Problems with Student Evaluations: Is Assessment the Remedy? * † ‡

Richard R. Hake
Indiana University (Emeritus), 24245 Hatteras Street, Woodland Hills, CA 91367

In his POD post "Re: Problems with Student Evaluations: Is Assessment the Remedy," Ed Nuhfer (2002) hit the nail on the head:

> One of the most encouraging solutions that I see out of this morass . . . the unending tired debate over student evaluations ◊) . . . is the assessment movement. Those who object to sophisticated assessments usually ask, "Why can't we just use grades as measures of learning?" Doesn't that just echo like "Why can't we just use student ratings of professors as measures of good teaching?. . . . My hope is that, a decade from now, members will look at our discussions about student ratings in the POD archives and realize just how far people can come in ten years if they commit to breaking out of primitive conventions. . .(e.g., using mere grades as measures of learning).

---

*Partially supported by NSF Grant DUE/MDR-9253965.

† The reference is Hake, R.R. 2002. "Problems with Student Evaluations: Is Assessment the Remedy?"; online in pdf form as ref. 18 at < **http://www.physics.indiana.edu/~hake** > , and in HTML form (through the courtesy of Russ Hunt) at < **http://www.stu.ca/~hunt/hake.htm** >. This paper originally appeared as a discussion-list post (Hake 2002e), and can be seen in its original somewhat less readable ASCII form at the URL given in that reference.

‡ Comments and suggestions are welcomed at <rrhake@earthlink.net>. All URL's were checked on 11/16/02.

◊ The number of hits on "student evaluations" in the "Search for" slots of their archive search engines (as of 25 Apr 2002) is shown by the numbers after each URL :

AERA-D < **http://lists.asu.edu/archives/aera-d.html** > . . . . . (139)
ASSESS < **http://lsv.uky.edu/archives/assess.html** > . . . . . .(37)
EVALTALK < **http://bama.ua.edu/archives/evaltalk.html** > . . . . . .(51)
Phys-L < **http://mailgate.nau.edu/archives/phys-l.html** >. . . . . (105)
PhysLrnR < **http://listserv.boisestate.edu/archives/physlrnr.html** >. . . . .(27)
POD < **http://listserv.nd.edu/archives/pod.html** > . . . . . (206)
STLHE-L < **http://listserv.unb.ca/archives/stlhe-l.html** > . . . . . .(197)

TOTAL HITS 762

At the risk of replaying to deaf ears the same old record (as e.g., Hake 2000; 2002a,b,c,d), I reiterate Lesson #3 of Hake (2002c) (see that article for the references):

> **L3. High-quality standardized tests of the cognitive and affective impact of courses are essential for gauging the relative effectiveness of non-traditional educational methods.**
>
> As far as I know, disciplines other than physics, astronomy (Adams et al. 2000; Zeilik et al. 1997, 1998, 1999), and possibly economics (Saunders 1991, Kennedy & Siegfried 1997, Chizmar & Ostrosky 1998, Allgood and Walstad 1999) have yet to develop any such tests and therefore cannot effectively gauge either the need for or the efficacy of their reform efforts. In my opinion, all disciplines should consider the construction of high-quality standardized tests of essential introductory course concepts.

Because most disciplines have failed to develop definitive tests to measure cognitive and affective course impacts, seemingly simplistic statements from the pro Student Evaluation of Teaching (SET) camp cannot always be immediately dismissed. For example:

1. Aleamoni (1987) addressed "Myth #5: Student rating forms are both unreliable and invalid" as follows: ". . . Most student forms have been validated by the judgement of experts that the items and subscales measure important aspects of instruction . . . (and also) . . . by statistical tools such as factor analysis. . . further evidence of validity comes from studies in which student ratings are correlated with other indicators of teacher competence, such as peer (colleague) ratings, expert judges' ratings, graduating seniors and alumni ratings, and student learning."

2. Michael Scriven (1988) [as quoted by D'Apollonia & Abrami (1997)] stated that "student ratings are not only A valid, but often *the only* valid, way to get much of the information needed for most evaluations." (*Emphasis* in the original.)

3. Marsh & Dunkin (1992) concluded: "SET's are clearly multidimensional, quite reliable, and reasonably valid."

4. Cashin (1995) stated "In general, student ratings tend to be statistically reliable, valid, and relatively free from bias or the need for control; probably more so than any other data used for evaluation."

