# School Finance 101

Data and thoughts on public and private school funding in the U.S.

## Take your SGP and VAMit, Damn it!

By Bruce Baker
*Posted on September 2, 2011*

In the face of all of the public criticism over the imprecision of value-added estimates of teacher effectiveness, and debates over whether newspapers or school districts should publish VAM estimates of teacher effectiveness, policymakers in several states have come up with a clever shell game. Their argument?

We don't use VAM… 'cuz we know it has lots of problems, we use Student Growth Percentiles instead. They don't have those problems.

WRONG! WRONG! WRONG! Put really simply, as a tool for inferring which teacher is "better" than another, or which school outperforms another, SGP is worse, not better than VAM. This is largely because SGP is simply not designed for this purpose. And those who are now suggesting that it is are simply wrong. Further, those who actually support using tools like VAM to infer differences in teacher quality or school quality should be most nervous about the newly found popularity of SGP as an evaluation tool.

To a large extent, the confusion over these issues was created by Mike Johnston, a Colorado State Senator who went on a road tour last year pitching the Colorado teacher evaluation bill and explaining that the bill was based on the Colorado Student Growth Percentile Model, not that problematic VAM stuff. Johnston naively pitched to legislators and policymakers throughout the country that SGP is simply not like VAM (True) and that therefore, SGP is not susceptible to all of the concerns that have been raised based on rigorous statistical research on VAM (Patently FALSE!). Since that time, Johnston's rhetoric that SGP gets around the perils of VAM has been widely adopted by state policymakers in states including New Jersey, and these state policymakers understanding of SGP and VAM is hardly any stronger than Johnston's.

This brings me back to my exploding car analogy. I've pointed out previously that if we lived in a society where pretty much everyone still walked everywhere, and then someone came along with this new automotive invention that was really fast and convenient, but had the tendency to explode on every third start, I think I'd walk. I use this analogy to explain why I'm unwilling to jump on the VAM bandwagon, given the very high likelihood of falsely classifying a good teacher as bad and putting their job on the line – a likelihood of misfire that has been validated by research. Well, if some other slick talking salesperson (who I refer to as slick Mikey J.) then showed up at my door with something that looked a lot like that automobile and had simply never been tested for similar failures, leading the salesperson to claim that this one doesn't explode (for lack of evidence either way), I'd still freakin' walk! I'd probably laugh in his face first. Then I'd walk.

Origins of the misinformation aside, let's do a quick walk through about how  and why, when it comes to estimating teacher effectiveness, SGP is NOT immune to the various concerns that plague value-added modeling. In fact, it is potentially far more susceptible to specific concerns such as the non-random assignment of students and the influence of various student, peer and school level factors that may ultimately bias ratings of teacher effectiveness.

**What is a value-added estimate?**

A value added estimate uses assessment data in the context of a statistical model, where the objective is quite specifically to estimate the extent to which a student having a specific teacher or attending a specific school influences that student's difference in score from the beginning of the year to the end of the year – or period of treatment (in school or with teacher). The best of VAMs attempt to account for several prior year test scores (to account for the extent that having a certain teacher alters a child's trajectory), classroom level mix of students, individual student background characteristics, and possibly school characteristics. The goal is to identify most accurately the share of the student's value-added that should be attributed to the teacher as opposed to all that other stuff (a ~~nearly~~ impossible task)

**What is a Student Growth Percentile?**

To oversimplify a bit, a student growth percentile is a measure of the relative change of a student's performance compared to that of all students and based on a given underlying test or set of tests. That is, the individual scores obtained on these underlying tests are used to construct an index of student growth, where the median student, for example, may serve as a baseline for comparison. Some students have achievement growth on the underlying tests that is greater than the median student, while others have growth from one test to the next that is less (not how much the underlying scores changed, but how much the student moved within the mix of other students taking the same assessments, using a method called quantile regression to estimate the rarity that a child falls in her current position in the distribution, given her past position in the distribution). For more precise explanations, see:
http://dirwww.colorado.edu/education/faculty/derekbriggs/Docs/Briggs_Weeks_Is%20Growth%20in%20Student%20Achievement%20Scale%20Dependent.pdf

