On The Mythologies of Student Growth Percentiles and Teacher Evaluation

*A series of blog posts from School Finance 101*

Bruce D. Baker
Rutgers University

Part 1: Take your SGP and VAMit, Damn it!

Posted on September 2, 2011 by schoolfinance101

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).¹

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.²

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

¹ For more precise explanations, see: http://dirwww.colorado.edu/education/faculty/derekbriggs/Docs/Briggs_Weeks_Is%20Growth%20in%20Student%20Achievement%20Scale%20Dependent.pdf
² Briggs and Domingue run various alternative scenarios to this effect here: http://nepc.colorado.edu/publication/due-diligence
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)*

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’ 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).

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

---

3 http://dirwww.colorado.edu/education/faculty/derekbriggs/Docs/Briggs_Weeks_Is%20Growth%20in%20St udent%20Achievement%20Scale%20Dependent.pdf
findings only to a simple value-added model of school effects with no background covariates, 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.

- **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.

- **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.).

- **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 vendors 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

---

4 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.”
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.

**Part 2: More on the SGP debate: A reply**

**Posted on September 13, 2011 by schoolfinance101**

A new post at Ed News Colorado, by the creators of Student Growth Percentiles, responded to my critique of Student Growth Percentiles, which was also posted on Ed News Colorado.

I must say that I agree with almost everything in this response to my post, except for a few points. First, they argue:

*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.*

No, I do not conflate the data and measures with their proposed use. Policy makers are doing that and doing that based on ill advisement from other policymakers who don’t see the important point – the primary purpose – as Betebenner, Briggs and colleagues explain. This is precisely why I use their work in my previous post – because it explains their intent and provides their caveats.

Policymakers, by contrast are pitching the direct use of SGPs in teacher evaluation. Whether they intended this or not, that’s what’s happening. Perhaps this is because they are not explaining as bluntly they do here, what the actual intent/design was.

Further, I should point out that while I have marginally more faith that a VAM could, in theory be used to parse out teacher effect than an SGP, which isn’t even intended to, I do not have any more faith than they do that a VAM actually can accomplish this objective. They interpret my post as follows:

*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.*

I’m not, as they would characterize, a VAM supporter over SGP, and any reader of this blog certainly realizes that. However, it is critically important that state policymakers be informed that SGP is not even intended to be used in this way. I’m very pleased they have chosen to make this the central point of their response!

---

And while SGP information might reasonably be used in another way, if used as a tool for ranking and sorting teacher or school effectiveness, SGP results would likely be more biased even than VAM results… and we may not even know or be able to figure out to what extent.

I agree entirely with their statement (but for the removal of “freakin”):

*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.*

The authors of the response make one more point that I find objectionable (because it’s a cop out!):

*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.*

What the authors accomplish with this point, is permitting policymakers to still assume (pointing to this quote as their basis) that they can actually use this kind of information, for example, for a fixed 90% share of high stakes decision making, regarding school or teacher performance, and certainly that a fixed 40% or 50% weight would be reasonable. Just not 100%. Sure, they didn’t mean that. But it’s an easy stretch for a policymaker.

If the measures aren’t meant to isolate system, school or teacher effectiveness, or if they were meant to but simply can’t, they should NOT be used for any fixed, defined, inflexible share of any high stakes decision making. In fact, even better, more useful measures shouldn’t be used so rigidly.

[Also, as I’ve pointed out in the past, when a rigid indicator is included as a large share (even 40% or more) in a system of otherwise subjective judgments, the rigid indicator might constitute 40% of the weight but drive 100% of the decision.]

So, to summarize, I’m glad we are, for the most part, on the same page. I’m frustrated that I’m the one who had to raise this issue in part because it was pretty clear to me from reading the existing work on SGP’s that many were conflating the measure with its use. I’m still concerned about the use, and especially concerned in the current policy context. I hope in the future that the designers and promoters of SGP will proclaim more loudly and clearly their own caveats – their own cautions – and their own guidelines for appropriate use.

Simply handing off the tool to the end user and then walking away in the face of misuse and abuse would be irresponsible.

**Addendum:** By the way, I do hope the authors will happily testify on behalf of the first teacher who is wrongfully dismissed or “de-tenured” on the basis of 3 bad SGPs in a row. That they will testify that SGPs were never intended to assume a causal relationship to teacher effectiveness, nor can they be reasonably interpreted as such.
Part 3: Inkblots and Opportunity Costs: Pondering the Usefulness of VAM and SGP Ratings

Posted on September 17, 2011 by schoolfinance101

I spent some time the other day, while out running, pondering the usefulness of student growth percentile estimates and value added estimates of teacher effectiveness for the average school or district level practitioner. How would they use them? What would they see in them? How might these performance snapshots inform practice?

