RESEARCH BRIEF

Research Services

Vol. 1206
April 2013

Dr. Aleksandr Shneyderman
Dr. Terry Froman

The Unfulfilled Promise of Value-Added

The Desire for Teacher Assessment Tools

It would be nice to have a dependable, defensible, and sensitive system of assessing teacher effectiveness for the purposes of identifying professional development needs, deciding on teacher retention or promotion, and determining appropriate compensation. The traditional measures of teacher compensation weigh heavily on years of experience, supplemented with references to advanced degrees and teaching in the special programs. Critics of this approach (and there are many) state that real teacher effectiveness is not directly connected to teacher experience, except perhaps in the beginning of teaching career, and that this approach does not take into account student learning outcomes. It would be nice, instead, to have a system of evaluation that relied on hard data and quantifiable evidence of classroom performance that would directly relate to teacher effectiveness. It would be extremely useful in monitoring educational progress, alerting us to special needs, establishing standards of performance, and enabling us to recognize and assign the most capable teachers to the schools with the greatest demands. It would be nice to have a reliable method of assessing teachers -- so nice that we might be inclined to look eagerly and perhaps even a little bit uncritically to any program that promises such a system of evaluation.

The Definition of Teacher Effectiveness

Our expectations for effective teachers are broad and challenging. We certainly expect them to foster achievement in the main academic subjects, although we may not be sure how teachers can best accomplish that goal. But, beyond the course content itself, we expect teachers to encourage critical thinking, help students work cooperatively together, and promote good study habits and interest in learning, at the same time that they teach self-discipline and mutual respect. As important as each of these aspects of teaching is, there is only one facet for which we have tangible data readily available. Our focus for evaluating teachers often narrows down to the clues we can filter from the results of annual achievement tests. Test scores in reading and math are typically available, and the list of other subjects is growing, though it is far from complete. Even when high-quality achievement tests are administered regularly, they rarely assess all of the academic skills we hope our students are acquiring. Nonetheless, if teachers are truly effective, we rightfully expect to see exceptional results in the test scores of their students. But we should keep in mind that we have settled for a rather restricted definition of teacher effectiveness.

Random Assignment in Research Design

If we confine our attention to student test scores, there is a simple research design that would allow us to tease out the effects of the teachers. It requires that we randomly assign teachers and students to schools and randomly assign students to teachers. This random assignment accomplishes a great deal. There are many variables that can have an effect on student test performance that are beyond the influence of the teacher. First, students start the year with the teacher at different achievement levels. If students were assigned to teachers at random, and if there were enough students to sufficiently represent each teacher's effects, we could assume that advantages and disadvantages of different starting levels would balance out. More importantly, we could assume that many more outside influences – such as family dynamics, student attitudes and interest, peer effects, and a host of other unmeasured, possibly as yet unidentified variables – would also be balanced out and have no systematic impact on our estimates of teacher effectiveness.

Alternatives to Random Assignment

Of course, random assignment of teachers and students is a logistical impossibility. Even if we could circumvent the expense, disruption and the practical problems of assignment to locations, ethical and legal concerns would prevent randomization of students and teachers. When randomization is not possible, the effects associated with grouping can be confused by possible differences in the levels of extraneous variables among the groups. Sometimes these extraneous variables can be measured and statistically controlled for, effectively eliminating them as possible competing explanations for group differences. This can be particularly effective when there are only a few, readily identified important extraneous variables for which covariate adjustments can be made. Special precautions may be necessary if the covariates are numerous or difficult to define and measure. Of course, even in the best of circumstances, all extraneous influences cannot be controlled for and there will always remain some threats to conclusions of teacher and school effects based on alternative explanations for observed differences.

Complex Structure of the Data

The inability to utilize random assignment is just one challenge to assessing teacher effectiveness via a statistical model. Another important statistical issue is the complex nature of the data. In general, students are nested within teachers – that is, variation of student test scores must be considered within the natural grouping of students for each teacher. Additionally, teachers are generally nested within school settings. In some cases, students are taught the same course by more than one teacher or take more than one class for a given course. Sometimes teachers work in more than one school. Trying to account for student, teacher, and school effects simultaneously greatly complicates any potential analysis. The statistical techniques used in Florida for teacher evaluation try to accommodate both covariate adjustment and complex data structure through the use of multilevel linear modeling, often referred to as "value-added" methods. These sophisticated methods have proven themselves in other research domains and seem to be ideally suited as sensible candidates for approaches to teacher evaluation. In fact, proponents of applying the value-added models to determine teacher effects believe that these models hold a great promise in identifying true teacher effectiveness.

