Assessment of Physics Teaching Methods*‡◊

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

I. Introduction

Although a tremendous amount of work has been devoted to assessment in education [for references see Hake (2002b)], very little effort has thus far been devoted to what I regard as one of the most crucial types of assessment, vis., *valid and reliable measures of student learning in introductory courses*. In his recent review of assessment, or lack thereof, in introductory undergraduate science courses in the U.S., Stokstad (2001) wrote: "*Physicists are out in front in measuring how well students learn the basics*, as science educators incorporate hands-on activities in hopes of making the introductory course a beginning rather than a finale." (My *italics*.) This lead position is due in no small part to pre/post testing using the *Force Concept Inventory* (FCI) [Hestenes et al. (1992), Halloun et al. (1995)]; the Mechanics Diagnostic (MD) test – precursor to the FCI [Halloun & Hestenes (1985 a,b)]; the *Force Motion Concept Evaluation* (FMCE) [Thornton & Sokoloff (1998)]; and the *Conceptual Survey of Electricity and Magnetism* (CSEM) [Maloney et al. (2001)].

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

‡ The reference is: Hake, R.R. 2002. "Assessment of Physics Teaching Methods," *Proceedings of the UNESCO-ASPEN Workshop on Active Learning in Physics*, Univ. of Peradeniya, Sri Lanka, 2-4 Dec. 2002; also online as ref. 29 at < http://www.physics.indiana.edu/~hake/ >. [ASPEN = ASian Physics Education Network < http://www.swin.edu.au/physics/aspen/welcome.htm >.]

◊ Although requested to write on "Assessment and Evaluation in Active Learning," I have taken the liberty of making two changes in the title:

(a) Deleting the word "Evaluation" since nothing is lost if the word "assessment" is defined so as to include both *formative* and *summative* assessment, or a combination of these. As commonly understood, summative assessment (often called "evaluation") is performed *at the completion* of a program so as to gauge its accomplishment, whereas formative assessment takes place *during* a program so as to obtain guidance for its development.

(b) Changing "in Active Learning" to "of Physics Teaching Methods," so as to avoid prejudging the effectiveness of an "active-student" method by calling it "active learning." Students may be both physically and mentally "active," but little substantive learning may take place – witness the often very active students in the traditional recipe lab. Also as a control, one should also assess various "non-active-student" methods.

II. Accomplishments of Physics Education Research

I think most physics education researchers (PER's), if not all physics instructors, currently agree that PER, with its: (a) standardized assessments such as the FCI, MD, FMCE, and CSEM [for other physics assessment instruments see NCSU (2002) and FLAG (2002)], and (b) half-century-old "normalized gain" g (Sec. III below), has shown rather conclusively that:

(1) the traditional mode of introductory physics instruction (passive student lectures, recipe labs, and algorithmic problem exams) is relatively ineffective in promoting students' conceptual understanding, even when employed by teachers who receive relatively high student evaluations [(Hake (2002e), Hunt (2002)], and

(2) "interactive engagement (IE) methods" *can* be much more effective than "traditional" (T) methods in promoting conceptual understanding of mechanics. For example, for mechanics test results (Hake 1998 a,b,c) for courses using IE and T methods, I obtained a Cohen (1988) effect size "d" of 2.43 (Hake 2002a), much higher than any found by Lipsey & Wilson (1993) in their meta-meta-analysis of psychological, educational, and behavioral treatments. Seven reasons for this "d disparity" between physics-education and other social-science research are given in Hake (2002a).

More recently normalized gain differences between T and IE courses that are consistent with the work of Hake (1998a,b,c; 2002a,b,c,d) have been reported by Redish et al. (1997); Saul (1998); Francis et al. (1998); Redish & Steinberg (1999); Redish (1999); Beichner et al. (1999); Cummings et al. (1999); Novak et al. (1999); Beichner et al. (2000); Bernhard (2000); Crouch & Mazur (2001); Johnson (2001); Meltzer (2002a,b,c); Meltzer & Manivannan (2002); Savinainen & Scott (2002a,b); Steinberg and Donnelly (2002); Fagan et al. (2002); and Van Domelen & Van Heuvelen (2002).

