Value-added Modeling and the Power of Magical Thinking

Henry Braun*

Abstract

This article explains the impetus in the United States behind the drive to extend test-based accountability to teachers and the growing interest in employing value-added models to generate the indicators to be used in teacher evaluation. The empirical literature finds that teaching quality is the most important school-related determinant of student. Yet, in the main, teacher evaluations are done poorly – if at all – and compensation has been largely determined by seniority and credentials. Policy makers see strengthening teacher accountability as a priority. In particular, they are looking to increase the role of outputs in comparison to inputs. However, given the technical problems associated with status-based indicators of teacher and school effectiveness, the focus has turned to indicators based on some measure of the progress students have made during the academic year. Value-added analysis relies on sophisticated statistical models to generate estimates of the relative effectiveness of teachers, based on a measure related to student progress. This article provides a brief introduction to value-added models and summarizes key research findings. Although value-added estimates have some desirable properties, they do not represent a simple, neat solution to a complex evaluation problem. In this spirit, the article concludes by describing some of the many concerns regarding the use of value-added scores for high-stakes decisions and suggests some ways to enhance the likelihood that teacher accountability will contribute constructively to the improvement of teaching.

Keywords: Accountability. Teacher evaluation. Value-added.

Modelagem de valor agregado e o poder do pensamento mágico

Resumo

Esse artigo trata do ímpeto nos Estados Unidos da América no processo de estender a responsabilização/prestação de contas (accountability) baseada em testes a professores e no crescente interesse em empregar modelos de valor acrescentado

* Professor of Education and Public Policy, Boston College (USA). *E-mail*: braunh@bd.edu

para gerar indicadores a serem utilizados na avaliação de professores. A literatura empírica mostra que a qualidade de ensino pelo professor é o fator escolar mais importante para o sucesso do aluno. Porém, em geral, as avaliações de professores são feitas de modo precário – quando feitas – e benefícios são prioritariamente determinados por tempo de serviço e títulos. Os formuladores de políticas veem o fortalecimento da responsabilização do professor como uma prioridade. Em particular, eles estão procurando aumentar o papel de resultados (outputs) em comparação com dados de entrada (inputs). Mas, devido aos problemas técnicos associados com indicadores derivados do status (status-based) de professores e da eficácia da escola, o foco foi orientado para indicadores baseados em alguma medida do progresso realizado pelos estudantes durante o ano letivo. Análises de valor acrescentado dependem de modelos estatísticos sofisticados para gerar estimadores da eficácia relativa dos professores, baseados em uma medida relacionada ao progresso dos alunos. Esse artigo fornece uma breve introdução aos modelos de valor acrescentado e faz um resumo dos principais resultados de pesquisas. Embora estimadores de valor acrescentado tenham algumas propriedades desejáveis, eles não representam uma solução simples e elegante para um problema complexo de avaliação. Nesse espírito, o artigo finaliza descrevendo algumas das muitas preocupações relativas ao uso de escores de valor acrescentado para decisões de grande consequência (high stakes) e sugere algumas maneiras para aumentar a possibilidade de que a responsabilização de professores irá contribuir construtivamente para a melhoria do ensino.
Palavras-chave: *Prestação de contas. Avaliação de professores. Valor acrescentado.*

El modelado de valor agregado y el poder del pensamiento mágico

Resumen

Este artículo analiza el entusiasmo existente en Estados Unidos con el proceso de extender la responsabilización (accountability) basada en tests a profesores y el creciente interés en emplear modelos de valor agregado para originar indicadores que se utilizarán en la evaluación de profesores. La literatura empírica muestra que la calidad de enseñanza del profesor es el factor escolar más importante para el éxito del alumno. Pero, en general, las evaluaciones de profesores se realizan de modo precario, cuando son hechas, y los beneficios se determinan prioritariamente por tiempo de servicio y por títulos. Los formuladores de políticas consideran el fortalecimiento de la responsabilización del profesor como una prioridad. En particular, buscan aumentar el papel de los resultados (outputs) comparados con los datos de entrada (inputs). Pero, debido a los problemas técnicos asociados con indicadores derivados del estatus (status-based) de profesores y de la eficacia de la escuela, el foco se orientó hacia indicadores basados en el progreso de los estudiantes durante el año lectivo. Se sabe que análisis de valor agregado dependen de modelos estadísticos

