# Why normalized gain should continue to be used in analyzing preinstruction and postinstruction scores on concept inventories

Vincent P. Coletta
*Department of Physics, Loyola Marymount University, Los Angeles, California 90045, USA*

Jeffrey J. Steinert
*Arizona School for the Arts, Phoenix, Arizona 85004, USA*

(Received 10 November 2018; revised manuscript received 5 August 2019; accepted 6 December 2019; published 6 February 2020)

Recently, Nissen *et al.* argued in this journal for the use of Cohen's $d$, in place of the more commonly used normalized gain, in the analysis of preinstruction and postinstruction scores on concept inventories used to measure the effectiveness of instruction. Their reason for advocating such a change is that they say normalized gains are "prescore biased." We provide five examples, including one cited by Nissen, that show no prescore bias when data are carefully analyzed, demonstrating that the problem with their analysis is omitted variable bias. We show that Cohen's $d$ is less informative than normalized gain when used as a single parameter measure of teaching effectiveness, even though, as Nissen points out, $d$ is more widely used in other fields. We believe that physics education researchers should continue to use normalized gain to assess educational effectiveness of pedagogy. However, because different student populations can have significantly different responses to the same pedagogy, in any interpretation of normalized gain, it is important to consider a measure of the abilities of the students. In analyzing normalized gains for the Force Concept Inventory (FCI), average scores on either Lawson's Test of Scientific Reasoning Ability or the SAT should be considered, because these scores are strongly correlated with normalized gain, indicating student abilities may have a greater impact on the gains achieved in a class than the specific pedagogy used.

DOI: 10.1103/PhysRevPhysEducRes.16.010108

## I. INTRODUCTION

Concept inventories are widely used in physics education research (PER) to test the efficacy of alternative pedagogical methods. These include Force Concept Inventory (FCI), Force and Motion Concept Evaluation (FMCE), and Conceptual Survey of Electricity and Magnetism (CSEM). In 1997 Hake introduced normalized gain as a measure of change when the same concept test is used to gauge student understanding at the beginning and again at the end of a physics course [1]. Hake analyzed FCI data from 62 courses with 6542 students and provided compelling evidence of the superiority of interactive engagement (IE) courses in physics to those taught using traditional methods. Hake's original use of normalized gain, which we denote[1] here by $g$, is a measure of change in average class scores, preinstruction to postinstruction, defined as

$$g = \frac{\text{class av \% post-class av \% pre}}{100\% - \text{class av \% pre}}.$$

Normalized gain is the change in the class average score divided by the maximum possible gain. This measure can yield the same value for classes with quite different averages. For example, prescores to postscores of 20% to 60%, 40% to 70%, and 60% to 80% all correspond to $g = 0.5$. Loosely speaking, normalized gain is the fraction of concepts learned by a class that were not known at the beginning of the course.

Nissen *et al.* claim that normalized gains are "prescore biased," and that Cohen's $d$ should be used in place of normalized gain, though normalized gain has been widely used in PER for over 20 years, since its introduction by Hake.

In his study [1], Hake found no significant correlation between $g$ and prescores ($r = 0.02$). In 2005 Hestenes reported no significant correlation between normalized gain and prescore for 12 000 high school students, again showing no prescore bias [3]. One of the attractions of using $g$ as a measure of learning achieved in a class is that it can be independent of the class prescore. However, this lack of correlation is not found consistently. In Hake's study, it appears to have been a consequence of his

[1]Hake's original notation for normalized gain was $\langle g \rangle$ rather than $g$. We use the same notation used by Nissen *et al.* [2].

including both traditionally taught and interactive engagement classes in his analysis. In Hestenes' study it was likely the result of the population being high school students with no prior knowledge of physics. In most IE college classes there is a significant correlation between prescore and normalized gain, but this apparent prescore *bias* is a consequence of analysis that fails to take into account other variables, a case of omitted variable bias.

We shall demonstrate in Sec. II that there is no real prescore bias when data are carefully analyzed. In Sec. III we point out problems with Nissen's own data used to support the argument of prescore bias. In Sec. IV we argue against the use of normalized change, a modification of normalized gain proposed by Marx and Cummings in 2007, but not widely used. Nissen *et al.* describe the Marx and Cummings paper and their argument that traditional normalized gain is sometimes *negatively* prescore biased. Nissen uses their argument as further evidence of prescore bias. We address that argument. In Sec. V we point out problems with the use of Cohen's $d$ in place of normalized gain.

