Office of the State Superintendent of Education

School-Level Model to Measure Student Achievement over Time Technical Report

Authored by:
Harold Doran
Andrew Swanlund
Mariann Lemke

American Institutes for Research

AMERICAN INSTITUTES FOR RESEARCH

TABLE OF CONTENTS

ACKNOWLEDGMENTS ..... *ii*
INTRODUCTION *1*
METHODS *2*
Description of Quantile Regression Model *3*
Quantile Model Specification *3*
Creating Student Growth Percentiles *3*
Aggregating Student Growth Percentiles to Form Median Growth Percentiles *4*
Measures of Dispersion and Precision of MGP Within Groups *4*
(MAD): Descriptive Measure of Dispersion *4*
Analytic and Bootstrap Standard Errors: Measures of Precision of the MGP *5*
Aggregating Medians Using Student Weights *5*
The Impact of Measurement Error in Quantile Regression *6*
RESULTS *6*
Within-Year Precision of the School MGPs *7*
Between-Year Stability of the School MGPs *8*
Aggregated MGP Over Time *11*
Analysis of Student Growth Percentiles by Student Group *12*
Relationship of Student Growth Percentiles to Prior Year Scores *17*
Stability of Student Growth Percentiles Over Time *19*
CONCLUSION *20*

APPENDIX A. SCHOOLWIDE GROWTH ADVISORY COMMITTEE (SGAC) MEMBERS

APPENDIX B. BUSINESS RULES

APPENDIX C. DATA OVERVIEW

ACKNOWLEDGMENTS

The authors would like to thank the following analysts at the American Institutes for Research for their contributions to this report: Dan Sherman, Nick Sorensen, Susan Ward, Paul Van Wamelen, and Vijay Sundaram.

In addition, Robin Chait at the Office of the State Superintendent of Education provided overall guidance for the project, and Tamara Reavis provided perspective from the Assessment and Accountability Office. Clara Hess of the DC Public Charter School Board also provided input.

Finally, members of the Student Growth Advisory Committee, consisting of representatives from local education agencies and a charter advocacy group, provided suggestions on the contents of this report. Members of the Committee included Joshua Boots (KIPP DC), Lydia Carlis (Apple Tree Institute), Steve Cartwright (DC Public Schools), Naomi Deveaux (FOCUS), Prudence Hallarman (DC Prep), Zac Morford (Friendship Public Charter School), Unique Morris (Septima Clark Public Charter School), Jeff Noel (FOCUS), and Megan Reamer (Capital City Public Charter School).

INTRODUCTION

As part of its Race to the Top (RTT) grant, the Office of the State Superintendent of Education (OSSE) in Washington, DC, developed a school-level model to measure growth in student achievement over time. Different education agencies in the District of Columbia are using the data from this model for different purposes, including informational reporting and as one component of a school-level accountability system.

The model used by OSSE is a "student growth percentile" (SGP) or "median growth percentile" (MGP) model and includes only test scores to measure growth (i.e., with no adjustments for student- or school-level contextual variables).

The DC Public Charter School Board (PCSB) voted to adopt this model, and a Race to the Top advisory group including educators, district-level staff, and representatives from OSSE and PCSB selected the same model for use across the District. The American Institutes for Research (AIR) collaborated with OSSE to implement this model.

A Schoolwide Growth Advisory Committee (SGAC), consisting of representatives from local education agencies (LEAs) and a charter advocacy group, provided input on the specifics of the model design and implementation. The names of the participants and their affiliations are provided in Appendix A. Table 1 shows the dates of the main SGAC meetings and the focus of each meeting.

Table 1: SCAG Meeting Dates and Meeting Focus

Meeting Date	Focus
July 20, 2011	Establish business rules
September 28, 2011	Review initial model results and select final model
October 7, 2011	Review year-to-year stability of estimates and recommend reporting approach
October 11, 2011	Provide information on final selected model (meeting open to all LEAs in DC)

After OSSE made a preliminary decision about the type of model to be used, AIR worked with OSSE and its partners to

- determine the specifics of the data to be included (i.e., the business rules, included in this report as Appendix B), and
- choose between the two model variants presented to the SGAC.

A model with two prior years of test scores was ultimately selected, and that model and its results are described in more detail in the following sections of this report and in Appendix C.

The SGAC also provided input on other project documents (e.g., communications materials) and participated in project update meetings.

This document provides an overview of the data used in the analysis, the technical details of the growth model, and a description of its results. The report first presents the methodology used to estimate student growth percentiles and then summarizes the results.

METHODS

The model implemented for OSSE is typically referred to as a student growth percentile model (Betebenner, 2009). Student growth percentiles (SGPs) are metrics that represent how a student's growth in measured achievement compares with that of other students with similar prior test scores. The metric is normative in that the SGP for student *i* is always relative to other students with the same prior score history.

The canonical expression for the SGP is:

$$SGP = \Pr(current\ achievement | prior\ achievement) * 100.$$

Because the model conditions the SGP based on prior test scores, the SGP is essentially locally normed relative to the achievement of other students with the same prior test scores. For example, we say that a student with an SGP of 60 has performed better in the current year than 60 percent of the students in the data with similar prior test score histories.

It is important to note that an SGP of 60 does not imply that a student grew more than 60 percent of *all* students in the tested population but rather in comparison to those with the same prior scores. Because the SGP is estimated conditionally on prior test scores, two students with an SGP of 60 with very different prior test scores could have very different scores in the current year, but they could still have the same SGPs relative to the students with whom they are compared. Consequently, SGPs are not directly comparable across students with different prior scores.

The typical method used to obtain SGPs is through the use of a quantile regression (QR).[1] In our approach, which is described below, we use the most recent test score as the outcome variable and prior test scores as covariates.

The QR method is similar to a least squares regression model. However, in QR the interest is the conditional value of the response variable at the τth percentile whereas in least squares regression the interest is only on the conditional mean. For example, $\tau = .5$ would be the conditional value of the response variable at the 50th percentile.

Least squares and QR have a technical difference in the methods used for optimization. Least squares methods provide closed-form solutions that cannot be implemented for QR. In estimating the QR model, the objective function is minimized differently from a least squares regression in that gradient optimization methods cannot be used. Rather, linear programming methods are used for optimization. Most commonly, the simplex algorithm applied to QR, as described by Barrodale and Roberts (1973), is used to find the optimal value of model

[1] A description of the QR procedure and its implementation in the SAS software package is available at www2.sas.com/proceedings/sugi30/213-30.pdf.

parameters. Although estimation methods are very different, interpretation of resulting coefficients and inferences between QR and least squares regressions are generally similar.

Description of Quantile Regression Model

Quantile regression is based on minimization of the objective function, $\rho_\tau(.)$:

$$\hat{\beta}(\tau) = \underset{\beta \in \Re}{\arg\min} \sum_{i=1}^{N} \rho_\tau(y_i - x'_i\beta)$$

with $\tau \in (0,1)$, y_i is the outcome variable for student *i*, and x'_i is the vector of observed prior scores and other measured characteristics for student *i*. The function $\rho_\tau(.)$ is a generalized absolute value function for a given percentile:

$$\rho_\tau(x) = (\tau - I(x < 0))x$$

where I is the indicator function (so $\rho_{.5}$is the absolute value function). The function is minimized in the SAS Quantreg procedure for the work described in this report.

Quantile Model Specification

The following model specification is used to estimate the quantile regression model for OSSE. The model below uses student-level data and is run separately by grade and subject:

$$y_{gi} = \mu_\tau + \beta_1(\tau) * y_{i,g-1} + \beta_2(\tau) * y_{i,g-2} + \beta_3(\tau) * M_1 + \beta_4(\tau) * M_2 + \varepsilon$$

where y_{gi} is the DC CAS test score for student *i* in grade *g*, $y_{i,g-1}$ is the test score for student *i* observed in grade $g - 1$, $y_{i,g-2}$ is the test score for student *i* observed in grade $g - 2$, $\beta(\tau)_k$ ($k \in \{1,2,\ldots,4\}$) are the estimated coefficients at quantile τ, ε is an error term, and

$$M_1 = \begin{cases} 1 \text{ if } y_{g-1} \text{ is missing} \\ 0 \text{ otherwise} \end{cases}$$
$$M_2 = \begin{cases} 1 \text{ if } y_{g-2} \text{ is missing} \\ 0 \text{ otherwise} \end{cases}$$

The indicator variables, M_1 and M_2, are used only to allow retention of students with a missing prior score. Students must have at least one of the two prior scores to be included in the model estimation.

Creating Student Growth Percentiles

Given a solution for $\beta(\tau)$ for each quantile $\tau = \{.01, .02, \ldots, .99\}$, fitted values for each of the quantiles, denoted as $\hat{y}_i(\tau)$, are produced as:

$$\hat{y}_i(\tau) = \boldsymbol{x}'_i\hat{\boldsymbol{\beta}}(\tau)$$

where $\boldsymbol{x}_i'$ is the transpose of the *i*th row of the model matrix $\boldsymbol{X}$ and $\widehat{\boldsymbol{\beta}}(\tau)$ is the vector of fixed coefficients at the solution $\beta(\tau)$. The student growth percentile is assigned by locating the largest τ where the observed outcome variable meets the following inequality:

$$SGP_i = y_i > \hat{y}_i(\tau)$$

For example, we may find that $y_i > \hat{y}_i(.5)$ and $y_i > \hat{y}_i(.51)$ in which case the SGP assignment would be $.51$. That is, we say the SGP is τ when we find the fitted value that satisfies the inequality above.