sofisticados para originar estimadores de la eficacia de los profesores, basados en una medida relacionada con el progreso de los alumnos. Este artículo ofrece una breve introducción a los modelos de valor agregado y resume los principales resultados de investigaciones. Aunque estimadores de valor agregado tengan algunas propiedades deseables, no representan una solución sencilla y elegante para un problema complejo de evaluación. Finalmente, el artículo describe algunas preocupaciones que surgen con el uso de scores de valor agregado para decisiones de gran consecuencia y sugiere algunas formas que ayuden a que la responsabilización de profesores pueda contribuir constructivamente para la mejora de la enseñanza.
Palabras-clave: *Responsabilización. Evaluación de profesores. Valor agregado*

Introduction

Over the last two decades, there has been increasing interest in improving the quality of public services through performance monitoring of governmental units and the individuals working in those units. Under the general rubric of "accountability", these initiatives have taken many forms, varying by country and by type of service. Education, as a highly visible public service, has not been immune to the *zeitgeist*. Although it is uncontroversial that government agencies, as well as publicly funded service providers, should be held accountable for their performance, it is very difficult to do this well; that is, for monitoring to achieve its intended goals without causing unwanted deterioration in other aspects of performance. The literature in public administration, and education in particular, is replete with warnings on the unintended consequences of poorly designed and/or poorly implemented accountability efforts (BIRD et al., 2005; LINN, 2004; MADAUS; RUSSEL; HIGGINS, 2009; ROTHSTEIN; JACOBSEN; WILDER, 2008).

In the United States, accountability in education has evolved in tandem with reform initiatives involving the development of new curricula, more demanding performance standards and greater reliance on standardized assessments to generate evidence regarding student learning. In 2001, under President George W. Bush, the U.S. Congress passed the No Child Left Behind Act (NCLB), which includes provisions for holding schools accountable for their students' levels of achievement in English/language arts and mathematics. The general opinion is that not only has it not resulted in accelerated student achievement, but also has had many negative consequences for both educators and students (SUNDERMAN, 2008).

One aspect of NCLB is of especial interest in this context. Simplifying somewhat, schools are held accountable based on the proportions of students meeting or exceeding pre-determined cut-scores in English/language arts and in mathematics. Of course, students' current achievement levels are the result of all the educational inputs received over their lifetimes, both due to formal schooling and other sources. Thus, holding this year's teachers responsible for the current achievement levels is

manifestly unfair – especially to schools where many students enter with substantial deficits. To add insult to injury, this test-based indicator plays a major role in determining how schools are evaluated, despite the empirically well-established finding that overreliance on a single indicator of performance inevitably leads to both distortion and corruption of the indicator (CAMPBELL, 1976).

Nonetheless, the administration of President Barack Obama has continued to promote test-based accountability and, through various competitive funding initiatives, such as the *Race to the Top*, has provided encouragement to states to overhaul their accountability programs for educators (teachers and principals) with a call to produce more credible and useful information. This call includes having student test scores contribute in some way to the evaluation of educators. A number of states are moving in this direction, with many state legislatures passing accountability laws that go beyond the federal government's guidelines. As the movement appears to be rapidly gaining ground, this is a propitious moment to take stock of this new phase of accountability, what forms it takes, and what are the likely results.

In light of the problems associated with the status-based indicators introduced under NCLB, there is great enthusiasm, in some quarters, for a statistical approach termed *value-added modeling* to determining educator effectiveness based on students' test scores. Although indicators derived from a value-added analysis have many advantages over status-based indicators, researchers have identified a number of technical issues that should give policy makers pause in employing value-added scores in high-stakes accountability systems and, especially, assigning them substantial weight in the evaluation process. Despite these warning flags, states are moving forward, but without putting in place systems for monitoring the broader impact of high-stakes accountability on schools and teachers. One interpretation is that they are relying on a kind of *magical thinking* that the identified problems with value-added analysis will somehow cancel each other out or that they will be of little concern once the system is implemented. In view of the potentially serious consequences of the evaluations, the plausibility of this approach merits strict scrutiny.

The paper begins with a brief history of educational accountability and then presents the essentials of value-added analysis. This is followed by a short discussion of some of the technical issues that have been raised, along with an argument that the ensemble of problems are neither likely to be self-canceling nor to exert a minor impact on the outcomes of the analysis. It concludes with some policy implications of this trend toward reliance on test-based indicators for accountability.

