DESIGN-BASED RESEARCH IN PHYSICS EDUCATION: A REVIEW*◊

Richard R. Hake

In this chapter I argue that some physics education research (PER) is design-based research (DBR) and that an important DBR-like facet of PER, the pre/post testing movement, has the potential to improve drastically the effectiveness of undergraduate instruction generally, the education of pre-service teachers in particular, and, as a net result, the education of the general population.

I. Some Physics Education Research Is Design-Based Research

In their resource letter on physics education research, McDermott & Redish (1999) list about 160 empirical studies, extending over almost three decades, that (a) focus on the learning of physics by students, (b) represent systematic research, and (c) give procedures in sufficient detail that they can be reproduced. My own effort in developing, testing, and disseminating Socratic dialogue inducing (SDI) laboratories is rather typical of the work reported by long-established physics education research groups and referenced by McDermott & Redish.

SDI laboratories emphasize hands- and heads-on experience with simple mechanics experiments and facilitate the interactive engagement of students with course material. They are designed to promote students' mental construction of concepts through:

(1) interactive engagement of students who are induced to think constructively about simple Newtonian experiments which produce conflict with their commonsense understandings;

(2) the Socratic method (e.g., Arons, 1997; Hake, 1992, 2002d) of the *historical* Socrates (Vlastos, 1990, 1991), not Plato's alter ego in the *Meno* (as mistakenly assumed by many - even some physicists), utilized by experienced instructors who have a good understanding of the material and are aware of common student preconceptions and failings;

(3) considerable interaction between students and instructors and thus a degree of individualized instruction;

(4) extensive use of multiple representations (verbal, written, pictorial, diagrammatic, graphical, and mathematical) to model physical systems;

(5) real world situations and kinesthetic sensations (which promote student interest and intensify cognitive conflict when students' direct sensory experience does not conform to their conceptions);

(6) cooperative group effort and peer discussions;

(7) repeated exposure to the coherent Newtonian explanation in many different contexts.

_____

As described in Hake (1987, 1991, 1992, 2000, 2002d, 2007c), Hake & Wakeland (1997), and Tobias & Hake (1988), SDI laboratories were inspired by the astute empirical observations of Arnold Arons (1973, 1974, 1983, 1986, 1997) who had the uncommon sense to "shut up and listen to what students say" in response to probing Socratic questions.

In numerous publications, I scientifically and iteratively developed (Hake, 1987), explored (Tobias & Hake, 1988; Hake, 1991; 1992), confirmed (Hake, 1998a, 1998b, 2002a, 2002b, 2005, 2006, 2007a), and disseminated (Hake, 2000, 2002b, 2002d, 2007b) SDI laboratories. My research and development involved active innovation and intervention in the classrooms of introductory physics classes for prospective elementary teachers (Hake, 1991), premedical students (Hake, 1987, 1992; Hake & Wakeland, 1997), and even nonphysical science professors (Tobias & Hake, 1988). Further, my research and development drew upon models from design and engineering, in that SDI laboratories were designed initially by taking into account my own teaching experience, the advice of the late Arnold Arons, and the physics education and cognitive science literature. Then, trial runs that exposed design failures and successes were carried out in regularly scheduled courses; this phase was followed by exploratory out-of-class research with paid student subjects involving videotape analysis of SDI laboratory sessions (Hake, 2000) and interviews with students.

Three redesigns, retests, and more exploratory, in- and out-of-class research and development over many cycles of application—all in typical engineering fashion—generated new ideas for physics teaching (Hake, 1987, 1992, 2007b; Tobias & Hake, 1988; Hake & Wakeland, 1997) and contributed to the transformation of the traditional recipe laboratory. I sought to understand learning and teaching while I was active as the instructor (Hake, 1987, 1992; Tobias & Hake, 1988; Hake & Wakeland, 1997). As explained in Hake (2002a) (in the section titled "Can Educational Research Be Scientific Research?"), my research and development were examples of use-inspired, basic scientific research, consistent with the theses of Shavelson and Towne (2002) and Stokes (1997). Such work contributed to the movement of at least some introductory mechanics courses from malfunction to function, as shown by pre/post test results (Hake, 1998a, 1998b; 2002a, 2002b, 2005, 2006, 2007a).

Considering the above two paragraphs, I submit that some PER qualifies as design-based research as characterized by Kelly (2003a).

BUT WAIT! Should not the major concern of education research be K-12, as appears to be the area of activity for most education specialists, psychologists, and cognitive scientists? Not necessarily. The NSF's (1996) report *Shaping the Future* hit the nail squarely on the head (my italics):

> Many faculty in Science, Mathematics, Engineering, & Technology (SME&T) at the postsecondary level continue to blame the schools for sending underprepared students to them. But, increasingly the higher education community has come to recognize the fact that *teachers and principals in the K-12 system are all people who have been educated at the undergraduate level, mostly in situations in which SME&T programs have not taken seriously enough their vital part of the responsibility for the quality of America's teachers.*

In my opinion, the DBR-like, pre/post testing movement, stimulated to some extent by physics education research, has the potential to improve undergraduate science instruction dramatically and thereby upgrade K–12 science education. Currently, prospective K–12 teachers derive little conceptual understanding from traditional, undergraduate, introductory science courses; then they tend to teach as they were taught, with similar negative results. As emphasized by Goodlad (1990):

> Few matters are more important than the quality of the teachers in our nation's schools. Few matters are as neglected....A central thesis of this book is that there is a natural connection between good teachers and good schools and that this connection has been largely ignored....*It is folly to assume that schools can be exemplary when their stewards are ill-prepared.* (My italics, pp. xi-xii.)

II. Pre/post Testing In Physics Education Research

The pre/post testing movement in PER was initiated by the landmark work of Ibrahim Halloun and David Hestenes (1985a, 1985b). Previously, in "Lessons from the Physics Education Reform Effort" (Hake, 2002a) I wrote:

> For over three decades, physics education researchers repeatedly showed that *traditional* (T) introductory physics courses with passive-student lectures, recipe laboratories, and algorithmic problem exams were of limited value in enhancing conceptual understanding of the subject (McDermott & Redish, 1999). Unfortunately, this work was largely ignored by the physics and education communities until Halloun and Hestenes (1985a, 1985b) devised the *Mechanics Diagnostic* (MD) test of conceptual understanding of Newtonian mechanics. Among the virtues of the MD, and the subsequent *Force Concept Inventory* (FCI) tests (Hestenes, Wells, & Swackhamer, 1992) are: (a) the multiple-choice format facilitates relatively easy administration of the tests to thousands of students, and (b) the questions probe for conceptual understanding of basic concepts of Newtonian mechanics in a way that is understandable to the novice who has never taken a physics course (and thus can be given as an introductory-course pretest), while at the same time are rigorous enough for the initiate.

Construction of the MD test involved laborious *qualitative* analysis of extensive interviews with students and the study of prior qualitative and quantitative work on misconceptions (McDermott & Redish, 1999). All this led to a "taxonomy of common sense concepts about motion" (Halloun & Hestenes, 1985b; Hestenes, Wells, & Swackhamer, 1992) and finally the formulation of a balanced and valid test that has proven consistently to be highly reliable, as judged by relatively high Kuder–Richardson reliability coefficients KR-20 in the 0.8 to 0.9 range (see, e.g., Halloun & Hestenes, 1985b; Hake, 1998a, 1998b).

