Abstract Title: Connecting the Pieces: Methodology for Developing Student Achievement Trajectories from Different Instruments
MSP Project Name: Math in the Middle Institute Partnership
Authors: Jennifer L. Green, Walter W. Stroup, Wendy M. Smith, & Ruth M. Heaton
Presenters: Jennifer L. Green & Wendy M. Smith

Summary:
Student achievement data is one tool for documenting the impact of MSP programs on student learning. These data are used to build trajectories of student performance over time. Detecting MSP program impact requires estimates of trajectories both before the MSP begins and after it is in place. While projects implement planned research designs from the start of the MSP, the "before" trajectory requires using available retrospective student achievement data. The data are often less-than-ideal – for example, different achievement tests may have been used in different years preceding the MSP's start. Z-score and binning methodology have been developed to allow estimation of meaningful student trajectories using data from different instruments.

Questions for dialogue at the MSP Learning Network Conference session

How can value-added models be effectively used to estimate teacher effects when student achievement data are from a variety of measures at different times?

How can MSP impact on teacher effects be estimated? Are these effects short-term or can they be sustained?

What issues arise when attempting to develop student trajectories from multiple measures of student achievement?

What are the benefits and limitations associated with Z-score and binning methods for estimating teacher effects when student achievement data are from a variety of assessments?

Conceptual Framework

NSF's MSP programs provide substantial content-based professional development to teachers. Through the MSP program, NSF is making a significant investment in the education of our nation's mathematics teachers. Increasingly, there is considerable interest in finding out which MSP projects appear to be effective in demonstrably increasing student success and thus are candidates for scaling up.

Student achievement data is one tool for documenting the impact of MSP programs on student learning. Value-added modeling techniques aim to estimate teacher and school effects and the changes to such effects that can be associated with specific interventions (such as MSP professional development). Value-added modeling methods provide opportunities to estimate the proportion of variability in achievement or student growth attributable to teachers, as well as estimate an individual teacher's effect on student learning. Of urgent concern is to develop value-added modeling techniques that can be effective with less-than-ideal (e.g., real) school district data on student achievement.

Professional development programs focus on improving teachers' abilities to provide quality instruction, but rigorous evaluations are needed to determine whether these programs are actually effective. Value-added modeling techniques provide opportunities to estimate the relationship between teacher development and student learning, but most require student achievement data to be on a single developmental scale over time (McCaffrey, Lockwood, Koretz, & Hamilton, 2003). Typically, available assessment data do not meet such requirements, limiting analyses that can be conducted.

The Math in the Middle (M^2) Institute, an NSF MSP program for middle-level mathematics teachers, has significantly impacted mathematics teaching and learning in Nebraska and resulted in institutional change that will impact mathematics education in the state for decades to come. The M^2 Institute defines student success fairly broadly: achievement test scores, rubric scores on alternative assessments, and habits of mind of mathematical thinkers. Students with habits of mind of mathematical thinkers are flexible in their thinking, persistent, have a toolbox of strategies they know how to use, and communicate mathematical reasoning clearly. For the purpose of this paper, we focus on how to make sense of the achievement data we have available from our partner school districts. Because Nebraska has had no common statewide mathematics test (until 2011), each school district in Nebraska was free to choose whatever student achievement measure it deemed appropriate; a variety of criterion- and norm-referenced tests were administered to various grade levels at various points during the school year across districts. Consequently, student achievement scores from different districts are not directly comparable. Considering these issues, attention is restricted to one of the larger participating school districts, Middleview Public Schools[1] (MPS). M^2 researchers have used the longitudinal student achievement data shared by MPS, a core partner district, to document the impact of the program on student achievement. Challenges arose, because MPS has a history of using different test for students at different grade levels. Consequently, scores are not on a single developmental scale. Methodology was needed to connect achievement data from different instruments and build trajectories of student learning over time. The situation of different instruments being used to measure student achievement is not unique to MPS; thus, statistical methods to handle this situation are needed.

In order to document MSP impact on student learning trajectories and teacher effects, all the data need to be placed on a common basis. Alternative value-added methodology, such as the use of Z-scores and binning by quantile, are two possible ways to analyze less-than-ideal longitudinal student achievement data collected from a mixture of norm- and criterion-referenced assessments to estimate the impact of a professional development program on student learning. Specifically, value-added models to estimate teacher effects were adapted to use with “messy” data, such as situations when different tests are administered in different grade levels at different times of year. The use of Z-scores and binning by quantile in a value-added context can create a coherent picture of student achievement trajectories across different assessments given at different points in time. Subsequently, these methods offer the potential to connect student achievement trajectories to measures of teaching quality, as well as to measures of student and teacher attitudes in order to create a coherent picture of mathematics teaching and learning.

