RUNNING HEAD: Reliability of SGPs

The Effect of Conditioning Years on the Reliability of SGPs[1]

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

University of Massachusetts Amherst

Paper presented at the annual meeting of the National Council of Measurement in Education, April 3-7, 2014, Philadelphia, PA.

[1] *Center for Educational Assessment Research Report Number 869*. Amherst, MA: Center for Educational Assessment, University of Massachusetts Amherst.

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 (Collins and Amrein-Beardsley, 2012; Soto, 2013). For SGPs to be useful, the estimates must have a minimal amount of systematic and random error. However, previous research has raised questions about the reliability of SGPs when computing SGPs based on one conditioning year (Sireci, Wells, & Bahry, 2013). The purpose of the present study was to examine the effect of the number of conditioning years on the reliability of SGPs to determine at what point SGPs may be considered reliable. A simulation study was conducted in which test scores for four grades were generated. SGPs were estimated using one, two, or three conditioning years. The results indicate that the number of conditioning years had a minimal impact on the reliability of SGP estimates. Furthermore, the amount of random error was substantial, bringing into question whether SGP estimates are appropriate at the student level.

The Effect of Conditioning Years on the Reliability of SGPs

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.

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 when conditioning on one previous grade via a simulation study. They found that although the SGP estimates exhibited small systematic error, the amount of random error at the student level was substantial. One of the limitations of the Sireci et al. study was that they 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 4th-grade students, or students new to a state), but for most students, two or more years of prior test scores are used. 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 of the SGP estimates when conditioning on one, two, and three years of test data.

Student Growth Percentiles (SGPs)

SGPs provide a normative indication of student growth by quantifying a student's current level of achievement relative to her/his academic peers as defined by previous years' test scores. To elaborate, suppose we wanted to determine the normative growth from Grade 5 to Grade 6 in math for a particular student. The first step would be to identify the student's academic peers based on her/his Grade 5 test scores. If the data set were very large (i.e., nearly infinite), then we could simply identify the student's academic peers based on students who had the same Grade 5 test score. The growth percentile would then be determined by the percentile based on the student's Grade 6 score relative to the scores in her/his academic peer group (i.e., those who had the same Grade 5 test score). Therefore, the percentile would represent how well the student performed relative to a group of students who performed similarly in the past. For example, if the student received a percentile of 70, then s/he performed as well or better than 70% of her/his academic peers who had a similar prior score history, which would indicate equal or greater "growth" than 70% of the relevant peer group.

Because the previously described procedure is impractical and inaccurate for data sets that are not infinite (or when defining academic peers based on multiple years of test scores), a statistical model based on regression is used to estimate SGPs (Betebenner, 2008, 2011). Quantile regression is an attractive model for this purpose because it does not assume a linear relationship or homoscedasticity. Quantile regression fits a curvi-linear relationship between the students' test scores in the current administration and the previous year's test scores for each percentile (i.e., 99 regression lines). Each curvi-linear relationship (which can be expressed as a regression line when using one conditioning year) represents the predicted percentile. 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 line closest to her/his data point.

Method

Data Simulation

A simulation study was conducted 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 that 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 $\left(\text{denoted } \tau\right)$ 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:

$$\boldsymbol{\Sigma} = \begin{bmatrix} 225.000 & & & \\ 180.450 & 225.000 & & \\ 184.275 & 193.275 & 225.00 & \\ 181.350 & 189.000 & 203.175 & 225.000 \end{bmatrix}. \tag{1}$$

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:

$$\mathbf{R} = \begin{bmatrix} 1 & & & \\ .802 & 1 & & \\ .819 & .859 & 1 & \\ .806 & .840 & .903 & 1 \end{bmatrix}. \tag{2}$$

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,

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

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. For the 95% CIs, the lower bound of the confidence interval for each simulee was based on the SGP estimate associated with the 2.5th percentile across replications and the upper bound was based on the SGP estimate associated with the 97.5th percentile. For the 68% CIs, the lower bound of the confidence interval for each simulee was based on the SGP estimate associated with the 16th percentile across replications and the upper bound was based on the SGP estimate associated with the 84th 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. 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 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.