It may also be true, for example, that $y_i > \hat{y}_i(.5)$, $y_i < \hat{y}_i(.51)$, and $y_i > \hat{y}_i(.52)$. In this scenario, the SGP would be .52. It is for this reason that the business rule requires the largest τ where the observed outcome variable meets the following inequality $y_i > \hat{y}_i(\tau)$.

Aggregating Student Growth Percentiles to Form Median Growth Percentiles

For each aggregate unit *j* ($j \in \{1,2,\ldots,J\}$), such as a school, the interest is a summary measure of growth for students within this group. Within group *j* we have $\{SGP_{j(1)}, SGP_{j(2)}, \ldots, SGP_{j(N)}\}$. That is, we have an observed SGP for each student within group *j*.

We then produce the median growth percentile for unit *j* as:

$$\theta_j = \text{median}(SGP_{j(i)})$$

Measures of Dispersion and Precision of MGP Within Groups

There is no well-established method for computing standard errors of a sample median. For this reason, we provide three measures of variability of the group median growth percentile (MGP):

- Median absolute deviation (MAD)
- Approximate (analytic) standard error
- Bootstrap standard error

(MAD): Descriptive Measure of Dispersion

The MAD is a useful descriptive measure of variability (of dispersion around the median) within a group, but the MAD is not useful for hypothesis testing. The two alternative standard error estimators are provided for possible hypothesis testing.

The MAD within unit *j* from a group median is:

$$MAD_j = \text{median}(|SGP_{j(i)} - \theta_j|).$$

Analytic and Bootstrap Standard Errors: Measures of Precision of the MGP

The approximate standard error of the MGP is computed within group *j* as:

$$se(\theta_j) \approx 1.25 \frac{sd(\theta_j)}{\sqrt{N_j}}.$$

where $sd(\theta_j)$ is the sample standard deviation of the SGPs in group *j*, and *N* is the number of students in group *j*. The analytic standard error is limited in that it assumes a symmetric distribution around the MGP, which is not tenable for MGPs near 0 or 100. Hence, this statistic is useful for MGPs near the median, but it is less useful for MGPs at the tails of the distribution.

The bootstrap standard error within unit *j* requires the following steps performed for each group *j*:

1. Take a random sample with replacement of *n* SGPs from within group *j* where $n < L$ where L is the number of students in group *j*.
2. Compute $\theta_j^t = \text{median}(SGP_i)$ using only this sample where the superscript *t* denotes the value obtained at iteration *t*.
3. Store θ_j from iteration *t*.
4. Repeat steps 1 through 3 *Q* times with *Q* = 100.
5. Compute the variance of the vector, $var(\theta_j) = \{\theta_j^1, \theta_j^2, \ldots, \theta_j^Q\}$.
6. The bootstrap standard error for group *j* is $se_boot(\theta_j) = \sqrt{var(\theta_j)}$.

The standard error from step 6 also assumes a symmetric distribution around the MGP, which as noted above may not always be tenable. To obtain the non-symmetric 95% confidence intervals around the MGP, we then apply the additional following step:

7. Sort the values in the vector $\{\theta_j^1, \theta_j^2, \ldots, \theta_j^Q\}$ from lowest to highest. Capture the value at the 5th percentile and the 95th percentile. These are used to form the appropriate non-symmetric 95% confidence limits on the MGP.

Aggregating Medians Using Student Weights

MGPs are computed for each group *j* for each year *t*. To lessen some of the observed year-to-year variability in the MGPs, we compute an aggregated MGP by combining them over time within a subject and weighting by the number of students who took the test. The weighted median and its standard error will be more efficient than a simple arithmetic average of the two medians as the MGP estimated on the basis of the larger sample size will have more weight in the final result. Aggregating medians as if they were means provides only an approximation and is one option for combining MGPs over time. Other options were considered by OSSE, and this method was chosen given its transparency relative to the other options.

For convenience, the subscript *s* for subjects has been dropped. Hence, this same calculation is performed separately for reading and math.

Define the weights for group *j* at time *t* as

$$w_{jt} = \frac{n_{jt}}{\sum_{t=1}^{T} n_{jt}},$$

where n_{jt} is the number of students in group j for time t.

We then aggregate the medians over time as

$$\varphi_j = \sum_{t=1}^{T} w_{jt}\theta_{jt}$$

where θ_{jt} is the group median MGP at time t. The variance of the estimate φ_j is a function of the weights and the variances of the component estimates and is computed as:

$$var(\varphi_j) = \sum_{t} w_{jt}^2 var(\varphi_{jt})$$

where $var(\varphi_{jt})$ is the bootstrap standard error squared for each group median.

The Impact of Measurement Error in Quantile Regression

The previously described model uses the observed test scores in the estimation of the quantile regression. However, this will introduce bias into the estimation of the model parameters if there is measurement error in the observed test scores (Wei & Carroll, 2009), which subsequently propagates to the SGPs and MGPs.

AIR is currently developing an error-in-variables approach for the quantile regression that mitigates the bias related to the measurement error in the predictor variables. However, because the current model does not incorporate an error-in-variables correction, it is important to note two consequences associated with estimating the QR based on the observed scores.

First, the parameter estimates used to produce the SGPs will be biased toward 0 by an unknown amount. Because the model coefficients are used to produce SGPs and because SGPs are also based on the observed scores, the bias from the measurement error will propagate to the SGPs and the MGPs.

Second, the standard errors of the coefficients are underestimates of the true variance since they will not include the measurement variance. In practice this means that the analytic and bootstrap standard errors are smaller than they would otherwise be if they accounted for both the sampling variance and measurement error. In practice, OSSE may observe year-to-year variation in the MGPs that is larger than the standard errors would indicate as a result of this underestimate.

RESULTS

AIR analyzed data from the 2009-10 and 2010-11 school years for the District of Columbia. A summary of results is provided here.

Within-Year Precision of the School MGPs

We first examine the precision of the school MGPs within a school year. The two caterpillar plots in Figures 1 and 2 show the school MGPs in reading and mathematics with a non-symmetric 95% confidence interval around each MGP. The *x*-axis sorts the schools from lowest to highest based on their MGPs in each plot. The *y*-axis plots the school MGP and the non-symmetric 95% confidence interval.

Both plots show that there is variability in school performance across the district with some schools having very low MGPs and other schools having high MGPs. However, we also see that the MGPs in both subjects are estimated with some imprecision. The median standard error across schools of the 2011 MGPs is 5.55 in math and 6.16 in reading.

As previously noted, MGPs are bounded between 0 and 100, and so the standard errors imply symmetry around the MGP, which is not always tenable. It is also useful to examine the typical width of the non-symmetric 95% confidence intervals for each subject as well. Table 2 below shows the median width of the 95% confidence interval in reading and math and the range of the confidence intervals by subject.

Table 2: Median Width of Confidence Intervals and Range (2011 SY)

Subject	Median Width	Range
Reading	16.1	6 to 48
Math	14.5	6 to 58

Figure 1: Schoolwide Mathematics MGP with 95% Confidence Interval (2011 SY)

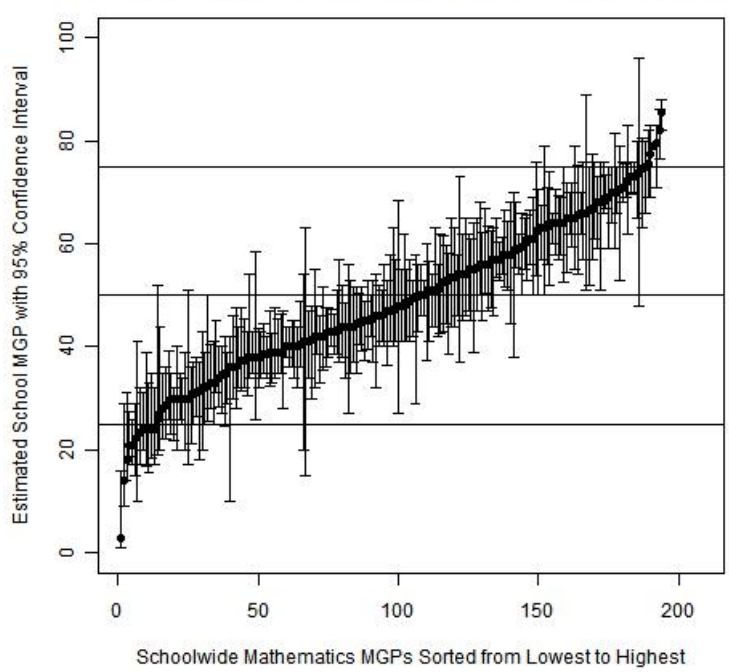
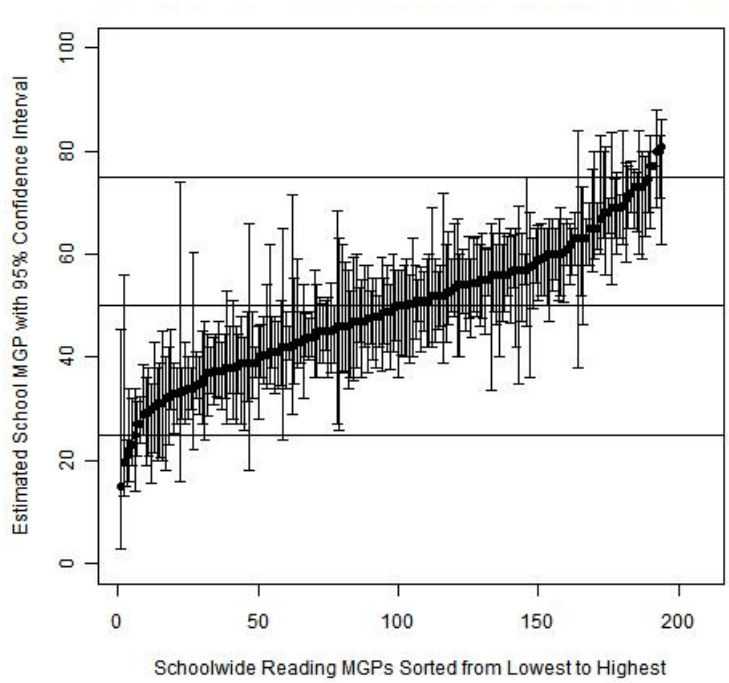


