A Technical Evaluation of the Student Growth Component of the Georgia Teacher and Leader Evaluation System

Daniel F. McCaffrey, Katherine E. Castellano, and J.R. Lockwood

Educational Testing Service

Final Report
November 26, 2014

A Technical Evaluation of the Student Growth Component of the Georgia Teacher and Leader Evaluation System

1. Background and Summary of Findings

The Georgia Department of Education's (GaDOE's) Race to the Top scope of work includes contracting an independent validation of its value-added growth model (GaDOE, 2011, p. 20). Accordingly, the GaDOE commissioned three independent studies by different research teams to examine aspects of the growth model used for teacher and leader evaluation in Georgia. The GaDOE selected topics for the investigation in consultation with its Educator Effectiveness Technical Advisory Committee (TAC). The studies examined issues that arose in the development and pilot of the state's educator effectiveness measurement system as part of its Race to the Top grant and which were discussed by the GaDOE and the TAC at multiple meetings.[1] This report, intended primarily for the GaDOE and those responsible for computing their growth measures, summarizes the findings of a study commissioned to disentangle the correlation the GaDOE found between students' prior achievement and educator ratings from the state's student achievement growth model. This section describes the study's motivating issue, including defining the growth model of interest, and summarizes the key findings that are presented in detail in the remainder of the report.

1.1 Motivating Issue

The GaDOE selected Student Growth Percentiles (SGP; Betebenner, 2009) as the measure of student achievement growth for assessing students and the growth component of its educator effectiveness measure. SGP are percentile ranks of students' current test scores in the

[1] Preliminary results of these studies were presented at a GaDOE Educator Effectiveness Technical Advisory Committee meeting in June 2014.

distribution of scores among students who scored similarly on prior year tests. For example, an SGP of 70 for a student conveys that the student scored higher on the current test than 70% of that student's peers who had similar test score histories. The methodology for SGP involves fitting nonlinear, quantile regressions for each conditional quantile .005 to .995 in increments of .01 for the current test scores given prior test scores. For detailed documentation of this methodology, see Betebenner (2009; 2011).

One of the appealing features of SGP is that they can easily be aggregated to higher levels, such as teachers or school leaders, to provide summaries of the performance of groups of students sharing common educational experiences. Typically, the mean or median SGP of students taught by a teacher or enrolled in a leader's school is used to summarize student growth. Both of these summary statistics are used for different purposes in Georgia (GaDOE, 2014): the median is used in the student growth model visualization tool, and the mean is used in the Georgia Teacher and Leader Effectiveness System. We use aggregated SGP (AGP) as a general term for either the mean or median SGP.

After generating AGP for teachers and schools for each content area and grade level for state standardized tests in 2011 to 2013, the GaDOE found moderate positive correlation between teacher or school AGP and the average prior achievement of the students linked to a teacher or school, respectively.[2] Accordingly, teachers (or leaders) whose students entered their classrooms (or schools) with high prior achievement tended to have higher AGP. There are two potential sources of such correlation:

[2] Some early GaDOE investigations found correlations between school-level AGP and average prior achievement of.4 or higher, which was larger than expected by the GaDOE. The sizes of the correlations varied across subjects and grade-levels, but in both mathematics and English language arts, there were correlations of over .3 for some grades and years, with some values as high as .5. Details on the correlations for the 2013 data are presented in Table 3.3.

1. Teacher or school sorting so that students with higher prior achievement are more likely to be attending schools that are more effective at promoting achievement growth, or assigned to teachers who are more effective at promoting achievement growth, than students with lower prior achievement.[3]
2. Statistical error in the AGP calculations that is correlated with students' prior achievement.

Statistical errors are differences between the AGP and true variations in teachers' performance. These statistical errors may be systematically related to the students in a teacher's class or classes due to bias in the SGP calculations resulting from measurement error in the achievement test scores. These errors then create a spurious correlation between the AGP and classroom or school average prior achievement. The question is how much each source (sorting or statistical error) contributes to the observed correlation. If the correlation is due to teacher sorting, it reveals important information about the distribution of teachers in the state.[4] If the correlation is a result of statistical error, then using AGP for teacher evaluations could result in errors in inferences about individual teachers that depend on the background characteristics of the teacher's students. The state would want to modify its SGP and AGP calculations to remove such statistical errors, if possible. Determining the source of the correlation is challenging, and we present explorations of the source below.

The GaDOE plans to use a modified version of the SGP methodology to measure student achievement growth. The standard SGP approach develops statistical models for the percentiles

[3] Although there are many aspects to teacher effectiveness, we use "effectiveness" to refer to educators' ability to promote student achievement growth.

[4] Teacher sorting can result in errors in AGP. When teachers are sorted, the AGP of more effective teachers will tend to be relatively too low, whereas the AGP of less effective teachers will tend to be relatively too high. The size of the errors is unknown because empirical studies cannot determine the level of teacher sorting. See Guarino, Reckase, Stacy, & Wooldridge (forthcoming), Walsh & Isenberg (forthcoming) or McCaffrey & Castellano (2014) for details.

of the current score distribution as a function of prior achievement (Betebenner, 2009; 2011). In the standard approach, the statistical models are updated annually when new test scores are released. We refer to this approach as "cohort SGP." Because the cohort SGP updates the statistical models every year, the SGP cannot reflect any general trends in achievement growth. That is, if across all students, achievement growth was greater in 2014 than in 2012, the distribution of cohort SGP in 2014 would not differ from the distribution in 2012. If educators were improving their practice and promoting greater achievement growth across different school years, cohort SGP would not capture this trend.

In an attempt to use SGP while also capturing any trends in educator effectiveness, the GaDOE will use "baseline-referenced SGP," rather than cohort SGP, for the student achievement growth component of its educator effectiveness measure, when the data necessary to calculate baseline-referenced SGP are available. For baseline-referenced SGP, or, simply "baseline SGP," the statistical models for the percentiles of the current year distribution are developed using a baseline cohort of students,[5] and the same models are then used annually with each new year of test score data to determine the percentile ranks of students' current test scores given their test score histories. Additional details on baseline-referenced SGP can be found in *A Guide to the Georgia Student Growth Model* (GaDOE, 2012).

Measurement error in the prior achievement test scores could potentially distort aggregate SGP as a measure of educator effectiveness (see Section 2) and create a spurious correlation between AGP and average student achievement or other student background variables. The GaDOE also implemented a simulation-extrapolation method (SIMEX; Carroll et al., 2006, Cook & Stefanksi, 1994) to correct for measurement error in the prior test scores in SGP calculations

[5] The baseline cohort for the GaDOE baseline-referenced SGP models included students from multiple adjacent school years.

and potentially mitigate the correlation between AGP and student background variables. Additional details on the SIMEX approach are provided in Section 2.2 of this report. The GaDOE applied this measurement error correction method to the baseline-referenced SGP; hence, we refer to them as "SIMEX-corrected, baseline-referenced SGP," or, simply, "SIMEX-baseline SGP." Although the GaDOE generally found that the positive correlation between AGP and mean prior achievement decreased for AGP computed from the SIMEX-baseline SGP, questions remained about the accuracy of the corrected AGP values and the interpretation of the correlation between the corrected AGP and average student background variables. Accordingly, we further investigate these concerns in this report.

1.2 Summary of Findings

This report summarizes an exploratory analysis of the correlation between AGP and average student background variables. The investigation includes an analytic study of measurement error and AGP and the SIMEX correction method. It also includes an empirical analysis of data from Georgia to explore further the correlation between AGP and student background variables.

The main findings from these investigations are:

1. Measurement error bias and AGP: Considering only measurement error in the *prior* achievement scores, measure error potentially can result in statistical errors in AGP in which teachers and leaders in economically disadvantaged schools may tend to have underestimated AGP, and teachers and leaders in schools serving economically advantaged students may tend to have overestimated AGP. Measurement error in the *current* score can also result in statistical error in AGP. These errors will tend to compress the expected value of educators' AGP toward 50, so that all else being equal,

the AGP of effective educators will tend to be underestimated and the AGP of ineffective educators will tend to be overestimated. Combining these two sources of measurement error may result in effective teachers and leaders in economically disadvantaged schools receiving AGP that are too low, and ineffective teachers and leaders in schools serving economically advantaged students receiving AGP that are too high. The combined effects of measurement error in prior and current test for other teachers and leaders will depend on other factors such as the reliability of the tests, the effectiveness of the individual educators, and the achievement and growth of their students.

2. SIMEX-corrected SGP: A review of the application of the SIMEX methodology to SGP and AGP found some potential limitations with this approach for removing errors introduced by test measurement error. However, the SIMEX measurement error correction implemented by the GaDOE reduced the correlation between AGP and averaged student background variables, including mean prior achievement.
3. Types of SGP: The correlation between AGP and mean prior achievement for empirical Georgia data were smallest for aggregated SIMEX-baseline SGP, followed by aggregated cohort SGP, and then strongest for aggregated baseline SGP. The medians tended to have slightly lower correlations than means within each SGP type.
4. Investigating Evidence of Bias in AGP for Georgia Data: From three empirical studies in which teachers were kept constant so as to disentangle the effect of bias in AGP from teacher sorting as the source of the moderate, positive correlation with mean prior achievement, we found evidence of teacher sorting and little evidence of spurious correlation between the AGP and student background characteristics among classes taught by the same teacher.

5. More Prior Years: Empirical analyses revealed adding more prior test scores in the SGP calculations reduced the spurious correlation between AGP and mean prior achievement.
6. Baseline vs Cohort SGP: Baseline-referenced SGP, as implemented by the GaDOE, only include two prior years of test scores and thus do not take advantage of the finding that including more prior years is beneficial (see #5). Moreover, baseline SGP rely on strong assumptions of test equating that need to be verified before implemented operationally.

We present these findings in detail in the remainder of this report, which we divide into four additional sections. In Section 2, we present our analytic evaluation of the impact of measurement error in student test scores on SGP and AGP and of the SIMEX method for correcting for that measurement error. In Section 3, we present findings from our empirical analyses of the Georgia data. In Section 4, we note some considerations for the use of baseline SGP in the AGP calculation. Lastly, in Section 5, we discuss implications for the implementation of AGP in Georgia's teacher and leader evaluation system.

2. Analytic Evaluation of AGP and Measurement Error

In Section 1, we identified two possible sources for the positive correlation the GaDOE found between AGP and mean prior achievement. One of these sources is statistical error in the AGP calculations that is correlated with students' prior achievement. These statistical errors may be systematically related to the students in a teacher's class or classes due to bias in the SGP calculations resulting from measurement error in the achievement test scores. In this section, we describe how measurement error in the test scores may bias SGP and, consequently, AGP.. Our conclusions follow from analytical derivations detailed in Appendix A. Before presenting our analytic results, we define key terms.

All tests have measurement error. Following well-established standard practice, we use "true score" to refer to a student's score if there was no measurement error, and "observed score" to refer to actual test scores that we observe with measurement error. SGP are defined as the percentile ranks of students' current observed test scores in the distribution of scores among students with similar observed test score histories. We call these the "observed SGP" or, simply, the "SGP." An alternative ranking of a student is the percentile rank of the current true score among students with similar true score histories. We call this the student's "true SGP." We cannot calculate the true SGP because we cannot observe the student's current or prior true scores. Because of test measurement error, the true SGP and the SGP will not be the same. Exactly how they relate to each other is difficult to assess because there are errors in both the current and prior year tests. We cannot observe achievement without some error, so we cannot rank students' true scores. Methods such as SIMEX use information on test score measurement error to try to obtain a better estimate of the true SGP than is provided by ranking student observed scores, as is done in the common methods of calculating SGP.

AGP are aggregated SGP. Consequently, AGP based on the true SGP will differ from AGP based on the observed SGP. Given that the goal is to use AGP to learn about teachers' or leaders' contributions to student achievement, ideally we would have AGP based on the true SGP ("true AGP") rather than AGP calculated from the observed SGP ("observed AGP" or "AGP").[6] Deviations of the observed AGP from the true AGP are errors that may distort inferences about educators.

In this section, we first discuss likely patterns for the differences between the true and observed SGP. We then discuss the implications of these differences for AGP and how they

[6] Although the direct goal of the AGP is to quantify student growth associated with teachers and leaders to use in the evaluation system, the goal of the evaluation system is to make determinations of teacher and leader effectiveness as demonstrated by the name "Leader or Teacher Keys Effectiveness System"

might lead to systematic errors in inferences about educators that may be correlated with student background variables or the true AGP. Subsequently, we discuss how SIMEX might change SGP and AGP. We need a model for the distribution of test scores to conduct our analytic evaluation of the statistical properties of SGP and AGP based on achievement, as we cannot directly observe these quantities. We assume that true scores for student achievement tests and the resulting observed scores are normally distributed.[7] This assumption allowed us to derive closed form analytic expressions for the statistical properties of SGP and AGP. Although observed achievement test scores are not normally distributed, the insights gained from studying SGP and AGP under this model for the test scores are likely to be valid for real data. The substantive results we derive using the normal distribution will hold for any symmetric distribution of the test scores although closed form expressions may not be possible for other models. Moreover, predictions of the behaviors of estimated AGP and SGP based on these insights have been borne out with the actual SGP provided by the GaDOE and then AGP we calculated from it. Details of our derivations are presented in Appendix A.

2.1 SGP

To understand how measurement error affects SGP, we first compare the ranking of *current observed scores* to the ranking of *current true scores* among students with *similar true score histories*. Using theoretical results, we derived analytic formulas (presented in Appendix A) for the statistical properties of the test score rankings under these conditions. We could not conduct empirical studies under these conditions because we cannot observe true scores. Next, we compare the ranking of current true scores among students with *similar true score histories* with the ranking of current true scores among students with *similar observed score histories*.

[7] We also assume the true scores and measurement errors have constant variance which is not generally true of the measurement error in test scores.

Again, we used theoretical results to derive formulas (presented in Appendix A) for the properties of the rankings. Finally, we put these pieces together to derive analytic formulas for the deviation of SGP based on current and prior observed scores from true SGP based on current and prior true scores.

