Some Physics Education Research is Design-Based Research

In this chapter I argue that some physics education research (PER) is design-based research (DBR). 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 preservice teachers in particular, and, as a net result, the education of the general population.

In their resource letter on physics education research, McDermott and 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 and 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:

- Interactive engagement of students who are induced to think constructively about simple Newtonian experiments that produce conflict with their commonsense understandings.
- 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.
- Considerable interaction between students and instructors and thus a degree of individualized instruction.
- Extensive use of multiple representations (verbal, written, pictorial, diagrammatic, graphical, and mathematical) to model physical systems.
- 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).
- Cooperative group effort and peer discussions.
- Repeated exposure to the coherent Newtonian explanation in many different contexts.

(1997), and Tobias and 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 (Hake, 1991, 1992; Tobias & Hake, 1988), confirmed (Hake, 1998a, 1998b, 2002a, 2002b, 2005, 2006, in preparation), and disseminated (Hake, 2000, 2002b, 2002d, 2007) 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 video-tape 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, 2007; Hake & Wakeland, 1997; Tobias & Hake, 1988) 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; Hake & Wakeland, 1997; Tobias & Hake, 1988). 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, in preparation).

Considering the above two paragraphs, I submit that some PER qualifies as design-based research as characterized by Kelly (2003a). 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 National Science Foundation’s report Shaping the Future hit the nail squarely on the head (my italics):

. . . Science, mathematics, engineering, and technology (SME&T) [programs] 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.

(1996: 35)

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: xi–xii):

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.

Pre/Post Testing in Physics Education Research

The pre/post testing movement in PER was initiated by the landmark work of Halloun and 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 Mechanics Diagnostic, 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 Mechanics Diagnostic 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 et al., 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., Hake, 1998a, 1998b; Halloun & Hestenes, 1985b).

Halloun and Hestenes (1985a, 1985b) then used the Mechanics Diagnostic 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:

- 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.
- 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.

(Halloun & Hestenes, 1985a: 1048)
Can multiple choice tests gauge higher-level cognitive outcomes such as the conceptual understanding of Newtonian mechanics? Wilson and Bertenthal think so, writing:

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

(2005: 94)

The Halloun and 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 Mechanics Diagnostic/Force Concept Inventory-induced changes in introductory physics courses at pacesetting Harvard University and the Massachusetts Institute of Technology. Harvard University’s Mazur wrote:

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.

(1997: 4)

In Table I of Crouch and Mazur (2001: 972), 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 were made in the implementation of peer instruction.

Belcher, describing his institute’s introductory physics course transition from traditional to interactive-engagement, wrote:
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.

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: 2), in its *Identifying and implementing educational practices supported by rigorous evidence: A user friendly guide* states:

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 \(<\%_{\text{post}}\) – \(<\%_{\text{pre}}\) 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: 1) dubious pronouncement 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, in preparation), 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 et al. (1982), Zimmerman and Williams (1982), Rogosa and Willett (1983, 1985), Rogosa (1995), Wittmann (1997), Zimmerman (1997), and Zumbo (1999). All this more recent work should serve as an antidote for the emotional, pre/post testing paranoia that grips many educational researchers.

Meta-analysis of Pre/Post Learning Gains

A reservation regarding meta-analysis has been raised by DBR pioneer Carl Bereiter. In Design research for sustained innovation, he 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.

(2002: 10)

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 Mechanics Diagnostic and Force Concept Inventory as used in introductory Newtonian mechanics instruction, shown graphically in Figure 26.1, has proven to be of direct interest to course designers, even though the data were not (and could not have been) obtained from randomized controlled trial studies.

In Figure 26.1, %<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: (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, and (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.

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 \(<g> = \frac{<\text{Gain}>}{\text{Max. Possible}<\text{Gain}>}\). Thus, for example, if a class averaged 40 percent on the pretest and 60 percent on the post-test, then the class-average normalized gain \(<g> = \frac{(60\% - 40\%)}{100\% - 40\%} = 20\%/60\% = 0.33. (The random guessing score is 20 percent.)

Regarding the average normalized gain \(<g>\), ever since the work of Hovland et al. (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 post-test (Cohen et al., 1999; Gery, 1972; Hake, 1998a, 1998b; Meltzer, 2002b). Justification for the use of \(<g>\) for the present data set resides in the fact that the correlation of \(<g>\) with \(<\%_{\text{pre}}\>\) for the 62 survey courses is a very low +0.02. In contrast, the average post-test score \(<\%_{\text{post}}\>\) and the average gain \(<g>\) are less suitable for comparing course effectiveness.

