Student Growth Percentiles and Shoe Leather

Damian W. Betebenner, Richard J. Wenning, Derek C. Briggs

Bruce D. Baker recently published a critique of The Colorado Growth Model and its use of Student Growth Percentiles in his School Finance 101 blog ("Take Your SGP and VAMit, Damn it!" available here). In his blog, he both mischaracterizes the SGP methodology and the policy context. Having participated in creating the Colorado Growth Model and leading the policy development associated with it, we thought it would be useful to clarify these misconceptions.

In work over the past decade with over two dozen State Education Agencies (SEAs) to develop models of student growth based upon state assessment results, one lesson that is repeatedly learned is that data, regardless of their quality, can be used well and can be used poorly. Unfortunately Professor Baker conflates the data (i.e. the measure) with the use. A primary purpose in the development of the Colorado Growth Model (Student Growth Percentiles/SGPs) was to distinguish the measure from the use: To separate the description of student progress (the SGP) from the attribution of responsibility for that progress.

There is a continuum of opinion about how large-scale assessment data and derived quantities can be used in accountability systems. On one extreme are those that believe large-scale assessment results are the ONLY “objective” indicator and thus any judgment about educator/education quality should be based on such measures. At the other extreme are those that hold that any use of large-scale assessment data is an abuse. Our experience in discussing these issues in numerous contexts with stakeholders ranging from parents to policy makers, students to superintendents, is that they fall in between these two extremes. We believe that the results of large-scale assessments, particularly when examined in a longitudinal fashion, can yield numerous insights (some quite profound) about the manner in which the education system is functioning.

In work with the Colorado Department of Education and numerous other SEAs we clearly state that all growth models (including the Colorado Growth Model) can be turned into a value-added model (VAM). A VAM is a type of growth model but not all growth models are necessarily VAM models. We propose that a VAM is, in fact, constituted by its use, not by any particular statistical model specification. A simple gain score model, for example, is often used as an example (usually a bad example) of a value-added model. Other examples abound in the literature (see, for example, McCaffrey, Bin & Lockwood, 2008).

After deriving quantities of individual growth it is natural (and responsible) to ask whether there are contexts or curricular programs where students demonstrate higher or lower rates of growth, on average, than others. This is where

investigations of growth start to become investigations of value-added. Believing that "value-added" is a hypothesis to be tested (Ho, 2011) and not a quantity derived from a model, the challenge in Colorado and other states we work with is to develop indicator systems that facilitate the investigation of what programs, districts, schools, teachers, and contexts promote (and fail to promote) the greatest growth amongst students in the state. Furthermore, going beyond traditional VAM approaches focused on attributing responsibility, to use student growth to investigate growth toward career and college readiness and issues of equal educational opportunity through the examination of growth gaps between demographic and other student subgroups of interest.

The causal nature of the questions together with the observational nature of the data makes the use of large-scale assessment data difficult "detective work". Indeed, good detective work requires shoe leather, looking at multiple sources of evidence, particularly as stakes become high, to ensure that conclusions about responsibility are warranted. We believe that the education system as a whole can benefit from such scrupulous detective work, particularly when all stakeholders hold a seat at the table and are collectively engaged in these efforts to develop and maintain an education system geared toward maximizing the academic progress of all students.

To be clear about our own opinions on the subject: The results of large-scale assessments should ***never*** be used as the sole determinant of education/educator quality. No state or district that we work with intends them to be used in such a fashion. That, however, does not mean that these data cannot be part of a larger body of evidence collected to examine education/educator quality. The dichotomy of appropriate/inappropriate does not and should not lead to an all or nothing dichotomy of data use. The challenge is to enable appropriate and beneficial uses while minimizing those that are inappropriate and detrimental.

Despite Professor Baker's criticism of VAM/SGP models for teacher evaluation, he appears to hold out more hope than we do that statistical models can precisely parse the contribution of an individual teacher or school from the myriad of other factors that contribute to students' achievement. Numerous published writings by scholars on the subject over the past decade (see, for example, Raudenbush (2004); Rubin, Stuart, & Zanutto (2004); Braun (2005), Lockwood, McCaffrey, Mariano, & Setodji (2007); Linn (2008); Rothstein, 2009; 2010; Betebenner & Linn (2010); Briggs & Domingue (2011)) have taken issue with this presumption.

Professor Baker emphasizes this with SGPs:

> Again, the whole point here is that it would be a **leap, a massive freakin' unwarranted leap** to assume a causal relationship between SGP and school quality, if not building the SGP into a

> model that more precisely attempts to distill that causal relationship (if any). [Emphasis in original]

We would add that it is a similar "massive … leap" to assume a causal relationship between *any* VAM quantity and a causal effect for a teacher or school, not just SGPs. We concur with Rubin et al (2004) who assert that quantities derived from these models are descriptive, not causal, measures. However, just because measures are descriptive does NOT imply that the quantities cannot and should not be used as *part* of a larger investigation of root causes.

