# EVALUATING TECHNICAL ISSUES WITH STUDENT GROWTH PERCENTILES AS PART OF GEORGIA'S TEACHER AND LEADER EFFECTIVENESS SYSTEMS: AN OVERVIEW OF GEORIA STUDENT GROWTH MODEL RESEARCH

**Scott Marion**[1]

**Center for Assessment**

**December 8, 2014**

[1] Thanks to Katie Buckley for helpful comments and edits.

# Introduction

The Georgia Department of Education (GaDOE) conducted an extensive process, involving multiple stakeholder groups and technical advisors, to decide on the growth or value-added model that GaDOE would use for school and educator accountability. Unless there was a compelling reason to use separate models for school accountability and educator evaluation, GaDOE was interested in fostering coherence across the two systems by adopting the same model for both schools and educators. For a variety of reasons, GaDOE selected the Student Growth Percentile (SGP) Model (Betebenner, 2008, 2009) as its growth model.

Georgia proposed in its RTTT application that it would contract with a researcher other than the developer of the state growth/VAM model to “validate” the model before using it fully operationally as described in this excerpt from the RTTT application.

> ***Validation of the Value-Added Growth Model****: Contract for independent validation of the value added growth model with a vendor other than the primary developer. This would cover an independent validation of the Value-Added Model approach and analysis, before taking the analysis and results “live” and communicating them more broadly with educators (district leaders, principals, teachers) and also with non-educators (researchers, parents, etc.).*

Validation is a major undertaking and cannot be fully completed until after the model is used operationally in the system. Therefore, it is more appropriate to think of this effort as an early evaluation of key technical features of the SGP model. There are many issues to pursue in examining the use of SGP and the aggregation of SGPs (mean or median SGPs) for teacher and leader evaluation in Georgia, including:

1. The extent to which aggregated SGPs can be viewed as measures of “true” differences in teacher quality,
2. The observation that aggregated SGPs and average prior achievement and/or other student- or class-level characteristics are moderately correlated,
3. The effects of “churn” in terms of student mobility at the classroom level,
4. The relationship of aggregate SGPs and alternative estimates of teacher quality from other value-added model specifications,
5. The use of the SIMEX approach to adjust for measurement error, and
6. The within-year and year-to-year reliability/consistency of MGP results.

## The Researchers

Given the time frame and the needs of the GaDOE, several classes of studies have been prioritized and five studies (#1-5 above) were conducted by three leading independent researchers,[2] all of whom are members of the GaDOE educator effectiveness technical advisory committee (TAC):

- Daniel McCaffrey, Principal Research Scientist, Educational Testing Service
- Derek Briggs, Professor of Education, University of Colorado, Boulder
- Henry Braun, Professor of Education, Boston College

While each researcher operated independently of one another, there was opportunity for the researchers to present preliminary findings to one another and the rest of the TAC to receive feedback on their studies.

## Summary of Studies and Findings

The five studies were designed to address complementary questions all focused on the general issue of whether SGPs and aggregate SGPs are a fair and useful indicator as part of Georgia's Teacher Keys Effectiveness System (TKES). The studies addressed questions dealing with the degree to which aggregate SGPs could distinguish among teachers considered high and lower quality by their principals and the relationship of aggregate SGPs with factors thought to be out of the control of the teacher such as the demographic characteristics and mobility rates of the students in the class.

While the primary purpose of the studies was to provide an external review of the technical quality of the SGP model, each researcher was also motivated by the need to better understand the relatively large correlations—compared to other states—between average prior achievement and aggregate SGPs at the classroom and school levels. On the one hand, this relationship could be the result of bias in the aggregate SGP measure; however it could also be that in Georgia, there is greater sorting of teachers among more and less advantaged schools. Each researcher sought to disentangle this issue further.

The five studies, collectively, provide information useful for disentangling these correlations by helping to shed light on whether the correlations between aggregate SGPs and average prior achievement were due to some features of the SGP model or large structural issues such as sorting of teacher quality across the state. Of course, this question is impossible to answer definitively without conducting experiments where teachers are randomly assigned to districts and schools, which would be an incredibly difficult experiment to conduct. Some key findings from the five studies include:

- Aggregate SGPs tend to produce results highly related to those from alternate value-added model specifications.
- The SIMEX correction is effective for adjusting for measurement error in the prior scores, but not for the current score.
- While the SIMEX correction helps to reduce the correlation between aggregate SGPs and prior scores, a non-trivial relationship remains.

