CBE—Life Sciences Education
Vol. 13, 41–48, Spring 2014

Research Methods

Is It the Intervention or the Students? Using Linear Regression to Control for Student Characteristics in Undergraduate STEM Education Research

Roddy Theobald* and Scott Freeman†

*Department of Statistics and †Department of Biology, University of Washington, Seattle, WA 98195-4322

Submitted July 19, 2013; Revised October 15, 2013; Accepted October 21, 2013
Monitoring Editor: Elisa Stone

Although researchers in undergraduate science, technology, engineering, and mathematics education are currently using several methods to analyze learning gains from pre- and posttest data, the most commonly used approaches have significant shortcomings. Chief among these is the inability to distinguish whether differences in learning gains are due to the effect of an instructional intervention or to differences in student characteristics when students cannot be assigned to control and treatment groups at random. Using pre- and posttest scores from an introductory biology course, we illustrate how the methods currently in wide use can lead to erroneous conclusions, and how multiple linear regression offers an effective framework for distinguishing the impact of an instructional intervention from the impact of student characteristics on test score gains. In general, we recommend that researchers always use student-level regression models that control for possible differences in student ability and preparation to estimate the effect of any nonrandomized instructional intervention on student performance.

INTRODUCTION

For the past several decades, discipline-based education researchers have focused on testing whether educational interventions in college science classrooms lead to improved student understanding and performance. Most interventions are given at the classroom level, meaning that all students in a given classroom receive the intervention. For example, all students in a class may be exposed to a new multimedia program (Aly *et al.*, 2004), asked to participate in reciprocal peer tutoring (Fantuzzo *et al.*, 1989), or taught in a workshop or studio format (Udovic *et al.*, 2002).

To evaluate the impact of educational interventions like these on student performance, researchers typically collect student test scores before and after the intervention—that is, from a pretest and a posttest. Although some researchers are interested in whether student scores improve after instruction (see Arwood, 2004; McConnell *et al.*, 2006; Nam and Ito, 2011), most are interested in demonstrating that student test scores improve more in treatment classrooms than in control classrooms—that is, in sections that do receive the intervention versus sections that do not.

What is the best way to analyze pre–post data in this setting? At least four different methods for determining whether learning gains differ in the treatment and control classrooms are commonly used in the science, technology, engineering, and mathematics (STEM) education literature: comparing 1) raw change scores (e.g., Udovic *et al.*, 2002); 2) normalized gain scores (Hake, 1998); 3) normalized change scores (Marx and Cummings, 2007); and 4) effect sizes (Andrews *et al.*, 2011). Unfortunately, none of these methods accounts for a fundamental problem: controlling for student equivalence, or lack thereof, in the classrooms being compared.

The problem of student nonequivalence is pervasive, because it is seldom possible to use randomization to control for differences in student ability or preparation (but see Fantuzzo *et al.*, 1989; Buzzell *et al.*, 2002; Aly *et al.*, 2004; Bilgin *et al.*, 2009). While nonrandomized designs are often unavoidable—it is very difficult to convince a registrar's

DOI: 10.1187/cbe-13-07-0136
Address correspondence to: Roddy Theobald (roddy@u.washington.edu).

office to randomly assign students to courses—they raise difficult questions about interpreting results. Namely, researchers who use the methods listed above have no way of knowing whether observed differences in learning gains between the treatment and control classes are due to the impact of the intervention itself or to differences between treatment and control classes—including the instructor, the instructional techniques used, and student characteristics—that are completely independent of the intervention.

In this paper, we use test score data from two sections of a college-level introductory biology course to illustrate how each of the four commonly used methods can lead to misleading conclusions. The two sections were taught by the same instructor, in the same term, using identical instructional techniques. However, due to a scheduling conflict during that term, the students enrolled in one of the sections had substantially better academic qualifications, on average, than students in the other section. We show that each of the four methods commonly used to assess educational interventions in college STEM classrooms would support the conclusion that an "instructional intervention" in the higher-performing section led to larger student learning gains, when in fact there was no intervention at all.

We propose a solution to the problem by introducing an approach that is ubiquitous in many other research areas but currently underused in the STEM education literature: multiple linear regression analysis. Specifically, we employ a student-level regression model that controls for observable differences between students in the treatment and control classes and demonstrate that it leads to the correct conclusion: differences in learning gains between the two sections are driven by differences in the composition of the students in the two sections, not by any intervention that was given in one section or the other. We argue that to estimate an unbiased intervention effect when analyzing data from nonrandom experimental designs, researchers must account for student background in a regression framework.

