Norm- and Criterion-Referenced Growth
An Overview of the Student Growth Percentiles Methodology

Damian W. Betebenner
Adam R. VanIwaarden
National Center for the Improvement of Educational Assessment (NCIEA)

May 2019

Abstract

This report provides details about the Student Growth Percentile (SGP) methodology. Topics addressed include an introduction to the concept of student growth and how SGPs approach questions that parents, teachers and other stakeholders have about how students are growing academically. The technical aspects of SGP calculation are covered in detail, and then expanded to show how the concept of adequate growth can be addressed and growth targets established using the SGP methodological framework.

In addition to the discussion of how all model parameters of interest are estimated and the beneficial properties of the model, the issue of bias in statistical models and student growth estimates resulting from the use of error-prone assessment data is addressed in detail. This report provides methodological details about the SIMEX method for correcting for measurement error problems as well as the Ranked SIMEX approach to improve the initial SIMEX correction.

1 Introduction - Why Student Growth?

Accountability systems constructed according to federal adequate yearly progress (AYP) requirements currently rely upon annual "snap-shots" of student achievement to make judgments about school quality. Since their adoption, such *status measures* have been the focus of persistent criticism (Linn, 2003; Linn, Baker, & Betebenner, 2002). Though appropriate for making judgments about the achievement level of students at a school for a given year, they are inappropriate for judgments about educational *effectiveness*. In this regard, status measures are blind to the possibility of low achieving students attending effective schools. It is this possibility that has led some critics of No Child Left Behind (NCLB) to label its accountability provisions as unfair and misguided and to demand the use of growth analyses as a better means of auditing school quality.

A fundamental premise associated with using student growth for school accountability is that "good" schools bring about student growth in excess of that found at "bad" schools. Students attending such schools - commonly referred to as highly effective/ineffective schools - tend to demonstrate extraordinary growth that is causally attributed to the school or teachers instructing the students. The inherent believability of this premise is at the heart of current enthusiasm to incorporate growth into accountability systems. It is not surprising that the November 2005 announcement by Secretary of Education Spellings for the Growth Model Pilot Program (GMPP) permitting states to use growth model results as a means for compliance with NCLB achievement mandates and the Race to the top competitive grants program were met with great enthusiasm by states (Spellings, 2005).

Following these use cases, the primary thrust of growth analyses over the last decade has been to determine, using sophisticated statistical techniques, the amount of student progress/-growth that can be justifiably attributed to the school or teacher - that is, to disentangle current *aggregate* level achievement from effectiveness (Ballou, Sanders, & Wright, 2004; Braun, 2005; Raudenbush, 2004; Rubin, Stuart, & Zanutto, 2004). Such analyses, often called *value-added* analyses, attempt to estimate the teacher or school contribution to student achievement. This contribution, called the *school* or *teacher effect*, purports to quantify the impact on achievement that this school or teacher would have, on average, upon similar students assigned to them for instruction. Clearly, such analyses lend themselves to accountability systems that hold schools or teachers responsible for student achievement.

Despite their utility in high stakes accountability decisions, the causal claims of teacher/school effectiveness addressed by value-added models (VAM) often fail to address questions of primary interest to education stakeholders. For example, VAM analyses generally ignore a fundamental interest of stakeholders regarding student growth: How much growth did a student make? The disconnect reflects a mismatch between questions of interest and the statistical model employed to answer those questions. Along these lines, Harris (2007) distinguishes value-added for program evaluation (VAM-P) and value-added for accountability (VAM-A) - conceptualizing accountability as a difficult type of program evaluation. Indeed, the current climate of high-stakes, test-based accountability has blurred the lines between program evaluation and accountability. This, combined with the emphasis of value-added models toward causal claims regarding school and teacher effects has skewed discussions about growth models toward causal claims at the expense of description. Research (Yen, 2007) and personal experience suggest stakeholders are more interested in the reverse: description first that can be used secondarily

as part of causal fact finding.

In a survey conducted by Yen (2007), supported by the author's own experience working with state departments of education to implement growth models, parents, teacher, and administrators were asked what "growth" questions were most of interest to them.

- **Parent Questions:**
  - Did my child make a year's worth of progress in a year?
  - Is my child growing appropriately toward meeting state standards?
  - Is my child growing as much in Math as Reading?
  - Did my child grow as much this year as last year?
- **Teacher Questions:**
  - Did my students make a year's worth of progress in a year?
  - Did my students grow appropriately toward meeting state standards?
  - How close are my students to becoming Proficient?
  - Are there students with unusually low growth who need special attention?
- **Administrator Questions:**
  - Did the students in our district/school make a year's worth of progress in all content areas?
  - Are our students growing appropriately toward meeting state standards?
  - Does this school/program show as much growth as that one?
  - Can I measure student growth even for students who do not change proficiency categories?
  - Can I pool together results from different grades to draw summary conclusions?

As Yen remarks, all these questions rest upon a desire to understand whether observed student progress is "reasonable or appropriate" (Yen, 2007). More broadly, the questions seek a description rather than a parsing of responsibility for student growth. Ultimately, questions may turn to who/what is responsible. However, as indicated by this list of questions, they are not the starting point for most stakeholders.

In the following paragraphs, student growth percentiles and percentile growth projections/-trajectories are introduced as a means of understanding student growth in both norm-referenced and criterion referenced ways. With these values calculated we show how growth data can be utilized in both a norm- and in a criterion-referenced manner to inform discussion about education quality. We assert that the establishment of a norm-referenced basis for student growth eliminates a number of the problems of incorporating growth into accountability systems providing needed insight to various stakeholders by addressing the basic question of how much a student has progressed (Betebenner, 2008; D. W. Betebenner, 2009).

2 Student Growth Percentiles

It is a common misconception that to quantify student progress in education, the subject matter and grades over which growth is examined must be on the same scale - referred to as a vertical scale. Not only is a vertical scale not necessary, but its existence obscures concepts necessary to fully understand student growth. Growth, fundamentally, requires change to be examined for a single construct like math achievement across time - *growth in what?*

Consider the familiar situation from pediatrics where the interest is on measuring the height and weight of children over time. The scales on which height and weight are measured possess properties that educational assessment scales aspire towards but can never meet.[1]

> An infant male toddler is measured at 2 and 3 years of age and is shown to have grown 4 inches. The magnitude of increase - 4 inches - is a well understood quantity that any parent can grasp and measure at home using a simple yardstick. However, parents leaving their pediatrician's office knowing only how much their child has grown would likely be wanting for more information. In this situation, parents are not interested in an absolute criterion of growth, but instead in a norm-referenced criterion locating that 4 inch increase alongside the height increases of similar children. Examining this height increase relative to the increases of similar children permits one to diagnose how (a)typical such an increase is.

Given this reality in the examination of change where scales of measurement are perfect, we argue that it is unreasonable to think that in education, where scales are at best quasi-interval (Lord, 1975; Yen, 1986) one can/should examine growth differently.

Going further, suppose that scales did exist in education similar to height/weight scales that permitted the calculation of absolute measures of annual academic growth for students. The response to a parent's question such as, "How much did my child progress?", would be a number of scale score points - an answer that would leave most parents confused wondering whether the number of points is good or bad. As in pediatrics, the search for a description regarding changes in achievement over time (i.e., growth) is best served by considering a norm-referenced quantification of student growth - *a student growth percentile* (Betebenner, 2008; D. W. Betebenner, 2009).

A student's growth percentile (SGP) describes how (a)typical a student's growth is by examining his/her current achievement relative to his/her *academic peers* - those students beginning at the same place. That is, a student growth percentile examines the current achievement of a student relative to other students who have, in the past, "walked the same achievement path" (see this presentation for a detailed description of the academic peer concept). Heuristically, if the state assessment data set were extremely large (in fact, infinite) in size, one could open the infinite data set and select out those students with the exact same prior scores and compare how the selected student's current year score compares to the current year scores of those students with the same prior year's scores - his/her academic peers. If the student's current

[1] The scales on which students are measured are often assumed to possess properties similar to height and weight but they don't. Specifically, scales are assumed to be interval where it is assumed that a difference of 100 points at the lower end of the scale refers to the same difference in ability/achievement as 100 points at the upper end of the scale. (See Lord, 1975; and Yen, 1986 for more detail on the interval scaling in educational measurement.)

year score exceeded the scores of most of his/her academic peers, in a norm-referenced sense they have done as well. If the student's current year score was less than the scores of his/her academic peers, in a norm-referenced sense they have not done as well.

The four panels of Figure B.1. depict what a student growth percentile represents in a situation considering students having only two consecutive achievement test scores.

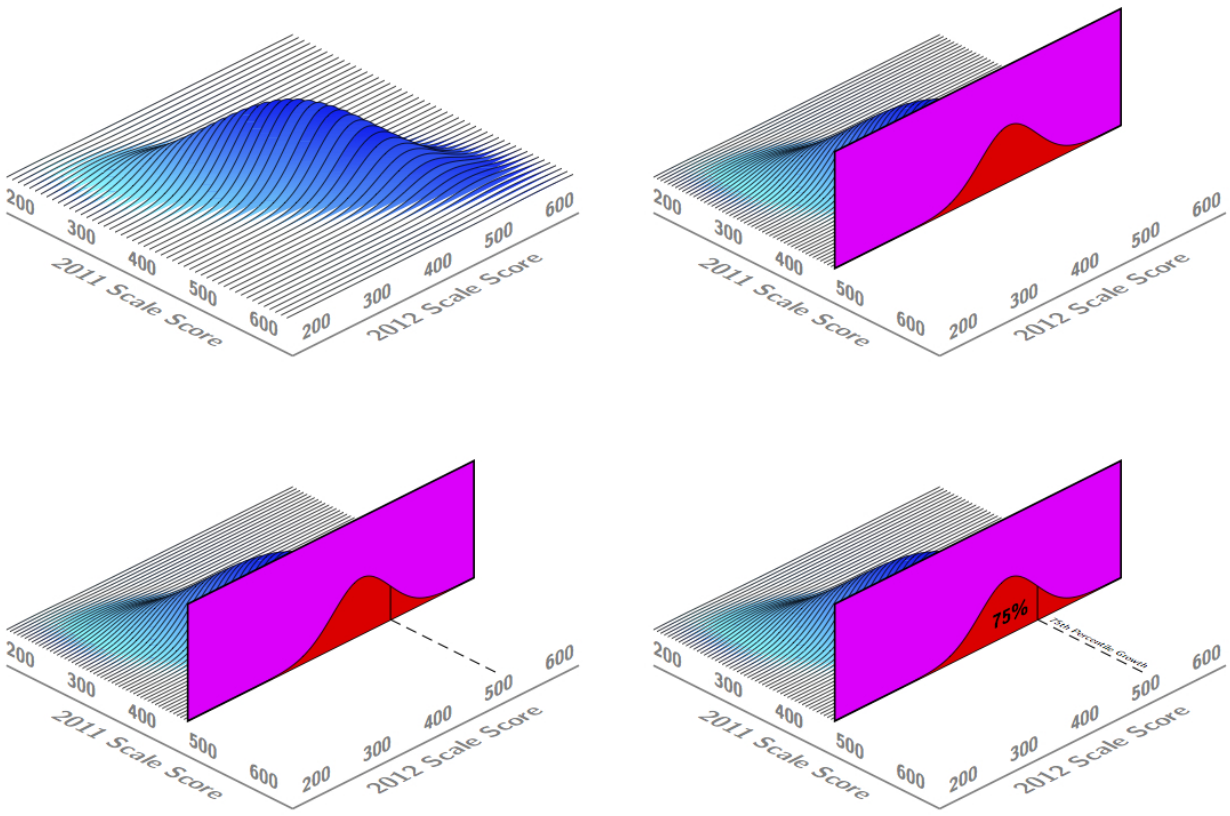
- **Upper Left Panel** Considering all pairs of 2011 and 2012 scores for all students in the state yields a bivariate (two variable) distribution. The higher the distribution, the more frequent the pair of scores.

- **Upper Right Panel** Taking account of prior achievement (i.e., conditioning upon prior achievement) fixes the value of the 2011 scale score (in this case at approximately 460) and is represented by the red slice taken out of the bivariate distribution.
- **Lower Left Panel** Conditioning upon prior achievement defines a *conditional distribution* which represents the distribution of outcomes on the 2012 test assuming a 2011 score of 460. This distribution is indicated by the solid red slice of the distribution.

- **Lower Right Panel** The conditional distribution provides the context against which a student's 2012 achievement can be examined and provides the basis for a norm-referenced comparison. Students with achievement in the upper tail of the conditional distribution have demonstrated high rates of growth relative to their academic peers whereas those students with achievement in the lower tail of the distribution have demonstrated low rates of growth. Students with current achievement in the middle of the distribution could be described as demonstrating "average" or "typical" growth. In the figure provided the student scores approximately 500 on the 2012 test. Within the conditional distribution, the value of 500 lies at the 75^{th} percentile. Thus the student's progress from 460 in 2011 to 500 in 2012 met or exceeded that of 75 percent of students starting from the same place. It is important to note that qualifying a student growth percentile as "adequate", "good", or "enough" is a standard setting procedure that requires stakeholders to examine a student's growth *vis-a-vis* external criteria such as performance standards/levels.

Figure B.1 also serves to illustrate the relationship between the state's assessment scale and student growth percentiles. The scale depicted in the panels of Figure B.1 is not vertical. Thus the comparisons or subtraction of scale scores for individual students is not supported. However, were such a scale in place, the figure would not change. With or without a vertical scale, the conditional distribution can be constructed.

Figure. B.1: Depiction of the distribution associated with 2011 and 2012 student scale scores together with the conditional distribution and associated growth percentile.



In situations where a vertical scale exists, the increase/decrease in scale score points can be calculated and the growth percentile can be understood alongside this change. For example, were the scales presented in Figure B.1 vertical, then one can calculate that the student grew 40 points (from 460 to 500) between 2011 and 2012. This 40 points represents the absolute magnitude of change. Quantifying the magnitude of change is scale dependent. For example, different vertical achievement scales in 2011 and 2012 would yield different annual scale score increases: A scale score increase of 40 could be changed to a scale score increase of 10 using a simple transformation of the vertical scale on which all the students are measured. However, relative to other students, their growth has not changed - their growth percentile is invariant to scale transformations common in educational assessment. Student growth percentiles norm-referencedly situate achievement change bypassing questions associated with the magnitude of change, and directing attention toward relative standing which, we would assert, is what stakeholders are most interested in.

