Assessment of Student Learning in Introductory Science Courses * ‡

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

OUTLINE

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

‡ The reference is: Hake, R.R. 2002. "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; updated on 6/01/02; online at < **http://www.pkal.org/events/roundtable2002/papers.html** >. Sections II (except for the last paragraph), III, and V are excerpted from Hake, R.R. "Lessons from the physics education reform effort." *Conservation Ecology* **5**(2): 28; online at < **http://www.consecol.org/vol5/iss2/art28** >. Please refer to that paper for details and references.

I. INTRODUCTION

Although a tremendous amount of work has been devoted to assessment in higher education [see, e.g., AAHE 2001a,b; AAHE Assessment Forum 2001a,b; Angelo 1999; Cambridge 2001; FLAG 2001; McTighe & Wiggins 1999; NCSU 2001; Suskie 2000; Wiggins 1998, 1999; Wiggins & McTighe 1998; Wright 2000; AERA-D 2002; ASSESS 2002; EvalTalk 2002) 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*.

For a review of outcome assessments (or lack thereof) in introductory undergraduate courses see the recent popular review by Stokstad (2001). Stokstad writes:

> ". . . too many instructors, say reformers, still engage in the stalest form of pedagogy: nonstop lectures to hundreds of faceless students who sit and listen passively. Supplementing the lectures are textbooks thick with facts and figures and thin on concepts and process. End of chapter homework problems and cookbook labs are solved by 'plugging and chugging' numbers into the equations . . . . One idea that astronomers . . .(such as Bruce Partridge of Haverford College, education officer of the American Astrophysical Association). . . . are gravitating toward is the principle that students understand a concept better it they construct it themselves, step by step, rather than being told what it is and asked to simply remember it. This so-called active learning has become a popular strategy for reforming all manner of introductory courses, from asking students to predict the outcome of a hypothetical situation to sharing information in labs and discussions.
>
> But do these approaches work? . . . The most common way to gauge the success or failure of efforts to reform intro courses is to see how thoroughly students digest the material being taught. But traditional measures can be misleading if they don't require students to understand the material. In a traditional chemistry course, for example, Mary Nakhleh of Purdue . . .found that about half of the students who solved test problems couldn't explain the underlying concepts. Traditional tests may hide that fact, warns, John Moore. . .(Moore 2001a,b). . . of the University of Wisconsin – Madison.
>
> So Moore, Nakhleh, and a handful of other researchers . . . (for Chemistry see, e.g., ASU 2001, Gutiwill-Wise 2001; Milford 1996; Robinson & Nurrenbern 2001; Wright et al. 1998 — for Biology see Lord 1997 Mintzes et al. 1999 Fisher et al. 2000) . . . . have tried to come up with more accurate tools, based on extensive interviews with students. 'These tests are not trivial to design,' says Edward 'Joe' Redish of the University of Maryland. One of the most widely used is the *Force Concept Inventory* (FCI) . . . . (Hestenes et al. 1992, Halloun et al. 1995) . . . . the first version of which was published in 1985 by David Hestenes of the University of Arizona in Tuscon."

In my opinion, Stokstad's (2001) excerpt coveys a reasonably accurate introduction to the current introductory-course learning assessment situation, except for the last sentence, which could be more rigorously phrased as: "One of the most widely used is the *Force Concept Inventory* (FCI) . . . . (Hestenes et al. 1992, Halloun et al. 1995) . . . . the precursor of the FCI, the *Mechanics Diagnostic* (MD) was published by Halloun and Hestenes (1985a,b) of Arizona State University in Tempe, Arizona."

II. ANALYSIS OF MECHANICS-TEST DATA

I have devoted some time to the collection and analysis of MD/FCI test data (Hake 1998a,b). Plots of the 1998 data are shown in Figs. 1 & 2 (from Hake 2002a). In the analysis it was useful to discuss the data in terms of a quantity that I called the "average *normalized* gain" <g>, *defined* as the actual average gain, %<Gain>, divided by the maximum possible actual average gain, %<Gain>$_{max}$:

$$\langle g \rangle = \%\langle Gain \rangle \; / \; \%\langle Gain \rangle_{max} \quad . . . . . \quad (1a)$$

$$\langle g \rangle = (\%\langle posttest \rangle - \%\langle pretest \rangle) \; / \; (100 - \%\langle pretest \rangle) \quad . . . . . \quad (1b)$$

where %<posttest> and %<pretest> are the final (posttest) and initial (pretest) class percentage averages.

For example, suppose that for a given class the test average before instruction was %<pretest> = 44%, and the test average after instruction was %<posttest> = 63%. Then the percentage average actual gain is

$$\%\langle Gain \rangle = 63\% - 44\% = 19\%.$$

The *maximum* possible actual gain for this class would have been

$$\%\langle Gain \rangle max = (100\% - 44\%) = 56\%.$$

Thus, for this example, the average *normalized* gain is

$$\langle g \rangle = \%\langle Gain \rangle \; / \; \%\langle Gain \rangle \; max = 19\%/56\% = 0.34,$$

that is, the class made an average gain of 34% of the maximum possible average gain.