## II. NORMALIZED GAIN IS NOT PRESCORE BIASED; OMITTED VARIABLES BIAS

We shall provide examples showing correlation between prescore and normalized gain, but we shall also show that such correlation is not the result of prescore being a true predictor of gain, but instead the result of other fundamental predictors that are strongly correlated with both prescore and gain. Failure to take account of these other predictors is an error referred to as omitted variable bias [4]. We shall show that one must be cautious in comparing $g$'s for classes with very different student populations, and that valid predictors of gain such as SAT scores should be considered.

### A. Examples 1 and 2: Harvard and LMU

In 2005, Phillips and one of us reported on class average FCI data [5] we collected from 38 classes at seven colleges and universities that used interactive engagement methods of instruction, and also on individual student data—preinstruction and postinstruction scores for 2948 students in 31 classes from four of the seven schools. We analyzed our class average data as well as class average data from the 35 additional IE college and university classes from Hake's study. We did not include in our dataset high school classes or any classes taught using traditional methods, as it was our intent to look for possible correlations of normalized gains with preinstruction FCI scores when IE methods are used in college classes. We found a significant positive correlation between normalized gain $g$ and class average preinstruction score for the 73 IE college and university classes for which we had class average data ($r = 0.39$, $p = 0.0006$). Average prescores ranged between 25% and 70%. Average normalized gain for those classes with average prescores around 25% was roughly 0.3, while average prescores of 70% had average gains of 0.6. This range of 0.3 to 0.6 is the approximate range of normalized gains in all IE classes. These results are sometimes taken to mean that normalized gains are "biased" in favor of higher prescore. However, we shall describe below data that contradict this claim, a result of omitted variable bias.

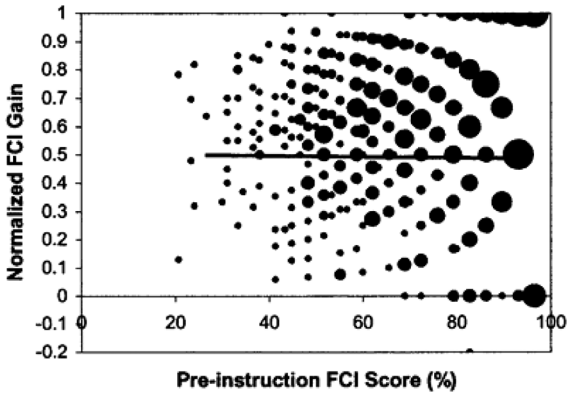


FIG. 1. At Harvard, individual students' normalized gain is not correlated with FCI prescore ($r = 0$, $N = 670$). For example, the average of the individual normalized gains for the 20 Harvard students with the lowest prescores was 0.6, the same as the average normalized gain for students with higher prescores [5]. The largest dot represents 19 students.

In 2002 Hake applied normalized gain to individual student scores [6], defining individual normalized gain,[2] which we denote here by $g_{\text{ind}}$, in terms of a student's individual prescores and postscores:

$$g_{\text{ind}} = \frac{\text{postscore\%} - \text{prescore\%}}{100\% - \text{prescore\%}}.$$

In our 2005 paper [5], Phillips and one of us analyzed individual student FCI prescores and postscores from our own school, Loyola Marymount University (LMU), and three other schools: Southeastern Louisiana University (SLU), University of Minnesota (UM), and Harvard University (HU). We found a significant positive correlation between $g_{\text{ind}}$ and prescore at LMU, SLU, and UM. However, for Harvard students there was no correlation between $g_{\text{ind}}$ and prescore, thus showing no prescore bias (Fig. 1).

In light of the Harvard data, we suspected that there were other population dependent factors that were at work, affecting the gains of students, and that the apparent dependence of gain on prescore was an artifact, a result of the prescore itself being dependent on a fundamental predictor. So we began administering Lawson's Classroom Test of Scientific Reasoning Ability [7] to our students to

[2]Hake reported that the average of $g_{\text{ind}}$ for a class generally differs from $g$ by less than 5% [6].

see whether their preinstruction scores on it were predictors of their normalized gains. In our initial study of 65 LMU students, we indeed found a much stronger correlation between Lawson prescore and normalized gain $g_{\text{ind}}$ ($r = 0.51$, $p < 0.0001$) than we found between FCI prescore and normalized gain $g_{\text{ind}}$ ($r = 0.33$, $p < 0.0001$), as reported in 2005 [5].

