Possible Palliatives for the Paralyzing Pre/Post Paranoia that Plagues Some PEP's

Richard Hake[1]
Indiana University, Emeritus

> *…gain scores are rarely useful, no matter how they may be adjusted or refined…investigators who ask questions regarding gain scores would ordinarily be better advised to frame their questions in other ways* (Cronbach & Furby, 1970).

Pre-post Paranoia

Pre/post testing is anathema to many members of the psychology-education-psychometric (PEP) community. This irrational bias stems in part from the dour appraisal of pre/post testing by Cronbach & Furby (1970), echoed down though the literature to present day texts on assessment such as that by Suskie (2004b). In my opinion, the reticence to employ pre/post testing in evaluation, as used so successfully in physics education reform (Hake, 2005, 2006a), is one reason for the glacial progress of educational research (Lagemann, 2000) and reform (Bok, 2005) in higher education.

Should We Measure Change? Yes!

In a recent *Carnegie Perspective*, Lloyd Bond (2005), a senior scholar at the Carnegie Foundation, wrote:

> If one wished to know what knowledge or skill Johnny has acquired over the course of a semester, it would seem a straightforward matter to assess what Johnny knew at the beginning of the semester and reassess him with the same or equivalent instrument at the end of the semester. It may come as a surprise to many that measurement specialists have long advised against this eminently sensible idea. Psychometricians don't like "change" or "difference" scores in statistical analyses because, among other things, they tend to have lower reliability than the original measures themselves. Their objection to change scores is embodied in the very title of a famous paper by Cronbach and Furby "How we should measure "change,"—or should we?"

As for the unreliability of "change scores," such charges by Lord (1956, 1958) and Cronbach and Furby (1970) have been called into question by, for example, Rogosa, Brandt, and Zimowski (1982), Zimmerman and Williams (1982), Rogosa and Willett (1983, 1985), Collins and Horn

[1] Partially supported by NSF Grant DUE/MDR-9253965.

(1991), Rogosa (1995), Wittmann (1997), Zimmerman (1997), and Zumbo (1999). All this more recent work should (but does not) serve as an antidote for those who would dismiss gain scores because of their supposed statistical unreliability.

Aside from the supposed unreliability of "change scores," seven other objections to pre/post testing, have been enumerated by Suskie (2004a), and countered by Hake (2004a) and Scriven (2004). Suskie's fourth objection (as listed by Hake) is:

> If we do indeed see a significant gain, we often can't be sure it's due to our courses/program and not to other experiences…[history]…or normal maturation. A student might have a concurrent part-time job, for example, that has improved her oral communication skills far more than her required speech course.

The View from U.S. Department of Education

"History" and maturation are among the nine threats to internal validity listed in Table 2.4 of Shadish et al. (2002), are discussed on pages 56-57 of that text, and are reiterated by U.S. Department of Education's (USDE's) "Coalition for Evidence-Based Policy" (CEBP) at USDE (2003):

> There is persuasive evidence that the randomized controlled trial, when properly designed and implemented, is superior to other study designs in measuring an intervention's true effect.
>
> 1. "Pre-post" study designs often produce erroneous results. Definition: 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.

But CEBP's criticism of pre/post testing is irrelevant for most of the recent pre/post studies in introductory astronomy, economics, biology, chemistry, computer science, economics, geoscience, engineering, and physics (see below). The reason is that comparison groups *have* been utilized—they are the introductory 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 control (Traditional [T]) groups are drawn from the same generic introductory course taken by relatively homogeneous groups of students, and (b) IE-course teachers in all institutions are drawn from the same generic pool of introductory course teachers who, judging from uniformly poor average normalized gains <g> they obtain in teaching traditional (T) courses, do not vary greatly in their ability to enhance student learning.

In fact, I suspect that the pre/post testing in the disciplines referred to above might pass muster at the USDE's "What Works Clearing House" (http://www.w-w-c.org/) as "*quasi-experimental studies* (Shadish et al., 2002) of especially strong design" (http://www.w-w-c.org/reviewprocess/standards.html).

