PRECISION, INTERPRETABILITY & UTILITY OF SGPs:

A response to *Why we should abandon student growth percentiles* by Sireci, Wells, and Keller

July 7, 2016

Damian W. Betebenner, Ph.D.

Charles DePascale, Ph.D.

Scott Marion, Ph.D.

Chris Domaleski, Ph.D.

Joseph Martineau, Ph.D

Center for Assessment

National Center for the Improvement
of Educational Assessment
Dover, New Hampshire

CENTER FOR ASSESSMENT

EXECUTIVE SUMMARY

On June 22nd, 2016 the University of Massachusetts, Amherst, Center for Educational Assessment (UMASS CEA) released a report authored by Stephen Sireci, Craig Wells, and Lisa Keller entitled "Why we should abandon student growth percentiles" (Sireci, Wells, & Keller, 2016). Student growth percentiles (SGPs) were developed beginning in 2007 by Damian Betebenner of the National Center for the Improvement of Educational Assessment (Center for Assessment). As the inventor and primary developer of the SGP measure, the Center for Assessment has worked with more than two dozen states to refine and adapt the SGP measure to specific state contexts including diagnostic reporting as well as its use in education accountability systems. The SGP methodology is open source as part of the Center for Assessment's commitment to broadly share expertise and to expose and correct any errors that occur quickly and openly.

The report by Sireci et al. puts forth six reasons why student growth percentiles should be abandoned. A detailed reading of the evidence supporting the six reasons shows numerous misunderstandings, distortions and basic errors. The Center for Assessment encourages rigorous critique of all of our work including the SGP methodology.[1] Unfortunately, the report by Sireci et al. fails in that regard.

OVERVIEW

Beginning in 2007, in joint work between the Center for Assessment and the Colorado Department of Education, student growth percentiles (SGPs) were developed as a growth measure for use with state large scale assessments (Betebenner, 2008, 2009, 2012; Shang, VanIwaarden, & Betebenner, 2015). Since then, the Center for Assessment has worked with more than two dozen states to refine and adapt the SGP measure to specific state contexts including diagnostic student reporting and in education accountability systems. SGPs were developed to maximize utility and interpretability providing stakeholders from parents to policy makers with a readily understandable measure to communicate student learning (i.e., growth).

On June 22nd, 2016 the University of Massachusetts, Amherst Center for Educational Assessment (UMASS CEA) released a report authored by Stephen Sireci, Craig Wells, and Lisa Keller entitled "Why we should abandon student growth percentiles" (Sireci et al., 2016). The report puts forth six reasons why student growth percentiles should be abandoned:

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.

The conclusions drawn from evidence cited in the report supporting the six reasons contain substantive errors leading to erroneous conclusions. Because of the overlapping nature of the six reasons put forth, our critique of Sireci et al. is broken into three sections:

1. Reliability/Precision of SGPs
2. Interpretation and Use of SGPs
3. Alignment with the *Standards for Educational and Psychological Testing*.

Reliability/Precision of SGPs

Sirici et al. discuss reliability/precision of SGPs for two different use cases: individual level SGPs and group level (aggregate) SGPs. Individual level SGPs are used most often for diagnostic reporting via individual student reports (see, for example, Figure 4). Aggregate level SGPs (e.g., median/mean SGPs) are often used for accountability purposes such as school or teacher evaluation purposes. We discuss these use cases separately as they differ both in terms of use as well as technically.