Nissen [2] *et al.* cite the correlation we found between $g_{\text{ind}}$ and FCI prescore as evidence for prescore bias of normalized gain. However, a multiple variable regression[3] on both Lawson score and FCI prescore ($r = 0.51$) reveals that FCI prescore is not a significant predictor of $g$ ($p = 0.99$), while Lawson score is ($p = 0.0001$). For the multiple variable regression, the correlation coefficient $r = 0.51$ has the same value found for single variable regression, with Lawson score as the only independent variable. Thus, careful analysis of our data that was cited by Nissen *et al.* [2] as evidence for prescore bias in using normalized gain, reveals that our data show no such bias. Omitting Lawson scores in one's analysis is an example of omitted variable bias, the omitted variable being the Lawson score. The correlation between $g$ and FCI prescore for these students was only a consequence of the correlation between FCI prescore and Lawson score ($r = 0.50$, $p = 0.00001$). Students with higher Lawson scores tended to have higher FCI prescores. This makes sense because it is reasonable that students with greater scientific reasoning ability would likely learn more in their high school classes.

### B. Example 3: Finland

Our Lawson Test correlations with $g$ have been replicated with remarkable consistency for other IE classes at high schools, colleges, and universities in the U.S. and Europe [8–12]. For example, Savinainnen [11] found for 136 Finnish students, that FCI $g_{\text{ind}}$ was correlated with Lawson prescore ($r = 0.53$, $p < 10^{-4}$) and in a multiple regression of $g_{\text{ind}}$ on both Lawson prescore and FCI prescore, the dependence of $g_{\text{ind}}$ on FCI prescore was not significant ($p = 0.08$), even though in a single variable regression, with prescore as the only independent variable, $g$ and prescore were significantly correlated ($r = 0.34$, $p < 10^{-4}$). Again FCI prescores were significantly correlated with Lawson prescores ($r = 0.43$, $p < 10^{-4}$), and this again seems to account for the misleading appearance of $g_{\text{ind}}$ depending on prescore. Savinainnen's results show no prescore bias in using normalized gain $g_{\text{ind}}$.

### C. Example 4: Edward Little High School

In 2007 Phillips and both of us investigated whether SAT combined math and verbal scores could also be used as a predictor of normalized gain, and we found that they could

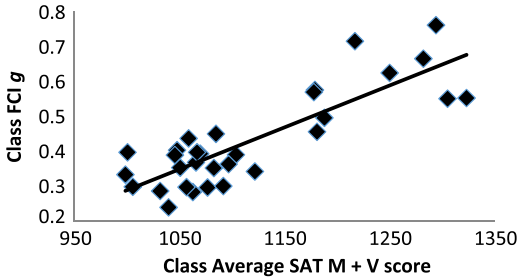


FIG. 2. Class average data for 31 classes at Edward Little High School with a total of 361 students ($r = 0.84$).

[13]. We reported correlations of $g_{\text{ind}}$ with both Lawson and SAT scores for LMU students and for students at Edward Little High School (ELHS) in Maine, where Steinert taught until 2006. The correlation coefficients between normalized gain $g_{\text{ind}}$ and SAT score were $r = 0.46$ for LMU ($N = 292$) and $r = 0.57$ for ELHS ($N = 335$).

An especially compelling indication of the effect of SAT score on conceptual learning in physics is provided by class average data from 31 honors and regular classes at ELHS from 1999 to 2006. Class average SAT scores ranged from about 1000 to about 1300 and corresponding FCI $g$ from below 0.3 to above 0.7 (Fig. 2). The correlation between $g$ and SAT score was exceptionally strong ($r = 0.84$, $p < 0.0001$). Normalized gain $g$ was also strongly correlated with prescore ($r = 0.64$, $p < 0.0001$). Apparently this correlation was a consequence of the strong correlation between prescore and SAT score ($r = 0.76$, $p < 0.0001$), as indicated by the fact that a multiple variable regression of $g$ on SAT *and* prescore yields the same correlation coefficient as the correlation between $g$ and SAT alone ($r = 0.84$), while the multiple variable regression indicated no significant correlation of $g$ with FCI prescore ($p = 0.99$). This shows once again no prescore bias for $g$.