Towards Valid and Reliable Diagnostic Tests

> *What we assess is what we value. We get what we assess, and if we don't assess it, we won't get it* (Lauren Resnick in Grant Wiggins, 1991).

Of course, pre/post testing is only as good as the tests employed. In some fields, *disciplinary experts* have engaged, or are engaging, in the arduous quantitative and qualitative research required to develop valid and consistently reliable tests that probe for understanding of the basic concepts. A model for such effort is the pioneering but under-appreciated effort of Halloun and Hestenes (1985a, 1985b) in developing the *Mechanics Diagnostic Test*, precursor to the widely used *Force Concept Inventory* (Hestenes et al., 1992). Such test development was among the themes of a recent NSF *Assessing Student Achievement (ASA) Conference* (http://www.drury.edu/multinl/story.cfm?ID=17783&NLID=306; for a report see PKAL, 2006).

Examples of areas in which student learning gains are now being assessed by means of pre/post testing are:

a. Newtonian mechanics (Halloun & Hestenes, 1985a, 1985b; Hestenes et al., 1992; Thornton & Sokoloff, 1998);

b. other physics subjects as indicated at NCSU (2005), FLAG (2005), OERL (2006);

c. Astronomy, Economics, Biology, Chemistry, Engineering (see the references in Hake, 2004b; Geoscience [Libarkin & Anderson (2005)]; and Calculus [Epstein (2006)].

The above mentioned *Force Concept Inventory* and *Mechanics Diagnostic* Tests were used as pre/post tests in a survey Hake (1998a, 1998b, 2002a, 2002b) of 62 university, college, and high-school courses with a total enrollment of 6,542 students. That study demonstrated that "interactive engagement" (IE) methods of instruction *can* yield average normalized gains <g> in conceptual understanding about two standard deviations (cf. Bloom's "The Two Sigma Problem," 1984) greater than in courses using "traditional" (T) methods.

Although ignored by most PEP's, and even by some physicists (e.g., McCray, DeHaan, & Schuck, 2003), Hake's meta-analysis has been noted by, among others, Rothman and Narum (1999), Stokstad (2001), Klmkowsky et al. (2003), Wood and Gentile (2003), Handelsman et al. (2004), Wieman and Perkins (2005), Heron and Meltzer (2005), Nuhfer (2006a, 2006b), and even DeHaan (2006). In addition, it and subsequent confirmatory pre/post studies have at least partially stimulated the reform of a small fraction of introductory physics courses throughout the U.S., including large enrollment courses at Harvard (Crouch & Mazur, 2001), North Carolina State University (Beichner & Saul, 2004), MIT (Dori & Belcher, 2004), University of Colorado at Boulder (Pollock, 2004), and California Polytechnic State University at San Luis Obispo (Hoellwarth et al., 2005).

The approximately two-sigma superiority of IE over T courses in introductory mechanics has been independently corroborated in hundreds of courses with widely varying types of instructors, institutions, and student populations (see e.g., the references in Hake, 2002a, 2002b), thus satisfying Shavelson and Town's (2002) fifth principle of good scientific practice (italics added):

> *Replicate and Generalize Across Studies*: By one replication we mean, at an elementary level, that if one investigator makes a set of observations, another

> investigator can make a similar set of observations under the same conditions…At a somewhat more complex level, *replication means the ability to repeat an investigation in more than one setting (from one laboratory to another or from one field site to a similar field site) and reach similar conclusions.*