The SGP estimates were used to classify simulees into one of three growth level descriptors: low, average, and high growth. Similar to the classifications on the statewide assessments such as 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. Although the correlation coefficients were larger when using more conditioning years, the differences were very small.

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 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 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 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. 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. 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. 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

Due to the increased use of SGPs, it is prudent to understand the conditions in which SGP estimates are useful, or conversely, contain so much error that they hinder interpretation. The present study 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 indicate that the number of conditioning years had a minimal impact on the amount of systematic and random error in SGP estimates. And, although the amount of systematic error was relatively small, the amount of random error was substantial, which raises questions about the appropriate use of SGPs at the student level. In other words, the large amount of random error in SGP estimates impedes inferences about student “growth” relative to her/his academic peers.

It is interesting to note that the amount of random error observed in this study is inconsistent with precision estimates provided in Betebenner (2013). Betebenner (2013) developed a method for determining the standard errors for SGP estimates using the CSEM in conjunction with the student scale scores. He observed that 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.

One possible reason for the inconsistent results is that Betebenner's method does not appear to take into account the correlational structure of the scale scores. The disattenuated correlation between test 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 test scores increases. For example, when the disattenuated 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 any difference in growth is determined by measurement error. The correlational structure used in this study to generate true scale scores was based on real data from MCAS and contained large correlation coefficients which can explain the large amount of random error present. Therefore, using only the CSEM in determining the precision estimates in SGPs may underestimate the actual amount of random error present.

The current study examined the amount of systematic and random error at the student level. It would be interesting to investigate the error in SGP estimates at aggregated levels such as at the teacher, school and district level. In such cases, the amount of random error may be more acceptable, potentially making the use of SGPs appropriate in those situations. However, the results of the current study argue against reporting SGPs at the student level, due to the amount of error in their estimation.

References

Betebenner, D. (2008). Norm- and criterion-referenced student growth. *Educational Measurement: Issues and Practice, 28(4)*, 42-51.

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/).

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.