Evolving Paradigms of Accountability

A now common criticism of current systems of teacher evaluation and compensation is that, for the most part, they are based on seniority and credentials

(such as advanced degrees). Current research suggests that while teacher effectiveness does improve over the first several years of practice, it tends to flatten out after seven to ten years. Similarly, with the exception of degrees in mathematics or the sciences, teachers' additional credentials are only weakly related to their students' test performance (GOLDHABER, 2008). The relationship between credentials and non-test outcomes is unknown. There has been growing dissatisfaction with a system that is heavily weighted towards "inputs" rather than "outputs" – a system that, evidently, has not led to general improvements in the quality of instruction.

At least at a rhetorical level, there is a broad consensus that the primary purpose of educator accountability should be to provide information that can signal strengths and weaknesses, leading to greater effectiveness through focused professional development. At the same time, there are some who argue that an essential function is to identify educators at the extremes of the distribution of effectiveness -- with those at the high end garnering rewards and those at the low end subject to sanctions and perhaps dismissal. To accomplish both these functions, the system must collect relevant and credible evidence. Presumably, such evidence should include indicators based on student performance.

In the U.S., however, there is general agreement that in most jurisdictions the teacher accountability system is not up to the task. Setting aside contractual constraints on using student performance in evaluations, useful indicators related to teachers' professional practice are difficult to obtain. Most principals are not well trained to carry out rigorous observations of teachers. In many cases, teachers are observed on a haphazard schedule, or not at all. Moreover, most rating scales do not permit even relatively crude distinctions among teachers; for example, in many jurisdictions teachers can be classified only as "satisfactory" or "not satisfactory". Not surprisingly, the vast majority of teachers are rated "satisfactory", even when there is considerable evidence that this is likely not the case. This is particularly problematic since a number of analytic studies have concluded that there are substantial differences among teachers in effectiveness and that these differences have discernible implications for students' learning trajectories (LADD, 2008).

Such considerations have prompted the Federal government to encourage states to develop accountability systems that incorporate indicators based on student learning, as well as those derived from professional practice. In this article, I will consider the former. In particular, I will focus on the issues that arise for those U.S. teachers in grades 4 through 8 whose students are tested at the end of each school year in English/language arts and mathematics. Typically, this comprises about 30% of the teachers in a state. (How the other 70% of teachers should be held accountable is a very challenging issue that various states have addressed in different and sometimes bizarre ways.)

In principle, there are many ways to transform students' test scores into indicators that can be used for teacher accountability. The simplest is to use the average score earned by the class. The problem is that students' scores are the result of all the experiences, in school and out, that students have had until the time of the test. As is the case with other status-based indicators, it is unfair to hold the current teacher entirely responsible for those histories. Moreover, class-average scores are highly correlated with students' demographic characteristics. One alternative is to compute an average "gain score": the difference between the average score at the end of this year and (say) the average score at the end of last year. This is attractive because it appears to focus on the change that occurred during the year in which the students were in the teacher's class. Unfortunately, this requires that test scores from different grades be placed on a common scale. This is not easy to do – even when feasible -- and not recommended for a number of technical reasons (BRIGGS; WEEKS; WILEY, 2008). Further, gain scores are generally moderately correlated with students' demographic characteristics.

As any educator will attest, how much a student learns during the academic year depends on many factors, the skill of the teacher being only one, albeit an important one. In addition to home- and community-related characteristics, there are a number of school- related factors that also contribute to some degree. These include the demographic compositions of the class and the school, the quality of instructional leadership and the degree of professional collaboration in the school, as well as other aspects of the school climate. If one wants to "isolate" the contribution of the teacher, then the contributions of these other factors must be eliminated to the extent possible.

To disentangle multiple effects, the usual advice is to conduct a *randomized experiment*, as is often done in medical trials (SCHNEIDER et al., 2007). In such an experiment, individuals are assigned to different treatments using some random mechanism. If the study is large enough, then the groups of individuals exposed to each treatment are very similar on all potential relevant factors (observed or not) – except for the treatments themselves. Consequently, any differences in outcomes among the groups can be reasonably attributed to differences in the effectiveness of the treatments. Such experiments are challenging to carry out in education, especially at the student or class level.

