ANALYZING CHANGE/GAIN SCORES*†

Richard R. Hake, Dept. of Physics, Indiana University
24245 Hatteras Street, Woodland Hills, CA, 91367 USA

In an American Educational Research Association (Division D)* mailing list posting of 22 Feb 1999 of the above title John Reece wrote: "I am looking for some good references on the analysis of gain/change scores. The type of scenario I'm interested in is either an experimental or quasi-experimental pre-test/intervention/post-test design. I'm particularly interested in the issues surrounding the use of ANCOVA versus direct analysis of gain scores."

In the survey of refs. 2-4, I did not make use of ANCOVA (ANalysis of COVAriance[1]) , but instead analyzed gains in terms the average normalized gain <g> for a course, defined as the ratio of the actual average gain <G> to the maximum possible average gain, i.e.,

$$\langle g\rangle \equiv \%\langle G\rangle / \%\langle G\rangle_{max} = (\%\langle S_f\rangle - \%\langle S_i\rangle)/(100 - \%\langle S_i\rangle), \quad\quad (1)$$

where $\langle S_f\rangle$ and $\langle S_i\rangle$ are the final (post) and initial (pre) class averages on the "Force Concept Inventory,"[5] a well-known test of conceptual understanding of Newtonian mechanics.

I operationally defined:

(a) "Interactive Engagement" (IE) methods as those designed at least in part to promote conceptual understanding through interactive engagement of students in heads-on (always) and hands-on (usually) activities which yield immediate feedback through discussion with peers and/or instructors, all as judged by their literature descriptions;

(b) "Traditional" (T) courses as those reported by instructors to make little or no use of IE methods, relying primarily on passive-student lectures, recipe labs, and algorithmic-problem exams;

(c) "Interactive Engagement" (IE) courses as those reported by instructors to make substantial use of IE methods;

(d) "High-g" courses as those with (<g>) > 0.7;

(e) "Medium-g" courses as those with 0.7 > (<g>) > 0.3;

(f) "Low-g" courses as those with (<g>) < 0.3.

---

*Originally posted on 3/13/99 at AERA-D - American Educational Research Association's Division D, Measurement and Research Methodology (archived in a somewhat garbled form at <http://lists.asu.edu/cgi-bin/wa?A2=ind9903&L=aera-d&P=R6855>). Some minor additions and corrections were made on 6/16/99 by R. Hake.]

†This work received partial support from NSF Grant DUE/MDR9253965.

The regularities for courses with a wide range of average pretest scores [$18 < (\langle S_i \rangle) < 71$] and with diverse student populations in high schools, colleges, and universities was noteworthy:

(A) All points for the 14 T courses (N = 2084) fell in the Low-g region. The data[3] yielded

$$\langle\langle g \rangle\rangle_{14T} = 0.23 \pm 0.04\text{sd} \qquad \text{(2a)}$$

[Double carets "$\langle\langle X \rangle\rangle_{NP}$ " indicate an average of averages, i.e., an average of $\langle X \rangle$ over N courses of type P, and sd ≡ standard deviation.]

(B) Eighty-five percent (41 courses, N = 3741) of the 48 IE courses fell in the Medium-g region and 15% (7 courses, N = 717) in the Low-g region. Overall, the data[3] yielded

$$\langle\langle g \rangle\rangle_{48IE} = 0.48 \pm 0.14\text{sd} \qquad \text{(2b)}$$

(C) No course points lay in the "High-g" region.

I inferred from features A, B, and C that a consistent analysis over diverse student populations with widely varying initial knowledge states, as gauged by $\langle Si \rangle$, could be obtained by taking the normalized average gain $\langle g \rangle$ as a rough measure of the effectiveness of a course in promoting conceptual understanding.

This inference was bolstered by the fact that the correlation of $\langle g \rangle$ with $\langle S_i \rangle$ for the 62 survey courses is a very low + 0.02. In contrast, the average posttest score $\langle S_f \rangle$ and the average gain $\langle G \rangle$ are less suitable for comparing course effectiveness over diverse groups since their correlations with $\langle Si \rangle$ are, respectively, + 0.55 and – 0.49. [It should be noted that a positive correlation of $\langle S_f \rangle$ with $\langle S_i \rangle$ would be expected in the absence of instruction.]

An "effect size, (d)"[6] of a sort can be calculated from the averaged $\langle\langle g \rangle\rangle$ values as follows:

$$d = [\langle\langle g \rangle\rangle_{48IE} - \langle\langle g \rangle\rangle_{14T}] / \langle SD \rangle = (0.48 - 0.23)/0.09 = 2.78 \qquad \text{(3)}$$

where $\langle SD \rangle$ is the average of the standard deviations for the $\langle g \rangle_{IE}$ and $\langle g \rangle_T$ distributions = (0.14 + 0.04)/2 = 0.09, for Interactive Engagement (IE) and Traditional (T) classes.

The average normalized gain $\langle g \rangle$ has become reasonably well established in the physics/astronomy literature[2-4, 7-12] as a sensible method of analyzing pre/post test results, even though most education researchers have never heard of it and prefer to analyze data in terms of "effect sizes" and/or ANCOVA.