Reliability/Precision of Individual Level SGPs

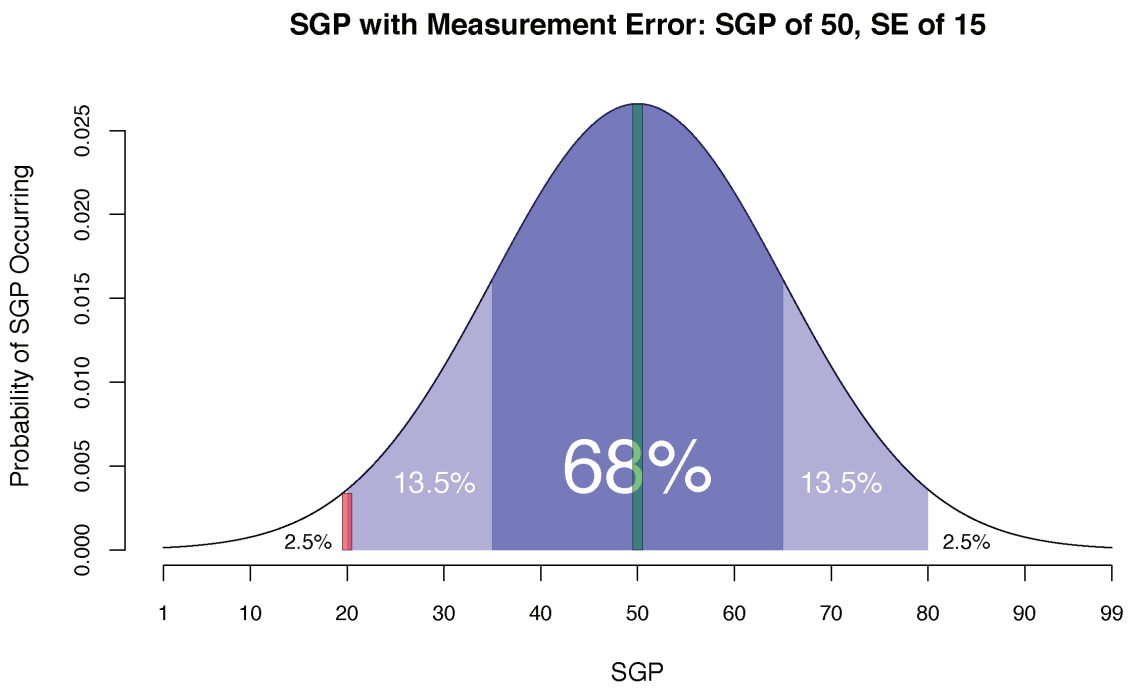
The most substantive reason put forward by Sireci et al. for not using SGPs is that, "estimates of the reliability of SGPs suggest they contain too much error to be useful." To support this claim the authors discuss simulation-based research showing a margin of error of 30 for an individual SGP. Based upon this amount of error, the authors suggest that one might as well simply "flip a coin" (Sireci et al., 2016, p. 6).[2]. Such a conclusion is not supported by the research cited and suggests a misunderstanding of how error impacts the utility and interpretation of the SGP measure.

A "margin of error" of 30 translates to a standard error of 15. In our work with state departments of education we have found, in general, standard errors of SGPs to range from 5 to 15, with a mean of 10. To illustrate misunderstandings present in Sireci et al., we accept the high end standard error of 15 reported by Sireci et al.

Figure 1 shows what the distribution of simulated SGPs would look like based upon an observed SGP of 50 and a standard error of 15. The height of the distribution indicates the relative frequency of each SGP. The claim by Sireci et al. that one might as well just flip a coin would correspond to a horizontally flat (i.e., uniform) distribution which Figure 1 clearly is not.

In Figure 1, an SGP of 50 (green bar) occurs much more frequently (7.38 times as frequently) as the SGP of 20 (red bar). Moreover, SGPs between 35 and 65 would be expected to occur 68% of the time whereas values in the tails occur much less frequently (16% of the time in each tail). A teacher could, for example, feel fairly confident that a student with an SGP of 50 does not have low growth (an SGP < 35) as such a result would only be expected to occur 16% of the time. All of these results derive from basic statistical concepts introduced in the first semester of a statistics/ measurement course. It is unclear how Sireci et al. reached their "flip a coin" conclusions presented in their report.

