September 2016

Precision of student growth percentiles with small sample sizes

Michael J. Culbertson

RMC Research Corporation

REL Central
Regional Educational Laboratory
at Marzano Research

Colorado • Kansas • Missouri • Nebraska
North Dakota • South Dakota • Wyoming

This Regional Educational Laboratory Central report was developed under Contract ED-IES-12-C-0007 from the U.S. Department of Education, Institute of Education Sciences. The content does not necessarily reflect the views or policies of the Institute of Education Sciences or the U.S. Department of Education, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

This report is available on the Regional Educational Laboratory Central at Marzano Research website at www.relcentral.org.

REL Central at Marzano Research
12577 E. Caley Ave
Centennial, CO 80111

RELCentral
Regional Educational Laboratory
at Marzano Research

SUMMARY

States in the Regional Educational Laboratory (REL) Central region serve a largely rural population with many states enrolling fewer than 350,000 students. A common challenge identified among REL Central educators is identifying appropriate methods for analyzing data with small samples of students. In particular, members of the REL Central Educator Effectiveness Research Alliance in Colorado, Kansas, South Dakota and Wyoming are interested in understanding how the precision of student growth percentiles (SGPs), a measure of student growth in their accountability systems, varies depending on sample sizes. To support the EERA members, this study investigates the precision of SGP estimates when SGP calculations are based on small sample sizes. In small samples, very few students have exactly the same prior achievement score. In order to increase the sample size at any given level of prior achievement, some states with small student populations have considered using a coarser measure of prior achievement, such as dividing students into four achievement levels, instead of using exact achievement scores. This study investigates how categorizing students coarsely by prior achievement level before SGP analysis affects precision. Findings suggest that SGP estimates are less precise for high- and low-achieving students than for students with average achievement when the total sample size is small. Moreover, categorizing students coarsely by prior achievement before estimating the SGP model results in an increase in the precision of SGP estimates for the highest and lowest achieving students; however, this technique also reduces the similarity of students whose growth is compared. Results for different sample sizes may help states plan their strategy for SGP implementation and communication with stakeholders, such as reporting SGP bands instead of single numbers or cautioning stakeholders about making comparisons between SGPs that are similar.

CONTENTS

WHY THIS STUDY?

States in the Regional Educational Laboratory (REL) Central region serve a largely rural population. They face a common challenge of identifying appropriate methods for analyzing data with small samples of students. In particular, members of REL Central's Educator Effectiveness Research Alliance in Colorado, Kansas, South Dakota and Wyoming are interested in understanding how the precision of student growth percentiles (SGPs) varies depending on sample sizes. Colorado, Kansas, South Dakota, and Wyoming make use of student growth percentiles (SGPs) as a measure of student growth in their educational accountability systems. SGPs describe how much students have grown relative to other students with similar prior achievement. While an SGP is used to measure an individual student's growth, individual SGPs are based on a model of student growth that is influenced by the performance of all students included in the analysis, and the exact SGP results obtained will vary from sample to sample (year to year). For example, a student with a score of 100 in grade 3 and 110 in grade 4 could have an SGP of 45 based on an SGP analysis of data from 2014, but a student with the same grades 3 and 4 scores could have an SGP of 46 based on an SGP analysis of data from 2015, due to differences in the 2014 and 2015 samples. The amount of this variability depends, among other things, on the number of students included in the analysis. This study investigates how variability in SGP estimates is related to the number of students included in the analysis, as well as how categorizing students coarsely by prior achievement before the analysis can affect the results.

Who is this report for?

This report is for state and local education agency leaders and policymakers who use SGP results to inform policy decisions or who are responsible for informing future plans for measuring student growth. There is no one correct way to conduct an SGP analysis, and different variations in the analysis design, such as the number of years of prior scores to include, can affect the specific SGPs that students receive. The results of this study may inform design decisions for SGP analyses, as well as how to contextualize SGPs for teachers, parents, and other local stakeholders.

What are student growth percentiles?