2.1.1 Conditioning on Prior True Scores but Ranking Current Observed Scores vs Current True Scores

For a given student in a given year, the test score measurement error might be positive or negative, and it will shift the ranking of the student's observed score up or down relative to the ranking of his or her true score. Rather than focusing on any individual year and student, we consider how a student's expected or mean SGP based on the current year observed score would compare with his or her true SGP and how this might vary for students with different levels of current true scores. The expected SGP equals the mean SGP over all the possible measurement errors that might result given the student's prior true score. When the data are normally distributed, we find that the expected SGP will be compressed toward 50 relative to the true SGP. That is

- For students whose true SGP is *greater than* 50, the expected value of their observed SGP will be *less than* the true SGP.
- For students whose true SGP is *less than* 50, the expected value of their observed SGP will be *greater than* the true SGP.
- For students with a true SGP equal to 50, their observed SGP will on average be correct.

In other words, using observed current scores rather than true current scores to calculate the SGP will tend to pull the SGP toward 50.

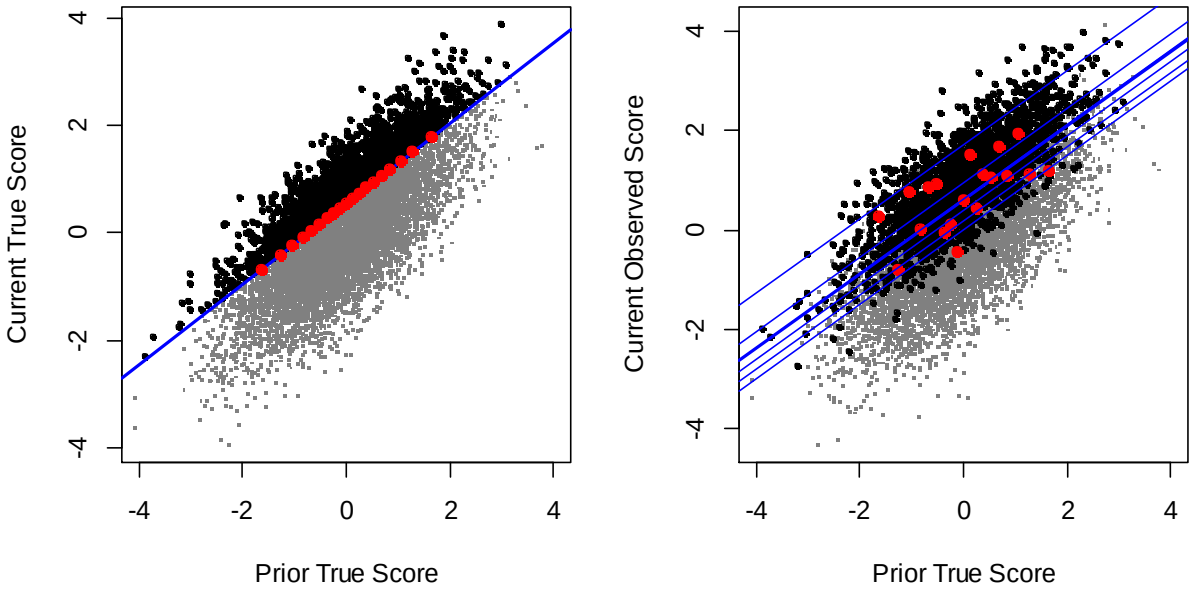


Figure 2.1. A demonstration of the impact of measurement error in the current scores on SGP with current true scores plotted against prior true scores (left-hand panel) and current observed scores plotted against prior true scores (right-hand panel). The bold blue line in the left-hand panel indicates the conditional 80th percentile of current true given prior true scores, and the students who fall along this line, indicated by the red dots, have true SGP of 80. The blue lines in the right-hand panel from bottom up represent the 50^{th}, 60^{th}, 70^{th}, 80^{th} (bold line) 90^{th}, and 99^{th} conditional percentiles of current observed versus prior true scores. The students with true SGP of 80 are again represented by red dots in the right-hand plot, but their observed SGP tend to be less than 80 as their current scores tend to fall below the 80^{th} conditional percentile line. This figure indicates that students with above average true SGP tend to have underestimated observed SGP.

Figure 2.1 demonstrates these results. This figure shows simulated data that was generated using the distributional assumptions given in Appendix A. The panel on the left contains a plot of the simulated current true scores versus prior true scores with a dark blue line indicating the 80^{th} conditional percentile of current true scores given prior true scores. Thus, the students (red dots) who fall along this line have an above average true SGP of 80; that is, compared to students with their same prior true scores, about 80 percent of these comparison students (gray dots) have current scores less than the current scores of these students of interest. These same students (red dots) are represented in the panel on the right that plots current

observed scores simulated from a test with reliability of .80 versus prior true scores. However, these students' scores no longer fall along the 80th conditional percentile line; rather, they are scattered. The blue lines in the right-hand plot are the 50th, 60th, 70th, 80th, 90th, and 99th conditional percentile lines of current *observed* scores given prior true scores, respectively, and they define the observed SGP. The (vertical) distances between the 50th and 60th conditional percentile lines are more compressed than between the lines for the larger percentiles, particularly for the 90th and 99th lines. Accordingly, for the students of interest (red dots), their observed SGP tend to be compressed toward 50, or less than their true SGP of 80, which is consistent with the bulleted results listed above.

Following general statistical practice, we call the difference between the true SGP and expected value of the observed SGP "bias." The direction and size of the bias depends on the SGP and student's current true score. In this scenario, the bias does *not* depend on the student's prior true score. The bias does depend, however, on the reliability of the test. As the reliability of the test decreases, the expected value of the observed SGP for students with a given true SGP will move closer to 50. We would not expect to rank students differently, but bias would tend to lead us to conclude students were more like the average student than is actually true. The exact derivations we used to determine the bias assumed normally distributed data. However, these conclusions will hold whenever the test scores follow a bell-shaped distribution and measurement error is at least roughly symmetric around true scores, and these features are generally true of test scores.

2.1.2 Ranking Current True Scores but Conditional on Prior Observed Scores vs. Prior True Scores

We now consider the ranking of students' current true scores among students with similar *true* score histories as compared with their ranking among students with similar *observed* score

histories. We consider matching students using a single test score (or true score). For a student with a prior true score that is above average, when we match this student to other students with the same prior observed score, more of those students will have lower prior true scores, rather than higher prior true scores. For students who had above average prior true scores, matching by observed scores will yield a comparison set of peers who tend to have lower true scores. The student's current true score will tend to rank higher among this set of peers than it would among a set of peers of equal prior true scores. Thus, we find:

- For students whose *prior true scores are above average*, their SGP (based on matching with prior observed scores) will tend to be too high or biased upward compared with their true SGP (that matches students on prior true scores).
- For students whose *prior true scores are below average*, their SGP (based on matching with prior observed scores) will tend to be too low or biased downward compared with their true SGP (that matches students on prior true scores).

Figure 2.2 demonstrates these results. As in Figure 2.1, this figure shows simulated data that was generated using the distributional assumptions laid out in Appendix A. The left panel contains a plot of the simulated current true scores versus prior observed scores from a test with reliability of .80. The scores of one student are highlighted by the yellow diamond. Students with similar prior observed scores are highlighted as red dots. These students have the same prior *observed* score as the target student.[8] About 39 percent of current true scores for these comparison students are less than the current true score of the target student. Accordingly, if we used prior observed scores for calculating the SGP, the target student would receive a 39. The panel on the right shows a plot of the current true score versus the prior true score. The score of

[8] Given these generated data are continuous, no two students have the exact same prior observed score. Thus, we used an interval of +/- 0.05 standard deviation units to identify students with the "same" observed prior score.

the target student is again shown as a yellow diamond. This student's observed and true prior scores were very similar. The target student's set of comparison set should be the students with scores along the gray vertical line. In contrast, the red dots show the comparison students based on the observed rather than the true prior score; that is, the same comparison students as shown in the right panel.

As we described above, in the right-hand panel, more of the red dots are below the gray line than above it: For students who have above average prior true scores, matching by their observed prior scores will yield a comparison set of peers who tend to have lower true prior scores. Along the right axis of this right panel, we plot the densities for the two possible comparisons groups—the one based on the observed prior scores (red curve) and the one based on the true prior scores (gray curve) with a horizontal gray line marking the target student's current score. The density of the true score comparison group is shifted upward so that our target student's true SGP would be 25, which, as expected from the bulleted results above, is smaller than the SGP of 39 based on the observed comparison students.

In practice, SGP calculations typically use multiple prior test scores. The true scores associated with the test scores are correlated—some students' true scores would tend to be above average year after year and other students' true scores will tend to be below average. However, measurement errors are not correlated across years. Consequently, matching students on multiple prior observed test scores as opposed to matching with a single test score tends to improve the quality of the match so that students are compared with peers whose true scores are more like his or her true scores. Accordingly, bias is reduced, but it can still exist.

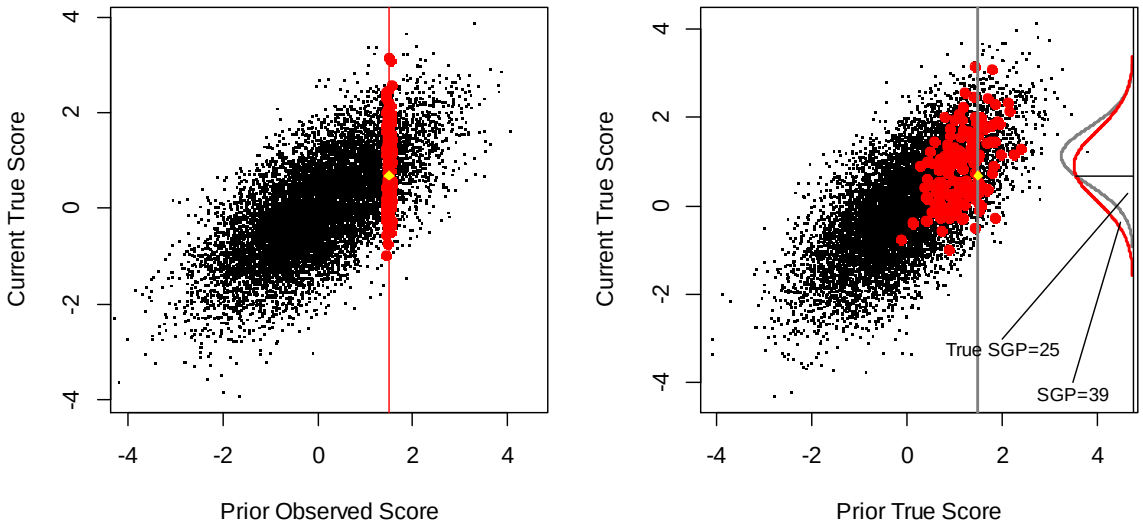


Figure 2.2. A demonstration of the impact of measurement error in prior scores on SGP with current true scores plotted against prior observed scores (left-hand panel) and prior true scores (right-hand panel). The yellow point represents a target student of interest for illustrative purposes, and the red points represent students who share the same prior observed score as the target student. The target student has an above average prior true score and the prior true scores of matched students tend to be lower than the target student's leading to an observed SGP of 39 for the target student, which is larger than this student's true SGP of 25, see densities on the right axis. This figure illustrates that students with above average prior (true) achievement have overestimated SGP.

2.1.3 Observed SGP vs. True SGP

The biasing effects of measurement error in both the current and prior scores described in the previous two sections occur simultaneously. Their net effect is what determines the bias in SGP for any given student. In some cases it is known that the two sources of bias are acting in opposite directions, and therefore the direction of the net bias is unclear. For example, for students with above average prior true scores and above average current true scores among peers with the same prior true scores, matching on prior observed scores will tend to bias their SGP upward, but measurement error in the current score will tend to bias it downward. The net bias will depend on the values of the student's current and prior true scores and the reliability of the tests, or be "Indeterminate" as indicated in the upper left-hand cell in Table 2.1. This table

summarizes the direction of the bias in SGP depending on the status of students' prior and current true scores. In cases where it is known that the two sources of bias are acting in the same direction, it is possible to determine the direction of the net bias. For example, for students with above average prior true scores and below average current true scores among peers with the same true score history, measurement error in their current and prior scores will tend to bias their SGP upward (upper right-hand cell in Table 2.1). Similarly, for students with below average prior true scores and above average current true scores relative to their peers with the same true score history, measurement error in the current and prior test will tend to bias their SGP downward (lower left-hand cell in Table 2.1). However, if they had below average current true scores compared with their peers, then the biases would go in opposite directions, producing indeterminate bias as given in the lower right-hand cell in Table 2.1.

Table 2.1. Summary of bias direction in SGP due to measurement error in current or prior test scores.

	Current True Score Among Peers with Equal Prior True Score	
Prior True Score	Above Average	Below Average
Above Average	Indeterminate	Bias Upward
Below Average	Bias downward	Indeterminate

Note: Indeterminate indicates bias from two sources is in different directions so the direction is unknown and will depend on the true scores and test reliability.

The bias due to measurement error can be large. For example if the correlation between the current and prior observed test scores is .78, the reliability for both the current and prior scores is .87, and the true scores and measurement error are normally distributed, then the average of the absolute value of the bias would be 9.9 percentile points. The values for the correlation and test reliability are consistent with values from historic Georgia test score data. For example, the average observed SGP would be about 52 among students whose true prior achievement is at the 25^{th} percentile of all students and whose true SGP is 60. The average of

observed SGP would be about 54 for students with prior achievement at the 40th percentile and with a true of SGP of 60. Similar results hold for students with above average prior achievement. Figure A.1 in Appendix A provides a complete summary of the bias across the range of prior true scores and true SGP.

2.1.4 Measurement Error Bias in AGP

Because students' average prior true scores vary among schools and classrooms, bias in the SGP that is associated with the prior true scores results in bias in AGP that is systematically related to the students' prior true scores. Test scores and student socio-demographic variables are correlated with students' prior true scores so that they also tend to be correlated with bias in the SGP. Among equally effective schools or equally effective teachers' classrooms, those with students with lower prior true scores (and, consequently, lower average observed scores) will tend to have lower AGP because of bias resulting from measurement error in the prior scores. We call this "prior score bias" in the AGP. Prior score bias tends to deflate the AGP of educators of disadvantaged students and inflate the AGP of educators of advantaged students.