So, on the one hand, we've got Value-Added Models, or VAMs, which attempt to construct a model of student achievement, and to estimate specific factors that may affect student achievement growth, including teachers, schools, and ideally controlling for prior scores of the same students, characteristics of other students in the same classroom and school characteristics. The richness of these various additional controls plays a significant role in limiting the extent to which one incorrectly assigns either positive or negative effects to teachers. Briggs and Domingue run various alternative scenarios to this effect here:
http://nepc.colorado.edu/publication/due-diligence

On the other hand, we have a seemingly creative alternative for descriptively evaluating how one student's  performance over time compares to the larger group of students taking the same assessments. These growth measures can be aggregated to the classroom or school level to provide descriptive information on how the group of students grew in performance over time, on average, as a subset of a larger group. But, these measures include no attempt at all to attribute that growth or a portion of that growth to individual teachers or schools. That is, sort out the extent to which that growth is a function of the teacher, as opposed to being a function of the mix of peers in the classroom.

What do we know about Value-added Estimates?

- They are susceptible to non-random student sorting, even though they attempt to control for it by including a variety of measures of student level characteristics, classroom level and peer characteristics, and school characteristics. That is, teachers who persistently serve more difficult students, students who are more difficult in unmeasured ways, may be systematically disadvantaged.
- They produce different results with different tests or different scaling of different tests. That is, a teacher's rating based on their students performance on one test is likely to be very different from that same teacher's rating based on her students performance on a different test, even of the same subject.
- The resulting ratings have high rates of error for classifying teacher effectiveness, likely in large part due to error or noise in underlying assessment data and conditions under which students take those tests.
- They are particularly problematic if based on annual assessment data, because these data fail to account for differences in summer learning, which vary widely by student backgrounds (where those students are non-randomly assigned across teachers).

What do we know and don't we know about SGP?

- They rely on the same underlying assessment data as VAMs, but simply re-express performance in terms of changes in relative growth rather than the underlying scores (or rescaled scores).
  - They are therefore susceptible to at least equal error of classification concern
  - Therefore, it is reasonable to assume that using different underlying tests may result in different normative comparisons of one student to another
  - Therefore, they are equally problematic if based on annual assessment data
- They do not even attempt (because it's not their purpose) to address non-random sorting concerns or other student and peer level factors that may affect "growth."
  - Therefore, we don't even know how badly these measures are biased by these omissions? Researchers have not tested this because it is presumed that these measures don't attempt such causal inference.

Unfortunately, while SGPs are becoming quite popular across states including Massachusetts, Colorado and New Jersey, and SGPs are quickly becoming the basis for teacher effectiveness ratings, there doesn't appear to be a whole lot of specific research addressing these potential shortcomings of SGPs. **Actually, there's little or none!** This dearth of information may occur because researchers exploring these issues assume it to be a no brainer that if VAMs suffer classification problems due to random error, then so too would SGPs based on the same data. If

VAMs suffer from omitted variables bias then SGP would be even more problematic, since it includes no other variables. Complete omission is certainly more problematic than partial omission, so why even bother testing it.

In fact, Derek Briggs, in a recent analysis in which he compares the attributes of VAMs and SGPs explains:

We do not refer to school-level SGPs as value-added estimates for two reasons. First, no residual has been computed (though this could be done easily enough by subtracting the 50th percentile), and second, **we wish to avoid the causal inference that high or low SGPs can be explained by high or low school quality** (for details, see Betebenner, 2008).

As Briggs explains and as Betebenner originally proposed, SGP is essentially a descriptive tool for evaluating and comparing student growth, including descriptively evaluating growth in the aggregate. But, it is not by any stretch of the imagination designed to estimate the effect of the school or the teacher on that growth.