To fully understand how many states intend to use growth percentiles to make determinations about whether a student's growth is sufficient, the next section details specifics of how student growth percentiles are calculated. These calculations are subsequently used to calculate percentile growth projections/trajectories that are used to establish how much growth it will take for each student to reach his/her achievement targets.

3 SGP Calculation

Quantile regression is used to establish curvilinear functional relationships between the cohort's prior scores and their current scores. Specifically, for each grade by subject cohort, quantile regression is used to establish 100 (1 for each percentile) curvilinear functional relationships between the students prior score(s) and their current score. For example, consider 7^{th} graders. Their grade 3, grade 4, grade 5, and grade 6 prior scores are used to describe the current year grade 7 score distribution.[2] The result of these 100 separate analyses is a single coefficient matrix that can be employed as a look-up table relating prior student achievement to current achievement for each percentile. Using the coefficient matrix, one can plug in *any* grade 3, 4, 5, and 6 prior score combination to the functional relationship to get the percentile cutpoints for grade 7 conditional achievement distribution associated with that prior score combination. These cutpoints are the percentiles of the conditional distribution associated with the individual's prior achievement. Consider a student with the following mathematics scores:

Table 1: Scale scores for a hypothetical student across 5 years in mathematics.

Grade 3	Grade 4	Grade 5	Grade 6	Grade 7
419	418	422	434	436

Using the coefficient matrix derived from the quantile regression analyses based upon grade 3, 4, 5, and 6 scale scores as independent variables and the grade 7 scale score as the dependent variable together with this student's vector of grade 3, 4, 5, and 6 grade scale scores provides the scale score percentile cutpoints associated with the grade 7 conditional distribution for these prior scores.

Table 2: Percentile cutscores for grade 7 mathematics based upon the grade 3, 4, 5, and 6 mathematics scale scores given in Table 1.

1^{st}	2^{nd}	3^{rd}	...	10^{th}	...	25^{th}	...	50^{th}	51^{th}	...	75^{th}	...	90^{th}	...	99^{th}
404.8	414.9	419.9	...	425.9	...	430.8	...	435.5	436.3	...	468.9	...	487.1	...	509.8

The percentile cutscores for 7^{th} grade mathematics in Table 2 are used with the student's *actual* grade 7 mathematics scale score to establish his/her growth percentile. In this case, the student's grade 7 scale score of 436 lies above the 50^{th} percentile cut and below the 51^{st} percentile cut, yielding a growth percentile of 50. Thus, the progress demonstrated by this student between grade 6 and grade 7 exceeded that of 50 percent of his/her academic peers - those students with the same achievement history. States can qualify student growth by defining ranges of growth percentiles. For example, some states designate growth percentiles

[2]For the mathematical details underlying the use of quantile regression in calculating student growth percentiles, see the *SGP Estimation* section

between 35 and 65 as being *typical*. Using Table 2, another student with the exact same grade 3, 4, 5, and 6 prior scores but with a grade 7 scale score of 404, would have a growth percentile of 1, which is designated as *low*.

This example provides the basis for beginning to understand how growth percentiles in the SGP Methodology are used to determine whether a student's growth is *(in)adequate*. Suppose that in grade 6 a one-year (i.e., 7^{th} grade) achievement goal/target of proficiency was established for the student. Using the lowest proficient scale score for 7^{th} grade mathematics, this target corresponds to a scale score of 500 Based upon the results of the growth percentile analysis, this one year target corresponds to 95^{th} percentile growth. Their growth, obviously, is less than this and the student has not met this individualized growth standard.

4 Percentile Growth Projections/Trajectories

Building upon the example just presented involving only a one-year achievement target translated into a growth standard, this section extends this basic idea and shows how multi-year growth standards are established based upon official state achievement targets/goals. That is, by defining a future (e.g., a 2 year) achievement target for each student, we show how growth percentile analyses can be used to quantify what level of growth, expressed as a per/year growth percentile, is required by the student to reach his/her achievement target. Unique to the SGP Methodology is the ability to stipulate *both* what the growth standard is as well as how much the student actually grew in a metric that is informative to stakeholders.

4.1 Defining Adequate Growth

Establishing thresholds for growth for each student that can be used to make adequacy judgments requires pre-established achievement targets and a time-frame to reach the target for each student against which growth can be assessed (i.e., growth-to-standard). Three years from the establishment of the target is a typical time frame many states have chosen for purposes of describing students growth to standard. Targets are initially established in the prior academic year, so that in the current year a student is considered to be *catching-up* to or *keeping-up* with proficiency. Other targets may also be considered (for example, *moving-up* to or *staying-up* with an advanced achievement level).

Using a three year target as an example, these adequacy categories are defined as:

- ***Catch-Up*** Those students currently not proficient (from the prior spring testing) are expected to be proficient within 3 years following the establishment of the achievement target or by the final grade, whichever comes sooner.[3]
- ***Keep-Up*** Those students currently at or above proficient are expected to remain at or above proficient in all of the 3 years following the establishment of the achievement target or by the final grade, whichever comes sooner.
- ***Move-Up*** Those students currently proficient are expected to reach advanced within 3 years following the establishment of the achievement target or by the final grade, whichever comes sooner.
- ***Stay-Up*** Those students currently advanced and are expected to remain advanced in all of the 3 years following the establishment of the achievement target or by the final grade, whichever comes sooner.

The previous definitions specify "3 years following the establishment of the achievement target" as the time frame. For example, an non-proficient 3^{rd} grader would be expected to be proficient by 6^{th} grade. The first check of the student's progress occurs in 4^{th} grade, when the student's growth over the last year is compared against targets calculated to assess their progress along a multi-year time-line. The question asked following the 4^{th} grade for the student is: Did the student become proficient and if not are they on track to become proficient within 3 years?

[3]The establishment of the achievement target occurs in the year prior, therefore the time frame of 3 years includes the current year as "year 1", which is the year in which the first growth adequacy judgment can be made for the student. The targets are then projected out two years beyond the current year to give a maximum time horizon of 3 years in which to make the adequacy judgement.

4.2 Calculation of Growth Percentile Targets

As mentioned previously, the calculation of student growth percentiles across all grades and students results in the creation of numerous coefficient matrices that relate prior with current student achievement. These matrices constitute an annually updated statewide historical record of student progress. For the SGP Methodology, they are used to determine what level of percentile growth is necessary for each student to reach future achievement targets. For example, imagine that the following coefficient matrices are produced for Mathematics in a state after the annual calculation of student growth percentiles using up to three prior years of test data:

- **Grade 4** Using grade 3 prior achievement.
- **Grade 5** Using grade 4 and grades 3 & 4 prior achievement.
- **Grade 6** Using grade 5, grades 4 & 5, and grades 3, 4, & 5 prior achievement.
- **Grade 7** Using grade 6, grades 5 & 6, grades 4, 5, & 6, and grades 3, 4, 5, & 6 prior achievement.
- **Grade 8** Using grade 7, grades 6 & 7, grades 5, 6, & 7, and grades 4, 5, 6, & 7 prior achievement.

To describe how these numerous coefficient matrices are used together to produce growth targets, consider, for example, a 4^{th} grade student in reading with 3^{rd} and 4^{th} grade state reading scores of 425 (Unsatisfactory) and 440 (Partially Proficient), respectively. The following are the steps that transpire over 3 years to determine whether this student is on track to reach proficient.

- **Spring Year 0 -** The growth target for Year 1 is established requiring students to reach state defined achievement levels within 3 years or by grade 8. In this example, the student under consideration was Partially Proficient in 3^{rd} grade (in Year 0) and is expected to be proficient by grade 6 in Year 3.
- **Spring Year 1 -** Because our example student was not proficient based on their prior year test score her initial status for the current year is a *catching-up* student. We want to see if the growth she demonstrated in Year 1 was adequate enough to make her proficient, or at least put her on a trajectory towards proficiency within the next two years. Employing the coefficient matrices derived in the calculation of Year 1 student growth percentiles:
  1. The coefficient matrix relating grade 4 with grade 3 prior achievement is used to establish the percentile cuts (i.e., one-year growth percentile projections/trajectories). If the student's actual Year 1 growth percentile exceeds the percentile cut associated with proficient, then the student's one year growth is enough to reach proficient.[4]
  2. The 2 year growth percentile projections/trajectories are calculated, extending from Year 0 to Year 2. The student's actual grade 3 scale score together with the 99 hypothetical one-year growth percentile projections/trajectories derived in the previous step are plugged into the Year 1 coefficient matrix relating grade 5 with grade 3 & 4 prior achievement. This yields the percentile cuts for the student indicating what

[4]Checking growth adequacy using one-year achievement targets is equivalent to confirming whether the student reached his/her one-year achievement target since the coefficient matrices used to produce the percentile cuts are based on current data.

consecutive two-year 1^{st} through 99^{th} percentile growth will lead to.[5] The student's Year 1 growth percentile is compared to the 2 year growth percentile cut required to reach proficiency. If the student's growth percentile exceeds this target, then the student is deemed on track to reach proficiency by the 5^{th} grade.

3. Last, the 3 year growth percentile projections/trajectories are established. The student's actual grade 3 scale score together with the 99 hypothetical 1 and 2 year growth percentile projections/trajectories derived in the previous two steps are plugged into the coefficient matrix relating grade 6 with prior achievement in grades 3, 4, & 5. This yields the percentile cuts for each student indicating what three consecutive years of 1^{st} through 99^{th} percentile growth will lead to in terms of future achievement. The student's observed Year 1 growth percentile is again compared to the percentile cut required to reach proficiency, and if it meets or exceeds it her growth is deemed adequate enough to reach proficiency by the 6^{th} grade.

- **Spring Year 1/Fall Year 2 -** The growth target for Year 2 is now established. The student in this example has now presumably completed grade 4 and beginning grade 5 in the Fall. She was again Partially Proficient in 4^{th} grade and is now expected to be on track to proficient by grade 7 in Year 4.
- **Spring Year 2 -** Employing the coefficient matrices derived in the calculation of Year 2 student growth percentiles:

1. The coefficient matrix relating grade 5 with grade 3 & 4 prior achievement is used to establish 99 percentile cuts (i.e., one-year growth percentile projections/trajectories). If the student's actual Year 2 growth percentile exceeds the cut associated with proficient, then the student's one year growth was enough to reach proficient.
2. The student's actual scores from grades 3 & 4 together with the 99 hypothetical one-year growth percentile projections/trajectories derived in the previous step are plugged into the coefficient matrix relating grade 6 with grade 3, 4, & 5 prior achievement. This yields 99 percentile cuts (i.e., 2 year growth percentile projections/trajectories) for the student indicating what consecutive two-year 1^{st} through 99^{th} percentile growth will lead to in terms of future achievement. The student's Year 2 growth percentile is compared to the 2 year growth percentile cut required to reach proficiency. If the student's growth percentile meets or exceeds it then the student is deemed on track to reach proficient.
3. The 3 year growth percentile projections/trajectories are established. The student's actual grades 3 & 4 scale scores together with the 99 hypothetical 1 and 2 year growth percentile projections/trajectories derived in the previous two steps are plugged into the coefficient matrix relating grade 7 with prior achievement in grades 3, 4, 5 & 6. This yields the percentile cuts for each student indicating what three consecutive years of 1^{st} through 99^{th} percentile growth will lead to in terms of future achievement. The student's observed Year 2 growth percentile is again compared to the percentile cut required to reach proficiency, and if it exceeds it her growth is deemed adequate enough to reach proficiency by the 7^{th} grade.

[5] Two or more year growth targets are estimated based upon the most recent student growth histories in the state. In this example, estimates for growth that will be needed in the 5th and 6th grades are based on students in 5th and 6th grades (concurrently) in Year 1.

This process repeats in a similar fashion as the student progresses from one grade to the next, year after year. The complexity of the process just described is minimized by the use of the R Software Environment (R Development Core Team, 2019) in conjunction with an open source software package SGP (Betebenner, VanIwaarden, Domingue, & Shang, 2019) developed by the National Center for the Improvement of Educational Assessment in consultation with the state department of education to calculate student growth percentiles and percentile growth projections/trajectories. Every year, following the completion of the test score reconciliation, student growth percentiles and percentile growth trajectories are calculated for each student. Once calculated, these values are easily used to make the yes/no determinations about the adequacy of each student's growth relative to his/her fixed achievement targets.

4.3 System-wide Growth and Achievement Charts

Operational work calculating student growth percentiles with state assessment data yields a large number of coefficient matrices derived from estimating Equation 4 (see the ***SGP Estimation*** section below). These matrices, similar to a lookup table, "encode" the relationship between prior and current achievement scores for students in the norm group (usually an entire grade cohort of students for the state) across all percentiles and can be used both to qualify a student's current level growth as well as predict, based upon current levels of student progress, what different rates of growth (quantified in the percentile metric) will yield for students statewide.

When rates of growth necessary to reach performance standards are investigated, such calculations are often referred to as "growth-to-standard". These analyses serve a dual purpose in that they provide the growth rates necessary to reach these standards and also shed light on the standard setting procedure as it plays out across grades. To establish growth percentiles necessary to reach different performance/achievement levels, it is necessary to investigate what growth percentile is necessary to reach the desired performance level thresholds based upon the student's achievement history.

Establishing criterion referenced growth thresholds requires consideration of multiple future growth/achievement scenarios. Instead of inferring that prior student growth is indicative of future student growth (e.g., linearly projecting student achievement into the future based upon past rates of change), predictions of future student achievement are contingent upon initial student status (where the student starts) and subsequent rates of growth (the rate at which the student grows). This avoids fatalistic statements such as, "Student X is projected to be (not) proficient in two years" and instead promotes discussions about the different rates of growth necessary to reach future achievement targets: "In order that Student X reach/maintain proficiency within two years, she will have to demonstrate n^{th} percentile growth consecutively for the next two years." The change in phraseology is minor but significant. Stakeholder conversations turn from "where will (s)he be" to "what will it take?"

Figure. B.2: Growth chart depicting future mathematics achievement conditional upon consecutive 10^{th}, 35^{th}, 50^{th}, 65^{th} and 90^{th} percentile growth for a student beginning the third grade at the cutpoint between achievement levels 1 and 2.