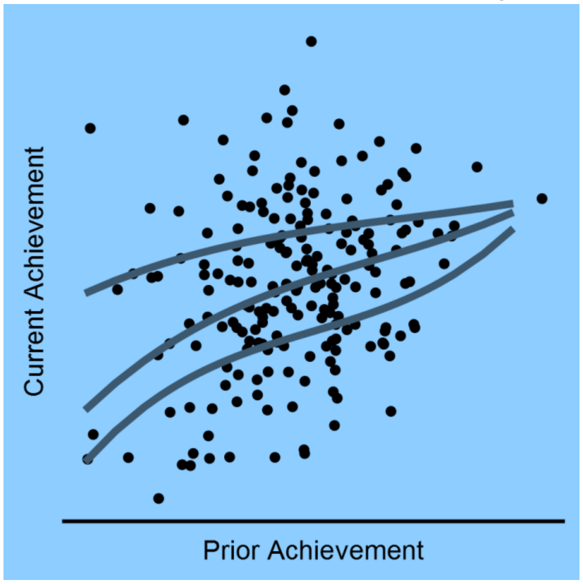
Student growth models track achievement scores of individual students over time in order to determine the extent to which students' learning is progressing (Goldschmidt et al., 2005). Conceptually, SGPs group students with similar past performance and then rank each student's academic growth within his or her group. Because students who start the year in different places academically might be expected to have different growth, SGPs compare students who have similar past academic achievement—each student's growth is measured relative to other students who had the same starting point at the beginning of the year. For example, a student with a score last year of 100 and an SGP of 45 received a score this year that was better than 45 percent of the other students who had the same score of 100 last year. SGPs thus measure *conditional* growth because the measurement depends on students' prior test scores.

Box 1. How are student growth percentiles calculated?

Many implementations of SGP calculations, including those used by states in the REL Central region, involve fitting 99 curves to achievement data so that the curves divide students with the same prior achievement (using one or more years of data) into percentiles based on current achievement. The estimated SGP for a given student is found by identifying the two curves between which the student's performance falls. The curves used in SGP calculations group students by prior achievement scores by assigning the same point on the curve to all students with the same prior achievement score. Grouping students in this way allows the computation of SGPs for students with very high or very low prior achievement scores, where there are relatively few students with the same prior achievement score who could be compared with one another directly. SGP calculations often use flexible curves instead of straight lines to account for different growth rates for high- and low-achieving students. For example, an SGP of 75 for a student with low achievement last year might correspond to an increase of 20 points because the student made up ground relative to state standards, but the same "high growth" for a student with high prior achievement might correspond to only 10 points because students with high achievement may have already mastered much of the material (figure 1).

In small samples, very few students have exactly the same prior achievement score. In order to increase the sample size at any given level of prior achievement, some states with small student populations have considered using a coarser measure of prior achievement, such as dividing students into four achievement levels, instead of using exact achievement scores. In this method, individual student scores would be replaced by a representative score for students in the same achievement category before finding the SGP curves. This method involves a trade-off: With fewer achievement levels, the sample size for each prior achievement level increases, but the similarity of students in the same category decreases because students with different prior achievement scores are placed in the same group.

Figure 1. The SGP model consists of a set of curves that divide the sample into percentiles



Note: For illustration purposes, only the 25th, 50th, and 75th curves are shown, dividing the sample into quartiles. In actual SGP calculations, 99 curves are used to divide students with the same prior achievement into percentiles based on their current achievement.

Source: Author's illustration with simulated data.

Why is the precision of student growth percentiles important?

Even if overall trends in student performance remain stable, SGP calculations will vary somewhat from year to year due to differences in the performance of the specific students used in the calculations each year. As a result, the SGPs calculated from a given sample are estimates of the "true" SGPs that would be known only if student achievement were measured without error.[1] Understanding the precision of these estimates is important for making justifiable comparisons between SGPs. For example, if a student had an estimated SGP of 65 and this estimate is precise to within 5 percentile points, the student's "true" SGP would likely fall between 60 and 70, and it would not be justifiable to conclude definitively that the student had greater growth than a student with an estimated SGP of 62.

In addition to measurement error in achievement scores, the number of students included in the analysis is one of the biggest factors that affects the precision of SGP estimates. In educational accountability, the sample size is generally predetermined by the number of students enrolled in a state and cannot be increased to improve precision. States in the REL Central region that use SGPs vary substantially in the size of their student populations (table 1). As such, it is important for each state to understand the precision of the SGP estimates given the state's own student population, because the precision of SGPs in one state may not reflect the precision of SGPs in another.

