Investigating the Amount of Systematic and Random Error in Mean Classroom-Level SGPs

Joshua J. Marland
Craig S. Wells
Stephen G. Sireci
Katherine Furgol Castellano
Abstract

Aggregate student growth percentiles (SGPs) are increasingly being used for educator and institutional accountability throughout the country, and research on their statistical properties is necessary to ensure appropriate use. In this study, true and observed SGPs were simulated, and the amount of systematic and random error was estimated to determine if aggregated SGPs can support their intended purposes across classrooms of differences sizes. Overall, the amount of systematic error was relatively small, while random error was substantially larger across all classroom sizes. Bias is more a function of true SGP, where random error is a function of both classroom size and true SGP. Classification was affected to a moderate extent across all classroom sizes because of the amount of error in the aggregate SGPs, but was most impacted for those in the top and bottom rating categories. Considerations should be made for random error when considering classifying teachers into rating categories, especially for those with small classrooms and who are close to the rating category cuts.
Investigating the Amount of Systematic and Random Error in Classroom-Level SGPs

Introduction

Student growth percentiles (SGPs; Betebenner, 2009) were initially developed to provide students with a “normative” measure about their growth, but have since expanded into being used for evaluating teachers, schools, and school districts for several federal and state accountability initiatives. 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 SGPs are being used in 22 states for various purposes.

Although SGPs are used by several states, the amount of random error present in student-level SGPs has been disconcerting (Sireci, Wells, & Bahry, 2013; Wells, Sireci, & Bahry, 2014). For example, Wells, Sireci, and Bahry (2014) examined the systematic and random error of student-level SGPs when conditioning on one, two, and three years of test data via a simulation study. They found that although SGPs exhibited small systematic error, the amount of random error was substantial (e.g., confidence intervals for students with SGPs around 50 ranged from 29 to 78), which calls into question their utility for interpreting students’ normative growth.

SGPs are also aggregated across students, for example, within a classroom for the purpose of evaluating teacher effectiveness, or across all students in a school for institutional accountability. As of 2012, the weight aggregate SGPs carried in educator evaluations ranged from 20 to 50 percent (Hull, 2013), which makes it important that the aggregated SGPs exhibit reasonably small systematic and random error to support inferences drawn regarding educator or institutional effectiveness. Although the amount of random error is expected to be smaller when

---

1 It is important to note that having a small amount of random and systematic error is not a sufficient condition to support valid inferences regarding teacher effectiveness. Additional evidence would need to be gathered as part of the validity argument supporting the valid use of SGPs for evaluating teacher effectiveness.
aggregating SGPs within a classroom, it is unclear if the amount of error is sufficiently small to support valid inferences regarding teacher effectiveness. Shang, VanIwaarden, and Betebenner (2015) found there to be greater random error in aggregate SGPs than bias, but both were non-negligible with and without a measurement error correction. The authors reported a squared bias of 10.41, variance of 12.97 and mean square error of 23.38 when calculating SGPs without a measurement error correction, and aggregating using a mean approach. McCaffrey, Castellano and Lockwood (2015) found similar results to Shang, et al, with greater random error in aggregate SGPs than bias with no substantive changes to results across estimation methods.

The purpose of the current study is to quantify the amount of random and systematic error exist in classroom SGPs, partition that error, and understand the implications of error on classifications of those being evaluated using aggregate SGPs. To accomplish these goals we simulate true SGPs at the student level and aggregate them at the classroom level across classes of different sizes. The details of our methodology are described next.

**Method**

A simulation study was conducted to examine the random and systematic error of classroom-level SGPs. True and observed scale scores were simulated to represent students’ test scores on a typical statewide assessment for grades 4 and 5. Observed scores were simulated using operational conditional standard errors of measurement for each scale score. Grade 5 true and observed SGPs were then calculated using the simulated data using grade 4 as the conditioning year. The data were simulated using a multilevel model to produce a nested structure that is observed in real data; that is, students were nested within classroom. Furthermore, to simulate realistic data, the parameters in the simulation were based on real test data. Bias, random error, and root mean square error (RMSE) were investigated across varying
classroom sizes to better understand the extent to which error is a function of true SGPs and
classroom size. One state’s recommended operational classification scheme was used to
determine rates of agreement between observed and true SGPs across 100 replications for every
classroom.