In the education context, teachers are the "treatments". Students are exposed to these treatments by virtue of being enrolled in a particular class. The assignment process is rarely random, especially if we consider how both students and teachers are sorted, first into schools and then into classes. This non-random sorting is sometimes referred to as *self-selection* (even if the individuals involved have no say in the matter). One consequence of self-selection is that differences in achievement

across classes cannot be simply attributed to differences in teachers' skills – as would be the case in a randomized experiment. This is the difficulty that average scores and average gain scores fail to surmount. Proponents of "value-added", however, argue that it can resolve this difficulty sufficiently well that the results can and should be used for purposes of accountability.

What is the value-added approach?

The term "value-added" refers to statistical approaches to estimating the specific contributions to the achievement of students of their current teachers, schools or programs – taking account of (i.e., eliminating the effects of) the differences among students with respect to other factors associated with achievement, such as prior test scores and demographic characteristics. As noted above, when estimating the specific contributions of teachers, relevant differences among schools must also be taken into account.

One way of thinking about the value-added approach is that it is an attempt to make fair comparisons among educational providers such as teachers or schools, even though they carry out their work in very different contexts. In other words, it is a kind of "statistical salvage" – using sophisticated statistical models and extensive data to yield estimates of educator effectiveness that are intended to be almost as credible as those that would have been obtained from a true randomized experiment. This is a very ambitious goal and far exceeds what simple indicators like current test scores or gain scores can achieve.

There are many different value-added approaches, but essentially they all rely on the same strategy: Take all the available data and build a statistical model that predicts for each student what her current test score in a particular subject would be if she were typical of students with similar prior test scores and background characteristics, and was taught by a typical teacher. The difference between the student's actual score and the predicted score is treated as the teacher's value-added for that student. Explicitly:

Teacher's value-added = Student's actual score -- Student's predicted score.
contribution to the student

Then the estimate of the teacher's value-added is the average of the value-added contributions for the students in her class. (In some models, these raw estimates are then adjusted using empirical Bayes methods in order to reduce volatility.) Similarly, a school's value-added in a particular subject/grade is the average of the value-added contributions of teachers for all the students in that subject/grade.

Because of the way they are constructed, value-added scores are normatively defined. This means that a teacher's value-added is determined with respect to all

the teachers contributing data to the model. This is known as the reference group for the analysis. (All 6th grade teachers of mathematics in a school district is an example of a reference group.) If the reference group changes, the value-added scores will change as well. When the value-added scores for the reference group are ordered from highest to lowest, they range from positive to negative, with zero near the center of the ranking. The usual interpretation is that a strongly positive score indicates the teacher is likely more effective than the average teacher, a score near zero indicates a teacher of about average effectiveness, and a strongly negative score indicates the teacher is likely less effective than the average teacher. Note that the set of value-added rankings does not suggest where one should establish cut-points to distinguish, say, between weak teachers and average teachers or between average teachers and strong teachers. Since such cut-points are required if the rankings are to be used for evaluation, a defensible procedure for establishing them also has to be developed.

Examining Value-added

A teacher's or school's value-added is an estimate of effectiveness that seems ideally suited for the "new accountability". The promise of being able to extract the contributions of a teacher to her students' test score trajectories is very attractive and fuels the enthusiasm of policymakers for value-added. But a closer look raises many concerns that, in the views of many methodologists, argue for caution (NATIONAL RESEARCH COUNCIL, 2010; HARRIS, 2011). What is the reality?

Clearly, the credibility of a value-added estimate of a teacher's effectiveness is highly dependent on the accuracy of the predicted scores of her students. If those predictions are inaccurate, then the estimates will also be inaccurate, undermining the claims of fairness. The accuracy of the predicted scores depends on many factors, including the amount and quality of the test scores and other data that serve as input to the statistical model. Information about family/community variables, such as socio-economic status and parental education, may be inaccurate and may poorly reflect the relative advantages or disadvantages among students in their school experiences.