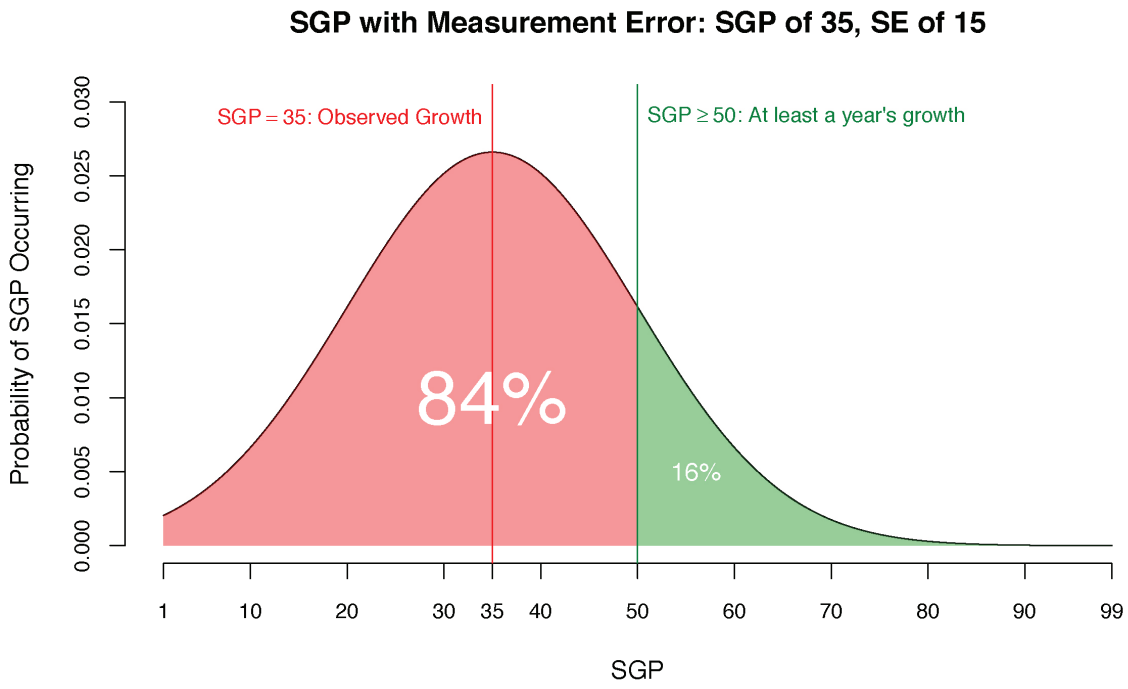
FIGURE 1: DISTRIBUTION OF SIMULATED SGPS BASED UPON AN OBSERVED SGP AND A STANDARD ERROR OF 15.



Beyond being far less random than a flip of a coin, if looked at appropriately, an SGP can be extremely useful. For example, SGPs are well suited for the screening of students demonstrating low growth. Students demonstrating low growth show low levels of learning relative to other students at their level of achievement. Such students may have missed something their peers got and are in need of remediation on that topic.

Consider, for example, a student demonstrating an SGP of 35, again with a worst case standard error of 15. Figure 2 depicts the distribution of simulated SGPs based upon this scenario. Even with a worst case standard error of 15, we can conclude with 84% certainty that a student with an observed SGP of 35 has a true SGP below 50, a value states often use to reflect a year's growth.[3] Such information is tremendously useful for teachers and parents to begin inquiring further about whether other evidence of student achievement (dis)confirms the low level of learning indicated by the SGP.

FIGURE 2: DISTRIBUTION OF SIMULATED SGPS BASED UPON AN OBSERVED SGP OF 35 AND A STANDARD ERROR OF 15.



Going further we can consider a student with an even lower growth percentile of 20. Figure 3 depicts the distribution of simulated SGPs based upon this scenario. For this student we have 98 percent certainty that they did not make at least a year's worth of learning (i.e., growth). This indicates an extremely high level of certainty that less learning has taken place over the course of the year than is typical for a student at that level of achievement.

