



Correcting the Normalized Gain for Guessing

John Stewart and Gay Stewart

Citation: *The Physics Teacher* **48**, 194 (2010); doi: 10.1119/1.3317458

View online: <http://dx.doi.org/10.1119/1.3317458>

View Table of Contents: <http://scitation.aip.org/content/aapt/journal/tpt/48/3?ver=pdfcov>

Published by the [American Association of Physics Teachers](#)

Articles you may be interested in

[Authors' response](#)

Phys. Teach. **48**, 212 (2010); 10.1119/1.3361980

[Interactive Engagement: How Much Is Enough?](#)

Phys. Teach. **48**, 108 (2010); 10.1119/1.3293658

[Introducing the USAYPT — Do research in your high school then debate your results with other schools](#)

Phys. Teach. **48**, 48 (2010); 10.1119/1.3274364

[Physics Lab Myths](#)

Phys. Teach. **46**, 560 (2008); 10.1119/1.3023664

[Assessment Strategies to Guide Instruction](#)

Phys. Teach. **45**, 122 (2007); 10.1119/1.2432095



Correcting the Normalized Gain for Guessing

John Stewart and Gay Stewart, University of Arkansas, Fayetteville, AR

The normalized gain, g , has been an important tool for the characterization of conceptual improvement in physics courses since its use in Hake's extensive study on conceptual learning in introductory physics.¹ The normalized gain is calculated from the score on a pre-test administered before instruction and a post-test administered after instruction and is defined as

$$g = \frac{\text{post-test} - \text{pre-test}}{100 - \text{pre-test}}, \quad (1)$$

where both the pre-test and post-test have a maximum score of 100. The statistic has been used in many published works since Hake's paper. It has become sufficiently important that extensions to the statistic² and investigations of its detailed properties³ have recently been published. This paper investigates the effect of students' guessing on the normalized gain and develops a correction for guessing for the pre-test and post-test. The normalized gain is found to be insensitive to the effects of guessing.

The normalized gain was one of the statistics used by external reviewers of educational projects at the University of Arkansas to interpret pre-test and post-test results for the Force Concept Inventory⁴ (FCI) and the Conceptual Survey in Electricity and Magnetism⁵ (CSEM) given to students in the introductory calculus-based physics sequence. The use of the statistic for students in the electricity and magnetism segment of the introductory sequence has long been a concern for the instructor of the course. The CSEM is a multiple-choice test with five possible answer choices per question. Blind guessing on the part of the students would result in an average score of 20%. The most recent five-year course average for the CSEM pre-test score is 30%. Extensive discussions with students indicate that the low scores are a result of a limited familiarity with the material and not due to strongly held misconceptions about electricity and magnetism. It seems likely that the 30% average pre-test score is substantially inflated by guessing. Since the normalized gain is calculated from the pre-test score, it also seems likely that guessing could substantially affect the normalized gain.

Method

To investigate the effect of guessing, the FCI and CSEM were administered with modified scoring rules. Many standardized tests employ some technique to minimize the effect of guessing. The SAT and the GRE penalize the student for each incorrect answer; a similar method was used to discourage guessing in this study. To allow comparison of scores for

which guessing is discouraged to scores with no penalty for guessing, the students were required to answer all questions. The students were also asked to select one of two scoring methods for each problem. If the first scoring option, denoted by "I'm sure," was selected, one point was given to a correctly answered question, but one quarter of a point was subtracted if the question was incorrect. If the second scoring option, denoted by "I'm guessing," was selected, one half of a point was given for a correct answer, but no penalty was imposed for an incorrect answer. Students selecting the second option were sufficiently unsure of the answer to sacrifice half the points to avoid a quarter-point penalty. Most students did not answer in a pattern, selecting only one of the options for all the questions, but seemed to make question-by-question decisions about which option to select.

The FCI and CSEM were given to students in the introductory calculus-based sequence at the University of Arkansas with the above scoring rules. Only students who completed both the pre-test and post-test were included in the study. Students taking the FCI in the introductory first-semester mechanics course were told that the pre-test would be counted if they did well but would not be counted if they did poorly. The FCI post-test was part of the final exam in the first-semester course. In the second-semester electricity and magnetism course, both the CSEM pre-test and post-test were graded as a quiz. The FCI and CSEM were administered with the above grading policy for two semesters, spring 2006 and fall 2006. A total of 316 students completed both the FCI pre-test and post-test and 205 students the CSEM pre-test and post-test.