[2] Only the principal investigator is listed here. Contributing researchers and authors are listed with each study.

- One set of studies examined the performance of teachers who moved between schools characterized by different levels of poverty or classes of students with different prior achievement to try to hold the teacher "constant." The researchers found evidence of teacher sorting and little relationship between aggregated SGPs and student background factors.
- Classes with higher mobility rates of students or "churn" appears to have a slight negative effect on the teacher's aggregate SGP over and above the negative effect related to having higher percentages of economically disadvantaged students in the class;
- One of the studies suggest that teachers rated by their principals as highly likely to improve student learning have noticeably higher aggregate SGPs than teachers rated by their principals as less likely to improve student learning.

While not free of technical challenges, these studies portray aggregate SGPs as a useful indicator in Georgia's TKES that is able to contribute valuable information for evaluating teachers in Georgia. In fact, most value-added models and aggregate SGPs produce such similar results that the technical and use challenges found with SGPs are similar to those found for VAM. A more detailed summary of each of the studies is presented below.

# STUDY SUMMARIES

## Study 1

### A Review of Comparisons of Aggregated Student Growth Percentiles and Value-Added for Educator Performance Measurement

Daniel McCaffrey and Katherine E. Castellano
Educational Testing Service

States using student growth as part of teacher evaluations are generally split in their choice of models for evaluating student growth for teachers in “tested” grades and subjects (those subject/grades with a state test and a state test in the same subject in a prior grade) between value-added models (VAM) and an aggregated SGP model. Both methods rely on comparing students’ current achievement to their past achievement, but differ with respect to how they use the data to yield measures of educator performance. Georgia is now using mean SGP of students linked to a teacher or leader, but had previously used the median in the first phase of the teacher evaluation pilot. While there was some interest in examining the differences and similarities of results when using value-added and student growth percentile models as applied to Georgia data, there is already a growing literature base in this area. Therefore, this first study provides a focused literature review regarding these comparisons and the implications for Georgia.

These authors reviewed the literature on comparisons of aggregated SGPs with VAM to assess the possible impact of Georgia’s decision. It should be noted that much of this literature focused on median SGPs and not mean SGPs like Georgia is currently using, although the relationship between the two types of aggregated SGPs is strong. The literature tends to find that aggregated SGPs and VAM are highly correlated and that a relatively small percentage of teachers change classification when aggregated SGPs are substituted for VAM measures. However, when aggregated SGPs and VAM differ, the differences are not randomly distributed across teachers. Several studies find that teachers of students with below average prior achievement or from low-income families tend to rank relatively lower using aggregated SGPs compared with other teachers than they would using VAM, although two studies also find the opposite: VAM had a stronger relationship to average prior achievement than did aggregated SGPs. Only one study compared the inter-temporal stability of aggregated SGPs and VAM and found aggregated SGPs to be more unstable across years, although the difference was not large. Another study found that the use of the mean SGP rather than the median SGP would reduce statistical errors in aggregated SGPs, which could improve year-to-year stability. This is one of the reasons why the GaDOE made the decision to switch from median to mean SGPs.

## Study 2

## An Evaluation of Technical Issues with the Student Growth Model Component of the Georgia Teacher and Leader Evaluation System

Daniel McCaffrey, Katherine Castellano, & J.R. Lockwood
Educational Testing Service

One of the major challenges with the implementation of aggregated SGPs for teacher and leader evaluation in Georgia is the relatively strong correlation between aggregated SGPs and prior achievement.  These correlations may reflect the real distribution of teacher/leader quality but may also reflect both error and bias in the model or some combination of both.  The simulated-extrapolation (SIMEX) correction for measurement error has helped reduce the correlations, but even with the SIMEX correction applied, there is still a noticeable relationship between prior achievement and aggregated SGPs. This study therefore comprises evidence from a set of analyses investigating the technical issues associated with using aggregated SGPs in Georgia's educator evaluation system, including the effects of measurement error on aggregated SGPs, the utility of the SIMEX method for dealing with this measurement error, the effects of using different types of SGPs, and empirically evaluating whether teacher sorting is occurring and the relationship of potential sorting to aggregate SGPs.