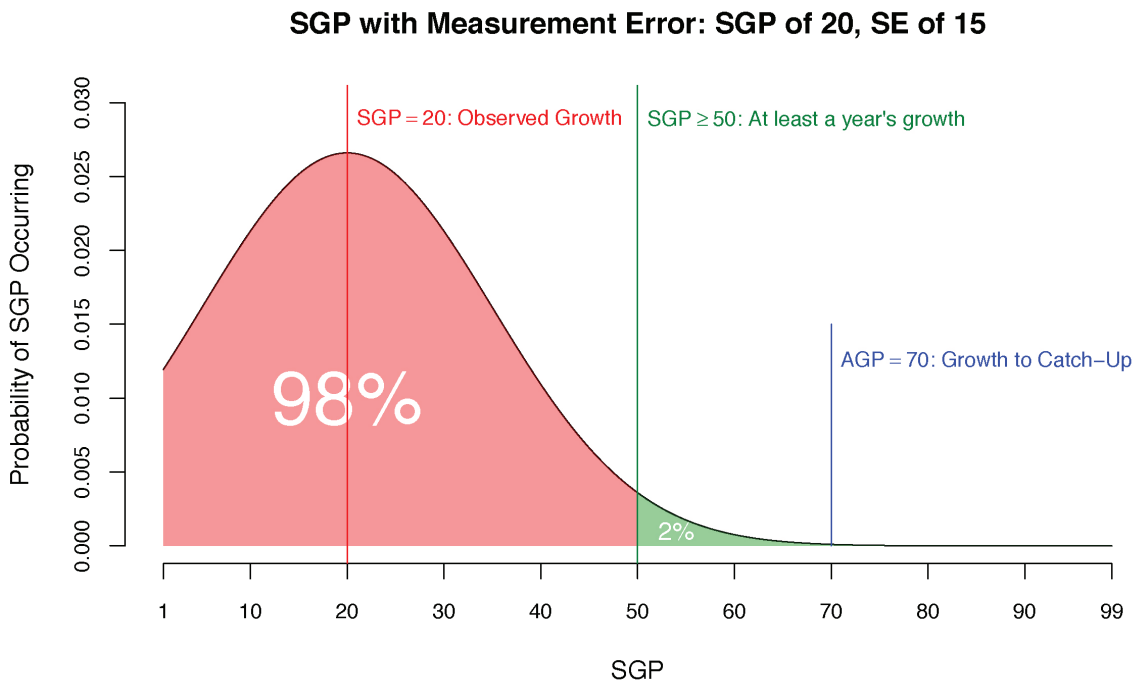
Given that one-fifth of all students in the state necessarily demonstrate growth at or below the 20th percentile (tens of thousands of students), this norm-referenced comparison can be a useful screening tool to assist parents, teachers and other stakeholders in identifying students having a tough year and in possible need of remediation.[4]

Additionally, if a criterion-referenced adequate growth percentile (AGP) is added, we can discuss the certainty of whether the student has met or exceeded a threshold for growth deemed necessary for them to reach or maintain proficiency or some other criterion-referenced achievement outcome. For non-proficient students, this is often referred to as *catch-up growth* and is in excess of 50th percentile growth as most students need more than a year's worth of growth to catch-up. In Figure 3 the students AGP is indicated by 70. With an observed SGP of 20 the chances that the students growth exceeds their criterion-referenced threshold of 70 is just 0.04%, 4 in 10,000. We can be sure that the student is not growing as much as we would like.

The scenarios associated with both Figures 2 and 3 illustrate several of the useful insights that can be derived from SGP data for tens of thousands of students taking annual state summative student assessments. Complementary insights can be derived for the thousands of high growth students as well, where evidence of a remarkable year of learning has taken place that will hopefully continue in the coming year.

In attempting to show that SGPs contain too much error to be useful, Sireci et al. make two errors. They incorrectly convey that error is uniformly distributed across SGPs when it is in fact normally distributed since the error derives from normally distributed scale score measurement error. Second, they assert that in order for data to be useful, it must have high reliability. SGP reliability is moderate and, as the results here demonstrate, is sufficient to yield many useful insights about individual student progress. Ultimately, reliability is like validity in that purpose must be considered: Just as one asks, "Valid for what purpose?" one also should ask, "Reliable enough for what purpose?"

FIGURE 3: DISTRIBUTION OF SIMULATED SGPS BASED UPON AN OBSERVED SGP OF 20 AND A STANDARD ERROR OF 15.



Reliability/Precision of Aggregate Level SGPs

Combined with the discussion of the reliability/precision of individual SGPs is a discussion of the precision of aggregate level SGPs. This issue arises frequently in SGP/Value-Added accountability decision making where imprecision is often included in the decision making process. The authors cite a study by Lash, Makkonen, Tran and Huang (2016) which utilized generalizability theory to report on the year-to-year stability of growth scores. The results are consistent with moderate year-to-year correlations of teacher level results found in other states.[5]

Lash et al. incorrectly confuse the stability of median SGPs from year-to-year with the precision of a median SGP within a given year. They summarize their results by concluding:

> For the annual teacher-level growth scores, the standard error of measurement was 12.22 for math and 11.31 for reading (see table 1). This means that 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 100 point score scale, and 44 points for reading.

Their results are for a prediction interval based upon year-to-year correlations, not the standard error of the median SGP in a given year. This is readily apparent if one considers that the standard error they provide is a constant for all schools and doesn't change as a function of the number of students a teacher instructs. This runs counter to common sense and the Central Limit Theorem: Median SGPs based upon a large sample size will be more precise than those based on a small sample size.

