Evaluating Value-Added Models for Teacher Accountability

DANIEL F. McCAFFREY
J.R. LOCKWOOD
DANIEL M. KORETZ
LAURA S. HAMILTON

Prepared for the Carnegie Corporation

RAND EDUCATION

The research described in this report was conducted by RAND Education for the Carnegie Corporation of New York.

Library of Congress Cataloging-in-Publication Data

Evaluating value-added models for teacher accountability / Dan McCaffrey ... [et al.].
p. cm.
"MG-158."
Includes bibliographical references.
ISBN 0-8330-3542-8 (pbk.)
1. Educational indicators—United States. 2. Teachers—Rating of—United States—Mathematical models. I. McCaffrey, Daniel F.

LB2846.E88 2004
371.14'4—dc22

2003026262

The RAND Corporation is a nonprofit research organization providing objective analysis and effective solutions that address the challenges facing the public and private sectors around the world. RAND's publications do not necessarily reflect the opinions of its research clients and sponsors.

Published 2003 by the RAND Corporation
1700 Main Street, P.O. Box 2138, Santa Monica, CA 90407-2138
1200 South Hayes Street, Arlington, VA 22202-5050
201 North Craig Street, Suite 202, Pittsburgh, PA 15213-1516
RAND URL: http://www.rand.org/
To order RAND documents or to obtain additional information, contact
Distribution Services: Telephone: (310) 451-7002;
Fax: (310) 451-6915; Email: order@rand.org

Summary

Background and Purpose

Value-added modeling (VAM), a collection of complex statistical techniques that use multiple years of students' test score data to estimate the effects of individual schools or teachers, has recently garnered a great deal of attention among both policymakers and researchers. For example, a recent bill drafted by the General Assembly of Pennsylvania proposes using student achievement results and value-added models to evaluate and reward administrators and teachers. In this bill, VAM-based estimates of teacher and school effects would affect salaries and career ladder stages as well as contract renewal for teachers and administrators.

There are at least two reasons why VAM has attracted growing interest. One reason is that VAM holds out the promise of separating the effects of teachers and schools from the powerful effects of such noneducational factors as family background, and this isolation of the effects of teachers and schools is critical for accountability systems to work as intended. The second is that early VAM studies purport to show very large differences in effectiveness among teachers. If these differences can be substantiated and causally linked to specific characteristics of teachers, the potential for improvement of education could be great.

The application of VAM to educational achievement holds considerable promise, but it also raises many fundamental and complex issues. Unfortunately, investigation and discussion of the issues raised by the use of VAM in education have been fragmented and incomplete. Although there have been reviews of particular approaches (e.g., Bock, Wolfe, and Fisher, 1996), no reviews have carefully compared recent VAM efforts or systematically discussed the wide variety of issues they raise. Moreover, while numerous methodological concerns have been raised by VAM researchers and by critics of the approach, much of the discussion remains unpublished, and the practical import of these concerns when VAM is applied to student achievement remains largely unclarified.

This monograph is one of the products of an effort by RAND Corporation researchers to begin a systematic review and evaluation of leading approaches to VAM. It had several goals: to clarify some of the most important issues, to begin evaluating their practical impact, to spur additional work on these issues, and to help inform the debate among both researchers and policymakers about the potential of VAM. In the monograph, we clarify the primary questions raised by the use of VAM for measuring teacher effects, review the most important recent applications of VAM, and discuss a variety of important statistical and measurement issues that might affect the validity of VAM inferences. Although parts of the monograph are technical in nature, we have avoided lengthy discussions of technical issues. Several more detailed discussions of technical issues are contained in an appendix to the monograph and have also been published elsewhere (Lockwood, Louis, and McCaffrey, 2002; McCaffrey et al., 2003).

What We Learned

Review of the Literature

The recent literature on VAM purports to show that teachers differentially affect student learning and growth in achievement. This literature suggests that teacher effects are large, accounting for a significant portion of the variability in growth, and that they persist for at least three to four years. A relatively small number of papers—several of them not published in the peer-reviewed literature—are the source of these claims. We critically evaluated the methods used in these papers and the validity of the resulting claims. We conclude that although the papers all have shortcomings, together they provide evidence that teachers have discernable, differential effects on student achievement, and that these effects appear to persist into the future. The shortcomings of the studies make it difficult to determine the size of teacher effects, but we suspect that the magnitude of some of the effects reported in this literature are overstated.

