On Student Growth Percentiles, imposing standards we don't have: A response to Sireci, Wells, and Keller

Andrew Ho
Harvard Graduate School of Education

June 25, 2016

Three days ago, Sireci, Wells, and Keller (2016; hereafter, the authors) released a research brief entitled, "Why we should abandon Student Growth Percentiles" (SGPs). They summarize recent research and state in bold font, "only one conclusion is justifiable from the research conducted on SGPs—they should be abandoned and not used in education."

There is no technical basis for this conclusion. It is fitting of a piece opposite the editorial page, not a research report. I also disagree with the authors' interpretation of the research. They exaggerate the standards of our field (AERA/APA/NCME, 2014) and impose standards that the field has not specified. Nonetheless, the authors raise issues that states and districts would be wise to address through ongoing research and improved reporting. I make three general recommendations consistent with the evidence the authors cite in their brief.

1. States and districts should improve the precision of teacher- and school-level accountability indices that use SGPs. Simple steps include assigning the SGP component a low weight, averaging across years, and using mean SGPs instead of median SGPs (medians more than double imprecision/unreliability, Castellano & Ho, 2015).

2. States and districts should consider whether and if so how to adjust for statistical bias imparted by measurement error. For example, McCaffrey, Castellano, and Lockwood (2015) consider creating percentile ranks of error-corrected SGPs.

3. States and districts should take Clauser, Keller, and McDermott's (2016) findings (that some users are misinterpreting SGPs) at face value and continue to invest in outreach to SGP users, to demystify the metric and facilitate accurate uses and interpretations.

All of these recommendations are consistent with improving the use of a new metric, not abandoning it. Longitudinal data should be used to inform and improve practice. Features of longitudinal metrics are not necessarily flaws. On the next page, I briefly review more specific points the authors raise.

The authors make 6 conflated points in support of their thesis. From this amalgam, I distill two key claims:

1) The reliabilities of individual and aggregate SGPs are insufficient for their use.
2) Some users are misinterpreting and misusing SGPs.

The first claim is neglects consensus in the field that the sufficiency of SGP reliability (or any score reliability) depends upon the intended interpretations and uses of SGPs. The authors do not explain why the reliability is insufficient for the intended use, and the field has no general benchmarks. The authors also mischaracterize, "a degree of error," as, "a coin flip," and, "it could be wrong," with, "it is wrong" (see also Rogosa, 2004). Although confidence intervals for individual SGPs are wide, there is a signal through the noise. As for their second claim, it is surely correct. It is not controversial to say that any public metric will be misinterpreted and misused. However, I would rather invest in increasing assessment literacy to address this rather than abandon ongoing efforts to make test scores more useful.

The authors make other claims that I consider to be unsupported or red herrings. They appeal to the 2014 standards as well as the 2014 ASA and 2015 AERA statements on value-added models, but they do not explain why existing state technical documentation (e.g., in Georgia, Massachusetts, and Colorado) for individual scores and SGPs do not count as validity evidence. Finally, their attack on SGPs as a norm-referenced metric is mystifying. As Betebenner (2009) has shown from the beginning, SGPs have always been able to reference both norms and criteria.

I conclude that the authors of this brief have raised a few important SGP-related issues, blended them with unsupported claims and red herrings, and prepended an unwarranted thesis statement.

June 27 Addendum:

In response to an email from the authors, I replied with the following technical elaboration:

The authors and I agree about the extent of the error, I think. Take McCaffrey, Castellano, and Lockwood's result from their 2015 paper (Figure 1). It shows the where we might expect the true SGPs to be given an observed SGP of 10, 25, and 50. As the authors suggest, the credible interval for an SGP of 50 spans from 20 to 80.

But this does not mean that all SGPs from 20 to 80 are indistinguishable. A standard error does not imply that all scores within a standard error are "statistically indistinguishable." We teach this in Statistics/Measurement 101.

So let's ask a different question. Is an 80 really greater than a 20? More formally, what are the chances that a student with an observed SGP of 80 actually has a higher true SGP than a student with an observed SGP of 20? We can derive that from densities like those in McCaffrey, Castellano, and Lockwood paper, and the answer is 92%. That's pretty good. It's not great (8 out of 100 times, an 80 isn't greater than a 20!), but it's pretty good. It is not "a coin flip." SGPs outside the 20/80 range have even more precision.

Again, SGPs have considerable imprecision, but they are not noise, and I believe with sound reporting and outreach, they can support wise interpretations and uses.

To conclude, again, while I think the authors have raised some good points, and while I also respect the opinion they have crafted in their brief as an opinion, it is not a conclusion that has a strong technical warrant, and I disagree with it.