By definition, effective teachers or leaders will have students who tend to score above average relative to peers with similar true score histories. Bias due to measurement error in current scores will tend to suppress the AGP for these teachers. The opposite is true for ineffective teachers or leaders: Bias due to measurement error in the current scores will tend to inflate their AGP. We call this "current score" bias in the AGP. Current score bias tends to depress the AGP of effective teachers and inflate the AGP of ineffective teachers. As shown in Table 2.2, effective teachers and leaders in economically disadvantaged schools will tend to be penalized by both current and prior score bias, and ineffective teachers and leaders in schools serving economically advantaged students will tend to benefit from both sources of bias. For the

two other possible cases of ineffective teachers in schools where students have low prior achievement, such as economically disadvantaged schools, (top left cell of Table 2.2) and effective teachers in schools students have high prior achievement, such as economically advantaged schools, (bottom right cell of Table 2.2), the bias is indeterminate.

Table 2.2. Summary of bias direction in AGP due to measurement error in current or prior test scores.

	Educator Effectiveness (Current Student Achievement)	
Prior Student Achievement	Ineffective (Low Current Scores Given Prior Scores)	Effective (High Current Scores Given Prior Scores)
Below Average	Indeterminate	Bias downward
Above Average	Bias Upward	Indeterminate

Note: Indeterminate indicates bias from two sources is in different directions so the direction is unknown and will depend on the true scores of the teacher's students, the level of the teacher effectiveness, and test reliability.

2.2 SIMEX

The SIMEX method (Carroll et al., 2006, Cook & Stefanksi, 1994) was designed to estimate coefficients from statistical models when variables used in the models have measurement error. Essentially, the SIMEX method has four steps:

1. Additional measurement error is added to the existing data by simulating random numbers from a distribution with a particular measurement error variance and adding them to the observed data so that the data become noisier. This step is repeated for a sequence of increasing values for the variance of the additional measurement error and repeated multiple times for each value of the sequence.
2. The model coefficients are estimated using each of the data sets with the simulated additional data measurement errors. The estimates are averaged across the multiple simulated data sets for each value of the variance of the additional measurement error.

3. The values of the estimated coefficients are modeled as a function of the measurement error variance (such as a linear or quadratic function).
4. Using the function from Step 3, the values of the estimated coefficients are projected or *extrapolated* to the case of no measurement error. These values from the projection serve as the final estimates.

For more details on this method, see Lederer and Kuchenhoff (2006). In this section, we discuss the GaDOE's application of this method to SGP and investigate its impact on bias in AGP.

Estimation of AGP involves an algorithmic process, which is more complex than estimating model coefficients in a statistical model. It involves the estimation of 100 quantile functions, the calculation of 100 percentiles for each student, the determination of the SGP for each student from the percentiles, and the aggregation of the SGP. However, the basic algorithm of the SIMEX method can be applied to the estimation of AGP, in that, simulated additional measurement error can be added to test scores. Estimates, SGP or AGP, depending on the approach taken with SIMEX, can be modeled as a function of the amount of measurement error and projected to the case of no measurement error.

The GaDOE and its consultants apply the SIMEX method to one component of the AGP process. They use SIMEX to correct the estimated percentile values for each student rather than the percentile ranks.[9] Specifically, they use the following procedure, according to our examination of the SGP package code (Betebenner et al., 2014) and documentation provided by the GaDOE:

1. Add additional measurement errors to the prior achievement scores to create multiple simulated datasets following the general SIMEX procedures outlined above.

[9] Note that "percentile" and "percentile rank" represent two different, though related, values. For instance, if we say that a student's current score of 280 is at or above 60 percent of current scores for students with the same prior test scores, the value 280 represent the 60th percentile and the value 60 represents the student's percentile rank.

2. For each of the simulated data sets, refit the 100 quantile functions using the data with the additional simulated measurement error.
3. Using these functions, estimate the percentiles for each student using the student's simulated prior achievement data with extra measurement error.
4. Average these values from Step 3 across simulated data sets for each value of the variance of the additional measurement errors, and project the percentiles to the case of no measurement error (extrapolation step).
5. Use these corrected percentiles to calculate the percentile rank of the student's observed current scores and aggregate these values to obtain their final corrected AGP.

They do not make any correction for measurement error in the current year scores. The adaptation of SIMEX used by the GaDOE only adjusts for measurement error in the prior scores and adjusts for those errors in an intermediate step of the entire AGP process.[10]

Consequently, two questions must be answered about this procedure to determine its effect on bias in the AGP:

1. Can the SIMEX method estimate without bias the percentiles of the current observed score distribution among students who have the same prior true score histories?
2. Can SIMEX provide an unbiased estimate of the percentile ranks of current true scores? That is, is the use of SIMEX sufficient to remove all the bias that can result from measurement error?

2.2.1 Can the SIMEX method estimate the percentiles of the current observed score distribution among students who have the same prior true score histories?

The primary challenge to applying SIMEX is the extrapolation step. If the function used for the extrapolation is not correct, then the final estimate of the percentiles will be biased. The GaDOE and its contractors approximated the projection function with a linear function of the

[10] Shang (personal communication, October 29, 2014) reports the SIMEX approach of simulating data, applying the entire AGP process to those data, and projecting estimated AGP yielded unsatisfactory results, whereas the approach used by Georgia of using SIMEX to correct the quantiles yielded more satisfactory results. Thus, although the SIMEX approach used by GaDOE may not remove all the bias due to measurement error, it may still perform better than alternatives.

measurement error variance. The projection function is almost certainly not a linear function.[11] For instance, it is not linear when the data are normally distributed (see Appendix A). Consequently, some amount of bias due to measurement error in prior year test scores is likely to remain in the percentiles. However, even when the projection function is mis-specified, SIMEX can reduce the bias in estimates. We cannot determine how much bias might remain. Other types of projection functions, such as polynomial functions, are often used in other SIMEX applications, but there is a tradeoff between using these more complex functions to reduce bias and introducing additional random statistical errors. The GaDOE contractors reported that a linear model for the projection best balanced the two goals of reducing bias and limiting random statistical errors in simulation studies.

Another challenge with applying SIMEX to AGP is that measurement errors in test score data has non-constant variance that depends on the unknown true scores. In limited explorations, we found that in general there is no way to use available data on test scores and their measurement error to simulate data in the simulation step of SIMEX so that it produces estimates that are free of bias from measurement error. We, however, found that applying SIMEX typically reduced bias compared with estimates that did not correct for measurement error. Given the challenges of selecting a projection function and simulating additional measurement errors with the correct variance, we suspect SIMEX has reduced but not removed bias in the percentiles. The GaDOE might commission additional simulation studies to understand better how much bias SIMEX can remove under some circumstances. It might also conduct sensitivity analyses using

[11] The GaDOE referred us to Shang, Van Iwaarden, and Betebenner (forthcoming) for justification of using linear extrapolation. In that paper, the authors report the results of a simulation study in which a linear projection function yielded smaller total error – the combined bias and random statistical error – but greater bias than the quadratic extrapolation. Additionally, in an earlier study, Shang (2012) reported that the linear extrapolation yielded greater bias but less random error than a quadratic extrapolation function.

its test score data to determine how sensitive results are to different extrapolation functions or methods for simulating additional measurement errors.

2.2.2 Can SIMEX provide an unbiased estimate of the percentile ranks of current true scores?

As noted above, the GaDOE application of SIMEX adjusts the percentiles that are then used to obtain the percentile ranks that are the individual students' SGP, and the SGP are then aggregated for the AGP. However, as described in Section 2.1.1, measurement error in the current year test scores biases the calculation of the SGP from the percentiles, even if the percentiles are calculated without error. The SIMEX implementation used by GaDOE makes no correction for the measurement error in the current test score. This measurement error in the current score will tend to compress the expected value of the SGP toward 50, as was discussed in Section 2.1.1. Using SIMEX can potentially remove bias that is correlated with students' prior achievement through its adjustment of the percentiles, but the expected value of SGP will still be compressed toward 50 by measurement error in the current scores.[12] The GaDOE could attempt an implementation of SIMEX that directly adjusts the percentile rank for measurement error in both the current and prior scores. Alternatively, the GaDOE could estimate the percentile ranks of current scores for students with similar test score histories using alternative methods. For example, they could use item responses to model the bivariate or multivariate distribution of true scores and from this distribution estimate the desired percentile ranks. Also, other statistical methods for estimating percentile ranks could be used. Examples of these methods are discussed in Lockwood and Castellano (forthcoming).

[12] Calculating percentile ranks from estimated percentiles results in bias in the SGP compared with true SGP even if the percentiles are estimated unbiasedly. This is because the percentile ranks are nonlinear functions of the estimated percentiles. The bias is unrelated to the level of the prior true score, but it is related to the level of current true score. The bias will again tend to compress the SGP toward 50 relative to the true SGP. For true SGP below 50, the bias will be positive, and for true SGP above 50, the bias will be negative.

For AGP, SIMEX can potentially remove the bias due to measurement error in AGP that is positively correlated with the teacher or leader's mean prior true scores and observed scores (averaged over the students in their classes or schools). However, SIMEX might exacerbate the bias that is negatively correlated with the effectiveness of the leader or teacher. It is impossible to know how large the various biases would be without aggregates of students' true SGP, which we cannot observe. As we discuss in Section 3, the SIMEX correction typically resulted in AGP with lower correlations with student background variables than the uncorrected AGP for real Georgia test score data. If SIMEX was reducing bias due to measurement error in prior test scores, we would expect such a reduction in the correlation between AGP and background variables. Other factors, such as a greater random statistical error in the AGP or a compression of SGP toward 50, could also reduce the correlation between AGP and student background variables, so the empirical evidence is not conclusive. It does, however, suggest a potential benefit from SIMEX that may warrant the use of this method, though we recommend the GaDOE pursue further analyses of their implementation of this method.

3. Empirical Analyses

We conducted a series of analyses using real Georgia student test score data to explore the source of the correlation between aggregate student background variables, including prior achievement, and teacher and leader AGP. First, we present relevant summary statistics, and second, we explicate our analyses, which followed two plans. The first plan was to disentangle the effect of bias from teacher sorting in AGP as the source of the moderate, positive correlation with mean prior achievement. For this plan, we compared AGP from classes that differed in the characteristics of the students, but in which the teaching effectiveness might be the same. The

variation in AGP across classes could then be attributed to bias rather than true differences in the teaching. We conducted three variations on this analysis.

Our second plan was to estimate AGP with different numbers of prior test scores used to match students to peers. As noted previously, controlling for more tests can compensate for test measurement error and mitigate bias. Thus, by using more scores we can potentially reduce bias due to measurement error without the potential biases that SIMEX might introduce. We explored the use of multiple prior years of the same subject scores as the current year score being modeled and the use of language arts, math, reading, and science scores from one or more years regardless of the subject of the current year score.

3.1 Constant Teachers with Varying Students

To isolate bias in the variation in AGP across classes as opposed to true differences in teaching quality, we conducted three studies: comparing the AGP of teachers who taught classes of varying prior achievement levels over adjacent years, modeling within-teacher variability by aggregate prior student performance and percentage of economically disadvantaged students, and comparing the AGP of teachers who taught accelerated and regular track mathematics courses. For the first two analyses, we used data from the entire state provided by the GaDOE. We first describe these data and then discuss each of the three analyses.

3.1.1 Summary of State Data

We used teacher-linked student data files provided by the GaDOE for three years: 2011, 2012, and 2013. These data included SGP for individual students calculated by the GaDOE. For these analyses, we used the GaDOE SGP data from the English Language Arts (ELA) and Mathematics (Math) tests. We present ELA results in tables labeled with an “a” and Math results

in tables labeled with a "b." Students with SGP for both mathematics and ELA contribute to both tables and teachers of both subjects also contribute to both tables.

These student-level datasets contained several variables, but we focused on a subset for our empirical analyses. Specifically, the variables of interest were the student identification number, teacher identification number, school identification number, prior grade-level student test scores (standardized), cohort SGP, baseline SGP, and SIMEX-baseline SGP. From these student-level files, we created teacher-level files by aggregating over the students linked to a given teacher (i.e., with the same teacher identification number). For each grade-level (grades 4 to 8) by subject area (ELA and Math), we created teacher-level datasets that contained the following variables: total number of students, mean prior test scores, and means and medians of each of the three types of available SGP.

Before performing any analyses with the state data, we used a few exclusion rules to ensure we analyzed AGP only for eligible teachers. Specifically, we dropped:

- non-teacher records (e.g., records with teacher identifiers equal to "Unknown" or "Contracted Services"; GaDOE confirmed all identifiers we classified as non-teachers),
- records with the Primaryfirst variable not equal to 1 (as directed by GaDOE),
- teachers within a particular grade-level and subject area if they taught fewer than 15 students (following the precedent in Georgia).

This last exclusion rule resulted in dropping about 25-35% of all teacher records within a given grade-level for both ELA and Math. As shown in Table 3.1a and Table 3.1b, a small number of teachers in all grades and years linked to over 150 ELA and Math students, respectively. For these teachers with more than 150 students, the number of linked students was as high as 363 for a single ELA teacher and 274 for a single Math teacher.

Some teachers linked to more than one school, and some to many schools. For instance, some teachers linked to students enrolled in as many as 17 different schools. However, by 2013, about 96 percent of teachers linked to students in only one school, and only a handful of teachers linked to 15 or more students from each of two or more schools. Students also linked to multiple teachers (up to as many as 8 for ELA and 6 for Math), although about 97 to 98 percent of students linked to just one or two teachers and 99 percent linked to three or fewer teachers. We were unable to investigate these linkages, so we accepted them as correct, but they might warrant further investigation. We do not think they had an impact on our analyses.

Table 3.1a. Distribution of teachers by number of linked students by grade-level and year, ELA.