Some definitions and explanations: Normalized Gain, Interactive Engagement, Traditional Teaching, and Multiple Choice Tests

a. The average *normalized* gain <g> is the average *actual* gain (<%post> - <%pre>) divided by the *maximum possible gain* (100% - <%pre>), where the angle brackets indicate the class averages. This half-century-old gain parameter was *independently* employed by Hovland et al. (1949), who called g the "effectiveness index"; Gery (1972), who called g the "gap-closing parameter"; and Hake (1998a, 1998b) who called g the "normalized gain". In Hake (1998a, 1998b), the use of the average *normalized* gain, rather than the average *actual* gain allowed meaningful comparison of *course effectiveness* for Harvard students (<%pre> about 70%) with that for non-honors and non-AP high-school students (<%pre> about 30%). It can be shown that such comparison requires that the test pose a performance ceiling effect (PCE) rather than an instrumental ceiling effect (ICE) as discussed in Hake (2006b). For a discussion of <g> in the context of the more psychometrically standard Item Response Theory (IRT) see Mislevy (2006).

b. Interactive Engagement (IE) courses are *operationally* defined, even despite the "anti-positivist vigilantes" (Phillips, 2000), as those designed at least in part to promote conceptual understanding through interactive engagement of students in heads-on (always) and hands-on (usually) activities which yield *immediate* feedback through discussion with peers and/or instructors, during virtually all class/section time.

c. "Traditional" (T) courses are *operationally* defined as those reported by instructors to make little or no use of "interactive engagement" (IE) methods, relying primarily on passive-student lectures, recipe labs, and algorithmic problem exams.

d. Why MCT's? So that the tests can be given to thousands of students in hundreds of courses under varying conditions in such a manner that meta-analyses can be performed, thus establishing general causal relationships in a convincing manner.

e. Can MCT's measure conceptual understanding and higher-order learning? 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 et al., 1992)…is an assessment that uses multiple-choice items to tap into higher-level cognitive processes.

f. Why are IE courses far more effective in promoting conceptual understanding than traditional passive-student methods? The superiority of IE methods is probably related to the "enhanced synapse addition and modification" induced by those methods. Bransford et al. (2000) wrote:

> …synapse addition and modification are lifelong processes, driven by experience. In essence, the quality of information to which one is exposed and the amount of information one acquires is reflected throughout life in the structure of the brain. This process is probably not the only way that information is stored in the brain, but it is a very important way that provides insight into how people learn.

For positive and negative views on the relevance of neuroscience to classroom instruction see the recent articles by, respectively, Willis (2006) and Bruer (2006).

The approximately two-sigma superiority of IE over T courses in introductory mechanics has been independently corroborated in hundreds of courses with widely varying types of instructors, institutions, and student populations (see e.g., the references in Hake, 2002a, 2002b), thus satisfying Shavelson and Town's (2002) fifth principle of good scientific practice.

Is There an Education Community Map?

Associated with the need for replication, good science requires "continual interaction, exchange, evaluation, and criticism so as to build a…*community map*" (Redish, 1999). The latter crucial feature of the scientific method has also been emphasized by, for example, Gottfried and Wilson (1997), Ziman (2000), Hake (2000a), Cromer (1997), Gere (1997), and Newton (1997), but does not generally characterize education research. In fact, whether or not an "education community" even exists is problematic. Lagemann (2000, p. 239) writes:

> …there are very few filters of quality in education. There is neither a Better Business Bureau nor the equivalent of the Federal Food and Drug Administration. Caveat emptor is the policy of this field. This is because education research has never developed a close-knit professional community, which is the prerequisite for the creation of regulatory structures that can protect both the welfare and safety of the public at large and the integrity of the profession. Such communities exist in some disciplines, for example, physics, and, to a lesser extent, psychology; they also exist in some professions, notably medicine and law. *But such a community has never developed in education* (italics added).

Whither Undergraduate Education?

Wood and Gentile cogently characterize the present state of affairs in undergraduate *science* education (references have been changed to match those in the reference list below, italics added):

> Unknown to many university faculty in the natural sciences, particularly at large research institutions, is a large body of recent research from educators and cognitive scientists on how people learn (Bransford et al., 2000). The results show that *many standard instructional practices in undergraduate teaching, including traditional lecture, laboratory, and recitation courses, are relatively ineffective at*

> *helping students master and retain the important concepts of their disciplines over the long term.* Moreover, these practices do not adequately develop creative thinking, investigative, and collaborative problem-solving skills that employers often seek. Physics educators have led the way in developing and using objective tests (Hestenes et al., 1992; Hake, 1998a; NCSU, 2006) to compare student learning gains in different types of courses, and chemists, biologists, and others are now developing similar instruments (Mulford & Robinson, 2002; Klymkowsky et al., 2003; Klymkowsky, 2006). These tests provide convincing evidence that students assimilate new knowledge more effectively in courses including active, inquiry-based, and collaborative learning, assisted by information technology, than in traditional courses (Hake, 1998a; NCSU, 2006).