Halloun and Hestenes (1985a, 1985b) then used the MD in *quantitative* classroom research involving massive pre/post testing of students in both calculus and noncalculus-based introductory physics courses at Arizona State University. Their conclusions were:

1. . . . the student's initial, qualitative, common-sense beliefs about motion and. . .[its]. . .causes have a large effect on performance in physics, but conventional instruction induces only a small change in those beliefs.

2. Considering the wide differences in the teaching styles of the four professors. . .[involved in the study]. . .the basic knowledge gain under conventional instruction is essentially independent of the professor. [pp. 1048 of Halloun & Hestenes (1985a)]

BUT WAIT! Can multiple choice tests gauge higher level cognitive outcomes such as the conceptual understanding of Newtonian mechanics? Wilson & Bertenthal (2005) think so, writing (p. 94):

Performance assessment is an approach that offers great potential for assessing complex thinking and learning abilities, but multiple choice items also have their strengths. For example, although many people recognize that multiple-choice items are an efficient and effective way of determining how well students have acquired basic content knowledge, many do not recognize that they can also be used to measure complex cognitive processes. For example, the *Force Concept Inventory* . . . [Hestenes, Wells, & Swackhamer, 1992] . . . is an assessment that uses multiple-choice items to tap into higher-level cognitive processes.

The Halloun & Hestenes (1985a, 1985b) research results were consistent with the findings of many researchers in physics education (McDermott & Redish, 1999), which suggested that traditional, passive-student, introductory physics courses, even those delivered by the most talented and popular instructors, imparted little conceptual understanding of Newtonian mechanics. But the Halloun and Hestenes research went far beyond earlier work because it offered physics teachers and researchers a valid and consistently reliable test that could be employed to gauge the effectiveness of *traditional* mechanics instruction, then to track continually the merit of the *nontraditional* methods with respect to (a) traditional methods, (b) one another, and (c) various modes of implementation. Thus, it could contribute to a steady, albeit very slow, iterative increase in the effectiveness of introductory mechanics instruction nationwide.

For example, consider the MD/FCI-induced changes in introductory physics courses at pacesetting Harvard University and the Massachusetts Institute of Technology. Harvard University's Mazur (1997) wrote (p. 4):

When reading this [Halloun & Hestenes, (1985a, 1985b, 1987); Hestenes, 1987], . . . my first reaction was "Not my students. . .!" Intrigued, I decided to test my own students' conceptual understanding, as well as that of physics majors at Harvard. . . .the results of the test came as a shock: The students fared hardly better on the Halloun and Hestenes test [1985a] than on their midterm exam. Yet the Halloun and Hestenes test is *simple*, whereas the material covered by the examination (rotational dynamics, moments of inertia) is of far greater difficulty, or so I thought.

In Table I, p. 972, of Crouch & Mazur (2001), note:

    1. The abrupt increase in the average normalized gain <g> (see below) from 0.25 in 1990 to 0.49 in 1991 when Mazur replaced his passive-student lectures (that netted very positive student evaluations—many administrators erroneously regard student evaluations as valid measures of students' learning!)—with the interactive engagement of peer instruction.

    2. The gradual increase in the average normalized gain <g> from 0.49 in 1991 to 0.74 in 1997 as various improvements (Crouch & Mazur, 2001) were made in the implementation of peer instruction.

MIT's John Belcher (2003), describing his institute's introductory physics course transition from traditional to interactive-engagement, wrote (p. 8):

    What is the motivation for this transition to such a different mode for teaching introductory physics? First, the traditional lecture/recitation format for teaching 8.01 and 8.02 has had a 40–50% attendance rate, even with spectacularly good lecturers (e.g., Professor Walter Lewin), and a 10% or higher failure rate. Second, there has been a range of educational innovations at universities other than MIT over the last few decades that demonstrate that any pedagogy using "interactive-engagement" methods results in higher learning gains as compared to the traditional lecture format (e.g., see Halloun & Hestenes, 1985a, 1985b; Hake, 1998a; Crouch & Mazur, 2001), usually accompanied by lower failure rates. Finally, the mainline introductory physics courses at MIT do not have a laboratory component.

The Harvard University and MIT results are consistent with those from hundreds of other introductory physics courses employing either traditional or interactive engagement methods, as evidenced by the meta-analysis discussed below.

BUT WAIT! Considering the canonical arguments regarding the invalidity of pre/post testing evidence, should not all the results cited above be viewed with grave suspicion? The Coalition for Evidence-Based Policy (2003), in its *Identifying and Implementing Educational Practices Supported by Rigorous Evidence: A User Friendly Guide* (Institute for Education Sciences, 2003) states (p. 2):

    Pre-post study designs often produce erroneous results. . . .A "pre-post" study examines whether participants in an intervention improve or regress during the course of the intervention, and then attributes any such improvement or regression to the intervention. The problem with this type of study is that, without reference to a control group, it cannot answer whether the participants' improvement or decline would have occurred anyway, even without the intervention. This often leads to erroneous conclusions about the effectiveness of the intervention.

In my opinion, the above objection is irrelevant for most of the pre/post testing studies considered here. The reason is that fairly well-matched control groups *have* been used; they are the courses taught by the traditional method. The matching is due to the fact that (a) within any one institution, the test (interactive engagement [IE]) and the control (traditional [T]) groups are drawn from the same, generic, introductory course taken by relatively homogeneous groups of students; and (b) IE- and T-course teachers in all institutions are drawn from the same generic pool of introductory course physics teachers who, judging from the uniformly poor, average, normalized gains <g> (see below) they obtain in teaching traditional courses, do not vary greatly in their ability to enhance students' learning.

Even if one were to maintain unrealistically that the traditional courses and instructors were not at all matched and that the pre/post testing of IE courses had no controls whatsoever, Table 2 of Lipsey and Wilson (1993) shows that, for the data they survey, the mean effect size of one-group pre/post testing studies is 0.76, compared to the control and comparison studies mean of 0.46. But both the difference in these two means and their individual magnitudes are small in comparison with the Cohen effect size of 2.43 (as calculated by comparing average <g>s for IE and T courses), or 2.16 (as calculated from the *actual gains* [<%post> – <%pre> of IE courses]) that I have reported (see Hake, 2002a) It is interesting that the same Lipsey and Wilson table shows that, for the meta-analyses surveyed by them, the average effect size is 0.46 for random studies and 0.41 for nonrandom studies. Thus, the non-random studies evidently tend to *underestimate* the effectiveness of programs, if we are to accept the Coalition for Evidence-Based Policy's (2003) dubious pronouncement (p. 1) that randomized controlled trials are the gold standard for gauging the effectiveness of an intervention:

> Well-designed and implemented randomized controlled trials are considered the "gold standard" for evaluating an intervention's effectiveness, in fields such as medicine, welfare and employment policy, and psychology. This section discusses what a randomized controlled trial is, and outlines evidence indicating that such trials should play a similar role in education.