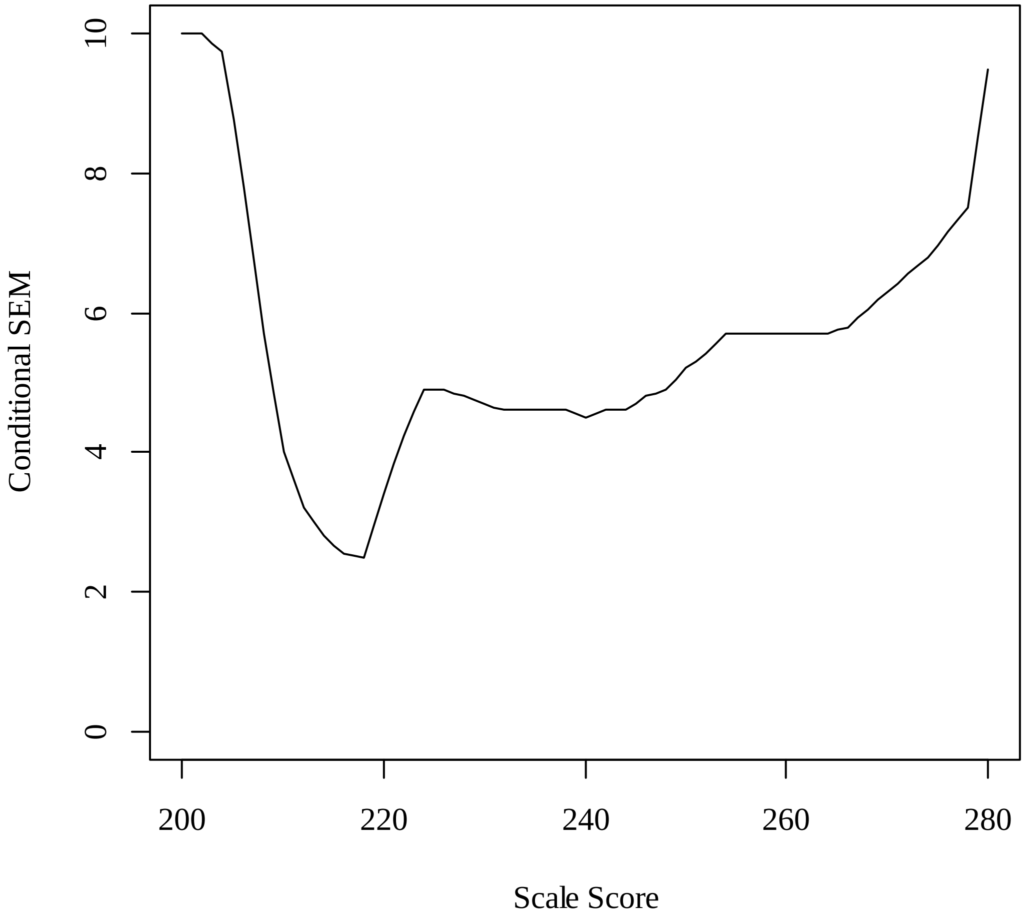


Figure 2. 68% and 95% CIs when conditioning on one, two, or three years.

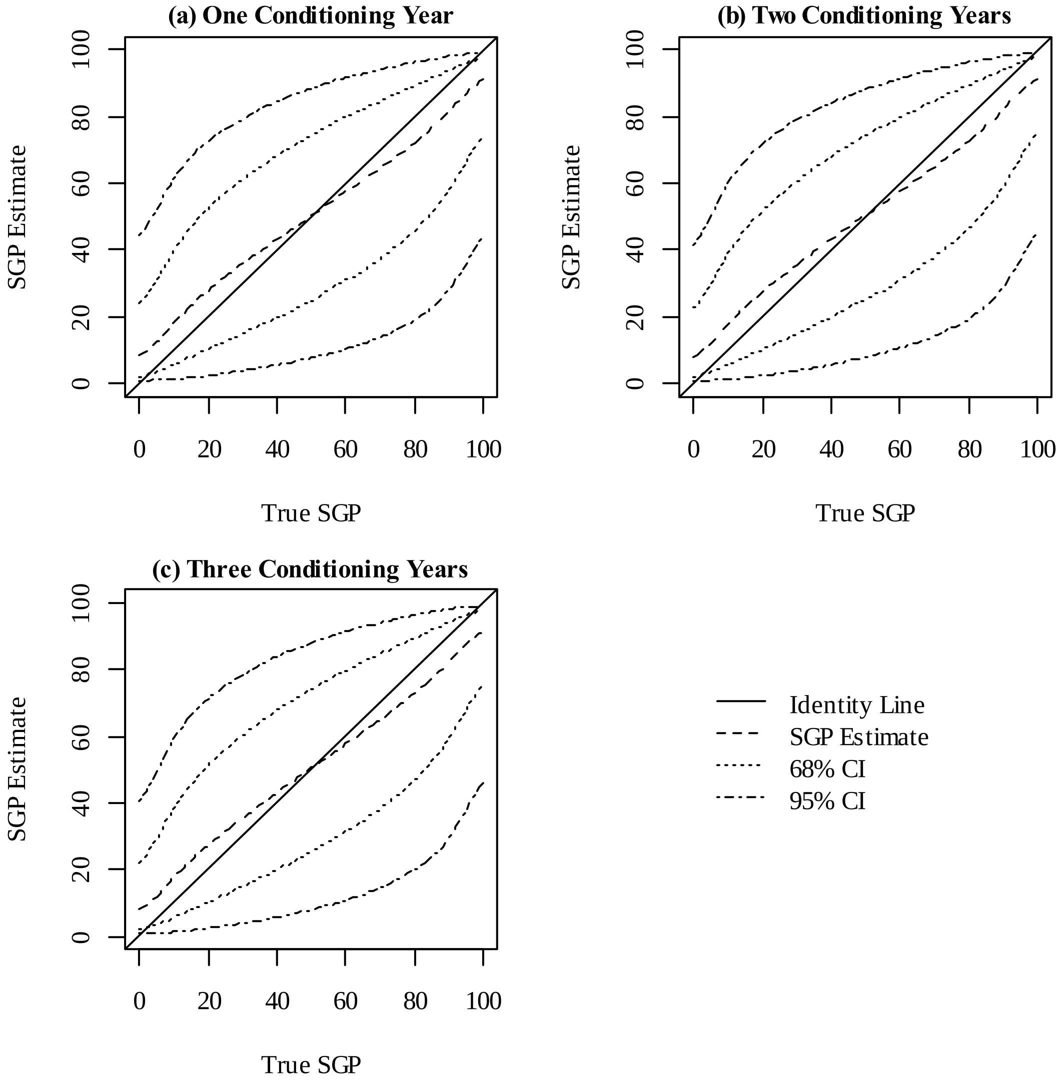


Figure 3. Comparison of the 68% and 95% CIs width across the three conditions.