Table 1. Number of students in each state by grade level

Grade	Colorado	Kansas	South Dakota	Wyoming
3	66,429	36,669	10,310	7,410
4	66,140	36,435	9,842	7,086
5	66,326	36,536	9,621	7,131
6	65,161	35,907	9,499	6,804
7	64,808	36,412	9,384	6,884
8	63,820	36,457	9,531	6,867
PK–12	876,999	496,440	130,890	92,732

Source: U.S. Department of Education, National Center for Education Statistics, Common Core of Data (CCD), 2013–14.

Prior research has investigated how the overall precision of SGP estimates is affected by total sample size (such as Castellano & Ho, 2013; McCaffrey, Castellano, & Lockwood, 2015; Monroe & Cai, 2015). However, the precision of SGP estimates depends not only on the total sample size, but also the number of students who have similar prior academic achievement. There are generally many more students who have average achievement than who have very high or very low achievement. As a result, the SGP of a student with average achievement will have greater precision than the SGP of a student with very high or very low prior achievement because groups of students with similar average prior achievement will be larger than groups of students with similar high or low prior achievement. As a result, it is important to consider the

[1] Note that a student does not have a single "true" SGP, since SGPs are highly dependent on how the SGP model is set up, such as the number of prior year scores included.

conditional precision of SGPs, which refers to the precision of SGP estimates for students who have the same level of prior achievement. Even when the total number of students is large and the overall precision of SGP estimates is high, the conditional precision of SGP estimates for students with very high or very low prior achievement may be substantially lower than for most students. This suggests that understanding conditional precision of SGP estimates is important for all states, even those with large student populations.

WHAT THIS STUDY EXAMINED

This study investigated the following research questions:

1. How does the precision of SGP estimates depend on sample size?
2. How does the precision of SGP estimates change when calculations are based on coarse prior achievement levels instead of exact achievement scores?

The methods used in this study are described briefly in box 2 and in more detail in the appendix.

Box 2. Methods

Computer simulation. This study made use of computer simulations to evaluate the precision of SGP estimates. With real-world data, it is not generally possible to know a student's "true" SGP due to measurement error; but with simulated data, the estimated SGP from the SGP calculations can be compared with the intended SGP based on the simulation design. First, student achievement was simulated for a sample of students. Second, an SGP model was estimated, treating the simulated data as if they were real-world data. Finally, estimated SGPs for pairs of prior- and current-year achievement scores were compared with the corresponding intended SGP based on the simulated student population. This process was repeated 1,080 times, and the results were averaged over replications. Sample sizes varied from 1,000 to 100,000 students. Prior achievement was either left as exact achievement scores or categorized into 100 (percentile), 20, or 10 (decile) achievement levels.

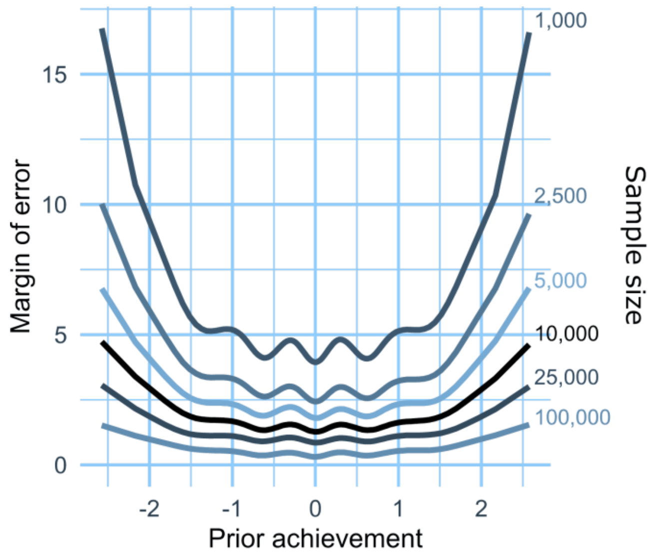
Margin of error. The margin of error of SGP estimates was used to give an indication of the estimate's precision. As the precision increases, the margin of error decreases—high precision means low error. To be confident that two SGPs are distinct, the difference between the SGPs should generally be at least twice as large as the margin of error. For example, if the margin of error is 10, there is a good chance that a student with an estimated SGP of 65 had a "true" SGP somewhere between 55 and 75 while a student with an SGP of 50 had a "true" SGP between 40 and 60. Since there is overlap in the range of likely values for students' "true" SGPs, one cannot be confident that the student with the SGP of 65 experienced greater growth than the student with an estimated SGP of 50. On the other hand, using the same margin of error of 10, it is clear that a student with an SGP of 65 experienced greater growth than a student with an estimated SGP of 40, because there is no overlap in the range of likely values for the "true" SGPs. A margin of error can be considered reasonable if it is small enough to allow meaningful distinctions to be made between students. For example, a margin of error of approximately 5 percentage points would allow making distinctions between students with SGPs that are at least 10 points apart, dividing student growth into approximately 10 meaningful groups. This study examined the conditional margin of error of SGP estimates given students' prior achievement scores.

