Suggestions for Administering and Reporting Pre/Post Diagnostic Tests * ‡

Richard R. Hake, Emeritus Professor of Physics
Indiana University <rrhake@earthlink.net>

I. Introduction

In a recent review[1], I give fourteen somewhat subjective lessons from my own interpretation of the physics-education reform effort. Lesson #6 (slightly edited) is:

"High quality standardized tests of the cognitive and affective impact of courses are essential for gauging the relative effectiveness of non-traditional educational methods . . . . so great is the inertia of the educational establishment . . . . that three decades of physics-education research demonstrating the futility of the passive-student lecture for enhancing the conceptual understanding of average students in introductory courses were ignored until high-quality Diagnostic Tests that could easily be administered to thousands of students became available. These multiple-choice tests are yielding increasingly convincing evidence that interactive-engagement methods enhance conceptual understanding and problem solving abilities far more than do traditional methods . . . . . As far as I know, disciplines other than physics and astronomy . . . . have yet to develop any such tests and therefore cannot effectively gauge either the need for or the efficacy of their reform efforts. In my opinion, all disciplines should consider the construction of high-quality standardized tests of essential introductory course concepts."

It should be understood, of course, that such quantitative data has its strengths and weaknesses, and must be guided and supplemented by complementary[2] qualitative research such as Socratic dialoguing with students,[3-5] interviews,[4] case studies,[6,7] and surveys.[6,7]

---

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

‡ The reference is: "Suggestions for Administering and Reporting Pre/Post Diagnostic Tests," 5/18/01, unpublished; online as ref. 14 at < http://physics.indiana.edu/~hake/ >
[TestingSuggestions.pdf, 5/18/01, 44K].

With the hope that more teachers will begin to research the effectiveness of their methods in promoting student learning,[8-17] and that new and better *Diagnostic Tests* will be developed in physics[16] and in other disciplines, I list below some testing and reporting suggestions that might assist teachers in using *formative* diagnostic testing (no relation to mandated high-stakes *summative* testing). These suggestions reflect the hard lessons I have learned in the pre/post testing of 1263 pre-med introductory-physics-course students at Indiana University over the years 1986 – 1995,[3-5] and the compilation of a 6542 student (62-course) survey[6,7] of pre/post test results. Although they are undoubtedly *not* appropriate for all classroom situations, they may at least indicate some of the problems that should be anticipated.

II . Administration of *Diagnostic Tests* (DT's) (An asterisk * preceding a suggestion indicates the suggestion is intended, at least in part, to promote confidentiality of the test.[18])

*A. When administering DT's to students, refer to the tests by home-made generic titles rather than the specific titles designated by the authors (e.g., "Mechanics Familiarity Survey" rather than "Force Concept Inventory"). The specific titles designated by the authors can provide the key to accessing the tests in the literature and on the web. For example a search for "Force Concept Inventory"[19] with the powerful Google search engine < http://www.google.com/ > netted 195,000 hits in 0.46 seconds.

B. To facilitate meaningful pre/post comparison, maximum time intervals ΔT given to students to complete the pretest and the posttest should be same. To facilitate more meaningful meta-analysis (an analysis of pre/post test results from many different courses – see e.g., refs. 1, 6, 7) an appropriate maximum time interval ΔT should be specified by the test designers and that interval should be rigidly enforced by all examiners. Unfortunately, such is not the case for administration of the Force Concept Inventory[19] (FCI) and/or its very similar precursor the Mechanics Diagnostic[21] (MD) test where ΔT´s of 25 minutes[22] for the FCI and 60 minutes[3] for the MD have been reported. Although it is commonly assumed that variation in ΔT for $\Delta T > 25$ minutes has little effect on the average of student scores on either the FCI pretest or posttest, there is, as far as I know, no evidence to support this assumption. (My guess is that the assumption is more likely to be true for the no-credit pretest where students usually finish in 30 minutes or less, and for the posttest *if* no course credit is given.)

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

(1) Proper incentive for students to exert effort on the test (see, e.g., the italicized sentence in Sec. IID below).