2011

Number of Students	Grade 4		Grade 5		Grade 6		Grade 7		Grade 8	
	N	%	N	%	N	%	N	%	N	%
<15	1601	26.39	1604	27.77	1104	32.51	1137	33.69	1037	32.67
15-150	4454	73.43	4156	71.97	2276	67.02	2213	65.57	2125	66.95
>150	11	0.18	15	0.26	16	0.47	25	0.74	12	0.38
Total	6066	100.00	5775	100.00	3396	100.00	3375	100.00	3174	100.00

2012

Number of Students	Grade 4		Grade 5		Grade 6		Grade 7		Grade 8	
	N	%	N	%	N	%	N	%	N	%
<15	1555	26.45	1505	26.67	1017	31.53	1065	33.41	967	31.59
15-150	4306	73.26	4120	73.00	2194	68.01	2112	66.25	2082	68.02
>150	17	0.29	19	0.34	15	0.46	11	0.35	12	0.39
Total	5878	100.00	5644	100.00	3226	100.00	3188	100.00	3061	100.00

2013

Number of Students	Grade 4		Grade 5		Grade 6		Grade 7		Grade 8	
	N	%	N	%	N	%	N	%	N	%
<15	1732	29.50	1612	29.07	919	31.13	976	32.98	922	31.99
15-150	4140	70.50	3932	70.90	2018	68.36	1970	66.58	1943	67.42
>150	0	0.00	2	0.04	15	0.51	13	0.44	17	0.59
Total	5872	100.00	5546	100.00	2952	100.00	2959	100.00	2882	100.00

Table 3.1b. Distribution of teachers by number of linked students by grade-level and year, mathematics.

2011

Number of Students	Grade 4		Grade 5		Grade 6		Grade 7		Grade 8	
	N	%	N	%	N	%	N	%	N	%
<15	1397	25.36	1399	27.25	902	31.15	892	30.80	878	31.39
15-150	4084	74.15	3706	72.19	1988	68.65	1994	68.85	1913	68.39
>150	27	0.49	29	0.56	6	0.21	10	0.35	6	0.21
Total	5508	100.00	5134	100.00	2896	100.00	2896	100.00	2797	100.00

2012

Number of Students	Grade 4		Grade 5		Grade 6		Grade 7		Grade 8	
	N	%	N	%	N	%	N	%	N	%
<15	1356	25.73	1268	26.04	825	30.07	817	29.64	804	29.90
15-150	3892	73.84	3580	73.53	1911	69.64	1932	70.10	1879	69.88
>150	23	0.44	21	0.43	8	0.29	7	0.25	6	0.22
Total	5271	100.00	4869	100.00	2744	100.00	2756	100.00	2689	100.00

2013

Number of Students	Grade 4		Grade 5		Grade 6		Grade 7		Grade 8	
	N	%	N	%	N	%	N	%	N	%
<15	1391	27.71	1293	28.09	747	28.75	763	28.89	794	30.25
15-150	3627	72.27	3308	71.87	1839	70.79	1866	70.66	1827	69.60
>150	1	0.02	2	0.04	12	0.46	12	0.45	4	0.15
Total	5019	100.00	4603	100.00	2598	100.00	2641	100.00	2625	100.00

A common characteristic of educator effectiveness measures that is often analyzed is the extent that they are stable over time, given it is unlikely that a teacher's effectiveness varies dramatically over time. Accordingly, we computed cross-year correlations for all three types of aggregated SGP—cohort SGP, baseline-referenced SGP, and SIMEX-corrected, baseline-referenced SGP (see Section 1.1 for definitions), aggregating by both the mean and median. We found that AGP were moderately stable across years. As shown in Table 3.2, the correlation between AGP from two adjacent school years ranged from .28 to .70 for ELA (Table 3.2a) and .43 to .71 for Math (Table 3.2b). For both ELA and Math, they were higher between 2011 and

2012 than between 2012 and 2013. As expected given that the mean SGP is more precise than the median, AGP based on medians have lower cross-year correlation. AGP based on the SIMEX-corrected SGP tended to have lower cross-year correlation than those derived from baseline or cohort SGP.

Table 3.2 also presents the cross-year correlation for the average prior test score, percentage economically disadvantaged, and counts of students for students linked to the teacher. These correlations are generally high, especially the percentage economically disadvantaged, indicating that teachers tend to teach in very similar conditions across adjacent school years. The consistency in students' teachers across years actually contributes to stability in AGP across years, as bias in the AGP that is correlated with student background variables will be correlated across years. Hence, lower cross-year correlation for the SIMEX-corrected AGP could be due to a reduction in the prior score bias in the AGP or to an increase in the estimation error.

Table 3.2a. Cross-year correlation of selected teacher-level statistics for ELA.

		2011 to 2012					2012 to 2013				
		Grade					Grade				
Statistic	Aggregation	4	5	6	7	8	4	5	6	7	8
Cohort SGP	Mean	0.49	0.47	0.67	0.54	0.53	0.40	0.38	0.52	0.50	0.47
	Median	0.43	0.41	0.63	0.48	0.46	0.35	0.34	0.48	0.47	0.41
Baseline SGP	Mean	0.50	0.45	0.70	0.59	0.51	0.41	0.36	0.57	0.56	0.41
	Median	0.45	0.40	0.65	0.53	0.45	0.37	0.32	0.54	0.51	0.36
SIMEX Baseline SGP	Mean	0.46	0.43	0.66	0.52	0.43	0.35	0.32	0.50	0.48	0.35
	Median	0.41	0.37	0.62	0.47	0.38	0.32	0.28	0.48	0.43	0.30
Prior Tests	Mean	0.73	0.74	0.85	0.85	0.86	0.75	0.75	0.85	0.84	0.86
Economically Disadvantaged	Mean	0.88	0.88	0.91	0.91	0.90	0.84	0.83	0.83	0.85	0.84
Count of students	Sum	0.75	0.80	0.78	0.76	0.79	0.62	0.64	0.72	0.74	0.72
Number of Teachers		2762	2764	1445	1367	1380	2762	2613	2575	1334	1220

Table 3.2b. Cross-year correlation of selected teacher-level statistics for mathematics.

		2011 to 2012					2012 to 2013				
		Grade					Grade				
Statistic	Aggregation	4	5	6	7	8	4	5	6	7	8
Cohort SGP	Mean	0.53	0.63	0.70	0.61	0.64	0.52	0.57	0.65	0.59	0.56
	Median	0.49	0.60	0.67	0.57	0.62	0.48	0.53	0.61	0.56	0.53
Baseline SGP	Mean	0.54	0.64	0.71	0.60	0.63	0.52	0.58	0.65	0.57	0.55
	Median	0.50	0.60	0.68	0.56	0.60	0.48	0.54	0.62	0.55	0.53
SIMEX Baseline SGP	Mean	0.50	0.62	0.70	0.57	0.62	0.48	0.56	0.62	0.55	0.53
	Median	0.46	0.59	0.66	0.53	0.59	0.43	0.53	0.59	0.54	0.50
Prior Tests	Mean	0.74	0.76	0.84	0.88	0.86	0.72	0.76	0.83	0.86	0.85
Economically Disadvantaged	Mean	0.89	0.88	0.91	0.90	0.90	0.83	0.83	0.82	0.81	0.81
Count of students	Sum	0.82	0.82	0.72	0.73	0.71	0.64	0.63	0.68	0.73	0.71
Number of Teachers		2443	2453	1255	1291	1288	2237	2125	1225	1201	1225

Given that the aim for this report is to investigate the correlation between AGP and mean prior achievement, to provide some context, we present these correlations for ELA and Math in the most recent year, 2013. Table 3.3a and Table 3.3b show these correlations for all three types of aggregated SGP and by grade-level for ELA and Math, respectively.

Table 3.3a. Correlations between AGP and mean prior achievement for ELA, 2013.

SGP Type	Stat	Grade				
		4	5	6	7	8
Cohort SGP	Mean	0.37	0.27	0.36	0.35	0.38
	Median	0.34	0.25	0.33	0.34	0.35
Baseline SGP	Mean	0.35	0.20	0.52	0.46	0.20
	Median	0.33	0.19	0.50	0.43	0.18
SIMEX-Baseline SGP	Mean	0.17	0.05	0.34	0.23	-0.15
	Median	0.15	0.05	0.35	0.20	-0.13
Number of Teachers		4140	3934	2033	1983	1960

Table 3.3b. Correlations between AGP and mean prior achievement for mathematics, 2013.

SGP Type	Stat	Grade				
		4	5	6	7	8
Cohort SGP	Mean	0.28	0.18	0.19	0.19	0.14
	Median	0.27	0.16	0.17	0.18	0.13
Baseline SGP	Mean	0.26	0.22	0.29	0.11	0.35
	Median	0.25	0.20	0.27	0.11	0.34
SIMEX-Baseline SGP	Mean	0.12	0.13	0.15	-0.06	0.25
	Median	0.11	0.11	0.13	-0.05	0.23
Number of Teachers		3628	3310	1851	1878	1831

First, Table 3.3 illustrates the issue of moderate, positive correlations between AGP and mean prior achievement. In general these correlations are consistent with relationships between value-added or AGP and background variables reported in other contexts. This table shows that the magnitude of this correlation depends on the type of SGP and, to some degree, the aggregation function. In order of weakest to strongest correlations between the AGP and mean prior achievement, the AGP are generally ranked: aggregated SIMEX-baseline SGP, aggregated cohort SGP, and aggregated baseline SGP. However, comparing across grade-levels and content areas, we see that the strength of these correlations and extent of the differences among the correlations for the different types of AGP depend on the unique characteristics of each grade-level by content area dataset.

These differences depend, in part, on the corresponding student-level correlations, which are presented in Table 3.4. For the cohort SGP, the correlations with student prior scores will be near 0 by construction. However, this does not hold for the baseline-referenced or SIMEX-corrected SGP. For instance, as shown in Table 3.3, for Grade 8, the correlations between mean prior achievement and mean SIMEX-baseline SGP (AGP) are approximately -.15 for ELA and +.25 for Math (values for medians are similar). The corresponding student-level correlations between SIMEX-baseline SGP and student prior scores are -.18 for ELA and +.04 for Math, as

shown in Table 3.4. These student-level correlations are nonzero, and, in this case, their signs correspond to those observed for the aggregated SGP.

Table 3.4a. Correlations between SGP and prior achievement for ELA, 2013.

SGP Type	Grade				
	4	5	6	7	8
Cohort SGP	-0.02	-0.01	0.00	0.00	0.00
Baseline SGP	-0.03	-0.06	0.07	0.03	-0.08
SIMEX-Baseline SGP	-0.14	-0.14	-0.01	-0.07	-0.18
Number of Students	113,657	114,416	114,982	114,748	112,812

Table 3.4b. Correlations between SGP and prior achievement for mathematics, 2013.

SGP Type	Grade				
	4	5	6	7	8
Cohort SGP	-0.02	0.00	0.00	0.00	0.00
Baseline SGP	-0.03	0.03	0.04	-0.06	0.10
SIMEX-Baseline SGP	-0.14	-0.04	-0.03	-0.14	0.04
Number of Students	112,768	112,700	113,833	113,989	112,172

We also compared each of the three types of SGP at the student-level within a given year. Figure 3.1 illustrates the pairwise differences using student records for grades 4, 7, and 8 for ELA in 2013. We chose these grade-levels to illustrate the range of differences in these pairwise comparisons. For grade 4, we observed the smallest differences among the three SGP types because both baseline- and cohort-referenced SGP are using the same number of prior scores—just prior Grade 3 scores. For grades 7 and 8, the figure clearly shows that the SIMEX-corrected and uncorrected baseline-referenced SGP are approximately the same, on average. However, these two baseline-referenced SGP differ substantially from the cohort-referenced SGP estimates

in grades 7 and 8, but they have average differences in opposite directions. In Section 4, we discuss important considerations in comparing baseline and cohort SGPs.

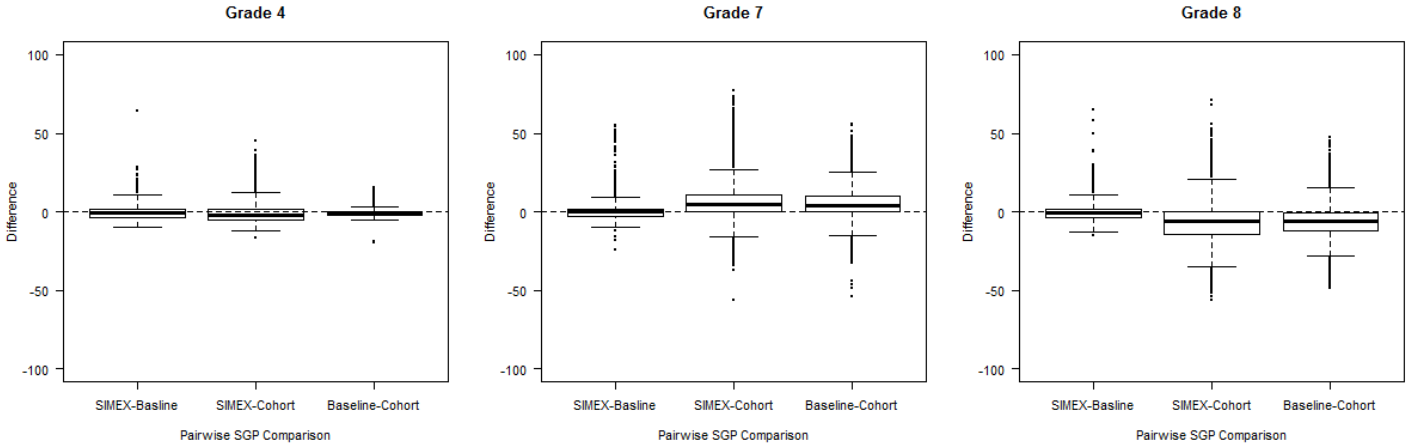


Figure 3.1. Comparing SGP types in select grade-levels for ELA, 2013.

3.1.2 Comparing Teachers across Time with Varying Classes

For the first of the three studies designed to isolate the effect of bias in AGP from the effect of teacher sorting, we compared the AGP from adjacent years for teachers in the state data whose students' average prior achievement changed substantially across years. Again, because the comparisons are made within the same teacher, teacher sorting should not contribute to variation in the AGP.

We stratified teachers by the degree to which the prior achievement of students entering their classrooms differed between two adjacent years. For teachers with data over adjacent years, we took the difference in the mean prior standardized achievement scores of their students and separated them into the following three groups:

1. Change in means is less than -.3 student test score standard deviations
2. Change in means is between -.3 and +.3 student test score standard deviations
3. Change in means is greater than +.3 student test score standard deviations

If AGP are biased, we expect the average differences in teacher AGP to increase by stratum in the order listed above.

We provide the average differences in AGP by stratum for 2011 versus 2012 in Figure 3.2 and for 2012 versus 2013 in Figure 3.3. We also provide corresponding tables of these average differences in Tables B.1 and B.2 in Appendix B. To increase readability of the figures, we present only the results for the mean-aggregations of each SGP type. We do not present those for the median SGP. The results for means and medians were similar.