III. The Nomalized Gain

The normalized gain "g" for a treatment is defined [Hovland et al. (1949), Gery (1972), Hake (1998a)] as g = Gain/[Gain (maximum possible). Thus, e.g., if a class averaged 40% on the pretest, and 60% on the posttest then the class-average normalized gain <g> = (60% - 40%)/(100% - 40%) = 20%/60% = 0.33. Ever since the work of Hovland et al. (1949) it's been know by pre/post cognoscente (up until about 1998 probably less than 100 people worldwide) that <g> *is a much better indicator of the extent to which a treatment is effective than is either gain or posttest*, for example, if the treatment yields <g> > 0.3 for a mechanics course, then the course could be considered as in the "interactive-engagement zone" (Hake 1998a, Meltzer 2002b).

Regrettably, the psychology/education/psychometric PEP community [see e.g., Pelligrino et al. (2001); Shavelson & Towne (2001); Fox & Hackermann (2002); Feuer et al. (2002)] remains largely oblivious of PER and the normalized gain. Paraphrasing Lee Schulman, as quoted by the late Arnold Arons (1986): "it seems that in education, the wheel (more usually the flat tire) must be reinvented every few decades." Unfortunately there seems to be little effort to build a "community map" [Redish (1999), Lagemann (2000), Ziman (2000), Shavelson & Towne (2001), Hake (2002a - "Can Educational Research be *Scientific* Research?")]. Extrapolating the historical record, around 2030 yet another investigator will come up with the idea of g, and fruitlessly attempt to interest the pre/post paranoiac (Hake 2001b) education community. Then around 2060 . . .

It should be remarked that most of the analysis of the FCI, MD, FMCE, CSEM, and other physics assessment tests have been done within the framework of "Classical Test Theory" in which only the number of correct answers is considered in the scoring. However more sophisticated analyses are being developed [e.g., Bao & Redish (2001) for the FCI, and Thornton (1995) for the FMCE]; and other psychometric approaches such as Item Response and Rasch analyses may offer advantages in some cases [see e.g., Hake (2002g) for a few references to the voluminous psychometric literature].

IV. Does the Normalized Gain Tell All?

Does the class average normalized gain <g> for the FCI, MD, or FMCE provide a definitive assessment of the *overall* effectiveness of an introductory physics class? *No*! It assesses "*only the attainment of a* ***minimal*** *conceptual understanding of mechanics.* In some first-semester or first-quarter introductory physics courses, subjects other than mechanics are often covered. The effectiveness of the course in promoting student understanding of those topics would not, of course, be assessed by the normalized gain on the FCI, MD, or FMCE. Furthermore, as indicated in Hake (1998b), among desirable outcomes of the introductory course that <g> *does not* measure directly are students':

(a) satisfaction with and interest in physics;
(b) understanding of the nature, methods, and limitations of science;
(c) understanding of the processes of scientific inquiry such as experimental design, control of variables dimensional analysis, order-of-magnitude estimation, thought experiments, hypothetical reasoning, graphing, and error analysis;
(d) ability to articulate their knowledge and learning processes;
(e) ability to collaborate and work in groups;
(f) communication skills;
(g) ability to solve real-world problems;
(h) understanding of the history of science and the relationship of science to society and other disciplines;
(i) understanding of, or at least appreciation for, "modern" physics;
(j) ability to participate in authentic research.

Affective aspects such as "a" (satisfaction with and interest in physics) can be assessed by *well designed* [e.g., Hake & Swihart (1979)] student evaluations. However, despite the arguments of some student-evaluation specialists [reviewed in Hake (2002e)], in my opinion student evaluations do *not* provide useful information on the cognitive impact of a course. In fact, the gross misuse of student evaluations as gauges of student learning is, in my view, one of the institutional factors that thwarts substantive educational reform (Hake 2002a, Lesson #12.)

May & Etkina (2002) have recently studied the relationship of "d" (students' ability to articulate their knowledge and learning processes) to their conceptual learning gains. In addition, as discussed in Hake (2002a, Lesson #3), the design and testing of instruments that approach the difficult assessment of factors such "b", "c", and "h" has been underway in physics education research for several years [see, e.g., Halloun (1997), Halloun & Hestenes (1998), Redish et al. (1998)]. Administration of the *Maryland Physics EXpectations* (MPEX) survey to 1500 students in introductory calculus based physics courses in six colleges and universities . .. . (showed). . . . “a large gap between the expectations of experts and novices and . . . . a tendency for student expectations to *deteriorate* rather than improve as a result of introductory calculus-based physics” (Redish et al. 1998). Here the term “expectations” is used to mean a combination of students’ *epistemological* beliefs about learning and understanding physics and students’ *expectations* about their physics course (Elby 1999). It may well be that students’ attitudes and understanding of science and education are irreversibly imprinted in the early years. If so, corrective measures await a badly needed shift of K-12 education away from rote memorization and drill (often encouraged by state-mandated standardized tests) to the enhancement of understanding and critical thinking [Hake (2000a,b; 2002a); Mahajan & Hake (2000); Benezet (1935/36)] .