Figure 2: Schoolwide Reading MGP with 95% Confidence Interval (2011 SY)



Between-Year Stability of the School MGPs

Figures 3 and 4 show the relationship of the MGP for a school between the 2009–2010 school year and the 2010–2011 school year. If the school estimates over time were perfectly stable, they would fall along a 45-degree line, and the correlation would be 1. The correlation between the results in reading over time is .60, and the correlation in the mathematics MGPs over time is .46.

As a heuristic, the graph is divided into six quadrants with lines at the 33rd percentile and 67th percentile on the x- and y-axes. These quadrants can be used to examine if schools have reversals in their performance over time. For example, we may observe a school with an MGP of 20 in year one but with an MGP of 80 in year 2.

There are at least two plausible explanations for the variation over time. First, schools change in their instructional effectiveness over time, and we would anticipate differences in student growth as a function of this change. Second, the within-year estimates have variability, as the prior section shows. Hence, it is reasonable to expect changes in performance over time as a function of the imprecision in the estimated statistic.

Figures 5 and 6 show the year-to-year variability in the MGPs as a function of school size. A "small" school is one that has 100 or fewer students, "medium" is 200 or fewer, and "large" is 200 or more students. The plots indicate that the observed year-to-year variation is slightly larger for smaller schools relative to larger schools.

Figure 3: Mathematics MGPs over Time

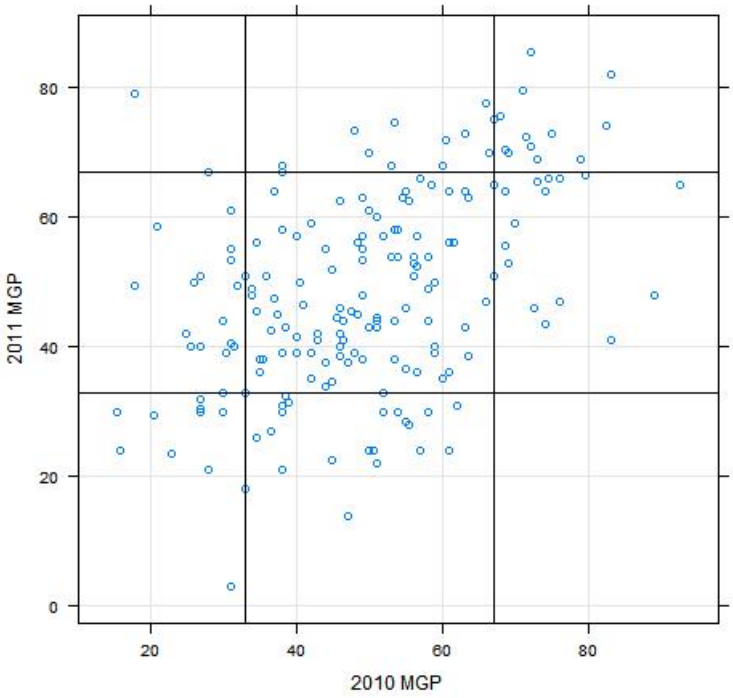


Figure 4: Reading MGPs over Time

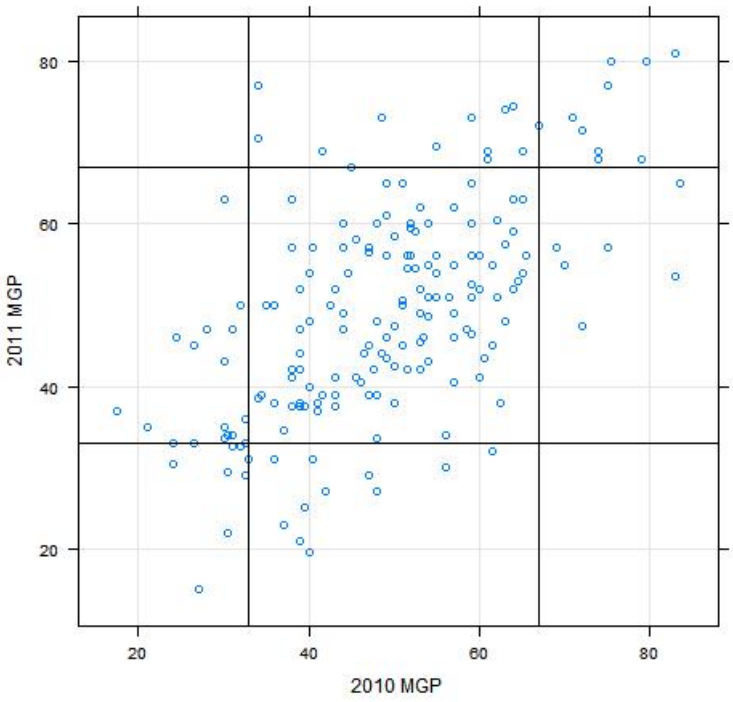


Figure 5: Mathematics MGPs over Time

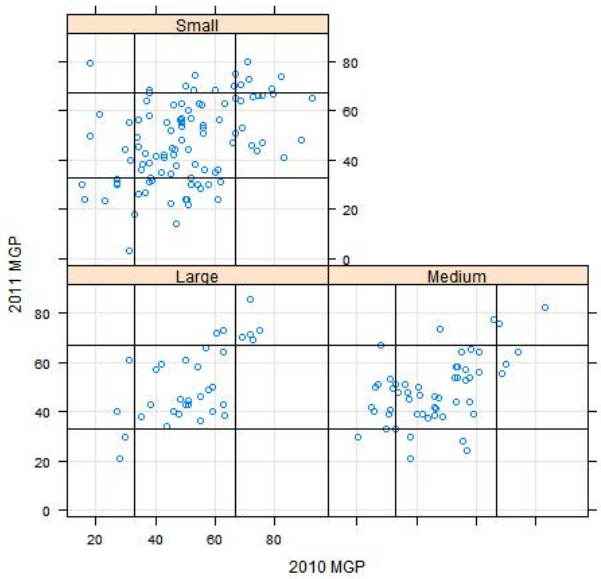
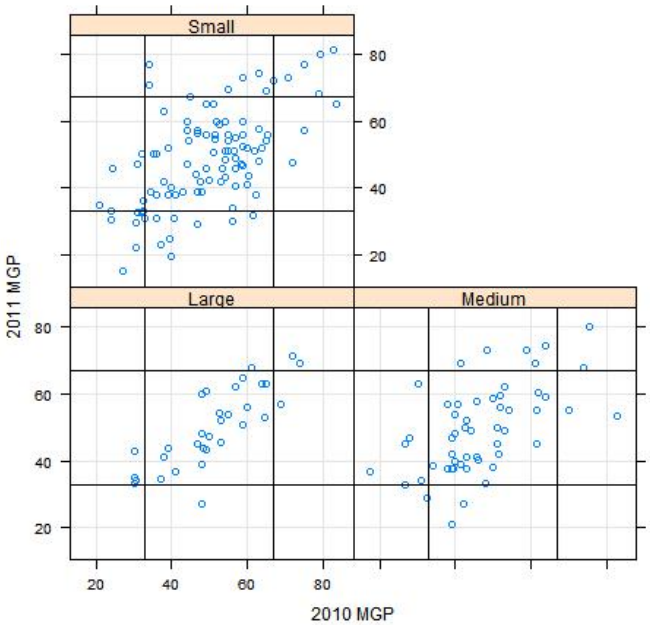


Figure 6: Reading MGP over Time



A second way to examine the year-to-year variability is via a frequency table of the changes in MGPs over time. Table 3 below shows the percentage of schools with MGP changes larger than 10 points, 20 points, 30 points, 40 points, or 50 points over time. For example, 32.7% of the schools have MGP changes of 10 or larger in reading, and 38.5% of the schools have MGP changes of 10 or larger in mathematics.