(2) Analysis of “matched” pre/post test data,[23] i.e., obtaining the average class pretest score by counting only the scores of those students who took the posttest, and thus allowing a more rigorous calculation of the class average normalized gain[6]:

$$\langle g \rangle = \langle \%G \rangle \ / \ \langle \%G \rangle_{max}$$
$$\langle g \rangle = (\langle \%post \rangle - \langle \%pre \rangle) / (100 - \langle \%pre \rangle) \ . \ . \ . \ . \ . \ . \ . \ . \ . \ . \ . \ . \ . \ . (1)$$

where G is the actual gain, and <%post> and <%pre> are the final (post) and initial (pre) class averages, and the angle brackets “< . . . . >” indicate an average over the students taking the tests.

It should be kept in mind that, for the FCI/MD, the experimental justification for using <g> as a comparative measure of course effectiveness over diverse student populations with widely varying average pretest scores is that in the 62-course survey of refs. 6 & 7, the correlation of the average normalized gain <g> with the average course pretest (<%pret>) was a very low +0.02. This low correlation is a reflection of the low correlation of single student g’s with their pretest scores within any given class. Unless, for a given DT, similar low correlations can be found from meta-analytic results over different courses, or at least for single student results within different courses, then the use of that DT for intercomparison of diverse groups is somewhat problematic.

Even for the FCI/MD, it remains an open question as to whether or not “hidden variables”[24] (e.g., average math proficiency, spatial visualization ability, motivation, socio-economic level, gender, ethnicity, scientific reasoning skills, IQ, SAT, GRE) of a class could have a significant effect on <g>.[24,25] I think that it is extremely unlikely that hidden variable effects could account for an appreciable fraction of the nearly two-standard-deviation difference in the average of <g>’s for the 48 Interactive Engagement and 14 Traditional courses found in the survey of ref. 6, although such effects could be significant in the comparison of only a few courses.[24]

(3) Knowledge of the normalized gain g for each single student in the class, thus allowing a calculation of the average of the single-student gains:

$$g_{ave} = (1/N)\ \Sigma_i\ g_i = (1/N)\ \Sigma_i\ [(post_i - pre_i)/\ (100\% - pre_i)],\ .\ .\ .\ .\ .\ .\ .\ (2)$$

where N is the number of students taking both the pre- and post-tests and the summation is over all N students.

(4) Analyses of single student normalized or actual gains in terms of single-student characteristics or performance on other tests.[24,25]

(5) Calculation of the correlation of individual student g's with their pretest scores.[26] (See Sec. IIID below.)

In practice, for the FCI, when the number of students taking the test is greater than about 20, g_{ave} is usually within 5% of <g>.[27] As explained in footnote #46 of ref. 6, the definitions of g_{ave} (Eq. (2) and <g> (Eq. 1) imply that [g_{ave} – <g>] is proportional to the g_j–weighted average of the deviations (%pre$_j$ – <%pre$_j$ >). Since the average of (%pre$_j$ – <%pre$_j$ >) is zero, a low [g(ave) - <g>] implies a low correlation between g_j and %pre$_j$ for individual students, just as there is a low correlation between <g> and <%pre> for survey courses, as discussed above in Sec. IIC(2).

*D. If possible, give the pretest on the *first* day of class. Take great care that all question sheets and answer sheets are returned and verify such return by counting those given out and those returned. Because the first day of class may be somewhat chaotic, it is advisable to have extra staff on hand to serve as monitors and help pass out and collect the question and answer sheets. If pre/post testing has been used before in the class, then monitors should be on the watch for non-enrolled people whose mission is to make off with copies of the DT. For the few inevitable stragglers who enter the class after the first day, arrange a second pretest at some mutually agreeable time, but preferably during the first week of class. *In order to promote serious effort on the test by students, explain that although their scores on the pretest will* ***no****t count towards the course grade, their scores will be confidentially returned to them and will assist both themselves and their instructors to know the degree and type of effort required for them to understand mechanics*. Of course, no mention should be made that the same or similar test will be given as a posttest. For conventional U.S. university and high-school classes, where the class period may be 40 – 60 minutes in length, allow students who have competed the test before the allotted time to place their answer and question sheets in a *monitored* collection box and then leave the classroom. The collection-box monitor should carefully check to see that each student submits both a completely-filled-in-and-signed answer sheet and a question sheet.

