A Response to Amrein-Beardsley (2008) "Methodological Concerns about the Education Value-Added Assessment System"

William L. Sanders
S. Paul Wright

Introduction

The Amrein-Beardsley paper attempts to give a summary of various issues associated with value-added assessment models both in general and with special focus on the methodology used in the Tennessee Value-Added Assessment System (TVAAS). SAS Institute Inc. now provides comparable analytical services to various schools, districts and states using various multivariate, longitudinal statistical approaches available from the SAS® EVAAS® application service provider group. The objective of this response is to give the rationale for the methodology that we use, to offer evidence for its robustness, and to note how this approach mitigates and dampens to near triviality many of the concerns expressed by Amrein-Beardsley.

Most of the questions raised by Amrein-Beardsley are not new. Some of the assertions question statistical concepts, others are psychometric questions, and some are policy concerns. However, most are directed toward the assertion that the whole approach which we deploy has not been adequately peer-reviewed. This response will address these assertions, specifically statistical questions, validity concerns, how to use the data in formative ways, and National Board-Certified Teachers and Student Achievement issues.

Statistical Questions

Missing data

Unlike what is implied in this paper, there is not one EVAAS model used in all applications; rather there are multiple models deployed according to the objectives of the analyses and the availability of the data. When we provide value-added analyses, there

are two general types of models fitted: multivariate response models (MRM) and univariate response models (URM). 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. *This approach is outlined in detail in one of the Amrein-Beardsley references (Sanders, et al., 1997).* With this approach, the entire observational vector of each student's test data is fitted simultaneously.

One of the many advantages of this approach is that selection bias is reduced to a point of no reasonable concern. The implicit assumption is that the data are missing not completely at random; rather the data are missing in a pattern that is consistent with the population of students served by this schooling entity—this has been referred to as missing-at-random (MAR). *Details of these distinctions can be found at the Carpenter & Kenward website:* www.missingdata.org.uk/jargon_web/. By fitting the entire observational vector, estimates from fitting the mixed model equations have little selection bias remaining. This has been found and reported on by Lockwood and McCaffrey (2007).

> *"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."*

When the data structures do not meet our requirements for a MRM analysis, we apply the URM models. In this approach, which is similar to traditional analysis of covariance, to minimize selection bias and to minimize the problem with errors of measurement of the predictor variables, we require that each student must have at least three prior scores. However, once this criterion is met, all prior achievement test data for

each student is used in the predictor variable set.[1] Notice that in both of these model types achievement gain is *not* used as a dependent variable as stated by Amrein-Beardsley. A more careful reading of all the references to the work cited would have revealed this factual error.

However, the problem of potential selection bias due to missing data is of serious concern in value-added assessment modeling generally. Many of the more simplistic approaches to value-added assessment ignore the problem which can lead to non-trivial biases of the estimates of the schooling influences on the rate of student academic progress. This is one of the reasons that in EVAAS modeling efforts we have chosen the more rigorous approaches. Additionally, these approaches avoid the need for any data imputation procedures.

A second aspect of the missing data problem is that of identifying which teachers taught which students. The author is correct that this is a major problem. Getting an accurate linkage between who actually taught each student for each subject for what percentage of the instructional time is a major challenge. When asked to provide classroom level value-added estimates, we require that appropriate linkages must be certified in some way by the supplier of the data. In the TVAAS case, individual teachers log-on to a web site—not built or maintained by SAS—to claim each student and certify that each student met the attendance requirement of the policy which is in place in that state. In another place, teachers confirm that the teacher-to-student linkages in the administrative database are accurate. The supposition by Amrein-Beardsley that we carelessly take for granted whatever comes from administrative databases is just not true.

Regression to the Mean

In this section of the Amrein-Beardsley paper the author asserts, in effect, that shrinkage estimation is a bad thing. The expression "bad thing" is used as a deliberate contrast to Robinson's (1991) article in *Statistical Science* with the title "That BLUP is a Good Thing: The Estimation of Random Effects." BLUP (best linear unbiased prediction)

[1] The ideas for accommodating fractured records for models of this type are outlined in Wright, et al. (2006).

is another name for shrinkage estimation. Another commonly used terminology is empirical Bayes estimation (Raudenbush and Bryk, 2002). The very existence of such a varying terminology is a consequence of the widespread use of shrinkage estimation in a variety of academic fields.