Then too, as discussed in Hake (2005, 2006, 2007a), the anti pre/post testing arguments by the psychometric authorities Lord (1956, 1958) and Cronbach and Furby (1970) that gain scores are unreliable have been called into question by, for example, Rogosa, Brandt, & Zimowski (1982), Zimmerman & Williams (1982), Rogosa & Willett (1983, 1985), Rogosa (1995), Wittmann (1997), Zimmerman (1997), & Zumbo (1999). All this more recent work should serve as an antidote for the emotional, pre/post testing paranoia that grips many educational researchers.

III. Meta-Analysis Of Pre/Post Learning Gains

A reservation regarding meta-analysis has been raised by DBR pioneer Carl Bereiter (2002) in his cogent discussion of "Design Research for Sustained Innovation." Bereiter wrote:

> Rather more successful than [attribute-treatment interactions (ATIs) to discover the optimal matching of persons to treatments (Cronbach, 1975)] has been meta-analysis (Glass, McGaw, & Smith, 1981), in which a number of different studies that are judged to involve the same variable are brought together into a statistically powerful test of the effects of the variable. Educational research journals regularly carry meta-analyses on topics ranging from the effects of computer use to the effects of phonemic awareness training. Meta-analysis, however, takes quantitative research an additional step away from design relevance. In combining results from a large number of experiments in the use of educational games, for instance, all the differences among games and in ways of using them are averaged out, leaving nothing to aid the person who would like to design a more effective educational game.

But Bereiter's hypothetical failure of the meta-analysis of the effects of heterogenous computer games does not justify the conclusion that all meta-analyses take "quantitative research an additional step away from design relevance." For example, my own meta-analysis (Hake, 1998a, 1998b, 2002a, 2002b) of pre/post test data for the MD and FCI as used in introductory Newtonian mechanics instruction, shown graphically in Figure 1, has proven to be of direct interest to course designers, even though the data were not (and could not have been!) obtained from a randomized controlled trial studies.
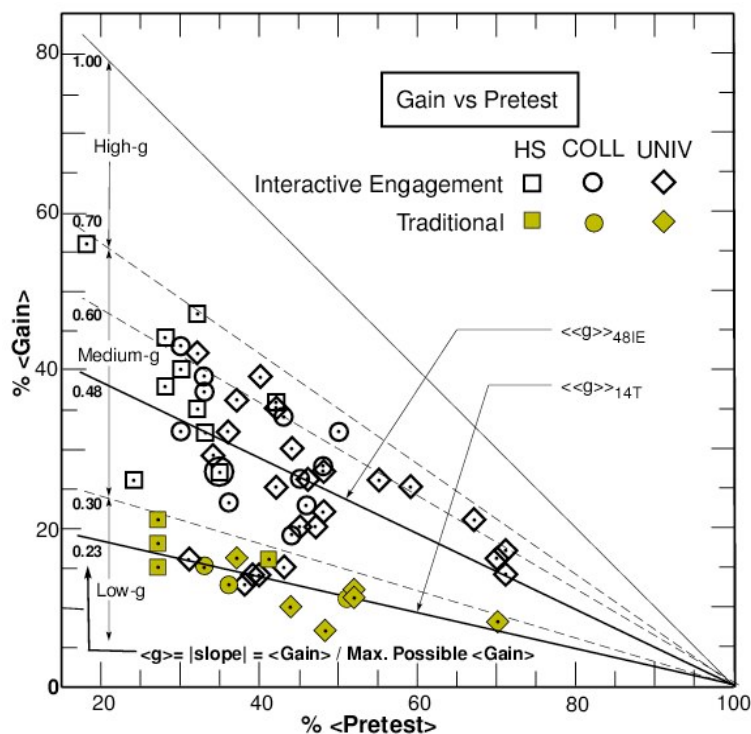
Institution: UNIV = university, COLL = college, HS = high school;
Method: IE = interactive engagement; T = traditional;
Symbols: <g> = course average *normalized* gain; max = maximum.

Figure 1. %<Gain> versus %<Pretest> scores on the conceptual *Mechanics Diagnostic* (MD) or *Force Concept Inventory* (FCI) tests for 62 courses enrolling a total N = 6542 students: 14 traditional (T) courses (n = 2084), which made little or no use of interactive engagement (IE) methods, and 48 IE courses (n = 4458), which made considerable use of IE methods. Slope lines for the average of the 14 T courses $<<g>>_{14T} = 0.23 \pm 0.04$ (SD) and the 48 IE courses $<<g>>_{48IE} = 0.48 \pm 0.14$ (SD) are shown. The negative-slope straight lines are lines of constant, normalized average gain <g> = <Gain>/maximum possible <Gain> =
(<%post > – <%pre >) / (100 – <%pre>). Thus, for example, 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. (The random guessing score is 20%.)

From Hake, 1998a. Copyright 1998 by American Institute of Physics. Reprinted with permission of the author.

In Figure 1: (a) the average normalized gain <g> is the actual gain (<%post> – <%pre>) divided by the maximum possible gain (100% – <%pre>) where the angle brackets indicate the class averages; (b) IE courses are defined *operationally* as those designed at least in part to promote conceptual understanding through the interactive engagement of students in heads-on (always) and hands-on (usually) activities that yield immediate feedback through discussion with peers and/or instructors; (c) T courses are defined *operationally* as those reported by instructors to make little or no use of IE methods, relying primarily on passive student lectures, recipe laboratories, and algorithmic problem examinations.

Regarding the average normalized gain <g>, ever since the work of Hovland, Lumsdaine, & Sheffield (1949/1965) it has been known by pre/post testing cognoscenti (which, up until about 1998, was 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* (Gery, 1972; Hake, 1998a, 1998b; Cohen, Cohen, Aiken, & West, 1999; Meltzer, 2002b). Justification for the use of <g> for the present data set resides in the fact that the correlation of <g> with <%pre> for the 62 survey courses is a very low +0.02. In contrast, the average posttest score <%post> and the average gain <G> are less suitable for comparing course effectiveness over diverse groups: the correlation of <%post> with <%pre> is +0.55, and the correlation of <G> with <%pre> is –0.49.

Regrettably, the insular (Hake, 2004b) psychology–education–psychometric community remains largely oblivious of normalized gain. Paraphrasing Lee Shulman, 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" (p. 24). Extrapolating the historical record, around 2030, yet another investigator will come up with the idea of <g> and attempt fruitlessly to interest the education community. Then, around 2060 yet another investigator . . . .

BUT WAIT! Could the average normalized gain <g> be a "flat tire" after all? Robert Mislevy (2005), while acknowledging the value of <g> in analyzing pre/post test gains Hake (1998a, 1998b), is uninterested in <g> because it, unlike Item Response Theory (Rudner, 2001), is not "grounded in the framework of probability-based reasoning."

But, in my opinion, Mislevy's objection must be balanced against the:
(1) *empirical* justification of <g> as an easy-to-use gauge of course effectiveness in hundreds of studies of classroom teaching in widely varying types of courses and institutions with widely varying types of instructors and student populations;

(2) evidently unsolved problem of how to employ IRT *to compare the effectiveness of courses* in which the initial average knowledge state of students is highly variable;

(3) difficulties that average faculty members might experience in using IRT to improve the effectiveness of their courses; and

(4) dearth of examples of the constructive employment of IRT in *higher* [as opposed to K-12 (Pellegrino, Chudowsky, & Glaser, 2001)] education research on classroom teaching.