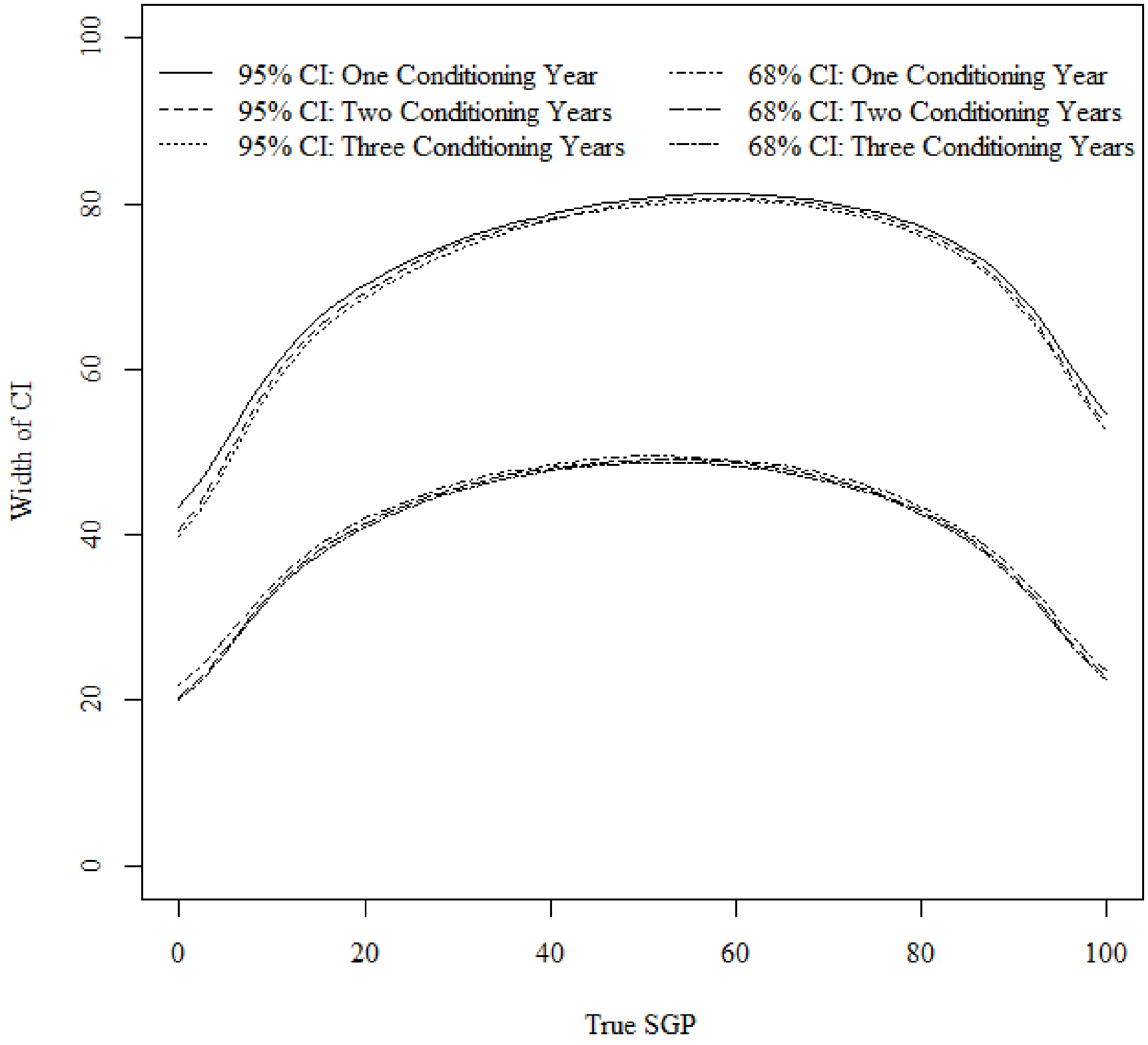
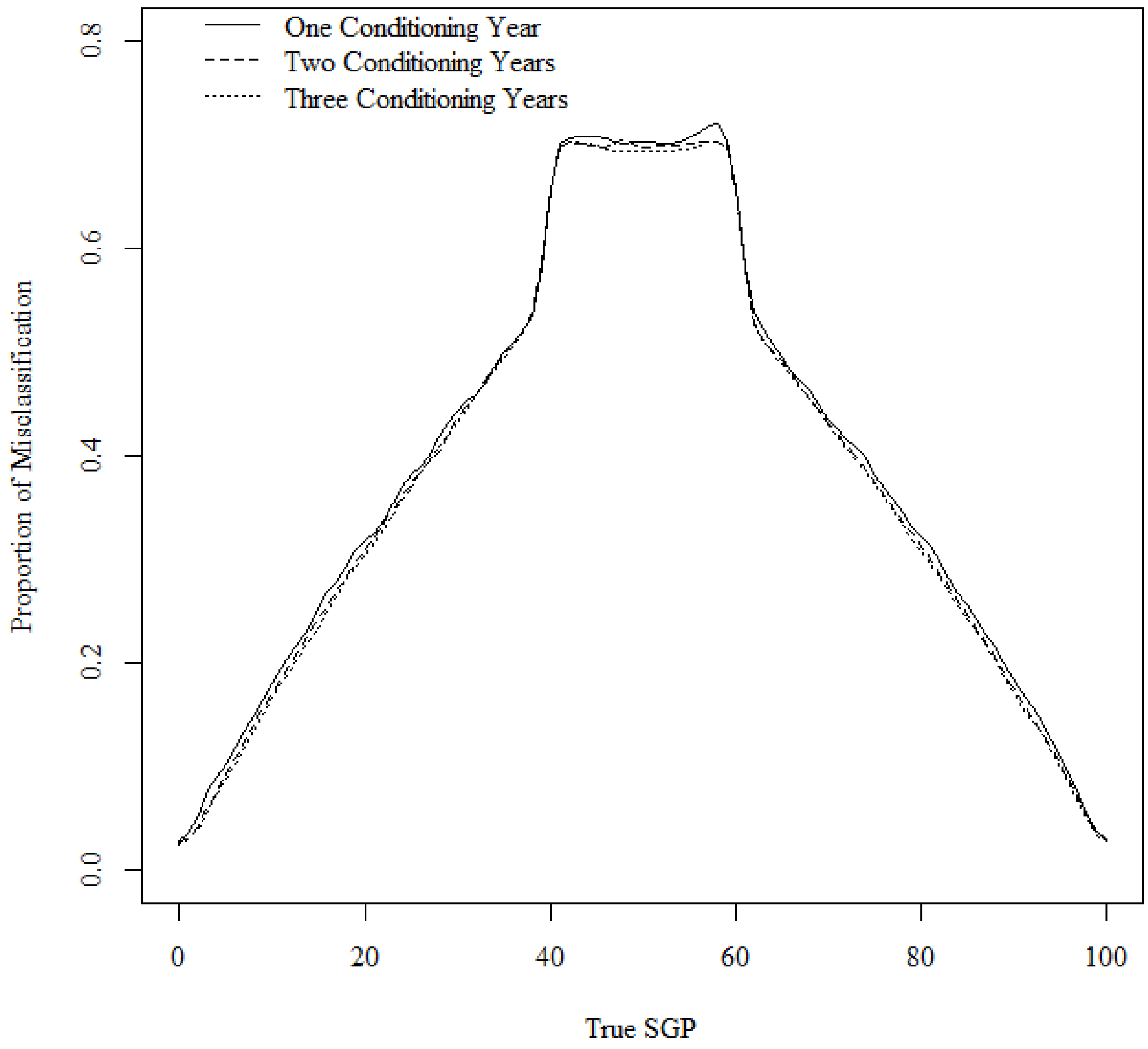


Figure 4. Misclassification rate relative to the true SGP values.


One Conditioning Year
Two Conditioning Years
Three Conditioning Years
Proportion of Misclassification
True SGP