REVIEWING EXISTING METHODS FOR ANALYZING PRE–POST DATA

Before introducing regression approaches for analyzing pre–post data, we provide a brief review of the four approaches commonly used in the undergraduate STEM literature to analyze pre/posttest data and discuss some relative advantages and disadvantages of each. However, we stress than *none* of these four methods accounts for possible differences in the student composition of the treatment and control courses.

Raw Change Scores

Udovic *et al.* (2002) compare student learning gains in a "workshop" introductory biology course, which included numerous active-learning activities, with learning gains in comparison courses taught primarily through lectures. Like Dori *et al.* (2007), Fallahi (2008), and Linsey *et al.* (2007, 2009), Udovic and colleagues use a *t* test to compare what we refer to as "raw change scores" between treatment and control classes. Raw change scores are simply the difference between the postscore and the prescore. Udovic and coworkers contend that if student scores in the treatment course improve more, on average, from the pretest to the posttest than do student scores in the control course, then the gains must be due to the intervention in the treatment courses. This procedure is identical—meaning that it will result in the same *p* value and conclusions regarding the effect of the intervention—to the two-way repeated-measures analysis of variance (ANOVA) used by Martin *et al.* (2007).

In both the treatment and control classes, the authors compute the average raw change for each of the 11 questions on their pre- and posttest. The mean raw change was higher in the treatment classes than in the control classes for all 11 questions, and the *t* test rejected the null hypothesis that the mean raw change scores were equal for seven of the 11 questions. The authors conclude that the active-learning strategies in the treatment courses had a significant impact on student learning gains.

Analyzing raw change is attractive in terms of simplicity but does not account for the observation that students with low scores on a pretest have more to gain than students who score higher. The problem arises because test scores are bounded—meaning that they have an upper limit. To account for differences in "ease of improving" from pre to post, researchers have used two methods for standardizing or normalizing gain scores, one at the classroom level and one at the student level.

Normalized Gain Scores

Hake (1998) compares student learning gains on the Force Concept Inventory across 62 different introductory physics courses. In 48 of these courses, instructors had made substantial use of interactive-engagement methods. Hake considers these the treatment classes, while the 14 courses that were based on traditional lecturing are the control classes. For each class, Hake calculates the "average normalized gain," symbolized $<g>$, as the ratio of the average gain from pretest to posttest to the maximum possible gain $\langle g \rangle = \left(\frac{\text{post}-\text{pre}}{100-\text{pre}}\right)$, where pre- and postscores are expressed as the average percent correct in each class in the study. He reports that the average normalized gain in the treatment courses was 0.48 ± 0.14 SD, while the average normalized gain in the control courses was 0.23 ± 0.04 SD. Although he did not perform a formal statistical hypothesis test, he concludes that interactive-engagement methods have a significant positive impact on student learning gains. Had he performed a *t* test of the average normalized gains, the *p* value would have been <0.001—more than enough evidence to make the same conclusion.

Reporting $<g>$, the normalized learning gain, became popular in the undergraduate STEM education literature for several reasons. First, by normalizing by the maximum gain possible in each class, it accounts for the fact that some classrooms have more room to gain than other classrooms. A class that scores an 80% on the pretest and a 90% on the posttest has an average normalized gain of 0.5, matching a class that scores 60% on the pretest and 80% on the posttest (i.e., each class gained exactly half of the amount it could have gained on the posttest). Second, the size of Hake's initial study made it possible for researchers to compare learning gains informally across classrooms, even if their own study did not include enough classes to make a formal statistical test possible. That is, researchers could compute $<g>$ for one or a few classrooms under study and make a judgment about whether the values are similar to those reported in the Hake study (e.g., Knight and Wood, 2005; McDaniel *et al.*, 2007; Tanahoung *et al.*, 2009). Finally, in studies with large numbers of treatment and control classrooms, $<g>$ can be used to formally test whether learning gains in the treatment classes are larger than in the control classes (e.g., Redish and Steinberg, 1999; LoPresto and Murrell, 2009). However, using $<g>$ results in low sample sizes and thus poor statistical power, because it uses the class as the unit of analysis instead of using individual students.