To understand the graphical interpretation of the "average *normalized* gain" <g>, consider the same example as above. Data for that class would be plotted in Fig. 1 as the point [%<pretest> = 44%, %<Gain> = 19%] at the tip of the white arrowhead. This point has an abscissa (100% - 44%) = 56% and ordinate 19%. The absolute value of the slope "s" of the purple dashed line connecting this point to the lower right vertex of the graph is |s| = ordinate/abscissa = 19%/56% = 0.34. Thus, this absolute slope

$$|s| = \%\langle Gain\rangle / \quad (100\% - \%\langle pretest\rangle)$$
$$= \%\langle Gain\rangle / \text{ (maximum possible } \%\langle Gain\rangle)$$
$$= \%\langle Gain\rangle / \%\langle Gain\rangle_{max}$$

is, of course, just the "average *normalized* gain" <g>. Thus, all courses with points close to the lower purple dashed line are judged to be of about equal average effectiveness, regardless of their average pretest scores. A similar calculation for the point [%<pretest> = 32%, %<Gain> = 47%] at the tip of the blue arrowhead yields <g> = 0.69. The *maximum* value of <g> occurs when %<Gain> is equal to $\%\langle Gain\rangle_{max}$ and is therefore 1.00, as shown in Fig. 1.

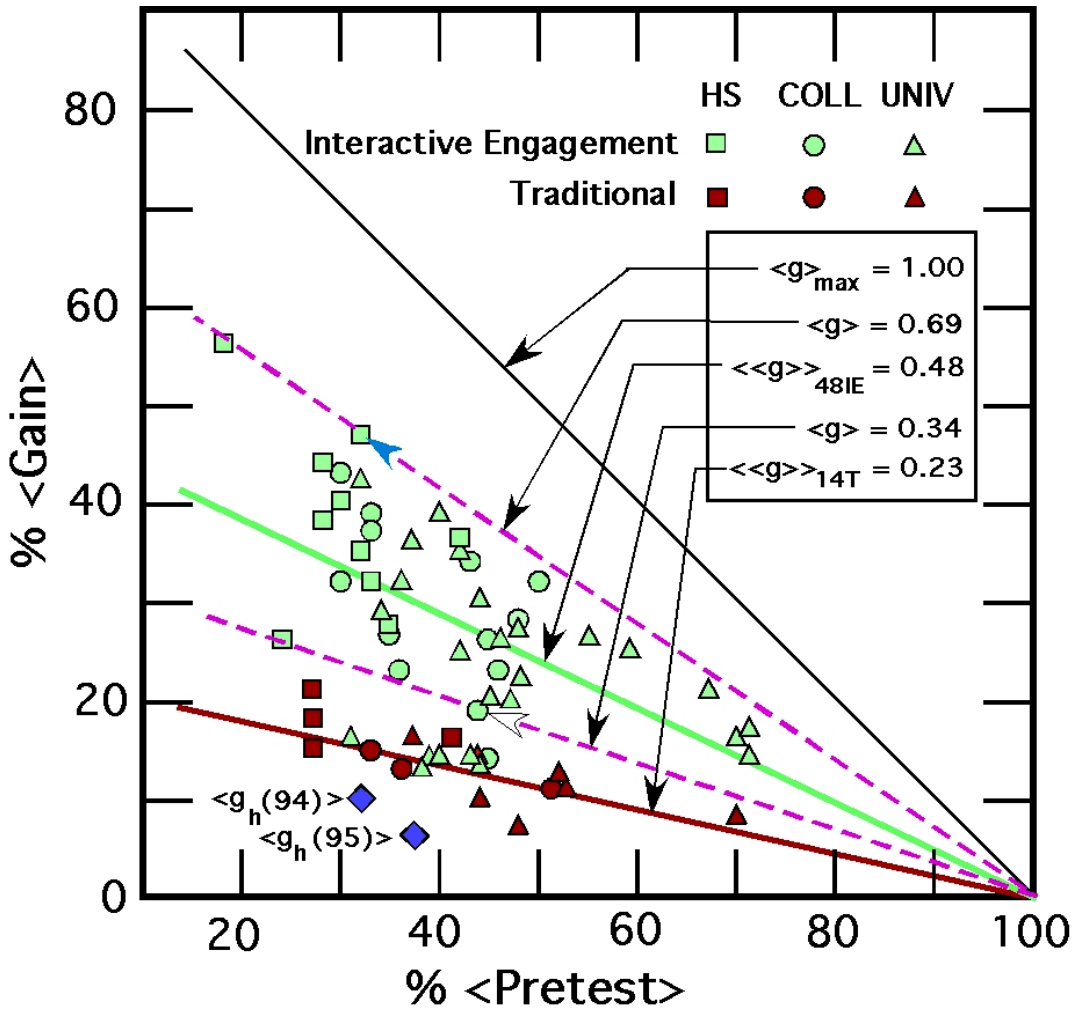

Fig. 1. The %<Gain> vs. %<Pretest> score for 62 courses, enrolling a total of 6542 students. Here, %<Gain> = %<posttest> – %<pretest>, where the angle brackets "<....>" indicate an "average" over all students in the course. Points for high school (HS), college (COLL), and university (UNIV) courses are shown in green for Interactive Engagement (IE) and in red for Traditional (T) courses. The straight negative-slope lines are lines of constant "average normalized gain" <g>. The two dashed purple lines show that most IE courses achieved <g>'s between 0.34 and 0.69. The definition of <g>, and its justification as an index of course effectiveness, is discussed in the text. The average of <g>'s for the 48 IE courses is $\langle\langle g\rangle\rangle_{48IE} = 0.48 \pm 0.14$ (standard deviation) while the average of <g>'s for the 14 T courses is $\langle\langle g\rangle\rangle_{14T} = 0.23 \pm 0.04$ (sd). Here, the double angle brackets "<<....>>" indicate an "average of averages." (Same data points and scales as in Fig. 1 of Hake 1998a.) The two blue-diamond points in the lower left of the graph are "hypothesized <g>'s" for Indiana University premed graduates of high-school physics courses as explained in Hake (2000a).