Interpretation and Use of SGPs

As mentioned previously, SGPs were designed by the Center for Assessment in close collaboration with the Colorado Department of Education. An initial design imperative of this partnership was to create a growth measure that is valid, reliable, and easy to understand using the vertically-scaled Colorado Student Assessment.[6] Following Colorado's adoption of the model in 2007, several other states, including Massachusetts, began investigating the model in 2008. In 2009, following peer review, the model was approved for use as part of the United States Department of Education Growth Model Pilot Program.[7] In 2010 the model received the prestigious annual award for *Outstanding Dissemination of Educational Measurement Concepts to the Public* by the National Council on Measurement in Education. Based upon model adoption, awards received, and discussions with thousands of stakeholders nationally, we feel confident that the model has achieved the original design imperative established in Colorado. NCIEA and states continue striving to improve the resources supporting the appropriate interpretation and use of SGP.

In contrast, Sireci et al. argue that SGPs are difficult to understand and are misleading to users. Their primary argument that SGPs are hard to understand is based upon the fact that the procedure used to calculate SGPs (quantile regression) is complex. This is a red herring based upon the fallacy that in order to understand something one needs to understand how it is calculated.

- The number π is difficult to calculate but has an easy to understand conceptual basis: The ratio of the circumference to the diameter of a circle.[8]
- The scaled scores reported in educational assessment involve very complex calculations yet are communicated widely to parents with the expectation that they can understand them.
- Most parents cannot calculate the height and weight percentiles provided by their doctor for their children yet they understand what they mean.

Regression analysis (whether linear or quantile) is a standard data modeling approach dating back more than 200 years for calculating the conditional mean/quantile as a function of independent variable(s). In documenting this for non-technical users, the heuristic of an "academic peer" is used to discuss the manner in which the regression analysis function relates values on the independent variable(s) to values on the dependent variable. This comports with the definition of regression analysis which seeks to understand, "as far as possible with the available data how the conditional distribution of the response y varies across subpopulations determined by the possible values of the predictor or predictors"(Cook & Weisberg, 1999, p. 27).[9]

Beyond the complexity of calculations, Sireci et al. (p. 4) go further and assert that efforts to help users understand SGPs by relating them to height and weight percentiles are misleading.

> 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.

It is unclear on what information the authors base their assertion. A cursory review of the literature on anthropomorphic growth charts shows that regression based techniques, including quantile regression, are the *status quo* in the field (Cole, 1994; Wei, 2004; Wei, Pere, Koenker, & He, 2006; Wei & He, 2006). Moreover, as referenced in Betebenner (2008), the methodology associated with the calculation of SGPs (Betebenner, Iwaarden, Domingue, & Shang, 2016) derives from this literature.

The issue of whether SGPs are percentiles at all is also raised by Sireci et al. (2016) (p. 3): "They also are not percentiles as most people think of them." Later, (p. 4), quoting Clauser et al. (p. 12): "SGPs are not percentiles as they are commonly understood, but instead likelihood estimates of a particular score pattern–not direct comparisons to a student's place within a peer group." It is not clear what the "commonly understood" definition of a percentile is or what "most people think of them". Whatever definition the authors use fails to recognize that percentiles, by definition, are statements of probability.[10] Just as the mean of a conditional distribution is still a mean and a quantile of a conditional distribution is still a quantile. A percentile of a conditional distribution is still a percentile. See Chapter 1.4 of Koenker (2005), for the mathematical elaboration. This is consistent with height and weight percentiles familiar to parents that utilize height and weight *conditioned* on age.

Unfortunately, misunderstandings by Sireci et al. (p. 6) extend beyond just definitions and into how data are used in practice.

> Clauser et al. (2016) 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.