But how about undergraduate education *generally*? I see no reason that student learning gains far larger than those in traditional courses could not eventually be achieved and documented in other disciplines from arts through philosophy to zoology *if* their practitioners would (a) reach a consensus on the *crucial* concepts that all beginning students should be brought to understand, (b) undertake the lengthy qualitative and quantitative research required to develop multiple-choice tests (MCT's) of higher-level learning of those concepts, so as to gauge the need for and effects of non-traditional pedagogy, and (c) develop Interactive Engagement methods suitable to their disciplines.

Formative versus Summative Assessment

It should be emphasized that such low-stakes *formative* pre/post testing as discussed above is the polar opposite of the high-stakes *summative* testing mandated by the U.S. Department of Education's "No Child Left Behind Act" for K-12 (USDE, 2005) that is now contemplated for higher education (USDE, 2006). As the NCLB experience shows, such testing often falls victim to *Campbell's Law* (Campbell, 1975; Nichols & Berliner, 2005):

> *The more any quantitative social indicator is used for social decision making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor.*

Why is the pre/post testing discussed above regarded as *formative*? Because both teachers' *action research* and education researchers' scientific research is carried out to improve classroom teaching and learning, NOT to rate instructors or students. Thus it's "formative" as defined by JCSEE (1994): "Formative evaluation is evaluation designed and used to improve an object, especially when it is still being developed."

Why Worry About Student Learning in Higher Education?

But why all the concern for enhancing student learning in higher education? Although international competitiveness is often cited by politicians, business leaders, and educational administrators, more crucial in my view is the need to overcome the monumental problems now threatening life on planet Earth. In "The General Population's Ignorance of Science Related Societal Issues: A Challenge for the University" (Hake, 2000b) I list a few (14) such problems and cite the imperative to (a) educate more effective science majors and science-trained professionals, and (b) raise the appallingly low level of science literacy among the general

population by properly educating prospective K-12 teachers.

Human history becomes more and more a race between education and catastrophe (H. G. Wells, 1920).

Reference & Footnotes (Tiny URL's courtesy http://tinyurl.com/create.php)

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

Bloom, B. S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher*, *13*(6), 4-16. Bloom wrote: "Using the standard deviation (sigma) of the control (conventional) class, it was typically found that the average student under tutoring was about two standard deviations above the average of the control class…The tutoring process demonstrates that *most* of the students do have the potential to reach this high level of learning. I believe an important task of research and instruction is to seek ways of accomplishing this under more practical and realistic conditions than the one-to-one tutoring, which is too costly for most societies to bear on a large scale. THIS IS THE *2 SIGMA* PROBLEM."

Bok, D. (2005). *Our underachieving colleges: A candid look at how much students learn and why they should be learning more*. Princeton University Press.

Bond. L. (2005). Carnegie perspectives: A different way to think about teaching and learning—Who has the lowest prices. Available at <http://www.carnegiefoundation.org/perspectives/sub.asp?key=245&subkey=569>.

Bransford, J. D., Brown, A. L., Cocking, R. R. (Eds.). (2000). *How people learn: brain, mind, experience, and school*. Nat. Acad. Press. Available at http://tinyurl.com/apbgf. The quote is from page 106 of the earlier 1999 edition.