Let’s just say I am skeptical that either VAMs (Value Added Models) or SGPs (Student Growth Percentiles) can provide useful insights to anyone who doesn’t have a pretty good understanding of the nuances of these kinds of data/estimates & the underlying properties of the tests. If I was a principal, would I rather have the information than not? Perhaps. But I’m someone who’s primary collecting hobby is, well, collecting data. That doesn’t mean it all has meaning, or more specifically, that it has sufficient meaning to influence my thinking or actions. Some does. Some doesn’t. Keeping some of the data that doesn’t have much meaning actually helps me to delineate. But I digress.

It seems like we are spending a great deal of time and money on these things for questionable return. We are investing substantial resources in simply maximizing the links in our data systems between individual student’s records and their classroom teachers of record, hopefully increasing our coverage to, oh, somewhere between 10% and 20% of teachers (those with intact, single teacher classrooms, serving children who already have a track record of prior tests – e.g. upper elementary classroom teachers).

At the outset of this whole “statistical rating of teachers” endeavor, it was perhaps assumed by some economists that we would just ram these things through as large scale evaluation tools (statewide and in large urban districts) and use them to prune the teacher workforce and that would make the system better. We’d shoot first… ask questions later (if at all). We’d make some wrong decisions, hopefully statistically more “right” than wrong, and we’d develop a massive model and data set for large enough numbers of teachers that the cost per unit (cost per bad teacher correctly fired, counterbalanced by the cost per good teacher wrongly fired) would be relatively low. We’d bring it all to scale, and scale would mean efficiency.

Now, I find this whole version of the story to be too offensive to really dig into here and now. I’ve written previously about “smart selection” versus “dumb selection” regarding personnel decisions in schools. And this would be what I called “dumb selection.”

But, it also hasn’t necessarily played out this way… thankfully… except perhaps for some large city systems like Washington, DC, and a few more rigidly mandated state systems (though we’re mostly in wait-and-see mode there as well). Instead, we are now attempting to be more “thoughtful” about how we use this stuff and asking teachers to ponder their statistical ratings for insights into how they interact with children? How they teach? And we are asking administrators to ponder teachers’ statistical estimates for any meaning they might find.

In my current role, as a researcher of education policy, I love equations like this:
I like to see the long lists of coefficients (estimates of how some measure in the model relates to the dependent variable) spit out in my Stata logs and ponder what they might mean, with full consideration of what I’ve chosen to include or exclude in the model, and whether I’m comfortable that the measures on both sides of the equation are of sufficient quality to really tell me anything… or at least something.

The other evening, I thought back to my teaching days (considered a liability as an education policy researcher), at whether I thought it would have been useful to me to simply have some rating of my aggregate effectiveness – simply relative to other teachers. Nothing specific about the performance of my students on specific content/concepts. Just some abstract number… like the relative rarity that my students scored X at the end of my class given that they scored X-Y at the end of last years class? Or, some generalized “effectiveness” rating category based on whether my coefficient in the model surpassed a specific cut score to call me “exceptional” or merely “adequate?” Something like this.

Would that be useful to me? to the principal? if I was the principal?

Given that I typically taught 2 sections of 7th grade life science and 2 of 8th grade physical science (yeah… cushy private school job), with class sizes of about 18 students each, which rotated through different times of day, I might also find it fun to compare growth of my various classes. Did the disruptive distraction kid really cause my ratings in one life science section to crash (you know who you are!)? Was the same kid able to bring her 8th grade teacher down the next year (hopefully not me again!)? Was the same kid able to bring her 8th grade teacher down the next year (hopefully not me again!)?

I asked myself… would those ratings actually tell me anything about what I should do next year (accepting that the data would come on a yearly cycle)? Should I go watch teachers who got better ratings? Could I? Would they protect their turf? Would that even tell me a damn thing? Besides, knowing what I do now, I also know that large shares of the teachers who got a better rating likely got that rating either because of a) random error/noise in the data or b) some unmeasured attribute of the students they serve (bias). Of course, I didn’t know that then, so what would I think?

My gut instinct is that any of these aggregate indicators of a teacher’s relative effectiveness, generated from complex statistical models, with, or without corrections for other factors, are little more than ink blots to most teachers and administrators. And I’m not convinced they’ll ever be anything more than that. They possess many of the same attributes of randomness or fuzziness of an ink blot. And while the most staunch advocate might wish them to appear as an impressionist
painting, I expect they are still most often seen as ink blots – not even a Jackson Pollock. More random than pattern. And even if/when there is a pattern, the average viewer may never pick it up.

I anxiously (though skeptically) await well crafted qualitative studies exploring stakeholders’ interpretations of these inkblots.

But these aren’t just any ink blots. They are rather expensive ink blots if and when we start trying to use them in more comprehensive and human resource intensive ways through local public schools and districts and if we weigh on them the burden that we MUST use them not merely to inform, but rather to DRIVE our decisions – and must find significant meaning in them to justify doing so. That is, if we really expect teachers and principals to log significant hours trying to derive meaning from them, after consultants, researchers, central office administrators and state department officials have labored over data system design, linking teachers to students, and deciding on the most aesthetically pleasing representation of teacher performance classifications for the individual reporting system. Using these tools as quick screening, blunt instruments is certainly a bad idea. But is this – staring at them for endless hours in search of meaning that may not be there – much better?