Especially at the classroom level, best linear unbiased prediction is the more desirable and preferred method for providing estimates of the impact of classrooms on the rate of student academic progress for the following reasons:

1. It has been proven theoretically that shrinkage estimation provides the maximum correlation between the estimate and the "true" effect (Searle, et al., 1992, pp. 263-4), and
2. The shrinkage estimates provide protection against spurious estimates due to too little data.

Others have found advantage to shrinkage estimation. 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."*

However, it is recognized that others, especially some economists, prefer to model classroom differences as fixed effects. McCaffrey, et al. (2008) framed the distinction between the two approaches by stating that without shrinkage there would be 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.

Extraneous Variables

This is the only section that, in some sense, singles out the models recommended by the EVAAS group as different from other value-added models. Unlike our advocacy that socioeconomic (SES) variables should not be explicitly included in value-added modeling efforts, it is indeed a fact that many other proposed value-added models include controls for a variety of student-level and higher-level (classroom, school, community) demographic and socioeconomic variables. The following are the reasons for our advocacy that these variables should *not* be included.

1. At the student level, if the entire multivariate, longitudinal data vector is fitted, then the inclusion of SES variables is not needed to insure fairness. To the extent that SES factors persist over time, then these influences are already contained in the observational vector. This has been empirically confirmed by Ballou, et al. (2004) and also more recently 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."*

   Additionally, by including 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 communities.

2. Whether to include adjustment for SES variables at the group level is much debated among serious investigators of value-added modeling. We recommend that these adjustments not be made. To the extent that teacher assignment patterns to schools are related to some degree to teacher effectiveness, then adjustment for group SES 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. 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 (see National Center for Education Statistics, *Monitoring Quality: An Indicators Report*, December 2000**,** http://nces.ed.gov/pubs2001/2001030.pdf.) This is one example of why an adjustment for group SES factors may hide influences that policy makers will need to address. The answer to whether or not to adjust for group SES variables depends on where the risks are to be placed. Even though we advocate for no adjustment, we certainly can make SES group adjustments if states and districts elect.

EVAAS® Projection Methodology

Amrein-Beardsley stated,

> "*EVAAS developers recently began using their value-added data to project how far students will go in their educational futures, for example, by predicting what students' scores will be on the ACT when they seek entrance to college (Olson, 2002). In these predictions, no mention is made of whether students will be followed to confirm that the predictions based on the EVAAS model come true, but projections are being made nonetheless (Sanders, 2003). This situation is another set of very troublesome implications and consequences."*

This is another assertion not based in fact but rather on a prejudged conclusion. First, we are proud that two states using our projection methodology have been approved in the growth model pilot program by the United States Department of Education. To achieve this approval, the methodology was reviewed by two different peer review teams. One of the peer review teams prior to its approval required an analysis to ascertain the reliability of the projections. By using historical data, the projections made from earlier years could be compared to the students' scores upon completion of future tests. It was documented that projections three years in advance, using all of each student's prior test scores, were more highly related to final outcome than a single score from the adjacent

year. Additionally, the methodology and software to produce the projections were reviewed by the Government Accounting Office (GAO) and found to produce the estimates as outlined in (Wright, Sanders, and Rivers, 2006).

Peer Review of Statistical Methods and Algorithms

The author's complaint is that *"the developers have not made this method completely open for peer review. Specifically, they hold as proprietary information the computational algorithms needed to manage and solve large systems of linear equations."* On the contrary, the statistical models are specified in detail in Sanders, Saxton and Horn (1997) and are explained quite well in McCaffrey, et al. (2004), both references that Amrein-Beardsley cites.

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), and by R programmers (Lockwood, et al., 2003; Bates 2007a, 2007b). In fact, McCaffrey, et al. (2008) have developed software to fit these models and have published comparisons with results from other value-added models. The author's assertion that the methodology deployed by the SAS EVAAS team has not been reviewed and has had no external evaluation is without factual basis.

On a lesser note, the assertion that Stroup (1995) used the same software to get the same numerical result is totally false; he created a dataset, fitted the model to that dataset using commercially available software, then sent us the data to run with our software. When Stroup compared the results of the calculations from the two series of analyses, he found nearly identical agreement in the results.