Bruer, J. T. (2006). Points of view: On the implications of neuroscience research for science teaching and learning: Are there any? A skeptical theme and variations: The primacy of psychology in the science of learning. *CBE-Life Sciences Education*, *5*, 104-110. Available at http://www.lifescied.org/cgi/reprint/5/2/104

Campbell, D. T. (1975). Assessing the impact of planned social change. In G. Lyons (Ed.), *Social research and public policies: The Dartmouth/OECD Conference* (pp. 3-45). Dartmouth College Public Affairs Center. Available at http://www.wmich.edu/evalctr/pubs/ops/ops08.pdf

Cromer, A. (1997). *Connected knowledge*. Oxford University Press.

Cronbach, L., & 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. *Am. J. Phys.*, *69*, 970-977. Available at http://tinyurl.com/d35z4

DeHaan, R. L. (2006). Comparing tertiary-level science instruction in China and the United States. To be published in the Proceedings of a Hong Kong Education Conference.

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). Available at http://tinyurl.com/cqoqt

Epstein, J. (2006). *The calculus concept inventory*. Available http://mathed.asu.edu/CRUME2006/Abstracts.html

FLAG. (2003). *Field-tested learning assessment guide*. Available at http://www.flaguide.org/. "...offers broadly applicable, self-contained modular classroom assessment techniques (CAT's) and discipline-specific tools for STEM [Science, Technology, Engineering, and Mathematics] instructors interested in new approaches to evaluating student learning, attitudes, and performance. Each has been developed, tested, and refined in real colleges and university classrooms." Assessment tools for physics and astronomy (and other disciplines) are available at http://www.flaguide.org/tools/tools.php

Gottfried, K., & Wilson, K. G. (1997). Science as a cultural construct. *Nature*, *386*, 545. They attack the "strong program" of the Edinburg school of social constructivists. Available at http://www.nature.com/nature/journal/v386/n6625/abs/386545a0.html. The abstract is: "Scientific knowledge is a communal belief system with a dubious grip on reality, according to a widely quoted school of sociologists. But they ignore crucial evidence that contradicts this allegation."

Giere, R. N. (1997). *Understanding scientific reasoning*. Holt, Rinehart, and Winston.

Hake, R. R. (1998a). Interactive-engagement vs traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *Am. J. Phys.*, *66*, 64-74. Available at http://www.physics.indiana.edu/~sdi/ajpv3i.pdf

Hake, R. R. (1998b). *Interactive-engagement methods in introductory mechanics courses*. Available at http://www.physics.indiana.edu/~sdi/IEM-2b.pdf. A crucial companion paper to Hake, 1998a.

Hake, R. R. (2000a). *Towards paradigm peace in physics education research*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, 24-28 April. Available at http://www.physics.indiana.edu/~sdi/AERA-Hake_11.pdf. Also at that location is a pdf version, available at http://www.physics.indiana.edu/~hake/ParadigmSlides.pdf

Hake, R. R. (2000b). The general population's ignorance of science related societal issues: A challenge for the university. *AAPT Announcer*, *30*(2), 105. Available at http://www.physics.indiana.edu/~hake/GuelphSocietyG.pdf. Based on an earlier libretto with the leitmotiv: "The road to U.S. science literacy begins with effective university science courses for pre-college teachers." The opera dramatizes the fact that the failure of universities throughout the universe to properly educate pre-college teachers is responsible for our failure to observe any signs of either terrestrial or extraterrestrial

intelligence.

Hake, R. R. (2002a). Lessons from the physics education reform effort. *Ecology and Society*, *2*, 28. Available at http://www.ecologyandsociety.org/vol5/iss2/art28/

Hake, R. R. (2002b). Assessment of physics teaching methods. *Proceedings of the UNESCO-ASPEN Workshop on Active Learning in Physics*, Univ. of Peradeniya, Sri Lanka, 2-4 Dec. Available at http://www.physics.indiana.edu/~hake/Hake-SriLanka-Assessb.pdf

Hake, R. R. (2004a). *Re: pre-post testing in assessment*. Available at http://listserv.nd.edu/cgi-bin/wa?A2=ind0408&L=pod&P=R9135&I=-3