The influence of measurement error on the relationship of aggregate SGPs to mean prior achievement is dependent on factors such as the reliability of the tests, the "effectiveness" of the individual educators, and the achievement and growth of their students. The correlation between aggregated SGPs and mean prior achievement for 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, but this is likely related to the lower reliability associated with medians than means.

The authors found that the combined measurement error in both the prior and current test scores may lead to teachers associated with high growing students in economically disadvantaged schools receiving aggregate SGPs that are too low, and teachers associated with low growing students in schools serving economically advantaged students receiving aggregate SGPs that are too high. The authors did find that applying the SIMEX measurement error correction reduced the correlations of aggregated SGPs with student background variables and average prior achievement, even though they noted some potential limitations with the SIMEX approach for correcting for measurement error.

The authors investigated the effects of including multiple prior test scores in the SGP calculations on the correlation between aggregated SGP and mean prior achievement and found that adding more prior test scores, even from different content areas, reduces the correlation between aggregate SGPs and prior achievement.  This benefit is greatest when moving from one to two prior scores, but noticeable reductions in the correlation are found when adding a third and even fourth prior score. The baseline-referenced SGPs implemented by GaDOE include two prior years in the baseline and may not realize the full advantage that additional prior scores provide.

The authors designed three very interesting empirical studies, one of which used the same set of teachers to address questions of statistical bias in aggregate SGPs in Georgia. These teachers were ones who switched contexts (e.g., from high to low poverty schools or from classes of students with low prior achievement to classes with higher prior achievement or vice versa) in adjacent years so the researchers were able to attempt to isolate the effect of bias in aggregated SGPs from teacher sorting as the source of the correlation with mean prior achievement. In general, the researchers found evidence that teacher sorting was occurring and they found little evidence of a relationship between the aggregated SGP and student background characteristics among classes with different types of students taught by the same teacher.

## Study 3

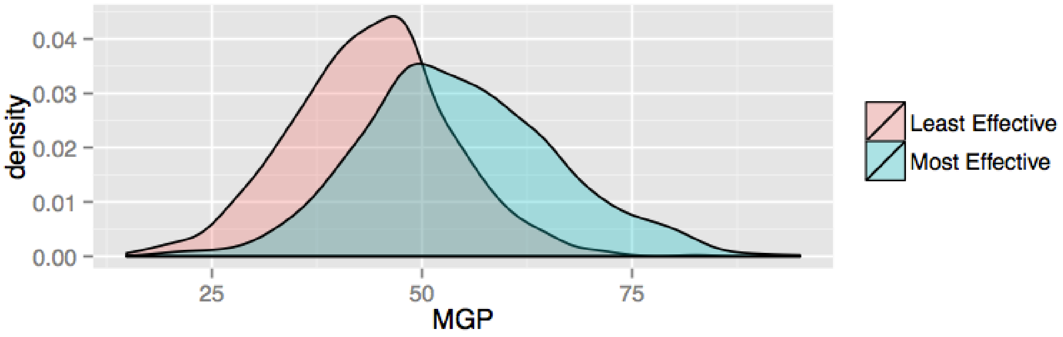
## Comparing Student Growth and Teacher Observation to Principal Judgments in the Evaluation of Teacher Effectiveness

Derek Briggs, Nathan Dadey, & Ruhan Kizil
University of Colorado, Boulder

As described above, for many of the studies, one of the major limitations to any analysis is that there is no way to really know the “truly" exceptional and ineffective teachers without having all teachers randomly assigned to districts, schools and classrooms. To attempt to find effective and ineffective teachers, researchers surveyed a set of school principals and used a set of questions to have them identify teachers who they consider highly effective and those who they consider ineffective. Specifically, the survey first asked the principals to rate the level of support each teacher in their school would need in order to have a positive impact on student achievement. This question was followed by two questions asking the principal to pick the teacher who they considered the most and least successful, respectively, at increasing student achievement. Surveys were sent to a stratified random sample of principals at 1,394 Georgia schools. Impressively, 1,013 principals responded for a response rate of 73%.