Figure 3.2a. Cross-year change (average difference) in AGP for ELA by change in mean prior achievement for teachers for grades 4 to 8*: 2012 - 2011

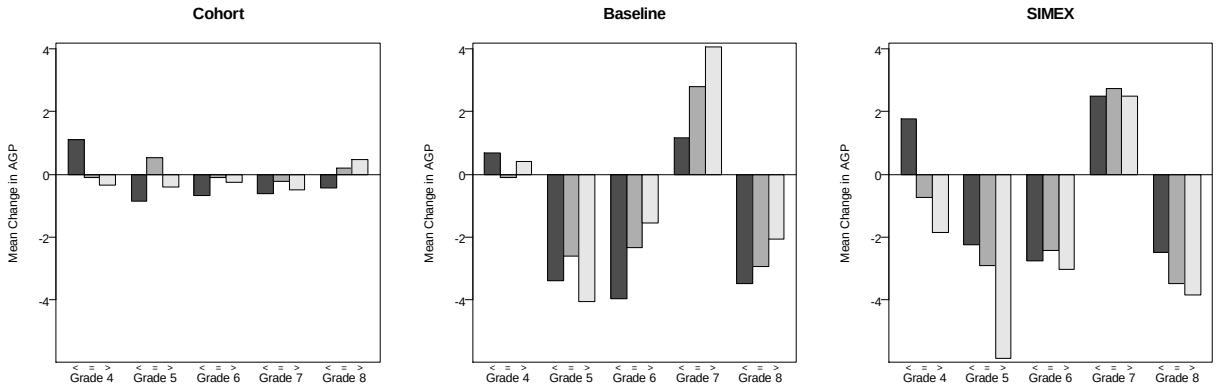


Figure 3.2b. Cross-year change (average difference) in AGP for mathematics by change in mean prior achievement for teachers for grades 4 to 8*: 2012 - 2011

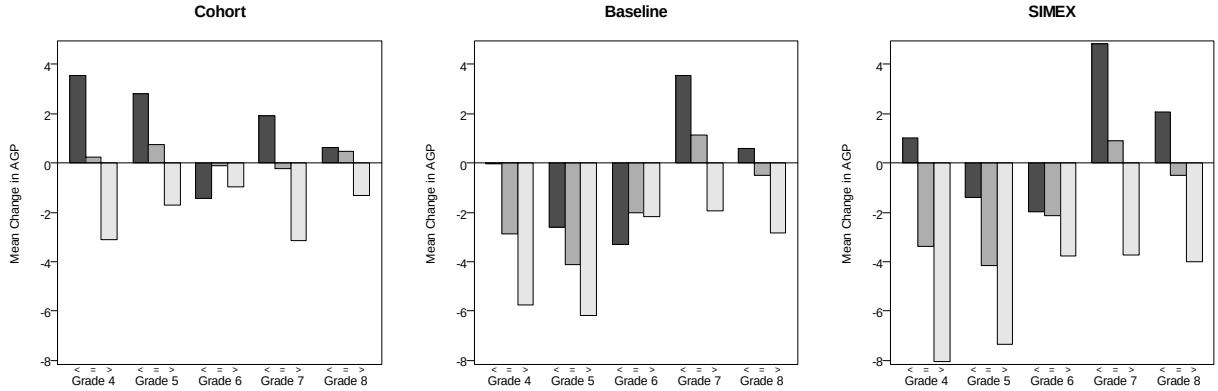


Figure 3.3a. Cross-year change (average difference) in AGP for ELA by change in mean prior achievement for teachers for grades 4 to 8*: 2013 - 2012

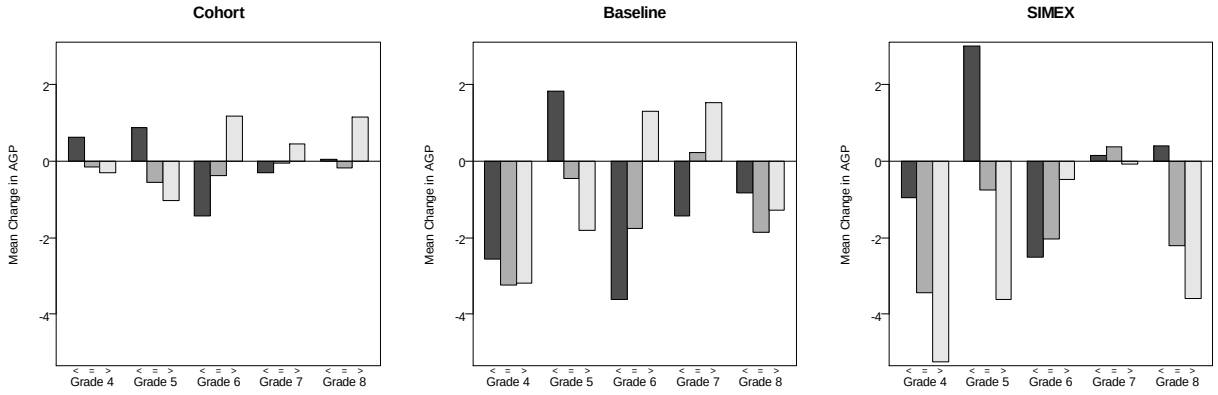
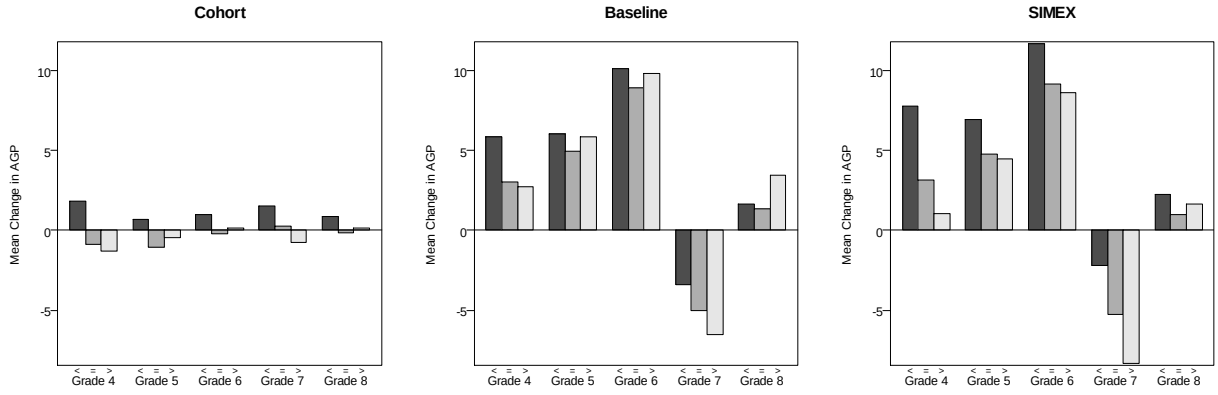


Figure 3.3b. Cross-year change (average difference) in AGP for mathematics by change in mean prior achievement for teachers for grades 4 to 8*: 2013 - 2012



*Note: Teachers are divided into three groups. The black bar is the difference in mean SGP for teachers whose students' mean prior achievement in year 2 was at least 0.3 SDs less than their students' means prior achievement in year 1 (difference in mean prior achievement, year 2 mean minus year 1 mean, was less than -0.3 SD). The dark gray bar is the difference in mean SGP for teachers whose students' mean prior achievement in year 2 was within 0.3 SD of their students' mean prior achievement in year 1 (difference in mean prior achievement, year 2 mean minus year 1 mean, was between +/- 0.3 SD). And the light gray bar is the difference in mean SGP for teachers whose students' mean prior achievement in year 2 was at least 0.3 SD greater than their students' means prior achievement in year 1 (difference in mean prior achievement, year 2 mean minus year 1 mean, was greater than +0.3 SD).

Figures 3.2 and 3.3 show that, in some cases, we do find that differences increase by stratum, such as for ELA grade 7 for baseline SGP in Figure 3.2a. However, this pattern is not

consistent across grade levels, content areas, or even SGP type. For example, the increasing mean differences by stratum for ELA grade 7 occurs only for the baseline SGP and not for the cohort or SIMEX-baseline SGP for which the mean differences fluctuate over the strata. In some cases, we observe the opposite pattern than representing bias, such as for the baseline and SIMEX-baseline AGP for Math grade 4 in Figure 3.3b. In general, the mean differences across the strata are larger (in absolute value) for the baseline and SIMEX-baseline AGP than the cohort AGP. However, the number of teachers at each grade-level within each of three strata varies, and, in some cases, the numbers of teachers is not large. For instance, there are only 86 grade 5 Math teachers who had students with much higher prior achievement (light gray bar) in their classes in 2013 than 2012 compared to 476 teachers who had much lower prior achievement (black bar) and 663 teachers with similar prior achievement students (dark gray bar) (see Table B.2b in Appendix B). Indeed, the maximum number of teachers within any of the strata is only about 1800 (Table B.1a). Accordingly, some of the odd patterns in mean AGP differences in Figures 3.2 and 3.3 could be due to the small samples.

This analysis did not find evidence of bias in AGP. However, we cannot control for any other differences across years for these groups of teachers or any differences among the teachers in the three strata, so the results should not be taken as strong evidence against bias either.

3.1.3 Adjusting AGPs

For the second analysis of this type, we pooled the AGP for teachers who had data for up to three school years and modeled variation in AGP within teachers as a function of variation in the average prior achievement of students and the percentage of students who are eligible for subsidized school meals. We again used the SGP calculated by the GaDOE to generate the AGP for this analysis. The idea behind this analysis is that by analyzing only the variation within the

teachers across years we remove the effects of teacher sorting because the relationship between teacher sorting and student growth can only be observed across teachers. Within teacher, teacher sorting cannot contribute to the relationship between annual AGP and average prior student achievement (e.g., Wooldridge, 2013). Thus, any remaining relationship between AGP and average prior student achievement must be due to other factors such as bias or possibly variation in teacher effectiveness that depends on the students in the class.[13]

In this analysis, we also calculated adjusted AGP. We averaged the AGP across years for each teacher and subtracted bias estimated from the within teacher model. We then calculated the correlation between the adjusted AGP and several aggregated student background variables, averaged over the three years, including student prior mean achievement, and the percentage of students eligible for subsidized school meals.

Tables 3.5 and 3.6 show these correlations when we define the teacher AGP as the mean-aggregated cohort SGP and as the mean-aggregated SIMEX-corrected, baseline-referenced SGP, respectively with the ELA results in Tables 3.5a and 3.6a and the Math results in Tables 3.5b and 3.6b. In the bottom portion of these tables, we provide the counts of teachers with 1, 2, or 3 years of data.[14]

For ELA, the adjusted AGP have weaker correlations with mean prior achievement (first rows of Tables 3.5a and 3.6a) than the unadjusted AGP (first row of Table 3.3a). Accordingly, subtracting out the bias, or dependence on mean prior achievement and percentage of economically disadvantaged students, mitigates the extent that teachers whose students come

[13] Variation in effectiveness that depends on the students is not the same as bias. Bias is an incorrect measure of the effectiveness. Variation in effectiveness that depends on the students reflects true differences in the effectiveness in which an individual teacher is more effective with some groups of students than others. Variation in effectiveness that depends on students is problematic for some uses of AGP because it would imply that inferences about a teacher would depend on what classes that teacher taught.

[14] We include the teachers with only 1 year of data in the within-teacher regressions as their data will contribute to the estimation of the coefficients of the predictors, mean prior achievement and percent economically disadvantaged.

into their classrooms with high achievement tend to obtain high AGP. However, the adjusted AGP still have moderate correlations with not only mean prior achievement but also several other student background variables, such as percentage economically disadvantaged and percentage Black students. Moreover, for Math, the adjusted AGP have as strong or stronger correlations with mean prior achievement (first rows of Tables 3.5b and 3.6b) than the unadjusted AGP (first row of Table 3.3b). Thus, this adjustment to the AGP does not necessarily reduce the extent that teachers with high mean prior achievement obtain high AGP. This could indicate that classroom level factors may be suppressing the correlation between teacher effectiveness and their students' background variables. However, given that the adjustment follows from a within-teacher model that relies of variation in within teachers in their students' background and this variation is very small in Math, the results may also be due, in part, to a lack of variation in AGP over time.

A possible limitation to this analysis is the strong correlation among the characteristics of teachers' students across years. This creates instability in the estimate of the adjustments. However, even with the instability of the adjustment it is unlikely that we would observe correlations as large as those in Tables 3.5 and 3.6 if there were no teacher sorting.