Just as important as the data included in the model is the data that should be in the model – but isn't. For example, many students may be missing one or more prior test scores. With many models, these students are dropped from the analysis. The number and characteristics of the "dropped students" can vary systematically from class to class and from school to school, perhaps making comparisons less fair. Similarly, aspects of the dynamics of the peer group in a particular classroom are poorly or not at all captured by available data. Another critical issue is student mobility. Both the extent and patterns of mobility can vary substantially across schools in a system. Teachers of classes that experience high mobility are at a double disadvantage. First, multiple transitions are very disruptive and often require the teacher to allocate her efforts in ways that are not optimal for the class as a whole.

Second, only a fraction of the students taught during the year contribute to the value-added calculation. The effort the teacher expended on the other students is lost to the model and this usually works to the disadvantage of the teacher.

There are other exogenous factors that can impact student achievement but are not captured by the model and, hence, confound estimates of teacher relative effectiveness. For example, in some areas parents can compensate for poor teaching by providing greater support – either by themselves or by hiring tutors. Many schools are sites for one or more interventions that may involve pre-school or after-school activities, in-class interventions, provision of medical, dental and psychological services, and the like. To the extent that the contributions of these efforts are not related (statistically) to the variables in the model, estimates of teacher value-added will be biased.

Many authors have pointed out that holding teachers accountable on the basis of the results of a statistical analysis is a form of causal inference. Now making causal inferences from an analysis that draws on data from an observational study rather than a randomized study is inherently problematic. In fitting a value-added model the goal is to statistically adjust for the differences among students and contexts so that one can make fair comparisons among teachers. Unfortunately, "... no statistical model, however complex, and no method of analysis, however sophisticated, can fully compensate for the lack of randomization" (BRAUN, 2005). As noted above, there are myriad ways in which the value-added approach fails to properly account for those differences. In some cases, it underadjusts and in others, it overadjusts. Most problematically, the extent and direction of the overall bias is difficult, if not impossible, to determine at the level of the individual teacher.

Reardon and Raudenbush (2009) discuss a number of fundamental assumptions that must be satisfied in order to justify making causal inferences from such models. One is that it must be possible to imagine that each student could earn a test score at any of the schools involved in the analysis and that score should not depend on what other students are enrolled in that student's class. The first assumption is called "manipulability" and the second is related to a property denoted "stable unit treatment value assumption" or SUTVA (RUBIN, 1986). As Reardon and Raudenbush (2009) point out, neither assumption is terribly credible in current educational contexts. They further note that the real issue is the degree of departure from the assumptions and the impact of the departure on the nature of the inferences made. Another concern is that the structure of the model itself may not properly capture the relationship between actual (current) test scores and the available predictors. This so-called "lack of fit" can also undermine the fairness of comparisons among teachers, particularly for those teachers whose classes enroll many students with unusual predictor profiles.

Finally, both the value-added models favored by econometricians (education production functions) and those preferred by statisticians (multi-level, mixed effects models) assume that the allocation of students to classes is based on fixed, or average, characteristics of the students. This is referred to as "static allocation". Recent research (ROTHSTEIN, 2009; ROTHSTEIN, 2010; KOEDEL; BETTS, 2009), however, casts serious doubt on that assumption. Although there is some evidence that averaging results over multiple cohorts can mitigate the problem, there remains concern that "dynamic allocation" introduces further bias into the value-added estimates.

Recall that the error associated with a statistical estimate has two components: bias and variance. With respect to the latter, the analysis of data from many jurisdictions reveals that estimates of teachers' value-added are not very stable; that is, they can vary substantially depending on the model (NEWTON et al., 2010; BRIGGS ; DOMINGUE, 2011) and the test that is used (CORCORAN, 2010; PAPAY, 2011). For example, using 4th and 5th grade data from Houston, Corcoran shows that 15% of the teachers whose value-added estimates using one reading test place them in the bottom fifth of all teachers, would be placed in the top two-fifths of all teachers based on their value-added estimates using a different reading test. Similarly, 17% of the teachers whose value-added estimates using the first reading test place them in the top fifth of all teachers, would be placed in the bottom two-fifths of all teachers based on their value-added estimates using that second reading test.

Perhaps more significant is that investigations of the stability of value-added estimates over time, which consistently yield low to moderate correlations (SASS, 2008; MCCAFFREY et al. 2009; GOLDHABER ; HANSEN, 2012). Using longitudinal data from North Carolina spanning 12 years, Goldhaber and Hansen (2012) examine the intertemporal stability of value-added estimates of teachers' relative effectiveness. Their findings are in agreement with the literature. Even when the volatility is dampened by averaging over three or more years, it remains quite substantial. Although they argue that this volatility does not present an absolute barrier to employing value-added estimates to hold teachers accountable for student progress, their use in high-stakes settings is certainly problematic.