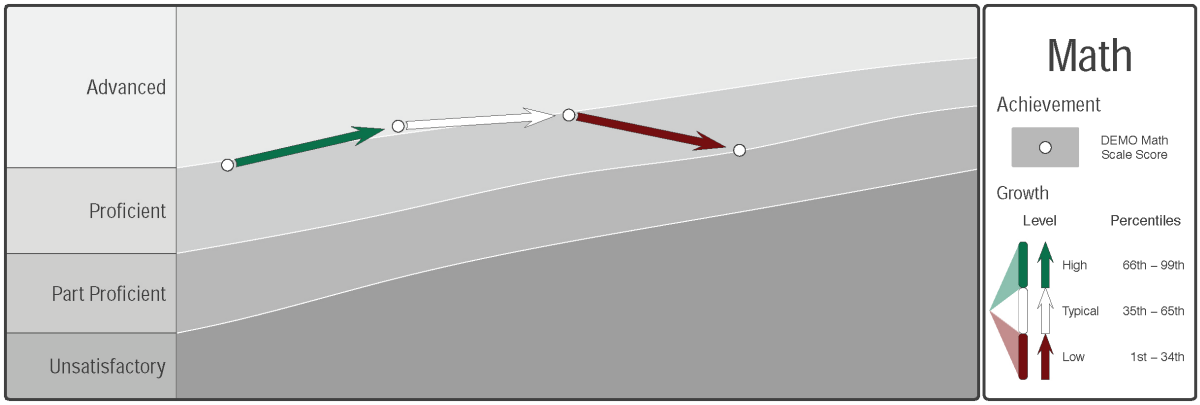
In the last sentence Sireci et al. conclude that high achieving students with low growth are incorrectly identified for remediation. On the contrary, a currently high achieving student could have been even higher achieving in the previous year. Even for high achievers, a low SGP implies that whatever learning occurred for the student was less than for the typical student. As such, the student could very well need remediation in particular topics. Figure 4 illustrates such a student in mathematics who dropped precipitously in math (SGP of 17) in the 2015-2016 school year but remained proficient. Had they demonstrated typical progress (SGP of 50), their

achievement would have remained at the upper end of the proficient range. Based upon an impoverished understanding of norm-referenced growth by Sireci et al. and Clauser et al., it is troubling that the researchers disparage a practice conducted by a majority of principals surveyed that is exactly what they should be doing. As Figures 2, and 3 in Section Reliability/ Precision of SGPs demonstrate, even after taking account of error of measurement in SGPs (what the authors refer to as "full of error"), there is useful information for evaluating school growth and students with exceptionally high and low growth.

Another troubling misunderstanding in Sireci et al. concern their belief that "SGPs encourage comparing students to one another rather than the knowledge and skill areas being taught." Norms, by their very nature, are comparative. This should not be taken as a criticism of norms. Angoff (1974) emphasized that norm- and criterion-referenced understandings are mutually supportive and not in conflict with one another. As emphasized in Betebenner (2008) (the title of which is *Norm- and Criterion-referenced student growth*) one needs *both* norms and a criterion-referencing to have a complete understanding of a phenomenon like student growth. To that end, states use SGPs to calculate adequate growth percentiles, indicating the amount of growth necessary for a student to reach/maintain achievement outcomes like those established by the state performance standards. This methodology has a long history as it was approved by the USED as part of the growth model pilot program as the Colorado Growth Model (Spellings, 2005).[11] Additionally, states who were reluctant to re-norm each year have fixed/anchored growth norms that avoid issues related to the zero-sum nature of norms.[12]

Interpretation and use were a focal point of SGP development both in terms of the measure itself as well as the reporting associated with it.

FIGURE 4: DIAGNOSTIC REPORT OF A HIGH ACHIEVING MATHEMATICS STUDENT DEMONSTRATING LOW GROWTH IN 2015-2016 INDICATING NEED FOR POSSIBLE REMEDIATION.



	Grade 3 2012–2013	Grade 4 2013–2014	Grade 5 2014–2015	Grade 6 2015–2016	Grade 7 2016–2017	
Scale Score Achievement Level	505 Proficient	547 Advanced	559 Proficient	521 Proficient		Achievement
Growth Percentile Growth Level		70 High	46 Typical	17 Low		Growth

Alignment with the *Standards for Educational and Psychological Testing*

Sireci et al. suggest that the use of SGPs violates the *Standards for Educational and Psychological Testing* (American Educational Research Association and American Psychological Association and National Council on Measurement in Education and Joint Committee on Standards for Educational and Psychological Testing (U.S.), 2014). The bulk of the evidence cited to support this has been refuted in the previous sections. Beyond that, such an assertion ignores much of the dissemination (both technical and non-technical) associated with SGP analyses done by the Center for Assessment that supports the standards. For example,

Standard 6.10 When test score information is released, those responsible for testing programs should provide interpretations appropriate to the audience. The interpretations should describe in simple language what the test covers, what the scores represent, the precision/reliability of the scores, and how scores are intended to be used.

Standard 9.6 Test users should be aware of potential misinterpretations of test scores; they should take steps to minimize or avoid foreseeable misinterpretations and inappropriate uses of test scores.