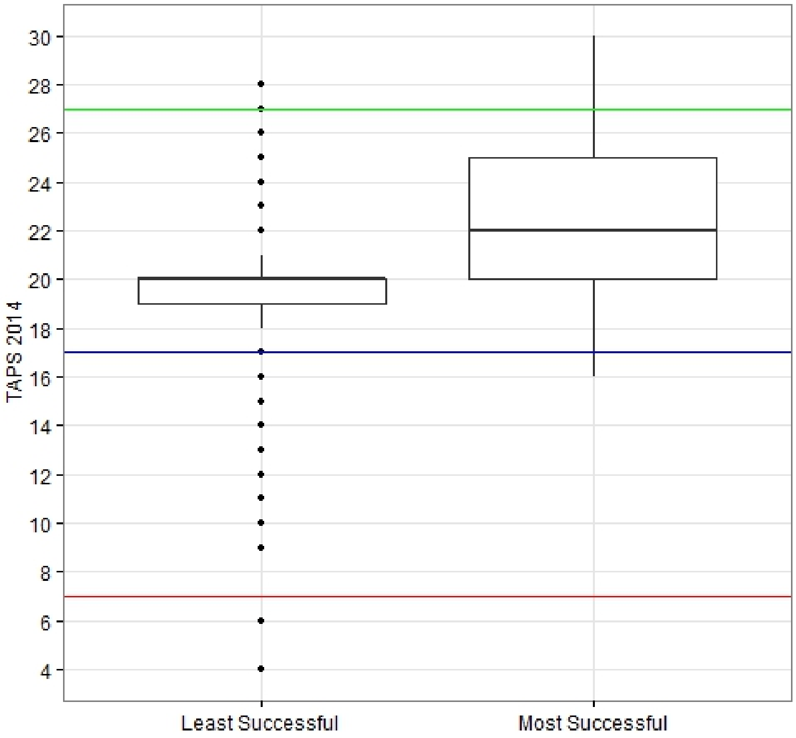
The researchers computed the mean SGP for all 2012-2013 SGPs linked to each teacher across all subject areas. For example, if an elementary teacher had SGPs in mathematics, reading, ELA, science, and social students, the teacher’s mean SGP for this study would be the average of all SGPs associated with that teacher. The researchers then compared these mean SGPs for teachers rated by their principals as most and least effective in terms of improving student achievement. The authors’ results are summarized nicely in the figure below (Figure 3 from the original paper).

While the two distributions may share a fair amount of overlap, those used to looking at such distributions will be impressed by the degree of non-overlap. In fact, teachers identified as “most effective” by their principal had a mean SGP of 54.4 and those identified as “least effective” had a mean SGP of 43.9, for a difference of 10.5. The average standard deviation of mean SGPs across the two groups of teachers was approximately 10.6 so that the effect size difference between these two groups of teachers was about 1.0. This means that the average “highly effective” teacher would be at approximately the 84th percentile of the distribution of less effective teachers.



There are some differences in these findings when mean SGPs are calculated using the SIMEX correction or calculated using a post-hoc regression equation to account for differences in classroom context. However, in both cases, the difference in mean SGPs between highly and least effective teachers was reduced slightly from the results presented above to differences of 10.1 and 9.4 percentiles depending on the method, still sizeable differences.

The authors’ final analysis looked at the relationship of these ratings to the TAPS scores received by teachers. TAPS is the tool and procedures used to evaluate teacher practices in ten categories generating total scores from 0 to 30. Teachers are classified as Level 3, which might be considered “proficient,” on TAPS if their total score is 17-26; almost all teachers in the state (approximately 94%) fall into this range. That said, the authors of this study noted that principal ratings of most and least successful teachers at improving student achievement were still related to TAPS scores. The figure below (Figure 8 in the original paper) clearly shows that the TAPS scores barely overlap for the teachers rated effective and ineffective at raising student achievement by their principals.