A histogram of the data of Fig. 1 is shown in Fig. 2.

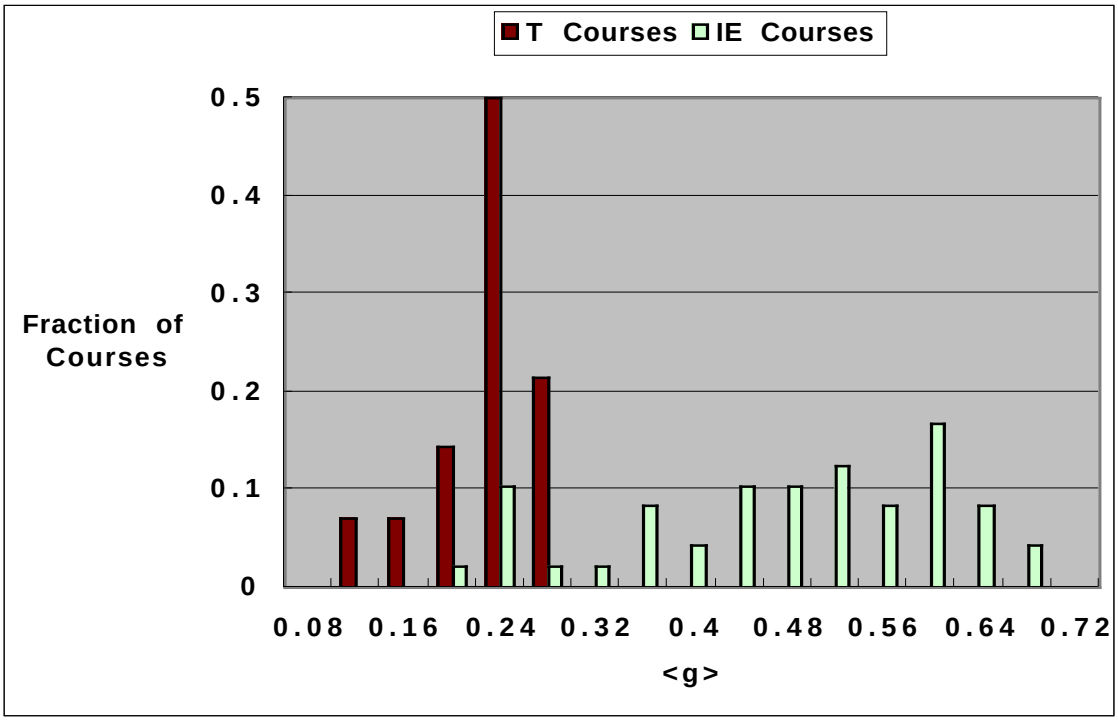

Fig. 2. Histogram of the average normalized gain <g>: red bars show the fraction of 14 Traditional (T) courses (2108 students) and green bars show the fraction of 48 Interactive Engagement (IE) courses (4458 students), both within bins of width <g> = 0.04, centered on the <g> values shown. (Same as Fig. 2 of Hake 1998a.)

Some suggestions for the collection and analysis of pre/post diagnostic tests data gleaned from Hake (1998a,b) are given in Hake (2001a).

For a recent discussion of the "normalized gain" and its history see Hake (2002f,g) (with references changed to the present list) : "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.* . . . Now g, per se, does not necessarily have anything to do with the "mistermed 'normal' curve." In my own survey (Hake, 1998a,b; 2002a) the <g> . . . distributions for both

interactive-engagement (IE) and traditional (T) courses are NOT Gaussian [see Fig. 2 . . . .] so Micceri and other statistics buffs might feel more comfortable using the name "effectiveness index" (Hovland 1949),"gap closing parameter" (Gery 1972), or "Hovland Measure" (Pendelton 1998). 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 is little effort to build a "community map" [Redish 1999; Langeman 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 . . . . . ."

III. CONCLUSIONS OF THE SURVEY

The conclusions of the survey (Hake 1998a,b) as excerpted from Hake (2002a) are listed below.

NOTE: most of the references in the excerpted conclusions **A - G** are *not* repeated here, but can be found in Hake 2002a.

A. The average normalized gain <g> affords a consistent analysis of pre/post test data on conceptual understanding over diverse student populations in high schools, colleges, and universities.