Standard 9.7 Test users should verify periodically that their interpretations of test data continue to be appropriate; given any significant changes in the population of test takers, the mode(s) of test administration, or the purposes of testing.

Standard 9.8 When test results are released to the public or to policy makers, those responsible for the release should provide and explain any supplemental information that will minimize possible misinterpretations of the data.

To encourage proper understand and use of the results by non-technical stakeholders, a major thrust of the work of the Center for Assessment on SGP is to assist states in producing reporting systems and SGP resources to help communicate results. The SchoolView data visualization platform is used by many states to help understand student achievement and growth simultaneously.[13] The visualization platform was a finalist for the Adobe Max Award at its 2009 Adobe Max convention. The open source SGP Package (Betebenner et al., 2016) creates numerous data visualizations including summary bubble plots and student reports (see Figure 4) to help nontechnical users understand student progress both at the individual and group level.

The Center for Assessment has also worked with states to develop video tutorials to help non-technical users understand student growth percentiles.

Mississippi An introduction to the Mississippi Growth Model

Georgia An introduction to the Georgia Student Growth Model

Virginia An introduction to the Virginia Growth Model

Hawaii An introduction to the Hawaii Growth Model

For more sophisticated audiences, technical reports are produced for each state's SGP analyses. These technical reports provide details of the model technical specification and document the technical characteristics of the model such that the SGPs produced are not, for example, subject to model bias or constrained by highest obtainable/lowest obtainable scale scores (i.e. ceiling/floor effects). In addition to state reports, the NCIEA partnered with CCSSO to work with states participating in the two testing consortia (PARCC and SBAC) to understand issues related to student growth when transitioning between assessment programs.

Georgia Georgia Student Growth Model

Massachusetts Massachusetts Growth Model

Colorado Colorado Growth Model

Hawaii Hawaii Growth Model

Center for Assessment/CCSSO Using Student Growth Percentiles During the Assessment Transition: Technical, Practical and Political Implications

Validation of SGPs for use in accountability have also been performed as part of state efforts. Georgia, for example, commissioned a study to validate SGPs for use in teacher evaluation (Briggs, Dadey, & Circi-Kizil, 2014). In claiming that there has been no research validating the use of SGP for either diagnostic or accountability purposes, it is not clear whether Sireci et al. bothered to look.

CONCLUSION

The Center for Assessment encourages rigorous review of all of our work including the SGP methodology. Sireci et al. make numerous strong claims against the use of SGPs for either diagnostic or accountability purposes. This review has rigorously examined these claims and in each case found them to be based upon misunderstandings and basic errors. Sadly, the likely result of such unsubstantiated and false assertions put forward by Sireci et al. will be general confusion among stakeholders on an issue of national importance to educators and policy makers.

NOTES

1. Rigorous research on SGPs does exist. See, for example, Castellano and Ho (2013), Briggs, Kizil and Dadey (2014), McCaffrey, Castellano, and Lockwood (2014), and Monroe and Cai (2015).

2. See also Vaznis (2016) for a quote comparing an SGP to a flip of a coin.

3. In such cases it is important to recognize that only one side of the distribution is utilized leading to a significant increase in our ability to understand the growth of students, even in the presence of measurement error.

4. Sireci et al., citing a recent paper by Clauser, Keller, and McDermott (Clauser, Keller, & McDermott, 2016) assert that using SGPs to identify students for remediation is wrong, especially when those students are high achieving. See Section Interpretation and Use of SGPs for further discussion of this incorrect assertion.

5. See, for example, Georgia Student Growth Model.

6. Sireci et al. claim that SGPs were developed due to states lacking a vertical scale with which to measure “growth”. Colorado’s state assessment was vertically-scaled. A vertical scale does not solve the issue of growth. See Subtraction isn’t a growth model, developed in conjunction with the Georgia Department of Education to assist non-technical users to understand student growth.

7. Colorado Growth Model Pilot for USED.

8. For a review of techniques used historically to calculate π see the Wikipedia entry for π.

9. See Jan De Leeuw's Editor Introduction (p. xi) to Richard Berk's Regression Analysis: A Constructive Critique. Also available here: Series Introduction by Jan De Leeuw

10. See Koenker (2005), especially Equations 1.7 and 1.8.

11. Documentation associated with the SGP criterion-referenced model (i.e., SGP growth-to-standard model) can be found at SGP Criterion-Referenced Model.