Data Generation

Generating true scale scores and SGP values. To generate the simulees’ true scores, a
two-level hierarchical linear model (HLM) with random-intercepts and slopes was used, where
level 1 represents the student-level scores, level 2 represents the classroom effect. The student-
level model (level 1) is presented in equation (1).

\[ Y_{ij} = \beta_{0j} + \beta_{1j} \times \text{Grade4}_{ij} + r_{ij} \]  

\( Y_{ij} \) represents student i’s 5th grade scale score in classroom j; \( \beta_{0j} \) and \( \beta_{1j} \) represent the intercept
and slope for students in classroom j and when regressing 5th grade scores on 4th grade scores;
\( \text{Grade4}_{ij} \) represents the grade 4 scale score for student i in classroom j; and \( r_{ij} \) represents the
level-1 residual for students in classroom j. Along with values for \( \beta_{0j} \) and \( \beta_{1j} \), the variance of the
student level residual, denoted \( \sigma^2 \), will be used in the simulation to generate scores.

The classroom-level model (level 2) is presented in equations (2) and (3).

\[ \beta_{0j} = \gamma_{00} + u_{0j} \]  

\[ \beta_{1j} = \gamma_{10} + u_{1j} \]  

\( \gamma_{00} \) represents the intercept for all students and \( u_{0j} \) represents the random intercepts for students
in classroom j. The variance of the intercepts, denoted \( \tau_{\beta_{0}} \), was used in the simulation. In
equation (3), \( \gamma_{10} \) represents the slope for students at the mean of classroom j when regressing 5th
grade scores on 4th grade scores. In addition, the correlation between the intercept and slopes, denoted, $\tau_{\beta_0, \beta_1}$, was used in the simulation.

The full model is shown in equation (4).

$$Y_{ij} = \gamma_{00} + \gamma_{10} \cdot \text{Grade4} + u_{0j} + u_{1j} \cdot \text{Grade4} + r_{ij}$$

(4)

Using real data from a large-scale, statewide assessment, we fit the previously described two-level multilevel model. Table 1 contains the parameter estimates that will be used to generate true scale scores. However, the variance at level-1 ($\sigma^2$) was manipulated so that the correlation between the grade 4 and grade 5 scale scores is approximately 0.85, which equals the disattenuated correlation coefficient in the real data.

Generating true scores with a nested structure is a two-step process. In step one, we sampled 5,000 intercepts and slopes for level 2 (i.e., classroom level) from a bivariate normal distribution with a mean vector defined by $\gamma_{00}$ and $\gamma_{10}$ and a covariance matrix defined by $\tau_{\beta_0}$, $\tau_{\beta_1}$, and $\tau_{\beta_0, \beta_1}$. For this study, the mean vector was 238 and 0.73 and the covariance matrix is given in equation (5).

$$\Sigma_{12} = \begin{bmatrix} 118 & 0.35 \\ 0.35 & 0.015 \end{bmatrix}$$

(5)

In step two, we first sampled 4th grade scale scores from a normal distribution with a mean of 240 and standard deviation of 15 (to mimic the mean and standard deviation of the state data we are emulating). We used the 4th grade scale score to predict the 5th grade scale score for classroom $j$ using a modification to equation (1), adjusting for the fact that the coefficient estimates are group-centered. We found a deviation for each student that represents the
difference between their Grade 4 score and the intercept for their classroom ($\beta_{0k}$). We used that deviation score in equation (1) to calculate Grade 5 scores.

$$\text{Deviation}_{ij} = \text{Grade4}_{ij} - \beta_{0j}$$

(6)

$$Y_{ij} = \beta_{0j} + \beta_{1j} \times \text{Deviation}_{ij} + r_{ij}$$

(7)

We then sampled $n$ students for a classroom from a normal distribution with the mean equal to the predicted value for classroom $j$ with a variance of 30. A variance of 30 was selected so that the correlation between the grade 4 and grade 5 scale scores was approximately 0.85, which equals the disattenuated correlation observed in the real data. Classroom sizes were sampled from a normal distribution with a mean of 20 and standard deviation of 5. The minimum $n$ for a classroom was 10 with a maximum of 36. The total $n$ of students in each data set was 100,158. True scale scores were rounded to the nearest integer and bounded between 200 and 280.