Hake, R. R. (2004b). *Re: Measuring content knowledge, POD posts of 14 &15 Mar 2004*. available at http://listserv.nd.edu/cgi-bin/wa?A2=ind0403&L=pod&P=R13279&I=-3 and http://listserv.nd.edu/cgi-bin/wa?A2=ind0403&L=pod&P=R13963&I=-3

Hake, R. R. (2005). *The physics education reform effort: A possible model for higher education?* Available at http://www.physics.indiana.edu/~hake/NTLF42.pdf. This is a slightly edited version of an article that was (a) published in the *National Teaching and Learning Forum* 15(1), December, available at http://www.ntlf.com/FTPSite/issues/v15n1/physics.htm, and (b) disseminated by the *Tomorrow's Professor* list http://ctl.stanford.edu/Tomprof/postings.html as Msg. 698 on 14 Feb 2006. For an executive summary see Hake (2006a).

Hake, R. R. (2006a). A possible model for higher education: The physics reform effort. *Spark* (American Astronomical Society Newsletter), June. Available at http://www.aas.org/education/spark/SparkJune06.pdf

Hake, R. R. (2006b). *Re: Ceiling effects: Performance and instrumental #3*. Available at http://listserv.nd.edu/cgi-bin/wa?A2=ind0607&L=pod&P=R5240&I=-3

Handelsman, J., Ebert-May, D., Beichner, R., Bruns, P., Chang, A., DeHaan, R., Gentile, J., Lauffer, S., Stewart, J., Tilghman, S. M., & Wood, W. B. (2004). Scientific teaching. *Science*, *304*(23), 521-522. Available at http://www.plantpath.wisc.edu/fac/joh/scientificteaching.pdf. See also the supporting material at http://scientificteaching.wisc.edu/resources.htm

Heron, P. R. L., & Meltzer, D. (2005). The future of physics education research: Intellectual challenges and practical concerns. *Am. J. Phys.*, *73*(5), 459-462. Available at http://www.physicseducation.net/docs/Heron-Meltzer.pdf

Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. *Phys. Teach.*, *30*, 141-158. Available 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. *Am. J. Phys.*, *73*(5), 459-463. Available at http://tinyurl.com/br88n

IES. (2006). *Institute for Education Sciences*. Available at

http://www.ed.gov/about/offices/list/ies/index.html?src=oc

Joint Committee on Standards for Educational Evaluation. (1994). *The program evaluation standards* (2nd ed.). Sage. A glossary of evaluation terms from this publication is online at http://ec.wmich.edu/glossary/prog-glossary.htf

Klymkowsky, M. W., Garvin-Doxas, K., & Zeilik, M. (2003). Bioliteracy and teaching efficiency: What biologists can learn from physicists. *Cell Biology Education, 2*, 155-161. Available at http://www.cellbioed.org/article.cfm?ArticleID=67

Klymkowsky, M. W. (2006). *Bioliteracy.net*. Available at http://bioliteracy.net/. "Our goal is to generate, test and distribute the tools to determine whether students are learning what teachers think they are teaching. We assume that accurate and timely assessment of student knowledge will pressure the educational world toward more effective teaching. WHY? (a) Because basic understanding of the biological sciences impacts our lives in more and more dramatic ways every year. (b) A wide range of important personal, social, economic and political decisions depend upon an accurate understanding of basic biology and the means by which science generates, tests and extends our knowledge.

Lagemann, E. C. (2000). *An elusive science: The troubling history of educational research*. University of Chicago Press.

Libarkin, J. C., & Anderson, S. W. (2005). Assessment of learning in entry-level geoscience courses: Results from the Geoscience Concept Inventory. *Journal of Geoscience Education*, *53*, 394-401. Available at http://www.nagt.org/files/nagt/jge/abstracts/Libarkin_v53p394.pdf

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.