Wright, Horn, and Sanders (WHS, 1997) conclude that teachers are the most important factor affecting student learning. In their replicated study design, they model gains in student tests score as a function of random teacher effects and a small set of student covariates including achievement. They standardize the contributions of all variables in the models using what they call a "z-score." They informally meta-analyze the results of the 30 replicated models and find that the z-score for teacher effects exceeds the standardized contribution of every other variable in 26 of 30 models. Via a simulation study, we find that the authors' standardized z-scores do not necessarily preserve the ranking of variables based on contribution to total variance in scores. In other words, while the WHS z-scores for teachers might dominate in 26 of 30 models, this does not imply that teacher effects explain more variance than all the other predictors. Furthermore, WHS provide no evidence that the estimated teacher effects and their corresponding variance components are unbiased by contributions of other inputs to education that are not accounted for in the model.

In another report, Rowan, Correnti, and Miller (RCM, 2002) find that residual classroom-level variance accounts for a significant proportion of the variability in growth in student achievement scores. The results are robust across subjects (reading or math), statistical models, and two cohorts of students from a nationwide sample of schools. Although classrooms account for meaningful portions of the variance in all models, the magnitude of the variance explained varies. While the results are impressive and strongly suggest that teacher (classroom) effects are nonzero, the authors do not provide details on missing data, the nature of the measure, or the distributions of student characteristics—so a full assessment of possible biases is impossible.

Rivkin, Hanushek, and Kain (RHK, 2000) take advantage of multiple cohorts of students, each with three years of test scores, to aggressively remove the effects on achievement of factors other than teachers. The authors find that teacher effects do exist and estimate that, as a lower bound, teachers account for about 3.2 percent of variance in achievement. In other words, a one-standard-deviation increase in teacher effectiveness is associated with about a 0.18-standard-deviation increase in scores. While their methods remove many possible confounding factors, the estimates are based on gains and differences of scores that are not on a single developmental scale. Changes in scores, therefore, do not necessarily correspond to growth in achievement—making the interpretation of results difficult. Also, the authors restricted their analyses to students who remained in the same school and completed testing for three consecutive years. Thus, the authors' findings suggest that teachers can matter for some students in some metrics, but a more generalizable interpretation of their results is impossible.

In 1996, Sanders and Rivers (SR) released a technical report purporting to show that teacher effects accumulate over time. They report that for math tests, students taught by the least effective teachers for three consecutive years would score 52 to 54 percentile points below similar students taught by the most effective teachers for three consecutive years. This dramatic finding has garnered enormous attention from researchers, policy makers and other interested parties. We evaluated the methods used by SR used via simulation and concluded that, based on scenarios that best match the numbers reported in SR and our experience with school data, the SR results would be unlikely to occur if teachers or classrooms had no effects. However, there is reason to expect a small positive bias in their estimates of the size of these effects. Thus, the SR results are consistent with the existence of persistent teacher effects but might overstate the size of such an effect.

Given the magnitude of the SR effects, the implications of their finding, and the controversy with their methodology, other authors have attempted to replicate the result with slight modifications. Rivers (1999) replicated the design with several important changes to address some of the criticisms of SR and still found persistent teacher effects. Mendro, Jordan, Gomez, Anderson, and Bembry (MJGAB, 1998) used data from students in the Dallas Independent School District to replicate the SR study. MJGAB again found large persistent teacher effects across multiple cohorts and on both mathematics and reading scores. Kain (1998) conducted an independent analysis of a subset of the MJGAB and found similar results. The MJGAB and Kain analyses control for many student characteristics, including neighborhood effects. Thus, their estimated teacher effect should be reasonably unconfounded by other sources. Even though all these studies have limitations, and MJGAB and Kain provide limited details of their studies, the consistency of findings across samples from different locations and different statistical models suggests to us that these papers together provide evidence that teacher effects do persist across years.