For the 62 courses of the survey (Hake 1998a,b) the correlation of

<g> with (%<pretest>) is + 0.02. . . . . (2)

This constitutes an experimental justification for the use of <g> as a comparative measure of course effectiveness over diverse student populations with widely varying average pretest scores, and is a reflection of the usually relatively small correlations of single student g's with their pretest scores within a given class (Hake 1998a, 2001b; Cummings 1999; Meltzer 2001).

The average posttest score (%<postest>) and the average actual gain (%<Gain>) are less suitable for comparing course effectiveness over diverse groups since their correlations with (%<pretest) are significant. The correlation of

(%<posttest>) with (%<pretest>) is + 0.55, ... (3)

and the correlation of

(%<Gain>) with %<pretest> is – 0.49, ... (4)

both of which correlations would be anticipated. Note that in the *absence* of instruction, a high positive correlation of (%<posttest >) with (%<pretest>) would be expected. The successful use of the normalized gain for the analysis of pre/post test data in this and other physics-education research calls into question the common dour appraisal of pre/post test designs (Lord 1956, 1958; Cronbach & Furby 1970; Cook & Campbell 1979). For a review of the pre/post literature (pro and con) see Wittmann (1997).

Regarding successful use of the normalized gain, in the section "Can Educational Research be Scientific Research" of Hake (2002a), I discuss the fact that normalized gain results for IE and T courses that are consistent with those of (Hake 1998a,b) have now been obtained by physics-education research (PER) groups at the Univ. of Maryland (Redish et al. 1997, Saul 1998, Redish & Steinberg 1999, Redish 1999); Univ. of Montana (Francis et al. 1998); Rennselaer and Tufts (Cummings et al. 1999); North Carolina State Univ. (Beichner et al. 1999); and Hogskolan Dalarna - Sweden (Bernhard 1999); Carnegie Mellon Univ. (Johnson 2001); and City College of New York (Steinberg & Donnelly 2002).

B. Fourteen *Traditional* (T) courses (2084 students) of the survey yielded

$$\mathbf{\langle\langle g\rangle\rangle_{14T} = 0.23 \pm 0.04sd} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (5)$$

Here "$\langle\langle g\rangle\rangle_{14T}$ means an average of <g> over 14 Traditional courses and sd = standard deviation. Considering the *elemental* nature of the MD/FCI questions (many physics teachers regard them as too easy to be used on examinations) and the relatively low <g> = 0.23 (i.e., *only 23% of the possible gain was achieved*), it appears that **traditional (T) courses fail to convey much basic conceptual understanding of Newtonian mechanics to the average student**. Here "traditional" (T) courses were *operationally* defined as "those reported by instructors to make little or no use of IE methods, relying primarily on *passive-student* lectures, recipe labs, and algorithmic-problem exams."

C. Forty-eight *Interactive Engagement* (IE) courses (4458 students) of the survey yielded

$$\mathbf{\langle\langle g\rangle\rangle_{48IE} = 0.48 \pm 0.14sd} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (6)$$

The $\langle\langle g\rangle\rangle_{48IE}$ is over twice that of $\langle\langle g\rangle\rangle_{14T}$. The difference $\langle\langle g\rangle\rangle_{48IE} - \langle\langle g\rangle\rangle_{14T}$ is over 6 sd's of $\langle\langle g\rangle\rangle_{14T}$ and almost two sd's of $\langle\langle g\rangle\rangle_{48IE}$, reminiscent of differences seen in comparing instruction delivered to students in large groups with one-on-one instruction (Bloom 1984). **This suggests that IE courses CAN be much more effective than T courses in enhancing conceptual understanding of Newtonian mechanics.**

In Hake (1998a) I *operationally* defined:

a. "Interactive Engagement (IE) methods" 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, all as judged by their literature descriptions;

b. "IE courses" as those reported by instructors to make substantial use of IE methods.

As indicated in Hake (2002b), disregard for the fact that in Hake (1998a), "traditional" T and "interactive engagement" IE courses were *operational defined* as specified above can lead to misinterpretation and unjustified criticism (Ehrlich 2002) of Hake (1998a).

D. A detailed analysis of random and systematic errors has been carried out (Hake 1998a) but will not be repeated here**.** Possible systematic errors considered were: (a) question ambiguities and isolated false positives (right answers for the wrong reasons); and uncontrolled variables in the testing conditions such as (b) teaching to the test and test question leakage, (c) the fraction of course time spent on mechanics, (d) post and pretest motivation of students, and (f) the Hawthorne/John Henry effects. It was concluded that "**it is extremely unlikely that random or systematic error plays a significant role in the nearly two-standard deviation difference in the <<g>>'s of T and IE courses**."

