RUNNING HEAD: Error in SGPs

Estimating the Amount of Error in Student Growth Percentiles[1]

Craig S. Wells, Stephen G. Sireci, and Louise M. Bahry

University of Massachusetts Amherst

[1] *Center for Educational Assessment Research Report No. 869*: Amherst, MA: Center for Educational Assessment.

Abstract

Student growth percentiles (SGPs) are currently used by several states and school districts to provide information about individual students as well as to evaluate teachers, schools, and school districts. For SGPs to be defensible for these purposes, they should be reliable. In this study we examine the amount of systematic and random error in SGPs by simulating test scores for four grades and estimating SGPs using one, two, or three conditioning years. The results indicate that, although the amount of systematic error was small, the amount of random error was substantial. Furthermore, the magnitude of the random error was considerable regardless of the number of conditioning years. For example, for a true SGP of 56, the width of an estimated 68% confidence interval was 50 percentile points (ranging from 29 to 78) when using three conditioning years. The results suggest SGP estimates are too imprecise to be reported at the student level.

Estimating the Amount of Error in Student Growth Percentiles

Student growth percentiles (SGPs; Betebenner, 2009) were developed to provide a measure of normative "growth" for the purpose of determining educational effectiveness. SGPs are currently used by many states and school districts for purposes ranging from providing information about individual students to parents and teachers, to evaluating teachers, schools, and school districts. According to Collins and Amrein-Beardsley (2012), during the 2011/2012 school year, 13 states were using or piloting SGPs for the purpose of evaluating teachers. Soto (2013) stated that SGPs are being used in 22 states. Thus, their popularity appears to be growing.

Although SGPs are popular, their use has far outpaced research on their utility. As we view the current state of affairs, SGPs are being reported, interpreted, and acted on, without a clear understanding of their reliability or validity. This is a dire situation, and puts state departments of education in violation of the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association, & National Council on Measurement in Education, 2014), which state, "A rationale should be presented for each intended interpretation of test scores for a given use, together with a summary of the evidence and theory bearing on the intended interpretation" (p. 23). To date, there has not been sufficient evidence to support the reliability or validity of SGPs for interpreting students' educational progress. In these situations, the AERA et al. *Standards* suggest "If validity for some common or likely interpretation for a given use has not been evaluated…that fact should be made clear and potential users should be strongly cautioned about making unsupported interpretations" (p. 23). Unfortunately, that does not seem to be the case, as SPGs are being regularly reported without any cautions, and without estimates of their precision.

Such practice represents another violation of the *Standards* that requires testing agencies to inform users about the precision of the scores they report (p. 43).

SGPs represent a new use of test scores; however, there is no comprehensive summary of evidence to support them for the purposes for which they are being used. In fact, the few studies that have been done in this area have not provided encouraging results. Soto (2013) found that teacher classifications based on the median SGPs of their students were not consistent across different samples of students. Castellano and Ho (2013) found that using ordinary least squares to estimate the conditional percentile ranks were superior to SGPs based on quantile regression when the scaled scores followed a multivariate normal distribution. Furthermore, they concluded that SGPs (and related conditional measures) do not actually answer the question "how much did my child grow?" and pointed out that students could actually *decline* in performance across years, but still receive a high SGP (p. 212). Such contradictions in the reporting of students' test performance is likely to confuse parents and educators, thus having the opposite effect than intended on improving the educational process.

It is troubling that although much has been written about how to calculate and use SGPs (e.g., Betebenner, 2006, 2009, 2012), very little has been written about their reliability and validity. Although we provide a brief and conceptual description of how SGPs are computed (see below), our focus is on a critical aspect of their quality—reliability and bias.

For SPGs to be useful, the estimates must be unbiased and reliable; that is, the estimates must be accurate and contain a relatively small amount of random error. Unfortunately, there is a dearth of research on the statistical properties of SGP estimates. Sireci, Wells and Bahry (2013) examined the systematic and random error in SGP estimates across three growth rate conditions (i.e., variability in growth across simulees) when conditioning on one previous grade via a

simulation study. They found that the systematic error was small regardless of the amount of variability in growth. However, the amount of random error at the student level was substantial even for the largest growth rate condition. For example, a simulated student with an SGP of 56, had an estimated confidence interval of 67 percentile points in the low variability growth condition (i.e., the confidence interval ranged from 18 to 84!).