Modeling Longitudinal Data to Estimate Teacher Effects

Estimating the effects of teachers by modeling longitudinal data on student achievement raises a number of important statistical and psychometric issues and requires decisions about how these issues should be addressed. Estimates may vary appreciably as a result of these decisions, and the resulting uncertainty of findings should be considered when interpreting VAM estimates. In this respect, VAM analyses of teacher effects are no different from other statistical models, estimates from which are often potentially sensitive to choices about the modeling approach. However, the analyses used to estimate teacher effects are complex and challenging, and the potential sensitivity of their results to modeling choices has not been well explored. In this monograph, we discuss some of the decisions that must be made about modeling achievement data to estimate teacher effects and the possible sensitivity of estimates of teacher effects to them. We break these decisions or issues into four groups: basic issues of statistical modeling, issues involving confounders or omitted variables, issues arising from the use of achievement test scores as dependent measures, and uncertainty about estimated effects.

Basic Issues of Statistical Modeling. Analysts generally have used one of three approaches to analyzing longitudinal data to estimate teacher effects. Two of these approaches break the longitudinal analysis into a sequence of models for single-year outcomes, which makes statistically inefficient use of information for the multiple years of data but is computationally simpler than the alternative full multivariate modeling of multiple years of data. Full multivariate analysis of the data is flexible and uses correlation among multiple years of data. This approach is likely to be preferable but is computationally demanding.

Another choice in statistical modeling is the specification of teacher effects as "fixed" or "random" effects. In the past, fixed effects were used in such efforts (Murnane 1975; Hanushek, 1972); recent applications (Sanders, Saxton, and Horn, 1997; Ballou, Sanders, and Wright, 2003; Rowan, Correnti, and Miller, 2002) use random effects specification. The two methods will tend to yield similar conclusions about the variability of teachers but will provide different estimates of individual teacher effects. The differences result from different strategies for dealing with inherent sampling error of estimated effects. The fixed-effects method uses a teacher's students to estimate his or her effect. The random-effects method "shrinks" the estimate based on the given teacher's students toward the overall mean for all students. On average, shrinking the estimate has optimal statistical properties across teachers but can be sub-optimal for teachers whose effects are far from the mean. Fixed-effects estimates can be highly sensitive to sampling error because teachers tend to teach only small numbers of students.

Omitted Variables, Confounders, and Missing Data. VAM uses data collected in an observational setting (as opposed to an experimental setting). The data collected from this observational setting can be subject to a number of problems. In particular, two types of problems arising from these circumstances have the potential to distort VAM estimates of teacher effects. The first type is confounding by influences other than teachers on student learning that are incorrectly modeled or are not modeled at all—for example, a model that does not properly distinguish the effects of teachers from other effects of the school in which the teacher works. The second type is incomplete data. In the case of VAM, incompleteness frequently arises in two areas: data for individual students over time and information on the linking of students to teachers. We believe these problems are among the greatest challenges facing VAM.

Models that fail to account for differences in student populations across schools can yield biased estimates of teacher effects. This is the case even for complex multivariate models that jointly model many student outcomes. Bias can occur when students attending different schools differ in ways that are likely to affect both achievement and growth in achievement, and the context of the school (e.g., the proportion of students eligible for free and reduced price lunches) affects these outcomes. Given that student populations tend to vary among schools—and our limited empirical findings suggest that context does affect growth in some settings—omitted variables appear to be a likely source of bias in most VAM applications. Although recent work on this topic (Ballou, Sanders and Wright, 2003) suggests that in some settings including student level covariates has little effect on estimated teacher effects, this work was unable to reach the same conclusion about context effects. In our own limited example, context effects had a great impact on estimated effects. Because true teacher effects might be correlated with the characteristics of the students they teach, current VAM approaches cannot separate any existing contextual effects from these true teacher effects.

Other effects that can be difficult to disentangle from the effect of the students' current teachers are those arising from schools, districts, or prior teachers. If terms for these effects are omitted from models, they are implicitly subsumed by teacher effects, which may bias what analysts conceive as true teacher effects. Alternatively, if such effects are included in models and teachers of differential effectiveness cluster at the school or district level, part of the true teacher effects will be attributed to schools or districts. Analysts must decide which potential error is more acceptable.