For the purpose of this study, the tests were graded using two separate methods. In the first grading method, all problems were graded with correct answers receiving one point and incorrect answers none. In the second grading method, problems where the student selected "I'm sure" received one point if correct and none if incorrect, whereas problems where the student selected "I'm guessing" received no points. Scores generated by the first grading method, where all problems were graded, will be called *raw* scores and represent the scores produced by the normal grading method for the FCI or CSEM. Scores generated by the second grading method will be called *corrected* scores and should better represent the score a student would receive on the test if guessing was eliminated. These two scoring methods were chosen to allow the comparison of the score resulting from the most common scoring scheme in which all questions are graded to the scores that would result if guessing was eliminated and the students only answered questions for which they knew the answer.

Table I. Comparison of raw scores, where every question is graded, to corrected scores, where all questions where the student answered “I’m guessing” are graded as incorrect. Percent change is calculated as $100\% \cdot (\text{corrected} - \text{raw})/\text{raw}$. The estimated score is calculated using Eq. (2). The estimate error is the difference between the corrected score and the estimated score ($\text{corrected} - \text{estimated}$).

| | Raw score | Corrected score | Change | % Change | Estimated score | Estimated error |
|----------------------|-----------|-----------------|--------|----------|-----------------|-----------------|
| FCI pre-test | 45.2 | 29.7 | -15.5 | -30 | 29.4 | 0.3 |
| FCI post-test | 73.5 | 69.4 | -4.1 | -6 | 66.3 | 3.1 |
| FCI normalized gain | 0.57 | 0.58 | 0.01 | 3 | 0.54 | 0.04 |
| CSEM pre-test | 29.7 | 9.0 | -20.7 | -69 | 14.1 | -5.1 |
| CSEM post-test | 65.9 | 56.7 | -9.2 | -14 | 57.2 | -0.5 |
| CSEM normalized gain | 0.52 | 0.52 | 0.00 | 1 | 0.50 | 0.02 |

Table II. Correction Equations: This table presents a set of linear equations that allow the calculation of the corrected pre-test and post-test scores from the uncorrected raw value of each quantity.⁶ Pre-test and post-test scores measured from 0 to 100.

| | N | correction | R ² |
|--------------------------------|------|---|----------------|
| FCI pre-test | 316 | $\text{corrected} = 1.03 \cdot \text{raw} - 14.1$ | 0.76 |
| FCI post-test | 316 | $\text{corrected} = 1.07 \cdot \text{raw} - 9.1$ | 0.92 |
| FCI pre-test/post-test pooled | 632 | $\text{corrected} = 1.4 \cdot \text{raw} - 16.8$ | 0.90 |
| CSEM pre-test | 205 | $\text{corrected} = 0.67 \cdot \text{raw} - 11.0$ | 0.41 |
| CSEM post-test | 205 | $\text{corrected} = 1.19 \cdot \text{raw} - 21.7$ | 0.79 |
| CSEM pre-test/post-test pooled | 410 | $\text{corrected} = 1.21 \cdot \text{raw} - 25.2$ | 0.89 |
| All pre-test/post-test pooled | 1042 | $\text{corrected} = 1.19 \cdot \text{raw} - 21.2$ | 0.89 |

Results

The raw and corrected averages for the FCI and the CSEM are presented in Table I. The results of the individual semesters were pooled to form a two-semester average. As expected, there is a substantial correction to the CSEM pre-test score, where a raw average of 30% yielded a corrected average of 9%. A smaller correction is observed for the FCI pre-test score. In general, the size of the correction decreases as the average score increases. Students who know more guess less. While the pre-test and post-test required significant correction for both the FCI and the CSEM, the normalized gain was virtually unchanged.

Linear models were fit to the pre-test and post-test to produce equations that allow the conversion of a raw score into a corrected score. These equations are presented in Table II. No equation is provided for the raw normalized gain since the raw normalized gain is an excellent estimate for the corrected normalized gain. The “FCI Pre-test/Post-test Pooled,” “CSEM Pre-test/Post-test Pooled,” and “All Pre-test/Post-test Pooled” results in Table II are of particular interest. The first is our best estimate of an equation to correct the FCI for a general population of students, the second an equation to correct the CSEM, and the third an equation to correct a general multiple-choice test. These first two equations result from fitting a line to the pooled pre-test/post-test data set of the FCI and CSEM. The equations correcting the FCI and CSEM are somewhat different. The FCI equation maps a raw score of 50% to 40%, whereas the CSEM equation maps a raw score of 50% to 35%. This difference was expected because the student’s increased personal experience prior to the course with the subject matter addressed by the FCI should produce more strongly held misconceptions. The third equation, to be used to correct multiple-choice evaluations other than the FCI and CSEM, results from fitting a data set that pools both the FCI and CSEM pre-test and post-test, and results in Eq. (2):

$$\text{corrected} = 1.19 \cdot \text{raw} - 21.2. \quad (2)$$

Equation (2) results from a linear regression of the raw scores on the corrected scores for a data set that pools the pre-test and post-test scores of both the FCI and the CSEM for both the spring and fall 2006 semesters. Since this equation is drawn from data that pools two different classes with two different student populations, different grading policies, and different testing conditions (pre-test and post-test), it a reasonable estimate of an equation that could be applied to any class. Equation (2) was used to calculate an “estimated” value for all quantities in Table I. These estimated values are superior to the raw score for all quantities except for the normalized gain. The raw normalized gain is the best estimate of the corrected normalized gain for both the FCI and the CSEM: the normalized gain can be used without correction. The insensitivity of the normalized gain to student guessing is further reason for the continued and expanded use of the statistic.