Figure 1 serves as an existence proof that a two-standard deviation difference between average to posttest normalized gains <g> on the FCI and MD tests between interactive engagement and traditional courses *can* be obtained. I calculated a Cohen (1988) effect size "d" of 2.43 (Hake, 2002c), as indicated above, much higher than any found by Lipsey and Wilson (1993) in their meta-meta-analysis of psychological, educational, and behavioral treatments. Seven reasons for the "d disparity" between my survey and other social science research are given in Hake (2002a):

(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) was administered to both IE and T classes.

(4) The tests employed are widely recognized for their validity and consistent reliability, have been designed carefully 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 correlated significantly 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 should have included in the above list:

(8) Possible preferential selection of outstanding IE courses.

In regard to "8", I stated in Hake (1998a):

As in any scientific investigation, bias in the detector [due to the mode of data collection–voluntary contributions that tend to preselect results that are biased in favor of outstanding courses] can be put to good advantage if appropriate research objectives are established. We do *not* attempt to assess the average effectiveness of introductory mechanics courses. Instead, we seek to answer a question of considerable practical interest to physics teachers [and to physics education researchers]: ***can** the classroom use of IE methods increase the effectiveness of introductory mechanics courses well beyond that attained by traditional methods*?

For the 48, interactive engagement courses of Figure 1, the ranking in terms of number of IE courses using each of the more popular methods follows:

1. *Collaborative Peer Instruction* (Johnson, Johnson, & Smith, 1991; Heller, Keith, & Anderson, 1992; Slavin, 1995; Johnson, Johnson, & Stanne, 2000): 48 (all courses) {CA} – for the meaning of "CA" and similar abbreviations below within braces "{...}", see the paragraph after this list.

2. *Microcomputer-Based Laboratories* (Thornton & Sokoloff, 1990, 1998; Thornton, 1995): 35 courses {DT}.

3. *Concept Tests* (Mazur, 1997; Crouch & Mazur, 2001; Fagen, Crouch, & Mazur, 2002); Lorenzo, Crouch, & Mazur, 2006); Rosenberg, Lorenzo, & Mazur, 2006): 20 courses {DT}; such tests for physics, biology, and chemistry are available on the Web, along with a description of the peer instruction method, at the Mazur group's Galileo (2007) and Interactive Tool Kit (2007) sites.

4. *Modeling* (Halloun & Hestenes, 1987; Hestenes, 1987, 1992; Wells, Hestenes, & Swackhamer, 1995): 19 courses {DT + CA}; a description is on the Web at <http://modeling.la.asu.edu/>.

5. *Active Learning Problem Sets* or *Overview Case Studies* (Van Heuvelen, 1991a, 1991b, 1995): 17 courses {CA}; information on these materials is online at <http://www.physics.ohio-state.edu/~physedu/>.

6. Physics education research-based text (referenced in Hake, 1998b, Table II) or no text: 13 courses.

7. *Socratic Dialogue Inducing Laboratories* (Hake 1987, 1991, 1992, 2000, 2002d, 2007b; Tobias & Hake, 1988; Hake & Wakeland, 1997; & Tobias & Hake, 1988): 9 courses {DT + CA}; }; a description and laboratory manuals are on the web at <http://www.physics.indiana.edu/~sdi>.

The notations within the braces {. . .} follow Heller (1999) in associating loosely the methods with learning theories from cognitive science. Here, "DT" stands for "developmental theory," originating with Piaget (Inhelder & Piaget, 1958; Gardner 1985; Inhelder, deCaprona, & Cornu–Wells, 1987; Phillips & Soltis, 1998); "CA" stands for "cognitive apprenticeship" (Collins, Brown, & Newman, 1989; Brown, Collins, & Duguid, 1989). All the methods recognize the important role of social interactions in learning (Vygotsky, 1978; Lave & Wenger, 1991; Dewey, 1938/1997; Phillips & Soltis, 1998). It should be emphasized that the above rankings are by popularity within the survey and have no necessary connection with the effectiveness of the methods relative to one another. In fact, it is quite possible that some of the less popular methods used in some survey courses, as listed by Hake (1998b), could be more effective in terms of promoting students' understanding than any of the popular strategies noted above.

Average normalized gain differences between T and IE courses that are consistent with the work of Hake (1998a, 1998b, 2002a, 2002b) and Figure 1 have been reported by: Redish, Saul, & Steinberg, 1997; Saul, 1998; Francis, Adams, & Noonan, 1998; Heller, 1999; Redish & Steinberg, 1999; Redish, 1999; Beichner et al., 1999; Cummings, Marx, Thornton, & Kuhl, 1999; Novak, Patterson, Gavrin, & Christian, 1999; Bernhard, 2000; Crouch & Mazur, 2001; Johnson, 2001; Meltzer, 2002a, 2002b; Meltzer & Manivannan, 2002; Savinainen & Scott, 2002a, 2002b); Steinberg & Donnelly, 2002; Fagan, Crouch, & Mazur, 2002; Van Domelen & Van Heuvelen, 2002;, and Belcher, 2003; Dori & Belcher, 2004; Hoellwarth, Moelter, & Knight, 2005; Lorenzo, Crouch, & Mazur, 2006; & Rosenberg, Lorenzo, & Mazur, 2006.

This consistency of the results of many investigators in various institutions working with different student populations with the results of Hake (1998a, 1998b, 2002a, 2002b) constitutes the most important single warrant for the validity of conclusion in Hake (1998a) that: "The conceptual and problem-solving test results strongly suggest that the classroom use of IE methods can increase mechanics-course effectiveness well beyond that obtained in traditional practice" (p. 71). Such gradual buildup of an agreed-upon "community map" (Redish, 1999; Ziman, 2000) is characteristic of the progress of traditional science, but it seems to be consistently undervalued in educational research.

Furthermore, that interactive engagement courses would be more effective in enhancing conceptual understanding of counterintuitive Newtonian laws than traditional courses with their passive-student lectures, recipe laboratories, and algorithmic problem sets certainly would be expected from previous physics education research (McDermott & Redish, 1999), including the astute ethnographically-based insights of the late Arnold Arons (1997) - for a discussion see Hake (2004a).

More exploratory research is required to increase the effectiveness of IE courses. None that I surveyed (Hake, 1998a, 1998b) achieved an average normalized gain $<g>$ greater than 0.69, only fair on an absolute scale. Additional research is needed to ascertain the conditions under which IE courses can be most effective and to test IE courses in a wider variety of environments. In my opinion, what are needed are new meta-analyses of mechanics course results accruing over the past decade and in the future, using new and more secure tests than the FCI or *Force Motion Conceptual Evaluation* (FMCE) of Thornton and Sokoloff (1988).