E. Conclusions A–C above are bolstered by an analysis (Hake 1999a) of the survey data in terms of Cohen's (1988) "effect size" d. The effect size is commonly used in meta-analyses (e.g., Light et al. 1990, Hunt 1997, Glass 2000), and strongly recommended by many psychologists (B. Thompson 1996, 1998, 2000), and biologists (Johnson 1999, Anderson et al. 2000, W.L. Thompson 2001) as a preferred alternative (or at least addition) to the usually inappropriate (Rozeboom 1960, Carver 1993, Cohen 1994, Kirk 1996) t-tests and p values associated with null-hypothesis testing. The effect size d is defined by Cohen (1988, p. 20, 44) as

$$d = |m_A - m_B| / [(sd_A^2 + sd_B^2)/2]^{0.5} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (7)$$

where m_A and m_B are population means expressed in the raw (original measurement) unit, and where the denominator is the root mean square of standard deviations for the A- and B-group means, sometimes called the "pooled standard deviation." For the present survey, Eq. (7) becomes

$$d = [<<g>>_{48IE} - <<g>>_{14T}] / [(sd_{48IE}^2 + sd_{14T}^2)/2]^{0.5} = 2.43 \quad \ldots\ldots\ldots \quad (8)$$

The above "d "can be compared with:

(a) Cohen's (1988, p. 24) rule of thumb – based on typical results in social science research – that d = 0.2, 0.5, 0.8 imply respectively "small," "medium," and "large" effects. But Cohen cautions that the adjectives "are relative, not only to each other, but to the area of behavioral science or even more particularly to the specific content and research method being employed in any given investigation."

(b) The course-enrollment N-weighted <d> = 0.57 obtained for 31 test/control group studies (2559 students) of achievement by Springer et al. (1999, Table 2) in a metaanalyisis of the effect of small-group learning among undergraduates in science, math, and engineering.

The effect size "d" of the present study is much larger than might be expected on the basis of Cohen's rule of thumb, or on the basis of the results of Springer et al. This difference may be related to the facts that in this survey, unlike most education-research meta-analyses (e.g., that of Springer et al., Slavin 1995, and Johnson et al. 2000):

(1) *all* courses covered nearly the same material (here introductory Newtonian mechanics);

(2) the material is conceptually difficult and counterintuitive;

(3) the *same* test (either MD or FCI – see above) was administered to both IE and T classes;

(4) the tests employed are widely recognized for their validity and consistent reliability, have been carefully designed to measure understanding of the key concepts of the material, and are far superior to the plug-in regurgitation type tests so commonly used as measures of "achievement";

(5) the measurement unit gauges the normalized learning *gain* from start to finish of a course, *not* the "achievement" at the end of a course;

(6) the measurement unit <g> is not significantly correlated with students initial knowledge of the material being tested;

(7) the "treatments" are all patterned after those *published by education researchers in the discipline being tested*.

I think that the Springer et al. meta-analysis probably understates the potential of small-group learning for advancing conceptual understanding and problem-solving ability.

F. Considering the elemental nature of the MD/FCI tests, current IE methods and their implementation need to be improved, **since none of the IE courses achieves <g> greater than 0.69.** In fact, seven of the IE courses (717 students) achieved <g>'s close to those of the T courses. Case histories of the seven low–<g> courses (Hake 1998b) suggest that implementation problems occurred that might be mitigated by:

(1) apprenticeship education of instructors new to IE methods,

(2) emphasis on the nature of science and learning throughout the course,

(3) careful attention to motivational factors and the provision of grade incentives for taking IE-activities seriously,

(4) recognition of and positive intervention for potential low-gain students,

(5) administration of exams in which a substantial number of the questions probe the degree of conceptual understanding induced by the IE methods,

(6) use of IE methods in all components of a course and tight integration of those components.

Personal experience with the Indiana IE courses and communications with most of the IE instructors in the survey suggest that similar implementation difficulties probably occurred to a greater or lesser extent in all the IE courses and are probably partially responsible for the wide spread in the <g>'s, apparent for IE courses in Figs. 1 & 2.

G. I have plotted (Hake 1998a) average post-course scores on the problem-solving *Mechanics Baseline* test (Hestenes & Wells, 1992) [available for 30 (3259 students) of the 62 courses of the survey] **vs those on the conceptual FCI.** There is a very strong positive correlation r = + 0.91 of the MB and FCI scores. This correlation and the comparison of IE and T courses at the same institution (Hake 1998a) **imply that *IE methods enhance problem-solving ability*.**

IV. DOES <g> TELL ALL?

Does the normalized gain <g> provide a definitive assessment of the *overall* effectiveness of an introductory physics class? **NO!** It assesses "*only the attainment of a **minimal** competence in 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. 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 2002c), 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.)

The design and testing of instruments for the 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), as discussed in Hake (2002a, Lesson #3). 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; 2002d,e; Mahajan & Hake 2000; Benezet 1935/36) .

V. SIX ASSESSMENT LESSONS

The primary assessment lessons from the physics education research effort [of the 14 lessons listed Hake (2002a)] are listed below.

NOTE: most of the references in the excerpted *lessons* **L2, 3, 5, 8, 9, 14** are *not* repeated here, but can be found in Hake 2002a.