WHAT THIS STUDY FOUND

The margin of error in SGP estimates is substantially larger for high- and low-achieving students than for students with average achievement when the sample size is small. For smaller sample sizes, the graph of prior achievement against the margin of error (figure 2) has a pronounced U shape, with relatively high margins of error for extremely high and low values of prior achievement and much lower margins of error for midrange values of prior achievement. In contrast, when the sample sizes were larger, the graphs of prior achievement against margin of error were much flatter, indicating that the margin of error was more consistent across levels of prior achievement.

When the SGP calculations were based on exact achievement scores, the average error in SGP estimates for students with high or low prior achievement was about four times larger than the average error in SGP estimates for students with average prior achievement, across all sample sizes. However, because the precision of all estimates increases for larger samples, the absolute size of the difference in error between average and high- or low-achieving students decreases for larger samples. For example, when SGPs were calculated based on a sample of 1,000 students, the margin of error for SGPs of students with high or low achievement was approximately 17, and the margin of error for SGPs of students with average achievement was approximately 4. However, with a sample of 100,000 students, the margin of error for SGPs of students with high or low achievement was approximately 1.5, and the margin of error for SGPs of students with average achievement was approximately 0.3. REL Central region states that use SGPs have sample sizes that provide reasonable margins of error using exact achievement scores. With 7,500 students, the margin of error of SGPs for students with high or low achievement was 5.5, when SGPs were based on exact achievement scores (table 2).

Figure 2. The margin of error of SGPs is larger for students with high or low achievement and decreases with increasing sample size

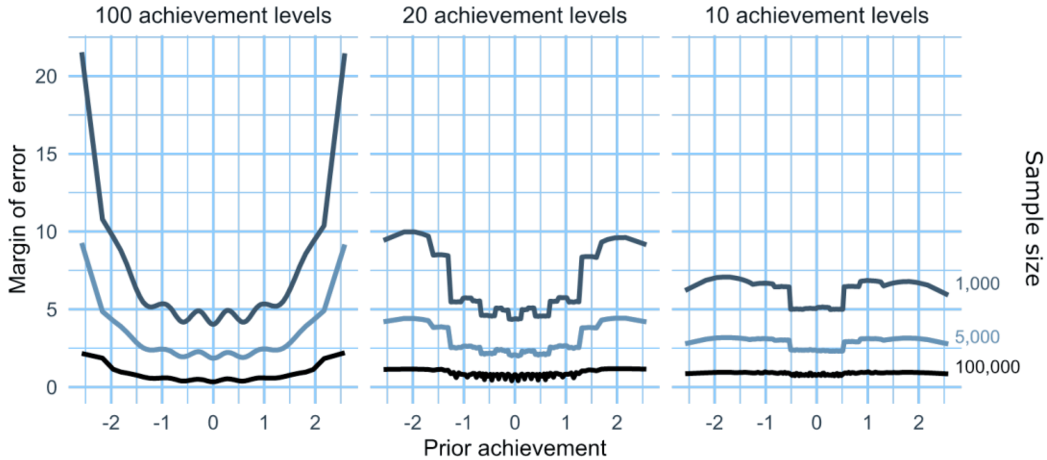


Note: Prior achievement is measured as a *z*-score. For an explanation of the undulation in the curves, refer to the end of the appendix.

Source: Author's analysis.