*E. Give the posttest *unannounced* near the final day of classes, and preferably as part of the final exam with significant course credit given for posttest performance. It would seem that giving course credit would motivate students to take the posttest more seriously and thereby demonstrate more adequately their understanding, especially if time devoted to the posttest subtracts from time spent on the rest of the final exam – see the next-to-last sentence of this paragraph. [In this connection, Henderson et al.[28] have reported data (based on N = 1818 introductory calculus-based physics students at the University of Minnesota during 1997 – 1999) indicating that there is no meaningful difference between FCI scores on for-credit and not-for-credit posttests, *providing that in the latter case exams that display an "obvious lack of seriousness" (about 2.4%) are discarded*. No details on how the time allotted to the posttest portion of the final exam was structured are given.] Again, take great care that all question sheets and answer sheets are returned and verify such return by counting those given out and those returned. In standard U.S. university classes the final exam period will normally be two or three hours. One strategy is to pass out the entire final exam [DT + other test(s)] at the start of the period. Tell students to start working *first* on the DT, that the DT will be collected at the end of the standard allotted time interval ΔT, and that they may start working on the other parts of the final exam before the DT's are collected if they so choose. DT collecters should carefully check to see that each student submits both a completely-filled-in-and-signed answer sheet and a question sheet.

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

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

*H. Avoid, if at all possible, in-class discussion of questions identical or almost identical to DT questions (an example of "teaching to the test"). Because in most disciplines there are probably many sources of good conceptual questions and problems (for introductory physics see ref. 7, footnote #66), there is little need to draw on the standardized tests for questions or problems to be used for ordinary class discussion and testing.

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

*J. Do not make DT questions or problems available on the web unless they are password protected such that only authorized instructors may gain access. Do not publish DT's in the open literature, as has been the common practice.[19-21,29-31] As indicated in ref. 1, carefully constructed DT's are national assets whose confidentiality should be as well protected as the MCAT (Medical College Admission Test).

*K. Because of the almost unavoidable slow diffusion of test questions and answers to student files, replace each DT at approximately 5- or 10-year intervals, such that it can be meaningfully calibrated against the previous test(s). (So far this has NOT been done for the now overused 1992/95 version of the FCI; in my opinion, as time goes on, research results based on the 1992/95 FCI[19] will become more and more doubtful.)

III . Reporting[8-17] of *Diagnostic Tests* (DT's)

A. Report at least:

(1) The class average <%pre> *with its standard deviation* (sd),

(2) the class average <%post> *with its sd*, and

(3) the class average normalized gain <g> (Eq. 1 above).

Unless standard deviations are reported, the effect size,[32.33] and errors[26,27] in <g> cannot be ascertained. As a statistic for comparison of courses and for meta-analyses, the class-average <g> is better, in my opinion, than g_{ave} because the latter: (a) must exclude students who score 100% on the pretest and thus achieve an infinite or indeterminate g, and (b) may introduce skewing due to outliers who score near 100% on the pretest and less on the posttest such their <g>'s are large and negative. The selective removal of outliers so as to avoid "b" by various different investigators with different outlier criteria will lead to a degree of uncertainty in comparing normalized gains of different courses.

B. Report if possible:

(1) Cohen's "effect size,"[32.33]

$$d = (\langle\%post\rangle - \langle\%pre\rangle) / [(sd_{pre}^{2} + sd_{post}^{2})/2]^{0.5} \quad \ldots\ldots\ldots\ldots\ldots\ldots (3)$$

where the denominator is the *root mean square* of standard deviations for the pre- and post-tests, sometimes called the "pooled standard deviation." Unless sd_{pre} and sd_{post} differ markedly, the pooled standard deviation will not differ greatly from the arithmetic mean of sd_{pre} and sd_{post}. Cohen's <d> is reported for the data of refs. 34-36.

(2) The *Kuder-Richardson reliability coefficients* KR-20[37,38] (or equivalent - for tests in which the answer is either correct or incorrect as in the FCI - *Cronbach's alpha*[38,39]) for the pre- and post-tests. These reliability coefficients have been reported in refs. 3, 4, 7, 21, 28, 29, 31 & 35.

