Left Behind-style testing regime to higher education” [see USDE (2005a,b)].
Thus it would appear to be high time for American faculty members to turn more of their attention to shifting the higher education paradigm from teaching to learning, both because it’s the right thing to do, and because not doing so may invite stifling oversight by state and national bureaucrats. Judging from the paper by Othman et al. (2006) “Accountability in Malaysian Higher Education” reported at this conference, a similar suggestion might be appropriate for Malayan university faculty.

References and Footnotes [Tiny URL’s courtesy <http://tinyurl.com/create.php>.


Bloom, B.S. 1984. “The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring,” Educational Researcher 13(6), 4-16 (1984). Bloom wrote: "Using the standard deviation (sigma) of the control (conventional) class, it was typically found that the average student under tutoring was about two standard deviations above the average of the control class. . . The tutoring process demonstrates that most of the students do have the potential to reach this high level of learning. I believe an important task of research and instruction is to seek ways of accomplishing this under more practical and realistic conditions than the one-to-one tutoring, which is too costly for most societies to bear on a large scale. This is the ‘2 sigma’ problem.”


Bok, D. 2005c. "Are colleges failing? Higher ed needs new lesson plans" *Boston Globe*, 18 December, freely online (probably only for a short time) at <http://tinyurl.com/da5v2>, and to educators at <http://tinyurl.com/aj95w> (scroll to the APPENDIX). Bok wrote: “... studies indicate that problem-based discussion, group study, and other forms of active learning produce greater gains in critical thinking than lectures, yet the lecture format is still the standard in most college classes, especially in large universities.”


Duderstadt, J.J. 2000. *A University for the 21st Century*. Univ. of Michigan Press; for a description see <http://tinyurl.com/9lhpl>. Duderstadt writes: “Few faculty members have any awareness of the expanding knowledge about learning from psychology and cognitive science. Almost no one in the academy has mastered or used this knowledge base. One of my colleagues observed that if doctors used science the way college teachers do, they would still be trying to heal with leeches.”


Hake, R. R. 2005a. “The Physics Education Reform Effort: A Possible Model for Higher Education,” online at <http://www.physics.indiana.edu/~hake/NTLF42.pdf> (100 kB), a slightly edited version of an article that was (a) published in the National Teaching and Learning Forum 15(1), December 2005, online to subscribers at <http://www.ntlf.com>, and (b) disseminated by the Tomorrow’s Professor list <http://ctl.stanford.edu/Tomprof/postings.html> as Msg. 698 on 14 Feb 2006.


Hersh, R.H. 2005. “What Does College Teach? It's time to put an end to 'faith-based' acceptance of higher education's quality,” *Atlantic Monthly* 296(4): 140-143, November; freely online to (a) subscribers of the *Atlantic Monthly* at <http://tinyurl.com/dwss8>, and (b) (with hot-linked academic references) to educators at <http://tinyurl.com/9nqon> (scroll to the APPENDIX).


Lederman, D. 2005b. “Graduated but Not Literate,” Inside Higher Ed, 16 December, online at <http://insidehighered.com/news/2005/12/16/literacy>. Lederman writes (my italics): “Not only does [the report on literacy NAAL (2005)] find that the average literacy of college educated Americans declined significantly from 1992 to 2003, but it also reveals that just 25 percent of college graduates — and only 31 percent of those with at least some graduate studies — scored high enough on the tests to be deemed ‘proficient’ from a literacy standpoint, which the government defines as ‘using printed and written information to function in society, to achieve one’s goals, and to develop one’s knowledge and potential.’ ‘This seems like another piece of hard evidence, a fairly clear indication, that the ‘value added’ that higher education gave to students didn’t improve, and maybe declined, over this period,’ said Charles Miller, the former University of Texas regent who is heading the U.S. education secretary’s Commission on the Future of Higher Education. ‘You have the possibility of people going through schools, getting a piece of paper for sitting in class a certain amount, and we don’t know whether they’re getting what they need. This is a fair sign that there are some problems here.’”


USDE. 2005. U.S. Dept. of Education, “Secretary Spellings Announces New Commission on the Future of Higher Education,” press release online at <http://tinyurl.com/cxgfz>: “Spellings noted that the achievement gap is closing and test scores are rising among our nation's younger students, due largely to the high standards and accountability measures called for by the *No Child Left Behind Act*. More and more students are going to graduate ready for the challenges of college, she said, and we must make sure our higher education system is accessible and affordable for all these students.”


Wolff. B. 2005. Review of Weber & Duderstadt (2004), *Planning for Higher Education* 33(4): 37-39; online at <http://tinyurl.com/82ppn>. Wolff wrote: “In each of the essays there is a written and implied understanding that a university of the future can only be successful ‘by adapting to market forces’ (p. 208).” But in a nation beset with life-threatening problems (political, social, environmental, and economic ) should not the university adapt to Barr and Tagg’s 1995 student-learning paradigm rather than to “market forces”? 