Table 3: Frequency Distribution of MGPs Changes Over Time

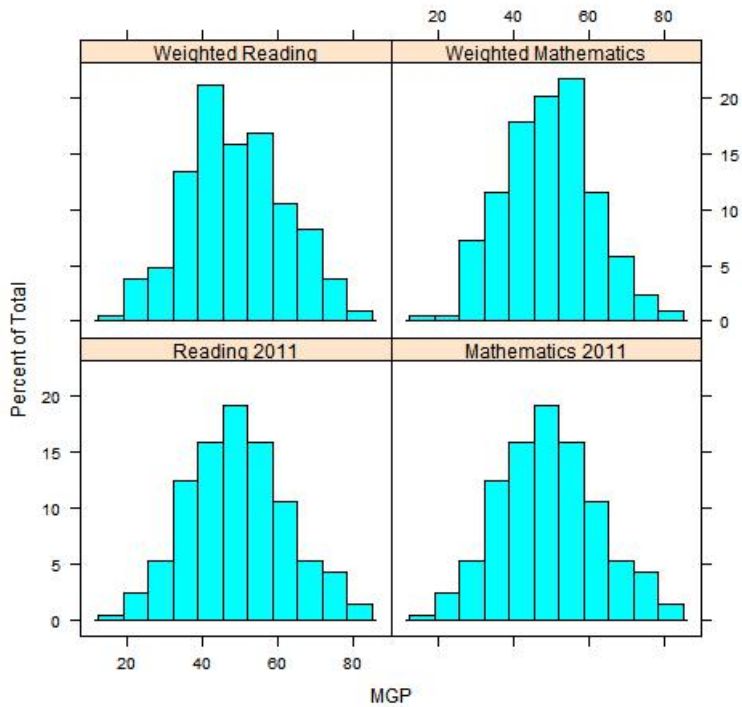
Subject	Less than 10	Greater than 10	Greater than 20	Greater than 30	Greater than 40	Greater than 50
Reading	57.72%	32.7%	7.7%	1.4%	.48%	0%
Mathematics	35.52%	38.5%	18.8%	5.3%	1.4%	.48%

Aggregated MGP Over Time

Given the observed year-to-year variability, it is useful to consider a combined MGP over time. That is, the individual point-in-time MGPs have year-to-year variation that is dampened by combining them to produce a combined MGP using the point-in-time values.

The histogram in Figure 7 shows the distribution of the school MGPs in 2011 and the combined MGPs over both years. The histograms are stacked, making it easy to compare the distribution of MGPs. The histograms show that the MGPs are similarly distributed in the aggregate relative to the 2011 estimates.

Figure 7: Conditional Distribution of School MGP



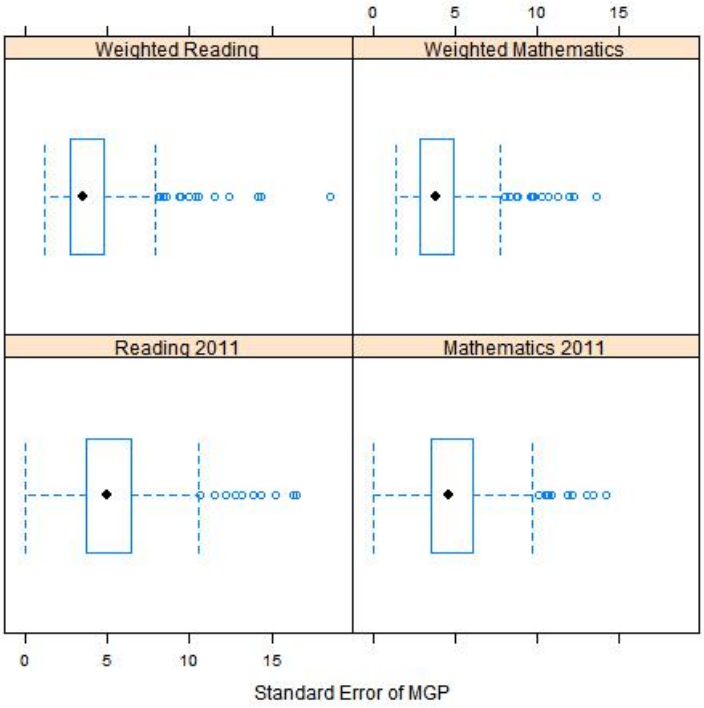
While the distribution of MGPs may appear similar, the standard errors of the combined MGPs are smaller than the 2011 standard errors. Hence, the precision of the estimates improves by aggregating over time.

The box plots in Figure 8 show the median standard errors and their distribution of the combined MGPs relative to the standard errors of the 2011 estimates and conditional on the weighted and 2011 MGP values. The dark dot in the center of each box is the median standard error; the solid

left and right lines around the dot are the 25th and 75th percentiles, respectively, and the whiskers to the left and right are the standard errors at the 5th and 95th percentiles, respectively.

The 2011 estimates are larger on average than the combined standard errors, and there is less variability in the standard errors. This indicates that standard errors are smaller for all schools in the combined MGP relative to the 2011 point-in-time values.

Figure 8: Conditional Distribution of MGP Standard Errors



Analysis of Student Growth Percentiles by Student Group

The plots in this section are descriptive and show the differential performance of various student groups across the District using the 2011 school year data. The intention in showing these plots is only to provide descriptive statistics on the SGPs. The plots showing the differences across demographic groups are not necessarily checks on how well the model behaves. However, the plots that follow the demographics can be used as checks on the model behavior.

Figure 9 shows the conditional distribution of student growth percentiles (SGPs) by gender. Female students appear to have slightly larger SGPs than male students in both subjects; however, the difference is small, and the variability is comparable between both groups.

Table 4: Median SGP by Gender

	Female	Male
Reading	51	48
Mathematics	50	48

Figure 9: Distribution of Student Growth Percentiles by Subject and Gender

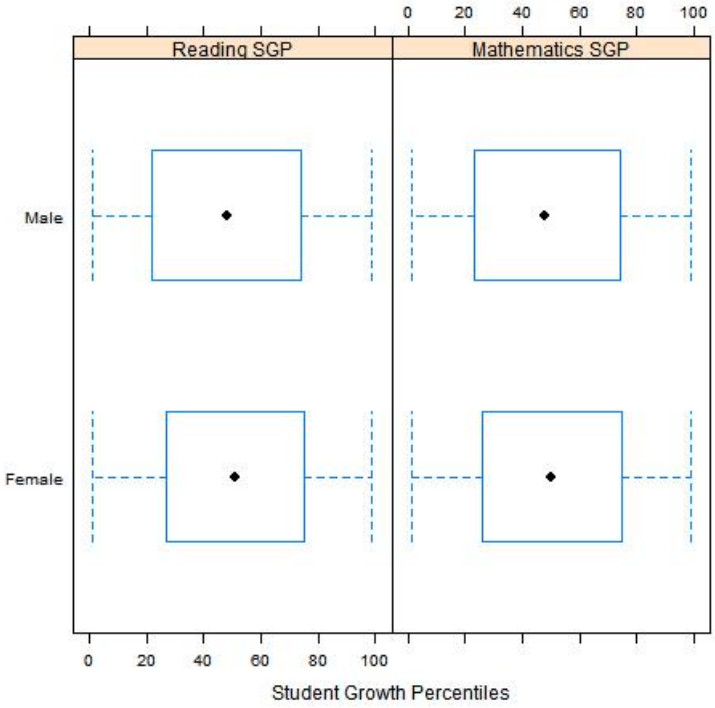


Figure 10 shows the conditional distribution of SGPs by race for reading and mathematics. The differences in both subjects are similar across groups. For instance, White students tend to have higher SGPs relative to the other groups, with Asian students having the second largest SGPs.

Table 5: Median SGP by Ethnicity

	Asian	African American	Hispanic	American Indian	White
Reading	61	48	52	54	64
Mathematics	62	48	53	54	62

Figure 10: Distribution of Student Growth Percentiles by Subject and Race

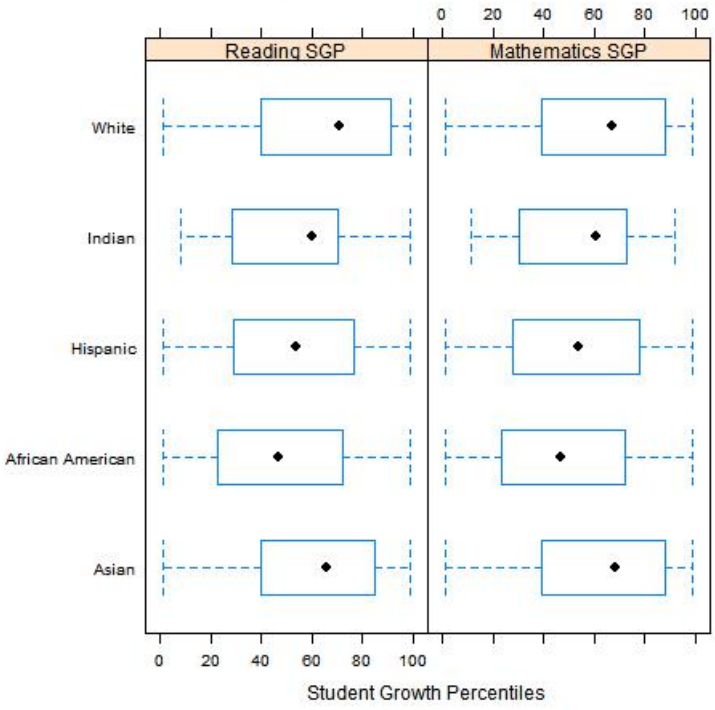


Figure 11 shows that students enrolled in a Limited English Proficiency (LEP) program have SGPs that are slightly larger than their non-LEP counterparts. One plausible hypothesis regarding this difference raised by the SCAG is that LEP students have initial test scores that may underrepresent their true performance. For example, an LEP student's baseline score may be too low relative to his or her true ability as the student's limited English proficiency might impede his or her ability to demonstrate true mastery of the tested content. However, after a year of learning in the United States, the student's English proficiency may improve, and he or she may be able to better represent an understanding of the tested content.