Normalized Change Scores

Marx and Cummings (2007) created a student-level alternative to Hake's normalized gain measure called the normalized change score, symbolized *c*. Instead of computing learning gains at the classroom level with $<g>$, Marx and Cummings advocate calculating learning gains at the student level, using the following formula:

$$c = \begin{cases} \dfrac{\text{post} - \text{pre}}{100 - \text{pre}} & \text{if post} > \text{pre} \\ \text{drop} & \text{if post} = \text{pre} = 100 \text{ or } 0 \\ 0 & \text{if } 0 < \text{post} = \text{pre} < 100 \\ \dfrac{\text{post} - \text{pre}}{\text{pre}} & \text{if post} < \text{pre} \end{cases}$$

For students who score higher on the posttest than the pretest, the student-level normalized change score is computed similarly to a classroom-level normalized gain. The last three possibilities deal with unusual circumstances: students who score 0 or 100 on both the pre- and posttest are dropped; students who score the same on the pre- and posttest get a 0; and students who score lower on the posttest than the pretest have this negative gain scaled by the possible number of points they could have lost.

Because normalized change scores compare learning gains for students rather than for classrooms, they have two substantial advantages over normalized gain scores. First, they can be used to compare the impact of interventions assigned within classrooms rather than across classrooms (see Smith *et al.*, 2011, for an example). Second, because the observations are at the student level rather than at the classroom level, the sample size is substantially larger compared with using normalized change scores, providing increased statistical power.

Normalized change scores have an important limitation, however. If students get a perfect score on the posttest, their *c* is 1 no matter whether their prescore was 1% or 99%. Similarly, if students score the same on the pre- and posttest, their score is 0, no matter whether their prescore was 1% or 99%. In these cases, the goal of normalizing for "ease of improvement" is lost.

Effect Sizes

Andrews and colleagues (2011) collect pre/posttest score data on the conceptual inventory of natural selection from a sample of introductory biology courses around the United States, and compare learning gains from courses in which instructors used different numbers of active-learning exercises per week. To quantify learning gains at the classroom level, they use a metric known as an effect size. Effect sizes are commonly used in meta-analyses because they put estimated treatment effects from different studies in a common scale. For example, researchers can calculate a standardized mean difference, which expresses the difference between groups in units of SD, using Cohen's *d* statistic or a variant called Hedges' *g*. With pre–post data from identical assessments, it is appropriate to use a modification of Cohen's *d* that accounts for the same students being tested twice (see Becker, 1988; Dunlap *et al.*, 1996). Thus, Andrews *et al.* calculate the effect size for each class as $d = \frac{\bar{X}_{\text{post}} - \bar{X}_{\text{pre}}}{s_g/\sqrt{2(1-r)}}$, where $\bar{X}_{\text{post}}$ and $\bar{X}_{\text{pre}}$ are the average scores on the post- and pretest, s_g is the SD of the raw gain scores, and *r* is the correlation between student scores on the pre- and posttests. Andrews and colleagues estimate a classroom-level linear regression using each class's effect size as the dependent variable, and find that the number of active-learning exercises used per week has no relationship to student learning gains.

Andrews and colleagues' (2011) use of linear regression is an important addition to the undergraduate STEM education literature, as it allows them to control for factors other than active learning—such as the instructor's position and years of teaching experience, class size, and student-rated course difficulty—that could influence learning gains in the treatment and control classrooms. However, Andrews and colleagues estimate their regression at the classroom level and do not have access to student characteristics that can be used as control variables. A large K–12 literature (e.g., Rockoff 2004; Rivkin *et al.*, 2005) demonstrates that observable student characteristics are often correlated not just with student performance but also with student learning gains. Thus, this approach—like the prior three methods we reviewed—does not account for the possibility that differences in student learning gains, or lack thereof, are due to differences in the characteristics of students in the treatment and control classrooms rather than to the effect of the intervention.

CONTROLLING FOR STUDENT NONEQUIVALENCE: THE PROBLEM

To illustrate the importance of controlling for observable student characteristics in the treatment and control classes when evaluating the impact of nonrandomized educational interventions, we apply the four methods above to pre- and posttest scores from two sections of an introductory biology course offered during the Summer of 2012 at the University of Washington. Each section was taught by the same instructor using the exact same materials and instructional strategies. Thus, without knowing anything about the student composition of the two classes, there is no a priori reason to expect different student performance in the two classes. Given that there is actually no treatment at all, this should be an example of a statistical test wherein the null hypothesis—that the treatment had no impact on student learning gains—should *not* be rejected. We will label one of these sections as the treatment (section A) and the other section as the control (section B).