One limitation of the Sireci et al. (2013) study was that scores were conditioned on only one previous year of test data. That practice is realistic for reporting growth for students who have scores from only the previous year (e.g., most 4^{th}-grade students, or students new to a state), but for most students, two or more years of prior test scores are available to be used in the estimation. It is possible that conditioning on multiple years of test data will result in a smaller amount of random error. Therefore, the purpose of the present study was to examine the reliability and bias of the SGP estimates when conditioning on one, two, and three years of prior test scores. This extension of the Sireci et al. (2013) study better reflects common practice across several states.

Student Growth Percentiles (SGPs)

SGPs attempt to provide an indication of student growth relative to "academic peers" (Betebenner, 2012, p. 5). Betebenner and others have described this as "normative growth," because it represents change in a student's standing relative to a subgroup of other students. For example, suppose we wanted to describe the change in the relative standing of a student in Grade 5 to her relative standing in Grade 6, based on her math test performance each year. The first step would be to identify her academic peers based on her Grade 5 test scores. If the data set were very large (i.e., nearly infinite), then we could simply identify her academic peers based on students who had the same Grade 5 test score. The growth percentile would then be determined

by computing the Grade 6 percentiles for all students wither her grade 5 test score (i.e., her "academic peers"). If she received a percentile of 70, then she performed as well or better than 70% of the students who had the same score as her on the grade 5 test. Advocates of SGPs add the interpretation that she exhibited equal or greater "growth" than 70% of her peer group. Whether this change in conditional status can be called growth is a matter of debate, which is outside the scope of this paper.

The previous description of SGPs provides a theoretical overview, but in practice, there is no empirical group of students selected as academic peers because such selection would be impractical for finite the sample sizes involved in most statewide testing programs, especially when conditioning on multiple years of test scores. Instead, quantile regression is used (Betebenner, 2008, 2011, 2012). Quantile regression does not assume a linear relationship between test scores across years or homoscedasticity, which makes it a more general procedure than linear regression. Linear regression typically estimates the mean criterion score for each level of the predictor. Quantile regression fits a curvi-linear relationship to estimate each percentile of the criterion score (rather than the mean) for each level of the predictor. Thus, instead of a single regression being conducted as in linear regression (to estimate the mean), 99 nonlinear regressions are conducted to estimate the 1st percentile, 2nd percentile, and so forth, for each level of the predictor or set of predictors (prior year's test score, or multiple years' test scores).

Each of the 99 quantile regressions represents a predicted percentile conditional on prior test scores. A student is assigned a percentile score based on the distance of her/his observed data point to the regression lines; that is, s/he is assigned the percentile corresponding to SGP line closest to her/his data point. Thus, there is not really a group of "academic peers" to which a

student is being compared. See Betebenner (2006, 2009) and Castellano and Ho (2013) for a more detailed description of how quantile regression is used to compute SGPs.

Precision Estimates for SGPs

Betebenner (2013) developed a simulation method for estimating the precision of student-level SGPs based on the impact of measurement error. The conditional standard error of measurement (CSEM) is used to add random error to the observed scores for the prior and current administration scores. In other words, each examinee receives a perturbed observed-score by sampling from a normal distribution with the mean equal to the examinee's observed score and the standard deviation defined by the value from the CSEM given the examinee's observed score (see Equation (1)).

$$X_{j_{\text{perturbed}}} \quad N\left(X_j, CSEM_j\right). \tag{1}$$

$X_{j_{\text{perturbed}}}$ represents the simulated score for examinee *j* that includes the effect due to measurement error; X_j represents the observed-score for examinee *j*; and $CSEM_j$ represents the conditional standard error of measurement for examinee *j*. 100 perturbed scores and their corresponding SGPs are generated for each examinee. The 100 SGPs are used to derive the standard error estimates associated with each examinee's SGP score.

Betebenner (2013) derived standard errors using test data from the Massachusetts Comprehensive Assessment System (MCAS) for the English Language Arts (ELA) and Mathematics exam, grades 4 through 8 and 10. The standard errors of the student-level SGPs varied across scale scores and the CSEMs. Overall, the standard errors were less than 10 with a median near 6 across several grades and subjects. One potential flaw of how the standard errors of the SGPs are computed using Betebenner's method is that it treats the observed scores as true scores when adding random error based on the CSEM. The reason that is important is, as Sireci

et al. (2013) observed, the variability in growth rates plays an important role in the amount of random error in SGPs. Therefore, adding random error to observed scores that already are influenced by random error increases the variability in growth rates which would lead to smaller estimates of random error (i.e., underestimated standard errors). One of the purposes of this study, therefore, is to determine if the simulation method provided by Betebenner (2013) provides accurate precision estimates in SGPs.