V. Suggestions for the Administration and Reporting of Diagnostic Tests

I list below some administrative and reporting suggestions that might assist teachers in using *formative or summative* diagnostic testing. These suggestions are, for the most part drawn from Hake (2001a), and reflect the hard lessons I have learned in the pre/post testing of 1263 pre-med introductory-physics-course students at Indiana University and the compilation of a 6542 student (62-course) survey (Hake 1998a, b) of pre/post test results. Although they are undoubtedly *not* appropriate for all classroom situations, they may at least indicate some of the problems that should be anticipated. (An asterisk * preceding a suggestion indicates that the suggestion is intended, at least in part, to promote confidentiality of the test.)

A . Administration of Diagnostic Tests (DT's)

*1. When administering DT's to students, refer to the tests by home-made generic titles rather than the specific titles designated by the authors (e.g.,"Mechanics Familiarity Survey" rather than *Force Concept Inventory* or *Force Motion Concept Evaluation*).

2. To enable meaningful pre/post comparison, maximum time intervals ΔT given to students to complete the pretest and the posttest should be same. To facilitate more meaningful meta-analysis, maximum time interval ΔT should be specified by the test designers and that interval should be rigidly enforced by all examiners.

3. Do *not* allow students to take either the pretest or the posttest anonymously, because non-anonymity allows:

(a) Proper incentive for students to exert effort on the test.

(b) Analysis of "matched" pre/post test data, i.e., obtaining the average class pretest score by counting only the scores of those students who took the posttest, and thus allowing a more rigorous calculation of the class average normalized gain <g>. [For a discussion of "matched data" see Hake (2002a); and also Hake (1998b), Table I, footnote "c" on page 7.]

(c) Knowledge of the normalized gain g for each single student in the class, thus allowing a calculation of the average of the single-student gains: "g-ave." [See Hake (1998a - Sec. V and also footnote #46), for a discussion of systematic and random errors in pre/post testing and the connection between low correlation of single students g's with their pretest scores, and the small difference between values of g-ave and <g>.]

(d) Analyses of single student normalized or actual gains in terms of single-student characteristics or performance on other tests.

(e) Calculation of the correlation of individual student g's with their pretest scores.

ASIDE: It should be kept in mind that the experimental justification for using <g> as a comparative measure of course effectiveness over diverse varying average pretest scores in Hake (1998a) was that the correlation of <g> with <pre> was a very low +0.02. Unless, for a given DT, similar low correlations can be found from meta-analytic results over different courses, or at least for single student results within different courses, then the use of that DT for intercomparison of diverse groups is somewhat problematic. Even if the correlation of <g> with <pre> is low, open questions remain for most DT's as to whether or not "hidden variables" (e.g., average math proficiency, spatial visualization ability, high-school physics, gender, motivation, socio-economic level, ethnicity, scientific reasoning skills, IQ, SAT, GRE) of a class could have a significant effect on <g>. [See e.g., Hake et al. (1994), Meltzer (2002a), Hake (2002f).]

*4. If possible, give the pretest on the *first* day of class. Take great care that all question sheets and answer sheets are returned and verify such return by counting those given out and those returned. *In order to promote serious effort on the pretest by students, explain that although their scores on the pretest will **no**t count towards the course grade, their scores will be confidentially returned to them and will assist both themselves and their instructors to know the degree and type of effort required for them to understand mechanics*.

*5. Give the posttest *unannounced* near the final day of classes, and preferably as part of the final exam with significant course credit given for posttest performance. Giving course credit probably motivates students to take the posttest more seriously and thereby demonstrate more adequately their understanding, especially if time devoted to the posttest subtracts from time spent on the rest of the final exam. *If no grade credit is given for performance on the posttest then* selective removal of "no-effort" tests [see e.g., Henderson (2002), Mallinckrodt (2001)] by different investigators with different no-effort criteria will lead to uncertainty in comparing normalized gains of different courses.

*6. Do *not* return DT's to students after either the pretest or the posttest.

*7. Post DT scores by ID without posting or disseminating questions or answers.

*8. Avoid in-class discussion of questions identical or almost identical to DT questions (an example of "teaching to the test").

*9. For the posttest, announce that instructors be willing to discuss the questions and/or problems only *privately* with students.