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

Education research in biology (Hake 1999a,b), chemistry (Herron & Nurrenbern 1999), and engineering (Felder et al. 2000a,b), although neither as extensive nor as systematic as that in physics (McDermott & Redish 1999, Redish 1999), is consistent with the latter in suggesting that in conceptually difficult areas, interactive engagement (IE) methods are more effective than traditional (T) passive-student methods in enhancing students' understanding. Furthermore, there is some preliminary evidence that learning in IE physics courses is substantially retained one to three years after the courses have ended (Chabay 1997, Francis et al. 1998, Bernhard 2000). I see no reason to doubt that enhanced understanding and retention would result from more use of interactive engagement methods in other science and even non-science areas, but substantive research on this issue is sorely needed – see L3 & L4.

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 . . .(widely recognized and utilized). . . 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.*

The lengthy and arduous process of constructing valid and reliable multiple choice tests has been discussed by Halloun & Hestenes (1985a), Hestenes et al. (1992), Beichner (1994), Aubrecht (1991), and McKeachie (1999). In my opinion [and contrary to the standpoint of Wiggins (1999)] such hard-won Diagnostic Tests that cover important parts of common introductory courses are national assets whose confidentiality should be as well protected as the MCAT (Medical College Admission Test). Otherwise the test questions may migrate to student files and thereby undermine education research that relies upon the validity of such tests. Suggestions for both administering Diagnostic Tests and reporting their results so as to preserve confidentiality and enhance assessment value have been given by Hake (2001b).

L5. The development of effective educational methods within each discipline requires a redesign process of continuous long-term classroom use, feedback, ASSESSMENT, research analysis, and revision.

Wilson and Daviss (1994) suggest that the "redesign process," used so successfully to advance technology in aviation, railroads, automobiles, and computers can be adapted to K-12 education reform through "System Redesign Schools." Redesign processes in the reform of introductory undergraduate physics education have been undertaken and described by McDermott (1991) and by Hake (1998a). In my opinion "redesign" at both the K-12 and undergraduate levels can be greatly assisted by the promising *Scholarship of Teaching & Learning* movement (Carnegie Academy 2000) inspired by Boyer (1990) and the Boyer Commission (1998).

L8. College and university faculty tend to overestimate the effectiveness of their own instructional efforts and thus tend to see little need for educational reform.

As examples of this tendency see Geilker (1997) [countered by Hilborn (1998)]; Griffiths (1997) [countered by Hestenes (1998)]; Goldman (1998); Mottman (1999a,b) [countered by Kolitch (1999), Steinberg (1999), and Hilborn (1999)]; and Carr (2000).

L9. Such complacency can sometimes be countered by the administration of high-quality standardized tests of understanding and by "video snooping."

a. Harvard's Eric Mazur (1997) was very satisfied with his introductory-course teaching - he received very positive student evaluations and his students did reasonably well on "difficult" exam problems. Thus it came as a shock when his students fared hardly better on the "simple" FCI than on their "difficult" midterm exam. As a result, Mazur developed and implemented his interactive- engagement *Peer Instruction* method as a replacement for his previous traditional passive-student lectures. This change resulted in much higher <g>'s . . . (average normalize gains). . . on the FCI as shown by comparison of the red and green triangular points with average pretest scores in the vicinity of 70% in Fig. 1.

b. Like Mazur, most Harvard faculty members are proud of their undergraduate science courses. However, the videotape *Private Universe* (Schneps & Sadler 1985) shows Harvard graduating seniors being asked "What causes the seasons?" Most of them confidently explain that the seasons are caused by yearly changes in the distance between the Sun and the Earth! Similarly most MIT faculty regard their courses as very effective preparation for the difficult engineering problems that will confront their elite graduates in professional life. However the videotape *Simple Minds* (Shapiro et al. 1997) shows MIT graduating seniors having great trouble getting a flashlight bulb to light, given one bulb, one battery, and one piece of wire.

L14. “Education is not rocket science, it’s much harder.”

George Nelson, astronaut and astrophysicist, as quoted by Redish (1999).

My own belief, conditioned by 40 years of research in superconductivity and magnetism, 28 years in physics teaching, and 16 years in education research, is that *effective* education (both physics teaching and education research) is harder than solid-state physics. The latter is, of course, several orders of magnitude harder than rocket science. Nuclear physicist Joe Redish (1999) writes: “The principles of our first draft of a community map for physics education are different in character from the laws we would write down for a community map of the physical world. They are much less like mathematical theorems and much more like heuristics. This is not a surprise, since the phenomena we are discussing are more complex and at a much earlier stage of development.”

References & Footnotes

AAHE. 2001a. "2001 AAHE's Assessment Research Forum," download the pdf to obtain *Enacting a Scholarship of Assessment - A Research Agenda*; online < **http://www.aahe.org/assessment/** >.

AAHE Assessment Forum. 2001b. "9 Principles of Good Practice for Assessing Student Learning," online at < **http://www.aahe.org/assessment/principl.htm** >.

AAHE. 2002. 2002 Assessment Conference - *Assessment: A Shared Commitment*, online at < **http://www.aahe.org/assessment/2002/** >.

AERA-D. 2000. American Educational Research Association, Measurement and Research Methodology discussion list with excellent searchable LISTSERV archives at < **http://lists.asu.edu/archives/aera-d.html** >.

Angelo, T.A. 1999. "Doing Assessment As If Learning Matters Most" *AAHE Bulletin*, May 1999; online at < **http://www.aahe.org/bulletin/angelomay99.htm** >.

