› White Paper

Addressing Common Concerns About Value-Added Modeling

THE POWER TO KNOW.

Contents

Introduction: SAS® EVAAS® Services and Models 1

Common Concerns About SAS® EVAAS® 1

Concern 1:
Value-added models rely on standardized tests, which have limitations themselves..... 1

Concern 2:
Missing student test data jeopardize the validity of the analyses. 2

Concern 3:
Potential for rewards and punishments is related to class size (shrinkage estimation). 3

Concern 4:
SAS® EVAAS® does not adjust for socioeconomic factors..... 3

Concern 5:
SAS® EVAAS® modeling lacks transparency and is too complex. 4

Concern 6:
SAS® EVAAS® statistical methods and algorithms have not been peer reviewed. 4

Concern 7:
SAS® EVAAS® predictions of student performance are not verified later. 5

Using EVAAS Results in Formative Ways 5

References 6

Authors

William L. Sanders
S. Paul Wright
June C. Rivers
Jill G. Leandro
November 2009

Introduction: SAS® EVAAS® Services and Models

SAS EVAAS provides analytical services, including value-added and projection analysis, for the assessment of schooling effectiveness at the district, school and, when requested, at the classroom level. The projection methodology predicts individual students' chances for success at future academic milestones. The results of these analyses, along with additional diagnostic information and querying capabilities, are made available via a secure Web application.

SAS EVAAS has provided these assessments for more than 15 years, and the concept of value-added has become a prominent part of the national education discussion during this time. More specifically, the U.S. Department of Education has emphasized the importance of value-added assessment models in evaluating teaching effectiveness as part of its criteria in distributing stimulus funds through the Race to the Top program. Such national emphasis has led proponents and opponents of value-added to speak up louder than ever before. As the most established of these systems, SAS EVAAS has taken the brunt of the detractors' criticisms. This document addresses common concerns raised about value-added models in general as well as the SAS EVAAS models specifically.

It should first be noted that there is not one, single EVAAS value-added model used in all applications; rather there are multiple models implemented according to the objectives of the analyses and the characteristics and availability of the test data. Two general types of value-added models are employed. The multivariate response model (MRM) is a multivariate, longitudinal, linear mixed model. In traditional statistical terminology it is essentially a multivariate repeated-measures ANOVA model. With this approach, the entire set of observed test scores belonging to each student is fitted simultaneously. When the data have been scaled or transformed to allow comparable expectations of progress – evaluated over many schools and/or districts – regardless of entering levels of groups of students, then the MRM approach is preferred. Details of this model as applied to districts, schools, and classrooms are given in Sanders, Saxton and Horn (1997).

When the data structures do not meet the requirements for a MRM analysis, a univariate response model (URM) is employed. This model is similar to traditional analysis of covariance (ANCOVA): student scores in a particular subject-grade-year serve as the response variable; these students' prior scores in multiple subjects-grades-years serve as covariates or predictor variables; the group or classification variable is an educational entity (district, school, classroom). The URM differs from traditional ANCOVA in that the group variable is treated as random rather than fixed.[1] In this respect, the URM has much in common with certain hierarchical linear models (HLMs) that have been used for value-added analyses. To minimize selection bias and to minimize problems caused by errors of measurement in the predictor variables, we require that each student must have at least three prior scores. However, all available prior achievement test scores for each student are used in the predictor variable set. A method for accommodating fractured records for models of this type is outlined in Wright, Sanders and Rivers (2006).

Addressing Common Concerns About Value-Added Modeling has been aimed primarily at the MRM as applied to classroom assessment, but most of what follows applies to all EVAAS value-added models.

Common Concerns About SAS® EVAAS®

Concern 1: Value-added models rely on standardized tests, which have limitations themselves.

Student test scores are the basic ingredient of all SAS EVAAS analyses. SAS EVAAS is not involved in, and has no control over, test construction. However, before using any tests in EVAAS modeling, we require assurances and conduct exploratory analyses to verify that tests have the required psychometric properties. We require evidence that the tests (1) are reliable, (2) are highly correlated with curricular objectives, and (3) have sufficient stretch in the reporting scale to measure the achievement of both very low and very high achieving students in a grade and subject. To date, we have found only one battery of state CRT tests that did not meet these three criteria (and this test is no longer in use).