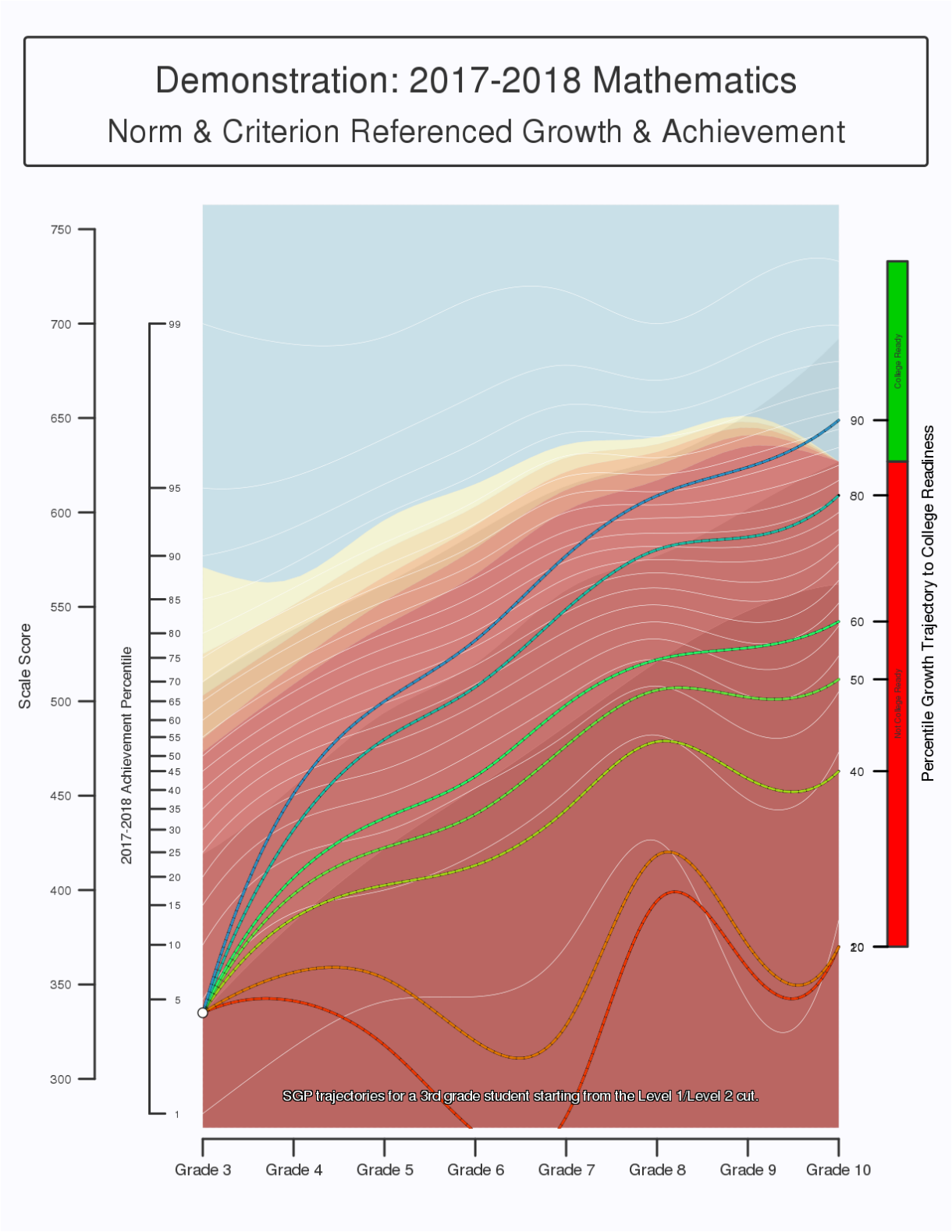
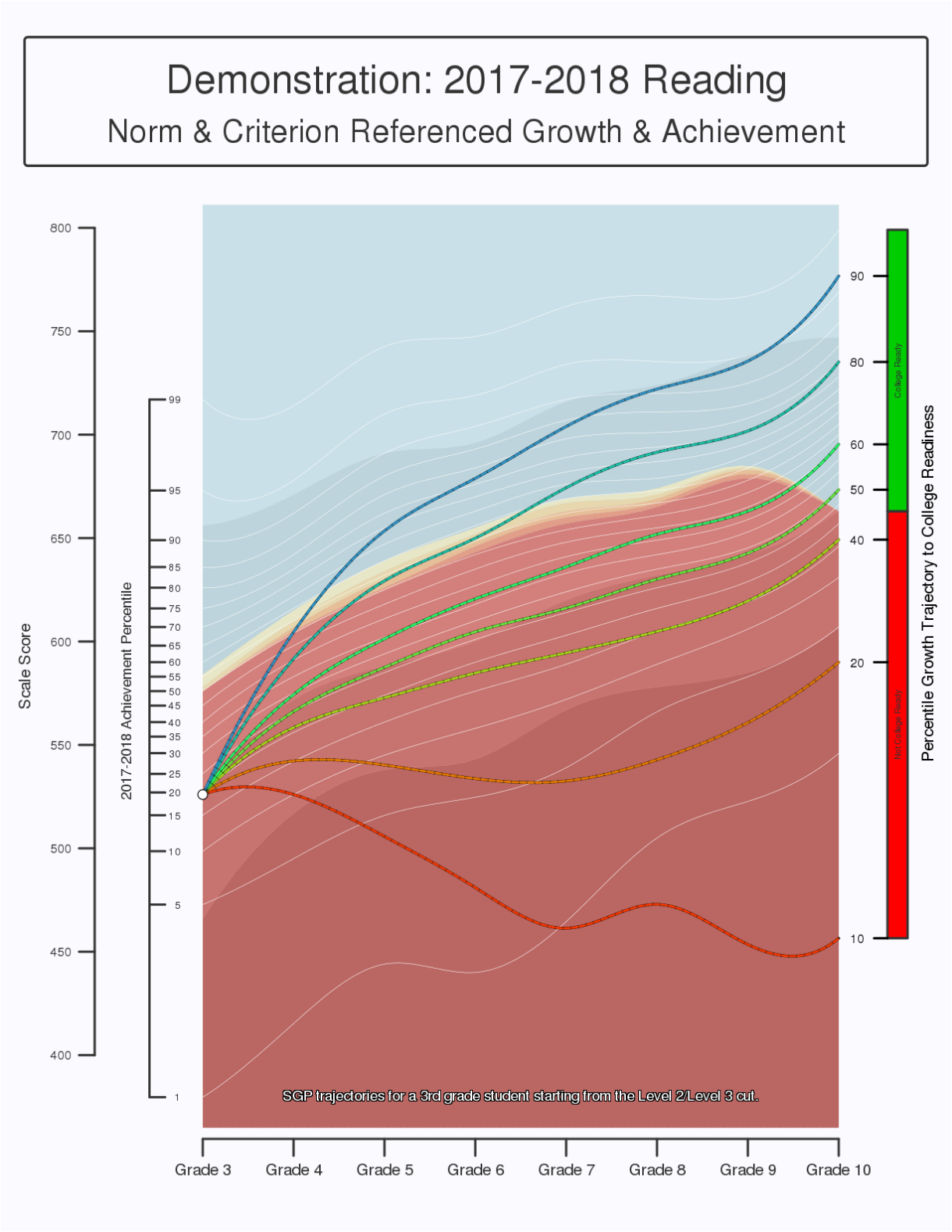


Figure. B.3: Growth chart depicting future reading achievement conditional upon consecutive 10^{th}, 35^{th}, 50^{th}, 65^{th} and 90^{th} percentile growth for a student beginning the third grade at the cutpoint between achievement levels 2 and 3.



Parallel growth/achievement scenarios are more easily understood with a picture. Using the results of a statewide assessment growth percentile analyses, Figures B.2 and B.3 depict future growth scenarios in mathematics for a student starting in third grade and tracking that student's achievement time-line based upon different rates of annual growth expressed in the growth percentile metric. The figures depict the four state achievement levels across grades 3 to 10 in shades of red to light blue (Unsatisfactory, Partially Proficient, Proficient and Advanced) together with the 2018 achievement percentiles (inner most vertical axis) superimposed in white. Beginning with the student's achievement starting point at grade 3, a grade 4 achievement projection is made based upon the most recent growth percentile analyses derived using prior 3^{rd} to 4^{th} grade student progress. More specifically, using the coefficient matrices derived in the quantile regression of grade 4 on grade 3 (see Equation 4), predictions of what 10^{th}, 35^{th}, 50^{th}, 65^{th}, and 90^{th} percentile growth lead to are calculated. Next, using these seven projected 4^{th} grade scores combined with the student actual 3^{rd} grade score, 5^{th} grade achievement projections are calculated using the most recent quantile regression of grade 5 on grades 3 and 4. Similarly, using these seven projected 5^{th} grade scores, the 5 projected 4^{th} grade scores with the students actual third grade score, achievement projections to the 6^{th} grade are calculated using the most recent quantile regression of grade 6 on grades 3, 4, and 5. The analysis extends recursively for grades 6 to 10 yielding the *percentile growth trajectories* in Figures B.2 and B.3. The figures allow stakeholders to consider what consecutive rates of growth, expressed in growth percentiles, yield for students starting at different points.

Figure B.2 depicts percentile growth trajectories in mathematics for a student beginning at the threshold between achievement level 1 and achievement level 2. Based upon the *achievement* percentiles depicted (the white contour lines), approximately 25 percent of the population of 3^{rd} graders rate as "Partially Proficient" or below. Moving toward grade 8, the percentage of Partially Proficient students increases to near 45 percent. The dashed, colored lines in the figure represent seven different growth scenarios for the student based upon consecutive growth at a given growth percentile, denoted by the right axis. At the lower end, for example, consecutive 10^{th} percentile growth leaves the student, unsurprisingly, mired in the Unsatisfactory category. Consecutive 10^{th}, through 60^{th} percentile growth also leave the student in the Partially Proficient category. Even consecutive 65^{th} percentile growth may not be enough to lift these students above Partially Proficient into the Proficient category. This demonstrates how difficult probabilistically, based upon current rates of progress, it is for students to move up in performance level in math statewide. Considering a goal of reaching proficient (next to top region) by 8^{th} grade, a student would need to demonstrate growth percentiles consecutively in excess of 65 to reach this achievement target indicating how unlikely such an event currently is. In light of policy mandates for universal proficiency, the growth necessary for non-proficient students to reach proficiency, absent radical changes to growth rates of students statewide, is likely unattainable for a large percentage of non-proficient students.

Figure B.3 depicts percentile growth trajectories in reading for a student beginning at the level 2/level 3 threshold in grade 3. In a normative sense, the performance standards in reading are more demanding than those in mathematics (particularly in the higher grades) with approximately 20-30 percent of students are Partially Proficient in grades 3 to 10. The dashed, colored lines in the figure represent growth scenarios for the hypothetical student based upon consecutive growth at a the given growth percentile. Compared with the growth required in mathematics, more modest growth is required to maintain proficiency in reading.

Typical growth (50^{th} percentile growth) appears adequate for such a student to move up into the proficiency category by the end of 10^{th} grade.

5 SGP Estimation

Calculation of a student's growth percentile is based upon the estimation of the conditional density associated with a student's score at time t using the student's prior scores at times $1, 2, \ldots, t-1$ as the conditioning variables. Given the conditional density for the student's score at time t, the student's growth percentile is defined as the percentile of the score within the time t conditional density. By examining a student's current achievement with regard to the conditional density, the student's growth percentile situates the student's outcome at time t taking account of past student performance. The percentile result reflects the likelihood of such an outcome given the student's prior achievement. In the sense that the student growth percentile translates to the probability of such an outcome occurring (i.e., rarity), it is possible to compare the progress of individuals not beginning at the same starting point. However, occurrences being equally rare does not necessarily imply that they are equally "good." Qualifying student growth percentiles as "(in)adequate," "good," or as satisfying "a year's growth" is a standard setting procedure requiring external criteria (e.g., growth relative to state performance standards) combined with the wisdom and judgments of stakeholders.

Estimation of the conditional density is performed using quantile regression (Koenker, 2005). Whereas linear regression methods model the conditional mean of a response variable Y, quantile regression is more generally concerned with the estimation of the family of conditional quantiles of Y. Quantile regression provides a more complete picture of both the conditional distribution associated with the response variable(s). The techniques are ideally suited for estimation of the family of conditional quantile functions (i.e., reference percentile curves). Using quantile regression, the conditional density associated with each student's prior scores is derived and used to situate the student's most recent score. Position of the student's most recent score within this density can then be used to characterize the student's growth. Though many state assessments possess a vertical scale, such a scale is not necessary to produce student growth percentiles.

