Addendum to:

The relationship between mathematics preparation and conceptual learning gains in physics: a possible "hidden variable" in diagnostic pretest scores

David E. Meltzer
Department of Physics and Astronomy, Iowa State University, Ames, Iowa 50011

NORMALIZED LEARNING GAIN: A KEY MEASURE OF STUDENT LEARNING

A single examination yields information about a student's knowledge state at one point in time. However, the primary interest of instructors is in learning, that is, the transition *between* states. In addition to being inadequate by itself for measuring that transition, performance on a single exam may be strongly correlated with a student's preinstruction preparation and knowledge.

In order to assess learning *per se* it is necessary to have a measure that reflects the transition between knowledge states and that has maximum dependence on instruction, with minimum dependence on students' preinstruction states. Therefore, it would be desirable to have a measure that is correlated with instructional activities, but that has minimum correlation with students' pretest scores. In addition, the ideal measure would be reliable in the sense that minor differences in test instrument should yield approximately the same value of learning gain. The question of how to measure learning gain is not a simple one, and it is subject to many methodological difficulties.[1] Here I will simply identify some of the most significant issues.

To the extent that pre- and post-testing makes use of an instrument that has both a minimum and maximum possible score (as most do), there are issues of "floor and ceiling" effects. That is, a student can get no less than 0% nor more than 100% correct on such an instrument. This immediately leads to a problem in quantifying learning gain. For, a student whose pretest score is, say, 20% may gain 80 percentage points, while a student who begins at 90% may gain no more than ten. If the first student gains ten percentage points, and the second gains nine, one would be hard pressed to conclude that the first student has really learned more. Quite likely, a somewhat different assessment instrument might have led to a very different conclusion. (For instance, if the questions were significantly harder so that the two students' pretest scores were 10% and 45%, respectively, the second student might very well have shown a much greater gain as measured simply in terms of percentage points.) As a consequence of the ceiling effect (and other reasons), it is common to observe a strong negative correlation between students' absolute gain scores (posttest score minus pretest score) and their pretest score: higher pretest scores tend to result in smaller absolute gains, all else being equal.

If, therefore, one wants a measure of learning gain that is reliable – such that simply modifying the testing instrument would not lead to widely disparate results – the absolute gain score is unlikely to suffice. The fact that the gain score tends to be correlated (negatively) with pretest scores is also an obstacle to isolating a measure of *learning* from confounding effects of preinstruction state. One way of dealing with this problem is to derive a measure that normalizes the gain score in a manner that takes some account of the variance in pretest scores. As is well known, the use of such a measure in

physics education research was introduced by Hake in Ref. 2; it is called the "normalized gain" g and it simply the absolute gain divided by the maximum possible gain:

$$g = \frac{posttest\ score - pretest\ score}{maximum\ possible\ score - pretest\ score}$$

In Hake's original study he found that, for a sample of 62 high-school, college, and university physics courses enrolling over 6500 students, the mean normalized gain $\langle g \rangle$ for a given course on the Force Concept Inventory was almost completely uncorrelated (r = +0.02) with the mean pretest score of the students in that course. (In this case, mean normalized gain $\langle g \rangle$ is found by using the mean pretest score and mean posttest score of all students in the class.) At the very least, then, $\langle g \rangle$ seems to have the merit of being relatively independent of pretest score, and one might therefore expect that the reliability of whatever measure of learning gain it yields would not be threatened simply because a particular class had unusually high or low pretest scores. That is, if a diverse set of classes has a wide range of pretest scores but all other learning conditions are essentially identical, the values of normalized learning gain measured in the different classes should not differ significantly.

The pretest-independence of normalized gain also suggests that a measurement of the $\langle g \rangle$ difference between two classes having very different pretest scores would be reproduced even by a somewhat different test instrument which results in a shifting of pretest scores. If one class had, say, a $\langle g \rangle = 0.80$ and the other had $\langle g \rangle = 0.20$, using another instrument which significantly alters their pretest scores might still be expected to yield more or less unchanged values of $\langle g \rangle$.

