The Effect of Demographic Factors in Value-Added Models

November 2012



In the following research brief, Hanover discusses the merits of adding demographic variables to value added models. In particular, we examine whether adding poverty measures reduces estimation errors.

HANOVER RESEARCH

EXECUTIVE SUMMARY AND KEY FINDINGS

INTRODUCTION

The topic of value added modeling (VAM) has generated much debate within the education community in recent years. While proponents of these systems have argued that these models offer an objective, data-driven approach to teacher and school evaluations, those who oppose this trend suggest that these models are biased and that they should not be used for "high stakes" decisions, such as compensation and tenure. In our earlier report titled "Stability of VAM Estimates of Teacher or School Effectiveness," a review of the literature revealed that most experts agree that while VAMs are accurate enough to serve as a formative assessment (e.g. for identifying teachers who may warrant closer evaluation by a principal or supervisor), they are not reliable enough to be the sole basis for important personnel decisions.

One major concern of researchers who have studied VAM is that it may not be possible to adequately control for the non-random assignment of students to teachers. In this report, we describe how districts have sought to minimize these potential biases and attempt to offer the District guidelines as South Carolina seeks to formulate its model. In particular, we attempt to determine the value of including student-level demographic data, including poverty measures, in the models.

KEY FINDINGS

- The research indicates that the results produced by simple models that do not control for student characteristics (e.g. race, poverty measures, etc.) are strongly correlated with those produced by more complex models that do include these controls, suggesting that the inclusion of these variables does not play a big role in determining the effectiveness of the model.
- Nevertheless, we do recommend the inclusion of variables such as poverty measures, considering that the potential benefits (i.e. minimizing potential biases) of doing so clearly outweigh the costs, which are negligible in most cases. This perhaps explains why most of the VAMs we encountered do include these controls.
- There seems to be some agreement that it is best to use more than one data point when controlling for pre-test achievement, as random fluctuations in scores can lead to imprecise estimates of teacher effects in these cases. There are districts that also use data from multiple cohorts to minimize potential biases.
- Some of the research we encountered indicates that the characteristics of a classroom may impact the rate at which students learn. This likely explains the growing trend towards incorporating classroom effects and characteristics into

these models. For instance, a VAM just introduced in the state of Florida attempts to control for these variables.

- One weakness of value added models is that it can be difficult to produce reliable estimates for teachers who have a small number of students. In cases like these, many districts rely on techniques such as Bayes estimation (or shrinkage), while some limit their analyses to teachers who have taught a certain minimum number of students.

SECTION I: EFFECTS OF STUDENT DEMOGRAPHICS

The inclusion of student-level demographic variables in VAMs has been a topic of much research over the years, with some experts stating that these measures should be included in VAMs given that they are known to impact student test score growth. Others argue that this may not be necessary considering that the models already include variables that are strongly correlated with poverty measures. Indeed, we know that it is at possible to at least partially control for a variable without directly including it in the model. For instance, the prior year test score of students (a variable often included in VAMs) is likely strongly correlated with poverty measures. Therefore, when the prior year test score is included, it is likely at least partially controlling for poverty.

Given this debate, this report summarizes the studies that have been conducted on this topic. In particular, we attempt to determine whether including these variables results in a significant change in the estimates produced by the models, as answering this question could help us determine whether the inclusion of these variables helps minimize biases. We also describe the types of models that are most likely to be used in practice.

INCLUDING STUDENT CHARACTERISTICS IN VAMS

In 2004, researchers studied a medium size district in Florida in an effort to answer exactly this question. [1] The authors of the study analyzed two years of student test score data using several different types of models. One was a simple model that did not control for any student-level variables (aside from the prior year test score). Two more complex models were also constructed, both of which made an attempt to control for student-level demographic variables (such as poverty measures and minority status). The authors found that the simpler model produced results that were highly correlated with those from the more complex models. Nonetheless, the researchers argued that the simple model could not be unequivocally recommended over the more complex ones. They argued that their results were in fact inconclusive, as there were some key differences between the two sets of models, particularly at the school-level. For instance, they concluded that while the simple model produced results that were biased against schools with an over-representation of poor and minority students, the more complex models resulted in measures that were biased against schools with an underrepresentation of these groups. They argued that choosing one over the other amounted to a policy choice.[2]

A study of high schools in the San Francisco Bay Area also confirmed a strong correlation between the results of a model that did not include student characteristics and one where

[1] Tekwe, C., et al. 2004. "An Empirical Comparison of Statistical Models for Value-Added Assessment of School Performance." Journal of Educational and Behavioral Statistics. http://familydata.health.ufl.edu/files/2010/10/tekwe_statistical_assessment.pdf
[2] Ibid.

the variables are included.[3] These researchers did find that the significant negative correlations between variables such as poverty status and teacher rankings were less strong in models that controlled for student demographics, though it is worth noting that the relationships were still statistically significant.[4] This finding seems to confirm the compositional or contextual effects outlined by other scholars. This line of research suggests that individual student achievement is affected not only by the student's own background, but also by the backgrounds of other students in their class.