In addition to enlarging the mechanics pre/post test data bank, PER groups have gone beyond the early survey work. For example, PER groups have shown that there may be significant differences in the effectiveness of various IE methods (Saul, 1998; Redish, 1999). In addition, there may be contributions to the average normalized gain $<g>$ from "hidden variables" such as averages over classes of gender, mathematics proficiency, spatial visualization ability, completion of high school physics courses, scientific reasoning skills, physics aptitude, personality type, motivation, socioeconomic level, ethnicity, intelligence quotient, scholastic aptitude test, and grade point average. One approach to this question is to investigate the relationship of *individual* student learning gains with such variables for *single courses* (Hake, 2002c; Meltzer, 2002b; Coletta & Phillips, 2005).

PER groups have developed diagnostic tests of students' cognitive and affective states before and after instruction in physics (including areas other than mechanics) and other disciplines (see, e.g., the listings at North Carolina State University, 2007and Field-Tested Learning Assessment Guide, 2007). They also have analyzed multiple-choice conceptual tests that go beyond classical test theory in which only the number of correct answers is considered in the scoring. These more advanced analyses can indicate incorrect models that students form during instruction in a single course or in a series of courses redesigned successively in attempts to improve their effectiveness. Such work has been reported by Bao and Redish (2001) for the FCI, by Thornton (1995) for the *Force Motion Conceptual Evaluation* (FMCE) (Thornton & Sokoloff, 1998), and may assist the study of transfer, i.e., the transfer of learning or capability from one area to another; see e.g., Chapter 3 "Learning and Transfer" of Bransford, Brown, & Cocking (2000), and also Hake (2004b). Finally, by the intensive study of how physics is learned (Redish, 1994, 1999, 2004), they have they have gone somewhat beyond the cognitive theories considered by Heller (1999).

IV. Conclusion

The physics education research discussed above qualifies as design-based research as gauged by the Kelly (2003a) criteria. In addition, PER has two attributes that seem to be missing in most DBR, as judged by the articles in Kelly (2003b): concern with undergraduate education—a major influence on the effectiveness of K–12 teaching—and rigorous assessment of the need for, and effects of, reform curricula. It is hoped that the provincialism of current education research (Hake, 2004b) which has hidden PER from DBR and DBR from PER can be reduced in order to make DBR a more interdisciplinary and synergistic effort.

V. Acknowledgments

References
Arons, A. B. (1973). Toward wider public understanding of science. *American Journal of Physics*, *41*, 769–782.

Arons, A. B. (1974). Toward wider public understanding of science: Addendum. *American Journal of Physics*, *42*, 157–58.

Arons, A.B. (1983). Achieving wider scientific literacy. *Daedalus*, Spring. This article forms the basis of Chapter 12 "Achieving Wider Scientific Literacy" in Arons (1997).

Arons, A. B. (1986). Conceptual difficulties in science. In M.R. Rice (Ed.)., *Undergraduate education in chemistry and physics*, *proceedings of the Chicago conferences on liberal education*, No. 1, (pp. 23–32). Chicago: University of Chicago.

Arons, A. B. (1997). *Teaching introductory physics*. New York: Wiley.

Bao, L., & Redish, E. F. (2001). Concentration analysis: A quantitative assessment of student states. *American Journal of Physics*, *69*, S45–S53. Retrieved 4 March 2007 from <http://www.physics.ohio-state.edu/~lbao/papers.htm>.

Beichner, R.J., Bernold, L., Burniston, E., Dail, P., Felder, R., Gastineau, J., Gjertsen, M., and Risley, J. (1999). Case study of the physics component of an integrated curriculum. *American Journal of Physics 67*(7), S16-S24. Retrieved 3 March 2007 from <http://www.ncsu.edu/PER/articles.htm>.

Beichner, R.J & Saul, J.M. 2004. Introduction to the SCALE-UP (Student-Centered Activities for Large Enrollment Undergraduate Programs) Project. In M. Vicentini and E.F. Redish, (Eds.), *Proceedings of the International School of Physics "Enrico Fermi" Course CLVI*. Varenna, Italy: IOS Press. Retrieved 3 March 2007 from <http://www.ncsu.edu/per/Articles/Varenna_SCALEUP_Paper.pdf>.

Belcher, J.W. (2003). Improving Student Understanding with TEAL [TEAL = Technology Enhanced Active Learning]. *The MIT Faculty Newsletter* XVI(2). Retrieved 3 March 2007 at <http://web.mit.edu/jbelcher/www/fnlEditedLinks.pdf>.

Bereiter, C. (2002). Design research for sustained innovation. *Cognitive Studies, Bulletin of the Japanese Cognitive Science Society*, *9*, 321–327. Retrieved on 4 March 2007 from <http://ikit.org/fulltext/2002Design_Research.pdf>.

Bernhard, J. (2000). Does active engagement curricula give long-lived conceptual understanding? *Proceedings of GIREP 2000*: *Physics Teacher Education Beyond 2000*, Barcelona. Retrieved 3 March 2007 from <http://staffwww.itn.liu.se/~jonbe/fou/didaktik/abstracts/girep2000_active.html>.

Bransford, J. D, Brown, A. L., & Cocking, R. R. (Eds.). 2000. *How people learn: brain, mind, experience, and school.* Washington, DC: National Academy Press. Retrieved 6 March 2007 from <http://books.nap.edu/books/0309070368/html/index.html>.

Coalition for Evidence-Based Policy (2003). *Identifying and Implementing Educational Practices Supported by Rigorous Evidence: A User Friendly Guide.* Retrieved 3 March 2007 from <http://www.ed.gov/rschstat/research/pubs/rigorousevid/rigorousevid.pdf>.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, P., Cohen, J., Aiken, L. S., & West, S. G. (1999). The problem of units and the circumstance for POMP. *Multivariate Behavioral Research 34*(3), 315-346. Retrieved on 6 March 2007 from <http://tinyurl.com/pcxt3>.

Coletta, V.P. and Phillips, J.A. (2005). Interpreting FCI scores: Normalized gain, preinstruction scores, & scientific reasoning ability. *American Journal of Physics*, *73*(12), 1172-1182. Retrieved on 6 March 2007 from <http://tinyurl.com/8eozb>.

Collins, A., Brown, J. S., & Newman, S. (1989). Cognitive apprenticeship: Teaching students the craft of reading, writing, and mathematics. In L. B. Resnick (Ed.), *Knowing, learning, and instruction: Essays in honor of Robert Glaser* (pp. 453–494). Hillsdale, NJ: Erlbaum.

Cronbach, L. J. (1975). Beyond the two disciplines of scientific psychology. American Psychologist, 30, 116-127. Abstract retrieved on 4 March 2007 from <http://tinyurl.com/3ce2w9>.

Cronbach, L. J., & Furby, L. (1970). How we should measure "change"—or should we? *Psychological Bulletin*, *74*, 68–80.

Crouch, C. H., & Mazur, E. (2001). Peer instruction: ten years of experience and results. *American Journal of Physics, 69*(9), 970-977. Retrieved 3 March 2007 from <http://tinyurl.com/sbys4>.

Cummings, K., Marx, J., Thornton, R., & Kuhl, D. (1999). Evaluating innovations in studio physics. *American Journal of Physics 67*(7), S38-S44.

