WHY WE SHOULD ABANDON STUDENT GROWTH PERCENTILES[1]

Stephen G. Sireci, Ph.D.
Craig S. Wells, Ph.D.
Lisa A. Keller, Ed.D.

Center for Educational Assessment
University of Massachusetts Amherst

Research Brief 16-1

[1] *Center for Educational Assessment Research Brief 16-1.* Amherst, MA: Center for Educational Assessment, University of Massachusetts.

Center for Educational Assessment
College of Education, University of Massachusetts Amherst
Amherst, MA 01003-9329, USA

WHY WE SHOULD ABANDON STUDENT GROWTH PERCENTILES

Educational tests provide useful information for parent, teachers, and policymakers. However, a relatively recent index, called *Student Growth Percentiles* (or SGPs for short) have found their way into score reports sent home to parents and are being used in 22 states for teacher accountability, school improvement plans, and other purposes. The use of this index exploded rapidly before its properties were understood. Only now is there sufficient research on their reliability—and the results are not good. It appears SGPs contain considerable error that prohibits its usefulness as a descriptive index of student progress or as a measure of educator accountability.

In this research brief, we review some recent research in this area, as well as the issues and problems surrounding SGPs. **Only one conclusion is justifiable from the research conducted on SGPs—they should be abandoned and not used in education.**

Based on our review of the research, we have identified 6 reasons why we should abandon SGPs. These reasons are:

1) SGPs are not what people think they are.
2) SGPs are unreliable.
3) Educators do not understand how to use SGPs.
4) There is no validity evidence to support the use of SGPs.
5) Current use of SGPs violates the *Standards for Educational and Psychological Testing,* and statements on value-added modeling issued by the American Educational Research Association and the American Statistical Association.
6) SGPs encourage comparing students to each other, rather than to the knowledge and skill areas they are being taught.

Before discussing each of these six reasons, we first provide a brief history of SGPs.

SGPs: A Brief History

SGPs were introduced as a descriptive index by Betebenner (2009)[2]. As the name suggests, it uses the concept of a *percentile* to describe the percentage of students at or above a certain level. Thus, like percentiles, SGPs range from 1 to 99. SGPs were proposed to solve two problems. First, many statewide tests in reading and math are not on the same scale from grade to grade. Therefore, talking about "how much" a student learned from say 4th grade to 5th grade, is not easily quantifiable. As we describe in the next section, SGPs avoid the "same scale" problem by depicting changes in how students "rank" from year to year. The second problem attempted to be addressed by SGPs is how to give students "credit" for learning if they did not increase according to the achievement levels created in the state. For example, if a

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

child received an achievement classification of "basic" in 4th grade, and also received a classification of "basic" in 5th grade, it is hard to quantify how much "growth" took place, if any, over the course of that year. Again, by ranking students relative to each other, SGPs attempt to link changes in the rankings to change in learning across years.

The logic in computing SGPs appears sensible at first blush, which is why SGPs were approved in Race-to-the-Top applications as a measure of growth. It was this approval that spearheaded the wide acceptance of SGPs. Unfortunately, it was approved as a measure of growth before policymakers really understood what SGPs represented, and before research on their statistical properties was conducted.

In the next sections, we provide further details for the reasons we should abandon SGPs. In discussing the first reason, "SGPs are not what people think they are," we illustrate how SGPs have been described in public documents, and then we describe what they are from a statistical perspective. This discussion provides a foundation for understanding the other problems inherent in SGPs.

Why We Should Abandon SGPs

Reason #1: SGPs are not what people think they are.

SGPs do not represent growth in terms of how much students have learned in a given subject area. They also are not percentiles as most people think of them. The misconceptions of SGPs can be seen by contrasting how they are described by those who promote them with what they actually are from a mathematical perspective. For example, Shang, VanIwaarden, and Betebenner[3] described SGPs as,

> A SGP represents the percentile rank of a student's current score relative to those students at the same grade level who share the same prior score(s) (p. 2).

Similarly, the Massachusetts Department of Elementary and Secondary Education describes SGPs as,

> An [SGP] is a measure of student progress that compares changes in a student's MCAS scores to changes in MCAS scores of other students with similar achievement profiles. The model establishes cohorts of students with "similar performance profiles" by identifying all students with the same (or very similar) MCAS scores in prior years; all MCAS data for a student since 2006 are used (where available) to establish academic peers" (p. 2)[4]

[3] Shang, Y., VanIwaarden, A., & Betebenner, D. W. (2015). Covariate measurement error correction for student growth percentiles using the SIMEX method. *Educational Measurement: Issues and Practice, 34*.

[4] Massachusetts Department of Elementary and Secondary Education (2009). *MCAS student growth percentiles: State report*. Malden, MA: Author. Downloaded June 13, 2016 from http://www.doe.mass.edu/mcas/growth/StateReport.doc.