Table 6: Median SGP by LEP Status

	Not LEP	**LEP**
Reading	49	50
Mathematics	49	51

Figure 11: Distribution of Student Growth Percentiles by Subject and LEP Status

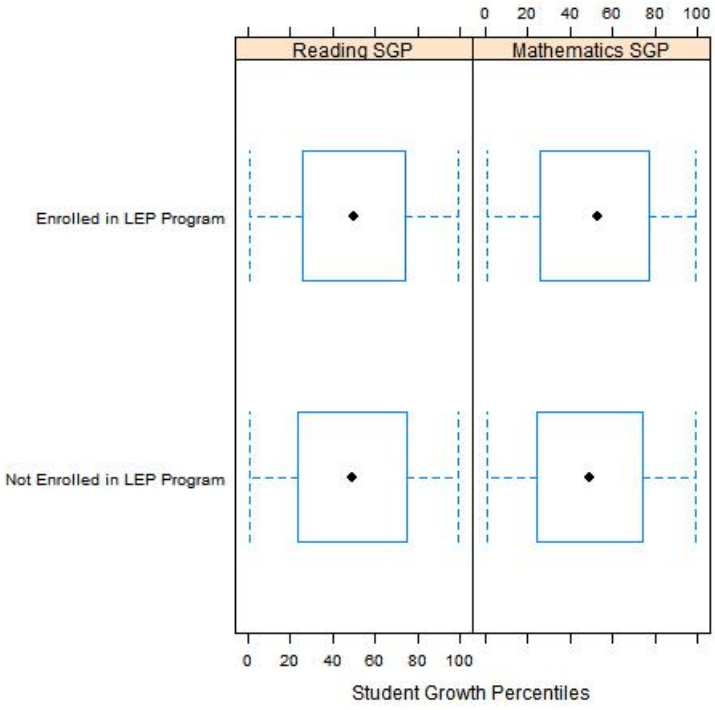
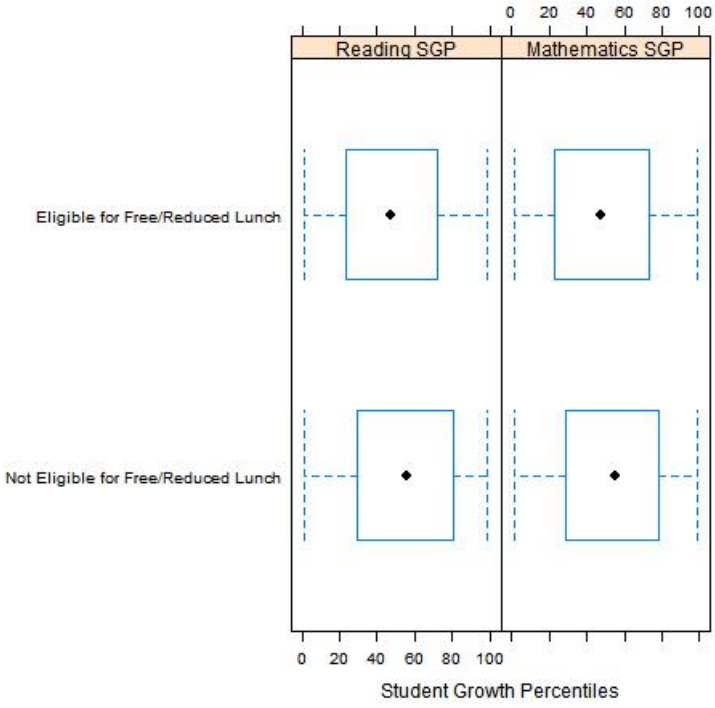


Figure 12 shows that students eligible for free or reduced price lunch (FRPL) have slightly lower SGPs relative to students not eligible for this program.

Table 7: Median SGP by FRPL Status

	Not Eligible for FRPL	Eligible for FRPL
Reading	54	48
Mathematics	53	48

Figure 12: Distribution of Student Growth Percentiles by Subject and FRPL Status

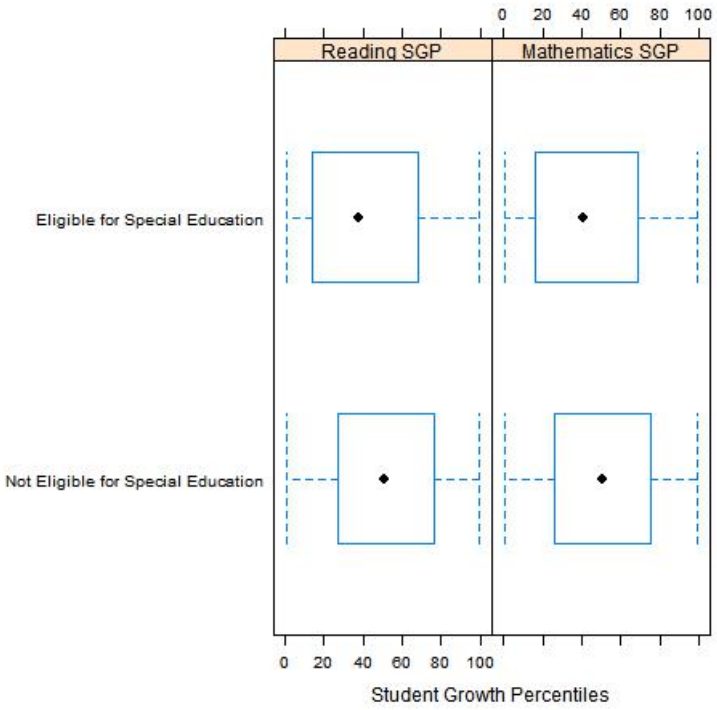


Last, Figure 13 shows that students enrolled in special education programs (SPED) tend to have lower SGPs than students not enrolled in special education programs.

Table 8: Median SGP by Special Education Status

	Not SPED	SPED
Reading	51	42
Mathematics	51	44

Figure 13: Distribution of Student Growth Percentiles by Subject and SPED Status



Relationship of Student Growth Percentiles to Prior Year Scores

It is useful to examine the relationship of the SGPs to the prior year scores for all students in both subjects via a scatterplot. This plot can illustrate whether there is any potential advantage or disadvantage in the student SGPs conditional on prior scores. For instance, if students with high test scores to begin with tend to have higher SGPs, the model would then favor higher performing students, and schools with these students would tend to have higher MGPs. Conversely, if students with lower test scores to begin with tend to have higher SGPs, the model would then favor lower performing students, and schools with these students would tend to have higher MGPs.

Figure 14 shows the relationship between SGPs and prior test scores in reading for all grades. There is no apparent systematic trend in the data. Hence, we can assume that high or low SGPs are not correlated with prior test scores. This suggests that students can earn a high or low SGP no matter what score they have in the prior school year. There does not appear to be any particular advantage or disadvantage in the growth estimates conditional on prior scores.