Method

Data Simulation

We conducted a simulation study to examine random and systematic error in SGP estimates when varying the number of conditioning years. Scale scores were simulated to represent students' test scores on a typical statewide assessment for four consecutive grades, hereafter referred to as Grades 3, 4, 5, and 6. SGPs were computed for simulees in Grade 6 using either Grade 5, Grades 4 and 5, or Grades 3, 4, and 5 as conditioning years. To ensure the simulation was realistic, the parameters in the simulation were based on real test data obtained from the Massachusetts Comprehensive Assessment System (MCAS) for English Language Arts (ELA) for grades 3 through 6 (see relevant details below).

Data Generation

Generating true scale scores and SGP values. Scale scores were sampled from a multivariate-normal distribution to represent true scale scores $\left(\text{denoted } \boldsymbol{\tau}\right)$ for 100,000 simulees from Grades 3 through 6 $\left(\text{i.e., } \boldsymbol{\tau} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})\right)$. The mean of the true scores in each grade was 240 $\left(\text{i.e., } \boldsymbol{\mu} = \left[240, 240, 240, 240\right]\right)$ and the variance-covariance structure $\left(\boldsymbol{\Sigma}\right)$ of the multivariate-normal distribution for Grades 3, 4, 5, and 6 was defined as follows:

$$\boldsymbol{\Sigma} = \begin{bmatrix} 225.000 & & & \\ 180.450 & 225.000 & & \\ 184.275 & 193.275 & 225.00 & \\ 181.350 & 189.000 & 203.175 & 225.000 \end{bmatrix}. \quad (2)$$

The diagonal of the variance-covariance matrix represents the variance of the scale scores and is similar to the variance observed in MCAS. To more easily understand the covariance structure in the simulated data, and because the covariance structure may play an important role in the amount of random error in SGP estimates, the standardized covariance matrix (i.e., correlation matrix denoted **R**) is provided below:

$$\mathbf{R} = \begin{bmatrix} 1 & & & \\ .802 & 1 & & \\ .819 & .859 & 1 & \\ .806 & .840 & .903 & 1 \end{bmatrix}. \quad (3)$$

The correlational structure of the multivariate-normal distribution was based on disattenuated correlation coefficients from real student data obtained from the 2010-2013 MCAS ELA (Massachusetts Department of Elementary and Secondary Education, 2013), grades 3 through 6, and thus, represents a realistic correlational structure among true scores in a large-scale assessment. The sampled true scores for each grade were rounded to the nearest integer and constrained to range from 200 to 280, which are the upper and lower limits of the MCAS score scale.

The true SGPs were determined via quantile regression using the true scale scores. For example, the true SGPs when conditioning on only Grade 5 were obtained by computing the percentile rank for the Grade 6 scores when conditioning on Grade 5 true scores. The R package *SGP* (Betebenner, VanIwaarden, & Domingue, 2013) was used to implement quantile regression.

True SGPs were obtained when conditioning on one previous year (Grade 5), two previous years (Grades 4 and 5), and three previous years (Grades 3, 4, and 5).

Generating observed scores. Observed scale scores for each grade were created to represent a typical amount of measurement error in scale scores for a large-scale statewide assessment. The observed scores (X_j) were sampled from a normal distribution with the mean equal to the simulee's true score and the standard deviation equal to the standard error of measurement conditioned on the true scale score (CSEM); that is,

$$\begin{aligned} X_{j,\text{Grade 6}} &\sim N\left(\tau_{j,\text{Grade 6}}, \text{CSEM}_j\right) \\ X_{j,\text{Grade 5}} &\sim N\left(\tau_{j,\text{Grade 5}}, \text{CSEM}_j\right) \\ X_{j,\text{Grade 4}} &\sim N\left(\tau_{j,\text{Grade 4}}, \text{CSEM}_j\right) \\ X_{j,\text{Grade 3}} &\sim N\left(\tau_{j,\text{Grade 3}}, \text{CSEM}_j\right) \end{aligned} . \quad (4)$$

The CSEM was based on the 2011, Grade 5 MCAS ELA test (Massachusetts Department of Elementary and Secondary Education, 2011) and is comparable to the CSEMs across other grades. Figure 1 displays the CSEM across the values on the score scale ranging from 200 to 280. The observed scale scores were rounded to the nearest integer and constrained to range from 200 to 280 for each grade. One-hundred data sets containing 100,000 observed scale scores for four grades were created.

SGP Estimation

Once the observed scores were generated, the R package *SGP* (Betebenner et al., 2013) was used to estimate the SGPs for Grade 6 for each data set using either Grade 5, Grades 4 and 5, or Grades 4, 5, and 6 as the conditioning variables. When estimating SGPs, the same Grade 5 scores were used when conditioning on Grade 5, Grades 4 and 5, or Grades 3, 4, and 5. In addition, the same Grade 4 scores were used when conditioning on Grades 4 and 5 as well as conditioning on Grades 3, 4, and 5.