Real longitudinal student achievement data will inevitably contain incomplete student achievement records. The accuracy of estimated teacher effects in the presence of incomplete records is sensitive to models for the nature of missing data and to the analytic approach. Little is currently known about the effects of missing data on VAM estimates of teacher effects. Similarly, the links between students and teachers might be incomplete, and the effects of these incomplete links on outcomes have received no investigation to date. If incomplete test score data and incomplete links between teachers and students do in fact result in bias, it could be a large problem. The factors that contribute to missing links and missing test scores are common: students are mobile, with large proportions transferring among schools every year.

Issues Arising from the Use of Achievement Tests as an Outcome. The student achievement measures that VAM uses to define and estimate teacher effects are limited in several ways. Testing is infrequent—typically only once a year—and the tests used to measure achievement cannot measure fully all topics related to achievement. In addition, the scale for measuring achievement is not predetermined by the nature of achievement but is chosen by the test developer. Changes to the timing of tests, the weight given to alternative topics, or the scaling of the test could change our conclusions about the relative achievement or growth in achievement across classes of students. These changes would change estimates of teacher effects. While our explorations suggest that some of these effects might not be large (for example, differential growth during the summer recess), the effects of other changes could have large impacts on estimated teacher effects and require further investigations.

Uncertainty About Estimated Effects. Accurate inferences about a teacher's effect require an estimate of the effect that is likely to be close to the real teacher effect. As we have discussed, a number of decisions related to both modeling and measurement contribute to possible errors and uncertainty of the estimate. Sampling error is another source of error in VAM estimates. Uncertainty must be very small to make useful inferences about some quantities of interest, such as teacher ranking, and real estimates are unlikely to have such small sampling error. However, estimates might be sufficiently precise for other inferences, such as identifying some teachers as distinct from the mean. In one small example, we estimated that about one-third of teachers had a very small probability of being equal to the average teacher, which suggests that, for some applications, sampling error in the estimates will not preclude identifying a fraction of teachers as above or below average. Our estimates were somewhat robust to the model for prior-year teachers, but we did not account for potentially formidable uncertainty in other factors, such as missing data, type of measurement, or the effects of omitted student characteristics.

What We Recommend

Using VAM to estimate individual teacher effects is a recent endeavor, and many of the possible sources of error have not been thoroughly evaluated in the literature. Our goal was to identify possible sources of error and bias and evaluate what is known at this point. We recommend that many of the possible errors we identified receive additional review in the literature. Some of the areas for future research include the following:

1. Develop databases that can support VAM estimation of teacher effect across a diverse sample of school districts or other jurisdictions.
2. Develop computational tools for fitting VAM that scale up to large databases and allow for extensions to the currently available models.
3. Link VAM teacher-effect estimates with other measures of teacher effectiveness to determine the characteristics or practices of effective teachers as a means of validating estimate effects and possibly identifying what produces effective teaching.
4. Empirically evaluate the potential sources of errors we identified to determine how these factors contribute to estimated teacher effects and to determine the conditions that exacerbate or mitigate the impact of these factors on teacher effects.

5. Estimate the prevalence of factors that contribute to the sensitivity of teacher-effect estimates by surveying school districts and by replicating VAM estimation effort across multiple locations and meta-analyzing the results.
6. Incorporate decision theory into VAM by working with policymakers to elicit decisions and costs associated with those decisions and developing estimators to minimize the losses.
7. Use research and auxiliary data to inform modeling choices.

Recommendations for the Use of VAM in Policy and Practice

The research base is currently insufficient to support the use of VAM for high-stakes decisions. We have identified numerous possible sources of error in teacher effects and any attempt to use VAM estimates for high-stakes decisions must be informed by an understanding of these potential errors. However, it is not clear that VAM estimates would be more harmful than the alternative methods currently being used for test-based accountability. At present, it is most important for policymakers, practitioners, and VAM researchers to work together, so that research is informed by the practical needs and constraints facing users of VAM and implementation of the models is informed by an understanding of what inferences and decisions the research currently supports.