In analogous fashion to the least squares regression line representing the solution to a minimization problem involving squared deviations, quantile regression functions represent the solution to the optimization of a loss function (Koenker, 2005). Formally, given a class of suitably smooth functions, $\mathcal{G}$, one wishes to solve

$$arg\ min_{g \in \mathcal{G}} \sum_{i=1}^{n} \rho_\tau(Y(t_i) - g(t_i)), \tag{1}$$

where t_i indexes time, Y are the time dependent measurements, and ρ_τ denotes the piecewise linear loss function defined by

$$\rho_\tau(u) = u \cdot (\tau - I(u < 0)) = \begin{cases} u \cdot \tau & u \geq 0 \\ u \cdot (\tau - 1) & u < 0. \end{cases} \tag{2}$$

The elegance of the quantile regression Expression 1 can be seen by considering the more familiar least squares estimators. For example, calculation of $arg\ min \sum_{i=1}^{n}(Y_i - \mu)^2$ over $\mu \in \mathbb{R}$ yields the sample mean. Similarly, if $\mu(x) = x'\beta$ is the conditional mean represented as a linear combination of the components of x, calculation of $arg\ min \sum_{i=1}^{n}(Y_i - x_i'\beta)^2$ over $\beta \in \mathbb{R}^p$ gives the familiar least squares regression line. Analogously, when the class of candidate functions $\mathcal{G}$ consists solely of constant functions, the estimation of Expression 1 gives the τ^{th} sample quantile associated with Y. By conditioning on a covariate x, the τ^{th} conditional quantile function is given by

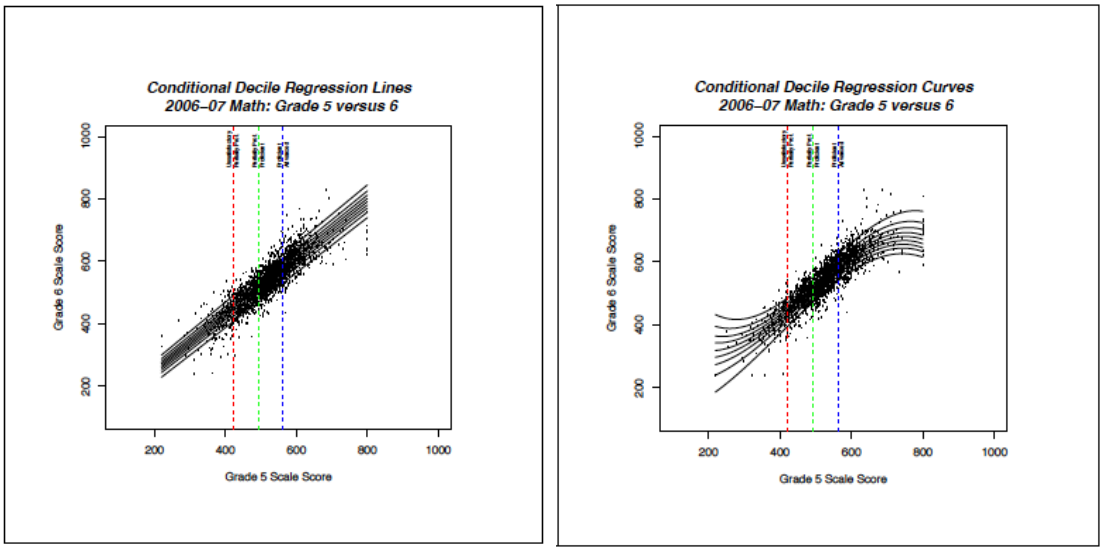
$$Q_y(\tau|x) = arg\ min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} \rho_\tau(y_i - x_i'\beta). \tag{3}$$

In particular, if $\tau = 0.5$, then the estimated conditional quantile line is the median regression line.[6]

Following Wei and He (2006), we parameterize the conditional quantile functions as a linear combination of B-spline cubic basis functions. B-splines are employed to accommodate non-linearity, heteroscedasticity and skewness of the conditional densities associated with values of the independent variable(s). B-splines are attractive both theoretically and computationally in that they provide excellent data fit, seldom lead to estimation problems (Harrell, 2001), and are simple to implement in available software.

Figure B.4 gives a bivariate representation of linear and B-splines parameterization of decile growth curves. The assumption of linearity imposes conditions upon the heteroscedasticity of the conditional densities. Close examination of the linear deciles indicates slightly greater variability for higher grade 5 scale scores than for lower scores. By contrast, the B-spline based decile functions better capture the greater variability at both ends of the scale score range together with a slight, non-linear trend to the data.

Figure. B.4: Linear and B-spline conditional deciles based upon bivariate math data, grades 5 and 6.



[6]For a detailed treatment of the procedures involved in solving the optimization problem associated with Expression 1, see (Koenker, 2005), particularly Chapter 6.

Calculation of student growth percentiles is performed using `R` (R Development Core Team, 2019), a language and environment for statistical computing, with `SGP` package (Betebenner et al., 2019). Other possible software (untested with regard to student growth percentiles) with quantile regression capability include SAS and Stata. Estimation of cohort referenced student growth percentiles is conducted using all available prior data, subject to certain suitability conditions. Estimation of baseline referenced student growth percentiles typically uses a restricted number of prior years' data (for example, some states have used a maximum of two prior years' data). Given assessment scores for t occasions, $(t \geq 2)$, the τ^{th} conditional quantile for Y_t based upon $Y_{t-1}, Y_{t-2}, \ldots, Y_1$ is given by

$$Q_{Y_t}(\tau|Y_{t-1}, \ldots, Y_1) = \sum_{j=1}^{t-1} \sum_{i=1}^{3} \phi_{ij}(Y_j)\beta_{ij}(\tau), \tag{4}$$

where $\phi_{i,j}$, $i = 1, 2, 3$ and $j = 1, \ldots, t-1$ denote the B-spline basis functions. Currently, bases consisting of 7 cubic polynomials are used to "smooth" irregularities found in the multivariate assessment data. A bivariate rendering of this is found is Figure B.4 where linear and B-spline conditional deciles are presented. The cubic polynomial B-spline basis functions model the heteroscedasticity and non-linearity of the data to a greater extent than is possible using a linear parameterization.

The B-spline basis functions require the selection of boundary and interior knots. Boundary knots are end points outside of the scale score distribution that anchor the B-spline basis. These are generally selected by extending the range of scale scores by 10%. That is, they are defined as lying 10% below the lowest obtainable (or observed) scale score (LOSS) and 10% above the highest obtainable scale score (HOSS). The interior knots are the *internal* breakpoints that define the spline.

The default choice in the `SGP` package (Betebenner et al., 2019) is to select the 20^{th}, 40^{th}, 60^{th} and 80^{th} quantiles of the observed scale score distribution. In general the knots and boundaries are computed using a distribution from several years of compiled test data (i.e. multiple cohorts combined into a single distribution) so that any irregularities in a single year are smoothed out. Subsequent annual analyses then use these same knots and boundaries as well.

Finally, it should be noted that the independent estimation of the regression functions can potentially result in the crossing of the quantile functions. This occurs near the extremes of the distributions and is potentially more likely to occur given the use of non-linear functions. The result of allowing the quantile functions to cross in this manner would be *lower* percentile estimations of growth for *higher* observed scale scores at the extremes (give all else equal in prior scores) and vice versa. In order to deal with these contradictory estimates, quantile regression results are isotonized to prevent quantile crossing following the methods derived by Chernozhukov, Fernandez-Val and Glichon (2010).

6 Discussion of Model Properties

Student growth percentiles possess a number of attractive properties from both a theoretical as well as a practical perspective. Foremost among practical considerations is that the percentile descriptions are familiar and easily communicated to teachers and other non-technical stakeholders. Furthermore, implicit within the percentile quantification of student growth is a statement of probability. Questions of "how much growth is enough?" or "how much is a year's growth?" ask stakeholders to establish growth percentile thresholds deemed adequate. These thresholds establish growth standards that translate to probability statements. In this manner, percentile based growth forms a basis for discussion of rigorous yet attainable growth standards for all children supplying a norm-referenced context for Linn's existence proof (Linn, 2003) with regard to student level growth.

In addition to practical utility, student growth percentiles possess a number of technical attributes well suited for use with assessment scores. The more important theoretical properties of growth percentiles include:

- **Robustness to outliers.** Estimation of student growth percentiles are more robust to outliers than is traditionally the case with conditional mean estimation. Analogous to the property of the median being less influenced by outliers than is the median, conditional quantiles are robust to extreme observations. This is due to the fact that influence of a point on the τ^{th} conditional quantile function is not proportional (as is the case with the mean) to the distance of the point from the quantile function but only to its position above or below the function (Koenker, 2005, p. 44).
- **Uncorrelated with prior achievement.** Analogous to least squares derived residuals being uncorrelated with independent variables, student growth percentiles are not correlated with prior achievement. This property runs counter to current multilevel approaches to measuring growth with testing occasion nested within students (Singer & Willett, 2003). These models, requiring a vertical scale, fit lines with distinct slopes and intercepts to each student. The slopes of these lines represent an average rate of increase, usually measured in scale score points per year, for the student. Whereas a steeper slope represents more learning, it is important to understand that using a norm-referenced quantification of growth, one cannot necessarily infer that a low achieving student with a growth percentile of 60 "learned as much" as a high achieving student with the same growth percentile. Growth percentiles bypass questions associated with magnitude of learning and focus on norm-referencedly quantifying changes in achievement.

- **Equivariance to monotone transformation of scale.** An important attribute of the quantile regression methodology used to calculate student growth percentiles is their invariance to monotone transformations of scale. This property, denoted by (Koenker, 2005) as *equivariance to monotone transformations* is particularly helpful in educational assessment where a variety of scales are present for analysis, most of which are related by some monotone transformation. For example, it is a common misconception that one needs a vertical scale in order to calculate growth. Because vertical and non-vertical scales are related via a monotone transformation, the student growth percentiles do not change given such alterations in the underlying scale. This result obviates much of the

discussion concerning the need for a vertical scale in measuring growth.[7]

Formally, given a monotone transformation h of a random variable Y,

$$Q_h(Y)|X(\tau|X) = h(Q_Y|X(\tau|X)). \tag{5}$$

This result follows from the fact that $\Pr(T < t|X) = \Pr(h(T) < h(t)|X)$ for monotone h. It is important to note that *equivariance to monotone transformation* does not, in general, hold with regard to least squares estimation of the conditional mean. That is, except for affine transformations h, $E(h(Y)|X) \neq h(E(Y|X))$. Thus, analyses built upon mean based regression methods are, to an extent, scale dependent.

[7] As already noted with regard to pediatrics, the existence of nice "vertical" scales for measuring height and weight still leads to observed changes being normed.

7 The Measurement Error Problem and Corrections

Measurement error (ME) is an inherent component of all standardized tests, and the impact that ME can have when test score data are used to compute student growth and teacher/school evaluation measures has been the focus of a growing body of academic research. Specifically in the area of Student Growth Percentile (SGP) measures of student progress, ME has been found to create bias that can disadvantage students with lower prior achievement and vice versa. This bias is transferred to aggregate measures of educator effectiveness when a disproportionate number of students with relatively low/high prior achievement are concentrated in a classroom or school (Lockwood & Castellano, 2015; Shang, VanIwaarden, & Betebenner, 2015).

Researchers have proposed several useful methods for correcting the effects of ME. The use of Simulation/Extrapolation (SIMEX) techniques has been found to effectively eliminate the ME induced bias related to prior achievement in SGPs (Shang et al., 2015), and this method is a currently available for SGP calculation in the SGP package (Betebenner et al., 2019) for the R statistical program. Currently several states utilize the SIMEX corrected measures in their student growth modeling and evaluation policies.

7.1 The SIMEX Method

The SIMEX method was proposed by Cook and Stefanski (1994) as a measurement error (ME) correction technique when the standard error of measurement (SEM) is known or can be reasonably well estimated. The SIMEX method is a functional approach that does not make strong assumptions about variable distributions (Battauz, Bellio, & Gori, 2011). Compared with other methods, SIMEX is much easier to implement for measurement error models that are less understood, such as that involving nonparametric quantile regression (QR). It relies on repeated random sampling to solve the problem, similar to bootstrap or jackknife, hence its simplicity and generality (L. Stefanski & Cook, 1995). For a detailed description and discussion of SIMEX see Carroll, Ruppert, Stefanski, & Crainiceanu (2006).

The basic idea of the method is to gauge the dependence of the ME effect on SEM through a series of experiments. Increasing amounts of simulated ME are added to observed values, and results from these experiments are then used to extrapolate the relationship of interest to the point where SEM is equal to zero. To explain further, let σ_{ui}^2 stand for the variance of the ME term, u_i. In the simulation phase, additional ME with known variance is generated and added to the observed test scores, w_i, to create increasingly error-prone "pseudo" data sets and then "pseudo" parameter estimates in the following steps. First, choose a set of monotonically increasing small numbers, denoted as λ. For example, let $\lambda = 0.5, 1, 1.5, 2$. Then, for each value of λ, produce an artificial error $\sqrt{\lambda}v_i$, where v_i is randomly generated from the distribution of u_i. The inflated ME, $u_i + \sqrt{\lambda}v_i$, would have a variance equal to $(1+\lambda)\sigma_{ui}^2$. Next, the "pseudo" data sets which are contaminated with the inflated ME are used to produce the "pseudo" parameter estimates with the chosen statistical model. In order to reduce sampling noise, the simulation and "pseudo" estimation are repeated for B times, and the sample mean of the B "pseudo" parameter estimates is calculated at each given λ. In the extrapolation stage, the averaged "pseudo" parameter estimates and the "naive" estimates (the original estimates obtained from the unperturbed data) are regressed on λ. Finally, when λ is set to be equal to -1, the predicted value of the extrapolant function would be the SIMEX estimate of the

error-free parameter.

In the SGP model, the interest lies in estimating $\widehat{SGP}_X$, and these quantities are derived from the fitted values of the model, not its regression coefficients. Following the example of Carroll et. al. (1999), the SIMEX process described above is carried out on the fitted values: "pseudo" fitted value estimates, $\hat{Q}_W^{(\tau)}(\lambda, b)$, for each of the $\tau = 1, 2, \ldots$ 99 percentiles are obtained with the repeatedly perturbed "pseudo" data sets. These values are averaged over B at each λ, regressed on λ, and finally extrapolated to $\lambda = -1$ to produce the SIMEX estimate $\hat{Q}_{(X,SIMEX)}^{(\tau)}$. In the case of quantile crossing, $\hat{Q}_{(X,SIMEX)}^{(\tau)}$ is sorted at the specific x_i, as recommended in Dette and Volgushev (2008) and Chernozhukov, Fernandez-Val and Glichon (2010).

The choices of λ, B, and extrapolation function demand explanations. Various authors provided rules-of-thumb (Carroll et al., 2006, etc.; see, for example, L. Stefanski & Cook, 1995). The commonly adopted values for λ are a few equally spaced numbers between 0 and 2; B is usually fixed at 100; and the extrapolant function is often specified to be linear, quadratic, or non-linear regressions. We conducted Monte Carlo experiments to compare linear with quadratic extrapolants under various λ specifications. Our results show that the linear extrapolation is generally a better choice than the quadratic. With very fine λ grid, such as $\lambda = 0, 2/25, 4/25, \ldots, 50/25$, the quadratic SIMEX estimator of SGP is slightly less biased than the linear one, but, with a much larger variability, its MSE is still considerably higher than that of the linear estimator. As for the choice of λ, a finer grid significantly improves the quadratic estimator but makes little difference for the linear one. The MSE of the SIMEX estimator decreases monotonically as B increases, but the return diminishes for $B > 30$. The detailed results are omitted. For additional information on the use of the SIMEX method to correct for covariate measurement error, see Shang, VanIwaarden and Betebenner (2015).