Data Analysis

We examined the relationship between the true and estimated SGPs, the 68% and 95% confidence intervals, and the classification accuracy across the conditions that varied with respect to the number of conditioning years. The Spearman rho correlation coefficient was computed for each replication to measure the strength of the relationship between the true and estimated SGPs. It was hypothesized that the correlation would increase as the number of conditioning years increased.

To examine the amount of random error present in the SGP estimates, empirical 68% and 95% confidence intervals (CI) were constructed for each simulee. These two intervals represent reasonable intervals testing agencies may choose for reporting confidence intervals for test scores. For the 95% CIs, the lower bound of the confidence interval for each simulee was based on the SGP estimate associated with the 2.5^{th} percentile across replications and the upper bound was based on the SGP estimate associated with the 97.5^{th} percentile. For the 68% CIs, the lower bound of the confidence interval for each simulee was based on the SGP estimate associated with the 16^{th} percentile across replications and the upper bound was based on the SGP estimate associated with the 84^{th} percentile.

We were particularly interested in examining the relationship between the lower and upper bounds of the confidence intervals relative to the true SGP values to determine if the amount of random error was a function of true SGPs. However, because simulees with the same true SGP had different lower and upper bound estimates due to measurement error, we determined the functional relationship between the lower and upper bound relative to the true SGP value using the Nadaraya–Watson kernel regression estimates. The *ksmooth* package in the computer program R was used to compute smoothed lower and upper bounds as a function of the

true SGP value. Moreover, we examined the systematic error by comparing the median[2] SGP estimates for each simulee to their respective true SGP value. It was hypothesized that the systematic and random error would be smaller when conditioning on more years.

It is important to note, that in addition to reporting students' SGPs on score reports that go home to parents, some states, such as Massachusetts, also classify students into SGP categories such as "lower growth" and "higher growth." To evaluate the consistency of such classifications, we used the SGP estimates to classify simulees into one of three growth level descriptors: low, average, and high growth. Similar to the classifications on the MCAS score reports, simulees with SGP estimates less than 40 were classified as "low growth," between 40 and 60 as "average growth," and above 60 as "high growth." The estimated growth classifications were compared to the true growth classifications for each replication to determine the proportion of simulees classified correctly or misclassified. It was hypothesized that the classification rate would increase as the number of conditioning years increased.

Results

Correlation Between SGP Estimates and True Values

The median Spearman rho correlation coefficient for the three conditioning year conditions (1 year, 2 years, and 3 years) were 0.69, 0.70, and 0.71, respectively. Within each of the three conditions, the Spearman rho correlation was nearly identical when rounded to the second place across all 100 replications, primarily because the sample size was large (i.e., N=100,000). Although the correlation coefficients were larger when using more conditioning years, the differences were very small.

[2] The median was used because the distribution of SGPs conditioned on true SGP value was skewed. For example, for large true SGP values, the distribution of SGP estimates was negatively skewed.

68% and 95% Confidence Intervals

Figures 2(a) to (c) report the 68% and 95% CIs as a function of the true SGP value across the three conditions. The dashed line represents the median SGP estimate conditioned on the true SGP value. The solid line is the identity line and represents the expected value of each SGP estimate. The dotted line represents the smoothed empirical 68% CI and the dashed-dotted line represents the smoothed empirical 95% CI.

The difference between the median SGP estimates (represented by the dashed line) and the identity line indicate the amount of systematic error present for each of the true SGP values. The systematic error was larger for extreme true SGPs with over-estimated values for small true SGPs and under-estimated values for large true SGPs. The amount of systematic error was comparable across the three conditions. Therefore, it appears that the number of conditioning years had a minimal effect on the amount of systematic error in the SGP estimates.

The width of the CIs indicates the magnitude of random error present for a specific true SGP. The 68% and 95% CIs were wide for much of the distribution, particularly for true SGP values near the 50th percentile. The 68% and 95% CIs were comparable across the three conditions. Therefore, it appears that the number of conditioning years had a minimal effect on the CIs.

To further understand the amount of random error relative to the true SGP values, Figure 3 shows the relationship between the true SGP values and the width of the confidence interval for each of the three conditions. It is apparent that the width of the CIs was slightly smaller when conditioning on more years. However, and more importantly, there was a considerable amount of random error in the SGP estimates, especially near the 50th percentile. For example, for a true SGP value of 56, the width of the 68% CI was 50 percentile points (the CI ranged from 29 to 78)

when using three conditioning years. The width of the 95% CI for a true SGP value of 56 was 81 percentile points (the CI ranged from 10 to 90) when using three conditioning years. Although the width of the CIs was smaller for true SGPs at the extremes, the width was still substantial. For example, the width of the 68% CI for a true SGP value of 10 was 33 (the CI ranged from 5 to 37) when using three conditioning years.