There are a number of excellent papers and books published over the last two decades that lay out the use and abuse of regression techniques in the social sciences, particularly with regard to making unsubstantiated causal claims. David Freedman's "Statistical Models and Shoe Leather" (1991), Richard Berk's "Regression Analysis: A Constructive Critique" (2003) are particularly good. Berk's book, in fact, details the importance of using regression analyses descriptively as part of a larger program to identify root causes. And this aligns with Linn's (2008, p. 21) call for descriptive accountability:

> "Accountability system results can have value without making causal inferences about school quality, solely from the results of student achievement measures and demographic characteristics. Treating the results as descriptive information and for identification of schools that require more intensive investigation of organizational and instructional process characteristics are potentially of considerable value. Rather than using the results of the accountability system as the sole determiner of sanctions for schools, they could be used to flag schools that need more intensive investigation to reach sound conclusions about needed improvements or judgments about quality."

The development of the Student Growth Percentile methodology was guided by Rubin et al's (2004) admonition that VAM quantities are, at best, descriptive measures. Taken seriously, we are tasked with constructing the best and most useful description possible. Believing that the quality of a description is judged primarily by its utility, the goal with the development and use of the SGP methodology is to maximize utility while maintaining the technical sophistication of a growth model that serves both norm- and criterion-referenced purposes (Betebenner, 2009). Given that all data, regardless of its quality, can be abused, the challenge is to produce an indicator system that maximizes the beneficial use cases of data.

We encourage the continued investigation of measures of student growth with the goal of producing indicator systems that address fundamental policy considerations and maximize utility without compromising technical quality.

Comparisons between models (especially those utilizing the full achievement history of student scores) often produce results that are highly correlated (> 0.8), making determinations of which model is "best" difficult if not impossible to resolve using technical criteria alone. For example, comparisons of SGPs with value-added model results have high correlations (Briggs & Betebenner, 2009; Wright, 2010).

Claims of model "bias" that Professor Baker refers to are often difficult to disentangle because, as McCaffrey, Bin, and Lockwood (2008) point out in their comprehensive comparison of VAM measures, there is no gold standard "teacher effect" or "school effect" against which to judge any of these measures. And scenarios where differential performance by demographic subgroup on a growth/value-added measure occur do not necessarily imply "bias" any more than scenarios with differential achievement level performance by demographic subgroup (e.g., percent at or above proficient) does. On the contrary, such growth gaps can be indicative of unequal educational opportunity. The determination of model validity is complex, involving judgments that are both technical and practical. This reality, we believe, reaffirms the wisdom of Box's (1987, p. 424) famous maxim: "All models are wrong, but some are useful".

Returning to the opening point, our work is directed toward the use of large-scale assessment results as an evidence base to promote and help facilitate the difficult detective work associated with investigations of quality and effectiveness in an education system. Ultimately, we contend, the goal is to use what we learn to improve the education system for the benefit of all children. To that end, the validity of an accountability system is determined by the consequences that derive from it.

> Assessment practices and systems of accountability are systemically valid if they generate useful information and constructive responses that support one or more policy goals (Access, Quality, Efficacy, Equity, and Efficiency) within an education system, without causing undue deterioration with respect to other goals. (Braun, 2008)

Large-scale assessment results are an important piece of evidence but are not sufficient to make causal claims about school or teacher quality. Black and white polemics about appropriate/inappropriate use of data often undercut valuable descriptions of the reality of a system in which large percentages of students are not receiving the education they deserve and we desire. Our goal is not to promote scapegoating for these unpalatable realities but to give stakeholders interpretable and actionable data that enable sound decision making, promote learning, and marshal a consensus for change.

References:

Baker, B. D. (2011). Take your SGP and VAMit, Damn it! http://schoolfinance101.wordpress.com/2011/09/02/take-your-sgp-and-vamit-damn-it/

Betebenner, D. W. (2009). Norm- and criterion-referenced student growth. Educational Measurement: Issues and Practice, 28(4):42-51.

Betebenner, D. W. & Linn, R. L. (2010). Growth in student achievement: issues of measurement, longitudinal data analysis, and accountability. Exploratory Seminar: Measurement Challenges Within the Race to the Top Agenda: Center for K-12 Assessment and Performance Management. www.k12center.org/rsc/pdf/**BetebennerandLinn**PolicyBrief.pdf

Berk, R. A. (2003). Regression Analysis: A Constructive Critique. Sage, Thousand Oaks, CA.