[1] The difference between random and fixed effects can be explained as follows: "If the effect level can reasonably be assumed to represent a probability distribution, then the effect is random. If an effect does not represent a probability distribution, it is fixed." Littell, Stroup and Freund (2002).

Concern 2: Missing student test data jeopardize the validity of the analyses.

A reality of student testing is that students move in and out of schools, and they miss tests throughout their academic career. While such student testing histories complicate the analyses, these challenges are not insurmountable and simply require a more sophisticated approach. In the test data received from many schools and districts, statistics show that missing data do not happen at random but in a pattern that is consistent with the population of students served by the schooling entity. In particular, lower achieving students are more likely to have missing test scores. This missing data can create selection bias. If not properly addressed, the inclusion of students with certain test scores and the exclusion of students without certain test scores can seriously bias the estimates of the schooling influences on the rate of student academic progress.

In many less sophisticated analyses, students with missing scores are simply omitted from the analysis. This virtually assures that there will be selection bias, and it also increases the uncertainty in the estimates. One of the major advantages of the multivariate, longitudinal modeling approach used by SAS EVAAS (and others) over the more simplistic models is that it does not require complete data for each student; all available data on each student are used. This reduces the uncertainty in the EVAAS estimates. It also minimizes the selection bias, roughly speaking, by using each student's observed scores (in multiple subjects over multiple years) to "invisibly predict" the missing scores. Technically speaking, SAS EVAAS models assume that scores are "missing at random" (MAR) while less sophisticated models rely on the stronger assumption of "missing completely at random" (MCAR). Details of this distinction can be found at the Carpenter & Kenward website: missing-data.org.uk. In an investigation of the bias reduction properties (or bias compression in their terminology) of multivariate, longitudinal models (such as the SAS EVAAS MRM), Lockwood and McCaffrey (2007) reach the following conclusion.

> This suggests that our general findings about the bias compression of the mixed models approach are not invalidated by the complexities of missing data, but it is likely that incompleteness in the test score data will in general degrade the bias compression to some extent. On the other hand, the mixed models approach makes use of all of the information available for each student in estimating the unknown parameters … and so might lead to particular efficiency gains relative to other approaches when missing data are substantial.

A second concern related to missing data is the problem of identifying which teachers taught which students. Getting an accurate linkage between who actually taught each student in each subject and for what percentage of the instructional time is a major challenge, especially in instances of team teaching and departmentalized instruction. SAS EVAAS modeling ensures its accuracy at the classroom level by requiring the supplier of the data to certify the appropriate linkages in some way. In one state, for instance, individual teachers log on to a website to claim each student and certify that each student met the attendance requirement of the policy that is in place in that state. This process addresses the need for accuracy and accountability in the student-teacher linkages such that appropriate assessments of effectiveness can be made at the classroom level.

To our knowledge, SAS EVAAS is the only supplier of value-added modeling services that accommodates team teaching and other forms of fractional attribution of student test scores to multiple teachers. (This has been done by others in a research setting, but not for production.) Details are given in Sanders, Saxton and Horn (1997).

Concern 3: Potential for rewards and punishments is related to class size (shrinkage estimation).

Many value-added models, including those of SAS EVAAS, use a statistical process called shrinkage estimation, more properly called best linear unbiased prediction (BLUP) and also known as empirical Bayes estimation (Raudenbush and Bryk, 2002). The very existence of such varying terminology is a consequence of its widespread use in a variety of applications. This methodology assumes that every teacher is "average" until the data show otherwise. The use of shrinkage estimation protects individuals from receiving a spurious estimate due to the unlucky accumulation of random errors. This is especially important in the case of teachers having small numbers of students. There are also statistical arguments in favor of shrinkage estimation. Theoretically, it is known that shrinkage estimation provides the maximum correlation between estimated and "true" effects (Searle, et al., 1992, pp. 263-264). In a recent presentation, McCaffrey, et al. (2008) compared results from 24 value-added models and reported the following:

> Multivariate mixed models, fixed effects with shrinkage and ANCOVA with shrinkage, all have high levels of consistent information relative to the noise. Shrinkage increases the correlation by reducing the noise relative to the consistent information about teachers.

McCaffrey, et al. (2008) also note that models that do not use shrinkage estimation would produce a disproportionate number of classroom estimates based upon very small numbers of students at the extremes of the distribution. These classroom estimates, which are based on very small numbers, will inevitably lead to a much lower repeatability between estimates in adjacent years, which in turn will lead to heightened suspicion about the value-added process itself. Thus, a distinct advantage of shrinkage estimates is the greater repeatability between estimates in adjacent years. Additionally, as reported by McCaffrey et al. (2009), estimates for a teacher's effectiveness, from three student cohorts of data, would approach a reliability coefficient of 0.8.