The relationships we see between normalized FCI gain, FCI prescore, and SAT score at ELHS provide an explanation for why we see significant correlation between $g$ and prescore at many schools, where there is a wide range of SAT scores, but no correlation at all at Harvard (Fig. 1), where students all have very high SAT scores—with 75% of the SAT combined math and verbal scores above 1410.

### D. Example 5: Arizona School for the Arts

From 2007 to the present, FCI and Lawson data have been collected from 36 high school physics classes, taught by one of us at Arizona School for the Arts (ASA). Class average Lawson scores varied greatly and so did the values of class FCI $g$, which were very strongly correlated with Lawson scores, as seen in Fig. 3. FCI normalized gain $g$ was completely uncorrelated with FCI prescore (Fig. 4). This is yet another example showing no prescore bias in normalized gain $g$.

[3]For a brief description of multivariable regression, see the Appendix.

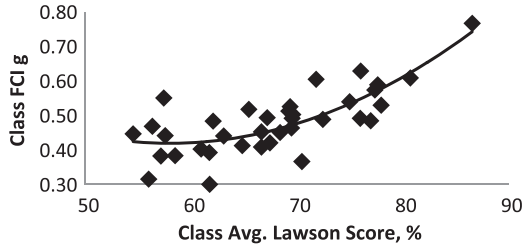


FIG. 3. Class average data for 36 classes at ASA, with a total of 803 students ($r = 0.78$). Here we have used a quadratic fit to the data, providing a better fit to the data than a linear function ($r = 0.78$ vs $r = 0.71$).

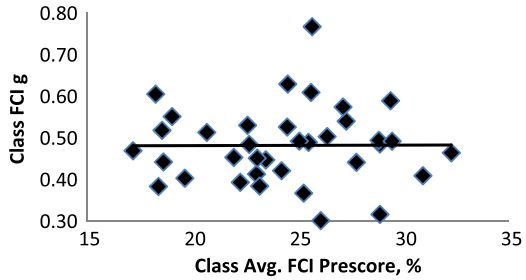


FIG. 4. Class average data for 36 classes at Arizona School for the Arts, with a total of 803 students ($r = 0.00$).

We also analyzed the relationships between individual FCI normalized gain $g_{\text{ind}}$, individual Lawson score, and individual FCI prescore for the 803 ASA students. Figures 5–7 show the graphs for pairs of these variables.

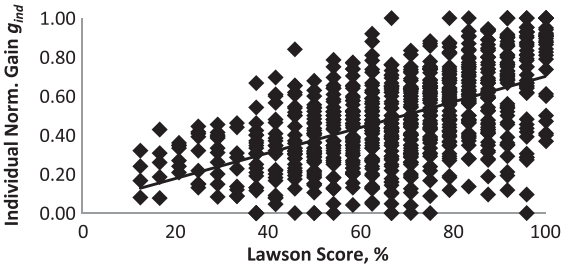


FIG. 5. Individual data for 803 students at Arizona School for the Arts, 2007–2018 ($r = 0.55$).

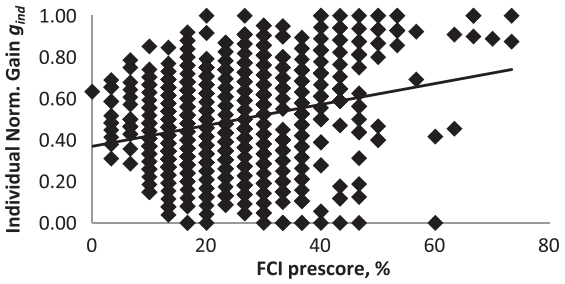


FIG. 6. Individual data for 803 students at Arizona School for the Arts, 2007–2018 ($r = 0.25$).

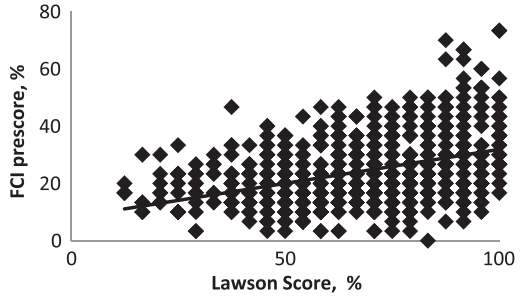


FIG. 7. Individual data for 803 students at Arizona School for the Arts, 2007–2018 ($r = 0.39$).