12. Georgia and Massachusetts pursued baseline referenced growth as part of their accountability systems: Georgia Student Growth Model & Massachusetts Growth Model.

13. Hawaii SchoolView Data Visualization

REFERENCES

American Educational Research Association and American Psychological Association and National Council on Measurement in Education and Joint Committee on Standards for Educational and Psychological Testing (U.S.). (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Angoff, W. H. (1974). Criterion-referencing, norm-referencing and the SAT. *College Board Review*, 92 , 2–5.

Berk, R. A. (2003). *Regression analysis: A constructive critique*. Thousand Oaks, CA: Sage.

Betebenner, D. W. (2008). Toward a normative understanding of student growth. In K. E. Ryan & L. A. Shepard (Eds.), *The future of test-based educational accountability* (pp. 155–170). New York: Taylor & Francis.

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

Betebenner, D. W. (2012). Growth, standards, and accountability. In G. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (pp. 439–450). New York: Routledge.

Betebenner, D. W., Iwaarden, A. V., Domingue, B., & Shang, Y. (2016). SGP: *Student growth percentiles & percentile growth trajectories*. [R Software Package, version 1.5-0.0]. (Available online from http://cran.r-project.org/package=SGP)

Briggs, D. C., Circi-Kizil, R., & Dadey, N. (2014). *Adjusting mean growth percentiles for classroom composition* (Technical Report). Boulder, CO: University of Colorado, Boulder. (Available from http://www.gadoe.org/Curriculum-Instruction-and-Assessment/Assessment/Documents/Briggs et al 2014 Adjusted.pdf)

Briggs, D. C., Dadey, N., & Circi-Kizil, R. (2014). *Comparing student growth and teacher observation to principal judgments in the evaluation of teacher effectiveness* (Technical Report). Boulder, CO: University of Colorado, Boulder. (Available from http://www.gadoe.org/Curriculum-Instruction-and-Assessment/Assessment/Documents/Briggs et al 2014 Survey.pdf)

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), 190–214.

Clauser, A. L., Keller, L. A., & McDermott, K. A. (2016). Principals' uses and interpretations of student growth percentile data. *Journal of School Leadership*, 36 , 6–33.

Cole, T. J. (1994). Growth charts for both cross-sectional and longitudinal data. *Statistics in Medicine*, 13 , 2477–2492.

Cook, R. D., & Weisberg, S. (1999). *Applied regression including computing and graphics*. New York: Wiley.

Koenker, R. (2005). *Quantile regression*. Cambridge: Cambridge University Press.

Lash, A., Makkonen, R., Tran, L., & Huang, M. (2016). *Analysis of the stability of teacher-level growth scores from the student growth percentile model* (Tech. Rep.). Washington, D.C.: National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory West. (REL 2016âAS104)

McCaffrey, D. F., Castellano, K. E., & Lockwood, J. R. (2014). *An evaluation of technical issues with the student growth model component of the georgia teacher and leader evaluation system.* (Technical Report). Princeton, NJ: Educational Testing Service.

Monroe, S., & Cai, L. (2015). Examining the reliability of student growth percentiles using multidimensional irt. *Educational Measurement: Issues and Practice*, 34 (4), 21–30.

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* (1), 4–14.

Sireci, S. G., Wells, C. S., & Keller, L. A. (2016). *Why we should abandon student growth percentiles*. (Research Brief No. 16-1). Amherst, MA: Center for Educational Assessment. (Retrieved June 24, 2016 from http://www.umass.edu/remp/news_SGPsResearchBrief.html)

Spellings, M. (2005, Nov). *Secretary Spellings announces growth model pilot* [Press Release]. U.S. Department of Education. (Retrieved June 15, 2016 from http://www2.ed.gov/admins/lead/account/growthmodel/proficiency.html)

Vaznis, J. (2016, May 31). *Plan to rate teachers based on test scores is under fire*. ([Online; posted 31-May-2016])

Wei, Y. (2004). *Longitudinal growth charts based on semi-parametric quantile regression*. (Unpublished doctoral dissertation, University of Illinois, Urbana-Champaign)

Wei, Y., & He, X. (2006). Conditional growth charts. *The Annals of Statistics, 34* (5), 2069–2097.

Wei, Y., Pere, A., Koenker, R., & He, X. (2006). Quantile regression methods for reference growth charts. *Statistics in Medicine*, 25 , 1369–1382.