Transparency

Amrein-Beardsley stated,

> *"Educators want to use relatively simple, understandable statistical models to analyze educational phenomena, but social complexity demands that statistical*

models be sophisticated enough to capture reality with integrity (Andrejko, 2004; Callendar, 2004). The EVAAS value-added model is caught on the horns of this dilemma."

Ignoring the fact the 'EVAAS value-added model' is referred to in the singular, this quote does point out a conundrum for value-added assessment 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. This has led some to advocate for simple value-added measures such as simple paired-mean gains, or simple regression models with only the previous score as a predictor variable.

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. To trade simplicity of calculation for reliable information, in our view, is a "devil's bargain." Policy makers 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.

Validity

In the Amrein-Beardsley paper, two of the four sections under the Validity heading (Content-Related Validity and Construct-Related Validity) have to do with test construction. They address psychometric issues, not statistical modeling issues. This is not to suggest that these issues are unimportant; they are. Before EVAAS analyses are conducted, assurances are obtained, and exploratory analyses are conducted to verify that the tests have the required psychometric properties. We require evidence that 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 student in a grade and subject.

These requirements are not compromised. An additional note is that with the multivariate, longitudinal analysis, which exploits the entire covariance structure over grades and subjects, the stability of the estimates for schools or classrooms with entering very low and high achieving students is greatly enhanced. In other words, the accumulation of the totality of the information greatly dampens the uncertainty around the extreme test scores. To date, we have only found one battery of state CRT tests which did not meet the three criteria: the original Texas TOSS test had dramatic ceiling effects and we deemed the data did not meet our requirements.

Using Data in Formative Ways

Once a multivariate, longitudinal data structure is completed, 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 summative 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 are produced. These are delivered via the web and can be accessed only by individual educators who have authorized passwords.

Some of these reports are:

- For each grade and subject, presentation of progress rates of students 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.
- Projections for each student 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 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, on-going formative assessment within the classroom? Of course not. However, the web delivery certainly enables 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 is a way to minimize the conundrum between rigorous analytical procedures and simple-to-understand reports for teachers' and principals' professional use.

National Board-Certified Teachers and Student Achievement

Amrein-Beardsley used nearly half of the pages of the paper to assert that our analysis and interpretation (Sanders, et al., 2005) of the NBPTS teacher study were wrong or flawed and to state that we consciously focused on the negative, ignoring the positive findings. In the Conclusions it was stated, *"A paradigm case in which the model was used to advance unfounded assertions was also examined."* It was also stated, "*So what does the reality of this one study suggest about the credibility of findings from other studies using the EVAAS model?"*

There can be only one conclusion as to why half of the pages in the Amrein-Beardsley paper were devoted to commentary on our NBPTS study—an attempt to discredit the findings that have been reported by us with various coauthors in other

publications. But the zeal expressed in this paper to discredit overlooks many important facts.

First, we were commissioned by the NBPTS governing board to do the NBPTS study. We did not seek this assignment; we were recruited to do the study. Our obligation was to produce a report for the governing board; it was not to produce a public document. This report was never made public by us.

Now for the objectives of our study. The NBPTS process has two primary purposes, (1) to select those teachers which by NBPTS standards have met the qualifications to be National Board Certified, and (2) to facilitate a learning experience for teachers as they go through this process. Because of earlier criticism that NBPTS had received from many quarters that definitive third party research had not been completed to ascertain if NBPTS certified teachers were indeed facilitating more student academic progress than other teachers (a selection effect), we were asked to focus primarily on this question. We initially recognized that if we applied our regular mixed model approach to the estimation of teacher effects, we would be vulnerable to criticism that others (at that time) could not duplicate our analyses. So instead, we *deliberately* opted to use models that could be fitted with many different commercially available statistical software packages. The author does not directly acknowledge this fact, rather elected to label this as the 'EVAAS model'.

To test the hypotheses that NBPTS teachers were eliciting more student progress, we made various comparisons between (a) NBPTS certified, (b) NBPTS failed, (c) NBPTS future candidates and (d) teachers who had never tried for board certification. We viewed these groups as "treatments," individual teachers as the sampling units, and students as the sub-sampling units. During the time of completing the final report, another study (Goldhaber & Anthony, 2004) was released. In the Goldhaber & Anthony analyses, it was clear that variation among teachers within "treatment" groups was not partitioned from the model residuals. Subsequently, we added models I and III to our report in an attempt to closely approximate the Goldhaber analyses. When the results

from our I and III models are compared with the Goldhaber & Anthony (2004) models, very similar results were obtained—significant differences with very small effect sizes.