The strongest correlation is between FCI $g_{\text{ind}}$ and Lawson score ($r = 0.55$, $p < 10^{-4}$). The correlation between FCI $g_{\text{ind}}$ and FCI prescore ($r = 0.25, <10^{-4}$) appears to be a consequence of the correlation between FCI prescore and Lawson score ($r = 0.39$, $p < 10^{-4}$) because a multiple variable regression of FCI $g_{\text{ind}}$ on Lawson and FCI prescores yields the same $r$ of 0.55 as when Lawson score is the only independent variable, with a $p < 10^{-4}$ for Lawson score and a $p$ of 0.16 for FCI prescore, indicating no significant correlation. This example shows no prescore bias in individual normalized gain $g_{\text{ind}}$.

## III. NISSEN'S DATA; MISSING DATA

Nissen [2] *et al.* reported their own data for 89 courses, showing correlation between gain and prescore ($r = 0.43$). Their study did not include analysis of either SAT or Lawson data. In light of the examples above, without such analysis one should not conclude that the results demonstrate prescore bias. Their conclusion is based on omitted variable bias.

Another problem with their analysis is the large quantity of missing data. They rejected data from 27 other courses with a total enrollment of 1116 students, because in those classes more than 60% of the data were missing. However, in their analysis of the 89 classes, they had complete pretests and post-tests for only 2626 of the 4551 students enrolled in those classes. They considered the 42% missing data acceptable because in 4 other published studies that reported the rate of missing data, the average was 37% missing data. Nineteen studies they looked at did not report how much data might have been missing. We suspect that in many of those cases missing data were not reported because they were insignificant. That has certainly been the case with data we have reported in the past, including our 2005 and 2007 papers [5,13]. Missing data have been well under 5%. When missing data rise to the level of 40%, we are concerned that the data could well be unrepresentative of the class as a whole. Nissen uses multiple imputation software to replace missing data, but if there is a systematic bias to the missing data, the results could be quite different than what would have been

obtained by complete data. As Jelicic, Phelps, and Lerner [14] wrote, "the best solution to missing data is not to have any." When pretests and post-tests are given in class, with a small point incentive for taking the post-test, our experience has been that very few students will have their results missing from the dataset.

Even if there was no systematic error introduced by the large amount of missing data in Nissen's analysis, that analysis suffers from missing variable bias, and therefore one should not accept their claim that normalized gain is prescore biased.

## IV. NORMALIZED CHANGE

Nissen *et al.* cited the 2007 work [15] of Marx and Cummings as support for their claim of prescore bias, and so we shall address that work in this section. Marx and Cummings advocated for replacing individual normalized gain $g_{\text{ind}}$ by individual normalized change $c$, defined in terms of an individual's pretest and post-test scores:

$$c = \begin{cases} \frac{\text{postscore\%}-\text{prescore\%}}{100\%-\text{prescore\%}} & \text{if postscore} > \text{prescore} \\ \frac{\text{postscore\%}-\text{prescore\%}}{\text{prescore\%}} & \text{if postscore} < \text{prescore} \end{cases}.$$

Note that a different denominator is used when $c$ is negative. If prescore = postscore, then $c$ is set equal to zero, unless prescore and postscore both equal either 0 or 100, in which case the data point is omitted. Marx and Cummings advocate using the class average of students' individual $c$'s, $c_{\text{avg}}$, to report class results.