7.2 The Effect of the SIMEX Method on Individual SGPs

The following sections examine the effect of the SIMEX method on individual SGP estimates and aggregations of them at the class/school level respectively. Results from several simulation studies are presented, followed by a discussion section. Findings from the application of the SIMEX method to actual state assessment data and practical issues to consider based for production level application of SIMEX are also provided.

The performance of the SIMEX method with the SGP model is tested in the following steps. Scale scores in `sgpData` of the `SGP` package (a toy dataset extracted from longitudinal assessment data) were treated as the "true scores" for a two cohorts of students. The first cohort consists of 6,977 4^{th} grade students with a single 3^{rd} grade prior scale score available. The other consists of 6,468 5^{th} grade students with 4^{th} and 3^{rd} grade priors scores available. Missing values were excluded. Normally distributed measurement errors were then generated using the conditional standard errors of measurement of the various grades provided in `SGPstateData` in the same package. The "observed" scores are the sum of the "true" scores and the generated measurement errors. SGP analysis on the "true" scores and the SGP with SIMEX analysis on the "observed" scores are then respectively to produce "true" SGPs, "observed" (or "naive") SGPs, and SIMEX corrected SGPs. To minimize sampling errors, 100 "observed" data sets were generated using the above method. Each student ultimately has a true SGP, 100 observed SGPs from the 100 generated data sets, and 100 SIMEX SGPs from the same data sets.

7.2.1 Quantifying the Performance of the SIMEX Method

Bias and Mean Squared Error (MSE) are the usual ways of quantifying the performance of statistical methods. In a simulation study where the major outcome is the estimation of a few model parameters, both are usually straight forward to calculate—bias is the difference between the estimated and the true parameter averaged over simulation replications, and MSE is the squared difference between the estimated and the true parameter averaged over simulation replications.

In the simulation study of the SGP analysis, however, each replication of the analysis produces results for thousands of students. To evaluate the estimators, a way to summarize across both students and replications is needed. The following statistics to compare SGPs estimated with and without SIMEX are derived for this purpose:

$$Total\ MSE\ = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{R}\sum_{r=1}^{R}(\widehat{SGP}_{i,r} - SGP_i)^2 \tag{6}$$

Where $i = 1, 2, \ ... \ , n$ indicates students, $r = 1, 2, \ ... \ , R$ indicates simulation replications, $\widehat{SGP}_{i,r}$ denotes the estimated SGP (either with or without SIMEX) for student i in replication r, and SGP_i denotes the "true" SGP for student i.

$$Mean\ Bias\ = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{R}\sum_{r=1}^{R}(\widehat{SGP}_{i,r} - SGP_i) \tag{7}$$

$$SD\ Bias = Standard\ Deviation\ _{across\ i}(\frac{1}{R}\sum_{r=1}^{R}\widehat{SGP}_{i,r} - SGP_i) \tag{8}$$

Total MSE is a comprehensive standard which accounts for both bias and variability. *Mean Bias* is the bias averaged over students and replications, but since *Mean Bias* of either estimator is likely to be very close to 0, we also calculated *SD Bias*, which is derived by taking the difference between the estimated and the true SGPs averaged across replications, and then calculate the standard deviation of these differences across students. *SD Bias* may be a better indicator of the magnitude of bias than *Mean Bias*, just like Standard Error of Measurement is a better indicator of the magnitude of measurement errors than the mean, which is usually 0.

7.2.2 Results

The following table presents the results of the simulation study.

Table 3:

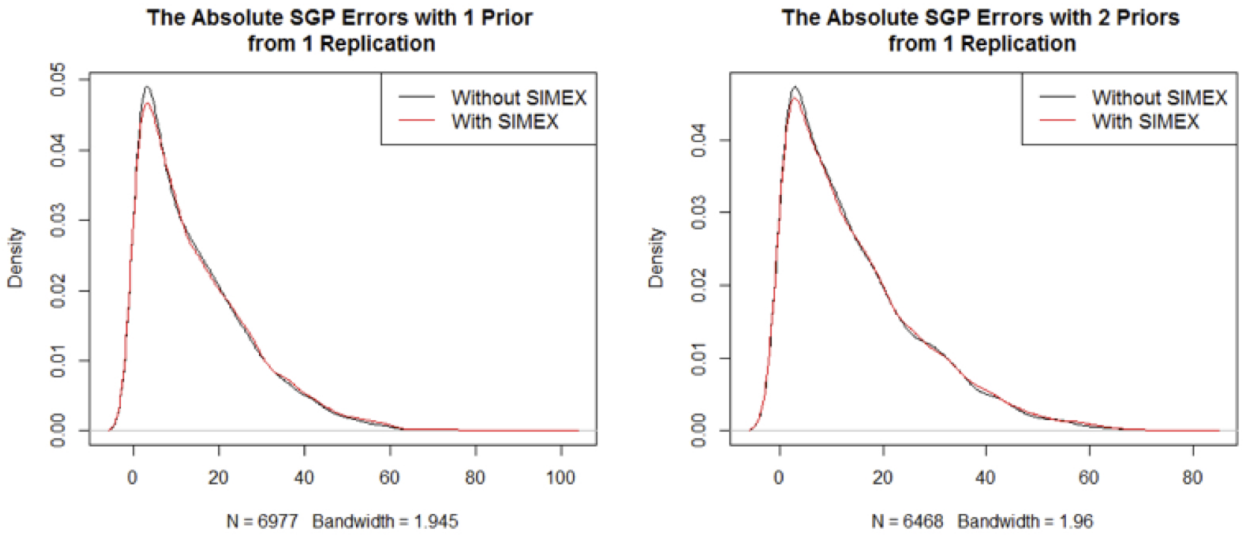
	SGP Without SIMEX			**SGP With SIMEX**		
	Total MSE	Mean Bias	SD Bias	Total MSE	Mean Bias	SD Bias
Gr 4 Gr 3	362.980	0.019	8.004	393.244	-0.107	6.612
Gr 5 Gr 4, 3	360.795	-0.001	7.512	379.277	0.088	6.602

7.2.3 *Why Are the Total MSEs So Large?*

The total MSEs are all very large. The square roots of the MSEs, generally considered a measure of the average magnitude of errors, are around 19, far greater than the *Mean Bias* or the *SD Bias*.

Figure B.5 shows the density of the absolute values of the SGP (with and without SIMEX) errors with 1 and 2 prior year scores respectively from a single replication of the simulation study. Plots of other replications are highly similar. Figure B.5 shows that, in any given replication of the simulation study, the absolute differences between estimated and true SGPs are highly skewed with a thin but long tail to the right. A small number of the errors in SGPs estimated with 1 prior are close to 100. Among SGPs estimated with 2 priors, it seems that there are less extreme outliers. SGPs estimated with SIMEX seem to mirror those without SIMEX closely, although the former tend to have more very large errors. The relatively small number of outliers can have dramatic influence on the MSE because, when squared, they become *much* larger and the skewness of the distributions of the squared errors will dramatically increase compared with those in Figure B.5.

Figure. B.5: Absolute values of SGP Errors with 1 and 2 Priors Used.



7.2.4 *Comparing the SIMEX Results with the Non-SIMEX Results*

Due to the skewness discussed in the section above, the root mean square error is not a good measure of the magnitude of errors in this context. The results of the SGP analysis with and without SIMEX are therefore based on two criteria—*SD Bias* and *MSE*. The comparison is straightforward: the SIMEX results have smaller biases but larger MSEs which indicates larger variability across replications. Furthermore it seems that in both scenarios (i.e. with 1 and 2 priors) the increase of variability outweighs the reduction of bias.

There is, however, something strange here. In earlier studies of the SIMEX method for correcting biased model coefficients, we noticed much bigger bias reductions than those presented in Table 3. The SIMEX estimators do generally have larger variability, but usually this

drawback is outweighed by the prominent gains in bias reduction. Why is the SIMEX method not doing well this time?

7.2.5 *Why Doesn't the SIMEX Method Perform Well in Correcting Individual SGP?*

The question about the validity of using the SIMEX method to correct individual level SGPs can be raised with the following rationale. Suppose that the model $\hat{y}_{GOOD} = a + bx_{OBSERVED}$ is estimated when $x_{OBSERVED}$ is measured with errors, but $\hat{y}_{BETTER} = \alpha + \beta x_{TRUE}$ is the actual model desired. Now the SIMEX method, or any other measurement error correction, moves us from a and b closer to α and β, and often times the goal is achieved at this stage. But in the SGP analysis, simply obtaining α and β is not enough. To get more accurate SGPs, we need to obtain $\hat{y}_{BETTER}$, which is not obtainable unless we have x_{TRUE}.

Therefore, when estimating individual SGPs with SIMEX, what is roughly estimated is $\alpha + \beta x_{TRUE}$, which is neither $\hat{y}_{GOOD}$ nor $\hat{y}_{BETTER}$. Thus some bias reduction is achieved because the observed scores are usually not far from the true scores, but the effect is much discounted.

To test this theory, the "true" scores were plugged into the SIMEX process. Table 4 presents the outcomes of SGP without SIMEX and with SIMEX combined with "true" scores from 100 simulation replications.

Table 4:

		SGP Without SIMEX			SGP With SIMEX and 'True' Score		
		Total MSE	Mean Bias	SD Bias	Total MSE	Mean Bias	SD Bias
Gr 4	Gr 3	362.980	0.019	8.004	248.836	-0.428	5.652
Gr 5	Gr 4, 3	360.795	-0.001	7.512	287.078	-0.057	6.225

In Table 4, the SIMEX method greatly reduces both bias and MSE when used in combination with true scores. This shows that the fundamental reason for increased MSE in Table 3 is the use of SIMEX in combination with observed scores.

7.2.6 *Can We Use Estimated True Scores?*

Results in Table 4 are obviously not obtainable in reality. As a possible alternative, the estimated true scores were substituted:

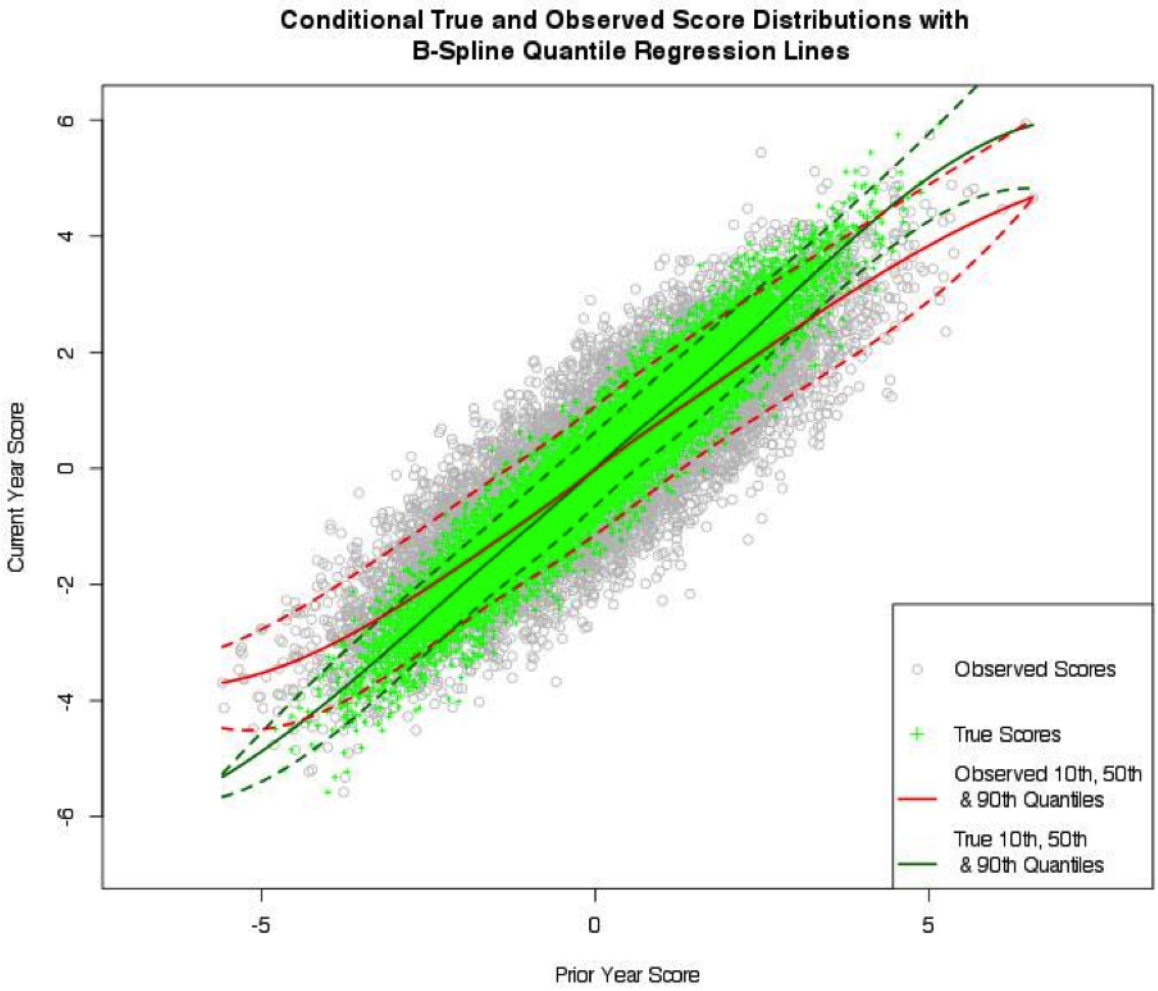
$$X_{Est.True} = \text{Reliability} \times (X_{OBSERVED} - \bar{X}_{OBSERVED}) + \bar{X}_{OBSERVED} \tag{9}$$

The results obtained are almost exactly the same as those in Table 3, which suggests that the inaccessibility of the true scores is a problem that is insurmountable without measurement replications. The transformation of the observed scores to estimated true scores does not change the rank order of the observations, and so their use in the SIMEX model yields very similar results to the use of observed scores in that model.

7.2.7 Visual representations of the SIMEX method corrections

Visualizations that highlight the impact of implementing the SIMEX method may provide a more intuitive understanding of the SIMEX SGP corrections and how they are obtained. Figure B.6 shows true and observed score distributions with quantile regression models fit to those two distributions. The simulated data points are bivariate normally distributed with homoscedastic, classical error structure in the prior year data. As expected, the slopes of quantile regression lines fit to the observed data are attenuated. Another notable feature is that the separation between the True 10^{th} and 90^{th} percentile lines is narrower than that between the corresponding lines fit to the observed data.