(3) The estimated systematic and random errors.[26]

(4) The correlation of individual student g's and pretest scores.[26,27] A significant *positive* correlation would suggest that the instruction tends to favor students who have *more* prior knowledge of the subject as judged by the pretest score ("Matthew effect" [40]); a significant *negative* correlation would suggest that the instruction favors students who have *less* prior knowledge of the subject as judged by the pretest score ("anti–Matthew effect"); and an insignificant correlation would suggest that the instruction is at about the right level for students who have an average prior knowledge of the subject as judged by the pretest score.

C. As a guide to other information that might be useful in a report of pre/post test results and complementary qualitative research, consider refs. 3 – 7, 19a, 21, 22, 29, 30, 35, 36, 41, and the questions in the *Mechanics Test Data Survey Form.*[42]

References and Footnotes

1. R.R. Hake, "Lessons from the Physics Education Reform Effort," submitted on 3/28/01 to *Conservation Ecology* < http://www.consecol.org/Journal/ >, a "peer-reviewed journal of integrative science and fundamental policy research." Online as ref. 10 at < http://www.physics.indiana.edu/~hake > [ConEc-Hake-O32601a.pdf, 3/26/01, 172K] (179 references, 98 hot-linked URL's).

2. R.R. Hake, "Towards paradigm peace in physics-education research," presented at the annual meeting of the American Educational Research Association, New Orleans, April 24-28, 2000; online at < http://www.physics.indiana.edu/~hake/ >.

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

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

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

6. R.R. Hake, "Interactive-engagement vs traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses," *Am. J. Phys.* **66**, 64-74 (1998); online at < http://www.physics.indiana.edu/~sdi/ >. See also ref. 7.

7. R.R. Hake, "Interactive-engagement methods in introductory mechanics courses," online at < http://www.physics.indiana.edu/~sdi/ > and submitted on 6/19/98 to the *Physics Education Research Supplement to AJP* (PERS).

8. Carnegie Academy for the Scholarship of Teaching and Learning < http://www.carnegiefoundation.org/CASTL/ >“. . . . represents a major initiative of *The Carnegie Foundation for the Advancement of Teaching*. Launched in 1998, the program builds on a conception of teaching as scholarly work proposed in the l990 report, *Scholarship Reconsidered* by former Carnegie Foundation President Ernest Boyer (ref. 9), and on the l997 follow-up publication, *Scholarship Assessed* by Glassick et al. (ref. 10):

(a) Boyer Commission, *Reinventing undergraduate education: A blueprint for America’ s research universities* (The Boyer Commission on Educating Undergraduates the Research University, 1998) < http://notes.cc.sunysb.edu/Pres/boyer.nsf >:“

(b) Journal of Scholarship of Teaching and Learning < http://www.iusb.edu/%7Ejosotl/ >.

(c) Programs for K-12 < http://www.carnegiefoundation.org/CASTL/k-12/index.htm >:

(d) Programs for Higher Education < http://www.carnegiefoundation.org/CASTL/highered/index.htm >.

9. E.L. Boyer, *Scholarship reconsidered: priorities for the professoriate*. (Carnegie Foundation for the Advancement of Teaching, 1990).

10. C.E. Glassick, M.T. Huber, and G.I. Maeroff, *Scholarship Assessed: Evaluation of the Professoriate* (Carnegie Foundation for the Advancement of Teaching, 1997).

11. N. Cantor (Provost, University of Michigan), “Reinvention: Why now? Why Us?” (A symposium on the The Boyer Commission Report: A Second Anniversary Retrospective, State University of New York at Stony Brook.) Online at , < http://www.umich.edu/~provost/speeches/boyer.html >.

12. C. Nelson, “Bibliography: How To Find Out More About College Teaching And Its Scholarship: A Not Too Brief, Very Selective Hyperlinked List - College Pedagogy IS A Major Area Of Scholarship!” online at < http://php.indiana.edu/~nelson1/TCHNGBKS.html >.

13. R.B. Barr, & J. Tagg, “From teaching to learning – a new paradigm for undergraduate education,” *Change*, Nov./Dec., 13-25 (1995)..