Figure 14: Reading SGPs and Prior Scores

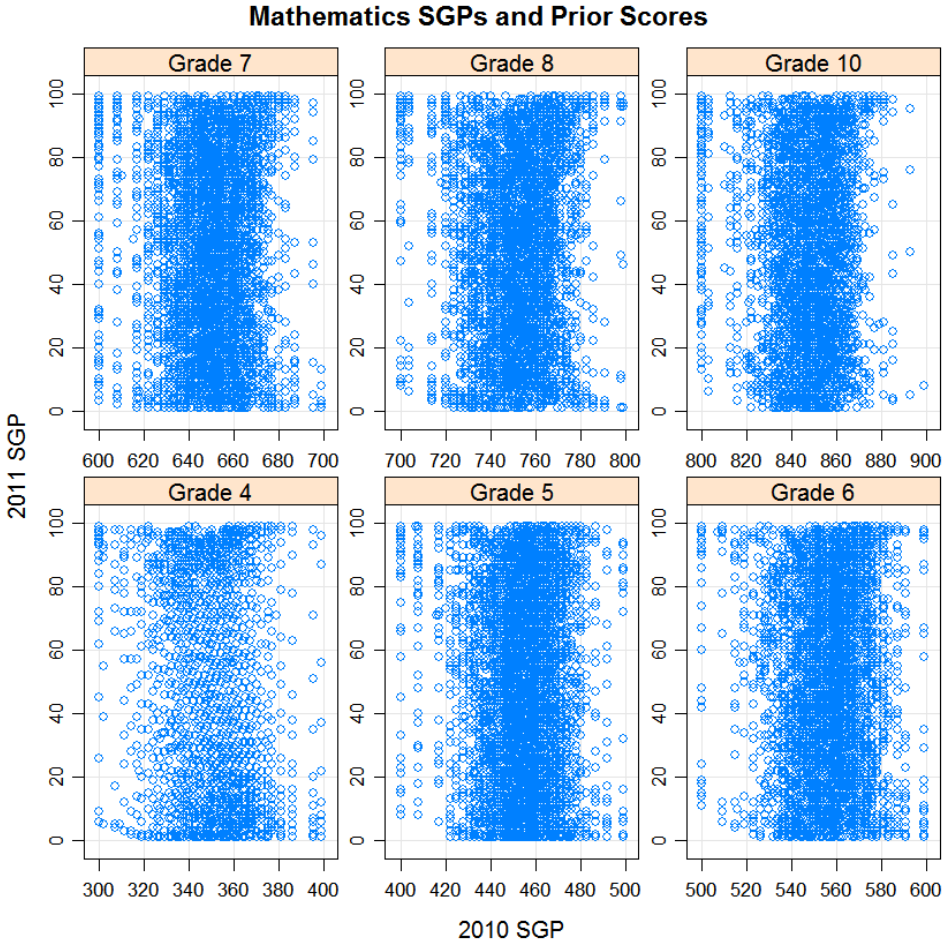
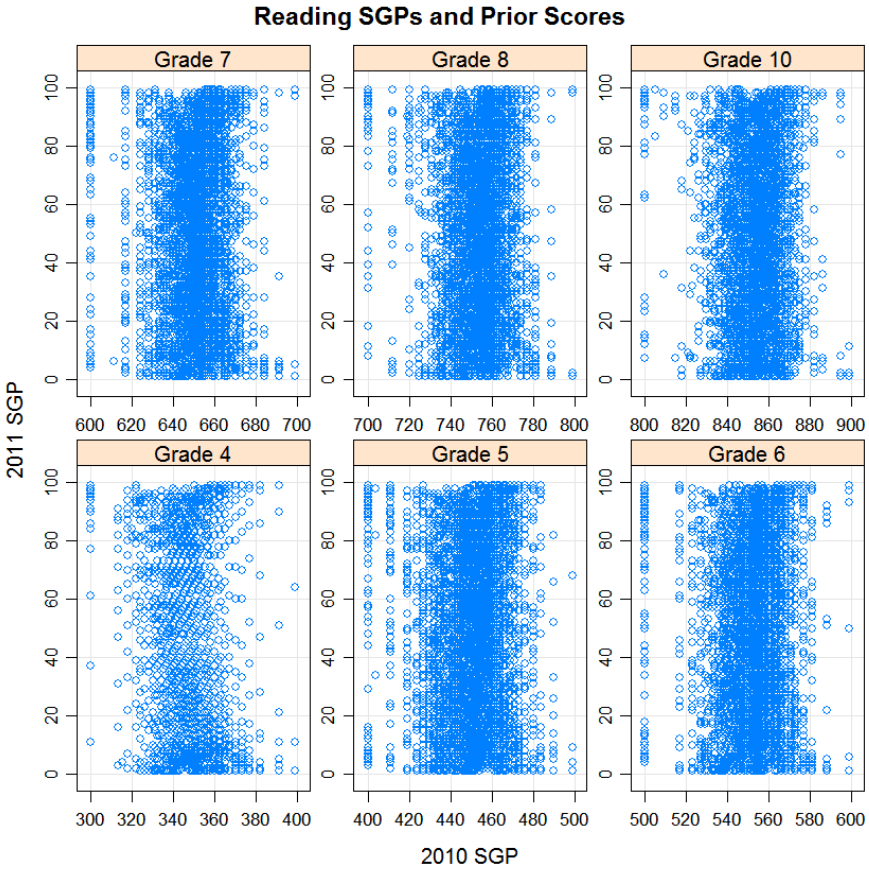


Figure 15 shows the relationship between SGPs and prior test scores in mathematics for all grades. Again, there is no apparent systematic trend in the data. Hence, we can assume that high or low SGPs are also not correlated with prior test scores in mathematics. This suggests that students can earn a high or low SGP no matter what score they have in the prior school year. There does not appear to be any particular advantage or disadvantage in the growth estimates conditional on prior scores.

Figure 15: Mathematics SGPs and Prior Scores



Stability of Student Growth Percentiles Over Time

Table 9 shows the correlation in the student-level SGPs from the 2009–2010 school year and the 2010–2011 school year. The table includes only grades 5 to 8 because grade 4 students do not have grade 3 SGPs, and grade 10 students do not have SGPs from grade 9. Therefore, only students in grades 5 to 8 have pairs of SGPs over two consecutive school years.

Grades 5, 6, 7, and 8 in both subjects show correlations near 0, indicating almost no correlation in the relationship of SGPs over time. This year-to-year variability is very large and suggests that the student growth measure varies drastically from year-to-year.

Table 9: Correlation in Student-Level SGPs from the 2009–1010 and 2010–2011 School Years

Grade	Mathematics	Reading
Grade 5	0.02	0.03
Grade 6	-0.04	-0.04
Grade 7	-0.02	-0.03
Grade 8	-0.01	-0.04

Table 10 below shows the mean absolute difference in the SGP from 2010 to 2011 in reading and math. The mean SGP difference for a student over time is around 33 SGP points. This means, for example, that a student at the 50th percentile in year 1 may have a growth percentile as low as 17 in year 2 or as high as 83 in year 2.

Table 10: Mean Absolute Difference in SGP from 2010 to 2011 in Reading and Mathematics

Grade	Mathematics	Reading
Grade 5	32 points	32 points
Grade 6	34 points	34 points
Grade 7	33 points	33 points
Grade 8	33 points	33 points

Given this variability in the year-to-year estimates of the SGP, it seems that this student-level statistic is not precise enough to use when making student-level decisions regarding a student's instructional progress.

CONCLUSION

The student growth percentile model implemented for schools in Washington, DC, represents a model chosen by OSSE and guided in its implementation by the SGAC.

The results of the model analysis suggest that the cross-year correlation in SGPs is small, suggesting volatility in these scores. School-level MGPs also show some cross-year variability. Given the year-to-year variation in MGPs, a two-year combined median is provided to use as the potentially more reliable statistic at the school level. This weighted median always has a standard error smaller than either of the components used to compute the weighted median.

While this report documents the implementation of the model for the 2009–2010 and 2010–2011 school years, OSSE will continue to examine this model and ways in which it can be improved in the future.

APPENDIX A. SCHOOLWIDE GROWTH ADVISORY COMMITTEE (SGAC) MEMBERS

Joshua Boots (KIPP DC)

Lydia Carlis (Apple Tree Institute)

Steve Cartwright (DC Public Schools)

Naomi Deveaux (FOCUS)

Prudence Hallarman (DC Prep)

Zac Morford (Friendship Public Charter School)

Unique Morris (Septima Clark Public Charter School)

Jeff Noel (FOCUS)

Megan Reamer (Capital City Public Charter School)

APPENDIX B. BUSINESS RULES

BUSINESS RULES

District of Columbia Schoolwide Growth Project

All student growth models rely on student achievement data and data describing how students are linked to courses, teachers, and schools. For a variety of reasons, these data often contain inconsistencies or are incomplete. Given this reality, states or districts adopting growth models must make decisions not just about which model to use, but about how to implement it once selected – the "business" or "operating rules" for the model. These decisions fall into several categories:

1. **Student Inclusions, Exclusions, and Attribution** – identification of students whose data will be included and excluded from analysis, along with rules for which students will be attributed to which schools, and how.
2. **School Exclusions –** identification of schools for which growth scores will not be estimated, or rules for how to generate scores in cases of school consolidation, etc.
3. **Prior Achievement** – specification of which prior achievement scores and how many years of scores to take into account in the growth model.

This document describes rules in each of these categories. Datafiles will be prepared for analysis based upon these rules.

Student Inclusions, Exclusions, and Attribution

Districts and states often perform extensive data checks and edits to make certain that their student data are as "clean" and accurate as possible, particularly in the case of student achievement scores used for accountability purposes. However, despite these best efforts, some student data may retain inaccuracies. In addition, some students may be in grades for which growth estimates cannot be computed. In general, AIR will include in analysis any student who attended an institution in DC identified as a school program (e.g. youth service center, alternative programs).

Table B–1 describes specific guidelines for exclusion of student data in the DC student growth analysis. Numbers of cases excluded will be documented and checked to see if any patterns emerge.

Table B–1: Student Data Exclusions[2]

	Scenario	Source of Rule Recommendation	Include/Exclude in School Growth Estimates (based on July 20 meeting)
S.1	Students with no current year test scores (e.g. students in grades 9, students whose scores may have been invalidated, students who may be medically exempt from testing)	AIR proposed	Exclude students from analysis
S.2	Students with multiple test score records in a single year with contradictory grades (e.g. enrolled in one school as grade 5 and in another as grade 6)	AIR proposed	Exclude students from analysis
S.3	Students with a current tested grade lower than a prior tested grade	AIR proposed	Exclude test scores from atypical grade progressions from analysis (replace with a missing flag and maintain student record in analysis – for example, a student with a 7th grade outcome score whose immediate prior score is from 8th grade and scores before that are from 6th grade and 5th grade is kept in analysis, but the 8th grade score is dropped)
S.4	Students who take the DC CAS Alternate Assessment	AIR proposed	Exclude students from analysis (too few students, different test scale)
S.5	Students with missing, invalid, or duplicate ID numbers	AIR proposed	Exclude students from analysis (not possible to merge student records over time)
S.6	Students who repeat a test(i.e. students who have the same tested grade in consecutive years))	AIR proposed	Exclude test scores from atypical grade progressions from analysis (replace with a missing flag and maintain student record in analysis – for example, a student with a 7th grade outcome score whose immediate prior score is from 7th grade and scores before that are from 6th grade and 5th grade is kept in analysis, but the 8th grade score is dropped)
S.7	Students with truly duplicate test records in a given year (no contradictory data)	AYP Rules	Exclude duplicate student records from analysis
S.8	Students with no prior test scores (or invalid prior test scores)		Exclude students with no prior test scores from up to four years back from analysis
S.9	Two or more students have the same name, ID, DOB but different scores in a given year	AYP Rules	Exclude students from analysis (not possible to merge student records over time)
S.10	Students who are flagged in the OSSE data file as "exclude from reports"	AYP rules	Exclude students from school-level reportig

[2] Note that the Student Growth Advisory Committee recommended a rule that students without Full Academic Year (FAY) status be included in analysis but excluded from school-level reporting. This rule will be implemented in future growth analyses. Analysis of the current two-year median scores in reading and mathematics showed that few schools' scores would differ with implementation of this rule.