*10. *Do not make DT questions or problems available on the web unless they are password protected* such that only authorized instructors may gain access. *Do not publish DT's in the open literature*, as has been the common practice. Carefully constructed DT's are international assets whose confidentiality should be as well protected as the MCAT (The U.S. Medical College Admission Test).

*11. Because of the almost unavoidable slow diffusion of test questions and answers to student files, *replace each DT at approximately 5- or 10-year intervals*, such that it can be meaningfully calibrated against the previous test(s). [So far this has NOT been done for the now overused 1992/95 versions of the FCI; in my opinion, *as time goes on, research results based on the 1992/95 FCI will become more and more doubtful*.]

B . Reporting of Diagnostic Tests (DT's)

1. Report at least the class average:

a. <%pre> *with its standard deviation* (sd),

b. <%post> *with its sd*, and

c. normalized gain <g>.

ASIDES:

(1) Unless standard deviations are reported, the effect size and errors in <g> cannot be ascertained.

(2) As a statistic for comparison of courses and for meta-analyses, the class average <g> is better, in my opinion, than g-ave because the latter: (a) must exclude students who score 100% on the pretest and thus achieve an infinite or indeterminate g; and (b) may introduce skewing due to outliers who score near 100% on the pretest and less on the posttest such their <g>'s are large and negative. The selective removal of outliers so as to avoid "(b)" by various different investigators with different outlier criteria will lead to a degree of uncertainty in comparing normalized gains of different courses.

2. Report if possible:

a. *Cohen's "effect size,"* d (Cohen 1988)

ASIDE: The effect size is commonly used in meta-analyses, and strongly recommended by many critics as a preferred alternative (or at least addition) to the *usually inappropriate* t-tests and p values associated with null-hypothesis testing. [See Hake (2002a) for references to the effect size and the extensive anti-p literature.] But as indicated in Hake (2002a): ". . . . it should be noted that *for **pre/post** comparisons of course effectiveness over diverse student populations with widely varying average pretest scores, "d" is a relatively poor metric* because unlike "<g>" :

(1) "d" depends on the actual bare (unrenormalized) gain that tends to be negatively correlated with the <pre>;

(2) <Gain>'s are confounded with sd's: given two classes both with identical *statistically* significant <Gain>'s, the more homogeneous class with the smaller rms pre/post sd's will be awarded the higher "d."

b. The *Kuder-Richardson reliability coefficients* KR-20 (or equivalent - for tests in which the answer is either correct or incorrect as in the FCI or FMCE - *Cronbach's alpha*) for the pre- and post-tests. [See Hake (2001a), Beichner (1994), Slavin (1992) for references.]

c. The *estimated systematic and random errors*. [See e.g., Hake (1998a).]

d. The *correlation of individual student g's and pretest scores*.

ASIDE: A significant *positive* correlation would suggest that the instruction tends to favor students who have *more* prior knowledge of the subject as judged by the pretest score ("Matthew effect"*); a significant *negative* correlation would suggest that the instruction favors students judged by the pretest score ("anti–Matthew effect"); and an insignificant correlation would suggest that the instruction is at about the right level for students who have an average prior knowledge of the subject as judged by the pretest score.

3. Further information

As a guide to other information that might be useful in a report of pre/post test results and complementary qualitative research, consider the work sited in the last paragraph of Sect. II above; and also Hake (1987, 1992, 1997, 1998a,b; 2001a; 2002a); Zeilik (et al. 1997, 1998, 1999); McDermott & Redish (1999), and Tobias & Hake (1988).

*Matthew, *First Gospel of the New Testament* (Gutenberg edition)"to him that hath shall be given, but from him that hath not shall be taken away even that which he hath."

References

Arons, A.B. 1986. "Conceptual Difficulties in Science" in *Undergraduate Education in Chemistry and Physics, Proceedings of the Chicago Conferences on Liberal Education*," No. 1, R.R. Rice. ed. (Univ. of Chicago), p. 50.

Bao, L., and E. F. Redish. 2001. "Concentration analysis: a quantitative assessment of student states." *Am. J. of Phys*. **69**(7): S45-S53; online at < http://www.physics.ohio-state.edu/~lbao/papers.htm >.

Beichner, R.J. 1994. "Testing student interpretation of kinematics graphs." *Am. J. Phys.* **62**(8):750-762. A good introduction for physicists to some of the basic statistics of educational research; online at < http://www.ncsu.edu/PER/articles.htm >.

Beichner, R.J., L. Bernold, E. Burniston, P. Dail, R. Felder, J. Gastineau, M. Gjertsen, and J. Risley. 1999. "Case study of the physics component of an integrated curriculum." *Physics Education Research Supplement to Am. J. Phys.* **67**(7): S16-S24; online at < http://www.ncsu.edu/PER/articles.htm >.