Explanatory Framework

Cross-classified models (Raudenbush & Bryk, 2002) and the Educational Value-Added Assessment System (EVAAS) model (Sanders, Saxton, & Horn, 1997) are currently recommended over other models to provide estimates of teacher effectiveness. The EVAAS model is a longitudinal linear mixed effects model that has each student serve as his or her own control, similar to the cross-classified model, which models individual growth curves. Using the EVAAS model, Sanders et al. (1997) have been able “to produce estimates of school and teacher effects that are free of socioeconomic confoundings and do not require direct measures of these concomitant variables” (Wright, Horn, & Sanders, 1997, p. 58). In fact, Sanders (2000) claims, “Our research work…clearly indicates that differences in teacher effectiveness is the single largest factor affecting [students’] academic growth” (p. 334); teachers are the dominant factor impacting student progress (Sanders, 2004; Wright et al., 1997). Darling-Hammond (2000) adds, “[E]ffects of well-prepared teachers on student achievement can be stronger than the influences of student

[1] Names are pseudonyms.

background factors, such as poverty, language background, and minority status" (Conclusions and Implications, ¶ 6).

Studies investigating value-added teacher effects provide evidence teachers have differing effects on student learning (Rivkin, Hanushek, & Kain, 2005; Rowan, Correnti, & Miller, 2002; Wright et al., 1997) that persist over time (Sanders & Rivers, 1996). Wright and Sanders (2008) distinguish between the layered and non-layered model in the construction of the coefficient matrix for teacher effects. In the non-layered model, each student's outcome in a given year is linked only to the current teacher. In contrast, the layered model links a student's achievement to current and previous teachers within a given time span. This approach accounts for the "correlation of future scores for students who [have] shared a past teacher" (Lockwood, McCaffrey, Mariano, & Setodji, 2007, p. 126).In both the cross-classified and the EVAAS teacher models, teacher effects persist undiminished into the future, so contributions of both current teachers and past teachers are accounted for in a student's set of scores. Consequently, the total teacher contribution to the variability of scores increases over time, even though the total variance may not, depending on the testing instrument used (McCaffrey, Lockwood, Koretz, Louis, & Hamilton, 2004). However, the EVAAS model, unlike the cross-classified model, does not place restrictions on the overall grade-specific means or the covariance structure of repeated measurements on the same student (McCaffrey et al., 2004; Wright, Sanders, & Rivers, 2006). The unstructured within-student covariance matrix allows each student to serve as his or her own control, making it unnecessary to account for factors affecting a student's level of achievement (Sanders et al., 1997). Yet, both of these models are computationally intensive and require scores be on a single developmental scale (McCaffrey et al., 2003). In general, using the EVAAS layered model to estimate teacher effects has advantages over other modeling approaches. The layered model accounts for both between- and within-student correlations to adjust for past and future student achievement scores. The model also uses all available scores, resulting in "less biased, more stable, more efficient estimates" (Wright & Sanders, 2008, p. 14) than either the gain score or the covariate adjustment model. The use of multiple years of data allows estimates to be adjusted, thereby accounting for external pulses occurring in a given year and rewarding teachers whose students perform better than expected in the future. Overall, Wright and Sanders (2008) argue the EVAAS layered model is a competitive option for estimating teacher effects, because of its flexibility and adherence to value-added philosophy.

In order to investigate how value-added models can be effectively used when student achievement data are from a variety of measures at different times, the Z-score and binning by quantile methods are used with MPS student achievement data. With the EVAAS model, changes in raw scores are not meaningful when test scores in successive years are not on a single developmental scale. To compensate for this problem, standard Z-scores can be used. In a given academic year, a student's Z-score indicates how many standard deviations the original score is away from the average score for a grade. Changes in Z-scores reflect changes in relative position across years for a group of students, but not necessarily changes in academic achievement, when measures are on different developmental scales (McCaffrey et al., 2003). The standardized scores allow for within-group comparisons across academic years.

In the EVAAS model, the variance of a student outcome inherently increases with the number of grades for which data have been collected on a student through an academic year. When the standard Z-score is used as the response variable, its variance is restricted to one. Sanders et al.'s (1997) proposed method of equally weighting each previous and current teacher's contribution to a student's score in a given academic year was adapted with new methodology to take into account the constraint imposed by the fact Z-scores, by definition, must have a variance of one.

Z-scores can be vulnerable to outliers and wide swings in standard deviations; the small population of Nebraska's rural districts exacerbates this problem. Non-parametric methods, in theory, are less vulnerable and more robust. Non-parametric alternatives to Z-scores based on binning, code scores into ranks based on their quantile rank. The basic idea of binning is to rank student achievement test scores for each year and then divide them into quantiles. A student making "expected" progress would tend to stay at approximately the same rank, relative to other students in the study population, from year to year. Steady movement up in rank over a period of years would indicate the student is making "above expected" progress. Uncategorized ranks tend to be too noisy to be useful for evaluating MSPs. However, classifying ranks into quantiles, e.g. deciles or quintiles, shows promise. Selection of the number of quantiles to some extent depends on the size of the student population involved in the study. Once the data are divided into quantiles, categorical hierarchical models incorporating the same layering structure used in other categorical models, can be used. The BIG advantage of binning, in addition to not being at the mercy of short term variation in standard deviation for small populations, is that it does not involve constraints on the variance and therefore does not require possibly arbitrary scaling of the coefficients of the Z matrix.