Classification Accuracy

Table 1 reports the average classification rates for the growth levels across the three conditions. The diagonal elements of each 3 x 3 table represent the correct classification rate whereas the off-diagonal elements indicate the erroneous classification rates. For example, when using one conditioning year, the average proportion in which the true SGP values were in Level 1 (i.e., between 1 and 39) and the SGP estimates were in Level 1 was .28. However, an average of 7% of the simulees who had true SGP values in Level 1, had SGP estimates in Level 2. Consistent with the previous results, the number of conditioning years had a minimal impact on the classification rates.

Although Table 1 provides useful aggregated information, it ignores the classification rates as a function of the true SGP values. To address this issue, Figure 4 portrays the relationship between the true SGP values and the erroneous classification rates across the three conditions. The misclassification rates were small for simulees at the extremes of the true SGP scale. However, for simulees with true SGPs near the middle of the scale, the misclassification rates were large. For example, simulees with a true SGP of 38 and three conditioning years were erroneously classified on average in 53% of the replications.

Discussion

In the present study we examined the amount of systematic and random error in SGP estimates as a function of the number of conditioning years via a simulation study. The conditions and parameter values of the simulation study were based on real data so that the results would be generalizable and directly inform practitioners of large-scale statewide assessments. The findings are troubling. The number of conditioning years had a minimal impact on improving the amount of systematic and random error in SGP estimates. Although the amount of systematic error was relatively small, the amount of random error was substantial, which raises serious questions about the validity of SGPs at the student level. The large amount of random error in SGPs impedes inferences about student "growth" relative to her/his academic peers. For example, for students with true SGPs of around 50, the SGP confidence interval based on just one SEM (i.e., 68% confidence interval) would range from 29 to 78, which includes all three "growth" classifications that students are currently classified into in Massachusetts. Bear in mind that hundreds of thousands of parents of school children in Massachusetts read score reports that classified their children as "lower" or "higher" growth. The results of the current study call such classifications of students into question.

The results suggest that reporting SGPs at the student level is likely to inhibit accurate interpretation of student progress. Many of the simulees in our study had "growth" classification consistency rates less than 50%. In contrast, the classification consistency rates for the achievement level classifications on all MCAS tests in 2014 ranged from .65 to .79. It is unclear why SGPs, which represent derivatives of test scores, have been reported without properly understanding, or estimating, their reliability or the consistency of the student classifications they provide.

It is important to note the amount of random error observed in this study is inconsistent with the estimates provided in Betebenner (2013), who estimated standard errors for SGP estimates using the CSEM in conjunction with the student scale scores. He concluded the standard errors were less than 10 with a median near 6 across multiple grades and subjects. The widths in the CIs observed in this study are much larger than would be predicted based on Betebenner (2013) and would correspond to standard errors that ranged from 10 to 20 percentile points.

Although there could be a problem in Betebenner's (2013) code for computing the standard errors of SGPs as a reviewer of this paper suggested, another possible reason for the discrepancy between our results and his is that Betebenner's method does not accurately represent the true growth rate which is primarily dictated by the correlational structure of the true scores. The correlation between true scores dictates the variability in growth, which plays an important role in the amount of random error in SGP estimates. The growth for students will be more uniform or homogeneous as the correlation between true scores increases. For example, when the correlation between the true scores approaches 1, the students will exhibit the same magnitude of growth. In such a case, the SGP estimates will be assigned randomly since the variability in observed scores for an academic peer group is determined by measurement error. The correlational structure used in this study to generate true scale scores was based on disattenuated correlation coefficients from real data from MCAS and contained large correlation coefficients, which can explain the large amount of random error present. Therefore, treating observed scores as true scores and using the CSEM to create subsequently perturbed observed scores (i.e., Betebenner's 2013 method) *underestimates* the actual amount of random error

present, and does not represent a realistic estimate of the amount of error in SGPs. A realistic estimate, as described in the present study, suggests the margin of error is too wide on SGPs to support their use in a typical statewide assessment program.

The current study represents only one investigation of the reliability of SGPs and looked only at the level of individual students. SGPs are aggregated at classroom and other levels and so the reliability of these aggregations should also be studied. Presently, many states and school districts are using mean and median SGPs at the classroom and school levels for teacher evaluation and school improvement planning. Clearly, research on the reliability of these aggregate measures also needs to be conducted.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, D.C.: American Educational Research Association.