Concern 4: SAS® EVAAS® does not adjust for socioeconomic factors.

A key principle of SAS EVAAS value-added modeling is to follow the progress of individual students. Consequently, the models include a student's entire testing history in multiple subjects (multivariate) over multiple years (longitudinal). However, socioeconomic (SES) and demographic (DEM) variables are not explicitly included in these models, either at the student level or at any higher (classroom, school, community) level, leading to concerns about unfairness.

At the student level, by including all of a student's testing history, each student serves as his or her own control. To the extent that SES/DEM influences persist over time, these influences are already represented in the student's data. This negates the need for SES/DEM adjustment. This was confirmed empirically by Ballou, et al. (2004) and by Lockwood and McCaffrey (2007) who conclude:

> William Sanders … has claimed that jointly modeling 25 scores for individual students, along with other features of the approach, is extremely effective at purging student heterogeneity bias from estimate teacher effects … The analytical and simulation results presented here largely support that claim.

On a philosophical level, the question educators should ask is whether they should have lower expectations for a student from a poor family than one from a rich family, even when the two students have identical test scores and academic history. By adjusting for these variables, one is directly assuming that there will be different expectations for two students with the same prior achievement pattern who come from different SES/DEM communities. The use of SES/DEM adjustments at the student level has largely been discouraged among statisticians and policymakers involved with value-added modeling, including the policies developed for Adequate Yearly Progress in growth model augmentations for No Child Left Behind.

Whether to include adjustment for SES/DEM variables at the group level (e.g., classroom, school, community) is much debated among educators and value-added modelers. To

educators it seems obvious that SES/DEM variables are important. Indeed, it is well documented, and easy to demonstrate with essentially any set of student test scores, that group average student achievement levels (e.g., average test scores) are highly (negatively) correlated with such measures as percent minority and percent in poverty. What is less well documented is whether student academic growth, when properly assessed, is likewise highly correlated with SES/DEM variables. What we at EVAAS have done over the years is to check to see whether this is in fact the case: We have repeatedly checked to see whether our value-added estimates are correlated with student characteristics (mainly poverty and ethnicity) summarized at the school or classroom level. We have found that the correlations vary from place to place, and they vary by academic subject; but the correlations are modest at worst and essentially zero at best. Furthermore, even when we see modest correlations, there is wide variation in school and classroom effectiveness across the distribution of, say, classroom poverty level. That is, there are very effective teachers (and very ineffective teachers) in classrooms full of poor students, and there are very effective (and very ineffective) teachers in classrooms full of affluent students. The gross unfairness that educators fear is not consistently present.

Consequently, we recommend that these adjustments not be made. Not only are they largely unnecessary, but they may be harmful. This is because patterns of assignment of teachers to schools are often related to teacher effectiveness. It has been documented in many studies that novice teachers are less effective than veteran teachers; it has also been documented that schools with a higher concentration of poor and minority students also get a disproportionate number of beginning teachers (Mayer, et al., 2000). In this scenario, adjustment for group SES/DEM factors will over-adjust the estimates and can camouflage the fact that students in certain schools are not getting an equitable distribution of the teaching talent. By excluding such adjustments, SAS EVAAS models are better able to highlight this disparity than models that make adjustments for SES/DEM variables.

Ultimately, the decision on whether or not to adjust for group SES/DEM variables depends on where the risks are to be placed. Even though we advocate for no adjustment, we certainly can make such adjustments if states and districts elect to use them.

Concern 5: SAS® EVAAS® modeling lacks transparency and is too complex.

Policymakers have long debated the trade-offs between simplicity and complexity of value-added models. To extract the most reliable estimates of the impact of various schooling entities on the rate of student academic progress, the models must be statistically complex. Yet many argue that if these estimates are to be used in a summative way, then the calculations must be simple enough that anyone with a minimum of instruction could duplicate the results. Such reasoning has led to less sophisticated approaches that are vulnerable to the problems of selection bias and increased uncertainty discussed above.