When SGP calculations were based on coarse prior achievement levels as compared to exact prior achievement levels, the margin of error of SGP estimates for the highest and lowest achieving students in small samples was substantially reduced (figure 3, right panel). When students were classified into prior achievement levels by deciles, the maximum margin of error was approximately 1.5 times the minimum for all sample sizes. In comparison, the maximum margin of error for SGPs based on exact prior achievement scores was approximately 4 times the minimum error. Even for small samples, the conditional margin of error of SGP estimates was relatively flat across different levels of prior achievement when students were classified by prior achievement deciles before calculating SGPs. However, categorizing students into coarse prior achievement levels decreases the similarity of students whose growth is compared. Given the sample sizes in REL Central region states, categorizing by deciles would only slightly decrease the margin of error. With a sample of 7,500, when students were categorized by deciles rather than exact achievement scores the margin of error decreased only to 2.6 (table 2).

Figure 3. Categorizing students into deciles by prior achievement substantially reduces the margin of error of SGPs for high- and low-achieving students



Note: Prior achievement is measured as a *z*-score.

Source: Author's analysis.

Table 2. Key characteristics of conditional margin of error of SGPs by sample size and number of coarse achievement levels

	Exact achievement scores			100 achievement levels			20 achievement levels			10 achievement levels		
Sample size	Max	Min	Mean	Max	Min	Mean	Max	Min	Mean	Max	Min	Mean
1,000	16.8	3.9	5.2	21.5	4.0	5.5	10.0	4.3	5.8	7.1	5.0	6.0
2,500	10.0	2.4	3.3	13.6	2.5	3.5	6.3	2.7	3.7	4.5	3.1	3.8
5,000	6.8	1.8	2.4	9.3	1.9	2.5	4.4	2.0	2.7	3.2	2.3	2.8
7,500	5.5	1.5	1.9	7.6	1.5	2.1	3.7	1.6	2.2	2.6	1.9	2.3
10,000	4.7	1.3	1.7	6.6	1.3	1.8	3.2	1.4	1.9	2.3	1.7	2.0
25,000	3.1	0.9	1.1	4.3	0.9	1.2	2.1	1.0	1.3	1.6	1.2	1.4
50,000	2.1	0.6	0.8	3.0	0.6	0.9	1.5	0.6	1.0	1.2	0.9	1.1
75,000	1.8	0.4	0.6	2.5	0.5	0.7	1.3	0.5	0.9	1.1	0.8	0.9
100,000	1.5	0.3	0.5	2.2	0.3	0.6	1.2	0.4	0.8	1.0	0.7	0.9

Source: Author's analysis.

IMPLICATIONS OF STUDY FINDINGS

Results from this study suggest that the sample sizes for the smallest REL Central states are likely sufficient for a reasonable margin of error for the highest and lowest achieving students. States should still caution stakeholders about making comparisons between SGPs that fall within the margin of error and consider carefully how to communicate to stakeholders that the margin of error in estimated SGPs differs for students with different levels of prior achievement. Reporting multiple conditional margins of error, for example in footnotes or technical notes, could confuse stakeholders with limited statistical experience.

Similar to the confidence band displays some states have adopted for reporting individual student assessment scores, reports that include a graphical display of the conditional margin of error for individual SGP estimates may be more readily understandable for a wide stakeholder audience. Alternatively, states could consider reporting SGPs in different units (for example, rounding estimated SGPs to the nearest 5 or 10, instead of reporting to the percentile, or reporting only broad growth designations, such as "low," "typical," and "high" growth) depending on the margin of error that corresponds to students' prior achievement level, using larger units for the highest and lowest achieving students and smaller units for students with average achievement levels.

Additionally, local education agencies that calculate SGPs based on small student samples could consider whether to classify students by coarse prior achievement levels (such as by decile) in order to decrease the margin of error in estimated SGPs for the highest and lowest achieving students. Agencies could also consider categorizing students only in the highest and lowest deciles, using scaled scores for the remaining students. However, categorizing students decreases the similarity of the students whose growth is compared to form SGPs.

LIMITATIONS OF THE STUDY

This study used a distribution for student achievement that assumes students at high and low levels of prior achievement grow at similar rates relative to their starting point. Results may differ for other patterns of student growth. This study also used an SGP model based on flexible curves, which is the model implemented by the popular SGP R package developed by Damian Betebenner (http://cran.r-project.org/package=SGP). Results may differ for SGP models that use straight lines, models that assume a particular conditional distribution for student achievement scores (such as the normal distribution, used in New York) instead of estimating percentile curves, or models that estimate SGPs for each prior achievement group separately instead of linking percentiles with curves.