From these descriptions, it sounds like the process of computing SGPs involves identifying students who had the same test scores in previous years, and then computing percentile ranks for this "cohort" of students in the current year. In fact, no cohorts are created in computing SGPs. Instead, a complex statistical model, called *quantile regression*, is used.

A complete description of quantile regression is beyond the scope of this brief, and so interested readers are referred to other sources[5]. However, for those familiar with linear regression, which uses a single line to predict future performance, quantile regression can be thought of as an extension, where 99 lines are computed to compute "conditional" or "predicted" percentiles for students. Thus, in a sense, quantile regression is 99 times more complex than linear regression. But the important point is not that the model is complex; the point is quantile regression does *not* create cohorts of students. These conditional, or predicted, percentiles are calculated using the entire population of students. Therefore, the descriptions of SGPs provided by state departments of education and others are not entirely accurate.

In conducting research on the use of SGPs and how they are calculated, Clauser, Keller, and McDermott (2016)[6] commented, "SGPs are not percentiles as they are commonly understood, but instead likelihood estimates of a particular score pattern [That is, they are] not direct comparisons to a student's place within a peer group" (p. 12).

Proponents of SGPs sometimes describe them as similar to the height and weight growth charts used by pediatricians. However, given how SGPs are calculated, these descriptions are particularly misleading. Physical growth charts for height and weight do not use quantile regression or any type of regression. They are simply percentiles computed from physical measurements of children at different ages. Unlike pediatric growth charts, the norm group for SGPs changes for each student each year. Several states are using SGPs to classify teachers into effectiveness categories. Can you imagine using pediatric growth charts to classify parents as "ineffective" or "effective" with respect to helping their children grow? That analogy alone should be sufficient for ending the use of SGPs in teacher evaluation.

To summarize our first reason for abandoning SGPs, we agree with Clauser et al. (2016) who commented,

> the term "growth" in this context is misleading. To most, the term growth implies that there has been a change in performance, typically positive, relative to some construct of interest. For example, suppose we consider the growth of a student in math

[5] See for example Betebenner, D.W. (2009). Norm- and criterion-referenced student growth. *Educational Measurement: Issues and Practice, 28*(4), 42-51; or Castellano, K. E., & Ho, A. D. (2013). Contrasting OLS and quantile regression approaches to student "growth" percentiles. *Journal of Educational and Behavioral Statistics, 38(2), 190–215*.

[6] Clauser, A.L., Keller, L.A., McDermott, K.A. (2016). Principals' uses and interpretations of student growth percentile data. *Journal of School Leadership, 26*(1), 6-33.

> achievement. If the student received a high SGP, it would be assumed that the student gained in his/her math achievement and a smaller number would mean that the student showed small improvements in his/her math achievement. This interpretation of growth is reasonable. However, this is not the proper interpretation of the SGP. The SGP is a norm-referenced rather than a criterion-referenced measure; this means that the value refers to a student's standing within a group rather than standing relative to the information being tested. As a result, the term "growth" as applied in the SGP is not the traditional notion of growth and the potential for misinterpretation is great. (pp. 12-13)

Reason #2: SGPs are not reliable.

There is a growing body of research that illustrates SGPs contain so much error that students would receive very different SGPs if they retook the same tests in the same years. The concept of "margin of error" is important in evaluating the reliability of a statistic. Statistics that are reliable give the same value over repeated measurements. For example, the bathroom scale is reliable because it gives the roughly same reading of our weight when we repeatedly weigh ourselves. Students' scores from standardized tests, such as the ACT, SAT, and statewide assessments also demonstrate good reliability, typically varying only a few points. Estimates of the reliability of SGPs suggest they contain too much error to be useful.

To estimate the amount of error in SGPs, several researchers have used *simulation methods* where the known statistics for statewide assessments are used to simulate actual growth for students across years. These simulations are repeated hundreds or thousands of times, and then the variation in the SGPs assigned to the same "students" is calculated.

The results of this research indicates SGPs are inherently unreliable. For example, if a student is reported to have an SGP of 50, the margin of error is about 30 points on either side, indicating that the "true" SGP for the student could be anywhere from 20 to 80, which is almost the entire SGP scale. This finding has been replicated by researchers from several different institutions[7].