Analysis

The normalized conceptual gain was constructed to allow the comparison of student populations with different backgrounds and therefore is expected to be somewhat insensitive to the pre-test score. The observed stability is a direct result of the invariance of the normalized gain to a certain class of linear transformation. A naive choice for a transformation to remove the effect of guessing is a member of this class of invariant transformations. If one wished to correct the pre-

test or post-test score without resorting to the experiment above, a natural correction to use is one that maps a score of 20% to a score of zero and leaves a score of 100% unchanged. This results in Eq. (3):

$$\text{corrected} = 1.25 \cdot (\text{raw} - 100) + 100 = 1.25 \cdot \text{raw} - 25. \quad (3)$$

Equation (3), a reasonable first attempt at a correction for guessing, is very similar to Eq. (2), our measured general correction for guessing. If Eq. (3) is used to correct both the pre-test and post-test score, and the corrected scores are then used to calculate a corrected normalized gain, one finds that the corrected normalized gain is exactly equal to the uncorrected normalized gain. In fact, the normalized gain is invariant under any linear transformation of both the pre-test and post-test score that has the form of Eq. (4):

$$\text{corrected} = m \cdot (\text{raw} - 100) + 100, \quad (4)$$

where m is any number except zero. Equation (4) is unique only in that it leaves a score of 100 unchanged.

The naive correction for guessing, Eq. (3), maps a score of 20% to zero and leaves a score of 100% unchanged. The general correction for guessing developed above, Eq. (2), maps a score of 20% to a score of 3%, implying that a student who records a score consistent with pure guessing still has some knowledge, and maps a score of 100% to 98%, showing even a perfect score is somewhat affected by guessing.

Discussion

The equations in Table II show substantial variation and none of the pre-test or post-test corrections are identical to Eq. (2), our general correction for guessing. This equation underestimates the correction required for very low pre-test scores, such as the CSEM pre-test. For best results, a scoring policy on multiple-choice instruments that discourages guessing should be adopted, particularly when low pre-test scores are expected. Because of its insensitivity to guessing, the normalized gain could still be used to compare results at different institutions even when a grading policy that discourages guessing is in place.

Conclusion

Pre-test and post-test results on conceptual instruments are substantially affected by guessing. The magnitude of the effect decreases as the score on the pre-test or post-test increases. The normalized gain is virtually unchanged by the effect of guessing and may be safely used even when pre-test and post-test scores require substantial correction.

Acknowledgments

This work supported in part by NSF Grants DUE-9455732 and PHY-0108787.

References

1. 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 (Jan. 1998).
2. Jeffery D. Marx and Karen Cummings, "Normalized change," *Am. J. Phys.* **75**, 87–91 (Jan. 2007).
3. Lei Bao, "Theoretical comparisons of average normalized gain calculations," *Am. J. Phys.* **74**, 917–922 (Oct. 2006).
4. David Hestenes, Malcolm Wells, and Gregg Swackhamer, "Force Concept Inventory," *Phys. Teach.* **30**, 141–158 (March 1992). The version of the FCI used in this study was the revised version included in *Peer Instruction* (E. Mazur, Prentice Hall, 1997).
5. David P. Maloney, Thomas L. O’Kuma, Curtis Hieggelke, and Alan Van Heuvelen, "Surveying students’ conceptual knowledge of electricity and magnetism," *Phys. Educ. Res., Am. J. Phys. Suppl.* **69**, S12–S23 (July 2001).
6. Graphs showing the relationship between the corrected score and the raw score were presented in a poster entitled "Correcting the Normalized Gain" at the 2007 Greensboro AAPT meeting and are available at the University of Arkansas PhysTEC site, physinfo.uark.edu/phystec/research/CorrectGainSummer2007JCS.pdf.

John Stewart is an assistant professor of physics at the University of Arkansas. He is a longtime participant in the University of Arkansas’ PhysTEC project. He is currently editor of the APS Forum on Education Newsletter – Teacher Preparation Section and is also editor of the PTEC collection on CompADRE.
johns@uark.edu

Gay Stewart is an associate professor of physics at the University of Arkansas. She is the site director for the University of Arkansas’ PhysTEC project and Preparing Future Faculty project. She is a member of the National Advisory Board for Physics Teaching Resource Agents and Councilor of the APS.
gstewart@uark.edu