Dewey, J. (1938/1997). *Experience and education.* New York: Free Press. (Original work published 1938).

Dori, Y. J. & Belcher, J. (2004). How does technology-enabled active learning affect undergraduate students' understanding of electromagnetism concepts? *The Journal of the Learning Sciences 14*(2). Retrieved 3 March 2007 from <http://tinyurl.com/cqoqt>.

Fagen, A. P., Crouch, C. H., & Mazur, E. (2002). Peer Instruction: Results from a Range of Classrooms. *Physics Teacher 40*, 206-209. Retrieved 3 March 2007 from <http://tinyurl.com/sbys4>.

Field-tested Learning Assessment Guide (FLAG). (2007). National Institute for Science Education, University of Wisconsin-Madison, College Level 1 Team, Retrieved on 4 March 2007 from <http://www.flaguide.org/>.

Francis, G. E., Adams, J. P., & Noonan, E. J. (1998). Do they stay fixed? *Physics Teacher 36*(8), 488- 491.

Galileo Project. (2007). Mazur Group, Harvard University. Retrieved on 4 March 2007 from <http://galileo.harvard.edu/>.

Gardner, H. (1985). *The mind's new science: A history of the cognitive revolution*. New York: Basic Books.

Gery, F. W. (1972). Does mathematics matter? In A. Welch (Ed.), *Research papers in economic education* (pp. 142–157). New York: Joint Council on Economic Education.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research*. Beverly Hills, CA: Sage.

Goodlad, J. I. (1990). *Teachers for our nation's schools*. San Francisco: Jossey–Bass.

Hake, R. R. (1987). Promoting student crossover to the Newtonian world. *American Journal of Physics*, *55*, 878–884. Retrieved 4 March 2007 as ref. 40 at <http://www.physics.indiana.edu/~hake>.

Hake, R. R. (1991). My conversion to the Arons-advocated method of science education. *Teaching Education*, *3*, 109–111. Retrieved 4 March 2007 as ref. 8 at <http://www.physics.indiana.edu/~hake>.

Hake, R. R. (1992). Socratic pedagogy in the introductory physics lab. *Physics Teacher*, *30*, 546–552. Retrieved 4 March 2007 as ref. 23 at <http://www.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. *American Journal of Physics*, *66*, 64–74. Retrieved on 4 March 2007 as reference 24 from <http://www.physics.indiana.edu/~hake>.

Hake, R. R. (1998b). Interactive-engagement methods in introductory mechanics courses. Retrieved on 4 March 2007 as reference 25 from <http://www.physics.indiana.edu/~hake>.

Hake, R. R. (2000, April). Towards paradigm peace in physics education research. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA. Retrieved on 3 March 2007 as ref. 3 at <http://www.physics.indiana.edu/~hake/>.

Hake, R. R. (2002a). Lessons from the physics education reform effort. *Ecology and Society*, *5*, online at <http://www.ecologyandsociety.org/vol5/iss2/art28/>.

Hake, R. R. (2002b, December). Assessment of physics teaching methods. Proceedings of the United Nations Educational, Scientific, and Cultural Organization's ASPEN (ASian Physics Education Network) workshop on active learning in physics, University of Peradeniya, Sri Lanka. Retrieved on 3 March 2007 as reference 29 from <http://www.physics.indiana.edu/~hake/>.

Hake, R. R. (2002c). Relationship of individual student normalized learning gains in mechanics with gender, high-school physics, and pretest scores on mathematics and spatial visualization. Retrieved on 3 March 2007 as reference 22 from <http://www.physics.indiana.edu/~hake>.

Hake, R. R. (2002d, December). Socratic dialogue inducing laboratory workshop. Proceedings of the United Nations Educational, Scientific, and Cultural Organization ASPEN (ASian Physics Education Network) workshop on active learning in physics, University of Peradeniya, Sri Lanka. Retrieved on 3 March 2007 as reference 28 from <http://www.physics.indiana.edu/~hake/>.

Hake, R. R. (2004a). The Arons advocated method. Retrieved on 3 March 2007 as reference 31 from <http://www.physics.indiana.edu/~hake>.

Hake, R. R. (2004b). Design-based research: A primer for physics education researchers. Retrieved on 3 March 2007 as reference 34 from <http://www.physics.indiana.edu/~hake>.

Hake, R.R. (2005, December). The Physics Education Reform Effort: A Possible Model for Higher Education, *National Teaching and Learning Forum 15(1)*. Retrieved on 4 March 2007 from <http://www.ntlf.com/FTPSite/issues/v15n1/physics.htm>.

Hake, R.R. (2006). Possible Palliatives for the Paralyzing Pre/Post Paranoia that Plagues Some PEP's, *Journal of MultiDisciplinary Evaluation*, Number 6, November. Retrieved 3 March 2007 from <http://evaluation.wmich.edu/jmde/JMDE_Num006.html>.

Hake, R.R. (2007a). Should We Measure Change? Yes! To appear as a chapter in Hake, R.R. (2007), *Evaluation of Teaching and Student Learning in Higher Education*, in preparation. Retrieved as ref. 43 on 3 March 2007 from <http://www.physics.indiana.edu/~hake>.

Hake, R. R. (2007b). Socratic dialogue inducing (SDI) lab website. Retrieved on 3 March 2007 from <http://www.physics.indiana.edu/~sdi>.

Hake R. R., & Wakeland, R. (1997). What's F? What's m? What's a?: A Non-Circular SDITST–Lab Treatment of Newton's Second Law. In J. Wilson (Ed.), *Conference on the introductory physics course*, (pp. 277–283). New York: Wiley.

Halloun, I. & Hestenes, D. (1985a). The initial knowledge state of college physics. *American Journal of Physics*, *53*, 1043-1055. Retrieved on 4 March 2007 from <http://modeling.asu.edu/R&E/Research.html>.

Halloun, I. & Hestenes, D. (1985b). Common sense concepts about motion. *American Journal of Physics, 53*, 1056-1065; Retrieved on 4 March 2007 from <http://modeling.asu.edu/R&E/Research.html>.

Halloun, I., & Hestenes, D. (1987). Modeling instruction in mechanics. *American Journal of Physics 55*, 455-462. Retrieved 3 March 2007 from <http://modeling.la.asu.edu/R&E/Research.html>.

Heller, K. J. (1999). Introductory physics reform in the traditional format: an intellectual Framework. *AIP Forum on Education Newsletter* (Summer): pages 7-9. Retrieved on 4 March 2004 from <http://units.aps.org/units/fed/newsletters/summer99.pdf> and as a talk at <http://groups.physics.umn.edu/physed/Talks/Heller1.pdf>.

Heller, P., Keith, R., & Anderson, S. (1992). Teaching problem solving through cooperative grouping. Part 1: group vs. individual problem solving. *American Journal of Physics, 60*(7), 627-636.

Heller, P., & Hollabaugh, M. (1992). Teaching problem solving through cooperative grouping. Part 2: designing problems and structuring groups. *American Journal of Physics, 60*(7), 637-644.

Hestenes, D. (1987). Toward a modeling theory of physics instruction. *American Journal of Physics, 55*, 440-454. Retrieved 3 March 2007 from <http://modeling.la.asu.edu/R&E/Research.html>.