It is also important to bear in mind that representations of the uncertainty in value-added scores only reflect estimates of model-based variance. These variance estimates do not reflect bias at all, even though the squared bias may well be of the same magnitude as the variance. Moreover, the confidence intervals that are calculated for each estimate are not adjusted for multiplicity; that is, they do not take account of the fact that many tests of significance are being conducted simultaneously and so a fraction will be found significant by chance. Hence, the representations of uncertainty are doubly optimistic about the capacity of the model to accurately distinguish teachers who are truly different from the average.

The Raw Material

The foregoing discussion has focused on a particular subset of issues that are relevant to a principled consideration of the appropriateness of including value-added results in the evaluation of teachers. However, the properties of the output of a value-added model are a complex function of the characteristics of the test scores that are entered into the model, as well as the model itself and the interactions between the two. There are numerous discussions in the literature of how the quality of the testing system can affect the appropriateness of the inferences and decisions made on the basis of the value-added scores (NATIONAL RESEARCH COUNCIL, 2010).

Ideally, developers of tests used for high-stakes would follow the *Standards for educational and psychological testing* (AMERICAN EDUCATIONAL RESEARCH ASSOCIATION; AMERICAN PSYCHOLOGICAL ASSOCIATION; NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION, 1999). However, these standards are simply guidelines for practice -- there is no body with the authority to monitor or enforce them. Consequently, assessments with varying degrees of quality are, in fact, employed to make decisions both about students, schools and, soon, about educators.

Assuredly, the *sine qua non* is that the tests should be construct valid (MESSICK, 1989). Suppose the test does not adequately reflect the content standards. Then student performance on the test does not fully represent what they know and can do, and consequently inferences that teacher value-added scores represent their relative effectiveness in teaching to the standards are faulty. Test scores may also suffer from construct irrelevant variance, which further undermines the credibility of the desired inferences.

However, some test-related issues are more germane to the use of test scores for accountability than for decisions about students. For example, the ways in which the raw student responses in a particular grade are scaled and then rescaled, either through vertical linking or some standardization procedure, impacts the value-added results (BRIGGS; WEEKS; WILEY, 2008). Other psychometric characteristics, such as the conditional standard error of measurement and, most importantly, severe floor- or ceiling-effects, also influence the validity of the inferences made on the basis of the value-added results.

Magical Thinking

At this point, the reader may wonder how it is possible to remain enthusiastic about the use of VAMs in high-stakes settings. Indeed, the set of potential problems is quite impressive. On the other hand, all indicators of teaching quality, such as those based on observations of teacher practice, also suffer from various defects and do raise validity concerns of equal or greater magnitude. So, for some commentators, it is quite appropriate for the results of a value-added analysis to be included in

an evaluation system, along with other fallible indicators (NATIONAL RESEARCH COUNCIL, 2010; HARRIS, 2011). For (non-technical) policy makers that conclusion is made all the more attractive when they are assured that a particular VAM produces accurate estimates of teachers' relative effectiveness.

In my darker moments, I regard this assertion as a variant of what anthropologists and psychologists call *magical thinking*. According to Zusne and Jones (1989), magical thinking is characterized by a belief that

a) transfer of energy or information between physical systems may take place solely because of their similarity or contiguity in time and space, or that;
b) one's thought, words, or actions can achieve specific physical effects in a manner not governed by the principles of ordinary transmission of energy or information.

In this context, magical thinking consists of believing that assertions accompanied by certain statistical incantations can overcome the deleterious effects of multiple serious threats to validity. Were it the case that there were only one or two such threats, then belief would (perhaps) be more understandable. But here we have a situation in which there are several such threats and it is almost impossible to calculate their impact on the accuracy and precision of the estimates, either singly or in combination. To believe that their cumulative impact is small and sufficiently uniform to be ignorable, is akin to magical thinking.

Moreover, there is no safety in numbers: Even if the rankings of teachers under different models are reasonably highly correlated (often not the case), they likely share a common bias; that is, departures from such assumptions as static allocation and SUTVA would impact the estimates from essentially all the models in use today. In conjunction with the value-added estimates' substantial volatility, it seems to me to be rather reckless to allocate a great deal of weight to those estimates.