It strikes me that there are a lot more useful things we could/should/might be spending our time looking at in order to inform and improve educational practice or evaluate teachers. And that the cumulative expenditure on these ink blots, including the cost of time spent musing over them, might be better applied elsewhere.

**Part 4: Piloting the Airplane with Musical Instruments & using SGPs to Evaluate Teachers**

*Posted on September 22, 2011 by schoolfinance101*

I’ve posted a few blogs recently on the topic of Student Growth Percentile Scores, or SGPs and how many state policymakers have moved to adopt these measures and integrate them into new evaluation systems for teachers. In my first post, I argued that SGPs are simply not designed to make inferences about teacher effectiveness.

The designers of SGP replied to my first post, suggesting that I was conflating the measures with their use by suggesting that these measures can’t and shouldn’t be used to infer teacher effectiveness. And in their response (more below), they explained in greater detail, what was essentially my main point – that SGPs are not designed or intended to infer teacher effectiveness from student achievement growth. They also argued that the policy makers they have advised on adopting SGPs understood that.

Well, let’s review what’s going on in New Jersey. In New Jersey, a handful of districts have signed on to the department of education’s Pilot teacher evaluation program, explained here: http://www.state.nj.us/education/EE4NJ/faq/

Specifically, here’s how NJDOE responds to the question over how standardized testing data, and SGPs based on those data would be used within the pilot evaluations:
Q: How much weight do standardized test scores get in the evaluations?

A: Standardized test scores are not available for every subject or grade. For those that exist (Math and English Language Arts teachers of grades 4-8), Student Growth Percentages (SGPs), which require pre- and post-assessments, will be used. The SGPs should account for 35%-45% of evaluations. The NJDOE will work with pilot districts to determine how student achievement will be measured in non-tested subjects and grades.

Now, here is a quote from Betebenner and colleagues’ response to my criticism of policymakers proposed uses of SGPs in teacher evaluation.

From Damian Betebenner & colleagues

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.

But, you see, using these data to “evaluate teachers” necessarily infers “attribution of responsibility for that progress.” Attribution of responsibility to the teacher! If one cannot use these measures to attribute responsibility to the teacher, then how can one possibly use these measures to “evaluate” the teacher? One can’t. You can’t. No-one can. No-one should!

Perhaps in an effort to preserve proprietary interests, Betebenner and colleagues in their reply to my original criticism also note:

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.

But this statement stands in direct conflict with the first above. If the tool is insufficient for – simply not even designed to – ATTRIBUTE RESPONSIBILITY FOR PROGRESS to either teachers or schools, then it simply can’t and SHOULDN’T BE USED THAT WAY! Be it for 10% or 90%.

The reality is that even though Betebenner and colleagues explain that they believe that the policymakers with whom they have consulted “get it” and would never consider misusing the measures in the ways I explained on my original post, that is precisely what is going on.

Also, I noted previously that this paragraph from their response is a complete cop out. I explained:

What the authors accomplish with this point, is permitting policymakers to still assume (pointing to this quote as their basis) that they can actually use this kind of information, for example, for a fixed 90% share of high stakes decision making, regarding school or
teacher performance, and certainly that a fixed 40% or 50% weight would be reasonable. Just not 100%. Sure, they didn’t mean that. But it’s an easy stretch for a policymaker.

If the measures aren’t meant to isolate system, school or teacher effectiveness, or if they were meant to but simply can’t, they should NOT be used for any fixed, defined, inflexible share of any high stakes decision making. In fact, even better, more useful measures shouldn’t be used so rigidly.

[Also, as I’ve pointed out in the past, when a rigid indicator is included as a large share (even 40% or more) in a system of otherwise subjective judgments, the rigid indicator might constitute 40% of the weight but drive 100% of the decision.]

Look. It’s pretty simple. If you want to pilot an airplane effectively, the plane needs to have the right instruments – flight instruments. If you’re coming in for a landing in dense fog in mountainous terrain, you look down to where your flight instruments should be, http://www.b737.org.uk/images/fltinsts_panel_nonefis.jpg, and there sits an alto saxophone instead (albeit a fine, Selmer Mark VI w/serial # in the 180s), you’re screwed. You might have a few minutes left to blow through the changes to Foggy Day, but your chances of successfully piloting the plane to a safe landing are severely diminished.

Okay, this analogy is a bit of a stretch. But it is not a stretch to acknowledge that SGPs were simply not designed to attribute responsibility for student progress to teachers. Meanwhile, VAM models try, but are unable to effectively, accurately or precisely attribute student progress to teachers. So, we have a choice of piloting the plane with either a) the wrong instruments (SGP) or b) instruments that don’t work very well (have high error rates & comparable problems of inference). When faced with choices this bad, it may be wise to take another course entirely. Don’t pilot the damn plane! It would be a shame to crash it with such a beautiful saxophone on board!