One argument they give for this new measure is that the calculation of $g_{\text{ind}}$ in some extreme cases can lead to strange results. We shall provide what we believe is a better way to deal with such cases. Marx and Cummings argue that $g_{\text{ind}}$ has a "low test score bias" because, for example, a prescore of 20% and a postscore of 0% leads to a negative gain of $-20/80 = -0.25$, whereas prescores and postscores of 80% and 0%, respectively, lead to a much greater magnitude negative gain of $-80/20 = -4$. Because higher prescores could in such cases lead to more negative gains than is possible with low prescores, they claim that $g_{\text{ind}}$ is biased in favor of lower prescores, a claim cited by Nissen *et al.* [2] as another reason for discontinuing use of normalized gain. It is doubtful that anyone has ever seen actual student data with such extreme numbers, but what is occasionally seen is student data in which some students miss one more question on the post-test than on the pretest, resulting in about a 3% drop in score. For a prescore of 80%, this would result in a $g_{\text{ind}}$ of $-3/20 = -0.14$, whereas for a prescore of 20%, this would result in a $g_{\text{ind}}$ of $-3/80 = -0.01$. We believe that a better way to deal with negative gains is to eliminate them by setting $g_{\text{ind}}$ equal to zero whenever the postscore is less than the prescore, so that individual normalized gains will always range between 0 and 1. This seems reasonable from a practical point of view because, while it is possible for a student to have learned nothing from a course, justifying a 0, it is hard to imagine a course (at least an IE course) in which someone actually knew less at the end of the course than at the beginning. A more likely interpretation of a lower postscore is that the student made some lucky guesses on the pretest, not that they knew less at the end of the class. Hence setting 0 as a minimum in computation of $g_{\text{ind}}$ seems reasonable. This seems to us a better method than using normalized change.

Marx and Cummings also consider the example of an individual's prescore being 100%, leading to a 0 in the denominator of the equation defining $g_{\text{ind}}$. This is a problem easily avoided by simply eliminating that data point. A student with a perfect prescore has no possibility of demonstrating learning by achieving a higher post-test score, and so it seems reasonable to delete that data point. Even if the prescore is not 100%, but is very high, there is little gain in score that can be achieved and little need to focus on students with such scores. It is reasonable to regard scores above 80% as an indication of strong Newtonian thinking, and we have long eliminated such scores from our student data as we indicated in our 2005 paper [5]. In most cases such scores are relatively rare. When scores above 80% are common, for example, at elite universities such as Stanford, the FCI is not a very useful test.

Nissen *et al.* argue in most of their paper that normalized gain is positively prescore biased, but in their consideration of the work of Marx and Cummings, they claim normalized gain is negatively prescore biased. In fact normalized gain is neither positively nor negatively prescore biased, as we have shown in Sec. II.

## V. NORMALIZED GAIN VS COHEN'S *d*

Nissen *et al.* argue for the use of Cohen's $d$ instead of normalized gain, where $d$ is defined as the difference between two means divided by the pooled standard deviation $s$. Nissen *et al.* proposed to use it to measure the effect size of the change in average scores on the same test, preinstruction to postinstruction:

$$d = \frac{av\,\text{postscore\%} - av\,\text{prescore\%}}{s}$$

We have already shown that the claim of prescore bias for normalized gain is not valid. We shall now show just how misleading use of Cohen's $d$ can be. Consider two classes, one with average pretest score of 20% and average post-test score of 40%, the other with pre and post average scores of 60% and 80%, respectively. Suppose for simplicity that each class has the same standard deviation $s$ for both prescores and postscores. Because both the percent gain, pre to post, and $s$ are the same, the values of $d$ for the two classes are identical, a value of 2.0 if $s$ happens to be

10%. In contrast, the average $g$'s for the two classes are quite different, 0.25 for the first class and 0.50 for the second. If the populations of the two classes are similar, say for example both with average SATs of 1200 and/or a Lawson average of 70%, this difference in $g$'s would indicate that one of the classes was significantly more effective than the other. The first class's normalized gain is consistent with results that are typically seen in traditional classes, while the second class's gain is typical of a class that makes effective use of interactive engagement methods, as shown by Hake [1]. Normalized gain has been used by many for the last 20 years to provide that valuable information, which is often used to guide instructors toward use of more effective IE methods, as it did one of us. That kind of revealing information is lost if one considers only Cohen's $d$.

## VI. DISCUSSION

We have provided evidence that normalized gain is not prescore biased. In doing so, it was necessary to consider other important predictors of gain. Failure to do so is an example of omitted variable bias.

These omitted variables must be considered in order to interpret the meaning of a class's normalized gain, either $g$ or $g_{\text{ind}}$. More specifically, in order to determine what the value of the gain might imply about the effectiveness of the pedagogy for the conceptual learning achieved by the students, it is necessary to consider the class's scores on either the Lawson test, SAT, or ACT,[4] because the reasoning abilities of the class, as reflected in these scores, may well be a stronger determinant of conceptual learning than the pedagogy that is used in the course. For example, as shown in Sec. II C (Fig. 2), Steinert's classes, which varied dramatically in reasoning abilities because some were honors classes and some were not, had dramatically different normalized FCI gains, even though he used the same methods to teach all classes.