**Generating observed scale scores.** Observed scale scores were sampled from a normal distribution with the mean equal to the simulee’s true scale score and the standard deviation equal to the standard error of measurement conditioned on the true scale score (i.e., CSEM). The same CSEM was used for 4th and 5th grade, and was based on real test data (Figure 1). One hundred replications were conducted.

The true and observed student-level SGPs were determined via quantile regression using the true and observed scale scores. The R package SGP (Betebenner, VanIwaarden, & Domingue, 2013) was used to implement quantile regression. The true and observed classroom-level SGPs were based on the mean of the student-level SGPs within a classroom because it is used in practice.

**Data Analysis**
The relationship between the classroom-level true and observed SGPs was examined using the Spearman rho correlation coefficient. The amount of systematic error (bias) was examined by comparing the mean classroom-level SGP to the true classroom-level SGP as a function of the true SGP and classroom size. We also examined the amount of random error across replications by calculating the standard deviation of SGPs, using the difference between classroom-level observed SGPs and true SGPs for each replication. Lastly, we calculated the root mean square error for each classroom to determine the combined amount of random and systematic error.

To determine the amount of systematic error (or bias) present in observed SGPs, we use the following calculation:

\[
Bias = \frac{\sum_{k=100} SGP_k - SGP_{true}}{k} 
\]  

(8)

Where we calculate the difference between mean observed SGPs and true SGPs for each of \( k \) replications, and then find the mean across all replications.

To investigate the amount of error in our observed SGPs across replications, we use the following equation:

\[
SD_{SGP} = \sqrt{\frac{\sum_{k=100} (SGP_k - \overline{SGP})^2}{n-1}} 
\]  

(9)

Where we calculate for each replication the difference between the observed SGP (\( SGP \)) and the average of SGPs across replications (\( \overline{SGP} \)). This gives us the standard deviation of observed SGPs within a classroom.

To calculate root mean square error, we use the following formula:

\[
RMSE = \sqrt{SD^2 + Bias^2} 
\]  

(10)
Where we find the square root of the sum of the squared standard deviation (random error) and bias (systematic error).

Lastly, to investigate the practical implications of random and systematic error in aggregated SGPs, we classified teachers into four rating categories based on a method offered as guidance from a state education agency. The approach classifies teachers into rating categories based on numeric cuts along the classroom-level SGP scale. Table 2 contains the cut scores that define each of the performance categories.

For the analysis, teachers were classified into the four rating categories based on their classroom-level true SGP, as well as for each of the 100 observed SGPs they received in the simulation. The number and proportion of classifications in agreement between true and observed SGPs across all replications was calculated, which results in a 5000 x 1 column vector (one proportion agreement value for each classroom).

**Results**

**Bias**

The Spearman rho correlation between classroom-level true and observed SGPs was .917. The mean bias within classrooms across all replications is .04 SGPs. As can be seen in Figure 2, bias is really a function of true SGP rather than classroom size with small differences between class size categories. Classrooms with fewer than 19 students have the greatest bias, with an average bias of .11 for classrooms with between 10 and 14 students, and .07 for those with 15 – 19 students. For classrooms with 20 – 24 students, the average bias is .02, and .01 for classrooms with greater than 24 students (Table 2). Class size differences in bias exists mostly at the extremes. However, we can see that the mean SGPs for classrooms with lower true SGPs are
over-predicted by about seven SGPs, where those with higher true SGPs are under-predicted by about the same amount.