Policy Implications

Although the preceding review is rather discouraging, there is general agreement that, in the aggregate, value-added estimates do contain useful information about the relative effectiveness of teachers. But for any individual teacher they can lead to inferences that are misleading. It should be kept in mind, though, that any indicator of teacher effectiveness is fallible; all indicators are subject to both systematic and random errors. The solution is not to discard them but, rather, to develop a strategy that makes the best use of all the information available.

My own view is that value-added estimates from certain well-researched models do have a place in an accountability system, particularly if the law demands that

"evidence of student learning" play a role in the evaluation. However, such estimates should be considered alongside indicators of teachers' professional practice. For now, value-added results should not be assigned a heavy weight in the final evaluation score. In fact, I would argue that value-added estimates should be first reviewed by the school principal and subject to challenge when it is felt that the model has failed to capture relevant aspects of the local context. (There is some anecdotal evidence that many teachers assigned extreme negative value-added scores teach in contexts that are not well represented by the regression model.) Of course, local reviews should be conducted on a principled basis and centrally audited. The principal should be held accountable for numerous unsuccessful challenges.

It is generally agreed that we cannot "fire our way to success"; that is, firing the bottom x% of teachers based on estimated value-added will not have a substantial impact on the distribution of students' scores. Thus, the chief goal of the accountability system must be to improve the overall quality of teaching. In this regard, test-based indicators drawn from end-of-year summative assessments generally have little to offer on how a teacher should improve her practice. Rather, it is those indicators drawn from the classroom practices of the teacher, in addition to evidence of her professional skills and dispositions, that point the way to targeted professional development. Thus, multiple indicators are not only necessary for fair evaluations, but also for the accountability system to contribute to the systematic and sustained improvement of teaching and learning. At the same time, it is widely acknowledged that the statistical characteristics of observational systems need further investigation and that there is much room for improvement in that arena as well (HILL; CHARALAMBOUS; KRAFT, 2012).

The rationale for holding schools and teachers accountable is a compelling one and truly ineffective teachers should certainly be identified and dealt with appropriately. This sentiment, spurred by the promises of vendors, fuels the enthusiasm for value-added methods. It is clear, though, that the promises of test-based indicators should be examined very critically. They should be employed thoughtfully and with due regard to potential negative consequences.

We do society no favor if, with the best of intentions, we introduce an accountability system that discourages the best teachers from working with the students who need them the most, that hastens the departure of good teachers, and dissuades promising, prospective teachers from entering the field altogether. At present, an accountability system that accomplishes flexible regulation in the service of constructive improvement of institutional outcomes is as rare as a unicorn – the main difference is that nowadays no one claims to have a unicorn in their backyard!

References

AMERICAN EDUCATIONAL RESEARCH ASSOCIATION; AMERICAN PSYCHOLOGICAL ASSOCIATION; NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION. *Standards for educational and psychological measurement*. Washington, DC: American Educational Research Association, 1999.

BIRD, S. et al. Performance indicators: good, bad and ugly. *J. Royal Statistical Society A*, London, v. 168, n. 1, p. 1-27, 2005.

BRAUN, H. I. *Using student progress to evaluate teachers:* a primer on value-added models. Princeton, NJ: Policy Information Center, Educational Testing Service, 2005.

BRIGGS, D.; WEEKS, J.; WILEY, E. The sensitivity of value-added modeling to the creation of a vertical score scale. *Education Finance and Policy*, Cambridge, v. 4, n. 4, p. 384-414, 2008.

BRIGGS, D. ; DOMINIGUE, B. *Due diligence and the evaluation of teachers*: a review of the value-added analysis underlying the effectiveness rankings of LAUSD teachers by the Los Angeles Times. Boulder, CO: National Education Policy Center, 2011.

CAMPBELL, D. *Assessing the impact of planned social change.* [S.l.]: the Public Affairs Center; Hanover, New Hampshire, USA : Dartmouth College, 1976.

CORCORAN, S. *Can teachers be evaluated by their students' test scores? Should they be? The use of value-added measures of teacher effectiveness in policy and practice.* Providence, RI: Annenberg Institute for School Reform, 2010.