The authors note that the dots in the least successful boxplot indicate outliers and the colored horizontal lines represent demarcations between performance levels.

Importantly, the authors concluded that very few principals were selecting teachers as least or most successful at increasing student achievement solely on the basis of information at their disposal about student growth percentiles.  They provided a convincing case that most of their principal respondents were making holistic judgments on the basis of both teacher practices and student outcomes when asked to rate their teachers, and, in fact, they argued that the observations of teacher practices were likely the driving force behind a principal's rating. If one believes that the sort of principal ratings used in this study help shed light on "true teacher effectiveness," then it is encouraging to see that SGPs and aggregate SGPs are strongly related to these ratings.

## Study 4

## Adjusting Mean Growth Percentiles for Classroom Composition

Derek Briggs, Ruhan Kizil, & Nathan Dadey

University of Colorado, Boulder

SGP models condition current achievement only on prior achievement scores, but many value-added models (VAM) employ statistical adjustments for student-level characteristics such as poverty and special education status and/or aggregate characteristics such as the mean prior achievement of the class or school embedded in the model. These are considered first order adjustments. However, the GaDOE educator effectiveness technical advisory committee has suggested that it may be possible to conduct "second order" adjustments of SGP and aggregated SGP results. Second order adjustments would involve making adjustments to the SGP results after the initial calculations have been completed, by, among other approaches, regressing the aggregate SGP results on factors such as average prior achievement. An advantage of such an approach, compared to first order adjustments, is that it makes the adjustment quite transparent compared with having such adjustments buried deep in complex statistical models. On the other hand, this means that the adjustments are a bit coarser because they are made at the aggregate and not individual level, and may lead to teachers who would not be considered effective when compared to the entire distribution of teachers being labeled as effective due to their reduced comparison group.

This study used the 2012-2013 data for teachers in grades 4, 5, 6, 7, and 8 with SGPs in reading or mathematics. Only teachers with at least 15 students with SGPs in a given subject area were included in the sample. This resulted in a sample of teachers ranging from 4,500 to 8,500 depending on subject area and grade level. The mean, subject-specific SGP was regressed on the percentage of students eligible for free and reduced price lunch (FRL), percentage of English language learner (ELL) students, percentage of special education students (SWD), and the mean prior grade achievement to yield an adjusted mean SGP. The study then compared the adjusted and unadjusted mean SGPs in terms of the correlations between the two sets of results as well as the consistency of classifications between the two sets of analyses.

The correlations between the adjusted and unadjusted mean SGPs ranged from a low of .92 to a high of .97, depending on grade level and subject. In other words, the two sets of results were almost perfectly correlated. However, the strength of the correlation does not mean that all teachers would be classified the same using either approach. The researchers evaluated the classification consistency of the two methods and found that between 87.8 and 93.8% of teachers, depending on grade and subject, were classified identically using adjusted and unadjusted mean SGPs. These findings would be very impressive if the misclassifications are random, because we could assume the misclassifications are simply due to error that is magnified when trying to classify teachers into four categories. However, the researchers investigated the teachers who had higher or lower MGPs depending on whether adjusted or unadjusted mean SGPs were used. They found that teachers whose classification improved when moving from unadjusted to adjusted mean SGPs tended to have higher than average percentages of students eligible for FRL and a much lower than average mean prior achievement. Those teachers whose classification declined in performance levels tended to have classes with noticeably higher than average prior achievement and lower than average percentages of students eligible for FRL. This is not surprising considering this is what the regression adjusted mean SGPs is designed to accomplish. Importantly, no teacher shifted classification by more than one performance level.

The authors concluded with a discussion of whether to adjust or not to adjust. They pointed out that difficult decisions would have to be made about which variables to include in the regression equation and whether different equations should be employed for different subjects and grades. Employing regression equations with different variables or with many variables will likely make the intended comparisons difficult to explain. The authors pointed out that while there may be some temptation to use adjusted mean SGPs, the overall effect of doing so was quite limited.