We will demonstrate that each of the methods above *does* lead to the conclusion that, for this particular pair of sections in this particular course, learning gains in the treatment class are higher than in the control class. This would ordinarily be taken as evidence that the "instructional intervention" in the treatment class had a significant impact on student learning gains. But given that there was no intervention at all, we explore whether the student composition of these two particular sections may have contributed to the incorrect conclusion. Throughout the analysis that follows, we interpret the results of all tests of statistical significance at the 90% confidence level—meaning that the p value must be <0.1 to reject the null hypothesis. We caution, however, against overreliance on conventional levels of statistical significance.

Data Overview

At the start of the term, students in each section took a diagnostic test (Shi *et al.*, 2010), converted to a 100-point scale, that was intended to measure their prior knowledge about the topics to be covered in the course. Then 2 wk into the course, students took an in-class exam on the same material—also graded on a 100-point scale—that covered material taught in the first 2 wk of the course. We will treat the diagnostic test as the pretest and the in-class exam as the posttest. The average pretest scores were 59.8 (SD = 18.1) in section A (the "treatment" section) and 59.3 (SD = 17.0) in section B (the "control" section), and are not significantly different between the two sections (the p value from a two-sample t test is 0.865). This is important because many authors (e.g., McDaniel *et al.*, 2007) assume that the treatment and control classes have similar incoming characteristics if the pretest scores are not significantly different. The average posttest scores were 72.0 (SD = 15.8) in the treatment section and 67.0 (SD = 15.0) in the control section, which a t test indicates is significantly different ($p = 0.050$). We now analyze these data using the four methods discussed in the preceding section.

Comparison of Raw Change Scores

In the treatment class, the average raw change score is 72.0 – 59.8 = 12.2 (SD = 15.0), while in the control class, the average raw change score is 67.0 – 59.3 = 7.7 (SD = 15.8). A t test of the null hypothesis that these average raw change scores are the same gives a p value of 0.077, which is statistically significant at the 90% confidence level. Thus, with this methodology, there is sufficient evidence to reject the null hypothesis and conclude that student learning gains were greater in the treatment class than in the control class.

Normalized Gain Scores

The normalized gain score in the treatment class is $\frac{72.0-59.8}{100-59.8} = 0.30$, while the normalized gain in the control class is $\frac{72.0-59.8}{100-59.8} = 0.19$. Because we are limited to only one treatment class and one control class, there is no way to statistically test whether these normalized gain scores are significantly different. That said, the magnitude of the difference may lead to the conclusion that learning gains were greater in the treatment class than in the control class.

Normalized Change Scores

The average normalized change score in the treatment class is 0.31 (SD = 0.29), while the average normalized change score in the control class is 0.19 (SD = 0.29). A t test of the null hypothesis that these average normalized changes scores are the same gives a p value of 0.012, which is statistically significant. Thus, with this methodology, there is sufficient evidence to reject the null hypothesis and conclude that student learning gains were greater in the treatment class than in the control class.

Effect Sizes

The correlation between student scores on the pre- and posttests is $r = 0.56$. Thus, the effect size for the treatment class is $\frac{12.2}{15.0/\sqrt{2(1-0.56)}} = 0.76$, while the effect size for the control class is $\frac{7.7}{15.8/\sqrt{2(1-0.56)}} = 0.46$. As with normalized gains, there is no way to test whether these effect sizes are significantly different with only one treatment and one control class. That said, the magnitude of the difference between the two classes may lead to the conclusion that learning gains in the treatment class were larger than learning gains in the control class.

Potential Explanation

Each of the above methods could lead to the conclusion that the intervention in the treatment class had a significant positive impact on student learning gains. But given that there was no intervention at all, there must be another explanation for the observed difference in learning gains. One possibility is that the differences occurred by chance. The p value for a t test comparing normalized change scores, for example, tells us that there is a 1.2% chance of observing differences this extreme by chance alone. Another more probable explanation, though, is that the student composition of the two sections is driving the differences.

To investigate this hypothesis, we collected data on two measures that should reflect student ability and preparation: incoming undergraduate grade point average (GPA) and final grade in the preceding course in the introductory biology sequence. Due to a scheduling conflict during this particular term, the two sections had substantially different incoming performance levels. Specifically, the average incoming GPA in the treatment class was 3.33 (SD = 0.42), which a t test shows is significantly higher ($p < 0.001$) than the average incoming GPA in the control class, 3.04 (SD = 0.43). Likewise, the prior biology grade averaged 3.09 (SD = 0.76) in the treatment class, which a t test indicates is significantly higher ($p < 0.001$) than the average prior biology grade in the control class, 2.69 (SD = 0.66).

