EDUCATION WEEK
Published Online: January 15, 2013
Published in Print: January 16, 2013, as **Putting Value-Added Evaluation to the (Scientific) Test**

Probing the Science of Value-Added Evaluation

By R. Barker Bausell

Back to Story

Value-added teacher evaluation has been extensively criticized and strongly defended, but less frequently examined from a dispassionate scientific perspective. Among the value-added movement's most fervent advocates is a respected scientific school of thought that believes reliable causal conclusions can be teased out of huge data sets by economists or statisticians using sophisticated statistical models that control for extraneous factors.

Another scientific school of thought, especially prevalent in medical research, holds that the most reliable method for arriving at defensible causal conclusions involves conducting randomized controlled trials, or RCTs, in which (a) individuals are premeasured on an outcome, (b) randomly assigned to receive different treatments, and (c) measured again to ascertain if changes in the outcome differed based upon the treatments received.

The purpose of this brief essay is not to argue the pros and cons of the two approaches, but to frame value-added teacher evaluation from the latter, experimental perspective. For conceptually, what else is an evaluation of perhaps 500 4th grade teachers in a moderate-size urban school district but 500 high-stakes individual experiments? Are not students premeasured, assigned to receive a particular intervention (the teacher), and measured again to see which teachers were the more (or less) efficacious?

Granted, a number of structural differences exist between a medical randomized controlled trial and a districtwide value-added teacher evaluation. Medical trials normally employ only one intervention instead of 500, but the basic logic is the same. Each medical RCT is also privy to its own comparison group, while individual teachers share a common one (consisting of the entire district's average 4th grade results).

From a methodological perspective, however, both medical and teacher-evaluation trials are designed to generate causal conclusions: namely, that the intervention was statistically superior to the comparison group, statistically inferior, or just the

"A value-added analysis constitutes a series of personal, high-

same. But a degree in statistics shouldn't be required to recognize that an individual medical experiment is designed to produce a more defensible causal conclusion than the collected assortment of 500 teacher-evaluation experiments.

stakes experiments conducted under extremely uncontrolled conditions."

How? Let us count the ways:

• Random assignment is considered the gold standard in medical research because it helps to ensure that the participants in different experimental groups are initially equivalent and therefore have the same propensity to change relative to a specified variable. In controlled clinical trials, the process involves a rigidly prescribed computerized procedure whereby every participant is afforded an equal chance of receiving any given treatment. Public school students cannot be randomly assigned to teachers between schools for logistical reasons and are seldom if ever truly randomly assigned within schools because of (a) individual parent requests for a given teacher; (b) professional judgments regarding which teachers might benefit certain types of students; (c) grouping of classrooms by ability level; and (d) other, often unknown, possibly idiosyncratic reasons. Suffice it to say that no medical trial would ever be published in any reputable journal (or reputable newspaper) which assigned its patients in the haphazard manner in which students are assigned to teachers at the beginning of a school year.

• Medical experiments are designed to purposefully minimize the occurrence of extraneous events that might potentially influence changes on the outcome variable. (In drug trials, for example, it is customary to ensure that only the experimental drug is received by the intervention group, only the placebo is received by the comparison group, and no auxiliary treatments are received by either.) However, no comparable procedural control is attempted in a value-added teacher-evaluation experiment (either for the current year or for prior student performance) so any student assigned to any teacher can receive auxiliary tutoring, be helped at home, team-taught, or subjected to any number of naturally occurring positive or disruptive learning experiences.

• When medical trials are reported in the scientific literature, their statistical analysis involves only the patients assigned to an intervention and its comparison group (which could quite conceivably constitute a comparison between two groups of 30 individuals). This means that statistical significance is computed to facilitate a single causal conclusion based upon a total of 60 observations. The statistical analyses reported for a teacher evaluation, on the other hand, would be reported in terms of all 500 combined experiments, which in this example would constitute a total of 15,000 observations (or 30 students times 500 teachers). The 500 causal conclusions published in the newspaper (or on a school district website), on the other hand, are based upon separate contrasts of 500 "treatment groups" (each composed of changes in outcomes for a single teacher's 30 students) versus essentially the same "comparison group."

—iStockphoto/MHJ and Vanessa Solis

• Explicit guidelines exist for the reporting of medical experiments, such as the (a) specification of how many observations were lost between the beginning and the end of the experiment (which is seldom done in value-added experiments, but would entail reporting student transfers, dropouts, missing test data, scoring errors, improperly marked test sheets, clerical errors resulting in incorrect class lists, and so forth for each teacher); and (b) whether statistical significance was obtained—which is impractical for each teacher in a value-added experiment since the reporting of so many individual results would violate multiple statistical principles.

Of course, a value-added economist or statistician would claim that these problems can be mitigated through sophisticated analyses that control for extraneous variables such as (a) poverty; (b) school resources; (c) class size; (d) supplemental assistance provided to some students by remedial and special educators (not to mention parents); and (e) a plethora of other confounding factors.

Such assurances do not change the fact, however, that a value-added analysis constitutes a series of personal, high-stakes experiments conducted under extremely uncontrolled conditions and reported quite cavalierly.

Hopefully, most experimentally oriented professionals would consequently argue that experiments such as these (the results of which could potentially result in loss of individual livelihoods) should meet certain methodological standards and be reported with a scientifically acceptable degree of transparency.

And some groups (perhaps even teachers or their representatives) might suggest that the individual objects of these experiments have an absolute right to demand a full accounting of the extent to which these standards were met by insisting that students at least be randomly assigned to teachers within schools. Or that detailed data on extraneous events clearly related to student achievement (such as extra instruction received from all sources other than the classroom teacher, individual mitigating circumstances like student illnesses or disruptive family events, and the number of student test scores available for each teacher) be collected for each student, entered into all resulting value-added analyses, and reported in a transparent manner.

R. Barker Bausell is a biostatistics professor emeritus of the University of Maryland's school of nursing and the author of Too Simple to Fail: A Case for Educational Change *(Oxford University Press, 2010).*

Vol. 32, Issue 17, Pages 22-23, 25