## Study 5

## Georgia MGP Churn Rate Study

Henry Braun & James Burraston
Boston College

Other studies in this research program investigated the relationship of mean SGPs with factors such as mean prior achievement, percentage of students eligible for FRL, special education services, and English language learner programs. Factors such as the percentage of students eligible for FRL are considered distal indicators in that poverty itself (to the extent that FRL is a valid indicator of poverty) does not cause low achievement, but is played out through other more proximal influences. One such factor is the generally higher rates of mobility experienced by poorer compared with more financially secure students. There is some evidence that schools and classrooms with high degrees of student turnover, or “churn,” may influence teachers’ practices and, ultimately, how much (relative) progress their students make, at least as determined by the results for those students who contribute to teachers’ mean SGPs. This study investigated the relationship of teacher mean SGPs to a measure of classroom churn, which is an indicator related to the number of student transitions (into a class or out of a class) that take place in the course of a semester or a year. In addition, the relationship of MGPs to a measure of classroom-level socio-economic disadvantage was also investigated.

Greater mobility is thought to place an increased burden on the teacher as he or she must deal with the disruptions caused by students leaving and new students arriving. Presumably, these disruptions also affect the learning of all students in the class. In addition, accountability system business rules determine how long students must be enrolled in a class for their test scores to contribute to the teacher’s rating. Arguably then, the greater the churn, the lower the ratio of the teacher’s effort (measured in student-days in her class) that counts toward his or her rating to his or her total effort over the academic year (again measured in student-days). Greater churn at the school level can also influence the context of learning for all classes. This report presents the study findings, discusses limitations and suggests the need for a follow-up study.

In order to carry out this study, the researchers needed more frequent measures of enrollment than the once or twice each year reported by most districts. With the help of the GaDOE, the researchers were able to identify 314 high school teachers in 120 schools that reported enrollment every nine (9) weeks. The sample was reduced to 293 teachers with the requirement that teachers have at least five (5) students with SGPs to be included in subsequent analyses. With four enrollment reporting periods, the researchers were able to compute churn indicators for the last three periods of the school year. The churn rate was defined as the total number of transitions divided by the total number of students that appeared in that class during the reporting period.

The researchers found that higher churn rates were associated with slightly lower mean SGPs and interestingly this effect was essentially independent of the effect of higher rates of FRL on mean SGPs. In other words, the two effects were almost additive, which surprised the researchers. The researchers analyzed the semester and year-long classes separately. They reported for semester-long classes that a 0.5 standard deviation increase in both churn rates and FRL is associated with a reduction in mean SGPs by 3.0 percentile points. They found a slightly larger effect for year-long classes with a 4.3 percentile point drop in mean SGP associated with a 0.5 standard deviation increase in both churn and economic disadvantage.

The authors note that these results should be considered preliminary in light of the very limited data available and the lack of representativeness of the schools included in this early study. More refined transaction data for measuring churn rates (even finer-grained than every nine weeks), will offer a stronger foundation for understanding the influence of student mobility on indicators of teacher quality.

## References

Betebenner, D. W. (2008). Toward a normative understanding of student growth. In K. E. Ryan & L. A. Shepard (Eds.), *The future of test-based educational accountability* (pp. 155–170). New York: Taylor & Francis.

Betebenner, D. (2009). Norm- and criterion-referenced student growth. *Educational Measurement: Issues and Practice,* Vol. 28(4), 42–51.

Braun, H. and Burraston, J. (2014). Georgia MGP churn rate study. Boston College.

Briggs, D.C, Dadey, N., & Kizil, R.C. (2014). Comparing student growth and teacher observation to principal judgments in the evaluation of teacher effectiveness. University of Colorado.

Briggs, D.C, Kizil, R.C. & Dadey, N. (2014). Adjusting mean growth percentiles for classroom composition. University of Colorado.

McCaffrey, D.F., Castellano, K.E., & Lockwood, J.R. (2014). An evaluation of technical issues with the student growth model component of the Georgia teacher and leader evaluation system. Educational Testing Service.

McCaffrey, D.F. and Castellano, K.E. (2014). A review of comparisons of aggregated student growth percentiles and value-added for educator performance measurement. Educational Testing Service.