Research has also been conducted on the margin of error associated with "aggregated" SGPs. That is, SGPs averaged across students within a classroom to infer something about the effectiveness of a teacher. A recent study requested by the Nevada State Department of

---

[7] McCaffery, D.F., Castellano, K. E., & Lockwood, J. R. (2015). The impact of measurement error of individual and aggregate SGP. *Educational Measurement: Issues and Practice, 34,* 15-21.
Lash, A., Makkonen, R., Tran, L., & Huang, M. (2016). *Analysis of the stability of teacher-level growth scores from the student growth percentile model* (REL 2016–104). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory West. Retrieved from: http://ies.ed.gov/ncee/edlabs.
Dimmick, J., Lawrence, B., Mooney, F., Quinn, E, Soraci, A., & Yarmalowicz, M. (2016, April). *Student Growth Percentiles as Conditional Probabilities,* Poster presented at the 30th New England Statistics Symposium, Yale School of Public Health, New Haven, CT.
Wells, C.S., Sireci, S.G., & Bahry, L. (2014). Estimating the amount of error in student growth percentiles. *Center for Educational Assessment Research Report No. 869*: Amherst, MA: Center for Educational Assessment.

Education and published as an Institute of Education Sciences found that about half of the variation in teachers' average SGPs was due to error[8]. Using generalizability theory to estimate margins of error at the teacher level, the authors concluded,

> the 95 percent confidence interval for a teacher's true score would span 48 points for math, a margin of error that covers nearly half the [99-point SGP] scale, and 44 points for reading. For example, one would be 95 percent confident that the true math score of a teacher who received a score of 50 falls between 26 and 74. (Lash et al., 2016, p. 4)

This conclusion is shocking, given that some states classify teachers with average SGPs below 40 as "ineffective" and those with SGPs above 60 as effective. As the results of the Lash et al. (2016) study illustrate, when a margin of error is placed on teachers' SGP ratings, one might as well flip a coin to decide if they are "ineffective" or "effective." In a separate study, similar results were found[9].

In summary, five separate studies[10] that investigated the reliability of SGPs came to the same conclusion—they contain way too much error to be valid measures of student progress or teacher effectiveness. In fact, the McCaffery et al. (2015) study concluded SGPs computed for teachers were systematically biased such that the most effective teachers were likely to have SGP scores lower than they should and the least effective teachers were likely to have SGP scores higher than they should—the exact opposite of the intent of teacher evaluation.

Reason #3: Educators do not understand how to use SGPs.

Clauser et al. (2016)[11]surveyed over 300 principals in Massachusetts to discover how they used SGPs and to test their interpretations of SGP results. They found over 80% of the principals used SGPs for evaluating the school, over 70% used SGPs to identify students in need of remediation, and almost 60% used SGPs to identify students who achieved exceptional gains. These results suggest SGPs are being used for important purposes, even though they are full of error. The study also found that 70% of the principals misinterpreted what an average SGP referred to, and 70% incorrectly identified students for remediation based on low SGPs, when they actually performed very well on the most recent year's test. Extrapolating from this Massachusetts study, **it is likely SGPs are leading to incorrect decisions and actions in schools across the nation.**

---

[8] Lash, A., Makkonen, R., Tran, L., & Huang, M. (2016). *Analysis of the stability of teacher-level growth scores from the student growth percentile model* (REL 2016–104). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory West. Retrieved from: http://ies.ed.gov/ncee/edlabs.

[9] Marland, J., Wells, C.S., Sireci, S.G., & Castellano, C. (2015). *Investigating the amount of systematic and random error in classroom-level SGPs*. Paper presented at the annual conference of the Northeastern Educational Research Association.

[10] Dimmick, Lawrence, Mooney, Quinn, Soraci, & Yarmalowicz, (2016); Lash, Makkonen, Tran, & Huang, M. (2016); McCaffery, Castellano, & Lockwood, (2015); Marland, Wells, Sireci, & Castellano, (2015); Wells, Sireci, & Bahry, (2014).

[11] Clauser, A.L., Keller, L.A., McDermott, K.A. (2016). Principals' uses and interpretations of student growth percentile data. *Journal of School Leadership, 26*(1), 6-33.

Reason #4: There is no validity evidence to support the use of SGPs.

In our review of the literature we did not find any empirical studies (i.e., studies that involved analysis of data) that provided positive results to defend the use of SGPs. It appears SGPs are being used across the country without any data to support their use.

Reason #5: Current use of SGPs violates the *Standards for Educational and Psychological Testing,* and statements on value-added modeling issued by the American Educational Research Association and the American Statistical Association

For over 60 years, the American Educational Research Association (AERA), the American Psychological Association, and the National Council on Measurement in Education have worked together to produce *Standards for Educational and Psychological Testing*, which provide guidance for those who develop, use, and evaluate tests[12]. These *Standards* assert that accountability indices based on aggregates of students' test scores "should be subjected to the same validity, reliability, and fairness investigations that are expected for the test scores that underlie the index" (p. 210). They also state,

> Users of information from accountability systems might assume that the accountability indices provide valid indicators of the intended outcomes of education…, that the differences among indices can be attributed to differences in the effectiveness of the teacher or school, and that these differences are reasonably stable over time and across students and items. These assumptions must be supported by evidence. (p. 206)

However, as mentioned earlier, there is no empirical evidence to support SGPs for score reporting or accountability purposes.