This observation underlines the central message of this paper. The gold standard for evaluating the impact of treatments of any kind—educational or otherwise—is a large randomized controlled trial. If sample sizes are large and if treatments are randomly assigned to the experimental subjects (or students, in this case), then there is no reason to expect the treatment and control groups to differ in any way, except that the treatment group received the treatment, while the control group did not. But in the context of evaluating the impact of interventions in undergraduate STEM classrooms, it is often not feasible to randomly assign students to treatment and control classes. The nonrandomized design that results opens up the possibility that the treatment and control classes will be substantially different, as our example shows. If these differences are correlated with student learning gains, then any of the methods above runs the risk of attributing observed differences to the impact of the treatment, when in reality they are due to differences in the composition of the groups being compared. This is true even if the treatment had no effect at all.

Given that many interventions in undergraduate STEM education cannot be randomized, is there a way to distinguish the impact of incoming student characteristics from the impact of the intervention itself? We argue that the answer is often *yes*, and that multiple linear regression can be a useful tool in any such analysis. We introduce this methodology in the next section, apply it to our data, and demonstrate that it leads to the correct conclusion: controlling for incoming student characteristics, there is no statistically significant difference in learning gains between the treatment and control classes in our example. Our goal in the following section is not to provide a rigorous theory of linear regression, but rather to motivate its use for evaluating the impact of educational interventions on student learning gains. We refer interested readers to chapters 3 and 4 of Gelman and Hill (2007) for an accessible discussion of broader considerations in linear regression.

CONTROLLING FOR STUDENT NONEQUIVALENCE: A SOLUTION

It is intuitive to think of a student's performance on a test as a function of many factors: the student's prior knowledge about the specific topics on the test, the student's understanding of the larger discipline, the student's work habits and study skills, and the intervention itself. A linear regression model formalizes this intuition by assuming that an outcome or dependent variable (in this case, a student's score on the posttest) is a linear function of explanatory (or control) variables and the intervention itself. Linear regression is not the only methodology that allows for this framework, but we will restrict our attention to it for simplicity.

A key step in a linear regression analysis is collecting data about control variables—measurements that can serve as proxies for factors that may influence the outcome variable, other than the treatment of interest. In a pre- and posttest setting, each student's score on the pretest is one obvious control variable, as the prescore controls for each student's prior knowledge about the specific topics on the test. In our example, we also collected data on each student's undergraduate GPA and grade in a previous biology class. The latter may be a reasonable proxy for each student's understanding of the broader field of biology, while both measures provide some information about each student's work habits and study skills.

Undergraduate GPA and previous biology grade are certainly not the only variables we could select to control for variation in student preparation and ability. In fact, researchers often have access to more student-level variables than are practical to use. Procedures like stepwise regression can assist researchers in selecting control variables that are most predictive of the outcome variable (see Freeman *et al.*, 2007, for an example). Researchers can also use professional judgment—based on the available variables, data from similar studies in the literature, and their own experience—in selecting control variables. We chose a measure of overall academic performance (undergraduate GPA) and a measure of performance specific to biology (previous biology grade). Although the correlation between these variables is high ($r = 0.84$), we chose both to account for the possibility that they capture different dimensions of student academic background—a decision that is borne out in the results in the next section. We also note that if we had data on many classrooms taught at different times by different instructors, we would also consider controlling for indicators such as time-of-day and class instructor, if the data suggested these indicators were relevant to the outcome being measured.

With these considerations in mind, we define the following variables for each student: X_{post} is the student's score on the posttest; X_{pre} is the student's score on the pretest; GPA is the student's undergraduate GPA; Grade is the student's grade in the introductory course; and Treatment is an indicator for whether the student was in the class that received the intervention (Treatment = 1 if the student is in the treatment class, Treatment = 0 if the student is in the control class). We recommend centering each of the control variables (in this case, Pre, GPA, and Grade) by subtracting the mean of each variable from each student's value (see Gelman and Hill, 2007, sections 4.1 and 4.2). One possible linear regression model that uses these variables is the following:

$$\begin{aligned} X_{\text{post}} &= \beta_0 + \beta_1 \times X_{\text{pre}} + \beta_2 \times \text{GPA} \\ &\quad + \beta_3 \times \text{Grade} + \beta_4 \times \text{Treatment} + \varepsilon \qquad (1) \end{aligned}$$