Hestenes, D. (1992). Modeling games in the Newtonian world. *American Journal of Physics, 60*(8), 732-748. Retrieved 3 March 2007 from <http://modeling.la.asu.edu/R&E/Research.html>.

Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. *Physics Teacher 30,* 141-158; online (except for the test itself) at <http://modeling.asu.edu/R&E/Research.html>. The 1995 revision by Halloun, Hake, Mosca, & Hestenes is online (password protected) at the same URL and is available in English, Spanish, German, Malaysian, Chinese, Finnish, French, Turkish, Swedish, and Russian.

Hoellwarth, C., Moelter, M.J., & Knight, R.D. (2005). A direct comparison of conceptual learning and problem solving ability in traditional and studio style classrooms. *American Journal of Physics 73*(5), 459-463.

Hovland, C. I., Lumsdaine, A. A., & Sheffield, F. D. (1949/1965). A baseline for measurement of percentage change. In C. I. Hovland, A. A. Lumsdaine, and F. D. Sheffield (Eds.) *Experiments on mass communication* New York: Wiley. (Original work published 1949).

Inhelder, B., deCaprona, D., & Cornu–Wells, A. (1987). *Piaget today.* East Sussex, UK: Erlbaum.

Inhelder, B., & Piaget, J. (1958). *Growth of logical thinking from childhood to adolescence: An essay on the construction of formal operational structures* (A. Parsons & S. Milgram, Trans.). London: Basic Books.

Interactive Learning Toolkit. (2007). Mazur Group, Harvard University. Retrieved on 4 March 2007 at <http://www.deas.harvard.edu/ilt>.

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

Johnson, D. W., Johnson, R. T., & Smith, K. A. (1991). *Cooperative learning: increasing college faculty instructional productivity.* Washington, D. C.: George Washington University.

Johnson, D. W., Johnson, R. T., & Stanne, M. B. (2000). *Cooperative learning methods: a meta-analysis.* Retrieved 3 March 2007 from <http://www.co-operation.org/pages/cl-methods.html>

Kelly, A. E. (2003a). Research as design. *Educational Researcher*, *32*, 3–4. Retrieved 4 March 2007 from <http://www.aera.net/publications/?id=393>.

Kelly, A. E. (Ed.). (2003b). Theme Issue: The role of design in educational research. *Educational Researcher*, *32*(1). Retrieved 4 March 2007 from <http://www.aera.net/publications/?id=393>.

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation.* Cambridge, UK: Cambridge University Press.

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist*, *48*, 1181–1209. Retrieved on 4 March 2007 from <http://www.vanderbilt.edu/cerm/MWL_web_bib.htm#methods%20and%20issues>.

Lord, F. M. (1956). The measure of growth. *Educational and Psychological Measurement*, *16*, 42 –437.

Lord, F. M. (1958). Further problems in the measurement of growth. *Educational and Psychological Measurement*, *18*, 437–454.

Lorenzo, M., Crouch, C.H., & Mazur, E. (2006). Reducing the gender gap in the physics classroom. *American Journal of Physics 74*(2), 118-122.). Retrieved 3 March 2007 from <http://mazur-www.harvard.edu/publications.php?function=search&topic=9>.

Mazur, E. (1997). *Peer instruction: A user's manual*. Upper Saddle River, NJ: Prentice–Hall. Description retrieved on 4 March from <http://mazur-www.harvard.edu/education/pi_manual.php>.

McDermott, L. C., & Redish, E. F. (1999). RL-PER1: Resource letter on physics education research. *American Journal of Physics*, *67*, 755–767. Retrieved 4 March 2007 from <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. *American Journal of Physics, 70*(12), 1259-1268. Retrieved 3 March 2007 as article 7 from <http://www.physicseducation.net/articles/index.html>.

Meltzer, D.E. (2002b). Normalized Learning Gain: A Key Measure of Student Learning. Retrieved 3 March 2007 as article 7 (addendum) from <http://www.physicseducation.net/articles/index.html>.

Meltzer, D.E., & Manivannan, K. (2002). Transforming the lecture-hall environment: The fully interactive physics lecture. *American Journal of Physics 70*(6), 639-654. Retrieved 3 March 2007 as article 6 from <http://www.physicseducation.net/articles/index.html>.

Mislevy, R. 2006: (a) "On approaches to assessing change," and (b) "Clarification"; both retrieved on 5 March 2007 from <http://www.education.umd.edu/EDMS/mislevy/papers/Gain/>.

National Science Foundation. (1996). *Shaping the Future, Volume II: Perspectives on Undergraduate Education in Science, Mathematics, Engineering, and Technology,* Advisory Committee to the NSF Directorate for Education and Human Resources, M.D. George *(*Chair), NSF Report #98-128. Washington, DC: National Science Foundation. Retrieved on 5 March 2007 from <http://www.nsf.gov/pubs/1998/nsf98128/nsf98128.pdf>.

North Carolina State University. (2007). *Assessment instrument information page*. Physics Education R & D Group. Retrieved on 4 March 2007 from <http://www.ncsu.edu/per/TestInfo.html>.

Novak, G. M., Patterson, E.T., Gavrin, A. D., & Christian, W. (1999). *Just-in-time teaching: blending active learning with web technology*. New York, NY: Prentice Hall.

Pellegrino, J.W., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing What Students Know: The Science and Design of Educational Assessment*. Washington, DC: National Academy Press. Retrieved 5 March 2007 from <http://www.nap.edu/catalog/10019.html>.

Phillips, D. C., & Soltis, J. F. (1998). *Perspectives on learning* (3rd ed.). New York: Teachers College Press.

Pollock, S. (2004). No single cause: Learning gains, student attitudes, and the impacts of multiple effective reforms. In J. Marx, P. Heron, & S. Franklin, (Eds.), *2004 Physics Education Research Conference: AIP Conference Proceeding, vol. 790* (pp. 137-140). Retrieved 3 March 2007 from <http://tinyurl.com/9tfk4>.

Redish, E. F. (1994). Implications of cognitive studies for teaching physics. *American Journal of Physics*, *62*, 796–803. Retrieved 4 March 2007 from <http://www.physics.umd.edu/perg/papers/redish/cogsci.html>.

Redish, E. F. (1999). Millikan lecture 1998: Building a science of teaching physics. *American Journal of Physics*, *67*, 562–573. Retrieved 3 March 2007 from <http://www.physics.umd.edu/rgroups/ripe/perg/cpt.html>.

Redish, E. F. (2004). A theoretical framework for physics education research: Modeling student thinking. In M. Vicentini and E.F. Redish, (Eds.), *Proceedings of the International School of Physics "Enrico Fermi" Course CLVI*. Varenna, Italy: IOS Press. Retrieved 4 March 2007 at <http://www.physics.umd.edu/perg/papers/redish/Redish%20VarennaPre.pdf>.

Redish, E. F., Saul, J. M., & Steinberg, R. N. (1997). On the effectiveness of active-engagement microcomputer-based laboratories." *American Journal of Physics 65*(1), 45-54. Retrieved 3 March 2007 from <http://www.physics.umd.edu/rgroups/ripe/perg/cpt.html>.