In Figure 3, bias is plotted as a function of true SGPs deciles and classroom size, so that variability in bias between and within each decile is more apparent. There appears to be a negative linear relationship between true SGP and bias across all four classroom size categories, with a decreasing amount of variability as class size gets larger.

**Random Error**

In Figure 4 (and Table 3), we can see that classrooms with the greatest number of students have the least amount of random error on average. The average standard deviation for classrooms with greater than 24 students is 6.8 SGPs, where it is 15.5 SGPs for classrooms with between 10 and 14 students. In practical terms, a teacher in a small classroom and an aggregate SGP of 50 could actually have an SGP between 20 and 80.

In Figure 5, random error is plotted as a function of true SGP and classroom size. Again, classrooms with fewer students have the greatest amount of random error throughout the true SGP scale. Random error is somewhat smaller as SGPs move out toward the extremes on the scale. In Figure 6, random error is again plotted as a function of true SGP deciles and class size, with substantially more variability in the amount of random error within a decile across class sizes. For instance, random error ranges from about 10 – 20 SGPs for the lowest true SGP decile for teachers with between 10 – 14 students, where it ranges from about 4 – 8 SGPs for classrooms in the lowest true SGP decile for teachers with greater than 24 students.

**Root mean square error (RMSE)**

RMSE is greatest for smallest classrooms, with an average RMSE of 16 for classrooms with fewer than 15 students, where it is only 7.7 for classrooms with greater than 24 students. In
Figure 6, the distribution of RMSE is almost bifurcated, where classrooms with true SGPs between 40 and 60 with more than 24 students have the lowest RMSE. Classrooms with SGPs toward the higher end and the fewest students have the highest RMSE.

**Classification**

Overall, mean agreement between true classroom-SGPs and each observed SGP across replications is 72.4 percent. In Figure 9, that average agreement is higher for true SGPs that are further away from the SGP cuts set forth by the state, each of which is plotted as a reference line. The proportion agreement is close to 100 percent for true SGPs below 22, after which agreement starts to decline all the way down to 20 percent for true SGPs near 35. For true SGPs just above the 35 cut, agreement increases slightly and then begins to decline again as true SGPs approach 50. However, classrooms in the top and bottom rating categories had the lowest average agreement across replications, compared with the two middle categories.

Those with true SGPs in the top rating category had an average agreement across all replications of 58.1 percent, where those in the bottom rating category had an agreement of 63.4 percent (Table 6). This results is compared with average agreement of 75.5 and 79.5 percent for the second and third categories, respectively. Similar patterns held when results are broken down by classroom size category, with greatest average agreement for the middle two rating categories. The lowest average agreement was for those in the highest true rating category with more than 24 students (55.1 percent), while the highest average agreement was for the same classroom size category, but for the third true SGP rating category (83.1 percent).

To further understand the magnitude of classification changes, we also calculated the number of categories classrooms change across all replications. As mentioned, 72.4 percent of teachers do not change categories at all, while 27.5 percent move up or down one category, and
just .1 percent move up or down two categories (Table 5). Classrooms with fewer students have
the lowest classification agreement across replications at 68.3 percent for classrooms with 10 –
14 students, compared with 74.3 percent agreement across replications for classrooms with
greater than 24 students.

Discussion

SGPs are widely used to evaluate teachers, but there has been little study of their
appropriateness for this purpose. In this study, we investigated the amounts of systematic and
random error in mean SGP scores at the classroom level across classes of various sizes. We
generated 4th and 5th grade true and observed Math scale scores using parameter estimates from a
two-level random intercepts and slopes model. One hundred observed score data sets and one
true score data set were created with five thousand classrooms with a total of 100,000 students in
each. Student-level true and observed SGPs were calculated using each of the data sets, and then
aggregated to the classroom level to investigate the amount of random and systematic error
present in mean SGPs. As a final step, classrooms were classified into one of four rating
categories to determine the extent to which error impacts stability of ratings across replications.