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.

ASSESS. 2002. Discussion list with excellent searchable LISTSERV archives at < **http://lsv.uky.edu/archives/assess.html** >.

ASU. 2001. Arizona State University Modeling Group, *Chemistry Concept Inventory*; password protected at < **http://hellevator.daisley.net** >.

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 as ref. 6 at < **http://wol.ra.phy.cam.ac.uk/sanjoy/benezet/** >. See also Mahajan & Hake (2000).

Cambridge, B. 2001. "Assessing the Real College Experience: The architects of the National Survey of Student Engagement talk about the meaning behind the numbers." *AAHE Bulletin* January 2001; online at < **http://www.aahe.org/bulletin/real_college.htm** >.

Ehrlich, R. 2002. "How do we know if we're doing a good job in physics teaching?" *Am. J. Phys.* **70**(1), 24-29.

Elby, A. 1999. "Another reason that physics students learn by rote." *Physics Ed. Res. Supplement to Am. J. Phys.* **67**(7): S52-S57.

EvalTalk. 2002. Discussion list with excellent searchable LISTSERV archives at < **http://bama.ua.edu/archives/evaltalk.html** >. Sponsored by American Evaluation Association < **http://www.eval.org/** >, "devoted to the application and exploration of evaluation in all its forms."

Fisher, K.M., J.H. Wandersee, & D.E. Moody. 2000. *Mapping Biology Knowledge*. Kluwer.

FLAG. 2001. *Field-tested Learning Assessment Guide for science, math, engineering, and technology instructors*. University of Wisconsin – Madison; online at < **http://www.wcer.wisc.edu/nise/cl1/flag/** >.

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

Gutwill-Wise, J.P. 2001. "The impact of active and context-based learning in introductory chemistry courses: An early evaluation of the modular approach." *J. Chem. Educ*. **78** (5), 684-690; abstract online at < **http://jchemed.chem.wisc.edu/Journal/Issues/2001/May/abs684.html** >; see the online supplement at < **http://jchemed.chem.wisc.edu/Journal/Issues/2001/May/PlusSub/JCESupp/supp684.html** > for the concept tests and interviews, with their coding keys, and the preclass and postclass attitudinal surveys used in this study.

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**(1):64-74; online at < **http://www.physics.indiana.edu/~sdi/** >.

Hake, R.R. 1998b. Interactive-engagement methods in introductory mechanics courses, submitted to *Physics Ed. Res. Supplement to Am. J. Phys*; online at < **http://www.physics.indiana.edu/~sdi/** >. In this sadly unpublished (*physics-education research has no archival journal!*) essential companion paper to Hake (1998a): (a) average pretest and posttest scores, (b) standard deviations, (c) instructional methods and materials used, and (d) institutions and instructors for each of the survey courses in Hake (1998a) are tabulated and referenced. In addition, the paper also gives case histories for the seven Interactive Engagement (IE) courses whose effectiveness, as gauged by pre- to posttest gains, was close to those of traditional courses, advice for implementing IE methods, and suggestions for further research.

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

Hake, R.R. 2000b. "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. 2001a. "Suggestions for Administering and Reporting Pre/Post Diagnostic Tests", unpublished; online as ref. 14 at < **http://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** >.

Hake, R.R. 2002b. Letter to the Editor, "Comment on 'How do we know if we're doing a good job in physics teaching?' by Robert Ehrlich [*Am. J. Phys*. **70**(1), 24-29 (2002)]," *Am. J. Phys.*, accepted for publication and online as ref. 15 at < **http://www.physics.indiana.edu/~hake** >.

Hake, R.R. 2002c. "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://listserv.nd.edu/cgi-bin/wa?A2=ind0204&L=pod&P=R14535** >.

Hake, R.R. 2002d. "Physics First: Precursor to Science/Math Literacy for All?" accepted for publication in the Summer 2002 issue of the American Physical Society's *Forum on Education Newsletter* < **http://www.aps.org/units/fed/index.html** > / "Forum newsletters" where "/" means "click on." Also online as ref. 17 at < **http://www.physics.indiana.edu/~hake** >. This is a truncated version of Hake (2002e).

Hake, R.R. 2002e. "Physics First: Opening Battle in the War on Science/Math Illiteracy?" submitted to *The Physics Teacher* < **http://www.aapt.org/tpt/** > on 6/3/02 >; online as ref. 18 at < **http://www.physics.indiana.edu/~hake** >.

Hake, R.R. 2002f. "Re: Normalized Gains," ASSESS/AERA-D/EvalTalk/PhysLrnR/POD post of 11 Apr 2002 20:25:41–0700; online at
< **http://listserv.nd.edu/cgi-bin/wa?A2=ind0204&L=pod&P=R4439** >.

Hake, R.R. 2002g. " Re: Normalized Gains - Correction"
ASSESS/AERAD/EvalTalk/PhysLrnR/POD post 12 Apr 2002 10:12:08-0700; online at
< **http://listserv.nd.edu/cgi-bin/wa?A2=ind0204&L=pod&P=R5051** >.