However, newer research has shown how egregiously bad the results from these simplistic approaches can be (Sanders, 2006; McCaffrey, et al., 2008). These simplistic approaches may over-identify either very ineffective or very effective teachers and lack the year-to-year reliability of more sophisticated value-added models (see Criticism 3). To trade simplicity of calculation for reliable information is a "devil's bargain." Policymakers should be advised that if these short-cut attempts at value-added assessment are deployed, then the lack of reliability in these estimates will be apparent the second and third year after deployment. Rather than focus on simplicity of calculation, SAS EVAAS models prioritize reliability of analysis and then focus on ease of interpretation and ease of usage.

Concern 6: SAS® EVAAS® statistical methods and algorithms have not been peer reviewed.

It has been asserted that SAS EVAAS developers have not made their methods and algorithms available for peer review. Broadly speaking, the models used in SAS EVAAS are linear mixed models for which algorithms have been employed for decades. Models of this type are widely available in publicly available software such as SAS®, SPSS®, and R. More narrowly, the specific statistical models used in SAS EVAAS are given in detail in Sanders, Saxton and Horn (1997) and are explained quite well in McCaffrey, et al. (2004). Apparently the models, and the algorithms necessary to solve them, are sufficiently well understood to have been implemented to varying extents by researchers at RAND (McCaffrey, et al., 2004; Lockwood, et al., 2007; McCaffrey, et al., 2008), by Raudenbush and Bryk (2002, Chapter 12), by R programmers (Lockwood, et al., 2003; Bates 2007) and others (Briggs and Weeks, 2008). It is clear that the SAS EVAAS models are well understood by many other value-added modelers.

Concern 7: SAS® EVAAS® predictions of student performance are not verified later.

SAS EVAAS can project future student performance on benchmark tests like the SAT, ACT or end-of-grade/end-of-course tests. This service of SAS EVAAS is distinct from the value-added modeling assessments, a distinction which critics often miss, resulting in confusion. It has been asserted that no one has followed up to confirm if those predictions came to pass. This is untrue.

To date, three states using the SAS EVAAS projection methodology have been approved in the growth model pilot program of No Child Left Behind by the United States Department of Education. To achieve this approval, the methodology was reviewed by four different peer review teams. One of the peer review teams, prior to its approval, specifically required an analysis using historical data to ascertain the reliability of the projections by comparing projections with subsequent observed scores. One of the findings of this analysis was that, by using all of each student's prior test scores from multiple grades and subjects to make the projections, one could predict three years in advance with more accuracy than predicting one year ahead using a single prior test score. Additionally, the methodology and software to produce the projections were reviewed by the Government Accountability Office (GAO), which verified that they produced the estimates as outlined in Wright, Sanders, and Rivers (2006).

Using EVAAS Results in Formative Ways

Once a multivariate, longitudinal data structure and modeling results are in place, there is a wealth of positive diagnostic information available for educational decision makers – teachers, principals, curricular specialists, superintendents, school board members, etc. The use of value-added measures as one component of accountability systems is important, but in our view, the diagnostic information is of greater importance. For those districts for which we are providing analytical services, a series of reports for each school is produced. These are delivered via the Web and can be accessed only by individual educators who have authorized passwords. These reports include the following.

For each grade and subject, progress rates of students are presented by prior achievement level, either for the whole school or any demographic subset, with comparisons with previous cohorts. This enables educators within each school to ascertain which subset of students is not making the appropriate progress.

For each student, projections are made to various academic endpoints. This enables local educators to identify which students, say, are not on trajectories to meet high school graduation requirements with sufficient time to plan different curricular and instructional strategies for these at-risk students, or to identify students who are meeting all proficiency requirements yet are not on a trajectory to be prepared for a more technical college major.

Some principals and superintendents have learned to use the flexible projection reports to plan for the number of "seats" that will be required to accommodate all students who are on trajectories to be successful in Algebra as 8th graders, based on the students' projections at the end of 6th grade. It has been found that in some schools, the number of "seats" available is considerably less than the number of students who could benefit from a more rigorous course.

Some teachers are finding it to be helpful to have all of the prior testing information available in an intuitively understandable Web interface for each student as they enter their classrooms.

Do any of these reports replace the need for good, ongoing formative assessment within the classroom? Of course not. However, the Web delivery does enable educators to access this body of information at their convenience and at their chosen location; and it gives, in simple-to-understand reports, reliable information based on rigorous analysis. This provides a way to minimize the conundrum between rigorous analytical procedures and simple-to-understand reports for teachers' and principals' professional use.