This study considered SGP models based on only a single year of prior achievement data. North Dakota and South Dakota use a single year of prior achievement data in their SGP models. Colorado and Kansas use up to seven years of prior achievement data, while other states use only two or three years of prior achievement data. Because including additional years of prior achievement data changes the cohort of students against which any given student's achievement is compared, SGPs depend heavily on the design model. Moreover, including additional years of prior data may decrease the number of similar students for a student with any given level of prior achievement on each year of prior data, which is likely to impact the precision of the SGP estimate.

APPENDIX: DATA AND METHODOLOGY

This section provides additional detail about the implementation of this study for a technical audience. It is not necessary to read this section in order to understand the main findings of this study. R code implementing the methodology of this study may be obtained online at https://git.io/sgp-error-2016.

Simulation and coarse achievement level categorization

Student achievement was simulated from a standard bivariate normal distribution with means of 0, standard deviations of 1, and 0.8 correlation. This study combines measurement error and sampling error from a super-population of students who could have been included. This choice reflects the fact that SGP analysts often have little control over either measurement error or sampling. Measurement error in achievement scores attenuates the observed correlation between scores; in this study, the correlation of 0.8 was considered as the attenuated correlation, as opposed to the correlation of true scores. This model assumes that measurement error does not depend on the true score—conditional error for estimated SGPs of students with high or low prior achievement may be greater than found in this study if measurement error variance follows a *U* shape.

For a given sample, the prior achievement variable was either left as is or categorized into left-continuous quantiles. That is, score x was categorized in quantile i if $(i-1)/n < \Phi(x) \leq i/n$, where $\Phi(x)$ is the standard normal cumulative distribution function and n is the number of categories (for example, $n = 100$ for percentiles). Categorized prior achievement scores were replaced with the median score for the category.

SGP model estimation

The SGP distribution was estimated via quantile regression using seventh-degree cubic B-splines (Betebenner, 2009).[2] B-spline knots were specified at the standard normal quantiles (0.001, 0.2, 0.4, 0.6, 0.8, 0.999). Regression equations for current achievement score given prior achievement were estimated for the 99 boundaries between percentiles (that is, the 0.01, 0.02, . . . , 0.99 quantiles). Note that this is slightly different from Betebenner's SGP R package, which estimates regression equations for the 0.005, 0.015, . . . , 0.995 quantiles.

Evaluation

Once the SGP model was estimated based on a given sample, SGP estimates were obtained for a 100 x 100 grid of (prior achievement, current achievement) pairs. The grid was constructed so that the mean of a statistic evaluated at the grid points approximates the expectation of that statistic over the population statistic. In particular, the prior achievement grid was fixed at the 0.005, 0.015, . . . , 0.995 quantiles of the

[2] Ordinary linear regression uses a linear function to specify the conditional mean of normally distributed outcome data. Quantile regression, on the other hand, uses a linear function to specify a given quantile, such as the conditional median. Seventh-degree cubic B-splines specify a flexible class of polynomials composed of the sum of seven cubic functions.

standard normal distribution. Current achievement values in the grid were based on the quantiles of the conditional distribution of current achievement, for a given prior achievement level. That is, the current achievement levels in row i of the grid, corresponding with the $x_i = \Phi^{-1}\left({}^{i}/_{100} - 0.005\right)$ prior achievement level, were the 0.005, 0.015, . . . , 0.995 quantiles of a normal distribution with mean $0.8x_i$ and standard deviation $\sqrt{1 - 0.8^2}$. The same grid was used for the exact prior score condition and the categorized prior achievement conditions so that the estimated error was based on expectation over the population distribution in all conditions.

For each grid point $\left(x_i, y_{ij}\right)$, predicted values for each of the 99 quantile functions $\hat{y}_{ik}$ were calculated from the estimated SGP model, and the estimated SGP for a student with scores $\left(x_i, y_{ij}\right)$ was calculated as $\widehat{SGP}_{ij} = 1 + \sum_{k=1}^{99} I\left(y_{ij} > \hat{y}_{ik}\right)$, where $I(\cdot)$ is the indicator function. For example, if the grid point y_{ij} was greater than 45 of the quantile boundaries, the estimated SGP was 46. Computing estimated SGPs from the predicted quantile functions in this way effectively “uncrosses” any quantile functions whose splines cross at the edges (Castellano & Ho, 2013).