Table 3.5a. Correlation of adjusted Mean SGP with averaged student background variables for ELA

	Grade				
Aggregated Variable	4	5	6	7	8
Mean Prior Score (ELA)	0.29	0.26	0.19	0.27	0.26
% Female	0.06	0.06	0.05	0.10	0.13
% Limited English Proficient	0.05	-0.06	0.00	0.02	0.01
% Students w/Disabilities	-0.04	-0.04	-0.08	-0.15	-0.11
% Economically Disadvantaged	-0.21	-0.19	-0.23	-0.15	-0.21
% Black	-0.18	-0.18	-0.18	-0.13	-0.17
% White	0.09	0.17	0.14	0.03	0.13
% Hispanic	0.06	-0.02	0.01	0.10	0.05
Number of Teachers w/1 yr	3520	2997	1784	1899	1691
Number of Teachers w/2 yrs	2013	1841	896	847	798
Number of Teachers w/3 yrs	1794	1855	986	917	968

Table 3.5b. Correlation of adjusted Mean SGP with averaged student background variables for Mathematics

	Grade				
Aggregated Variable	4	5	6	7	8
Mean Prior Score (ELA)	0.35	0.23	0.15	0.31	0.16
% Female	0.01	0.03	0.04	0.12	0.12
% Limited English Proficient	0.06	0.01	0.04	0.06	0.04
% Students w/Disabilities	-0.04	-0.01	-0.03	-0.11	-0.06
% Economically Disadvantaged	-0.25	-0.16	-0.23	-0.22	-0.13
% Black	-0.26	-0.11	-0.20	-0.23	-0.12
% White	0.16	0.05	0.14	0.13	0.09
% Hispanic	0.07	0.03	0.07	0.11	0.04
Number of Teachers w/1 yr	3341	2678	1511	1538	1400
Number of Teachers w/2 yrs	1878	1617	802	823	709
Number of Teachers w/3 yrs	1519	1578	883	879	939

Table 3.6a. Correlation of adjusted Mean SIMEX-corrected SGP with averaged student background variables for ELA

	Grade				
Aggregated Variable	4	5	6	7	8
Mean Prior Score (ELA)	0.27	0.26	0.18	0.29	0.19
% Female	0.05	0.06	0.04	0.12	0.15
% Limited English Proficient	0.04	-0.05	0.01	0.01	0.08
% Students w/Disabilities	-0.03	-0.04	-0.07	-0.20	-0.14
% Economically Disadvantaged	-0.20	-0.17	-0.19	-0.16	-0.09
% Black	-0.16	-0.16	-0.14	-0.11	-0.08
% White	0.08	0.15	0.11	0.03	0.04
% Hispanic	0.05	-0.03	0.00	0.08	0.03
Number of Teachers w/1 yr	3520	2997	1784	1899	1691
Number of Teachers w/2 yrs	2013	1841	896	847	798
Number of Teachers w/3 yrs	1794	1855	986	917	968

Table 3.6b. Correlation of adjusted Mean SIMEX-corrected SGP with averaged student background variables for Mathematics

	Grade				
Aggregated Variable	4	5	6	7	8
Mean Prior Score (Math)	0.35	0.33	0.18	0.31	0.19
% Female	0.00	0.04	0.06	0.11	0.13
% Limited English Proficient	0.06	0.01	0.04	0.05	0.04
% Students w/Disabilities	-0.05	-0.03	-0.08	-0.11	-0.11
% Economically Disadvantaged	-0.23	-0.19	-0.20	-0.23	-0.15
% Black	-0.25	-0.15	-0.18	-0.24	-0.13
% White	0.15	0.08	0.11	0.15	0.10
% Hispanic	0.07	0.04	0.07	0.09	0.02
Number of Teachers w/1 yr	3341	2678	1511	1538	1400
Number of Teachers w/2 yrs	1878	1617	802	823	709
Number of Teachers w/3 yrs	1519	1578	883	879	939

3.1.4 Accelerated versus Regular Math Tracks

For the last of these three studies of constant teachers-varying students, we used data from a special study for three large Georgia school systems for middle school grades. Specifically, we used a sample of teachers who taught students in advanced or accelerated Math classes and

during the same school year also taught students in the regular Math tracks. The students in the advanced classes had substantially higher prior achievement than the students in the other classes. Thus, if measurement error is creating spurious correlation between student background variables and AGP, then the AGP should be higher for the accelerated classes. However, if we compare the AGP for the accelerated class to the AGP for the regular class from the same teacher, teacher sorting cannot be source of the correlation. Hence, we can estimate the contribution of measurement error bias to differences in the AGP by comparing the AGP for the accelerated and regular classes from the same teacher.

We conducted this analysis using data from 123 middle school math teachers from three large school systems in Georgia where we had detailed data on course titles and numbers that allowed us to determine whether a course was a regular or accelerated Math class. We used up to 4 years of prior Math scores in the calculation of the SGP used to generate the AGP. We also repeated the analysis using fewer years of prior test scores and using the baseline SGP. All analyses were conducted in the R statistical computing environment (R Core Team, 2014) using the SGP package (Betebenner et al., 2014) to calculate SGP. We use the Georgia-specific parameters provided with the SGP package to calculate baseline SGP.

In this sample, students in the accelerated courses score .51 standard deviation units higher on average than the students in the regular courses. Moreover, the AGP were correlated with the prior achievement of the students in the classes. However, for AGP based on SGP calculated with up to four years of prior scores, across teachers, the average of the difference between each teacher's AGP for his or her accelerated class and the AGP for his or her regular class was just 0.004. That is, when comparing AGP from the accelerated class to the regular class for the same teacher there was no evidence of the AGP for accelerated classes being higher.

There is no evidence of a spurious positive correlation between AGP and average prior test scores, within a teacher. However, when we averaged each teacher's AGP from both his or her accelerated and regular classes and compared this average with the average prior test scores for his or her classes, we find the is correlation is .22. That is, teachers who teach students with higher prior test scores have higher AGP across both their classes. Because we have no evidence from the within teacher comparisons of bias, positive correlation between teachers' average AGP and prior test score is consistent with teacher sorting, with more effective teachers being assigned classes with higher achievement.

3.2 Changing the Number of Prior Years

As discussed previously, using multiple test scores to create peer groups for students can reduce the impact of measurement error in prior scores on their SGP and subsequently on their teacher's (or leader's) AGP. Thus, if the correlation between AGP and the average student prior test scores or other aggregate student characteristics is the result of bias due to measurement error, we can expect to see the correlation decrease when we use more prior year tests in the calculation of the SGP used in the AGP. To see if this was the case, we used data on 870 middle school mathematics teachers and their 48,717 students from three large Georgia school systems. The data contained up to five years of language arts, math, reading, science, and social studies CRCT scores for the 2011 and four proceeding school years. Because social studies scores were not available for all years, we include only language arts, math, reading, and science scores in our analysis. All the students in these data completed all four tests in a year in which they completed any tests. That is if a student has a math test score in any year, he or she has scores from all four subjects.

Using the SGP package (Betebenner et al., 2014) in R, we calculated 16 separate SGP for each student using from 1 to 16 prior scores to define students with the same test score history. For each set of SGP, we calculated the mean SGP for each teacher by grade-level. Teachers who taught students at multiple grade levels will be in the analysis at each grade level. We estimated the correlation between each of the AGP and the average most-recent, prior year, math test scores for the teacher's students, and the percentage of those students eligible for subsidized meals. We report the result for eight SGP: the four SGP calculated using from one to up to four years of prior math tests and the four SGP calculated using from one up to four years of all four subject test scores.

The results are presented in Tables 3.7 (average prior test scores) and 3.8 (percent eligible for subsidized meals). In both tables, the results clearly show a pattern of the correlation between the AGP and the student background variables decreasing (in absolute value) as the number of tests increases. For a given number of years or prior scores, the correlation between AGP and student background variables is always smaller when all four tests are used in calculating the SGP than when only math tests are used. Two years of prior scores on all four tests, yields correlations that are typically about equal to those from AGP calculated with four prior math tests.

These results clearly suggest that measurement error is contributing to the correlation between AGP and student background variables. Using more tests, and in particular tests from multiple subjects, in the calculation of the SGP can reduce the correlation. Given that baseline SGP are restricted to two prior years of test scores, the potential for using tests of other subjects to reduce measurement error bias warrants further investigation. In particular, the standard errors of the SGP need to be explored because modeling with many tests might add to the estimation

error in the SGP. These results do not contradict our previous finding of no evidence of measurement error contributing to spurious differences between the regular and advanced classes. In that analysis, we used four prior test scores when estimating the SGP. When we reduced the number of tests used in the calculation of the SGP, we find some differences between the AGP for the advanced and regular classes for the same teacher.

Table 3.7. Correlation between average prior student test scores and AGP calculated with varying numbers of prior test scores by student's grade level.

		Number of Prior Years			
Grade	Scores	1	2	3	4
6	Math Only	.37	.29	.24	.23
	All 4 Subjects	.29	.23	.20	.19
7	Math Only	.28	.16	.14	.12
	All 4 Subjects	.20	.11	.09	.07
8	Math Only	.27	.19	.16	.14
	All 4 Subjects	.22	.15	.14	.13

Table 3.8. Correlation between percentage of students eligible for subsidized meals and AGP calculated with varying numbers of prior test scores by student's grade level.

		Number of Prior Years			
Grade	Scores	1	2	3	4
6	Math Only	-.48	-.42	-.37	-.35
	All 4 Subjects	-.32	-.28	-.25	-.24
7	Math Only	-.35	-.29	-.28	-.25
	All 4 Subjects	-.26	-.20	-.20	-.19
8	Math Only	-.29	-.24	-.22	-.20
	All 4 Subjects	-.24	-.2	-.19	-.19

4. Cohort vs. Baseline SGP

The GaDOE has chosen to use baseline SGP for its educator evaluation. Accordingly, we consider this SGP estimator specifically here and how it compares against the cohort SGP. Baseline SGP, as implemented by the GaDOE, use, at a minimum, a baseline of two cohorts of students to calculate the quantile function used in the SGP calculations and then every year the

percentiles of the current score distribution for students with equal prior scores are calculated using the same baseline cohort quantile functions. The motivation for using baseline SGP is that these values are not normative by construction. If student growth is improving relative to the baseline cohort, the mean (baseline) SGP can be greater than 50, whereas the mean cohort SGP will always be 50.

However there are also drawbacks to using the baseline SGP. First, as demonstrated by our analysis, the correlation between AGP and aggregate student background variables decreases as more prior years of test scores are used in the calculation of the SGP, but the baseline SGP used by the state are restricted to two prior years of scores. Consequently, we found that the aggregate baseline SGP often had higher correlations with student background variables than the other approaches.

Second, baseline SGP rely on very strong assumptions. The underlying idea relies on a counterfactual—a quantity we cannot directly observe which could occur under alternative conditions. For baseline SGP, we have two cohorts of students, the target cohort for whom we want to calculate SGP and the baseline cohort who we want to use as a reference. The baseline cohort precedes the target cohort. For each student in the target cohort, we want to know the percentile rank of that student's current year score relative to students in the baseline cohort who had the same prior achievement history. However, the students in the target cohort took one set of current and prior year tests and students in the baseline cohort took a different set of current and prior year tests. Thus, we cannot identify students in the baseline cohort who had the same prior year achievement on the same test as a student in target cohort because no students in the baseline cohort took the tests of the student in the target cohort. Furthermore, we cannot compare the current year scores of students in the target cohort to scores on the same test for students in

the baseline cohort because students in the baseline cohort did not take the same current year test as the students in the target.

We must rely on test equating to use the baseline cohort to calculate the SGP of students in the target cohort. Moreover, we must assume that test equating is sufficient so that the percentile rank of a current year score for a given achievement history among students in the baseline cohort on the baseline cohort tests equals the percentile rank of the current year score for a given achievement history among students in the baseline cohort had they taken the tests taken by the students in the target cohort. Test equating does not typically check on such quantities; rather, test equating typically involves comparing the score distribution in the current administration to the score distribution in the previous year's administration. That is, equating tests attempts to set scores, so that a score on a grade 4 test in 2014 can be treated as equal to the same score on the grade 4 test in 2013 had the same student taken both tests. The same holds for other grades and years. This is not the same as setting scores so that the distribution of grade 4 test scores in 2014 given a grade 3 test scores from 2013 would have the same distribution as the 2013 grade 4 test scores given the same grade 3 test score from 2012, had the same student taken both sets of tests. However, the use of baseline SGP assumes such equivalence of test score distributions must hold between the tests taken by the baseline and target cohorts. If these strong assumptions (regarding maintaining conditional distributions) fail, then baseline SGP could give distorted pictures of student growth. Accordingly, if the GaDOE is interested in using baseline SGPs, they should verify these assumptions (assuming they have not already done so).

The final shortcoming of the baseline SGP is the somewhat suspicious patterns in the SGP across grade levels and years. As seen in Figure 3.1, the difference between baseline and cohort SGP change across grades. For grade 4 baseline and cohort SGP look very similar. In

grade 8, the average difference is -6.34 in 2013. However the distribution in the differences between baseline and cohort SGP on average shifts down each year for grade 8 students so the mean differences are -1.61 and -4.70 for 2011 and 2012 (see Table B.3 in Appendix B). Since the mean of the cohort SGP is 50 each year, the average baseline SGP is decreasing each year. Taken on face-value, this result implies that across the state grade 8 students grew substantially less in 2013 than they did during the baseline. Similar though somewhat less dramatic downward shifts exists for grades 5 and 6. The means are -0.90, -3.93 and -3.87 for grade 5 for 2011, 2012, and 2013 and the corresponding values for grade 6 are -0.20, -.2.36, and -3.79. However, the difference grew each year in grade 7 – which, if taken on face value, indicates students were growing more in grade 7 than during the baseline. The grade 7 means are 1.98, 5.18, and 5.33 for 2011, 2012, and 2013. These results are possible but they are suspicious. Also, the largest differences between cohort and baseline SGP are quite large, which suggests instability in the models. Such differences might occur if there were differences in equating, the tests themselves (e.g., test blueprints, item types), conditions of measurement (e.g., changes in accommodations or mode of delivery of test), and alignment between the assessment and instruction (e.g., a lag between adoption of new content standards and administration of assessments aligned to new content standards) from year to year.

The counterfactual that the baseline SGP attempt to estimate would be very valuable, so there is clearly strong motivation for attempting to use these measures. However, the use of the baseline SGP to estimate the counterfactual relies on very strong and untested assumptions. The state might want to consider ways to monitor the baseline SGP and its test equating to identify possible anomalies.

5. Implications

Our empirical analyses revealed that the moderate correlation between AGP and student prior achievement that motivated this study persist in data from the 2013 school year. As shown in Table 3.3a, the correlation between the state's preferred baseline AGP and average prior achievement was as large as .52 and exceeded .30 for three of five grade levels for ELA. Although smaller in mathematics, this correlation did exceed .30 for eighth grade teachers (see Table 3.3b). The source of this correlation is unknown. One possible source is error in the SGP that results in spurious correlation between AGP and the background characteristics of leaders or teachers' students. Another is sorting in which more effective teachers or leaders are assigned to classes or schools of more advantaged students with higher prior achievement. We cannot fully determine the source, but our investigations shed some light on the possible sources.

First, our analytic results demonstrate that measurement error in the test scores used to calculate SGP and subsequently AGP can result in bias in the AGP that is correlated with students' average prior achievement. Thus, bias in AGP due to measurement error is one potential source of the observed correlation. Consequently, a correction for measurement error may be useful for reducing the correlation if it is in fact due to such a bias. Given the complexity of the process used to calculate AGP, GaDOE chose the SIMEX measurement error correction.