	Scenario	Source of Rule Recommendation	Include/Exclude in School Growth Estimates (based on July 20 meeting)
	(including for instance, students placed in private settings)		
S.11	LEP/NEP students who have been in the U.S. for more than 12 months	AYP Proficiency Rules (AYP calls for including these students in participation and proficiency calculations)	Include available student scores in analysis and reporting (note: students may have been exempted from reading test in prior years but may have math scores)
S.12	LEP/NEP students who have been in the U.S. for less than 12 months	AYP Proficiency Rules (AYP calls for including these students in participation but NOT in proficiency calculations)	Include students in analysis; exclude students with "NewtoUS" flag from school-level reporting (note: students may be exempted from reading test but may have math scores, or may take reading test even though not required to)
S.13	Exited LEP/NEP students (became fully English proficient within the last two years)	AYP Proficiency Rules	Include students in analysis and reporting
S.14	Exited students with disabilities	AYP Proficiency Rules	Include students in analysis and reporting

School Exclusions

Table B–2 provides proposed rules for generating schoolwide growth scores for schools.

Table B–2: School Exclusions

	Scenario	Source of Rule Recommendation	Inclusion/Exclusion Rule (based on July 20 meeting)
Sch.1	Schools that have merged or consolidated with other schools (without new school status)	AYP Rules	Include in reporting if schools have a valid school code (and are not excluded from school-level reporting for other reasons)
Sch.2	A new school (school may be treated as new if 50% or more of the grade spans or population have changed in the school)	AYP Rules	Include in reporting if schools have a valid school code (and are not excluded from school-level reporting for other reasons)
Sch.3	Alternative education program, youth service center, or technical education schools	AIR proposed rule based on AYP student inclusion rules	Include in reporting (if not excluded from school-level reporting for other reasons)

Prior Achievement

All growth models take students' prior achievement into account. Determining which prior achievement scores to use in predicting performance in a particular subject is a key decision to

be made in implementing a growth model. Given the characteristics of DC's test data – in particular, the lack of vertical scales -- the growth model to be used will not be measuring growth in the strictest sense of the term (i.e. score point gains from year to year). Instead, we will be predicting students' achievement based on their prior achievement. Prior achievement will be in the same subject and covering the same general content. In DC, tests at the elementary and high school grade levels include a variety of content aimed at measuring a broad set of knowledge and skills, such as "math" or "reading" each year. In these cases, determining which test should serve as a predictor of future achievement may be relatively straightforward.

In addition to determining which prior test scores to use as predictors, it is also necessary to determine how many years of prior achievement ought to be included as a predictor. The benefit of including additional years of data is that it may improve the precision of the prediction and reduce bias. However, including many years of data will complicate the model, and because test scores are correlated from year to year, may provide limited additional information. It will also increase the likelihood of missing scores. AIR will test a model including two prior years of test scores (where available) and a model including one prior year of achievement to check if the addition of more test prior test score data provides additional explanatory value.

Table B–3 provides rules for the use of prior achievement scores. For growth models with one year of prior achievement data, we will simply take the most recent prior year instead of the two most recent.

Table B–3: Prior Achievement Scores

Test	Prior Achievement Predictor
DC CAS Grade 4 Mathematics	DC CAS Grade 3 Mathematics (only 1 prior score available)
DC CAS Grade 4 Reading	DC CAS Grade 3 Reading (only 1 prior score available)
DC CAS Grade 5 Mathematics	DC CAS Grade 4 Mathematics DC CAS Grade 3 Mathematics
DC CAS Grade 5 Reading	DC CAS Grade 4 Reading DC CAS Grade 3 Reading
DC CAS Grade 6 Mathematics	DC CAS Grade 5 Mathematics DC CAS Grade 4 Mathematics
DC CAS Grade 6 Reading	DC CAS Grade 5 Reading DC CAS Grade 4 Reading
DC CAS Grade 7 Mathematics	DC CAS Grade 6 Mathematics DC CAS Grade 5 Mathematics
DC CAS Grade 7 Reading	DC CAS Grade 6 Reading DC CAS Grade 5 Reading
DC CAS Grade 8 Mathematics	DC CAS Grade 7 Mathematics DC CAS Grade 6 Mathematics
DC CAS Grade 8 Reading	DC CAS Grade 7 Reading DC CAS Grade 6 Reading
DC CAS Grade 10 Mathematics	DC CAS Grade 8 Mathematics DC CAS Grade 7 Mathematics
DC CAS Grade 10 Reading	DC CAS Grade 8 Reading DC CAS Grade 7 Reading

APPENDIX C. DATA OVERVIEW

DATA OVERVIEW

We provide here background on the test scores used as inputs to the model, as well as other relevant data characteristics. AIR conducted three types of analyses: a data accuracy check, a merging accuracy check, and descriptive analyses aimed at identifying any characteristics of the test score scales that should be taken into consideration in the growth model.

The first section examines data accuracy. It provides descriptive statistics for key variables relating to student achievement and tested grade (for 2006-07 through 2010-11), data merge rates (for 2009-10 and 2010-11) and school assignment (for 2008-09, 2009-10, and 2010-11). Importantly, it points out areas where the data were consistent and inconsistent with expectations. The second section, which covers the merging of the data across years, illustrates the extent to which students can be linked to multiple years of test score data.

DATA INTAKE ACCURACY CHECK

Student data analysis began with checking accuracy and flagging unexpected or implausible values and ranges. The quality of student reading and mathematics scores, enrollment in the tested grades, and school codes is described below, disaggregated by academic year.

READING DC CAS SCALE SCORES

Tables C–1 through C–5 below provide descriptive statistics for the reading scale scores for the DC CAS for grades 3-8 and 10.

By design, DC CAS scale scores follow a consistent pattern i.e., grade 3 scores range from 300 to 399; grade 4, scores range from 400 to 499, and so on up to grade 8. Grade 10 ranges from 900 to 999. Data from all five years follow this pattern and are consistent with expectations.

In general, the mean scores for each grade are near the middle of the scale (e.g., 651.10 for grade 6 in 2010-11). The standard deviations are also quite consistent across grades.

Table C–1: Reading DC CAS Scale Score Descriptive Statistics: 2010-11

Grade	Number of Observations	Mean	Standard Deviation	Minimum	Maximum
3	4,786	348.24	16.38	300	399
4	4,826	450.87	15.79	400	499
5	4,725	552.89	14.57	500	599
6	4,359	651.10	14.11	600	699
7	4,418	753.95	13.84	700	799
8	4,277	853.71	14.82	800	899
10	4,266	952.11	16.2	900	999

Table C–2: Reading DC CAS Scale Score Descriptive Statistics: 2009-10

Grade	Number of Observations	Mean	Standard Deviation	Minimum	Maximum
3	4,894	348.81	16.81	300	399
4	4,781	451.17	15.67	400	499
5	4,452	552.76	14.9	500	599
6	4,474	650.94	14.47	600	699
7	4,327	753.85	13.62	700	799
8	4,477	853.72	14.87	800	899
10	4,100	951.93	13.86	900	999

Table C–3: Reading DC CAS Scale Score Descriptive Statistics: 2008-09

Grade	Number of Observations	Mean	Standard Deviation	Minimum	Maximum
3	5,018	351.41	13	300	395
4	4,537	451.75	15.19	400	499
5	4,679	553.29	14.33	500	599
6	4,450	653.31	12.94	600	699
7	4,581	752.04	14.13	700	799
8	4,477	853.18	13.19	800	899
10	3,834	952.8	13.27	900	999

Table C–4: Reading DC CAS Scale Score Descriptive Statistics: 2007-08

Grade	Number of Observations	Mean	Standard Deviation	Minimum	Maximum
3	4,679	351.44	13.77	300	399
4	4,606	451.83	14.61	400	498
5	4,615	552.73	14.46	500	599
6	4,677	650.87	13.72	600	699
7	4,603	751.51	13.22	700	799
8	4,969	851.07	13.95	800	899
10	3,831	951.8	13.37	900	999

Table C–5: Reading DC CAS Scale Score Descriptive Statistics: 2006-07

Grade	Number of Observations	Mean	Standard Deviation	Minimum	Maximum
3	4,827	349.97	13.01	300	399
4	4,708	449.59	14.38	400	499
5	4,864	551.32	13.75	500	599
6	4,767	650.89	13.87	600	699
7	5,148	749.89	13.22	700	799
8	4,963	849.21	14.77	800	899
10	3,982	949.77	15.06	900	999

MATHEMATICS DC CAS SCALE SCORES

Tables C–6 through C–10 present the descriptive statistics for the mathematics scale scores for the DC CAS for grades 3-8 and 10.