Beichner, R.J., J.M. Saul, R.J. Allain, D.L. Deardorff, D.S. Abbott. 2000. "Introduction to SCALE-UP: Student Centered Activities for Large Enrollment University Physics," *Proceedings of the 2000 Annual meeting of the American Society for Engineering Education*; online at < http://www.ncsu.edu/PER/articles.htm >.

Benezet, L.P. 1935-1936. "The teaching of arithmetic I, II, III: The story of an experiment," *Journal of the National Education Association* **24**(8), 241-244 (1935); **24**(9), 301-303 (1935); **25**(1), 7-8 (1936). The articles were (a) reprinted in the *Humanistic Mathematics Newsletter* #6: 2-14 (May 1991); (b) placed on the web along with other Benezetia at the Benezet Centre; online at < http://www.inference.phy.cam.ac.uk/sanjoy/benezet/ >. See also Mahajan & Hake (2000).

Bernhard, J. 2000. "Does active engagement curricula give long-lived conceptual understanding? Proceedings of GIREP 2000: *Physics Teacher Education Beyond 2000*, Barcelona; online at < http://www.itn.liu.se/~jonbe > / "Publications" / "Physics EducationResearch," where "/" means "click on."

Cohen, J. 1988. *Statistical power analysis for the behavioral sciences*. Lawrence Erlbaum, 2nd ed., pages 20, 44.

Crouch, C.H. & E. Mazur. 2001. "Peer Instruction: Ten years of experience and results," *Am. J. Phys*. **69**: 970-977; online at < http://mazur-www.harvard.edu/library/pubs.taf?function=search >.

Cummings, K., J. Marx, R. Thornton, and D. Kuhl. 1999. "Evaluating innovations in studio physics," *Am. J. of Phys*. **67**(7): S38-S44.

Elby, A. 1999. "Another reason that physics students learn by rote." *Physics Ed. Res. Supplement to Am. J. Phys.* **67**(7): S52-S57; abstract and introduction is online at < http://www2.physics.umd.edu/~elby/papers/distortions/abstract.html >.

Fagan, A.P., C.H. Crouch, & E. Mazur. 2002. *Phys. Teach*. **40**(4): 206-209; online at < http://mazur-www.harvard.edu/library/pubs.taf?function=search >.

Feuer, M.J., L. Towne, & R.J. Shavelson. 2002. "Scientific, Culture, and Educational Research," *Educational Researcher* **31**(8): 4-14; online at < http://www.aera.net/pubs/er/eronline.htm >.

Fox, M.A. & N. Hackermann, eds. 2002. *Recognizing, Evaluating, Rewarding, and Developing Excellence in Teaching of Undergraduate Science, Technology, Engineering, & Mathematics*. National Academy Press; online in prepublication form at: < http://www.nap.edu/catalog/10024.html >.

FLAG. 2002. *Field-tested Learning Assessment Guide*; online at < http://www.flaguide.org >.

Francis, G.E., J.P. Adams, E.J. Noonan. 1998. "Do they stay fixed?" *Phys. Teach*. **36**(8): 488-491.

Gery, F.W. 1972. "Does mathematics matter?" in A.Welch, ed., *Research papers in economic education*. Joint Council on Economic Education. pp. 142-157.

Hake, R.R. 1987. "Promoting student crossover to the Newtonian world," *Am J. Phys*. **55**(10): 878-884.

Hake, R.R. 1992. "Socratic pedagogy in the introductory physics lab," *Phys. Teach*. **30**: 546-552; updated version (4/27/98) online as ref. 23 at < http://www.physics.indiana.edu/~hake >.

Hake, R.R. 1997. *Mechanics Test Data Survey Form*; online as ref. 5 at < http://physics.indiana.edu/~hake/ >

Hake, R.R. 1998a. "Interactive-engagement vs traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses," *Am. J. Phys*. **66**: 64-74; online as ref. 24 at < http://www.physics.indiana.edu/~hake >.

Hake, R.R. 1998b. "Interactive-engagement methods in introductory mechanics courses," online as ref. 25 at < http://www.physics.indiana.edu/~hake >, a crucial companion paper to Hake (1998a): average pre/post test scores, standard deviations, instructional methods, materials used, institutions, and instructors for each of the survey courses of Hake (1998a) are tabulated and referenced. In addition the paper includes: (a) case histories for the seven IE courses of Hake (1998a) whose effectiveness as gauged by pre-to-post test gains was close to those of T courses, (b) advice for implementing IE methods, and (c) suggestions for further research.