GOLDHABER, D. Teachers matter, but effective teacher policies are elusive. In: LADD, H.; FISKE, E. (Ed.). *Handbook of Research in Education Finance and Policy.* New York: Routledge, 2008.

GOLDHABER, D.; HANSEN, M. *Is it just a bad class? Assessing the long-term stability of estimated teacher performance.* [S.l]: National Center for Analysis of Longitudinal Data in Education Research, 2012. (Working Paper 73).

HARRIS, D. *Value-added measures in education*. Cambridge, MA: Harvard Education Press, 2011.

HILL, H.; CHARALAMBOUS, Y. ; KRAFT, M. When rater reliability is not enough: teacher observation systems and a case for the generalizability study. *Educational Researcher*, Thousand Oaks, CA, v. 41, n. 2, p. 56-64, 2012.

KOEDEL, C.; BETTS, J. *Does student sorting invalidate VAMs of teacher effectiveness? An extended analysis of the Rothstein critique*. Columbia, MO: University of Missouri, 2009. (U. of Missouri Working Paper 09-02).

LADD, H. F. Teacher Effects: what do we know? In: DUNCAN, G.; SPILLANE, J. (Ed.). *Teacher quality:* broadening and deepening the debate. Evanston, IL: Northwestern University, 2008. Available from: <http://tqn.sesp.northwestern.edu/>.

LINN, R. Accountability models. In: FUHRMAN, S.; ELMORE, R. (Ed.). *Redesigning Accountability Systems for Education*. New York: Teachers College Press, 2004.

MADAUS, G.; RUSSELL, M.; HIGGINS, J. *The paradoxes of high-stakes testing*. Charlotte, NC: Information Age Publishing, 2009.

MCCAFFREY, D. et al. The intertemporal variability of teacher effect estimates. *Education Finance and Policy,* Cambridge, MA, v. 4, n. 4, p. 572-606, 2009.

MESSICK, S. Validity. In: LINN, R. (Ed.). *Educational Measurement*. 3rd ed. New York: Macmillan, 1989. p. 13-103.

NATIONAL RESEARCH COUNCIL. *Getting value out of value-added*. H. I. Braun, N. Chudowsky and J. Koenig (Eds.). Washington, DC: [s.n.] , 2010.

NEWTON, X. A. et al. Value-added modeling of teacher effectiveness: An exploration of stability across models and contexts. *Education Policy Analysis Archives*, [S.l.], v. 18, n. 23, 2010.

PAPAY, J. P. Different tests, different answers: the stability of teacher value-added estimates across outcome measures. *American Educational Research Journal*, Thousand Oaks, CA, v. 48, n. 1, p. 163-193, 2011.

REARDON, S. ; RAUDENBUSH, S. Assumptions of value-added models for estimating school effects. *Education Finance and Policy*, Cambridge, MA, v. 4, n. 4, 2009.

ROTHSTEIN, J. Student sorting and bias in value-added estimation: selection on observables and unobservables. *Education Finance and Policy,* Cambridge, MA, v.4, n. 3, p. 537-571, 2009.

_____. Teacher quality in educational production: tracking, decay, and student achievement. *Quarterly J. of Economics*, [S.l.], v. 125, n. 1, p. 175-214, 2010.

ROTHSTEIN, R., JACOBSEN, R.; WILDER, T. *Grading education: Getting accountability right*. New York: Teachers College Press, 2008.

RUBIN, D. Which ifs have causal answers? *J. American Statistical Association*, [S.l.], v. 81, p. 961-962, 1986.

SASS, T. R. *The stability of value-added measures of teacher quality and implications for teacher compensation policy*. Brief 4. [S.l.]: National Center for Analysis of Longitudinal Data in Education Research, 2008.

SCHNEIDER, B. et al. *Estimating causal effects:* using experimental and observational designs. Washington, DC: American Educational Research Association, 2007.

SUNDERMAN, G. (Ed.). *Holding NCLB accountable:* achieving accountability, equity & school reform. Thousand Oaks, CA: Corwin Press., 2008.

ZUSNE, L.; JONES, W. *Anomalistic Psychology:* a Study of Magical Thinking. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Assoc., 1989. Available from: <http://www.skepdic.com/magicalthinking.html>. Accessed: 4th February, 2012.

Recebido em: 19/12/2012
Aceito para publicação em: 18/04/2013