DC CAS scale scores for math are designed in a similar fashion to the reading scores. Data from all five years follow this pattern and are consistent with expectations. Mean scores for each grade are also near the middle of the scale (e.g., 752.52 for grade 7 in 2010-11). The standard deviations are also quite consistent across grades.

Table C–6: Mathematics DC CAS Scale Score Descriptive Statistics: 2010-11

Grade	Number of Observations	Mean	Standard Deviation	Minimum	Maximum
3	4,812	352.59	19.27	300	399
4	4,857	455.26	16.05	400	499
5	4,795	556.30	16.53	500	599
6	4,388	650.65	16.82	600	699
7	4,448	752.52	17.19	700	799
8	4,321	850.66	15.86	800	899
10	4,241	945.53	18.79	900	999

Table C–7: Mathematics DC CAS Scale Score Descriptive Statistics: 2009-10

Grade	Number of Observations	Mean	Standard Deviation	Minimum	Maximum
3	4,917	352.27	18.46	300	399
4	4,807	454.84	15.56	400	499
5	4,468	556.74	15.81	500	599
6	4,494	649.40	16.80	600	699
7	4,339	750.94	16.87	700	799
8	4,485	847.71	16.58	800	899
10	4,086	944.70	18.30	900	999

Table C–8: Mathematics DC CAS Scale Score Descriptive Statistics: 2008-09

Grade	Number of Observations	Mean	Standard Deviation	Minimum	Maximum
3	5,050	355.50	19.26	300	399
4	4,569	456.39	16.68	400	499
5	4,698	556.19	16.99	500	599
6	4,463	649.13	16.88	600	699
7	4,581	749.10	17.63	700	799
8	4,477	846.02	16.11	800	899
10	3,826	945.53	19.04	900	999

Table C–9: Mathematics DC CAS Scale Score Descriptive Statistics: 2007-08

Grade	Number of Observations	Mean	Standard Deviation	Minimum	Maximum
3	4,677	352.76	18.29	300	399
4	4,607	454.48	15.81	400	499
5	4,615	554.82	17.17	500	599
6	4,680	647.63	17.39	600	699
7	4,606	746.66	16.54	700	799
8	4,972	843.87	16.54	800	899
10	3,829	944.67	19.03	900	999

Table C–10: Mathematics DC CAS Scale Score Descriptive Statistics: 2006-07

Grade	Number of Observations	Mean	Standard Deviation	Minimum	Maximum
3	4,849	351.14	17.26	300	399
4	4,726	451.13	16.2	400	499
5	4,875	550.16	16.6	500	599
6	4,775	644.92	16.65	600	699
7	5,142	743.13	17.14	700	799
8	4,949	842.42	16.16	800	899
10	3,922	941.92	18.77	900	999

ENROLLMENT IN THE TESTED GRADES

Tables C–11 and C–12 show the frequency and percentage of students in each tested grade for the 2006-07, 2007-08, 2008-09, 2009-10, and 2010-11 school years, for reading and mathematics, respectively. No anomalies are detected.

Table C–11: Student Counts by Tested Grade and Year – Reading DC CAS

	2007		2008		2009		2010		2011	
Grade	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**
3	4,827	14.5%	4,679	14.6%	5,018	15.9%	4,894	15.5%	4,786	15.1%
4	4,708	14.2%	4,606	14.4%	4,537	14.4%	4,781	15.2%	4,826	15.2%
5	4,864	14.6%	4,615	14.4%	4,679	14.8%	4,452	14.1%	4,725	14.9%
6	4,767	14.3%	4,677	14.6%	4,450	14.1%	4,474	14.2%	4,359	13.8%
7	5,148	15.5%	4,603	14.4%	4,581	14.5%	4,327	13.7%	4,418	14.0%
8	4,963	14.9%	4,969	15.5%	4,477	14.2%	4,477	14.2%	4,277	13.5%
10	3,982	12.0%	3,831	12.0%	3,834	12.1%	4,100	13.0%	4,266	13.5%
Total	33,259	100%	31,980	100%	31,576	100%	31,505	100%	31,657	100%

Table C–12: Student Counts by Tested Grade and Year – Mathematics DC CAS

	2007		2008		2009		2010		2011	
Grade	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**
3	4,849	14.6%	4,677	14.6%	5,050	15.9%	4,917	15.6%	4,812	15.1%
4	4,726	14.2%	4,607	14.4%	4,569	14.4%	4,807	15.2%	4,857	15.2%
5	4,875	14.7%	4,615	14.4%	4,698	14.8%	4,468	14.1%	4,795	15.0%
6	4,775	14.4%	4,680	14.6%	4,463	14.1%	4,494	14.2%	4,388	13.8%
7	5,142	15.5%	4,606	14.4%	4,581	14.5%	4,339	13.7%	4,448	14.0%
8	4,949	14.9%	4,972	15.5%	4,477	14.1%	4,485	14.2%	4,321	13.6%
10	3,922	11.8%	3,829	12.0%	3,826	12.1%	4,086	12.9%	4,241	13.3%
All	33,238	100%	31,986	100%	31,664	100%	31,596	100%	31,862	100%

SCHOOL IDENTIFICATION CODES

The number of unique school codes varied slightly across years, as is shown in Table C–13 below.

Table C–13: Unique School Codes Across Years

Year	Number of Unique School Codes
2010-11	198
2009-10	211
2008-09	204

Over 90 percent of schools remained open for the past two years, as Table C–14 illustrates:

Table C–14: School Codes Across Years

Year	Number of Unique School Codes	Number of Unique School Codes in all Three Years	Percent of Unique School Codes in all Three Years	Number of Unique School Codes in Both 2010-11 and 2009-10	Percent of Unique School Codes in Both 2010-11 and 2009-10
2010-11	204	184	90.2%	189	92.6%

Large enrollment changes are uncommon. Of the 189 schools in both the 2010-11 and 2009-10 data, only 9 (4.8 percent) experienced an increase or a decrease in enrollment greater than 50 percent.

DATA MERGING

In order to estimate student growth models, it is imperative to be able to link students through a unique identifier to their prior achievement scores. Students without prior achievement data are typically excluded from growth models.

Tables C–15 and C–16 below display the merge rates for students by grade for the reading assessment. Table C–15 shows the merge rate for 2010-11 data (with one and two years of prior achievement scores). Table C–16 shows the merge rate for 2009-10 data. Tables C–17 and C–18 show similar data for the mathematics assessment. As expected, rates for grades 4 to 8 are high, although below 100%, reflecting student mobility. As there is no grade 2 test, zero percent of students in grade 4 have a second prior score. Rates are lower in 10th grade for which the pre-score is not for the immediately-preceding grade and year. These results suggest that about 20 percent of 10th grade students will not be included in the schoolwide growth models for grade 10, which could be due to a variety of situations that may occur at the transition point between middle and high school (e.g., student moves, transitions from other school systems).

Table C–15: Test Score Data Merge Rates by Grade: Reading DC CAS – 2010-11

Grade	Number of Valid 2010-11 Scores	Students with at least 1 prior score		Students with at least 2 prior scores	
		Count	Percent	Count	Percent
4	4,826	4,413	91.4%	0	0.0%
5	4,725	4,326	91.6%	4,036	85.4%
6	4,359	3,960	90.8%	3,683	84.5%
7	4,418	4,081	92.4%	3,783	85.6%
8	4,277	3,886	90.9%	3,602	84.2%
10	4,266	3,398	79.7%	2,932	68.7%

Table C–16: Test Score Data Merge Rates by Grade: Reading DC CAS – 2009-10

Grade	Number of Valid 2009-10 Scores	Students with at least 1 prior score		Students with at least 2 prior scores	
		Count	Percent	Count	Percent
4	4,781	4,415	92.3%	0	0.0%
5	4,452	4,079	91.6%	3,728	83.7%
6	4,474	4,107	91.8%	3,718	83.1%
7	4,327	3,956	91.4%	3,640	84.1%
8	4,477	4,082	91.2%	3,691	82.4%
10	4,100	3,288	80.2%	2,921	71.2%

Table C–17: Test Score Data Merge Rates by Grade: Mathematics DC CAS – 2010-11

Grade	Number of Valid 2010-11 Scores	Students with at least 1 prior score		Students with at least 2 prior scores	
		Count	Percent	Count	Percent
4	4,857	4,429	91.2%	0	0.0%
5	4,795	4,395	91.7%	4,100	85.5%
6	4,388	3,973	90.5%	3,701	84.3%
7	4,448	4,098	92.1%	3,796	85.3%
8	4,321	3,898	90.2%	3,599	83.3%
10	4,241	3,383	79.8%	2,920	68.9%

Table C–18: Test Score Data Merge Rates by Grade: Mathematics DC CAS – 2009-10

Grade	Number of Valid 2009-10 Scores	Students with at least 1 prior score		Students with at least 2 prior scores	
		Count	Percent	Count	Percent
4	4,807	4,437	92.3%	0	0.0%
5	4,468	4,099	91.7%	3,725	83.4%
6	4,494	4,126	91.8%	3,715	82.7%
7	4,339	3,952	91.1%	3,624	83.5%
8	4,485	4,067	90.7%	3,672	81.9%
10	4,086	3,284	80.4%	2,910	71.2%