Hake, R.R. 1998c. "Interactive-engagement vs Traditional Methods in Mechanics Instruction," *APS Forum on Education Newsletter*, Summer, p. 5-7, online as ref. 26 at < http://www.physics.indiana.edu/~hake >. Some criticisms of Hake (1998a) and of physics-education reform generally are countered.

Hake, R.R. 2000a. "The general population's ignorance of science-related societal issues - a challenge for the university." *AAPT Announcer* **30**(2):105; online as ref. 11 at < http://www.physics.indiana.edu/~hake/ >.

Hake, R.R. 2000b. "Is It Finally Time to Implement Curriculum S?" *AAPT Announcer* **30**(4):103 (2000); online as ref. 13 at < http://www.physics.indiana.edu/~hake/ >.

Hake, R.R. 2001a. "Suggestions for Administering and Reporting Pre/Post Diagnostic Tests"; online as ref. 14 at < http://www.physics.indiana.edu/~hake >.

Hake, R.R. 2001b. "Pre/Post Paranoia," AERA-D/PhysLrnR post of 17 May 2001 16:01:56 -0700; online at < http://lists.asu.edu/cgi-bin/wa?A2=ind0105&L=aera-d&P=R19884 >.

Hake, R.R. 2002a. "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. 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; an updated version of 1 June 2002 is online at < http://www.pkal.org/template2.cfm?c_id=354 > and as ref. 15 at < http://www.physics.indiana.edu/~hake/ >. General information on the Roundtable and a hot-linked listing of all the papers is at < http://www.pkal.org/template1.cfm?c_id=345 >.

Hake, R.R. 2002c. "Whence Do We Get the Teachers (Response to Madison)". *PKAL Roundtable on the Future: Assessment in the Service of Student Learning*, Duke University, March 1-3; updated on 6/17/02; online at (a) < http://www.pkal.org/template2.cfm?c_id=361 > (Caution - as of 9 December 2002 this version was garbled and contained only two pages of the five-page comment; (b) as ref. 16 at < http://www.physics.indiana.edu/~hake/ >.

Hake, R.R. 2002d. "Comment on 'How do we know if we are doing a good job in physics teaching?' by Robert Ehrlich," *Am. J. Phys*. **70**(10): 1058-1059; online as ref. 17 at < http://www.physics.indiana.edu/~hake >.

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 as ref. 18 at < http://www.physics.indiana.edu/~hake >.

Hake, R.R. 2002f. "Relationship of Individual Student Normalized Learning Gains in Mechanics with Gender, High-School Physics, and Pretest Scores on Mathematics and Spatial Visualization," submitted to the Physics Education Research Conference; Boise, Idaho; August 2002; online as ref. 22 at < http://www.physics.indiana.edu/~hake >.

Hake, R.R. 2002g. "References for Psychometrically Naive Physicists," PhysLrnR post of 6 Nov 2002 14:15:05-0800; online at < http://listserv.boisestate.edu/cgi-bin/wa?A2=ind0211&L=physlrnr&O=A&X=68074C7EB1BE3AAF49&Y=rrhake@earthlink.net&P=4719 >.

Hake R.R. & J.C. Swihart. 1979. "DIagnostic Student COmputerized Evaluation of Multicomponent Courses (DISCOE), *Teaching & Learning* (Indiana University), January; online as ref. 4 at < http://www.physics.indiana.edu/~hake >.

Hake, R. R., R. Wakeland, A. Bhattacharyya, and R. Sirochman. 1994. "Assessment of individual student performance in an introductory mechanics course," *AAPT Announcer* **24**(4):76.

Halloun, I. 1997. "Views About Science and Physics Achievement: The VASS Story." In *The Changing Role of Physics Departments in Modern Universities: Proceedings of the ICUPE*, E.F. Redish and J.S. Rigden, eds. (American Institute of Physics), pp. 605–613; online at < http://modeling.asu.edu/R&E/Research.html >, click on "VASS Results (ICUPE 96)."

Halloun, I. & D. Hestenes. 1985a. "The initial knowledge state of college physics students." *Am. J. Phys.* **53**:1043-1055; online at < http://modeling.asu.edu/R&E/Research.html >.

Halloun, I. & D. Hestenes. 1985b. "Common sense concepts about motion." *Am. J. Phys.* **53**:1056-1065; online at < http://modeling.asu.edu/R&E/Research.html >.