Betebenner, D.W. (2006). *Growth, standards, and accountability*. Unpublished manuscript available at http://www.nciea.org/publications/growthandStandard_DB09.pdf.

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. *GJ Cizek, Setting Performance Standards: Foundations, Methods & Innovations*, 439-450.

Betebenner, D. W. (2013). *An analysis of the precision associate with MCAS SGPs*.

Betebenner, D. W., VanIwaarden, A. & Domingue, B. (2013). SGP: An R Package for the Calculation and Visualization of Student Growth Percentiles & Percentile Growth Trajectories. (R package version 1.0-3.0. URL http://schoolview.github.com/SGP/).

Castellano, K. E., & Ho, A. D. (2013). Contrasting OLS and quantile regression approaches to student "growth" percentiles. *Journal of Educational and Behavioral Statistics, 38*, 190-215.

Collins, C. & Amrein-Beardsley, A. (2012, April). *Putting growth and value-added models on the map: A national overview*. Paper presented at the 2012 annual meeting of the American Educational Research Association, Vancouver, British Columbia, Canada.

Massachusetts Department of Elementary and Secondary Education (2011). *2011 MCAS and MCAS-Alt technical report*. Retrieved from: http://www.doe.mass.edu/mcas/tech/.

Massachusetts Department of Elementary and Secondary Education (2013). Massachusetts Comprehensive Assessment System - Student-Level Files. [data file and codebook]. Retrieved from: http://www.doe.mass.edu/infoservices/research/.

Sireci, S. G., Wells, C. S., & Bahry, L. (2013, April). *Student growth percentiles: More noise than signal?* Paper presented at the 2013 annual meeting of the American Educational Research Association, San Franciso, CA.

Soto, A. (2013). *Measuring teacher effectiveness using students' test scores*. Unpublished Dissertation, University of Massachusetts Amherst.

Table 1. Average SGP Classification Rates.

		Growth Levels Based on SGP Estimates		
Number of Condition Years	Growth Levels Based on True SGPs	Level 1	Level 2	Level 3
One Conditioning Year	Level 1	.28	.07	.05
	Level 2	.07	.06	.08
	Level 3	.05	.07	.27
		Level 1	Level 2	Level 3
Two Conditioning Years	Level 1	.28	.07	.04
	Level 2	.07	.06	.07
	Level 3	.04	.07	.28
		Level 1	Level 2	Level 3
Three Conditioning Years	Level 1	.28	.07	.04
	Level 2	.07	.06	.07
	Level 3	.04	.07	.28

Note: Level 1 corresponds to "lower growth," (SGP < 40) and Level 3 corresponds to "higher growth" (SGP > 60) as reported by at least one statewide testing program.

Figure 1. Conditional standard error of measurement (CSEM) used to simulate observed scale scores.