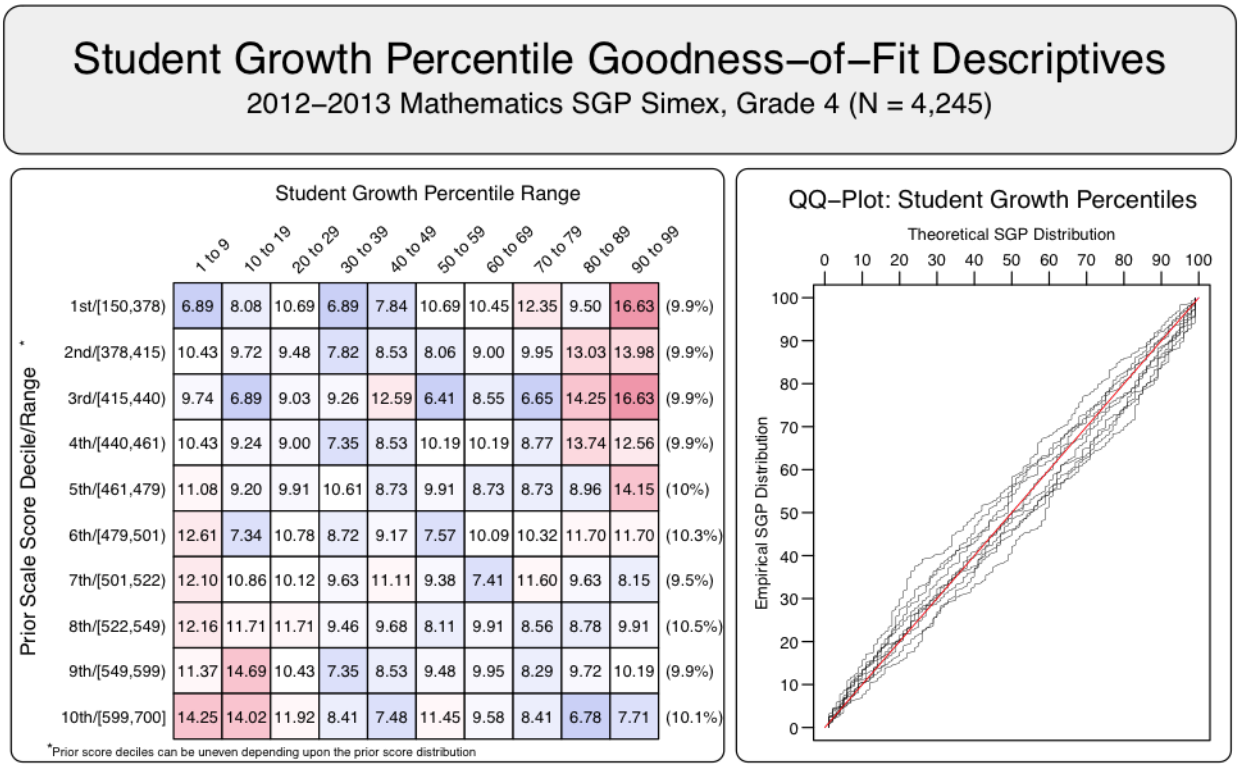
Figure. B.6:



The SIMEX method attempts to use quantile regression to produce a "map" of the observed score distribution that more closely resembles the map of the unobserved, True score map. If the observed data (grey circles in Figure B.6) were fit with the True quantile regression models (green lines), a disproportionately large amount of high SGPs would be expected at the lower tail of the prior score distribution and, conversely, more low SGPs at the upper end of the prior score distribution. Additionally, the compressed quantile lines in the SIMEX model of

the distribution would likely create fewer than expected SGPs in the middle percentile ranges. This would result in clustering of SGPs where the model cannot differentiate between similar (but not identical) score histories.

Figure. B.7: Goodness of Fit plot for SIMEX SGPs obtained from one 4^{th} grade simulation.



These two expected features are present in Figure B.7, which is a goodness of fit plot produced from a 4th grade Math SIMEX simulation. These plots obtained from the simulations conducted for this report suggest that the SIMEX method has produced a "better" map of what the True conditional distribution *may* look like.

7.2.8 Baseline Referenced SGPs

In addition to the issues discussed above for individual cohort SGPs, there are some theoretical issues that should be taken into account for the decision to use SIMEX corrections for baseline referenced SGPs. An ideal baseline model would show perfect fit (10 percent of kids in each cell of the left hand table in Figure B.7). From this baseline model we hope to find evidence of system wide improvement when future scores are fit to it. That is, the baseline-referenced model can be seen as a predictive model in which we use past *observed* score distributions to predict the distribution of current *observed* scores. Unlike predictive models that attempt to predict future observations with the greatest accuracy possible, we hope that fitting the current data with this model will show misfit due to systemic changes in the student population. In particular, we hope to see higher growth in the present than what was expected (typical) in the past. This would be show up as red cells on the right side of the

fit table in Figure B.7. Starting from a baseline that is already shifted due to measurement error "correction" can muddle this picture.

The measurement error literature suggests that predictive models should not use correction methods for measurement error when error prone data will be used to produce predicted values. Carroll et al. (2006, p. 38) provide this explanation,

> Generally, there is no need for the modeling of measurement error to play a role in the prediction problem. If a predictor, X, is measured with error and one wants to predict a response based on the error-prone version W of X, then ... it rarely makes any sense to worry about measurement error. The reason for this is quite simple: W is an error-free as a measurement of itself! If one has an original set of data $[Y, W]$, one can fit a convenient model to Y as a function of $[W]$. Predicting Y from $[W]$ is merely a mater of using this model for prediction, that is substituting known values of W into the regression model for Y on $[W]$; the prediction errors from this model will minimize the expected squared prediction errors in the class of all linear unbiased predictors. Predications with W naively substituted for X in the regression of Y on $[X]$ will be biased and can have large prediction errors.

In the SIMEX SGP analyses, using observed scores in place of true scores in a SIMEX corrected baseline model is equivalent to making "[p]redications with W naively substituted for X in the regression of Y on $[X]$". The large prediction errors would be evidenced in the initial non-uniform distribution of SGPs expected in the SIMEX model as shown in Figure B.7. These "prediction errors" would confound any information about system wide improvement (or decline) that we hope to see through model misfit as detailed above.

On the other hand, Carroll et al. (2006) also specify an exception: "The one situation requiring that we correctly model the measurement error occurs when we develop a prediction model using data from one population but we wish to predict in another population. A nave prediction model that ignores measurement error may not be transportable" (p. 39). We do not presently believe that this situation applies to the baseline-referenced analysis. We begin by assuming that current and future annual cohorts come from the baseline cohort population. If the model shows that this assumption is not reasonable, then we potentially have evidence that the system may be changing. Further research and simulations may provide some empirical evidence to clarify the issue.

7.3 The Effect of the SIMEX Method on Aggregated SGPs at the Class or School Level

Our conclusion that the SIMEX method should not be used to estimate individual SGPs does not automatically lead to the conclusion that the SIMEX method is not useful for correcting SGP aggregations (e.g. median growth percentiles, or MGPs), either. On the one hand, the aggregated mean/median prior scores of a class/school should be much closer to the true aggregated scores compared with individual scores, which means that the problem mentioned in preceding sections is partially solved and the SIMEX correction could be beneficial. On the other hand, the aggregated mean/median SGPs of a class/school should also be close to the true aggregated SGPs, which raises the question whether the SIMEX correction is necessary.

To investigate the utility of the SIMEX method at the school level, we examined three scenarios—randomly assigned schools, perfectly sorted schools, and actual schools which are usually somewhere in the middle. In each scenario, we looked at the bias of aggregated scale scores and SGPs, the MSE of MGPs estimated with and without SIMEX, and the correlation between MGP (with and without SIMEX) and prior aggregated scores. From now on, we use the word "school" to refer to a collective student body, which may be a class or a school. Since we simulate schools of small and large sizes (as small as 10 students, and as large as 100), the investigation may generalize to both classes and schools.

7.3.1 Randomly Assigned Schools

Because measurement errors are random and independent of each other *aggregated* scale scores or SGPs are much more accurate than *individual* scale scores or SGPs. We test this claim empirically in the simulation described in the first section of this appendix. Students are randomly sampled without replacement into schools of various sizes. We calculated the difference between mean observed and true scores, the difference between median observed and true scores, and the difference between individual observed and true scores. Figure B.8 plots the errors of Grade 3 scores with school sizes of 10, 20, and 50. Figure B.9 plots the errors of aggregated and individual SGPs of Grade 4. The data are drawn from a single replication of the simulation study. Plots of other grades and other replications are similar. The same plots drawn in the other scenarios, i.e. perfectly sorted schools and actual schools, are also highly similar and are not presented.

Figure. B.8: Errors of Aggregated and Individual Scale Scores of Grade 3.

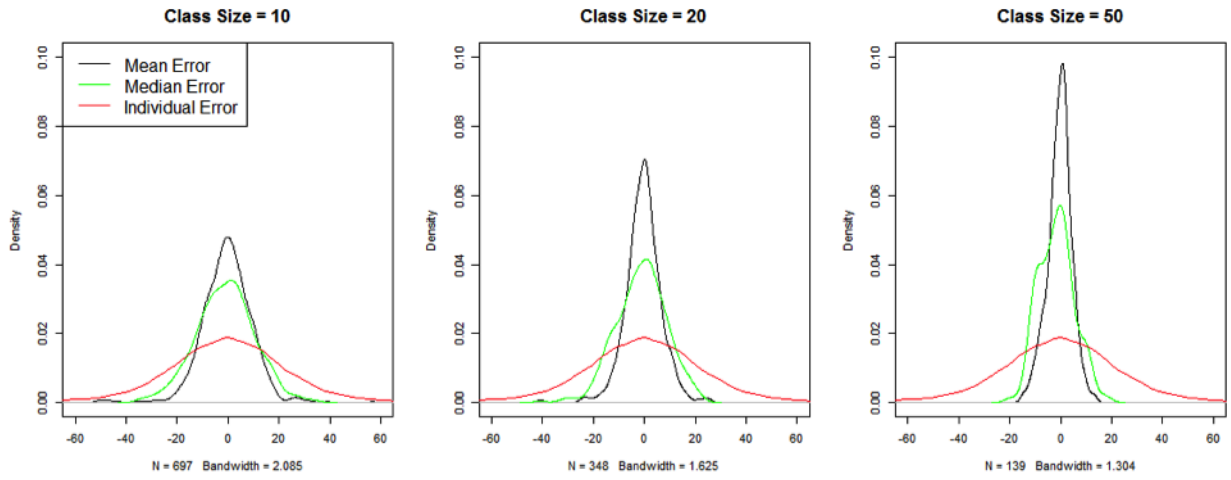


Figure. B.9: Errors of Aggregated and Individual SGPs of Grade 4.

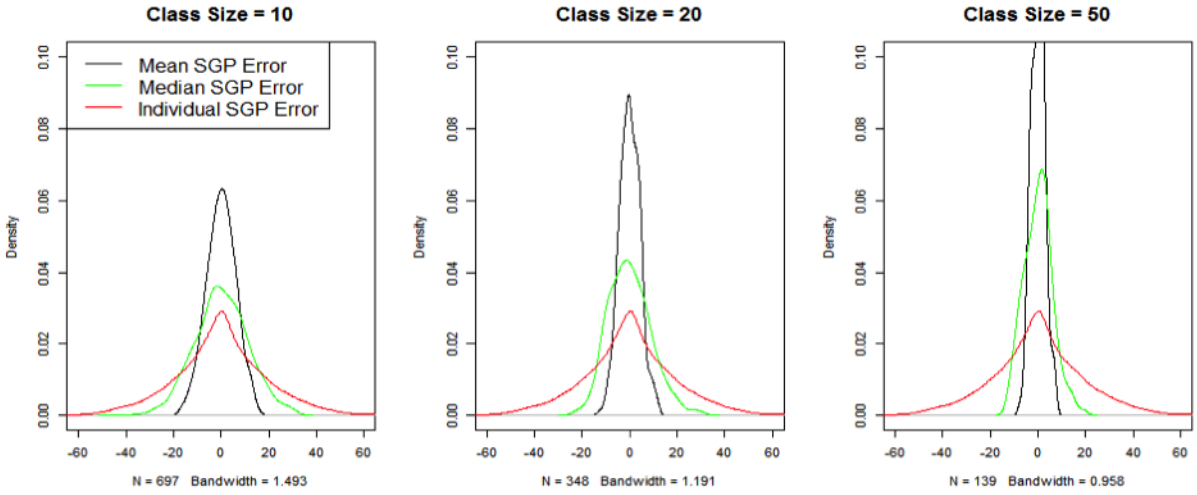


Figure B.10 and B.11 show that, when error-prone scores and SGPs are aggregated, even in schools of just 10 students, the magnitudes of errors are greatly reduced. The error reduction improves considerably as school sizes go up.

The next question is, can the SIMEX method improve the accuracy and precision of aggregated SGPs when students are randomly assigned to schools? Table 5 presents MSE of mean and median SGPs of schools of various sizes.

Table 5: Mean Squared Errors of Aggregated SGPs of Randomized Schools of Various Sizes

	Mean School SGP				Median School SGP			
	MSE 10*	MSE 20	MSE 50	MSE 100	MSE 10	MSE 20	MSE 50	MSE 100
Grade 4 ~ 3								
No SIMEX	36.39	17.82	7.15	3.57	131.21	83.29	39.07	18.49
SIMEX	39.29	19.45	7.79	3.96	142.2	89.92	42.82	20.48
Grade 5 ~ 4 and 3								
No SIMEX	35.95	17.94	7.10	3.50	129.90	83.99	35.50	18.89
SIMEX	37.78	18.85	7.49	3.69	137.00	88.66	38.32	20.11

** MSE 10 Means "Mean Squared Error for schools of 10 Students"*

Table 5 shows that MSE of aggregated SGP_SIMEX are very close to and always slightly larger than MSE of non-corrected SGPs. It also shows that mean SGP seems to have much smaller MSE than median SGPs.