[4]A recent study by John Stewart showed significant correlations between ACT scores and normalized gains on both the FMCE and CSEM [16].

## APPENDIX: MULTIPLE VARIABLE REGRESSION

Simple linear regression is an attempt to find a linear relationship between a single predictor variable and a response variable. Our goal in using it is to determine the extent to which variation in the response variable can be predicted by variation in the predictor variable. Multiple linear regression (MLR) is a statistical tool that allows one to simultaneously examine relationships between many different variables, to attempt to relate two or more predictor variables to a single response variable. In this context, we use MLR to propose a linear relationship between normalized gain, FCI prescore, and Lawson score:

$$g_{\text{ind}} = \beta_0 + \beta_1(\text{pre-FCI}) + \beta_2(\text{Lawson}),$$

which we then fit using various datasets. The standard practice is to estimate the regression coefficients (the betas) using the method of ordinary least squares (OLS): minimizing the sum of the squared differences between the *estimated* value of normalized gain and the normalized gain in the actual dataset. We indicate estimated parameters with a hat:

$$\hat{g}_{\text{ind}} = \hat{\beta}_0 + \hat{\beta}_1(\text{pre-FCI}) + \hat{\beta}_2(\text{Lawson}) + \epsilon.$$

The regression residual $\epsilon$ is the difference between $g$ and its estimate at each value of FCI prescore and Lawson score. The regression $r^2$, which ranges between 0 and 1, is the fraction of the sample variance of the response variable predicted by the predictor variables. The $p$ value for each of the coefficients is the probability that there is no relationship between that predictor variable and the response variable and that the computed coefficient is a result of random error.

[1] R. 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 (1998).
[2] J. M. Nissen, R. M. Talbot, A. N. Thompson, and B. Van Dusen, Comparison of normalized gain and Cohen's $d$ for analyzing gains on concept inventories, Phys. Rev. Phys. Educ. Res. **14**, 010115 (2018).
[3] D. Hestenes (private communication).
[4] J. H. Stock and M. W. Watson, *Introduction to Econometrics*, 2nd ed. (Pearson, Boston, 2007), p. 187.
[5] V. P. Coletta and J. A. Phillips, Interpreting FCI scores: Normalized gain, preinstruction scores, and scientific reasoning ability, Am. J. Phys. **73**, 1172 (2005).
[6] R. R. Hake, Relationship of individual student normalized learning gains in mechanics with gender, high-school physics, and pretest scores on mathematics and spatial visualization, in *Physics Education Research Conference, Boise, Idaho, 2002* (to be published).
[7] A. E. Lawson, The generality of hypothetico-deductive reasoning: Making scientific thinking explicit, Am. Biol. Teach. **62**, 482 (2000).

[8] M. A. Dubson and S. J. Pollock, Can the Lawson test predict student grades? AAPT Announcer **36**, 90 (2006).
[9] K. Diff and N. Tache, From FCI To CSEM To Lawson test: A report on data collected at a community college, AIP Conf Proc. **951**, 85 (2007).
[10] P. Pamela and J. Saul, Interpreting FCI normalized gain, pre-instruction scores, and scientific reasoning ability, AAPT Announcer **36**, 89 (2006).
[11] P. Nieminem, A. Savinainen, and J. Viiri, Relations between representational consistency, conceptual understanding of the force concept, and scientific reasoning, Phys. Rev. Phys. Educ. Res. **8**, 010123 (2012).
[12] B. A. Pyper, Changing scientific reasoning and conceptual understanding in college students, AIP Conf Proc. **1413**, 63 (2011).
[13] V. P. Coletta, J. A. Phillips, and J. J. Steinert, Interpreting force concept inventory scores: Normalized gain and SAT scores, Phys. Rev. Phys. Educ. Res. **3**, 010106 (2007).
[14] H. Jelicic, E. Phelps, and R. M. Lerner, Use of missing data methods in longitudinal studies: The persistence of bad practices in developmental psychology, Develop. Psych. **45**, 1195 (2009).
[15] J. D. Marx and K. Cummings, Normalized change, Am. J. Phys. **75**, 87 (2007).
[16] J. Stewart (private communication).