Our analytic derivations suggest that, under ideal conditions, using SIMEX to correct for measurement error in the percentiles can remove the bias in SGP that is correlated with student prior achievement. However, this fix does not correct for measurement error in the current scores. Current score measurement error can result in compressing SGP relative to true SGP toward 50 and could potentially compress AGP for teachers and leaders. Moreover, the ideal conditions for SIMEX require using the correct projection function for the extrapolation step and

adding additional measurement errors with the correct variance during the simulation step of the SIMEX procedure. The GaDOE used a linear projection function which is unlikely to be correct and the correct variance to use is unknown.

Although, SIMEX was not applied under ideal conditions, we believe that using the correction is likely to reduce the bias and its contribution to the correlation between AGP and mean prior achievement. The potential of SIMEX to mitigate bias has been demonstrated in the literature in various contexts. We also note that the SIMEX-baseline AGP have weaker correlation with prior achievement in the 2013 Georgia data than the other estimates. The use of SIMEX does, however, have tradeoffs. Using SIMEX can add to the random statistical error in AGP. Even though random statistical errors will not yield systematic relationships between AGP and student prior achievement, they can lead to errors in classifications of educators and instability across time. The GaDOE attempted to balance between the potential to reduce bias and the potential to increase random statistical error in AGP with its implementation of SIMEX. For example, it used a linear projection function and estimated corrected percentiles rather than using a more complex projection function or applying the entire AGP process to the simulated data. Given this balance, the GaDOE might want to work to develop standard error estimates for the SIMEX-baseline AGP, to monitor stability across years and to continue to compare the results to baseline and cohort AGP to check that SIMEX correction continues to reduce correlation with prior achievement and that it does not increase the random statistical error substantially.

Given that SIMEX appears to have some benefits but questions remain about the accuracy of the method, the GaDOE might also commission additional simulation studies to understand better how much bias SIMEX can remove under some circumstances. Such studies

should build on the work by Shang et al. (forthcoming) by generating data under different distributions and testing alternative SIMEX approaches. Because of questions about the best way to approximate the variance of measurement error for the simulation step and the extrapolation function, the GaDOE might also conduct sensitivity analyses using its test score data to determine how sensitive results are to different extrapolation functions or methods for simulating additional measurement errors.

Our empirical results find almost no relationship between variation in AGP across years or classes for the same teacher and variation in average student prior achievement in the corresponding years or classes. Also, when we adjust the AGP for aggregate student background variables, we find correlations between the adjusted AGP and average prior achievement is nearly as large or even larger than the correlations between unadjusted AGP and average prior achievement. These results suggest that there is some amount of teacher sorting that exists among teachers in Georgia.

Our results do not suggest GaDOE make any immediate and dramatic changes to their methods but additional analyses might be useful for continued improvement of the system. GaDOE might explore using more test scores, such as tests from multiple subjects, when calculating SGP. GaDOE might consider using adjusted AGP like those in Section 3.1.3. If AGP contain bias that is correlated with student backgrounds, this method could possibly remove it. Such a method would create some logistical problems in combining data across years and conducting additional modeling. Also, as discussed in Section 3.1.3, variation in average student characteristics among classes from the same teacher within year or across years is limited and this could add random statistical errors in adjusted AGP. Continued exploration of this method, including methods for calculating the standard errors of the adjusted AGP might prove useful.

The GaDOE might also explore some alternative methods to calculating the percentile ranks of current achievement among students with the same prior achievement histories. For example, they could use item responses to model the bivariate or multivariate distribution of true scores and from this distribution estimate the desired percentile ranks. This might involve extending item response theory models (Hambleton & Swaminathan, 1985) typically used with one year of test scores for scaling scores, to include items from tests from multiple years. Alternatively, other statistical methods for estimating percentile ranks could be used. Monroe, Cai and Choi (2014) provide an example of such an approach, and Lockwood and Castellano (forthcoming) discuss additional methods using a similar approach. Lockwood and Castellano (forthcoming) also suggest using other statistical methods for estimating the cumulative distribution and percentile ranks for the current score conditional on the past scores. For example, the problem of finding ranks for test scores is analogous to the analysis of discrete time to event data and well-established statistical methods for that problem (Singer & Willett, 2003) may provide an alternative approach to SGP.

The use of baseline SGP relies on strong assumptions about the equating of scores. It is not clear how the GaDOE can directly test the underlying assumptions of this approach using its existing data. It may be reasonable to assume that the distribution of student achievement growth does not change dramatically from one cohort to the next across school years. That is, the growth distribution for grade 8 students in 2013 is most likely not dramatically different than distributions for grade 8 students in 2012 or 2014. If the SGP distributions show large swings from year to year this might be an indication of a problem with equating or scaling.

SGP and AGP calculations are complex. The analyses presented here explore some of the facets of these measures but not all of them. The research community is actively investigating

SGP and AGP and generating new results about the properties of the methods and improvements to them. The GaDOE should continue to monitor those results and continue to conduct its own analyses as part of its ongoing work to develop and maintain accurate measures of teacher effectiveness.

Appendix A
Analytic Derivations

A.1. Background

In this appendix, we derive the expected value of a student's observed SGP conditional on that student's true current and prior year achievement. For a given student, bias equals the expected SGP minus the true SGP. Conditioning on the student is the same as conditioning on the the student's true level of achievement. The derivations assume that true achievement for two years is bivariately normally distributed and that measurement errors are also normally distributed. Normality of true scores and errors may not hold for test scores, but the results under normality are tractable and provide intuition into the functioning of SGPs that will apply more broadly. We also assume that the samples used to estimate the quantile functions are large so that statistical error in those can be ignored.

A.2. Notation

We let (X_1, X_2) denote the true prior and current year achievement for a student. Let (X_1, X_2) be distributed bivariate normal with zero means and variances $\sigma^2_{X_1}$ and $\sigma^2_{X_2}$ and covariance $\sigma_{X_{12}}$. This means that conditional on X_1, $X_2 \sim N(\beta X_1, \sigma^2_\epsilon = (1-\beta^2\sigma^2_{X_1}/\sigma^2_{X_2})\sigma^2_{X_2})$, where $\beta = \sigma_{X_{12}}/\sigma^2_{X_1}$ and the correlation between X_1 and X_2 equals $\beta\sigma_{X_1}/\sigma_{X_2}$. Let $\epsilon = X_2 - \beta X_1$.

We assume that each year the observed scale score Y_1 or Y_2 equals true achievement plus normal measurement error U_1 or U_2. The variance of the measurement error is $(1-\lambda_1)\sigma^2_{Y_1}$ for year 1 and $(1-\lambda_2)\sigma^2_{Y_2}$ for year 2. The test reliability is λ_1 and λ_2 in years 1 and 2, respectively. The distribution of Y_2 given X_1 is $N(\beta X_1, \sigma^2_\epsilon + (1-\lambda_2)\sigma^2_{Y_2})$ and the distribution of Y_2 given Y_1 is $N(\lambda_1\beta Y_1, \sigma^2_\epsilon + (1-\lambda_2)\sigma^2_{Y_2} + \lambda_1(1-\lambda_1)\beta^2\sigma^2_{Y_1})$.

Let $\Phi(z)$ equal the standard normal cumulative distribution function (CDF) evaluated at z. The p^{th} quantile (p between zero and 1) of the conditional distribution of X_2 given X_1 equals $p(X_1) = z_p\sigma_\epsilon + \beta X_1$, z_p is the p^{th} quantile of the standard normal distribution, that is $\Phi(z_p) = p$.

A.3. Expected Value of SGP

The true SGP for a student, which we denote SGP_0, equals the percentile rank of X_2 conditional on X_1, so that

$$SGP_0 = \Phi\left(\frac{X_2 - \beta X_1}{\sqrt{\sigma^2_\epsilon}}\right) = \Phi\left(\frac{\epsilon}{\sqrt{\sigma^2_\epsilon}}\right). \quad (1)$$

Observed Current and True Prior Achievement

If we observed X_1 but calculated the SGP at Y_2 rather than X_2, then the SGP is

$$SGP_1 = \Phi\left(\frac{X_2 - \beta X_1 + U_2}{\sqrt{\sigma_\epsilon^2 + (1-\lambda_2)\sigma_{Y_2}^2}}\right). \tag{2}$$

The goal is to derive the formula for the expected value of SGP_1 conditional on (X_1, X_2) and to compare this with SGP_0 to determine the bias. Conditional on X_1 and X_2 only U_2 is random, and it is normally distributed with mean zero and variance $(1-\lambda_2)\sigma_{Y_2}^2$. Given standard results (see for example, Zacks, 1981),

$$E[SGP_1|X_1, X_2] = \Phi\left(\frac{\epsilon}{\sqrt{\sigma_\epsilon^2 + 2(1-\lambda_2)\sigma_{Y_2}^2}}\right). \tag{3}$$

The numerator in the argument to the standard normal CDF in Equation 3 equals the numerator of the argument to the normal CDF for the true SGP in Equation1. However, the denominator of the argument in Equation 3 is larger than the denominator of the argument of the true SGP. Hence, SGP_1 is biased. It underestimates the SGP for students who are above the median and overestimates the SGP for students who are below the median. The size of the bias depends on ϵ. However, the expected value of SGP_1 preserves the ranks of students. Measurement error compresses the expected value of the SGP so that any percentile $p > 50$, fewer than p percent of the students will have expected SGP greater than p, and for $p < 50$, fewer than p percent of the students will have expected SGP less than p. The proportion below p will be $\Phi\left(\frac{z_p\sqrt{\sigma_\epsilon^2+(1-\lambda_2)\sigma_{Y_2}^2}}{\sqrt{\sigma_\epsilon^2}}\right)$. The distribution of expected SGP will be somewhat concentrated at the median rather than uniformly distributed. This is a problem if decisions are made based on the actual values of the SGP in an absolute way. For example, if observed SGP_1 were used to provide services for students with values less than .25, then some students intended to receive services because their true SGP is less than .25 would tend not to receive services.

True Current and Observed Prior Achievement

If we observed X_2 and Y_1, then the observed SGP is

$$SGP_2 = \Phi\left(\frac{\beta X_1 + \epsilon - \lambda_1\beta Y_1}{\sqrt{\sigma_\epsilon^2 + \lambda_1(1-\lambda_1)\beta^2\sigma_{Y_1}^2}}\right)\Phi\left(\frac{\epsilon + U_2}{\sqrt{\sigma_\epsilon^2 + \lambda_1(1-\lambda_1)\beta^2\sigma_{Y_1}^2}}\right). \tag{4}$$

Measurement error biases our estimate of the conditional distribution of the current year achievement given the prior achievement, so that numerator of the argument to Φ includes Y_1 because the slope, β, is

attenuated. The expected value of SGP_2 equals:

$$E[SGP_2|X_1, X_2] = \Phi\left(\frac{(1-\lambda_1)\beta X_1 + \epsilon}{\sqrt{\sigma_\epsilon^2 + \lambda_1(1-\lambda_1)\beta^2\sigma_{Y_1}^2 + \lambda_1^2(1-\lambda_1)\beta^2\sigma_{Y_1}^2}}\right) \quad (5)$$

$$= \Phi\left(\frac{(1-\lambda_1)\beta X_1 + \epsilon}{\sqrt{\sigma_\epsilon^2 + \lambda_1(1-\lambda_1^2)\beta^2\sigma_{Y_1}^2}}\right). \quad (6)$$

Both the numerator and the denominator of the argument to the CDF differ from the argument to Φ in Equation 1. For students with X_1 greater than the mean, $(1-\lambda_1)\beta X_1 > 0$, and they will rank higher than they would have if we had correctly estimated the conditional mean for the X_2 distribution. The expected SGPs are also more concentrated; this will exacerbate the distortion from the bias in the conditional mean.

Observed Current and Observed Prior Achievement

If we estimate SGP with Y_1 and Y_2, we obtain

$$SGP_3 = \Phi\left(\frac{\beta X_1 + \epsilon + U_2 - \lambda_1\beta Y_1}{\sqrt{\sigma_\epsilon^2 + (1-\lambda_2)\sigma_{Y_2}^2 + \lambda_1(1-\lambda_1)\beta^2\sigma_{Y_1}^2}}\right), \quad (7)$$

and its expected value conditional on the true achievement levels is

$$E[SGP_3|X_1, X_2] = \Phi\left(\frac{(1-\lambda_1)\beta X_1 + \epsilon}{\sqrt{\sigma_\epsilon^2 + 2(1-\lambda_2)\sigma_{Y_2}^2 + \lambda_1(1-\lambda_1^2)\beta^2\sigma_{Y_1}^2}}\right). \quad (8)$$

The relationship of this value to the true SGP is indeterminant and depends on ϵ, the test reliability, and the correlation between X_1 and X_2.

SIMEX Measurement Error Correction

Suppose we use SIMEX to correct for measurement error in the estimated quantiles. To do so, we would simulate data with additional measurement error with variance γ times variance of U_1 added to every observed value of Y_1 to obtain $Y_{1\gamma}$. We would then use the standard SGP estimation method to estimate the quantiles for observed values of $Y_{1\gamma}$. We would repeat this many times for a given value of γ and then repeat the entire process for multiple γ values. Finally, we would model the average quantiles at each value of γ as a function of γ and project back to $\gamma = -1$ or no measurement error. For given value of γ and given observed prior score Y_1, the average of the estimated quantiles approximately equals[1]

$$p_\gamma(Y_1) = z_p\sqrt{\sigma_\epsilon^2 + (1-\lambda_2)\sigma_{Y_2}^2 + (1-\lambda_\gamma)\lambda_\gamma\beta^2\sigma_{Y_1}^2} + \lambda_\gamma\beta E[Y_{1\gamma}|Y_1], \quad (9)$$

where $Y_{1\gamma}$ are the simulated scores with extra noise. The reliability of the scores with added measurement error is $\lambda_\gamma = \sigma_{X_1}^2/(\sigma_{X_1}^2 + (1+\gamma)\sigma_{U_1}^2) = 1/[1 + (1+\gamma)(1-\lambda_1)]$. Clearly this value converges to 1 as γ goes

[1]The average will equal $p_\gamma(Y_1)$ exactly as the number of simulated data sets gets large.

to -1. Also, $E[Y_{1\gamma}|Y_1] = Y_1$. Thus, the average of the estimated quantiles is a nonlinear function of γ that converges to $z_p\sqrt{\sigma_\epsilon^2 + (1-\lambda_2)\sigma_{Y_2}^2} + \beta Y_1$ as γ goes to -1. This is an unbiased estimate of the quantile based on the true prior scores which equals $z_p\sqrt{\sigma_\epsilon^2 + (1-\lambda_2)\sigma_{Y_2}^2} + \beta X_1$. However, it does not equal the true quantile.