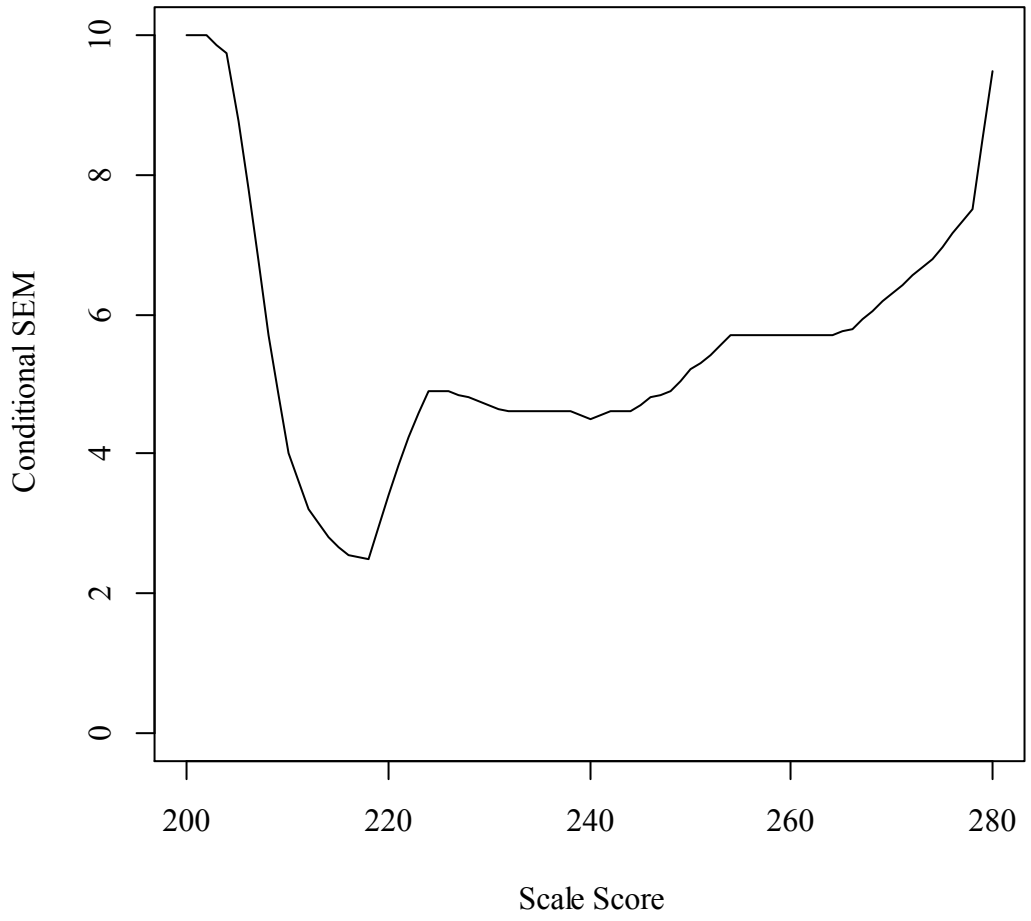

Figure 2. 68% and 95% CIs when conditioning on one, two, or three years.

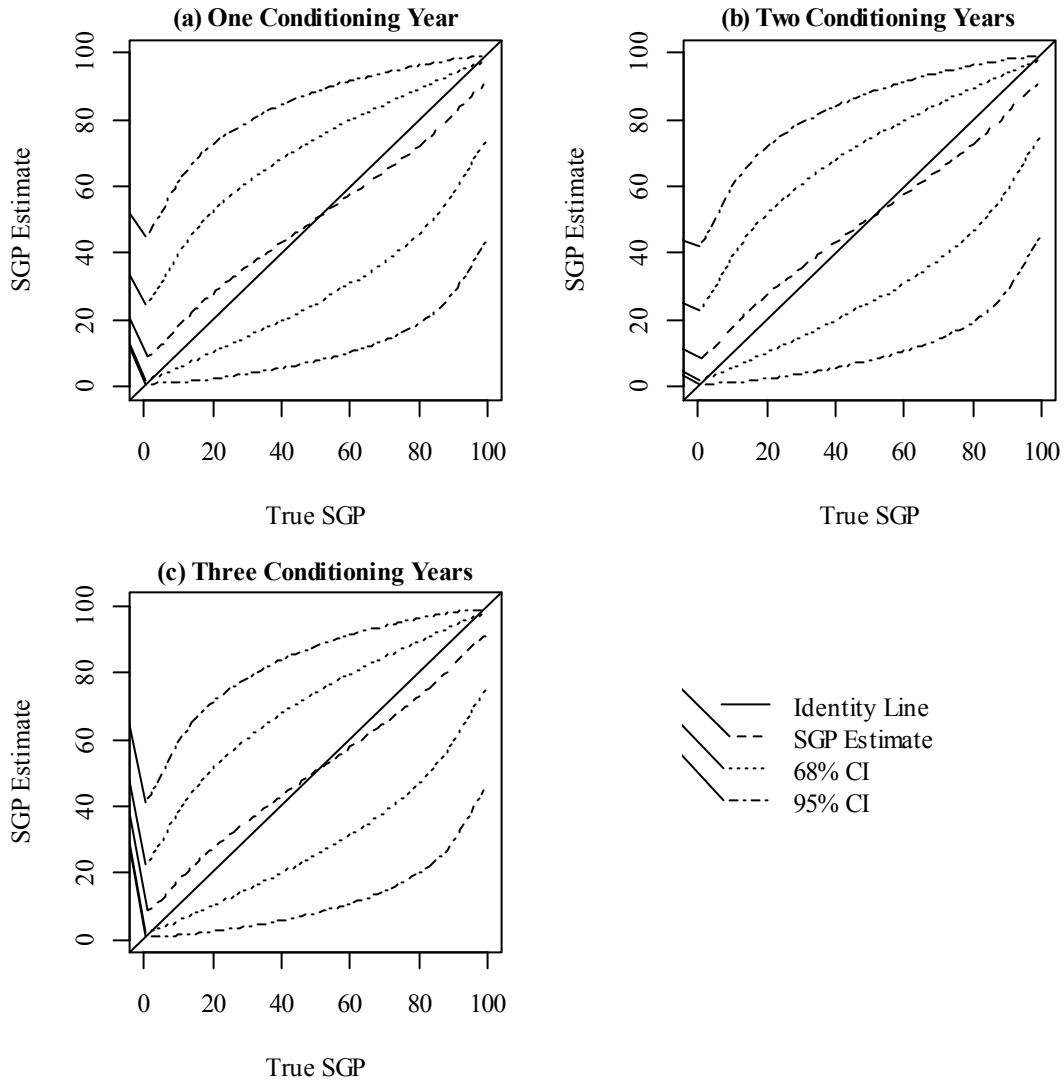

Figure 3. Comparison of the 68% and 95% CIs width across the three conditions.