Figure 26.1 \(<%\text{Gain}>\) Versus \(%\text{Pretest}>\) Scores on the Conceptual Mechanics Diagnostic (MD) or Force Concept Inventory (FCI) Tests for 62 Courses.

over diverse groups: the correlation of \(<%_{\text{post}}\) with \(<%_{\text{pre}}\) is +0.55, and the correlation of \(<G>\) with \(<%_{\text{pre}}\) is −0.49.

Regrettably, the insular (Hake, 2004b) psychology-education-psychometric community remains largely oblivious of normalized gain. Paraphrasing Shulman, as quoted by the late Arons (1986: 24): “it seems that in education, the wheel (more usually the flat tire) must be reinvented every few decades.”

Could the average normalized gain \(<g>\) be a “flat tire” after all? Mislevy (2006), 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:

- **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.**
- **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.**
- **Difficulties that average faculty members might experience in using IRT to improve the effectiveness of their courses.**
- **Dearth of examples of the constructive employment of IRT in higher (as opposed to K–12 [Pellegrino et al., 2001]) education research on classroom teaching.**

Figure 26.1 serves as an existence proof that a two-standard deviation difference between average to post-test normalized gains \(<g>\) on the Force Concept Inventory and Mechanics Diagnostic 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 Mechanics Diagnostic or Force Concept Inventory) 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 26.1, the ranking in terms of number of IE courses using each of the more popular methods follows:

1. **Collaborative Peer Instruction:** 48 (all courses) [CA] (Heller et al., 1992; Johnson et al., 1991, 2000; Slavin, 1995).
3. **Concept Tests:** 20 courses [DT] (Crouch & Mazur, 2001; Fagen et al., 2002; Lorenzo et al., 2006; Mazur, 1997; Rosenberg et al., 2006).
5. **Active Learning Problem Sets or Overview Case Studies:** 17 courses [CA] (Van Heuvelen, 1991a, 1991b, 1995).
6. **Physics Education Research-based Text or No Text:** 13 courses (referenced in Hake, 1998b, Table II).

Average normalized gain differences between T and IE courses that are consistent with the work of Hake (1998a, 1998b, 2002a, 2002a) and Figure 26.1 have been reported (Beichner et al., 1999; Belcher, 2003; Bernhard, 2000; Crouch & Mazur, 2001; Cummings et al., 1999; Dori & Belcher, 2003; Fagen et al., 2002; Francis et al., 1998; Heller, 1999; Hoellwarth et al., 2005; Johnson, 2001; Lorenzo et al., 2006; Meltzer, 2002a, 2002b; Meltzer & Manivannan, 2002; Novak et al., 1999; Redish, 1999; Redish & Steinberg, 1999; Redish et al., 1997; Rosenberg et al., 2006; Saul, 1998; Savinainen & Scott, 2002a, 2002b; Steinberg & Donnelly, 2002; Van Domelen & Van Heuvelen, 2002).

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: 71) 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.” Such gradual build-up 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 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 Force Concept Inventory or Force Motion Conceptual Evaluation (FMCE) of Thornton and Sokoloff (1998).

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 (Redish, 1999; Saul, 1998). 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 (Coletta & Phillips, 2005; Hake, 2002c; Melzter, 2002b).

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 (e.g., the listings at North Carolina State University, 2007 and Field-Tested Learning Assessment Guide, National Institute for Science Education, 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 Force Concept Inventory, 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 Bransford et al. (2000, Chapter 3), and also Hake (2004b). Finally, by the intensive study of how physics is learned (Redish, 1994, 1999, 2004), they have gone somewhat beyond the cognitive theories considered by Heller (1999).

**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.
Acknowledgments

This work received partial support from National Science Foundation Grant DUE/MDR9253965. I thank Anthony E. Kelly for suggesting that I write this article and for his valuable comments.

Notes

1 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 (Gardner 1985; Inhelder & Piaget, 1958; Inhelder et al., 1987; Phillips & Soltis, 1998); “CA” stands for “cognitive apprenticeship” (Brown et al., 1989; Collins et al., 1989). All the methods recognize the important role of social interactions in learning (Dewey, 1938/1997; Lave & Wenger, 1991; Phillips & Soltis, 1998; Vygotsky, 1978). It should be emphasized that the 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.

2 Tests available for physics, biology, and chemistry (Galileo, 2007; ILT-BQ Consortium, 2006).

3 A description is on the web at http://modeling.la.asu.edu/.

4 Information on these materials is online at http://www.physics.ohio-state.edu/~physedu/.

5 A description and laboratory manuals are on the web at http://www.physics.indiana.edu/~sdi.

References