The βs in Eq. 1 are regression coefficients that describe the relationship between each variable and the student's postscore:

- β_0 is the intercept, or the expected postscore for a student with an average prescore, GPA, and prior grade, who did not receive the treatment (note that if the control variables are not centered, β_0 is the expected postscore for a student with prescore of 0, GPA of 0, etc., which are not meaningful values);
- β_1 is the expected increase in the postscore for each additional point on the student's prescore;
- β_2 is the expected increase in the postscore for each additional GPA point;
- β_3 is the expected increase in the postscore for each additional grade point from the previous course; and
- β_4 is the expected increase in the postscore for students who received the intervention relative to students who did not receive the intervention.

In contrast to approaches like type I ANOVA that estimate the effect of each variable sequentially, a linear regression estimates each of these coefficients simultaneously. Thus, each of these regression coefficients should be interpreted as "all else equal," meaning that they represent the marginal effect of changing one variable *while holding all the other variables constant*. The error term ε captures the reality that the regression equation does not perfectly predict each student's postscore.

Statistical software packages provide an estimate for each regression coefficient and the p value from the t test of the null hypothesis that the coefficient equals zero. For example, consider the coefficient of interest in Eq. 1: β_4, or the "treatment effect" for students who received the intervention relative to students who did not receive the intervention.

Table 1. Estimated regression coefficients from linear regression Eq. 1

Coefficient	Estimate	SE	p Value from t test[a]
Intercept ($\hat{\beta}_0$)	69.61	1.31	<0.0001***
Prescore ($\hat{\beta}_1$)	0.28	0.06	<0.0001***
GPA ($\hat{\beta}_2$)	12.31	3.82	0.0016**
Grade ($\hat{\beta}_3$)	4.37	2.30	0.0595$^+$
Treatment ($\hat{\beta}_4$)	−0.42	1.91	0.8272

[a] Significance levels from two-sided t test: $^+$, $p < 0.1$; *, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$.

The null hypothesis is that this coefficient equals zero—that is, the intervention had no effect. Linear regression provides both an estimate of this treatment effect and a test of whether the treatment effect really is significantly different from zero, *controlling for the influence of each of the other variables in the model.* Note that if differences in learning gains between the treatment and control classes can be explained by the control variables and not by the intervention itself, then the treatment effect should not be significantly different from zero. On the other hand, if the intervention does have a significant impact on student performance, the null hypothesis should be rejected, and (if the regression model is correctly specified) the estimated treatment effect should quantify the average effect of the intervention on student test scores.

Linear regression makes some important assumptions. While it is beyond the scope of this paper to discuss all of them in depth (see Gelman and Hill, 2007, section 3.6, for more details), there are a few that are particularly important for the present application. The first is that the error term ε is normally distributed. This assumption can be problematic if the maximum score on the test creates a "ceiling effect" that artificially limits the scores of the best students in the class. In this situation, these students will consistently score lower than the model predicts, because there is a violation of the normality assumption. Another assumption is that the influence of the control variables truly is linear. There is no compelling reason, other than mathematical convenience, to assume that the influence of a student's prescore, GPA, and prior grades on his or her postscore is truly additive as opposed to multiplicative or otherwise nonlinear.

These assumptions are important, and there are many methods to test and relax them (see Gelman and Hill, 2007, chapters 3–6). Here, however, we focus on standard linear regression.

DATA ANALYSIS USING MULTIPLE LINEAR REGRESSION

We now return to the introductory biology data and illustrate that linear regression leads to the correct conclusion that there is no evidence that the "intervention" has a significant impact on student performance, controlling for other differences between students in the treatment and control classes. We estimate the linear regression equation specified in Eq. 1, and report the estimated coefficients $\hat{\beta}_k$ in Table 1. (Note that the "hat" over each regression coefficient indicates that it is an estimate of the parameter β_k in Eq. 1.)

These estimates can be interpreted as follows.