Halloun, I. & D. Hestenes. 1998. "Interpreting VASS Dimensions and Profiles." *Science & Education* **7**(6): 553-577: online (password protected) at < http://modeling.la.asu.edu/R&E/Research.html >; click on "Interpreting VASS."

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 >. Available in English, Spanish, German, Malaysian, Chinese, and Finnish.

Henderson, C. 2002 "Common Concerns About the Force Concept Inventory," *Phys. Teach.* **40**(9): 542-547; online at < http://www.aapt.org/tpt/toc_dec02.html >.

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

Hovland, C. I., A. A. Lumsdaine, and F. D. Sheffield. 1949. "A baseline for measurement of percentage change." In C. I. Hovland, A. A. Lumsdaine, and F. D. Sheffield, eds. 1965. *Experiments on mass communication*. Wiley (first published in 1949).) Reprinted as pages 77-82 in P. F. Lazarsfeld and M. Rosenberg, eds. 1955. *The language of social research: a reader in the methodology of social Research.*" Free Press.

Hunt, R. 2002. "Some Resources on Student Evaluations of Teaching"; online at < http://www.stu.ca/~hunt/evalbib.htm >.

Johnson, M. 2001. Facilitating high quality student practice in introductory physics . *Physics Education Research Supplement to American Journal of Physics)* **69**(7): S2-S11.

Lagemann, E.C. 2000. *An elusive science: the troubling history of education research*. Univ. of Chicago Press.

Lipsey, M.W. & D.B. Wilson. 1993. "The Efficacy of Psychological, Educational, and Behavioral Treatment: Confirmation From Meta-Analysis," *American Psychologist* **48**(12): 1181-1209; online at < http://www.vanderbilt.edu/cerm/MWL_web_bib.htm#methods%20and%20issues >. See also Wilson & Lipsey (2001).

Wilson, D.B. & M.W Lipsey. 2001. "The Role of Method in Treatment Effectiveness Research: Evidence from Meta-Analysis." *Psychological Methods* **6**(4): 413-429; online at < http://www.vanderbilt.edu/cerm/MWL_web_bib.htm#methods%20and%20issues >.

Mahajan, S. & R.R. Hake. 2000. "Is it time for a physics counterpart of the Benezet/Berman math experiment of the 1930's?" *Physics Education Research Conference 2000: Teacher Education*; online as article #6, at the Benezet Centre < http://www.inference.phy.cam.ac.uk/sanjoy/benezet/ >.

Mallinckrodt, J. 2001. "Re: Using the FCI," PhysLrnR post of 3 Oct 2001 19:08:11–0700; online at < http://listserv.boisestate.edu/cgi-bin/wa?A2=ind0110&L=physlrnr&F=&S=&X=068F037F97314CA315&Y=rrhake@earthlink.net&P=1479 >.

Maloney, D., T.L. O'Kuma, C.J. Hieggelke, & A. Van Heuvelen. 2001. "Surveying students' conceptual knowledge of electricity and magnetism," *Physics Education Research Supplement to Am. J. Phys* **69**(7): S12-S23.

May, D.B. & E. Etkina. 2002. "College physics students' epistemological self-reflection and its relationship to conceptual learning," *Am. J. Phys*. **70**(12): 1249-1258.

McDermott, L.C. & E.F. Redish. 1999. "RL-PER1: Resource letter on physics education research," *Am. J. Phys*. **67**(9): 755-767; online at < http://www.physics.umd.edu/rgroups/ripe/perg/cpt.html >.

Meltzer, D.E. 2002a. "The relationship between mathematics preparation and conceptual learning gains in physics: A possible 'hidden variable' in diagnostic pretest scores," *Am. J. Phys.* **70**(12): 1259-1268; also online as article #7 at < http://www.physics.iastate.edu/per/articles/index.html >.

Meltzer, D.E. 2002b. "Normalized Learning Gain: A Key Measure of Student Learning," Addendum to Meltzer (2002a); online as article #7 (addendum) at < http://www.physics.iastate.edu/per/articles/index.html >.

Meltzer, D.E. 2002c. "Issues Related to Data Analysis and Quantitative Methods in Physics Education Research," in *Proceedings of the Physics Education Research Conference 2002*; Boise, Idaho; August 2002; in press.

Meltzer, D.E. & K. Manivannan. 2002. "Transforming the lecture-hall environment: The fully interactive physics lecture," *Am. J. Phys*. **70**(6): 639-654; also online as article #6 at < http://www.physics.iastate.edu/per/articles/index.html >.