14. National Institute for Science Education (NISE) < http://www.wcer.wisc.edu/NISE/ >, "Field-tested Learning Assessment Guide (FLAG): For Science, Math, and Engineering Instructors" < http://www.wcer.wisc.edu/nise/cl1/flag/flaghome.asp >, esp. "Introduction to CAT's" (Classroom Assessment Techniques) at < http://www.wcer.wisc.edu/nise/cl1/flag/cat/catframe.asp >.

15. For a listing of physics diagnostic tests and their locations see the *Assessment Instrument Information Page* < http://www.ncsu.edu/per/TestInfo.html >, Physics Education R & D Group, North Carolina State University.

16. D.C. Phillips & N.C. Burbules, *Postpositivism and educational research* (Rowman & Littlefield, 2000), esp. Chap. 4: "Can, and should, educational inquiry be scientific?"

17. E.F. Redish, "Millikan lecture 1998: Building a Science of Teaching Physics," *Am. J. Phys.* **67**(7), 562-573 (1999); L.C. McDermott & E.F. Redish, "RL-PER1: Resource letter on physics education research," *Am. J. Phys.* **67**(9), 755-767 (1999.); both these papers are online at < http://www.physics.umd.edu/rgroups/ripe/perg/cpt.html >.

18. In ref. 1, I wrote: "The lengthy and arduous process of constructing valid and reliable multiple choice tests has been discussed by Halloun & Hestenes (1985a), Hestenes et al. (1992), Beichner (1994), Aubrecht (1991), and McKeachie (1999) . . . . (for references see ref. 1) . . . . In my opinion such hard-won Diagnostic Tests that cover important parts of common introductory courses are national assets whose confidentiality should be as well protected as the MCAT (Medical College Admission Test). Otherwise the test questions may migrate to student files and thereby undermine education research that relies upon the validity of such tests." For an earlier discussion of confidentiality and lack thereof for the Force Concept Inventory (FCI) see ref. 6, footnote #48. That these concerns are well founded is indicated in Sec. IIA of the present work: a Google search for "Force Concept Inventory" netted 195,000 hits, suggesting that the FCI may now be well-known to many student web-surfers.

19. (a) D. Hestenes, M. Wells, and G. Swackhamer, "Force Concept Inventory," *Phys. Teach.* **30**, 141-158 (1992); (b) I. Halloun, R.R. Hake, E.P. Mosca, and D. Hestenes, *Force Concept Inventory* (Revised, 1995) in ref. 20 and password protected at < http://modeling.la.asu.edu/modeling.html >.

20. E. Mazur, *Peer Instruction: A User's Manual* (Prentice Hall, 1997); online at < http://galileo.harvard.edu/ >. "Professionals" may obtain free copies of the book (a) from Prentice Hall campus representatives, or (b) by downloading the Adobe Acrobat portable document file at < http://galileo.harvard.edu/ >.

21. I. Halloun and D. Hestenes, “The initial knowledge state of college physics students,” *Am. J. Phys.* **53**, 1043-1055 (1985); “Common sense concepts about motion,” *ibid.* **53**, 1056-1065 (1985).

22. K. Cummings, J. Marx, R. Thornton, D. Kuhl, “Evaluating innovations in studio physics,” *Phys. Educ. Res. Sup. to AJP* **67**(7), S38-S44 (1999).

23. For a discussion of “matched data” see ref. 1; and also ref. 7, Table I, footnote “c” on page 7.

24. D. Meltzer, "Are There "Hidden Variables" in Students' Initial Knowledge States Which Correlate with Learning Gains?" *AAPT Announcer* **28**(4), 81 (1999); “Relationship between Mathematics Preparation and Conceptual Learning Gains” *AAPT Announcer* **30**(2), 111 (2000); “Re: validity of <g>,” PhysLrnR < http://listserv.boisestate.edu/archives/physlrnr.html > post of 5/14/00. In three of four classes tested with the *Conceptual Survey in Electricity and Magnetism* (ref. 31), Meltzer measured significant positive correlations between single student g’s and math-skills pretest scores. For the three classes (N = 182) the number-of-student-weighted correlation was +0.37.