By contrast, Hake found that the absolute gain scores *were* significantly (negatively) correlated with pretest score ($r = -0.49$). Suppose then that Class A (pretest score = 20%) had an absolute learning gain of 16 points, while Class B (pretest score = 90%) only had an eight-point gain, i.e., half as much as A. We would suspect that there might actually have been a *greater* learning gain in Class B, but that it was disguised by the ceiling effect on the test instrument. Suppose an assessment instrument that tested for the same concepts but with harder questions changed the pretest scores to 10% for A, and 45% for B. We might now find A with a gain of 18 points and B with a gain of 44 points – and so now Class *B* is found to have the greater learning gain. Yet in this example, both classes would be found to have unchanged $\langle g \rangle$ values ($\langle g \rangle_A = 0.20$; $\langle g \rangle_B = 0.80$). Thus we can argue that $\langle g \rangle$ is the more reliable measure because the relationship it yields is reproduced with the use of a test instrument that differs from the original only in an insignificant manner.

Now, this is a mock example with made-up numbers and untested assumptions. However it is consistent with the findings of Hake's survey and I would argue that it is a plausible scenario. Some empirical support might be drawn from the data reported at RPI (Ref. 3, Table II). For seven sections of the "standard" studio course, the mean pretest score and mean absolute gain score as measured on the FCI were 49.8% and 9.6%±2.2% (s.e.), respectively (s.e. ≡ standard error). The mean pretest score for the same seven sections as measured on the FMCE was lower (35.4%) and, as we might expect from the discussion above, the mean absolute gain score was indeed higher: 13.8%±1.4% (s.e.). These mean gain scores are significantly different according to a one-tailed, paired two-sample *t*-test ($t = 2.23$, $p = 0.03$), and the higher FMCE gain might naively be interpreted

as a substantially higher learning gain (13.8/9.6 = 144%). However, despite the significantly different pretest scores, the mean normalized gains as determined by the two instruments were statistically indistinguishable: $\langle g \rangle_{FCI} = 0.18 \pm 0.04$(s.e.); $\langle g \rangle_{FMCE} = 0.21 \pm 0.02$(s.e.); ($t = 0.84$, $p = 0.22$). The FMCE and the FCI do not measure exactly the same concept knowledge and so one might argue that the measured learning gains really *should* be different. We have insufficient data to dispute that argument in this case and so this example should be seen as primarily illustrative.

Probably the most persuasive empirical support for use of normalized gain as a reliable measure lies in the fact that $\langle g \rangle$ has now been measured for literally tens of thousands of students in many hundreds of classes worldwide with extremely consistent results. The values of $\langle g \rangle$ observed for both traditional courses and those taught using interactive-engagement methods both fall into relatively narrow bands which are reproduced with great regularity, for classes at a very broad range of institutions with widely varying student demographic characteristics (including pretest scores).[4] This provides a strong argument that normalized gain $\langle g \rangle$ is a valid and reliable measure of student learning gain.

References

[1]Walter R. Borg and Meredith D. Gall, *Educational Research*, *An Introduction* (Longman, New York, 1989), 5th ed., pp. 728-733.

[2]Richard R. Hake, "Interactive engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses," Am. J. Phys. **66,** 64-74 (1998); http://www.physics.indiana.edu/~sdi/.

[3]Karen Cummings, Jeffrey Marx, Ronald Thornton and Dennis Kuhl, "Evaluating innovations in studio physics," Phys. Educ. Res., Am. J. Phys. Suppl. **67**, S38-S44 (1999).

[4]Edward F. Redish and Richard N. Steinberg, "Teaching physics: figuring out what works," Phys. Today **52**, 24-30 (1999); Richard R. Hake, "Lessons from the physics education reform effort," Conservation Ecology **5**, 28 (2002), http://www.consecol.org/vol5/iss2/art28