The Tennessee Value-Added Assessment System (TVAAS), which relies on a model that does not include additional student-level data, has received some criticism within the education community in recent years. Some believe that not including demographic data amounts to a significant flaw in the model, as such factors not only influence the student's starting point, **but they also affect the rate at which the students learn**. The creators of the model attempted to address these criticisms by analyzing how the model is affected when various student-level variables are included.[5]

The authors concluded that the inclusion of socioeconomic and demographic variables at the student level had very little effect on the model.[6] Teacher effects from both models were highly correlated, and both models singled out similar teachers as being significantly above or below average. The authors do concede, however, that not controlling for these variables can be problematic when using a less sophisticated model than TVAAS (i.e. models that rely on only one year of pre-intervention test score data; TVAAS relies on as many as five years of data). They also note that applying TVAAS methodology at the school-level may not be advisable. They note that "we cannot be confident that the TVAAS controls for contextual variables [i.e., demographic and socioeconomic variables at the school level] in the same way that it controls for the influence of student-level SES and demographics."[7]

A study by McCaffrey et al attempts to address this issue further by constructing a more general value added model to analyze the viability of using VAM to measure school and teacher effects.[8] They argue that most of the current VAM's are essentially a restricted version of this more general model, thereby allowing it to serve as a framework with which to compare other VAMs.

When applying their general model to a charter school serving a diverse group of students, they found that the average number of students who were eligible for free/reduced price lunch was highly correlated with student test scores and gains in test scores. They also

[3] Newton, X., et al. 2010. "Value-Added Modeling of Teacher Effectiveness: An Exploration of Stability across Models and Contexts." *Education Policy Analysis Archives.* http://gse.berkeley.edu/admin/events/docs/epaa.pdf
[4] Ibid.
[5] Dale Ballou et al. 2004. "Controlling for Student Background in Value-Added Assessment of Teachers." *Journal of Educational and Behavioral Statistics*.
http://web.missouri.edu/~podgurskym/Econ_4345/syl_articles/ballou_sanders_value_added_JEBS.pdf
[6] Ibid.
[7] Ibid.
[8] McCaffrey, D. et al. 2004. "Models for Value-Added Modeling of Teacher Effects." *RAND Corporation*.
http://www.rand.org/pubs/reprints/2005/RAND_RP1165.pdf

found that free/reduced price eligibility at the school level predicted scores even after controlling for individual student lunch status. These findings led the authors to conclude that using VAM to measure school and teacher effects is problematic, even when controlling for student-level data.[9]

DO MOST VAMS INCLUDE DEMOGRAPHIC VARIABLES?

It might also be instructive to analyze what types of models are most often used in practice. A study conducted by Mathematica Policy Research in 2010 revealed that **all but two of the VAMs they reviewed from the literature controlled for student-level characteristics**.[10] The most common control variables were gender, race/ethnicity, disability/special education status, free or reduced-price lunch status, and English language proficiency level. The researchers also encountered studies which controlled for variables such as parental education, number of hours of television watched during the week, and family income. While they do not claim that there is a clear empirical basis for one type of model over the other, they do state that "there is a compelling argument that one should control for everything that the teacher cannot affect, so recorded student characteristics are included in the models especially when it is virtually costless to do so."[11]

In Figure 1 (see following page), we summarize the models used by several high profile districts and states. Interestingly, the state of Florida, which first introduced its model in 2011, does not include direct controls for measures such as student poverty, though it does attempt to control for certain classroom characteristics.

[9] McCaffrey, Op. Cit.
[10] Lipscomb, S. et al. 2010. "Teacher and Principal Value-Added: Research Findings and Implementation Practices." *Mathematica Poilcy Research*.
http://www.mathematica-mpr.com/publications/PDFs/education/teacherprin_valueadded.pdf
[11] Ibid.