Lastly, we look at the correlation between aggregated SGP and prior scale scores. Table 6 presents the correlations between aggregated observed SGP (No SIMEX) and prior scores, the correlations between SGP_SIMEX and prior scores, and the correlations between error-free SGP (true) and prior scores for schools of various sizes. Table 6 shows that aggregated SGPs hardly correlate with aggregated prior scores in randomly assigned schools, and that the SIMEX correction tends to make the correlation estimations more erroneous rather than accurate, if only at a small scale.

Table 6: Correlations of Aggregated SGPs and prior scores of Randomized Schools of Various Sizes

	Correlation of *Mean* SGP and prior score				**Correlation of *Median* SGP and prior score**			
	Size 10*	Size 20	Size 50	Size 100	Size 10	Size 20	Size 50	Size 100
Grade 4 ~ 3								
No SIMEX	0.061	0.018	0.014	-0.004	0.045	0.028	-0.004	0.036
SIMEX	0.040	-0.084	-0.083	-0.117	-0.029	-0.040	-0.067	-0.033
True	0.071	0.025	0.016	-0.026	0.057	0.027	0.107	-0.062
Grade 5 ~ 4 and 3								
No SIMEX	0.003	0.062	0.122	0.040	-0.016	0.013	0.104	0.051
SIMEX	-0.058	0.002	0.066	-0.019	-0.062	-0.029	0.062	0.020
True	0.003	0.078	0.167	0.031	-0.039	0.000	0.130	0.059

* *Size 10 Means schools of 10 Students*

7.3.2 Perfectly Sorted Schools

For the second scenario, we assume that the assignment of students to different schools is completely based on the ranking of their prior true scores. Tables 7 and 8 present the MSE and correlations calculated in this scenario similar to that of Tables 5 and 6. These tables show that MSE of aggregated SGPs increases dramatically when students are sorted compared with random assignment into schools. The large MSEs are somewhat mitigated when SGPs are conditioned on more than one prior scores.

Correlations between aggregated SGPs and prior scores are also greatly inflated, suggesting underestimation of MGP for under-achieving schools and overestimation of MGP at the other end. The good news is that the SIMEX method seems to be effective to a certain degree in reducing MSE, especially for larger schools. The SIMEX method also performs well in reducing the correlation between MGP and prior scores. In fact, it almost completely removes the inflation of the correlations. This confirms our earlier expectation that the SIMEX method might not be able to eliminate errors, but it can, under the right circumstances, quite effectively turn systematic bias into random errors.

Table 7: Mean Squared Errors of Aggregated SGPs of Sorted Schools of Various Sizes

	Mean School SGP				Median School SGP			
	MSE 10*	MSE 20	MSE 50	MSE 100	MSE 10	MSE 20	MSE 50	MSE 100
Grade 4 ~ 3								
No SIMEX	161.88	95.51	47.97	32.29	376.52	238.27	126.00	84.25
SIMEX	159.43	85.32	40.57	24.12	388.88	231.30	114.85	69.78
Grade 5 ~ 4 and 3								
No SIMEX	161.93	81.13	36.93	25.73	350.78	206.40	97.86	67.87
SIMEX	161.87	78.91	32.44	19.32	360.49	209.31	90.24	57.81

* *MSE 10 Means "Mean Squared Error for schools of 10 Students"*

Table 8: Correlations of Aggregated SGPs and prior scores of Sorted Schools of Various Sizes

	Correlation of *Mean* SGP and prior score				**Correlation of *Median* SGP and prior score**			
	Size 10*	Size 20	Size 50	Size 100	Size 10	Size 20	Size 50	Size 100
Grade 4 ~ 3								
No SIMEX	0.305	0.420	0.577	0.645	0.263	0.364	0.515	0.602
SIMEX	0.018	0.057	0.113	0.133	0.011	0.047	0.100	0.120
True	0.005	-0.058	0.077	0.107	0.007	-0.039	0.006	0.029
Grade 5 ~ 4 and 3								
No SIMEX	0.203	0.298	0.415	0.496	0.173	0.252	0.356	0.443
SIMEX	0.021	0.047	0.070	0.084	0.013	0.036	0.054	0.076
True	0.005	0.009	-0.011	-0.220	-0.015	0.005	-0.111	-0.192

* *Size 10 Means schools of 10 Students*

7.3.3 Actual Classes/Schools

Neither the randomized or perfectly sorted schools scenarios are very realistic. To get a sense of what may happen in reality, we turn to the `sgpData_LONG` dataset contained in the `SGP` R package, which is a toy dataset based on real schools in a different state. We consider scale scores in this dataset as "true" scores, and perturb them in the same way as in the previous simulations. For each perturbed data, individual and aggregated SGP and SGP_SIMEX are estimated. We calculated the total MSE, the Mean Bias, and the SD Bias, defined in equations (1), (2), and (3). Results from the Reading test in Grades 4 and 5 conditioning on 1 and 2 priors respectively are presented in Table 9. Correlations of aggregated SGP and prior scores are presented in Table 10.

Table 9: Mean Squared Errors of Aggregated SGPs of Actual Schools of Various Sizes

	Mean School SGP			Median School SGP		
	Total MSE	Mean Bias	SD Bias	Total MSE	Mean Bias	SD Bias
Grade 4 ~ 3						
Without SIMEX	5.574	-0.196	2.360	23.366	-0.104	5.137
With SIMEX	4.741	-0.075	2.177	28.364	0.034	5.329
Grade 5 ~ 4 and 3						
Without SIMEX	5.948	-0.263	2.433	27.101	-0.266	5.199
With SIMEX	4.519	0.094	2.125	25.679	0.084	5.060

Table 9 shows that MSEs of mean SGP from actual schools are quite small to begin with, and the SIMEX method is able to further reduce it to a limited extent. Bias and MSEs of median SGPs are much larger, and the effect of the SIMEX method on median SGPs seems to be mixed. Table 10 shows that there are considerable correlations between observed aggregated SGPs and prior scores, and that even the error-free SGPs correlate significantly with prior scores. This indicates that measurement error is not the only source of the correlation. When conditioned on more than one prior score, the correlations are much reduced. When the SIMEX method is applied, the correlations are further reduced to a level that is quite close to the correlation between error-free SGPs and prior scores.

Table 10: Correlations of Aggregated SGPs and prior scores of Actual Schools of Various Sizes

	Correlation of *Mean* SGP and prior score		Correlation of *Median* SGP and prior score	
	Mathematics	Reading	Mathematics	Reading
Grade 4 ~ 3				
No SIMEX	0.384	0.403	0.372	0.404
SIMEX	0.294	0.291	0.281	0.299
True	0.292	0.301	0.284	0.330
Grade 5 ~ 4 and 3				
No SIMEX	0.220	0.256	0.172	0.252
SIMEX	0.152	0.138	0.105	0.140
True	0.159	0.121	0.093	0.119

7.4 SIMEX implementation in the SGP package

The `SGP` package (Betebenner et al., 2019) allows the user to specify many of the parameters used in the production of SIMEX SGP estimates. The `calculate.simex` argument of the `studentGrowthPercentiles` function requires the user to specify the following SIMEX parameters in a list with the following named elements:

- `state` identifies the two letter state abbreviation under which the test specific CSEMs are located in `SGPstateData`, or
- `csem.data.vnames` identifies the variable name of the variable in the longitudinal data that contains the score specific CSEM value, if not provided in the `SGPstateData`
- `lambda` and `simulation.iterations` specify the desired values of λ and B respectively, and
- `extrapolation` to select a "linear" or "quadratic" extrapolant function.

The user may also request optional functionality, including

- `simex.sample.size` to specify a sample size of the data to be used in the production of the coefficient matrices[8],

[8] Because the time taken to produce a coefficient matrix increases exponentially as the number of students

- `save.matrices` to choose to save the coefficient matrices produced during each simulation experiment (`TRUE` or left `NULL` if not desired), and
- `simex.use.my.coefficient.matrices` to use previously computed coefficient matrices, if available (`TRUE` or left `NULL` if not), to produce fitted value estimates.
- `dependent.var.error` to add error to the dependent variable as well as the independent (covariate) variable(s).

When the `calculate.simex` argument is `TRUE` in the high-level function `analyzeSGP` (rather than providing a list as described above) the package defaults are used. These defaults are to set λ to 0.5, 1, 1.5, 2 and B as 75, the sample size is set at 5,000, and the linear extrapolant is used. When computing cohort referenced SIMEX SGPs new coefficient matrices will be produced, used and saved. Previously computed coefficient matrices are used for baseline referenced SGPs (see the section below regarding SIMEX baseline matrix construction).

Internally, the `studentGrowthPercentiles` function first uses the "naive" coefficient matrices (either calculated previously or during the current run) to obtain the "naive" fitted values from the unperturbed observed test scores. The (non-zero) values of λ are then iterated over, simulating B new data sets from the observed values each time. New coefficient matrices are produced if requested using each of the B data sets.[9] The function then finds the appropriate coefficient matrix that was either just produced, or is available in the `panel.data` list provided by the user or included in the package's `SGPstateData`. Once the appropriate matrices have been located, the fitted value predictions at each percentile value are produced and then averaged over the B simulation iterations. Once these averages are obtained for each value of λ, the extrapolant function is applied to them to estimate the predicted value at $\lambda = -1$ is extrapolated for each student. These (extrapolated) predicted value estimates and the original observed scores are then used to produce SGP values in the typical manner.[10]

7.4.1 Baseline referenced coefficient matrix production

As with the construction of the "naive" baseline coefficient matrices, SIMEX adjusted baseline coefficient matrices can also be produced using a "super-cohort" of students. That is, students with the same grade progression and/or course sequence from multiple academic year cohorts are combined into a single baseline norm group.

These combined cohort norm groups can range in size from several thousand in some of the less typical course progressions to hundreds of thousands in common norm groups. Initial tests of the SIMEX SGP routine on the larger super-cohorts proved to be impossible in the `R` environment, causing numerous technical issues including total memory consumption and other memory related seg-faults. In order to deal with these issues, the ability to take a random sample of the student score histories for construction of the matrices was added to the `SGP` package. The sample size of 25,000 students was found to be adequate for providing

increases a sample size smaller than the population can allow for satisfactory coefficient matrices to be produced in a more time efficient manner. When specified, the student population must be greater than the argument value. Note that the sample is only used to produce these matrices, and all students still receive SIMEX corrected SGP estimates.

[9] This includes producing the knots and boundaries used in the quantile regressions.

[10] From documentation sections not included here - the observed score is compared to all 100 predicted values. A students' SGP is equal to the highest percentile at which the students observed score is greater than or equal to the corresponding predicted (fitted) value.

consistent results, and the number of simulation iterations, B, was increased to 50 in order to compensate for added error from the sampling process (up from 30, the number identified earlier as the point at which the returns to reducing the MSE diminished).

The production of these SIMEX adjusted baseline matrices required several days to produce despite the use of parallel processing. After their construction, the matrices are added to the `SGPstateData` object and used to produce SIMEX adjusted SGPs for any relevant academic year.

7.5 Ranking SIMEX Corrected SGPs

Although SIMEX corrected SGPs are useful in reducing measurement error induced bias, they are not without technical limitations. At an individual level, the corrected SGPs have larger errors than the uncorrected, or "standard", SGPs (McCaffrey, Castellano, & Lockwood, 2015; Shang et al., 2015). McCaffrey, et al. (2015) first suggested that ranking the SIMEX SGP values may present a possible alternative that would have the beneficial properties of both SGP estimate types. Castellano and McCaffrey (2017) recently investigated the properties of the percentile ranked SIMEX SGP (RS-SGP) at the aggregate level for MSGP estimates of educator effectiveness. They found that the majority of the error variance in standard MSGP values is due to "sampling variability" (i.e. a classroom is considered a sample of all possible students), but that a substantial amount was also due to bias caused by ME. SIMEX correction can remove much of this bias initially. By subsequently taking the percentile ranks of the SIMEX corrected values and then aggregating these percentile ranks, the excess variance from the SIMEX estimation is removed. Furthermore they found that, at the individual level, the distribution of the RS-SGPs is also more uniformly distributed (similar to the standard SGP) rather than the SIMEX SGPs typical U-shaped distribution. The uniform distribution of the individual SGP values is a desirable characteristic because it suggests that the full range of SGP growth values (1-99) is equally likely to be attained.

Given the potential promise of RS-SGP, it is now calculated along with the SIMEX values in the `SGP` package[11]. Further insights from that implementation are discussed next.

7.5.1 SGP Package Implementation of Ranked SIMEX SGP

In their study, Castellano and McCaffrey report simply taking the percentile rank of the computed SIMEX SGP values to get the RS-SGP. However, unlike the SIMEX SGP values computed through data simulations in the `SGP` package, they compute their values using a closed-form equation. This produces continuous SIMEX SGP values, which allow for a more detailed ranking than using the integer values computed in the `SGP` package. Although their process helps to better understand the theoretical groundings of the various SGP estimates, it is only appropriate under particular assumptions about the data and ME structures that do not hold in the real-world situations.

Without a continuous value, the percentile ranking[12] of a set of numbers that is already on a percentile scale does not produce results that differ substantially from the original in absolute value or distribution. Therefore a solution was required in the simulation process that would

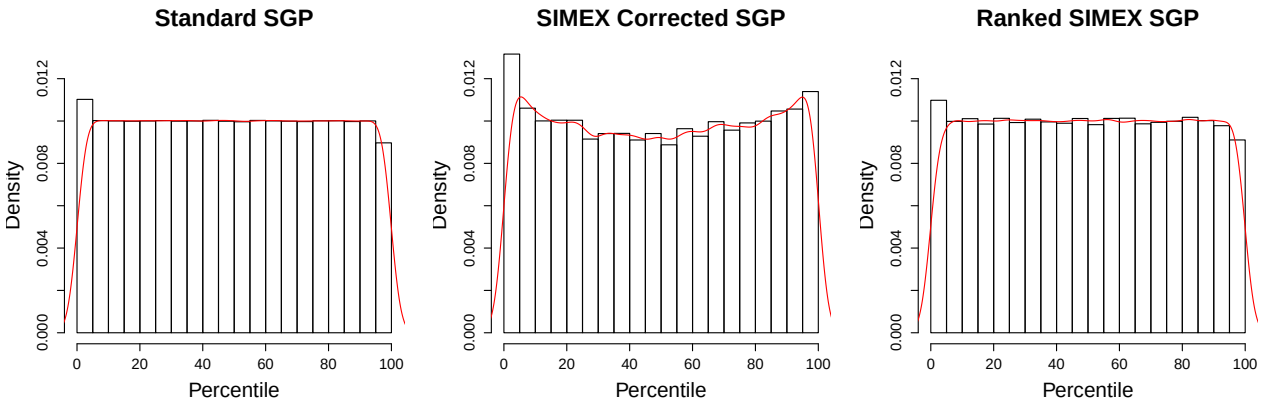
[11] SGP versions 1.7-0.0 and later

[12] Calculated as (rank(SIMEX SGP)/N) x 100 where N is the number of students. The result is rounded to the nearest integer.

allow for a more continuous SIMEX SGP to be established. In following up with the authors they suggested that more granular SIMEX SGPs be established in the simulations, however this would require the already computationally and time intensive process to take 10 times longer. Furthermore, previously calculated SIMEX values would no longer be reproducible. A simpler solution was used that allows the estimated SIMEX values to be placed on a $1/8^{th}$ interval by calculating arithmetic midpoints between each percentile's predicted score values[13]. This resulted in RS-SGP values that were more uniformly distributed in initial tests with real and simulated data.