5. Marsh and Roche (1997) claimed that "there is little evidence of the validity of any other sources of data . . . . (on teaching effectiveness)."

The question is "**VALID FOR WHAT**?" I think SETs can be "valid" in the sense that can be useful for gauging the *affective* impact of a course and for providing diagnostic feedback *to teachers* [see, e.g., Hake & Swihart (1979)] to assist them in making mid-course corrections. However IMHO, SETs are *not* valid in their widespread use by *administrators* to gauge the *cognitive* impact of courses [see, e.g., Williams & Ceci (1997); Hake (2000; 2002a,b); Johnson (2002)]. *In fact the gross misuse of SET's as gauges of student learning is, in my view, one of the institutional factors that thwarts substantive educational reform* (Hake 2002c, Lesson #12)**.**

Although there are many SET researchers (see, e.g. Abrami et al. 1990; Aleamoni 1987 ; d'Apollonia & Cohen 1997; Cohen 1981; Cashin 1995; Marsh & Roche 1997; Marsh & Dunkin 1992) who claim that SETs are valid indicators of students' cognitive condition (for a review see Hake 2000), their conclusions are almost always based on measuring student learning or "achievement" by course grades or exams and *not* by pre/post testing . . . (*pre/post* even despite the Lordly Cronbachian objections of some education/psychology specialists – see Hake (2001). . . with valid and reliable instruments such as the *Force Concept Inventory* of Hestenes et al. (1992) and Halloun et al. (1995) **[**see, e.g., Hake **(**2002c**)]**.

With regard to the problem of using course performance as a measure of student achievement or learning, Peter Cohen's (1981) oft-quoted meta-analysis of 41 studies on 68 separate multisection courses purportedly showing that:

> the average correlation between an overall instructor rating and student achievement was +0.43; the average correlation between an overall course rating and student achievement was +0.47 . . . the results . . . provide strong support for the validity of student ratings as measures of teaching effectiveness

was reviewed and reanalyzed by Feldman (1989) who pointed out that McKeachie (1987)

> has recently reminded educational researchers and practitioners that the achievement tests assessing student learning in the sorts of studies reviewed here. . . (e.g., those by Cohen (1981, 1986, 1987). . . typically measure lower-level educational objectives such as memory of facts and definitions rather than higher-level outcomes such as critical thinking and problem solving . . .[he might have added conceptual understanding] . . . that are usually taken as important in higher education.

Striking back at SET skeptics, Peter Cohen (1990) opined:

> Negative attitudes toward student ratings are especially resistant to change, and it seems that faculty and administrators support their belief in student-rating myths with personal and anecdotal evidence, which (for them) outweighs empirically based research evidence.

However, as far as I know, neither Cohen nor any other SET champion has countered the fatal objection of McKeachie *that the evidence for the validity of SET's as gauges of the cognitive impact of courses rests for the most part on measures of students' lower-level thinking as exhibited in course grades or exams.* At least in physics it is well-known (see, e.g., Hake 2002c) that students in *traditional* mechanics courses can achieve A's through rote memorization and algorithmic problem solving, while achieving *normalized* gains in conceptual understanding of only about 0.2 (i.e., pre-to-post gains that are only about 0.2 of the maximum possible gain).

Williams & Ceci (1997) write:

> 1. "in searching for better and fairer means of evaluating teaching effectiveness and providing better bases for reappraisal of one's teaching, we need to experiment with alternative methods of soliciting students' opinions," and
>
> 2. "teaching faculty should be given the opportunity to train in techniques . . . (of presentation style). . . that can enhance their student ratings. . .(as shown by Williams & Ceci 1997). . . , especially if such ratings are to be used by administrators in recommendations for tenure and promotion."