The estimated SGPs were compared with the theoretical quantiles from the conditional population distributions, $\Phi\left(y_{ij}|x_i\right)$. Due to the configuration of the grid, the theoretical quantiles in the exact scaled scores condition are simply 0.005, 0.015, . . . , 0.995. For models based on categorized prior achievement data, the conditional distribution of current achievement, given that prior achievement fell into quantile $i = 1, 2, \ldots, n$, requires integration of the bivariate normal density function over the values of x in quantile i, namely from $q_{i-1} = \Phi^{-1}\left({}^{i-1}/_{n}\right)$ to $q_i = \Phi^{-1}\left({}^{i}/_{n}\right)$. The theoretical SGP of y_{ij} given that x_i fell into quantile i can be computed from the bivariate normal cumulative distribution function $\Phi(x, y)$ as:

$$SGP_{ij} = 100n\left[\Phi\left(q_i, y_{ij}\right) - \Phi\left(q_{i-1}, y_{ij}\right)\right].$$

The squared difference between estimated and theoretical SGPs was calculated for each point in the evaluation grid, and the mean of the squared differences was computed across the 1,080 replicate datasets in each categorization-by-sample-size condition. Finally, the mean was calculated across grid points with the same prior achievement level x_i, and the square root was taken to obtain the conditional root-mean-square error, plotted in figures 2 and 3. This error was rescaled by 1.64 to yield a margin of error at 90 percent confidence. The error in this study includes both the bias and variance of the estimated SGPs. While separating error in bias and variance is instructive in theoretical statistical analyses, the total error is more salient in an applied policy context, where bias and variance cannot be manipulated independently.

The number of replications in this study, 1,080, was sufficient for the Monte Carlo error to be less than 2 percent in almost all study conditions and no more than 4 percent (Koehler, Brown, & Haneuse, 2009).

Undulation in conditional root-mean-square error curves

The undulation in conditional margin of error at the base of the curves in figure 2 results from the flexibility of the curves used in the SGP model. In the simulated bivariate normal distribution, the true percentile curves are straight lines, but the SGP model provides flexibility in these curves to account for potential differences in learning rates at different achievement levels, which may be encountered in real student data. This additional flexibility, which is not needed under the simulation conditions, allows the SGP model to fit the particular characteristics of any given sample more closely than it should according to the population distribution, slightly increasing the error at particular points. The valleys in the conditional margin of error curves correspond with fixed points called "knots" used in the technical specification of the flexible percentile curves. In these results, the difference between the local valleys and peaks is about 20 percent, and for moderately sized samples, the absolute difference is small.

Whether student growth is linear or nonlinear is an empirical question, and states should investigate the actual patterns of student growth in their population. Small states, in particular, may benefit from using a simpler straight-line SGP model if their population is well approximated by bivariate normal distributions, because flexible curves generally require more data for precise estimates.

REFERENCES

Betebenner, D. (2009). Norm- and criterion-referenced student growth. *Educational Measurement: Issues and Practice*, *28*, 42–51. doi:10.1111/j.1745-3992.2009.00161.x

Castellano, K., & Ho, A. (2013). Contrasting OLS and quantile regression approaches to student “growth” percentiles. *Journal of Educational and Behavioral Statistics*, *38*, 190–215. doi:10.3102/1076998611435413

Goldschmidt, P., Roschewski, P., Choi, K., Auty, W., Hebbler, S., Blank, R., & Williams, A. (2005). *Policymakers’ guide to growth models for school accountability: How do accountability models differ?* Washington, DC: Council of Chief State School Officers.

Koehler, E., Brown, E., & Haneuse, S. J.-P. A. (2009). On the assessment of Monte Carlo error in simulation-based statistical analyses. *The American Statistician*, *63*, 155–162. doi:10.1198/tast.2009.0030

McCaffrey, D. F., Castellano, K. E., & Lockwood, J. R. (2015). The impact of measurement error on the accuracy of individual and aggregate SGP. *Educational Measurement: Issues and Practice*, *34*, 15–21. doi:10.1111/emip.12062

Monroe, S., & Cai, L. (2015). Examining the reliability of student growth percentiles using multidimensional IRT. *Educational Measurement: Issues and Practice*, *34*, 21–30. doi:10.1111/emip.12092