Redish, E.F. & Steinberg, R.N. (1999). Teaching physics: figuring out what works. *Physics Today 52*(1), 24-30. Retrieved 3 March 2007 from <http://www.physics.umd.edu/rgroups/ripe/perg/cpt.html>.

Rogosa, D. R. (1995). Myth and methods: "Myths about longitudinal research" plus supplemental questions. In J. M. Gottman (Ed.), *The analysis of change* (pp. 3–66). Mahwah, NJ: Erlbaum. Examples retrieved 4 March 2007 from <http://www.stanford.edu/~rag/Myths/myths.html>.

Rogosa, D.R., Brandt, D., & Zimowski, M. 1982. A growth curve approach to the measurement of change. *Psychological Bulletin, 92*, 726-748.

Rogosa, D. R., & Willett, J. B. (1983). Demonstrating the reliability of the difference score in the measurement of change. *Journal of Educational Measurement*, *20*, 335–343.

Rogosa, D. R. & Willet, J. B. (1985). Understanding correlates of change by modeling individual differences in growth. *Psychometrika, 50*, 203-228. Abstract retrieved on 4 March 2007 from <http://www.springerlink.com/content/ru93345312046086/>.

Rosenberg, J., Lorenzo, M., & Mazur, E. (2006). Peer instruction: Making science engaging. In J. J. Mintzes and W. H. Leonard (Eds.) *Handbook of College Science Teaching* (pp. 77-85). Arlington, VA: NSTA Press. Retrieved on 5 March 2007 from <http://mazur-www.harvard.edu/sentFiles/Mazur_22532.pdf>.

Rudner, L.M. (2001). Item response theory. Retrieved 6 March 2007 from <http://edres.org/irt/>. This site offers a goldmine of information on IRT and its practical use, including free software.

Saul, J. M. (1998). *Beyond problem solving: Evaluating introductory physics courses through the hidden curriculum*. Unpublished doctoral dissertation, University of Maryland.

Savinainen A., & Scott, P. (2002a). The force concept inventory: A tool for monitoring student learning. *Physics Education 37*, 45-52. Retrieved 3 March 2007 at <http://kotisivu.dnainternet.net/savant/>.

Savinainen A., & Scott, P. (2002b). Using the force concept inventory to monitor student learning and to plan teaching. *Physics Education 37*, 53-58. Retrieved 3 March 2007 from <http://kotisivu.dnainternet.net/savant/>.

Shavelson, R. J., & Towne, L. (2002). *Scientific research in education*. Washington, DC: National Academy Press. Retrieved 3 March 2007 from <http://www.nap.edu/catalog/10236.html>.

Slavin, R. E. (1995). *Cooperative learning: theory, research, and practice*. Second edition. Boston, MA: Allyn & Bacon.

Steinberg, R. N., & Donnelly, K. (2002). PER-based reform at a multicultural institution. *Physics Teacher 40*(2), 108-114. Retrieved 3 March 2007 from <http://scitation.aip.org/dbt/dbt.jsp?KEY=PHTEAH&Volume=40&Issue=2>.

Stokes, D.E. (1997). *Pasteur's quadrant: Basic science and technological innovation*. Washington, DC: Brookings Institution Press.

Thornton, R. K. (1995). Conceptual dynamics: Changing student views of force and motion. In Bernardini, C. C. Tarsitani, & M. Vicentini (Eds.), *Thinking physics for teaching* (pp. 157–183). London: Plenum.

Thornton, R. K. (1995). Conceptual dynamics: Changing student views of force and motion. In C. Bernardini, C. Tarsitani, & M. Vicentini (Eds.), *Thinking physics for teaching* (pp. 157–183). London: Plenum.

Thornton, R. K., and Sokoloff, D. R. (1990). Learning motion concepts using real-time microcomputer-based laboratory tools. *American Journal of Physics 58*(9), 858-867.

Thornton, R. K., & Sokoloff, D. R. (1998). Assessing student learning of Newton's Laws: The force and motion conceptual evaluation and the evaluation of active learning laboratory and lecture curricula. *American Journal of Physics, 66*, 338–352.

Tobias, S., & Hake, R. R. (1988). Professors as physics students: what can they teach us? *American Journal of Physics, 56*(9), 786-794. Retrieved 4 March 2007 as ref. 41 at <http://www.physics.indiana.edu/~hake>.

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

Van Heuvelen, A. (1991a*).* Learning to think like a physicist: a review of research-based instructional strategies. *American Journal of Physics, 59*(10), 891-897.

Van Heuvelen, A. (1991b*).* Overview, case study physics. *American Journal of Physics*, *59*(10), 898-907.

Van Heuvelen, A. (1995). Experiment problems for mechanics. *Physics Teacher 33*, 176-180.

Vlastos, G. (1991). *Socrates, Ironist and Moral Philosopher*. Cambridge, UK: Cambridge University Press. Information retrieved on 6 March 2007 from <http://www.cambridge.org/us/catalogue/catalogue.asp?isbn=9780521314503>.

Vlastos, G. (1994). *Socratic Studies*. Myles Burnyeat (Ed.). Cambridge, UK: Cambridge University Press. Information retrieved on 6 March 2007 from <http://www.cambridge.org/catalogue/catalogue.asp?isbn=9780521447355>.

Vygotsky, L. S. (1978). M. Cole, V. John-Steiner, S. Scribner, & E. Souberman, (Eds.) *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.

Wells, M., Hestenes, D., & Swackhamer, G. (1995). A modeling method for high school physics instruction. *American Journal of Physics, 63*(7), 606-619. Retrieved 3 March 2007 from at <http://modeling.la.asu.edu/R&E/Research.html>.

Wilson, M.R. & Bertenthal, M.W. (Eds.). (2005). *Systems for State Science Assessment*. Washington, DC: National Academy Press. Retrieved on 5 March 2005 from <http://www.nap.edu/catalog.php?record_id=11312>.

Wittmann, W. W. (1997). *The reliability of change scores: Many misinterpretations of Lord and Cronbach by many others; revisiting some basics for longitudinal research.* Retrieved 4 March 2007 from
<http://www.psychologie.uni-mannheim.de/psycho2/psycho2.en.php3?language=en/> / "Publications" / "Papers and preprints" where "/" means "click on."

Ziman, J. (2000). *Real science: What it is, and what it means.* Cambridge, UK: Cambridge University Press.

Zimmerman, D. W., & Williams, R. H. (1982). Gain scores in research can be highly reliable. *Journal of Educational Measurement*, *19*, 149–154. Abstract retrieved 4 March 2007 from <http://mypage.direct.ca/z/zimmerma/earlier.html> .

Zimmerman, D. W. (1997). A geometric interpretation of the validity and reliability of difference Scores. *British Journal of Mathematical and Statistical Psychology, 50*, 73-80.

Zumbo, B. D. (1999). The simple difference score as an inherently poor measure of change: Some reality, much mythology. In B. Thompson, (Ed.), *Advances in Social Science Methodology*, Volume 5 (pp. 269-304). Greenwich, CT: JAI Press. Retrieved on 4 March 2007 from <http://tinyurl.com/kuf3t>.