I would suggest that (a) faculty teaching, (b) student learning, and (c) the goals of higher education might all be better served if faculty would pay less attention to suggestions #1 & #2 of Williams & Ceci, and more attention to the development of valid and reliable tests of the cognitive and affective impact of their courses, in accord with Lesson #3 above. Such effort is currently almost non-existent in academia and would probably require fulfillment of Lesson #4 of Hake (2002c):

> **L4. Education research and development (R&D) by disciplinary experts (DE's), 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 DE's should take advantage of the insights of (a DE's doing education R&D in other disciplines, (b) cognitive scientists, (c) faculty and graduates of education schools, and (d) classroom teachers . . . .**

> The education of disciplinary experts in education research requires Ph.D. programs at least as rigorous as those for experts in traditional research. The programs should include, in addition to the standard disciplinary graduate courses, some exposure to: the history and philosophy of education, computer science, statistics, political science, social science, economics, engineering – see Lesson 11, and, most importantly, cognitive science (i.e., philosophy, psychology, artificial intelligence, linguistics, anthropology, and neuroscience). . . . In the U.S. there are now about a dozen Ph.D. programs (Physical Science Resource Center 2001, UMd-PERG 2001) in physics education within physics departments and about half that number of interdisciplinary programs between physics and education or cognitive psychology. In my opinion, *all scientific disciplines should consider offering Ph.D. programs in education research.*

References

Abrami, P.C., S. d'Apollonia, & P. Cohen. 1990. "Validity of Student Ratings of Instruction: What We Know and What We Do Not." *Journal of Educational Psychology* **82**: 219-231.

Aleamoni, L.M. 1987. "Student Rating Myths Versus Research Facts," *Journal of Personnel Evaluation in Education* **1**: 111.

Cashin, W.E. 1995 "Student Ratings of Teaching, IDEA Paper No. 32,"Kansas State University Center for Faculty Evaluation and Development; online at
< **http://www.idea.ksu.edu/products/Papers.html** >.

Cohen, P.A. 1981. "Student ratings of Instruction and Student Achievement: A Meta-analysis of Multisection Validity Studies," *Review of Educational Research* **51**: 281. For references to Cohen's 1986 and 1987 updates see Feldman (1989).

Cohen, P.A. 1990. "Bring research into practice," in M. Theall & J. Franklin, eds. *Student ratings of instruction: Issues for improving practice: New Directions for Teaching and Learning,* No. 43, pp. 123-132. Jossey Bass.

d'Apollonia, S. & P.C. Abrami, 1997. "In response . . [to Williams & Ceci (1997)]. . . *Change*, September/October 1997.

Feldman, K.A. 1989. "The Association Between Student Ratings of Specific Instructional Dimensions and Student Achievement: Refining and Extending the Synthesis of Data from Multisection Validity Studies," *Research on Higher Education* **30**: 583. 30: 583.

Hake R.R. & J.C. Swihart. 1979. "Diagnostic Student Computerized Evaluation of Multicomponent Courses," *Teaching and Learning*, Vol. V, No. 3 (Indiana University, January 1979, updated 11/97; online as ref. #4 at < **http://www.physics.indiana.edu/~hake/** >.

Hake R.R. 2000. "Student Evaluations" PhysLrnR/POD post of 16 Jul 2000 16:53:53-0700; online at < **http://listserv.nd.edu/cgi-bin/wa?A2=ind0007&L=pod&P=R11705** >.

Hake, R.R. 2001. "Pre/Post Paranoia," POD post of 17 May 2001 16:24:54-0700; online at < **http://listserv.nd.edu/cgi-bin/wa?A2=ind0105&L=pod&P=R10317&m=4493** >.

Hake, R.R. 2002a. "Re: Can all teachers be improved?", POD post of 19 Mar 2002 16:33:49-0800; online at < **http://listserv.nd.edu/cgi-bin/wa?A2=ind0203&L=pod&P=R10552** >.

Hake, R.R. 2002b. "Assessment of Student Learning in Introductory Science Courses," 2002 *PKAL Roundtable on the Future: Assessment in the Service of Student Learning*, Duke University, March 1-3; online at < **http://www.pkal.org/events/roundtable2002/papers.html** > (when the PKAL server is functioning), and as reference 15 at < **http://www.physics.indiana.edu/~hake** >.

Hake, R.R. 2002c. "Lessons from the physics education reform effort." *Conservation Ecology* **5**(2): 28; online at < **http://www.consecol.org/vol5/iss2/art28** >. *Conservation Ecology*, is a FREE "peer-reviewed journal of integrative science and fundamental policy research" with about 11,000 subscribers in about 108 countries.