Berk, R. A. & Freedman, D. A. (2003). Statistical assumptions as empirical commitments. In T. G. Blomberg and S. Cohen (eds.), *Law, Punishment, and Social Control: Essays in Honor of Sheldon Messinger*, 2nd ed. (2003), Aldine de Gruyter, pp. 235–254. www.stat.berkeley.edu/~census/berk2.pdf

Box, G. E. P. & Draper, N. R. (1987). *Empirical Model-Building and Response Surfaces*, Wiley

Braun, H. I. (2008). Viccissitudes of the validators. Presentation made at the 2008 Reidy Interactive Lecture Series, Portsmouth, NH, September, 2008. www.cde.state.co.us/cdedocs/OPP/HenryBraunLectureReidy2008.ppt

Braun, H. I. (2005). Using student progress to evaluate teachers: A primer on value-added models. Technical report, Educational Testing Service, Princeton, New Jersey. www.ets.org/Media/Research/pdf/PICVAM.pdf

Briggs, D. C. & Betebenner, D. (2009). *Is Growth in Student Achievement Scale Dependent?* Paper presented at the invited symposium ―Measuring and Evaluating Changes in Student Achievement: A Conversation about Technical and Conceptual Issues‖ at the annual meeting of the National Council for Measurement in Education, San Diego, CA, April 14, 2009.

Briggs, D. & Domingue, B. (2011). *Due Diligence and the Evaluation of Teachers: A review of the value-added analysis underlying the effectiveness rankings of Los Angeles Unified School District teachers by the* Los Angeles Times. Boulder, CO: National Education Policy Center. http://nepc.colorado.edu/publication/due-diligence.

Freedman D. (1991) "Statistical Models and Shoe Leather," in P. V. Marsden (ed.) *Sociological Methodology*, Volume 21, Washington, D. C.: The American Sociological Association.

Ho, A. (2011). Supporting Growth Interpretations using Through Course Assessments. Center for K-12 Assessment and Performance Management at ETS. www.k12center.org/rsc/pdf/TCSA_Symposium_Final_Paper_**Ho**.pdf

Linn, R. L. (2008). Educational accountability systems. In The Future of Test Based Educational Accountability, pages 3–24. Taylor & Francis, New York.

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

McCaffrey, D, Han, B., & Lockwood, J. (2008). From Data to Bonuses: A Case Study of the Issues Related to Awarding Teachers Pay on the Basis of Their Student's Progress. National Center on Performance Incentives Working Paper Working Paper 2008-14. www.performanceincentives.org/**data**/files/.../McCaffrey_et_al_2008.pdf

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

Raudenbush, S. (2004). Schooling, statistics, and poverty: Can we measure school improvement? (Technical report). Princeton, NJ: Educational Testing Service. www.ets.org/Media/**Education**_Topics/pdf/angoff9.pdf

Rothstein, J. (2009). Student sorting and bias in value-added estimation: Selection on observables and unobservables. *Education Finance and Policy*, *4*(4), 537–571.

Rothstein, J. (2010). Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement. *Quarterly Journal of Economics*, *125*(1), 175–214.

Rubin, D. B., Stuart, E. A., and Zanutto, E. L. (2004). A potential outcomes view of value-added assessment in education. Journal of Educational and Behavioral Statistics, 29(1):103–116. www.ucce.ucdavis.edu/files/workgroups/6798/**Rubin**EtAl.pdf

Wright, P. S. (2010). An Investigation of Two Nonparametric Regression Models for Value-Added Assessment in Education, White paper. www.**sas**.com/resources/whitepaper/wp_16975.pdf

Dr. Damian W. Betebenner is a Senior Associate with the National Center for the Improvement of Educational Assessment (NCIEA). Since joining the NCIEA in 2007, his work has centered exclusively on the research and development of student growth models for state accountability systems. He is the analytic architect of the student growth percentile (SGP) methodology developed in collaboration with the Colorado Department of Education as the Colorado Growth Model.

Richard Wenning served until June 2011 as the Associate Commissioner of the Colorado Department of Education (CDE) and led CDE's Office of Performance and Policy. His responsibilities included public policy development and the design and implementation of Colorado's educational accountability system, including the Colorado Growth Model.

Professor Derek C. Briggs is chair of the Research and Evaluation Methodology Program at the University of Colorado at Boulder, where he also serves as an associate professor of quantitative methods and policy analysis. In general, his research agenda focuses upon building sound methodological approaches for the valid measurement and evaluation of growth in student achievement. His daily agenda is to challenge conventional wisdom and methodological chicanery as they manifest themselves in educational research, policy and practice.