Figure 1: Description of High-Profile Value Added Models

	Description	Inclusion of Poverty	Amount of Data Used
Dallas[12]	Controls for *preexisting student differences* (ethnicity, gender, language proficiency, socioeconomic status, and prior achievement levels), *school level variables* (including mobility, percent minority and socioeconomic status), and *educational variables* (such as student attendance rates and dropout rate).	Yes	1 year of pre-test data. Single cohorts.
Florida[13]	Includes student characteristics (such as Students with Disabilities status, attendance, gifted status, and ELL status) and classroom characteristics (class size, homogeneity of students' entering test scores in class).	No	Up to 2 years of pre-test data.
Los Angeles[14]	Takes into account gender, poverty, number of years in the district, ELL status, educational attainment of the student's parents, class size, student mobility, and 5 levels of English proficiency. It also makes an adjustment for what are called peer effects, the collective characteristics of a class.	Yes	Up to 3 years of cohorts. Multiple pre-tests may also be used.
New York[15]	Variables include race, gender, socioeconomic status, and even whole-class characteristics like the size of the class and how many students are new to the city.	Yes	As many as 4 years (1-year effects are also measured). Multiple pre-tests may also be used.
Tennessee[16]	To reduce the effects of one year of test data, the system also uses data from subsequent years (e.g. a 3rd grade teacher may be evaluated based on how students do in later grades).	No	As many as 5 years of pre and post test data.
Washington, D.C.[17]	Controls for characteristics at student level (e.g. free/reduced lunch status, limited English Proficiency, special education status, and attendance), but not at classroom level.	Yes	1 year of pre-test data. Single cohorts.

Note: Los Angeles and New York systems are not currently fully implemented. They have been included purely for informational purposes.

[12] "Dallas." University of Pennsylvania. http://www.cgp.upenn.edu/ope/21_dallas.html
[13] "Florida's Value Added Model." Florida Department of Education. http://www.fldoe.org/committees/pdf/PresentationValue-addedModel.pdf
[14] "Los Angeles Teacher Ratings." *Los Angeles Times*. http://projects.latimes.com/value-added/faq/
[15] "NYC Teacher Data Initiative: Technical Report on the NYC Value-Added Model." New York City Department of Education. http://schools.nyc.gov/NR/rdonlyres/A62750A4-B5F5-43C7-B9A3-F2B55CDF8949/87046/TDINYCTechnicalReportFinal072010.pdf
[16] Ballou, D. 2010. "Value-Added Assessment: Lessons from Tennessee." http://dpi.state.nc.us/docs/superintendents/quarterly/2010-11/20100928/ballou-lessons.pdf
[17] "Guide to Value Added." District of Columbia Public Schools. http://www.dc.gov/DCPS/Files/downloads/In-the-Classroom/Value-Added%20Guidebook%20(singles).pdf

CONCLUSION

Based on our analysis of the research that has been done on the topic, it does not seem as if there is a broad consensus as to whether the exclusion of student-level characteristics leads to substantial biases in VAMs. As we have seen, the results produced by the two sets of models (the base models, and those that do include the control variables) do seem to produce similar results. However, it appears that the potential benefits of inclusion (i.e. reducing the biases of non-random placement of students) outweigh the costs, which are negligible in most cases.

SECTION II: OTHER BEST PRACTICES

In the previous section, we conclude that the inclusion of student-level demographic data is advisable given that the potential benefits would outweigh the costs, even if the results produced by the two sets of models would likely be similar. In this section, we describe other best practices that we have identified after reviewing the research that has been conducted on the topic.

INCLUDE MORE THAN ONE YEAR OF PRE-TEST DATA

There seems to be a pretty broad consensus that it is preferable to use multiple years of pre-test data, when possible.[18] Using data from multiple years reduces the random fluctuations that are often present when only one year of data is used. As William Sanders, one of the originators of VAMs, recently noted, "When any one student takes a math test, on any one day, there is a huge uncertainty around that score. It could be the kid got lucky this year, and guessed two or three right questions. Or the kid this morning could not have been feeling well. Consequently that score on any one day is not necessarily a good reflection of a kid's attainment level." For this reason, Sanders uses pre-test data from multiple years in his model.[19]

WHEN POSSIBLE, USE MULTIPLE COHORTS TO EVALUATE TEACHERS

For districts with a low degree of student and staff turnover, using multiple cohorts to produce teacher effect estimates could increase sample sizes and reduce the random fluctuations that occur when teachers are evaluated on the basis of a small number of students. When this methodology is used, teachers who happen to be assigned a large number of low-growth students in one year are less likely to be penalized (assuming that the distribution of low-growth students evens out over a longer period). This seems to be supported by the literature, as sorting bias is less pronounced when three to four cohorts of students are combined in the VAMs.[20]

CONSIDER USING MULTIPLE ASSESSMENTS TO CONTROL FOR PRE-TEST ACHIEVEMENT

As noted earlier, relying on one data point to control for pre-test scores can be problematic, as random variation in scores can lead to imprecise estimates for teacher effects. In addition to using multiple years of test data, it is also possible to use data from multiple assessments in order to correct for this potential pretest measurement error. Several

[18] We should acknowledge that accumulating multiple years of data may be difficult for districts with a high degree of turnover.