- The estimated intercept ($\hat{\beta}_0 = 69.61$) means that the expected score on the posttest for a student with an average prescore, average GPA, and average prior grade, who did not receive the intervention is 69.61.
- The coefficient on the prescore ($\hat{\beta}_1 = 0.28$) means that we expect a student's score on the posttest to increase by 0.28 points for each additional point the student scores on the pretest, all else equal (i.e., holding all the other variables constant).
- The coefficient on GPA ($\hat{\beta}_2 = 12.31$) means that we expect a student's score on the posttest to increase by 12.31 points for each additional point in the student's GPA, all else equal.
- Finally, the coefficient on the prior grade ($\hat{\beta}_3 = 4.37$) means that we expect a student's score on the posttest to increase by 4.37 points for each additional point in the student's grade from the previous biology course, all else equal.

For each coefficient, the null hypothesis that the coefficient equals zero is rejected at the 90% confidence level, so we have sufficient evidence to conclude that each of these control variables has an independent, significant correlation with student performance on the posttest. This is extremely important, as it means that even when controlling for a student's score on the pretest, a student's GPA and prior grades are still predictive of his or her score on the posttest. This may be due to students with higher GPAs having better study skills and work habits, and therefore preparing more effectively for the posttest. As in many studies, the posttest in our example was announced on the syllabus and awarded course points, while the pretest was not—a situation that may increase the impact of differences in motivation or preparation. Alternatively, it is possible that students who received a better grade in the previous biology course have a better understanding of the broader discipline, which helped them prepare for and answer questions on the posttest.

Finally, the p value of the estimate $\hat{\beta}_4$—from the t test of the null hypothesis of no treatment effect—is 0.827, which means that there is not nearly enough evidence to reject the null hypothesis that the "intervention" is significantly correlated with student performance, controlling for the influence of other student-level characteristics. Given that there was no intervention in the treatment class at all, it is reassuring that the linear regression model leads to this conclusion. This reinforces our central message: It is essential to control for potential student nonequivalence between the treatment and control groups when evaluating the impact of a nonrandomized educational intervention.

TOWARD INTEGRATION OF LINEAR REGRESSION IN UNDERGRADUATE STEM RESEARCH

We have shown that existing methods of evaluating interventions in college science classrooms can lead to erroneous conclusions when the interventions are not randomly assigned to students, and that linear regression can help mitigate this problem by controlling for observable characteristics that are also correlated with student learning gains. We caution that estimates from a linear regression do *not* justify a causal interpretation, except under strict assumptions, and that randomization of the intervention is still the best way to establish a treatment effect.

In our motivating example, we have used a multiple linear regression model to illustrate the simplicity and utility of a regression framework. However, there are many other reasons that education researchers should be drawn to this framework. Although we choose not to control for gender and ethnicity in our regression model, regression can also be used to test whether women, minorities, or any other affinity group are gaining more or less in our classrooms, all else equal. Regression models can also include interaction terms that test whether the intervention has a differential impact on different types of students. Researchers who currently use normalized change scores can simply use these values as the outcome variable in a linear regression. (When doing so, though, we recommend not including prescore as a predictor variable, as prescore is already included in normalized change.) Finally, while it is beyond the scope of this paper to discuss more complex regression methods, an even more rigorous approach could use generalized linear models to model nonlinear relationships between student characteristics and test scores, analyze student responses at the individual-question level, or produce unbiased estimates in the presence of a ceiling effect.

The undergraduate STEM education literature has made remarkable strides in recent years, but the methods commonly used to estimate the impact of instructional interventions lead to troubling questions about whether these treatment effects really are due to the interventions. It is possible that none of the results in the studies we reviewed would have changed if the researchers had controlled for student characteristics in a regression framework, but we hope we have illustrated that linear regression should be a component of any analysis of a nonrandomized instructional intervention. It is time for this growing literature to take the next step and ensure that reported treatment effects are the result of the intervention itself, not the students.

ACKNOWLEDGMENTS

We thank Ben Wiggins for discussion and for access to the pre–post testing data on introductory biology students, and Sara Brownell, Sarah Eddy, Elli Theobald, Mary Pat Wenderoth, and two anonymous reviewers for comments that improved the manuscript. The study was funded by a grant from the National Science Foundation (DUE-0942215) and conducted under Human Subjects Division application #44438.

REFERENCES

Aly M, Elen J, Willems G (2004). Instructional multimedia program versus standard lecture: a comparison of two methods for teaching the undergraduate orthodontic curriculum. Eur J Dent Educ *8*, 43–46.

Andrews TM, Leonard MJ, Colgrove CA, Kalinowski ST (2011). Active learning *not* associated with student learning in a random sample of college biology courses. CBE Life Sci Educ *10*, 394–405.