Analytic Complications

Value-added methods can be very complicated and difficult even for professional statisticians. For the assessment of teachers in Florida the complexity is compounded by the statewide size of the database, the large number of covariates, the overlapping of students within teachers, and the numerous levels and types of standardized tests. In fact, the specific analytic procedures are so complicated that they are almost impossible for the layperson to understand. So many statistical adjustments are made within the procedures that the results may bear little semblance to our common-sense intuitions. In addition, the complex and unidimensional nature of the value-added outcomes makes it especially challenging to offer professional advice for improvement or to gain teacher confidence and cooperation in the evaluation process.

Limited Application

There is not one value-added model for teacher evaluation. The particulars of the statistical procedures are recalculated for each specific student test. For just the grade-level reading and mathematics tests used in Florida, there are well over a dozen independent analyses performed. It has proven to have been highly optimistic to assume that these separate analyses would be harmonized in a way that would allow combination, categorization, and common interpretation. In fact, the attempt to aggregate separate teacher value-added estimates and distill the results to a few mandated categories has been highly contentious. Even if it were possible to manage these separate value-added estimates, they would only cover the teachers of the reading and mathematics subjects for which standardized tests exist. That would leave many of the teachers without any value-added evaluations. Attempts to reconcile this deficiency by assigning average VAM estimates of other teachers strike almost everyone as patently absurd. It should be no surprise that there is no support for judging teachers on the performance of students who they did not teach or on the performance of their students on subjects that they did not teach. Yet, to leave half of our teachers without performance-based evaluations seems equally unacceptable.

Noisy Effectiveness Estimates

There are many potential problems and there have been many fronts of attacks on the value-added procedures as they are practiced in the state of Florida. But, we should not be distracted from the most important issues. The final results of the value-added procedures, the explicit numerical effectiveness scores, are very fuzzy and imprecise – the noise-to-signal ratio is high, especially in reading. We can see this in two specific areas. The consistency of teacher effectiveness estimates from year to year is quite low, and the confidence range for these estimates is alarmingly wide. Given that the great majority of teachers are practicing in the same grade levels, in the same schools, with the same books and curricula, with the same types of students from year to year, and with similar teaching methods, they have a right to expect any truly representative measure of their effectiveness to be rather stable. However, in practice, the year-to-year correlation of teacher effectiveness estimates in reading is only about 0.3 based on the 2011 and 2012 M-DCPS data. In addition, the 95% confidence intervals around teacher value-added estimates are unacceptably wide: when expressed in terms of percentile standings for an average teacher, in reading they extend from 7th to 93rd percentile.

The teacher value-added estimates are so noisy that the folks responsible for the calculation of the value-added results have stated in their last technical report that, in reading, “it is not possible to distinguish good teachers from bad teachers.” This stunning acknowledgment appears to be a clear warning to the state officials that the value-added results in reading should not be used for high-stakes decisions until the model can be sufficiently improved. In fact, the inexactitude of value-added effectiveness estimates is widely recognized. Almost every statistician involved with value-added procedures has warned us against their use in high-stakes teacher evaluations. This should be enough, in itself, to dissuade us from using value-added assessments for making high-stakes decisions regarding teacher evaluations.

Recommendations for the Future

As was stated at the beginning of this paper, it would be nice if we had a dependable and defensible system of assessing teacher effectiveness. We now have enough information on the value-added procedures to see that, despite valiant and commendable efforts, it is not currently possible to satisfy our ultimate assessments desires through this approach. Unfortunately, there is not some easily identified alternative that would provide the precision and soundness we need for making critical and uncompromising decisions about teacher compensation and retention. Supplementing teachers' value-added estimates with observational school principal ratings also falls short of these strict requirements.

Given the limitations of the value-added model, perhaps it is best to use the outcomes of the model without triggering high-stakes decisions, such as teacher dismissal. Using value-added estimates of teacher effectiveness in combination with classroom observations may prove to be very useful in identifying best teaching practices and providing information on professional development needs allowing for more targeted teacher improvement efforts. The greatest goal of teacher evaluation is to improve the academic achievement of students. There is no reason that value-added approaches cannot be a respected and constructive partner in that endeavor.