Figure 10 shows the distribution of the three types of SGP estimates for an 8^{th} Grade Mathematics SGP analyses in a large state: uncorrected ("standard"), SIMEX corrected and Ranked SIMEX. Note that these results are from analyses that use up to three years of data (two prior and the current year), which the authors indicate will also greatly reduce ME bias.

Figure 10: Comparison of the Uniformity of Distributions for 8^{th} Grade Mathematics Estimates.



By definition, the standard SGP is uniformly distributed *given any prior test score*, suggesting that any level of growth is equally likely regardless of prior achievement. This is a critical distinction, and Castellano and McCaffrey do not discuss the conditional uniformity of the RS-SGP. We find that this uniformity is not met in either the application of the closed-form equations to simulated data or in our initial tests with real data in the `SGP` package, although the RS-SGP distribution is much closer to uniform than that of the SIMEX SGPs.

The following figures are "Goodness of Fit" charts that are produced using the `SGP` package for each of the three SGP estimate types, and they can help investigate the SGP distribution in more detail. The "Student Growth Percentile Range" panel at bottom left shows the empirical distribution of SGPs given prior scale score deciles in the form of a 10 by 10 cell grid. Percentages of student growth percentiles between the 10^{th}, 20^{th}, 30^{th}, 40^{th}, 50^{th}, 60^{th}, 70^{th}, 80^{th}, and 90^{th} percentiles were calculated based upon the empirical decile of the cohort's prior year scaled score distribution. Perfect uniform distribution conditional on prior score would be indicated by a "10" in each cell. Deviations from perfect fit are indicated by red and blue shading. The further above 10 the darker the red, and the further below 10 the darker the blue.

[13]SGP estimates are found by predicting 100 scores for each student - one for each percentile. The position (1-99) of the predicted score that is closest to a student's observed score is their estimated SGP.

The bottom right panel of each plot is a Q-Q plot which compares the observed distribution of SGPs with the theoretical (uniform) distribution. An ideal plot here will show black step function lines that do not deviate from the ideal, red line which traces the 45 degree angle of perfect fit (as is seen here in the first plot for the standard SGP).

These plots display typical distributions of each SGP variant from the same 8^{th} Grade Mathematics SGP analyses as depicted above. The Standard SGPs are nearly perfectly distributed conditional upon prior achievement. The SIMEX and, to a lesser extent, RS-SGP distributions are skewed towards higher percentiles at the lower levels of achievement and lower growth for the higher prior achievement deciles.

Figure 11: Goodness of Fit Plot for ***Standard*** 8^{th} Grade Mathematics SGPs.

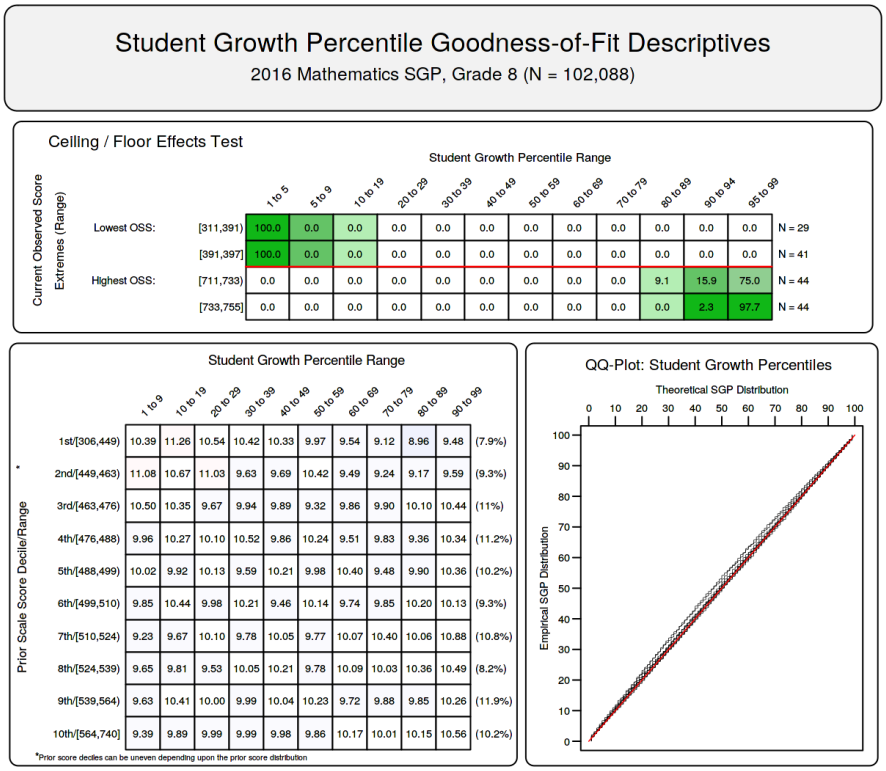


Figure 12: Goodness of Fit Plot for ***SIMEX*** 8^{th} Grade Mathematics SGPs.

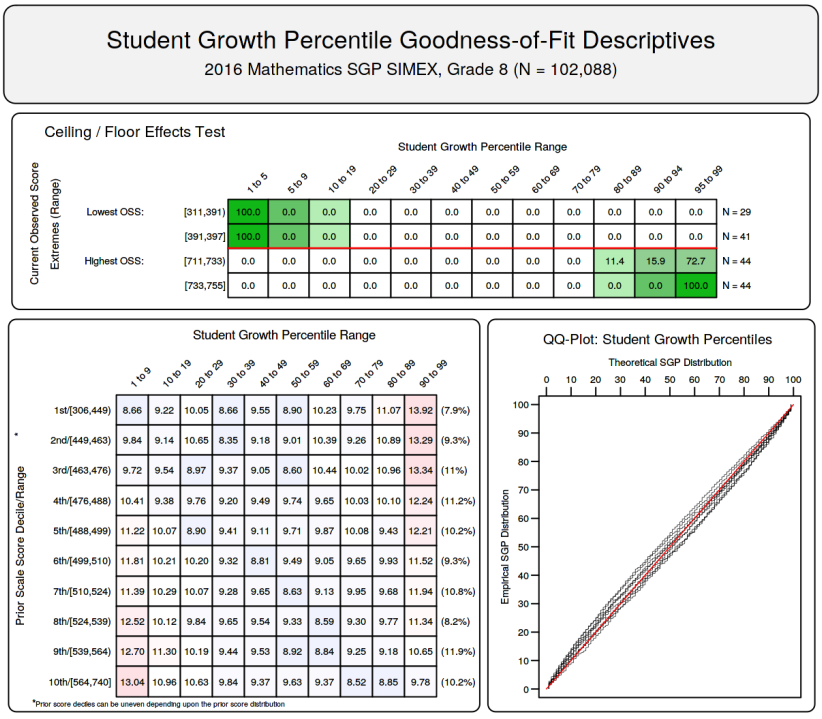
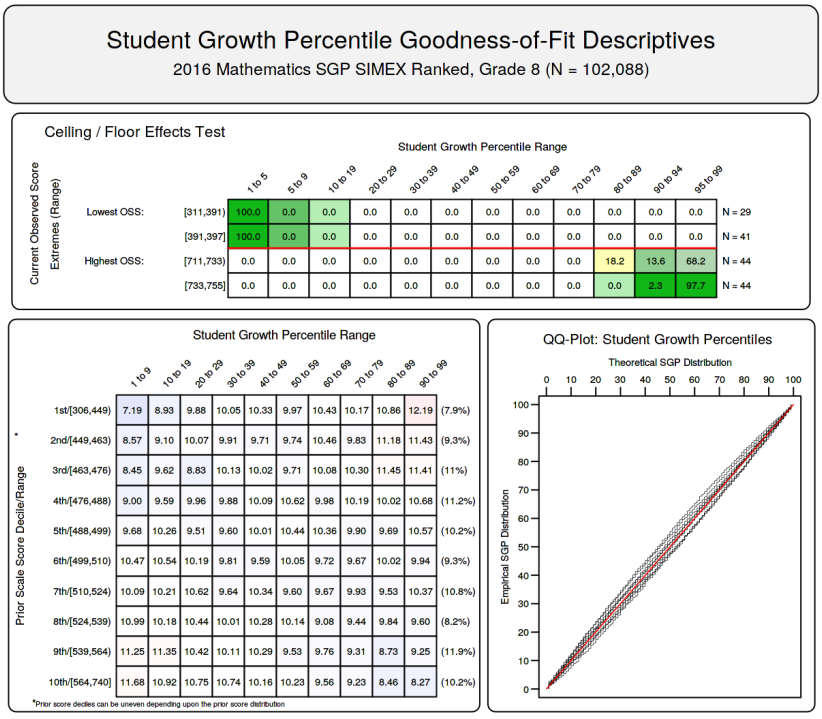


Figure 13: Goodness of Fit Plot for ***Ranked SIMEX*** 8^{th} Grade Mathematics SGPs.



References

Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for student background in value-added assessment for teachers. *Journal of Educational and Behavioral Statistics*, *29*(1), 37–65.

Battauz, M., Bellio, R., & Gori, E. (2011). Covariate measurement error adjustment for multilevel models with application to educational data. *Journal of Educational and Behavioral Statistics*, *36*(3), 283–306. SAGE Publications.

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

Betebenner, D. W., VanIwaarden, A., Domingue, B., & Shang, Y. (2019). *SGP: Student growth percentiles & percentile growth trajectories*. Retrieved from sgp.io

Braun, H. I. (2005). *Using student progress to evaluate teachers: A primer on value-added models*. Princeton, New Jersey: Educational Testing Service.

Carroll, R., Ruppert, D., Stefanski, L., & Crainiceanu, C. (2006). *Measurement error in nonlinear models; a modern perspective*. Boca Raton, FL: Chapman & Hall.

Castellano, K. E., & McCaffrey, D. F. (2017). The accuracy of aggregate student growth percentiles as indicators of educator performance. *Educational Measurement: Issues and Practice*, *36*(1), 14–27. Retrieved from http://dx.doi.org/10.1111/emip.12144

Chernozhukov, V., Fernndez-Val, I., & Galichon, A. (2010). Quantile and probability curves without crossing. *Econometrica*, *78*(3), 1093–1125. Wiley Online Library.

Cook, J. R., & Stefanski, L. A. (1994). Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical Association*, *89*, 1314–1328.

Dette, H., & Volgushev, S. (2008). Non-crossing non-parametric estimates of quantile curves. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *70*(3), 609–627. Wiley Online Library.

Harrell, F. E. (2001). *Regression modeling strategies*. New York: Springer.

Harris, D. N. (2007). *The policy uses and "policy validity" of value-added and other teacher quality measures*. Princeton, NJ: Educational Testing Service.

Koenker, R. (2005). *Quantile regression*. Cambridge: Cambridge University Press.

Linn, R. L. (2003). *Accountability: Responsibility and reasonable expectations*. Los Angeles, CA: Center for the Study of Evaluation, CRESST.

Linn, R. L., Baker, E. L., & Betebenner, D. W. (2002). Accountability systems: Implications of requirements of the No Child Left Behind Act of 2001. *Educational Researcher*, *31*(6), 3–16.

Lockwood, J. R., & Castellano, K. E. (2015). Alternative statistical frameworks for student growth percentile estimation. *Statistics and Public Policy*, *2*(1), 1–9. Retrieved from http://dx.doi.org/10.1080/2330443X.2014.962718

Lord, F. M. (1975). The "ability" scale in item characteristic curve theory. *Psychometrika*, *20*, 299–326.

McCaffrey, D. F., Castellano, K. E., & Lockwood, J. R. (2015). The impact of measurement error on the accuracy of individual and aggregate sgp. *Educational Measurement: Issues and*

Practice, *34*(1), 15–21. Retrieved from http://dx.doi.org/10.1111/emip.12062

R Development Core Team. (2019). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org

Raudenbush, S. W. (2004). What are value-added models estimating and what does this imply for statistical practice? *Journal of Educational and Behavioral Statistics*, *29*(1), 121–129.

Rubin, D. B., Stuart, E. A., & Zanutto, E. L. (2004). A potential outcomes view of value-added assessment in education. *Journal of Educational and Behavioral Statistics*, *29*(1), 103–116.

Shang, Y., VanIwaarden, A., & Betebenner, D. W. (2015). Covariate measurement error correction for student growth percentiles using the simex method. *Educational Measurement: Issues and Practice*, *34*(1), 4–14. Retrieved from http://dx.doi.org/10.1111/emip.12058

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis*. New York: Oxford University Press.

Spellings, M. (2005). Secretary spellings announces growth model pilot. *Secretary Spellings Announces Growth Model Pilot*. Press Release, U.S. Department of Education.

Stefanski, L., & Cook, J. (1995). Simulation-extrapolation: The measurement error jackknife. *Journal of the American Statistical Association*, *90*(432), 1247–1256. Taylor & Francis Group.

Wei, Y., & He, X. (2006). Conditional growth charts. *The Annals of Statistics*, *34*(5), 2069–2097.

Yen, W. M. (1986). The choice of scale for educational measurement: An IRT perspective. *Journal of Educational Measurement*, *23*, 299–325.

Yen, W. M. (2007). Vertical scaling and No Child Left Behind. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 273–283). New York: Springer.