Hake R.R. & J.C. Swihart. 1979. "Diagnostic Student Computerized Evaluation of Multicomponent Courses (DISCOE), *Teaching & Learning*
(Indiana University), January; online at < **http://www.physics.indiana.edu/~h**ake >.

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,* ed. by E.F. Redish and J.S. Rigden, (American Institute of Physics), pp. 605–613; online at
< **http://www.inco.com.lb/halloun/hallounTEST.html** >.

Halloun, I. & D. Hestenes. 1985a. "The initial knowledge state of college physics students." *Am. J. Phys.* **53**:1043-1055; online at
< **http://www.inco.com.lb/halloun/hallounTEST.html** >.

Halloun, I. & D. Hestenes. 1985b. "Common sense concepts about motion." *Am. J. Phys.* 53:1056-1065; online at < **http://www.inco.com.lb/halloun/hallounTEST.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** >.

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

Hestenes, D., M. Wells, & G. Swackhamer, 1992. "Force Concept Inventory." *Phys. Teach.* **30**: 141-158.

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.

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

Lord, T.R. 1997. "A Comparison Between Traditional and Constructivist Teaching in College Biology," *Innovative Higher Education* **21**(3), 197-216 (1997); abstract online at < **http:www.UGA.edu/ihe/IHEabstracts.html** >.

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 ref. 6 at < **http://wol.ra.phy.cam.ac.uk/sanjoy/benezet/** >.

McTighe, J. & G.P. Wiggins. 1999. *The Understanding by Design Handbook. Supervision and Curriculum Development.* Supervision and Curriculum Development.

Milford, D.R. 1996. "An Inventory for Measuring College Students' Level of Misconceptions in First Semester Chemistry," Purdue Masters Degree thesis; online at < **http://faculty.pepperdine.edu/dmulford/thesis/Title.html** >.

Mintzes, J.L., J.H. Wandersee, & J.D. Novak, *Assessing Science Understanding: A Human Constructivist View.* Academic Press. 1999.

Moore, J.W. 2001a. "Testing, Testing" editorial, *J. Chem. Educ.* **78 (7)**: 855; online at < **http://jchemed.chem.wisc.edu/Journal/Issues/2001/Jul/abs855.html** >.

Moore, J.W. 2001b. "Testing the Teacher? Or Teaching the Test?" editorial, *J. Chem. Educ*. **78** (8): 991; online at < **http://jchemed.chem.wisc.edu/Journal/Issues/2001/Aug/abs991.html** >.

NCSU. 2001. North Carolina State University, *Internet Resources for Higher Education Outcomes Assessment*; online at < **http://www2.acs.ncsu.edu/UPA/assmt/resource.htm** >.

Pendleton, W.W. 1998. "Re: measuring change," AERA-D post of 4 May 1998 12:11:16-0400; online at < **http://lists.asu.edu/cgi-bin/wa?A2=ind9805&L=aera-d&P=R509** >.

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** >.

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** >.

Robinson, W.R. and S.C. Nurrenbern. 2001. "Conceptual Questions & Challenge Problems"; online at < **http://JCHEMED.chem.wisc.edu/JCEWWW/Features/index.html** > / "Conceptual Questions & Challenge Problems" where "/" means "click on." See especially the *Chemical Concepts Inventory*. The authors provide "examples of conceptual questions, an introduction to several different types of conceptual questions, and a discussion of what constitutes a conceptual question." Details are provided at:

a. A discussion of conceptual questions.
b. Types of conceptual questions.
c. An introduction to writing conceptual questions.
d. A library of conceptual questions.
e. Other sources of conceptual questions.
f. A multiple choice misconceptions inventory.

Shavelson, R.J. & L. Towne, editors. 2001. *Scientific Research in Education*. National Academy Press; online at < **http://www.nap.edu/catalog/10236.html** >.

Stokstad, E. 2001. "Reintroducing the Intro Course." *Science* **293**: 1608-1610, 31 August 2001: "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"; 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

Suskie, L. 2000. "Fair Assessment Practices" *AAHE Bulletin* May 2000; online at < **http://www.aahe.org/bulletin/may2.htm** >.

Wiggins, G.P. 1999. "Assessing Student Performance: Exploring the Purpose and Limits of Testing." Jossey-Bass.

Wiggins, G.P. 1998 "Educative Assessment." Jossey-Bass.

Wiggins, G.P. & J. McTighe 1998 *Understanding by Design.* Association for Supervision and Curriculum Development; see also McTighe, J. & G.P. Wiggins (1999) .

Wright J.C., S.B. Millar, S.A. Kosciuk, D.L. Penberthy, P.H. Williams, and B.E. Wampold. 1998. "A Novel Strategy for Assessing the Effect of Curriculum Reform on Student Competence," *J. Chem. Ed*. **75**(8), 986 (1998); abstract online at < **http://jchemed.chem.wisc.edu/Journal/Issues/1998/Aug/abs986.html** >.

Wright, B.D. 2000. "Assessing Student Learning," in *Learning from Change: Landmarks in Teaching and Learning in Higher Education from Change Magazine 1969-1999*, Deborah DeZure, ed. Stylus: "Traces the development and continuity of the assessment movement in higher education."

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.