[19] Butrymowicz, S. and Garland, S. 2012. "How New York City's Value-Added Model Compares to What Other Districts, States Are Doing." The Hechinger Report. http://hechingerreport.org/content/how-new-york-citys-value-added-model-compares-to-what-other-districts-states-are-doing_7757/

[20] Koedel, C. and Betts, J. 2009. "Does Student Sorting Invalidate Value-Added Models of Teacher Effectiveness? An Extended Analysis of the Rothstein Critique." Working Paper.

recent models have incorporated multiple pretest scores into VAMs in order to control for pretest achievement. While some use pretest data from the previous year only, others incorporate data collected from multiple years.[21]

THINK ABOUT INCORPORATING CLASSROOM EFFECTS

Many have argued that the characteristics of a classroom may impact the rate at which students learn. For instance, a student may not be as likely to improve if they are in a classroom where the majority of students are living under poverty. Proponents of this theory therefore argue that it is not sufficient to simply control for the demographics of the individual student, and that districts must also control for the effects of classroom characteristics.

A study on the Los Angeles Unified School District used estimated teacher effects from the pre-experimental period to predict student performance in the experimental period. The study revealed that teacher effect estimates that controlled for mean classroom characteristics yielded the best prediction accuracy.[22] This study, along with others, likely explains the growing trend towards incorporating classroom effects into these models. For instance, the Florida, Los Angeles, and New York models outlined in the previous section all control for these variables. These models generally use classroom averages of variables (such as class size, number of students on free/reduced price lunch status, etc.) as controls.

CONSIDER ADJUSTING ESTIMATES TO ACCOUNT FOR SMALL SAMPLES

One weakness of traditional value added models is that teachers with a low number of students are often overrepresented in the high and low ends of the performance distribution. The estimates for these teachers are often imprecise, as strong or weak test scores may have been affected by other factors that are outside of the control of teachers (e.g. guessing answers correctly, illness, etc.). This is concerning because outliers such as these are more likely to have a significant impact on a teacher's estimate if they have a low number of students.

One way to account for this is to use an adjustment known as empirical Bayes estimation or shrinkage. When incorporating this adjustment, the teacher's estimate is a weighted average of his or her own initial estimate and the mean estimate of all the teachers in the sample, with the more precise estimates (i.e. those with a greater sample) receiving greater weight. Under this approach, teachers are assumed to be average in performance until there is enough evidence to reject this hypothesis.[23]

[21] Libscomb, Op cit.
[22] Kane, T. and Staiger, D. 2008. "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation," *National Bureau of Economic Research*, http://economics.stanford.edu/files/staiger.pdf
[23] Lipscomb, Op. cit.

A review of the literature by Mathematica Policy Research also found that many studies limit analyses to teachers who have taught a certain minimum number of students. Their review of the literature indicated the cutoff, which is typically chosen in an ad-hoc manner, ranged from 5 to 20 students.[24] We should acknowledge, however, that this may not be feasible given the evaluation requirements to which many districts must abide.

FINAL THOUGHTS

Our research has indicated that there is not yet a clear, broad consensus as to what model best measures teacher effects while also minimizing estimation errors. Nonetheless, there are some areas where there is some agreement among experts. For instance, most experts agree that districts should attempt to account for the random fluctuations that occur when a model relies on just one data point to control for pre-test achievement. Our hope is that the guidelines presented in this section will help the District account for this consideration.

[24] Lipscomb, Op. cit.

Project Evaluation Form

Hanover Research is committed to providing a work product that meets or exceeds member expectations. In keeping with that goal, we would like to hear your opinions regarding our reports. Feedback is critically important and serves as the strongest mechanism by which we tailor our research to your organization. When you have had a chance to evaluate this report, please take a moment to fill out the following questionnaire.

http://www.hanoverresearch.com/evaluation/index.php

Caveat

The publisher and authors have used their best efforts in preparing this brief. The publisher and authors make no representations or warranties with respect to the accuracy or completeness of the contents of this brief and specifically disclaim any implied warranties of fitness for a particular purpose. There are no warranties which extend beyond the descriptions contained in this paragraph. No warranty may be created or extended by representatives of Hanover Research or its marketing materials. The accuracy and completeness of the information provided herein and the opinions stated herein are not guaranteed or warranted to produce any particular results, and the advice and strategies contained herein may not be suitable for every member. Neither the publisher nor the authors shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages. Moreover, Hanover Research is not engaged in rendering legal, accounting, or other professional services. Members requiring such services are advised to consult an appropriate professional.

1750 H Street NW, 2nd Floor
Washington, DC 20006

P 202.756.2971 F 866.808.6585
www.hanoverresearch.com