The results suggest random error in aggregate SGPs poses more of a threat to the
effective use of the measures in educator evaluation than systematic error does. In this study,
however, systematic error was greatest for classrooms with really low or high true SGPs, with
bias averaging to about seven SGPs for those classrooms near 20 or 80 on the true SGP scale.
Across all classrooms, bias averaged to just .04, where classrooms with the fewest students had
slightly higher bias with an average of .11 SGPs compared with about .01 for classrooms with
greater than 24 students. Random error was about 10 SGPs across all classrooms, which means
that a classroom with an average SGP of 50 could be as low as about 30 and as high as 70. The
range grows large for small classrooms, where those with between 10 and 14 students at 50 could have an average SGP between 20 and 80.

The practical implications of this level of error means many educators may be misclassified, which can potentially impact their overall evaluation. In this study, 27.4 percent of educators were misclassified by one category, meaning they could have been one category higher or lower on their aggregate SGP measure. A misclassification rate this high argues against using aggregated SGPs for high-stakes purposes such as rewarding or sanctioning teachers. A negligible proportion of classrooms moved more than one category either way (~.1 percent). The variability in SGPs could be mitigated somewhat through the use of confidence intervals, where educators are classified into a no-stakes category when there is too much uncertainty to make a meaningful decision. Policymakers advocating for the use of SGPs should also consider the extent to which misclassification is a function of where the educator is along the SGP scale. In this study, those in the top and bottom rating category had lower average agreement across replications, where those in the middle two had higher average agreement. This could be due to the RMSE being greater toward the ends of the true SGP scale.

**Limitations**

This study utilized simulated data, albeit simulated based on empirical data from one state. There were aspects to the data generation process that could have more closely mirrored the realities of operational data collected from a state. Fourth grade true scale scores were generated using a random normal distribution with a mean of 240 and standard deviation of 15, rather than with a nested structure like 5th grade true scale scores.

The number of students in each classroom, or attributable to an educator, was also not perfectly realistic, in that some educators may be responsible for more than 36 students. This is
often the case in realistic settings. In addition, the minimum number of students in a classroom was ten, so the results are really only generalizable to those classrooms. There are teachers in 8 students, 1 teacher, 1 paraprofessional settings, for which states may choose to provide an aggregate SGP (if combined with ELA scores, for instance).

Lastly, classification results were based on one state’s approach that used four specific classification categories. Given that states are taking many different approaches to using SGPs in evaluation systems, classification consistency will be affected by the number of categories and the cut points that define them. However, this study does closely mirror the practical realities that exist in many states choosing to use aggregate SGPs for evaluating educators.

**Summary**

More research is needed on the reliability and validity of SGPs. The present study represents one investigation of the reliability of aggregated SGPs and is modeled after the way many states are using them to evaluate teachers. Previous research has suggested the amount of error in student-level SGPs is substantial and expressed concern about reporting and interpreting them at the student level (e.g., Wells et al., 2014). The present study raises similar concerns for aggregated SGPs, which are being used for higher-stake decisions. Our results suggest more research is needed on SGPs, with particular attention paid to the amount of random error and its implications for classification decisions.
References


Appendix

Tables & Figures

Table 1

*Two-Level Model, 4th grade Math score predicting 5th grade Math score*

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\gamma_{00}$</td>
<td>238.02</td>
</tr>
<tr>
<td>$\gamma_{10}$</td>
<td>0.729</td>
</tr>
<tr>
<td>$\tau_{\beta_i}$</td>
<td>118.41</td>
</tr>
<tr>
<td>$\tau_{\beta}$</td>
<td>.349</td>
</tr>
<tr>
<td>$\tau_{\beta_0,\beta_i}$</td>
<td>.262</td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td>51.14</td>
</tr>
</tbody>
</table>

Figure 1

*CSEM by scale score (same for both Grade 4 and 5)*
Figure 2

*Systematic error (bias) as a function of true SGP and classroom size*

![Figure 2: Systematic error (bias) as a function of true SGP and classroom size](image)

Figure 3

*Systematic error as a function of true SGP decile and classroom size*

![Figure 3: Systematic error as a function of true SGP decile and classroom size](image)
Figure 4