I have gone through the data tables of ref. 3 and calculated the d's for all data for which SD's were reported. The results are:

$$\langle d \rangle_{9T}(N = 1620) = 0.88 \qquad \text{(4a)}$$

$$\langle d \rangle_{24IE}\ (N = 1843) = 2.18 \qquad \text{(4b)}$$

Eq. (4b) is comparable to the d = 1.91 reported by Zeilik *et al.*[7a] for a single introductory astronomy course (N = 221) given in Spring 1995 at the University of New Mexico. The course employed IE methods and achieved an average normalized gain $<g>_{IE} = 0.48$.

It is interesting to compare the d = 2.78 of Eq. (3) with the lower d = 0.51 obtained in a meta-analysis of small-group learning by Springer *et al.*[13]: "Based on 49 independent samples, from 37 studies encompassing 116 separate findings, students who learned in small groups demonstrated greater achievement (d=0.51) than students who were exposed to instruction without cooperative or collaborative grouping."

In the Springer *et al.* study, as for much research reported in the educational literature, (a) in many cases there was no pretesting to disclose initial knowledge states of the test or control groups, (b) the quality of the "achievement tests" was not critically examined (were they of the plug-in-regurgitation type so common in introductory physics courses?). I think that the Springer *et al.* meta-analysis probably understates the effectiveness of cooperative learning in advancing conceptual understanding and problem-solving ability.

It might be of considerable interest to analyze and compare the results of a single, large-N, pre/post study in terms of normalized gain, effect size, and ANCOVA.

References and Footnotes

1. See, e.g., R.E. Slavin, *Research Methods in Education* (Allyn and Bacon, 2nd ed., 1992).

2. 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) and on the Web at <http://carini.physics.indiana.edu/SDI/>, and also the Harvard Galileo server at <http://galileo.harvard.edu/> under "Hands On Methods." See also refs. 3 and 4.

3. R.R. Hake, "Interactive-engagement methods in introductory mechanics courses," on the Web at <http://carini.physics.indiana.edu/SDI/> and submitted on 6/19/98 to the *Physics Education Research Supplement to AJP* (for information on this new journal see <http://www.physics.umd.edu/rgroups/ripe/perg/pers/>.

4. R.R. Hake, "Interactive-engagement vs Traditional Methods in Mechanics Instruction," *APS Forum on Education Newsletter*, Summer 1998, p. 5-7, also at <http://carini.physics.indiana.edu/SDI/>. Some criticisms of ref. 2 and of physics-education reform generally are countered.

5. D. Hestenes, M. Wells, and G. Swackhamer, "Force Concept Inventory," *Phys. Teach*. **30**, 141-158 (1992). I. Halloun, R.R. Hake, E.P. Mosca, and D. Hestenes, Force Concept Inventory (Revised, 1995) in E. Mazur, *Peer Instruction: A User's Manual* (Prentice Hall, 1997) and also password protected at <http://modeling.la.asu.edu/modeling.html>.

6. See, e.g., J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, ed. 2 (Erlbaum, 1988); R.J. Light, J.D. Singer, and J.B. Willett, *By Design: Planning Research on Higher Education*, (Harvard Univ. Press, 1990).

7. (a) M. Zeilik, C. Schau, and N. Mattern, "Misconceptions and Their Change in University-Level Astronomy Courses," *Phys. Teach.* **36**(2), 104-107 (1998); (b) M. Zeilik, C. Schau, and N. Mattern, S. Hall, K.W. Teague, and W. Bisard, "Misconceptions and Their Change in University-Level Astronomy Courses," *Am. J. Phys.* **65**(10), 987-996 (1997); (c) M. Zeilik, C. Schau, and N. Mattern, "Conceptual Astronomy II: Replicating conceptual gains, Probing attitude changes across three semesters" *Am. J. Phys.,* in press, 1999. The latter paper analyzes pre/post test results in terms of both normalized gain and effect size.

8. G.E. Francis, J.P. Adams, and E.J. Noonan, "Do They Stay Fixed?" *Phys. Teach.* **36**(11), 488 (1998).

9. J. Bernhard, "Do They Stay Fixed?" to be submitted to *Phys. Teach.*; on the web at <http://www.du.se/~jbe/fou/didaktik/papers/fixed.pdf>.

10. K. Cummings, J. Marx, R. Thornton, and D. Kuhl, "Innovations in Studio Physics at Rensselaer," *Physics Education Research Supplement to AJP*, in press.

11. E.F. Redish and R.N. Steinberg, "Teaching Physics: Figuring Out What Works," *Phys. Today* **52**(1), 24-30 (1999), on the web at <http://www.physics.umd.edu/rgroups/ripe/papers/PT/pt.htm>.

12. E.F. Redish, "Millikan Award Lecture: Building a Science of Teaching Physics," *Am. J. Phys*., to be published, on the web at <http://www.physics.umd.edu/rgroups/ripe/papers/millikan.htm>.

13. L. Springer, M.E. Stanne, and S. Donovan, "Effects of cooperative learning on undergraduates in science, mathematics, engineering, and technology: A meta-analysis," (Research Monograph No. 11, University of Wisconsin-Madison, National Institute for Science Education) *Review of Educational Research* (in press, 1998); on the web at <http://www.wcer.wisc.edu/nise/CL1/resource/R2.htm>.