Hake, R.R. 2002d. "Re: Outcome Measures #2" POD/EvalTalk/PhysLnR/AERA-D post of 31 Jan 2002 22:07:00-0800; online at < **http:listserv.nd.edu/cgi-bin/wa?A2=ind0202&L=pod&P=R2** >.

Hake, R.R. 2002e. "Re: Problems with Student Evaluations: Is Assessment the Remedy?" AERA-D/ASSESS/EvalTalk/Phys-L/PhysLrnR/POD/STLHE-H post of 25 Apr 2002 16:54:24-0700; online at < **http://lists.asu.edu/cgi-bin/wa?A2=ind0204&L=aera-d&P=R4664** >.

Halloun, I., R.R. Hake, E.P Mosca, D. Hestenes. 1995. Force Concept Inventory (Revised, 1995); online (password protected) at < **http://modeling.asu.edu/R&E/Research.html** >.

Hestenes, D., M. Wells, & G. Swackhamer. 1992. "Force Concept Inventory." *Phys. Teach.* 30:141-158. For the 1995 revision see Halloun et al. (1995).

Johnson, V.A. 2002. "An A Is an A Is an A. . . .And That's the Problem." *New York Times*, 23 April; online at
< **http://www.nytimes.com/2002/04/14/edlife/14ED-VIEW.html?pagewanted=print&pos** >.

Lang, S. 1997."Cornell study finds student ratings soar on all measures when professor uses more enthusiasm. Study raises concerns about the validity of student evaluations. *Cornell Science News*, Sept.; online at
< **http://www.news.cornell.edu/releases/Sept97/student.eval.ssl.html** >.

Lewis, R. 1998., "Student Evaluations: Widespread And Controversial," *The Scientist* **12**(9): 12, Apr. 27; online at < **http://www.the-scientist.com/yr1998/apr/prof_980427.html** >.

McKeachie, W.J. 1987. 'Instructional evaluation: Current issues and possible improvements." *Journal of Higher Education* **58**(3): 344-350.

Physical Science Resource Center. 2001. American Association of Physics Teachers; online at < **http://www.psrc-online.org/** > / "Resource Center"/ "Physics Education Research", where "/" means "click on."

Marsh, H.W. & L.A. Roche. 1997. "Making students' evaluations of teaching effectiveness effective," *American Psychologist* **52**: 1187-1197.

Marsh, H.W. & M. Dunkin. 1992. "Students' evaluations of university teaching: A multidimensional perspective" in *Higher education: Handbook on theory and research*, Vol. 8. (Agathon, 1992) pp. 143-234. [Reprinted in R. P. Perry & J. C. Smart (eds.), *Effective Teaching in Higher education: Research and Practice* (Agathon, 1997) pp. 241-320.

> The abstract reads: "One purpose of students' evaluations of teaching (SET's) is for research on teaching itself. Because SET's do not reflect the valid effects of presage and context measures, it is argued. . .(evidently by others). . . that SET's may not be a fair source in the evaluation of teaching in higher education. This study examines this problem, with an emphasis on the need for a multidimensional approach for measuring effective teaching. The reliability, stability, and generalizability factors are also discussed, as are potential biases in both peer and student related evaluations. *It is concluded that* SET's *are clearly multidimensional, **q**uite reliable, and reasonably valid*. However, caution is suggested in using SET's as a systematic approach, as SET's are yet only one indicator of effective teaching in higher education. (My *emphasis*.)

Nuhfer, E. 2002. "Re: Problems with Student Evaluations: Is Assessment the Remedy?" POD post of 24 Apr 2002 13:18:01-0600; online at
< **http://listserv.nd.edu/cgi-bin/wa?A2=ind0204&L=pod&O=D&P=13406** >.

Scriven, M. 1988. "The Validity of Student Ratings," *Instructional Evaluation* **9**: 5-18.

UMd-PERG. 2001. Univ. of Maryland Physics Education Research Group, listing of physics education groups with web homepages: online at
< **http://www.physics.umd.edu/perg/homepages.htm** >.

Williams, W.M. & S.J. Ceci. 1997 "How'm I Doing," *Change*, September/October, pp. 13-23.

> "Today, all instructors would be well advised to ask their students frequently 'How'm I doing?' and listen carefully to the answer. As in politics, however, the answer may have more to do with style than substance." See also Lewis (1998) and Lang (1997).