*Random Error as a Function of Classroom Size*

![Graph showing random error as a function of classroom size.](image1)

Figure 5

*Random Error as a Function of Average True SGP and Classroom Size (mspline smoothing)*

![Graph showing random error as a function of true SGP and classroom size.](image2)
Figure 6

*Random Error as a function of true SGP decile and classroom size*

![Random Error as a function of true SGP decile and classroom size](image1)

Figure 7

*Root mean square error as a function of true SGP and classroom size (mspline smoothing)*

![Root mean square error as a function of true SGP and classroom size](image2)
Figure 8

Root mean square error as a function of true SGP decile and classroom size

Table 3

Random Error, Bias and RMSE as a function of classroom size

<table>
<thead>
<tr>
<th></th>
<th>Random Error (SD)</th>
<th>Bias</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 - 14 students</td>
<td>15.5</td>
<td>0.11</td>
<td>16.0</td>
</tr>
<tr>
<td>15 - 19</td>
<td>11.0</td>
<td>0.07</td>
<td>11.6</td>
</tr>
<tr>
<td>20 - 24</td>
<td>8.6</td>
<td>0.02</td>
<td>9.3</td>
</tr>
<tr>
<td>&gt; 24 students</td>
<td>6.9</td>
<td>-0.01</td>
<td>7.7</td>
</tr>
<tr>
<td>Total</td>
<td>10.0</td>
<td>0.04</td>
<td>10.7</td>
</tr>
</tbody>
</table>

Table 4

Classification Approach

<table>
<thead>
<tr>
<th>Effectiveness Category</th>
<th>SGP Cut Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bottom Rating Category</td>
<td>1 – 34</td>
</tr>
<tr>
<td>Category 2</td>
<td>35 - 49</td>
</tr>
<tr>
<td>Category 3</td>
<td>50 – 64</td>
</tr>
<tr>
<td>Top Rating Category</td>
<td>65 - 99</td>
</tr>
</tbody>
</table>
Table 5:

Classification changes as a function of classroom size

<table>
<thead>
<tr>
<th>Category</th>
<th>No Change</th>
<th>+/- One Category</th>
<th>+/- Two Categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 - 14 Students</td>
<td>68.0</td>
<td>31.7</td>
<td>0.3</td>
</tr>
<tr>
<td>15 - 19</td>
<td>71.6</td>
<td>28.3</td>
<td>0.1</td>
</tr>
<tr>
<td>20 - 24</td>
<td>73.9</td>
<td>26.1</td>
<td>0.0</td>
</tr>
<tr>
<td>&gt; 24</td>
<td>74.4</td>
<td>25.6</td>
<td>0.0</td>
</tr>
<tr>
<td>All Classrooms</td>
<td>72.4</td>
<td>27.5</td>
<td>0.1</td>
</tr>
</tbody>
</table>

Table 6:

Average agreement across replications by true SGP rating category and classroom size

<table>
<thead>
<tr>
<th>Category</th>
<th>Bottom Category</th>
<th>Category 2</th>
<th>Category 3</th>
<th>Top Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 - 14 Students</td>
<td>62.7</td>
<td>70.6</td>
<td>72.3</td>
<td>60.6</td>
</tr>
<tr>
<td>15 - 19</td>
<td>63.2</td>
<td>74.2</td>
<td>78.1</td>
<td>58.7</td>
</tr>
<tr>
<td>20 - 24</td>
<td>64.3</td>
<td>77.5</td>
<td>81.0</td>
<td>57.5</td>
</tr>
<tr>
<td>&gt; 24</td>
<td>62.7</td>
<td>78.3</td>
<td>83.1</td>
<td>55.9</td>
</tr>
<tr>
<td>All Classrooms</td>
<td>63.4</td>
<td>75.5</td>
<td>79.5</td>
<td>58.1</td>
</tr>
</tbody>
</table>
Figure 9

Proportion agreement as a function of true SGP and classroom size

Figure 10

Proportion agreement as a function of true SGP and classroom size (mspline smoothing)