Statisticians: When We Teach, We Don't Practice What We Preach

Andrew Gelman and Eric Loken

Here's an example of inconsistent behavior. A statistician challenges a speeding ticket in court by arguing that the radar evidence was inconclusive and questioning whether the instrument was properly calibrated. Later in the day, the statistician watches a baseball game and is impressed with pitches clocked at 99 or 100 miles per hour. On the one hand, the statistician believes the TV network's radar gun can accurately peg, to the nearest mile per hour (mph), the speed of a small ball that is only visible for about 1 second. On the other, the statistician claims a police officer can't prove an SUV was traveling 20 mph over the speed limit on an open stretch of road.

We want to explore another example of inconsistent behavior that's far more consequential. As statisticians, we give firm guidance in our consulting and research on the virtues of random sampling, randomized treatment assignments, valid and reliable measurements, and clear specification of the statistical procedures that will be applied to data. With self-assured confidence that we occupy the moral high ground, we share horror stories about convenience samples, selection bias, multiple comparisons, and other problems that arise when those less enlightened about proper methodology don't follow the rules.

But are we really consistent in all aspects of our professional lives? How do we approach teaching? The following generalizations apply to most of us:

> We assign grades based on exams that would almost surely be revealed to be low in both reliability and validity if we were to ever actually examine their psychometric properties. Despite teaching the same courses year after year, we rarely use standardized tests.
>
> We almost never use pretests at the beginning of the semester, either to adjust for differences between students in different sections of a course or even for the more direct goal of assessing what has actually been learned by students in our classes.
>
> We evaluate teachers based on student evaluations which, in addition to all their problems as measuring instruments, are presumably subject to huge nonresponse biases. Would we tolerate client satisfaction surveys as the only measure of hospital quality?
>
> We try out new ideas haphazardly. Not only do we not do randomized experiments, we generally do not perform any systematic comparisons of treatments at all. As one high-level administrator put it to us recently, "It would be good if we introduced our new teaching methods based on something more than a 'hunch'."

The statistical field of quality control emphasizes the process of monitoring and improving a system, rather than focusing on individual cases. When we teach, however, we tend to focus on what seems to work or not work in an individual course, rather than on improving the process or the sequence. Consider how entrenched the freshman science sequence is at many large universities.

The contradiction is especially clear because we actually teach the stuff we believe in our classes and expect the students to parrot it back. However, we do not, in general, conduct our classes in a manner consistent with the principles we teach.

How would we seek to evaluate and improve statistics teaching—if we were following the advice we routinely give to our students and scientific collaborators?

We would clearly define our "treatment"—our teaching method. To the extent the treatment has natural variation, we would measure that variation. We would design alternative treatments ahead of time, ideally based on published research, and use pilot studies to tune any new teaching idea before trying it out on a live class.

We would assign different treatments at random to different groups of students and, after a class is over, compare student learning in treatment and control groups.

We would give students a pretest at the beginning of the term to adjust for differences between groups and to help study information in students dropping out of a class or switching sections.

What does this all have to do with ethics? We think of this as an ethical issue because it is an inconsistent application of best practices to all facets of professional service. If medical treatments, say, were decided on the spot by practitioners with no systematic evaluations, no controlled experimentation, no sampling to learn about populations, and no attempt at quality control, we as statisticians would consider this an ethical failing, or at the very least a level of public health malpractice revealing serious ignorance of research design.

Medicine is important, but so is education. To the extent that we believe the general advice we give to researchers, the unsystematic nature of our educational efforts indicates a serious ethical lapse on our part, and we can hardly claim ignorance as a defense. Conversely, if we don't really believe all that stuff about sampling, experimentation, and measurement—it's just wisdom we offer to others—then we're nothing but cheeseburger-snarfing diet gurus who are unethical in charging for advice we wouldn't ourselves follow.

Our view is that we do believe in our message of statistical data collection and analysis, but that when it comes to our own classes, we never quite feel like we have the time to do it right. Perhaps one failing of our statistical education is that it emphasizes clean solutions (simple random samples, experiments with perfect compliance, and precisely specified statistical models). When the ideal procedure isn't possible, we are all too ready to see potential criticisms, and thus we can be inclined to entirely avoid systematic data collection. But we suspect this is a mistake—a random sample with nonresponse is generally better than a haphazard sample with no probability selection at all, a broken randomized experiment is likely better than a retrospective comparison, and a flawed measurement is better than taking no measurement at all.

Being empirical about teaching is hard. Lack of incentives aside, we feel like we move from case study to case study as college instructors and that our teaching is a multifaceted craft difficult to decompose into discrete malleable elements. But even those of us who work with smaller classes can take systematic measurements of what our students knew before and after each class and we can roll out innovations more carefully to assess their effectiveness. Even if variation is high enough and sample sizes low enough that not much could be concluded, we suspect that the very acts of measurement, sampling, and experimentation would ultimately be a time-efficient way of improving our classes.

Those of us working in larger settings with hundreds or thousands of students each year can carry out item analysis and measurement evaluation. Testing in large classes generates masses of data that typically go unanalyzed. If we wanted to optimize our service to our profession, we would apply our own advice to improve student achievement, engagement, and retention. Being more empirical in evaluating our teaching efforts could yield happier and better students—and more of them. At the very least, we'd have a clear conscience about giving it our best shot.

In making our practice more research-based and our teaching more practically focused, it would make sense to involve the entire educational team, including members of college and university administrations who set curricula, permanent faculty who organize courses, adjuncts and teaching assistants who perform much of the grading and face-to-face teaching, and writers of textbooks and educational materials.

About the Authors

Andrew Gelman is a professor of statistics and political science and director of the Applied Statistics Center at Columbia University. He has received many awards, including the Outstanding Statistical Application Award from the American Statistical Association and the award for best article published in the *American Political Science Review*. He has coauthored many books; his most recent is *Red State, Blue State, Rich State, Poor State: Why Americans Vote the Way They Do*.

Eric Loken is a research associate professor of human development at Penn State, where he teaches graduate courses in regression and measurement. He studies latent variable models with applications in health and education and has co-founded two web-based assessment companies. The most recent is Criteria Corp, a pre-employment testing company.