Arwood L (2004). Teaching cell biology to nonscience majors through forensics, or how to design a killer course. Cell Biol Educ *3*, 131–138.

Becker BJ (1988). Synthesizing standardized mean-change measures. Br J Math Stat Psych *41*, 257–278.

Bilgin I, Åženocak E, Sözbilir M (2009). The effects of problem-based learning instruction on university students' performance of conceptual and quantitative problems in gas concepts. Eurasia J Math Sci Tech Ed *5*, 153–164.

Buzzell PR, Chamberlain VM, Pintauro SJ (2002). The effectiveness of web-based, multimedia tutorials for teaching methods of human body composition analysis. Adv Physiol Ed *26*, 21–29.

Dori YJ, Hult E, Breslow L, Belcher JW (2007). How much have they retained? Making unseen concepts seen in a freshman electromagnetism course at MIT. J Sci Ed Tech *16*, 299–323.

Dunlap WP, Cortina JM, Vaslow JB, Burke MJ (1996). Meta-analyses of experiments with matched groups or repeated measures designs. Psychol Methods *1*, 170–177.

Fallahi CR (2008). Redesign of a life span development course using Fink's taxonomy. Teach Psychol *35*, 169–175.

Fantuzzo JW, Dimeff LA, Fox SL (1989). Reciprocal peer tutoring: a multimodal assessment of effectiveness with college students. Teach Psychol *16*, 133–135.

Freeman S, O'Connor E, Parks JW, Cunningham M, Hurley D, Haak D, Dirks C, Wenderoth MP (2007). Prescribed active learning increases performance in introductory biology. CBE Life Sci Educ *6*, 132–139.

Gelman A, Hill J (2007). Data Analysis Using Regression and Multilevel/Hierarchical Models, New York: Cambridge University Press.

Hake RR (1998). Interactive-engagement versus traditional methods: a six-thousand-student survey of mechanics test data for introductory physics courses. Am J Phys *66*, 64–74.

Knight JK, Wood WB (2005). Teaching more by lecturing less. Cell Biol Educ *4*, 298–310.

Linsey J, Talley A, White C, Jensen D, Wood K (2009). From Tootsie Rolls to broken bones: an innovative approach for active learning in mechanics of materials. Adv Eng Educ *1*, 1–23.

LoPresto MC, Murrell SR (2009). Using the Star Properties Concept Inventory to compare instruction with lecture tutorials to traditional lectures. Astron Educ Rev *8*, 010105.

Martin T, Rivale SD, Diller KR (2007). Comparison of student learning in challenge-based and traditional instruction in biomedical engineering. Ann Biomed Eng *35*, 1312–1323.

Marx JD, Cummings K (2007). Normalized change. Am J Phys *75*, 87–91.

McConnell DA *et al.* (2006). Using conceptests to assess and improve student conceptual understanding in introductory geoscience courses. J Geosci Educ *54*, 61–68.

McDaniel CN, Lister BC, Hanna MH, Roy H (2007). Increased learning observed in redesigned introductory biology course that employed web-enhanced, interactive pedagogy. CBE Life Sci Educ *6*, 243–9.

Nam Y, Ito E (2011). A climate change course for undergraduate students. J Geosci Educ *59*, 229–241.

Redish EF, Steinberg RN (1999). Teaching physics: figuring out what works. Phys Today *52*, 24–30.

Rivkin SG, Hanushek EA, Kain JF (2005). Teachers, schools, and academic achievement. Econometrica *73*, 417–458.

Rockoff J (2004). The impact of individual teachers on student achievement: evidence from panel data. Am Econ Rev *94*, 247–252.

Shi J, Wood WB, Martin JM, Guild NA, Vicens Q, Knight JK (2010). A diagnostic assessment for introductory molecular and cell biology. CBE Life Sci Educ *9*, 453–461.

Smith MK, Wood WB, Krauter K, Knight JK (2011). Combining peer discussion with instructor explanation increases student learning from in-class concept questions. CBE Life Sci Educ *10*, 55–63.

Tanahoung C, Chitaree R, Soankwan C, Sharma MD, Johnston ID (2009). The effect of interactive lecture demonstrations on students' understanding of heat and temperature: a study from Thailand. Res Sci Tech Educ *27*, 61–74.

Udovic D, Morris D, Dickman A, Postlethwait J, Wetherwax P (2002). Workshop Biology: demonstrating the effectiveness of active learning in an introductory biology course. BioScience *52*, 272–281.