25. R.R. Hake, R. Wakeland, A. Bhattacharyya, and R. Sirochman, “Assessment of Individual Student Performance in an Introductory Mechanics Course,” *AAPT Announcer* **24**(4), 76 (1994). Scatter plots of FCI gains (posttest – pretest) vs pretest scores for all students in a class delineate relatively high-g (low-g) students for whom the course was (was not) effective. We discuss various diagnostic tests (mechanics, mathematics, and spatial visualization) given to incoming students which might be used to recognize potential “low gainers” and thus initiate helpful intervention. We measured correlation coefficients between single student g’s and test scores for spatial visualization ability and math skills of, respectively, +0.23 and +0.32.

26. See ref. 6, Sec. V and also footnote #46, for a discussion of systematic and random errors in pre/post testing and the connection between low correlation of single students g’s with their pretest scores, and the small difference between values of g_{ave} and <g>.

27. R.R. Hake, “Errors in the Normalized Gain,” 1/7/99, unpublished, available as [Errors-g.pdf, 32K] by request. This paper is a precursor to the final survey manuscript (ref. 6). I analyzed data for 10 different populations ($26 < N < 210$) and found correlations between individual student normalized gain and pretest ranging between +0.57 and –0.42, with an average of +0.31 (sd = 0.13). I found that the differences in g due to the two different types of averaging were small (all less than 4%), and that the correlations between individual student normalized gain and pretest were roughly proportional to the g difference due to the different averaging methods as explained here and in ref. 26 above.

28. C. R. Henderson, K. Heller, & P. Heller, "Common Concerns about the Force Concept Inventory," *AAPT Announcer* **29**(4), 99 (1999); online at < http://www.physics.umn.edu/groups/physed/Talks/talks.html >. Henderson et al. also present evidence that (a) giving the FCI as a pretest does not bias posttest results, and (b) there are gender differences in FCI scores as shown by a plot of pretest vs posttest scores for 392 females and 1233 males. Similar gender differences are apparent in the FCI/MD *normalized* gains at Harvard < http://galileo.harvard.edu/galileo/lgm/pi/testdata.html > and Indiana University (ref. 25).

29. R.J. Beichner, "Testing student interpretation of kinematics graphs," *Am. J. Phys*. **62**, 750-762 (1994).

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

31. D.P. Maloney, T.L. O'Kuma, C.J. Hieggelke, & A. Van Heuvelen, "Surveying Students' Conceptual Knowledge of Electricity and Magnetism," submitted to *Phys. Educ. Res. Sup. to AJP*.

32. J. Cohen, *Statistical power analysis for the behavioral sciences*. (Lawrence Erlbaum, 2nd ed., 1988), p. 44.

33. For discussion and references on the "effect size" and its growing emphasis in the non-physical-science literature in place of t-tests, p values, and null-hypothesis testing, see ref. 1. For a useful summary of effect size information see L.E. Becker, Colorado University – Colorado Springs, Psychology Dept., Psychology 590, Lecture Notes on Effect Size, online at < http://www.web.uccs.edu/lbecker/Psy590/es.htm >.

34. R.R. Hake, "Analyzing change/gain scores," 6/19/99, unpublished; online at < http://www.physics.indiana.edu/~sdi/ >, pdf format (16K): An analysis of the data of ref. 6 & 7 in terms of "effect size," so commonly considered in the non-physical-science literature.

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

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

37. G.F. Kuder & M.W. Richardson, "The theory of estimation of test reliability," *Psychmetrika* **2**, 151-160 (1937).

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

39. L. Cronbach, *Essentials of Psychological Testing* (Harper & Row, 3rd ed., 1970).

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

41. J.M. Saul, 1998, *Beyond Problem Solving: Evaluating Introductory Physics Courses Through the Hidden Curriculum.* (Ph.D. thesis, Univ. of Maryland, 1998); abstract online at < http://www.physics.ucf.edu/People/Faculty/saul.html >.

42. R.R. Hake, *Mechanics Test Data Survey Form*, 3/20/97, unpublished; online as ref. 5 at < http://physics.indiana.edu/~hake/ > [SurveyForm032097.pdf, 22K].