Again, Briggs in his conclusion section of his analysis of relative and absolute measures of student growth explains:

However, there is an important philosophical difference between the two modeling approaches in that Betebenner (2008) has focused upon the use of SGPs as a descriptive tool to characterize growth at the student-level, while the LM (layered model) is typically the engine behind the teacher or school effects that get produced for inferential purposes in the EVAAS. (value-added assessment system)
http://dirwww.colorado.edu/education/faculty/derekbriggs/Docs/Briggs_Weeks_Is%20Growth%20in%20Student%20Achievement%20Scale%20Dependent.pdf

To clarify for the non-researcher, non-statisticians, what Briggs means in his reference to "inferential purposes," is that **SGPs, unlike VAMs are not even intended to "infer" that the growth was caused by differences in teacher or school quality.** Briggs goes further to explain that overall, SGPs tend to be higher in schools with higher average achievement, based on Colorado data. Briggs explains:

These result suggest that schools that higher achieving students tend to, on average, show higher normative rates of growth than schools serving lower achieving students. Making the inferential leap that student growth is solely caused by the school and sources of influence therein, the results translate to saying that schools serving higher achieving students tend to, on average, be more effective than schools serving lower achieving students. The correlations between median SGP and current achievement are (tautologically) higher reflecting the fact that students growing faster show higher rates of achievement that is reflected in higher average rates of achievement at the school level.

Again, the whole point here is that it would be a **leap, a massive freakin' unwarrented 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).

It's a fun and interesting paper and one of the few that addresses SGP and VAM together, but intentionally does not explore the questions and concerns I pose herein regarding how the descriptive results of SGP would compare to a complete value added model at the teacher level, where the model was intended for estimating teacher effects. Rather, Briggs compares the SGP findings only to a simple value-added model of school effects with no background covariates,[1] and finds the two to be highly correlated. Even then Briggs finds that the school level VAM is less correlated with initial performance level than is the SGP (where that correlation is discussed above).

So then, where does all of this techno-babble bring us? It brings us to three key points.

1. First, there appears to be no analysis of whether SGP is susceptible to the various problems faced by value-added models largely because credible researchers (those not directly involved in selling SGP to state agencies or districts) consider it to be a non-issue. SGPs weren't ever meant to nor are they designed to actually measure the causal effect of teachers or schools on student achievement growth. They are merely descriptive measures of relative growth and include no attempt to control for the plethora of factors one would need to control for when inferring causal effects.
2. Second, and following from the first, it is certainly likely that if one did conduct these analyses, that one would find that SGPs produce results that are much more severely biased than more comprehensive VAMS and that SGPs are at least equally susceptible to problems of random error and other issues associated with test administration (summer learning, etc.).
3. Third, and most importantly, policymakers are far too easily duped into making really bad decisions with serious consequences when it comes to complex matters of statistics and measurement. While SGPs are, in some ways, substantively different from VAMS, they sure as heck aren't better or more appropriate for determining teacher effectiveness. **That's just wrong!**

And this is only an abbreviated list of the problems that bridge both VAM and SGP and more severely compromise SGP. Others include spillover effects (the fact that one teacher's scores are potentially affected by other teachers on his/her team serving the same students in the same year), and the fact that only a handful of teachers (10 to 20%) could be assigned SGP scores, requiring differential contracts for those teachers and creating a disincentive to teach core content in elementary and middle grades. Bad policy is bad policy. And this conversation shift from VAM to SGP is little more than a smokescreen intended to substitute a potentially worse, but entirely untested method with a method for which serious flaws are now well known.

**Note:** To those venders of SGP (selling this stuff to state agencies and districts) who might claim my above critique to be unfair, I ask you to show me the technical analyses conducted by a qualified fully independent third party that shows that SGPs are not susceptible to non-random assignment problems, that they miraculously negate bias resulting from differences in summer learning even when using annual test data, that they have much lower classification error rates when assigning teacher effectiveness ratings, that teachers receive the same ratings regardless of which underlying tests are used and that one teacher's ratings are not influenced by the other teachers of the same students. Until you can show me a vast body of literature on these issues specifically applied to SGP (or even using SGP as a measure within a VAM), comparable to that already in existence on more complete VAM models, don’t waste my time.

[1] Noting: “while the model above can be easily extended to allow for multivariate test outcomes (typical of applications of the EVAAS by Sanders), background covariates, and a term that links school effects to specific students in the event that students attend more than one school in a given year (c.f., Lockwood et al., 2007, p. 127-128), we have chosen this simpler specification in order to focus attention on the relationship between differences in our choice of the underlying scale and the resulting schools effect estimates.”