McCray, R. A., DeHaan, R. L., Schuck, J. A. (Eds.). (2003). *Improving undergraduate instruction in science, technology, engineering, and mathematics: Report of a workshop*. Committee on Undergraduate STEM Instruction, National Research Council, National Academy Press. Available at http://www.nap.edu/catalog/10711.html. Physicists/astronomers attending the workshop were Paula Herron, Priscilla Laws, John Lehman, Ramon Lopez, Richard McCray, Lillian McDermott, Carl Wieman, Jack Wilson, and Mike Zeilik.

Mislevy, R. (2006). (a) On approaches to assessing change, and (b) "Clarification". Available at http://www.education.umd.edu/EDMS/mislevy/papers/Gain/

NCSU. (2006). *Assessment instrument information page*. Physics Education R & D Group, North Carolina State University. Available at http://www.ncsu.edu/per/TestInfo.html

Nichols, S. L., & Berliner, D. C. (2005). *The inevitable corruption of indicators and educators through high-stakes testing*. Arizona State Univ., Education Policy Studies Laboratory. Available at http://tinyurl.com/7butg

Newton, R. G. (1997). *The truth of science: Physical science and reality*. Harvard University

Press.

Nuhfer, E. (2006a). A fractal thinker looks at measuring change: Part 1 - pre-post course tests and multiple working hypotheses - Educating in fractal patterns XVI. *National Teaching and Learning Forum*, *15*(4). Available at http://www.ntlf.com/. If your institution doesn't have a subscription, IMHO it should.

Nuhfer, E. (2006b). A fractal thinker looks at measuring change: Part 2 - pre-post assessments - Are all interpretations equally valid? Educating in fractal patterns XVII. *National Teaching and Learning Forum*, *15*(6). Available at http://www.ntlf.com/

OERL. (2006). *Online evaluation resource library*. Available at http://oerl.sri.com/

Phillips, D. C. (2000). *Expanded social scientist's bestiary: a guide to fabled threats to, and defenses of, naturalistic social science*. Rowman & Littlefield. See especially Chapter 9 on "Positivism."

PKAL. (2006). *Project Kaleidoscope, volume IV: What works, what matters, what lasts: Assessing student achievement*. Available at http://www.pkal.org/collections/AssessingStudentAchievement.cfm. A report on the NSF's *Assessing Student Achievement (ASA) Conference* of 19-21 October 2006.

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

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

Rogosa, D. R. (1995). Myth and methods: 'Myths about longitudinal research' plus supplemental questions. In J. M. Gottmann (Ed.), *The analysis of change* (pp. 3-66). Erlbaum. Examples from this paper are available at http://www.stanford.edu/~rag/Myths/myths.html

Shadish, W.R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin.

Rothman, F. G. & Narum, J. L. (1999). *Then, now, and in the next decade: A commentary on strengthening undergraduate science, mathematics, engineering, and technology education*. Available at http://www.pkal.org/documents/ThenNowAndInTheNextDecade.cfm

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental*

designs for generalized causal inference. Boston, MA: Houghton Mifflin.

Shavelson, R. J. & Towne, L. (Eds.). (2002). *Scientific research in education*. National Academy Press. Available at http://www.nap.edu/catalog/10236.html

Stokstad, E. (2001). Reintroducing the intro course. *Science*, *293*, 1608-1610. Stokstad wrote: "Physicists are out in front in measuring how well students learn the basics, as science educators incorporate hands-on activities in hopes of making the introductory course a beginning rather than a finale."

Suskie, L. (2004a). Re: pre- post testing in assessment. Available at http://tinyurl.com/akz23

Suskie, L. (2004b). *Assessing student learning*. Anker Publishing.

Scriven, M. (2004). Re: pre- post testing in assessment. Available at http://tinyurl.com/942u8

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. *Am. J. Phys*., *66*(4), 338-352.