In addition to the AERA et al. *Standards*, the American Statistical Association also cautioned against the use of SGPs for teacher evaluation[13]. Including SGPs as estimates from value-added models (VAMs), they cautioned,

> Estimates from VAMs should always be accompanied by measures of precision and a discussion of the assumptions and possible limitations of the model. These limitations are particularly relevant if VAMs are used for high-stakes purposes. (p. 1)

The AERA also issued a statement on the use of VAMs and SGPs for evaluating teachers[14] and stated the use of SGPs for teacher evaluation should include estimates of error, reliability and validity

---

[12] American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). *Standards for Educational and Psychological Testing*, Washington, DC: American Educational Research Association.

[13] American Statistical Association (2014). *ASA statement on using value-added models for educational assessment.* Downloaded June 16, 2016 from http://www.amstat.org/policy/pdfs/asa_vam_statement.pdf

[14] American Educational Research Association (2015). AERA statement on use of value-added models (VAM) for the evaluation of educators and educator preparation programs. *Educational Researcher*. Available for download at http://www.aera.net/Newsroom/NewsReleasesandStatements/AERAIssuesStatementontheUseofValue-AddedModelsinEvaluationofEducatorsandEducatorPreparationPrograms/tabid/16120/Default.aspx

evidence to support their use, and ongoing monitoring of reliability and validity evidence. They concluded,

> AERA recommends that VAM (which include…student growth percentile models) not be used without sufficient evidence… that support all claims, interpretative arguments, and uses (e.g., rankings, classification decisions). (p. 4)

Reason #6: SGPs encourage comparing students to each other, rather than to the knowledge and skill areas they are being taught

The science of educational testing has made great progress over the last 30 years, including moving away from "norm-referenced" tests that compare students to one another to "criterion-referenced" tests that describe how well students have mastered the subject matter taught at a particular grade level. Percentiles are used in norm-referenced testing to provide statements like "Geraldine performed as well or better than 56% of the 3rd-graders in the state on this test." While such information may be useful, it does not say how well Geraldine did with respect to the knowledge and skills she was supposed to acquire in 3rd grade. Plus, if all or most of the state did poorly on the test, it is hard to judge how good a percentile of 56 actually is.

In contrast, criterion-referenced tests, such as all statewide tests created since the No Child Left Behind era, set achievement mastery standards such as "basic," "proficient," or "advanced" to describe how well students are doing with respect to proficiency standards established in each grade by a state department of education or other agency. Such information is more informative to parents, and to those planning instruction for students.

The use of SGPs on these statewide assessments encourages comparing students to one another, when the tests are actually designed for comparing students to performance standards in a specific subject area (e.g., 3rd-grade mathematics). As one principal who participated in the Clauser et al. (2016) study commented "I have huge concerns about the statistical validity of the student growth percentiles. After claiming for years that we were getting away from norm referenced testing, this seems a huge step backwards."

Summary

In this research brief, we explained the genesis of SGPs, why they are used, and why they should not be used. Our review of the research discovered a substantial body of research, from several independent sources, that leads to the same conclusion—**SGPs should not be used for reporting "growth" at the student level, or for teacher evaluation purposes**. We found no empirical research to support their use.

SGPs are a new measure proposed for describing student progress and evaluating teachers. Only now are there sufficient data to evaluate their statistical properties. **These data unanimously point to one conclusion—SGPs are too problematic to be used for educational purposes.**

Given these results, we recommend the use of SGPs be abandoned. If states and other organizations continue to use SGPs in light of the results found in the literature, they need to clearly articulate the basis on which they defend such use.

Author contacts:

Stephen G. Sireci, Ph.D. is Professor of Educational Policy, Research, and Administration, and Director, Center for Educational Assessment, University of Massachusetts Amherst. He is a Fellow of AERA and APA. His specialties include **computerized-adaptive testing, assessing students with disabilities and English learners, test development, and test evaluation.** He was a major author on the Congressionally-mandated evaluation of the **National Assessment of Educational Progress (NAEP)** in 2009. sireci@acad.umass.edu; (413)545-0564; Twitter: @stevesireci.

Craig S. Wells, Ph.D. is Associate Professor of Educational Policy, Research, and Administration, and Associate Director, Center for Educational Assessment, University of Massachusetts Amherst. His research areas include non-parametric item response models, detection of differential item functioning, assessment of model fit, and evaluating educational statistics. Csw@educ.umass.edu; (413)577-1726.

Lisa A. Keller, Ed.D. is Associate Professor of Educational Policy, Research, and Administration, and Associate Director, Center for Educational Assessment, University of Massachusetts Amherst. Her specialties include equating tests, evaluating assessments for fairness, and generalizability theory. lkeller@umass.edu; (413)545-1528.