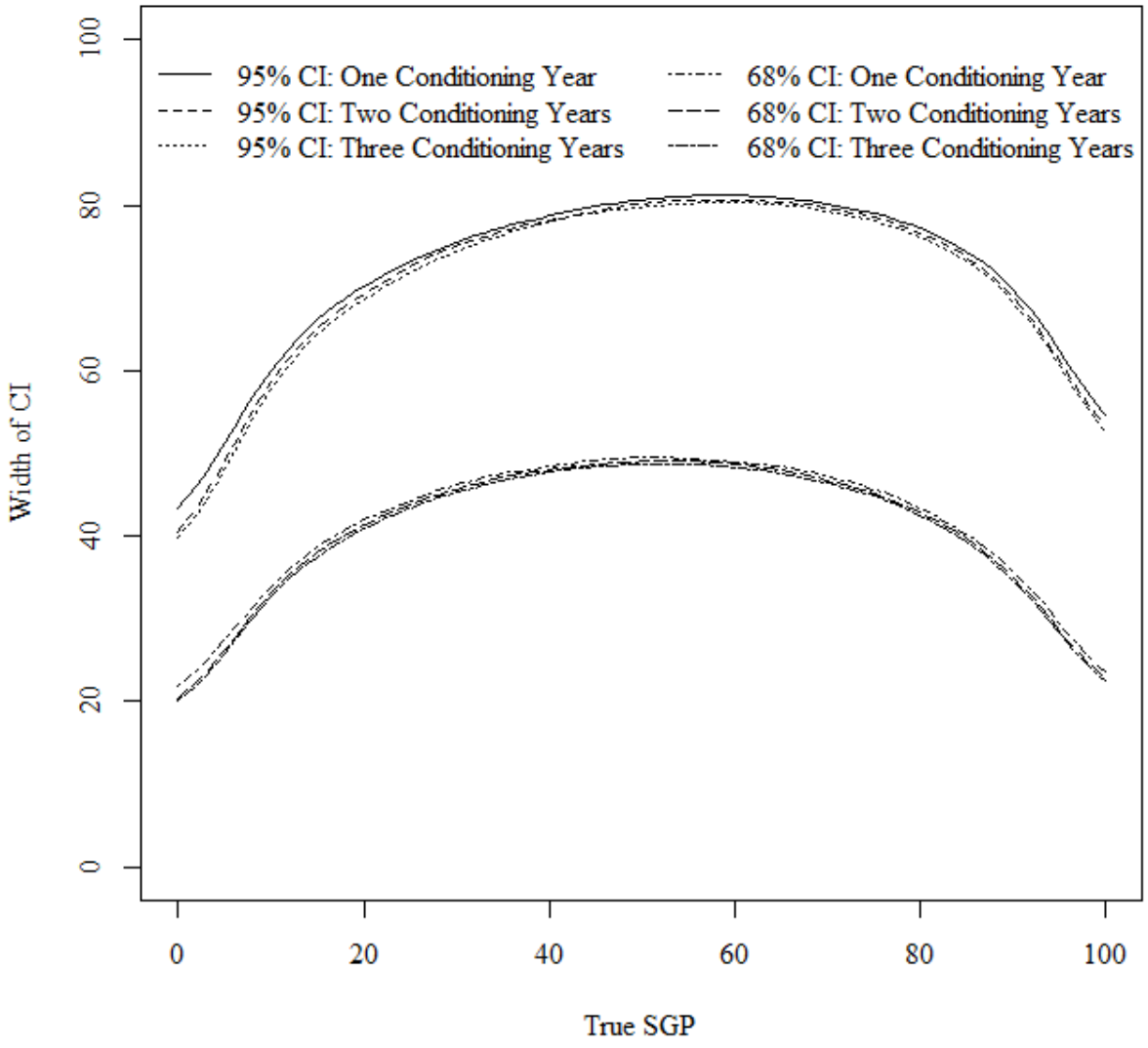

Figure 4. Misclassification rate relative to the true SGP values.