U.S. Department of Education. (2003). *Identifying and implementing educational practices supported by rigorous evidence: A user friendly guide*. Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. Available at http://www.ed.gov/rschstat/research/pubs/rigorousevid/rigorousevid.pdf. The Guide's authoring group, the Coalition for Evidence-Based Policy (CEBP) <http://coexgov.securesites.net/index.php?keyword=a432fbc34d71c7> was formerly a part of the Institute of Education Sciences [IES (2006)], in turn a part of the USDE [for the structure of this bureaucratic colossus see <http://www.ed.gov/about/offices/or/index.html?src=ln>]. The CEBP is now sponsored by the "council for excellence in government" <http://coexgov.securesites.net/index.php>, with "the mission to promote government policymaking based on rigorous evidence of program effectiveness." The CEBP's Board of Advisors <http://coexgov.securesites.net/index.php?keyword=a432fbc71d7564> includes luminaries such as famed Randomized Control Trial (RCT) authority Robert Boruch (University of Pennsylvania); political economist David Ellwood (Harvard); former FDA commissioner David Kessler (Univ. of California - San Francisco); past American Psychological Association president Martin Seligman (University of Pennsylvania); psychologist Robert Slavin (Johns Hopkins); economics Nobelist Robert Solow (MIT); and progressive-education basher Diane Ravitch. Unfortunately, no physical scientists, mathematicians, philosophers, or K-12 teachers are members of the CEBP.

U.S. Department of Education. (2005). *No Child Left Behind Act*. Available at http://www.ed.gov/nclb/landing.jhtml?src=pb

U.S. Department of Education. (2006). *A test of leadership: Charting the future of U.S. higher education*. Available at http://www.ed.gov/about/bdscomm/list/hiedfuture/reports/pre-pub-report.pdf. The report states: We believe that improved accountability is vital to ensuring the success of all the other reforms we propose. Colleges and universities must become more transparent about cost, price, and student success outcomes, and must

willingly share this information with students and families. Student achievement, which is inextricably connected to institutional success, must be measured by institutions on a "value-added" basis that takes into account students' academic baseline when assessing their results. This information should be made available to students, and reported publicly in aggregate form to provide consumers and policymakers an accessible, understandable way to measure the relative effectiveness of different colleges and universities.

Wieman, C., & Perkins, K. (2005). Transforming physics education. *Phys.Today*, *58*(11), 36-41. Available at http://www.colorado.edu/physics/EducationIssues

Wiggins, G. (1991). Toward assessment worthy of the liberal arts: The truth may make you free, but the test may keep you imprisoned. Appendix to the MAA's *Getting Started With Assessment*. Available at http://www.maa.org/saum/gettingstarted.html

Willis, J. (2006). RESEARCH WATCH II: Add the science of learning to the art of teaching to enrich classroom instruction. *National Teaching and Learning Forum*, *15*(5). Available at http://www.ntlf.com/FTPSite/issues/v15n5/research2.htm

Wilson, M. R., & Bertenthal, M. W. (Eds.). (2005). *Systems for state science assessment*. Nat. Acad. Press. Available at 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. Available at http://www.psychologie.uni-mannheim.de/psycho2/psycho2.en.php3?language=en

Wood, W. B., & Gentile, J. M. (2003). Teaching in a research context. *Science*, *302*, 1510. Available at http://www.sciencemag.org/content/vol302/issue5650/index.shtml

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

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.

Zimmerman, D. W., & Williams, R. H. (1982). Gain scores in research can be highly reliable. *Journal of Educational Measurement*, *19*, 149-154. The abstract, available at http://mypage.direct.ca/z/zimmerma/earlier.html reads: The common belief that gain scores are unreliable is based on certain assumptions about the values of parameters in a well known formula for the reliability of differences. In this paper we show that a reliability coefficient calculated from the formula can be high, provided one makes other assumptions about the values of pretest and posttest reliability coefficients and standard deviations. Furthermore, there is reason to believe that the revised assumptions are more realistic than the usual ones in testing practice.

Zumbo, B. D. (1999). The simple difference score as an inherently poor measure of change: Some reality, much mythology. In Thompson (Ed.), *Advances in social science methodology*, Volume 5, (pp. 269-304). JAI Press. Available at http://tinyurl.com/kuf3tb