References

Ballou, D., Sanders, W., and Wright, P. (2004). "Controlling for Student Background in Value-Added Assessment of Teachers." *Journal of Educational and Behavioral Statistics*, Vol. 29, No. 1, pp. 37–66.

Bates, D. M. (2007). *Computational Methods for Mixed Models*. Retrieved from cran.r-project.org/web/packages/lme4/vignettes/Theory.pdf.

Briggs, D. C., and Weeks, J. P. (2008). *The Persistence of Value-Added School Effects*. Paper presented at the annual meeting of the American Educational Research Association, March 27, 2008. Retrieved from colorado.edu/education/sites/default/files/attached-files/BW_AERA_Persistence_032708.pdf on Nov. 10, 2009.

Littell, R.C., W.W. Stroup and R.J. Freund (2002). *SAS® Linear Models*, Fourth Edition, p. 92. Cary, NC: SAS Institute Inc.

Lockwood, J. R., Doran, H., C., and McCaffrey, D. F. (2003). "Using R for Estimating Longitudinal Student Achievement Models." *R News: The Newsletter of the R Project*, Vol. 3, No. 3, pp. 17-23.

Lockwood, J. R., and McCaffrey, D. F. (2007). "Controlling for Individual Heterogeneity in Longitudinal Models, with Applications to Student Achievement." *Electronic Journal of Statistics*, Vol. 1, pp. 223-252. RAND reprint available at rand.org/pubs/reprints/RP1266.html.

Lockwood, J. R., McCaffrey, D. F., Mariano, L. T., and Setodji, C. (2007). "Bayesian Methods for Scalable Multivariate Value-Added Assessment." *Journal of Educational and Behavioral Statistics*, Vol. 32, No. 2, pp. 125-150.

Mayer, D. P., Mullens, J. E., and Moore, M. T. (2000). *Monitoring School Quality: An Indicators Report*, NCES 2001–030. Washington, DC: U.S. Department of Education, National Center for Education Statistics. John Ralph, Project Officer. Retrieved from nces.ed.gov/pubs2001/2001030.pdf on Nov. 10, 2009.

McCaffrey, D. F., Han, B. and Lockwood, J. R. (2008). *From Data to Bonuses: A Case Study of the Issues Related to Awarding Teachers Pay on the Basis of the Students' Progress*. Paper presented at the conference on Performance Incentives: Their Growing Impact on American K-12 Education, February 28-29, National Center on Performance Incentives at Vanderbilt University's Peabody College.

McCaffrey, D. F., Lockwood, J. R., Koretz, D., Louis, T. A., and Hamilton, L. (2004). "Models for Value-Added Modeling of Teacher Effects." *Journal of Educational and Behavioral Statistics*, Vol. 29, No. 1, pp. 67-101. RAND reprint available at rand.org/pubs/reprints/RP1165.html.

McCaffrey, D. F., Sass, T. R., Lockwood, J. R., and Mihaly, K. (2009). "The Intertemporal Variability of Teacher Effect Estimates." *Education Finance and Policy*, Vol. 4, No. 4, pp. 572-606.

Raudenbush, S. W., and Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*. Thousand Oaks, CA: Sage Publications.

Sanders, W. L. (2006). *Comparisons Among Various Educational Assessment Value-Added Models*. Paper presented at The Power of Two – National Value-Added Conference, October 16, 2006, Columbus, Ohio.

Sanders, W. L., Saxton, A. M., and Horn, S. P. (1997). "The Tennessee Value-Added Assessment System: A Quantitative, Outcomes-Based Approach to Educational Assessment." Pages 137-162 in J. Millman (Ed.), *Grading Teachers, Grading Schools: Is Student Achievement a Valid Evaluation Measure?* Thousands Oaks, CA: Corwin Press.

Searle, S. R., Casella, G., and McCulloch, C. E. (1992). *Variance Components.* New York: Wiley.

Wright, S. P., Sanders, W. L., and Rivers, J. C. (2006). "Measurement of Academic Growth of Individual Students toward Variable and Meaningful Academic Standards." Pages 385-406 in R. Lissitz (Ed.), *Longitudinal and Value Added Models of Student Performance*. Maple Grove, MN: JAM Press.

To contact your local SAS office, please visit: sas.com/offices

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies. Copyright © 2014, SAS Institute Inc. All rights reserved.
107101_S125192.0514

THE POWER TO KNOW.