While we have been using the Z-score approach for some time, and have initial results from this method, we are in the early stages of exploring and refining the binning approach. Early results indicate the binning method is better at identifying impacts of our professional development program on student achievement than prior methods.

Although both Z-score and binning approaches have limitations, each is an appropriate alternative to using raw data when analyzing less-than-ideal student achievement data across a mixture of norm- and criterion-referenced tests over time. Methodology developed addresses issues arising when using a layered, longitudinal linear mixed model to analyze gains in standardized scores, including weighting considerations for variance components. Additional studies should consider other weighting alternatives and investigate the impact of such variance component weighting schemes on the estimation of teacher effects. Because curricula and test content vary across grades, as do mobility rates, future research should also explore whether notable changes in a student's Z-scores from year to year are associated with changes in mobility rates, curricula and/or test content. It is also important to carefully consider what data are needed and how much baseline data should be obtained when estimating the impact of a professional development program. Ideally, these methods can be extended to other value-added modeling approaches, as well as other professional development programs, and could eventually be used to establish potential relationships between changes in a teacher's mathematical knowledge for teaching mathematics and changes in student achievement.

By developing viable models to estimate teacher effects from complex, heterogeneous rural environments and creating models that connect measures of student achievement and attitude to measures of teaching quality, teacher attitude, and teacher networks, we hope to address what most MSP programs are attempting to do. Many MSP programs offer professional development to teachers, in the hopes they will influence teaching quality, teacher attitude, and teacher networks, which in turn will increase student achievement and improve student attitudes. Our work to develop useful statistical models will serve the entire MSP community and those who work to effectively evaluate the work of MSP programs, as well as potentially impact other federally-funded and non federally-funded education programs that are trying to achieve similar outcomes.

References

Darling-Hammond, L. (2000). Teacher quality and student achievement: A review of state policy evidence. *Education Policy Analysis Archives*, *8*(1). Retrieved April 7, 2008, from http://epaa.asu.edu/ojs/article/view/392

Lockwood, J. R., McCaffrey, D. F., Mariano, L. T., & Setodji, C. (2007). Bayesian methods for scalable multivariate value-added assessment. *Journal of Educational and Behavioral Statistics*, *32*(2), 125-150.

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003). *Evaluating value-added models for teacher accountability*. Santa Monica, CA: RAND Corporation.

McCaffrey, D. F., Lockwood, J. R., Koretz, D., Louis, T. A., & Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics*, *29*(1), 67-101.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica, 73*(2), 417-458.

Rowan, B., Correnti, R., & Miller, R. J. (2002). What large-scale, survey research tells us about teacher effects on student achievement: Insights from the Prospects study of elementary schools. *Teachers College Record, 104*(8), 1525-1567.

Sanders, W. L. (2000). Value-added assessment from student achievement data: Opportunities and hurdles. J*ournal of Personnel Evaluation in Education*, *14*(4), 329-339.

Sanders, W. L. (2004). *A summary of conclusions drawn from longitudinal analyses of student achievement data over the past 22 years (1982-2004)*. Paper presented at the Governors Education Symposium, Ashville, NC.

Sanders, W. L., & Rivers, J. C. (1996). *Cumulative and residual effects of teachers on future student academic achievement.* Knoxville, TN: University of Tennessee Value-Added Research and Assessment Center.

Sanders, W. L., Saxton, A. M., & Horn, S. P. (1997). The Tennessee value-added assessment system: A quantitative outcomes-based approach to educational assessment. In J. Millman (Ed.), *Grading teachers, grading schools: Is student achievement a valid educational measure?* (pp. 137-162). Thousand Oaks, CA: Corwin Press.

Wright, S. P., Horn, S. P., & Sanders, W. L. (1997). Teacher and classroom context effects on student achievement: Implications for teacher evaluation. *Journal of Personnel Evaluation in Education*, *11*(1), 57-67.

Wright, S. P., & Sanders, W. L. (2008, April). *Decomposition of estimates in a layered value-added assessment model*. Paper presented at the National Conference on Value-Added Modeling, Madison, WI.

Wright, S. P., Sanders, W. L., & Rivers, J. C. (2006). Measurement of academic growth of individual students toward variable and meaningful academic standards. In R. Lissitz (Ed.), *Longitudinal and value added models of student performance* (pp. 385-406). Maple Grove, MN: JAM Press.