NCSU. 2002. *Assessment Instrument Information Page*, Physics Education R & D Group, North Carolina State University; online at < http://www.ncsu.edu/per/TestInfo.html >.

Novak, G.M., E.T. Patterson, A.D. Gavrin, & W. Christian. 1999. *Just-In-Time Teaching: Blending Active Learning with Web Technology*. Prentice Hall.

Pelligrino, J.W., N. Chudowsky, R. Glaser, eds. 2001. *Knowing What Students Know: The Science and Design of Educational Assessment*, National Academy Press; online at < http://www.nap.edu/catalog/10019.html >, esp. Chapter 4 "Contributions of Measurement and Statistical Modeling to Assessment," and references therein.

Redish, E.F. 1999. "Millikan lecture 1998: building a science of teaching physics." *Am. J. Phys*. **67**(7): 562-573; online at < http://www.physics.umd.edu/rgroups/ripe/perg/cpt.html >.

Redish, E. F., J. M. Saul, and R. N. Steinberg. 1997. "On the effectiveness of active-engagement microcomputer-based laboratories," *Am. J. Phys.* **65**(1): 45-54; online at < http://www.physics.umd.edu/rgroups/ripe/perg/cpt.html >.

Redish, E.F. & R.N. Steinberg. 1999. "Teaching physics: figuring out what works," *Phys. Today* **52**(1): 24-30; online at < http://www.physics.umd.edu/rgroups/ripe/perg/cpt.html >.

Redish, E.F., J.M. Saul, R.N. Steinberg. 1998. "Student expectations in introductory physics." *Am. J. Phys.* **66**(3):212-224; online at < 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, Univ. of Maryland.

Savinainen A. & P. Scott. 2002a. "The Force Concept Inventory: a tool for monitoring student learning," *Phys. Educ*. **37**: 45-52; online a < http://kotisivu.mtv3.fi/oma/physics/downloads.html >.

Savinainen A. & P. Scott. 2002b. "Using the Force Concept Inventory to monitor student learning and to plan teaching," *Phys. Educ*. **37**: 53-58; online at < http://kotisivu.mtv3.fi/oma/physics/downloads.html >.

Shavelson, R.J. & L. Towne. 2002. *Scientific Research in Education*, National Academy Press, online at < http://www.nap.edu/catalog/10236.html >.

Slavin, R.E. 1992. *Research Methods in Education.* Allyn and Bacon, 2nd ed.

Steinberg, R. N., and K. Donnelly. 2002. "PER-based reform at a multicultural institution," *Phys. Teach.* **40**(2):108-114; online at Available online at < http://www.aapt.org/tpt/toc_feb02.html >.

Stokstad, E. 2001. "Reintroducing the Intro Course." *Science* **293**: 1608-1610, 31 August 2001: online at < http://www.sciencemag.org/ >. Non-AAAS members may access the article by taking a few minutes to complete a free limited-access registration.

Thornton, R.K. 1995. "Conceptual Dynamics: Changing Student Views of Force and Motion," in *Thinking Physics for Teaching*, Bernardini, C, C. Tarsitani, & M. Vicentini, eds. Plenum.

Thornton, R.K. & D.R. Sokoloff. 1998. "Assessing student learning of Newton's Laws: The force and motion conceptual evaluation and the evaluation of active learning laboratory and lecture curricula," *Am. J. Phys.* **66**(4): 338-352. See also Thornton (1995).

Tobias, S. & R.R. Hake. 1988. "Professors as physics students: what can they teach us? *Am. J. Phys*. **56**(9): 786-794.

Van Domelen, D.J. & A. Van Heuvelen. 2002. "The effects of a concept-construction lab course on FCI performance," *Am. J. Phys*. **70**(7): 779-780.

Zeilik, M., C. Schau, N. Mattern, S. Hall, K.W. Teague, & W. Bisard. 1997. "Conceptual astronomy: A novel model for teaching postsecondary science," *Am. J. Phys.* **65**(10): 987-996.

Zeilik, M., C. Schau, N. Mattern. 1998. "Misconceptions and their change in university-level astronomy courses," *Phys. Teach.* **36**(2):104-107.

Zeilik, M., C. Schau, N. Mattern..1999. "Conceptual astronomy. II. Replicating conceptual gains, probing attitude changes across three semesters," *Am. J. Phys.* **67**(10): 923-927.

Ziman, J. 2000. *Real Science: What it is, and what it means*. Cambridge University Press. See, especially Sec. 9.3 "Codified knowledge," pages 258-266.