If the SGP is then calculated by comparing Y_2 to these corrected quantiles, the SGP is equals the quantile rank of Y_2 in a normal distribution that has mean βY_1 and variance $\sigma_\epsilon^2 + (1-\lambda_2)\sigma_{Y_2}^2$. This yields the SIMEX corrected SGP of

$$SGP_4 = \Phi\left(\frac{\beta X_1 + \epsilon + U_2 - \beta Y_1}{\sqrt{\sigma_\epsilon^2 + (1-\lambda_2)\sigma_{Y_2}^2}}\right), \tag{10}$$

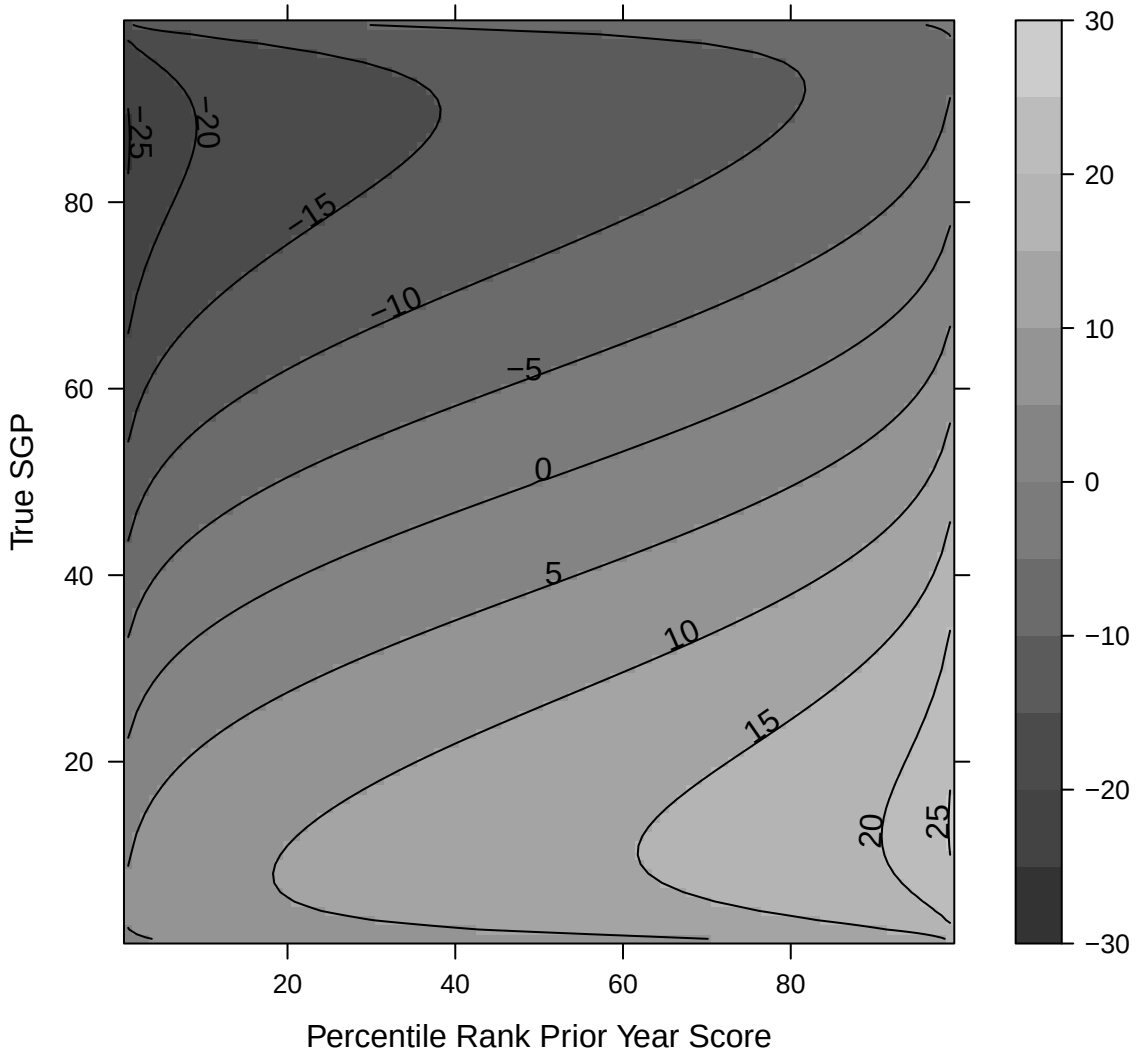
and its expected values conditional on the true achievement levels is

$$E[SGP_4|X_1, X_2] = \Phi\left(\frac{\epsilon}{\sqrt{\sigma_\epsilon^2 + 2(1-\lambda_2)\sigma_{Y_2}^2 + (1-\lambda_1)\beta^2\sigma_{Y_1}^2}}\right). \tag{11}$$

The argument in Φ has the same numerator as the argument in Φ for the true SGP, but the denominator in the argument in Equation 11 is too large. Consequently the SIMEX corrected SGP, SGP_4 is biased. However, it will rank order students correctly, although the expected values are more compressed than if we had actually estimated the SGP with X_1, that is they are more compressed than the expected value of SGP_1 (Equation 3).

Figure A.1 demonstrates the potential size of the bias in SGP_3. The formula for SGP_3 can be simplified to involve only standard normal random variables, the correlation between the current and previous year observed scores, and the reliability of the current and prior year tests. We used values of .775 for the correlation between the current and prior year test scores and .87 for the reliability of both the current and prior year test. These value are consistent with historical values for Georgia mathematics and ELA test scores. As shown in the figure, the bias can range anywhere from zero for students with median prior achievement and a true SGP of 50 to more than +/-20 percentile points for students with extremely low prior true score achievement and very high true growth or students with very high prior true score achievement and low true growth.

Figure A.1: Bias (gradient shading) in the observed SGP (SGP_3) due to measurement error in the prior and current year test scores as a function of the true SGP and the percentile rank of the prior score. Calculations assume that the correlation between the observed current and previous year scores was 0.775 and the reliability of both the current and prior year test was 0.87. Values of the correlation and reliability were chosen to be consistent with historical values form Georgia mathematics and ELA test score data.



Appendix B
Supplemental Tables

Tables B.1 to B.2 correspond to Section 3.1.2.

Table B.1a. Cross-year change (average difference) in AGP for ELA: 2012 - 2011

SGP Type	Change in Average Prior Test Score	Grade				
		4	5	6	7	8
Cohort SGP	< −.3 SD	1.10	-0.86	-0.68	-0.62	-0.42
	−.3 to +.3 SD	-0.09	0.52	-0.09	-0.23	0.19
	> +.3 SD	-0.34	-0.39	-0.26	-0.50	0.46
Baseline SGP	< −.3 SD	0.68	-3.38	-3.96	1.15	-3.49
	−.3 to +.3 SD	-0.09	-2.61	-2.34	2.79	-2.94
	> +.3 SD	0.40	-4.07	-1.56	4.07	-2.07
SIMEX-Baseline SGP	< −.3 SD	1.76	-2.26	-2.75	2.48	-2.50
	−.3 to +.3 SD	-0.74	-2.91	-2.43	2.72	-3.48
	> +.3 SD	-1.84	-5.88	-3.03	2.50	-3.85
Number of Teachers	< −.3 SD	444	408	153	172	140
	−.3 to +.3 SD	1813	1829	1094	988	1024
	> +.3 SD	505	527	198	207	216

Table B.1b. Cross-year change (average difference) in AGP for Mathematics: 2012 – 2011

SGP Type	Change in Average Prior Test Score	Grade				
		4	5	6	7	8
Cohort SGP	< −.3 SD	3.55	2.82	-1.42	1.93	0.62
	−.3 to +.3 SD	0.26	0.74	-0.10	-0.21	0.47
	> +.3 SD	-3.11	-1.69	-0.97	-3.14	-1.32
Baseline SGP	< −.3 SD	-0.05	-2.59	-3.28	3.55	0.59
	−.3 to +.3 SD	-2.86	-4.12	-2.01	1.12	-0.50
	> +.3 SD	-5.74	-6.19	-2.18	-1.93	-2.84
SIMEX-Baseline SGP	< −.3 SD	1.00	-1.38	-1.99	4.83	2.07
	−.3 to +.3 SD	-3.39	-4.14	-2.13	0.92	-0.51
	> +.3 SD	-8.05	-7.33	-3.77	-3.72	-4.00
Number of Teachers	< −.3 SD	422	200	159	165	92
	−.3 to +.3 SD	1536	1412	897	939	753
	> +.3 SD	485	841	199	187	443

Table B.2a. Cross-year change (average difference) in AGP for ELA: 2013 – 2012

SGP Type	Change in Average Prior Test Score	Grade				
		4	5	6	7	8
Cohort SGP	< −.3 SD	0.63	0.87	-1.42	-0.31	0.04
	−.3 to +.3 SD	-0.15	-0.56	-0.38	-0.05	-0.17
	> +.3 SD	-0.30	-1.03	1.17	0.44	1.15
Baseline SGP	< −.3 SD	-2.56	1.83	-3.62	-1.43	-0.82
	−.3 to +.3 SD	-3.25	-0.45	-1.77	0.22	-1.85
	> +.3 SD	-3.19	-1.81	1.31	1.53	-1.28
SIMEX-Baseline SGP	< −.3 SD	-0.96	3.00	-2.50	0.16	0.39
	−.3 to +.3 SD	-3.45	-0.75	-2.03	0.37	-2.21
	> +.3 SD	-5.24	-3.62	-0.49	-0.08	-3.58
Number of Teachers	< −.3 SD	542	531	195	198	212
	−.3 to +.3 SD	1674	1667	996	893	927
	> +.3 SD	397	377	143	129	134

Table B.2b. Cross-year change (average difference) in AGP for Mathematics: 2013 – 2012

SGP Type	Change in Average Prior Test Score	Grade				
		4	5	6	7	8
Cohort SGP	< −.3 SD	1.86	0.72	1.02	1.55	0.90
	−.3 to +.3 SD	-0.87	-1.08	-0.22	0.27	-0.15
	> +.3 SD	-1.28	-0.48	0.18	-0.74	0.12
Baseline SGP	< −.3 SD	5.86	6.02	10.17	-3.39	1.68
	−.3 to +.3 SD	3.03	4.96	8.94	-5.03	1.36
	> +.3 SD	2.75	5.89	9.85	-6.51	3.49
SIMEX-Baseline SGP	< −.3 SD	7.78	6.93	11.70	-2.19	2.27
	−.3 to +.3 SD	3.15	4.81	9.20	-5.28	0.97
	> +.3 SD	1.08	4.46	8.65	-8.34	1.66
Number of Teachers	< −.3 SD	499	826	205	183	476
	−.3 to +.3 SD	1376	1117	864	883	663
	> +.3 SD	362	182	156	135	86

Table B.3. Mean differences between baseline-references and cohort SGP by year and grade

	Year		
Grade	2011	2012	2013
4	1.97	2.11	-0.92
5	-0.90	-3.93	-3.87
6	-0.20	-2.36	-3.79
7	1.98	5.18	5.33
8	-1.61	-4.70	-6.34

References

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

Betebenner, D. (2011). *Technical Overview of the Student Growth Percentile Methodology: Student Growth Percentiles and Percentile Growth Projections/Trajectories*, Dover, NH: National Center for the Improvement of Educational Assessment Technical Report. Retrieved from http://www.nj.gov/education/njsmart/performance/SGP Technical Overview.pdf.

Betebenner, D. W., Van Iwaarden, A., Domingue, B., and Shang, Y. (2014), *SGP: An R package for the calculation and visualization of Student Growth Percentiles & percentile growth trajectories*, R package version 1.2-0.0.

Georgia Department of Education, (2011). *Race to the Top: State of Georgia scope of work*. Retrieved from http://www2.ed.gov/programs/racetothetop/state-scope-of-work/georgia.pdf.

Georgia Department of Education, (2012) *A Guide to the Georgia Student Growth Model.* Retrieved from http://www.gadoe.org/Curriculum-Instruction-and-Assessment/Assessment/Documents/SGP%20Guide%20122112.pdf.

Georgia Department of Education, (2014). *Methods of combining SGP*. Retrieved from http://www.gadoe.org/Curriculum-Instruction-and-Assessment/Assessment/Documents/Methods%20of%20Combining%20SGPs.pdf

Guarino, C. M., Reckase, M. D., Stacy, B. W., & Wooldridge, J. M. (forthcoming). A comparison of growth percentile and value-added models of teacher performance. *Statistics and Public Policy*.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications* (Vol. 7). Springer.

Lederer, W. & Kuchenhoff, H. (2006). A short introduction to the SIMEX and MCSIMEX. *The Newsletter of the R Project, 6/4,* 26-31.

Lockwood, J. R., & Castellano, K. E. (forthcoming). Alternative statistical frameworks for Student Growth Percentile estimation. *Statistics and Public Policy.*

McCaffrey, D., & Castellano, K. E. (2014). *A review of comparisons of aggregated student growth percentiles and value-added for educator performance measurement*. Report for the Georgia Department of Education.

Monroe, S., Cai, L., & Choi, K. (2014). *Student growth percentiles based on MIRT: Implications of calibrated projection.* (CRESST Report 842). Los Angeles, CA: University of

California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

R Development Core Team. (2014). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing: Vienna, Austria.

Shang, Y. (2012). Measurement error adjustment using the SIMEX method: An application to student growth percentiles. *Journal of Educational Measurement*, 49, 446–465.

Shang, Y., Van Iwaarden, A., and Betebenner, D. (forthcoming). The condition and application of covariate measurement error correction for Student Growth Percentiles. *Educational Measurement: Issues and Practice.*

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford: Oxford University Press.

Walsh, E. and Isenberg, E. (forthcoming). How does a value-added model compare to the Colorado growth model? *Statistics and Public Policy*.

Wooldridge, J. M. (2013). "Fixed Effects Estimation". *Introductory Econometrics: A Modern Approach* (Fifth international ed.). Mason, OH: South-Western. pp. 466–474.