However, there is a more important finding from our study!—variation among teachers who were NBPTS certified was huge and nearly overlapped the distribution of teachers within the other categories! That is, if two candidates are vying for a position with the same credentials except for NBPTS certification, what is the likelihood that if the certified teacher is chosen, the more effective teacher will be hired? Our answer, based on the results of our study, is only slightly better than a coin flip. Thus, models II and IV, which partition the teacher within group variability from the residual and use it as the error term for testing hypotheses, found very few significant differences.

It is our strong contention that we were not making "*unfounded assertions"*; rather the "*unfounded assertions"* perhaps are coming from those reports that are not appropriately recognizing the variability among NBPTS teachers. Amrein-Beardsley stated,

> *"Including all effect sizes, students of NBCT's made about three fourths of a month's greater gains in math (mean ES = 0.08) and about one third of a month's greater gains in reading (mean ES = 0.03) than students of non-NBCT's."*

Even if the means are true, this statement is misleading because it gives no indication as to what is the likelihood of an individual teacher achieving the results of the NBCT population average. Given the variability in effectiveness of NBCT teachers from our study, it is indeed only slightly better than a coin flip. As believers in the NBPTS concepts, we have recommended to the leadership of NBPTS that serious investigation proceed to ascertain what adjustments need to be made to the selection process that will insure that there is a greater separation among the distributions.

Summary and General Comments

The Amrein-Beardsley paper purports to give a summary of general questions concerning the use of value-added models with specific focus on the methodology which is used in SAS EVAAS analyses. To that end, the saddening part of the Amrein-

Beardsley paper is the omission of references to recent papers completed by third party investigators that have confirmed the robustness of the multivariate, longitudinal mixed model approach. There are serious investigators who are working to improve these analytical processes and to have a deeper and better understanding of advantages and limitations of these approaches. As one of the investigators who had read the Amrein-Beardsley paper recently commented, "It is as if this paper had been written seven or eight years ago." Presently, there is sufficient evidence obtained by several investigators to inform policy makers as to how appropriately constructed value-added measures can be an important tool in educational outcome assessment.

If this paper is viewed as an opinion piece, then the author has expressed opinions. But it is also very clear that the author has limited insight into the present state of sophisticated value-added modeling. It is unfortunate that citations from this paper may be taken as fact. We have tried to respond to clear errors expressed by the author while ignoring some of the author's opinions that are not directly related to the title given to the paper.

References

Bates, D. M. (2007). Linear Mixed Model Implementation in lme4. Retrieved from http://cran.us.r-project.org/web/packages/lme4/vignettes/Implementation.pdf.

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

Ballou, D., Sanders, W. L., 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.

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.

Goldhaber, D., and Anthony, E. (March 8, 2004). Can Teacher Quality Be Effectively Assessed. The Urban Institute. Available online at http://www.urban.org/UpLoadedPDF/410958_NBPTSOutcomes.pdf

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.

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.

McCaffrey, D. F., Han, B. and Lockwood, J. R. (2008). From Data to Bonuses: A Case Student 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.

National Center for Education Statistics, *Monitoring Quality: An Indicators Report*, December 2000.

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

Robinson, G. K. (1991). That BLUP is a Good Thing: The estimation of Random Effects. *Statistical Science*, Vol. 6, No. 1, pp. 15-51.

Sanders, W.L., Ashton, J.J, &Wright, S.P. (2005), Comparison of the Effects of NBPTS Certified Teachers with Other Teachers on the Rate Of Student Academic Progress. Arlington, VA: National Board for Professional Teaching Standards. Available at http://www.nbpts.org/UserFiles/File/SAS_final_NBPTS_report_D_-_Sanders.pdf.

Sanders, W. L., Saxton, A. M., and Horn, S. P. (1997). The Tennessee Value-Added Accountability 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.

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.

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

Stroup, W. W. (1995). *Assessment of the Statistical Methodology Used in the Tennessee Value-Added Assessment System*. Knoxville: Tennessee Value-Added Research and Assessment Center.

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.