Correcting for Measurement Error in Student Growth Percentiles:

The SIMEX correction method implementation in the SGP software.

Adam R. VanIwaarden

Damian W. Betebenner

National Council on Measurement in Education (NCME) Annual Meeting

April 15, 2018

1 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 currently available for SGP calculation in the SGP package (Betebenner, VanIwaarden, Domingue, & Shang, 2017) for the R statistical program. Currently several states utilize the SIMEX corrected measures in their student growth modeling and evaluation policies.

However, the SIMEX corrected SGPs also have 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). Because of this "bias-variance trade-off" the SIMEX correction is preferred for aggregated SGP results, such as the Mean SGP (MSGP) values typically used for teacher or school evaluation purposes, but the standard SGP is the preferred student level estimate. 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[1]. The progress, problems and further insights from the implementation of the SIMEX ranking are the primary focus of this report. However, before proceeding, it is useful to review the process used to compute the SIMEX measurement error (ME) adjusted SGPs.

[1] SGP versions 1.7-0.0 and later

2 The SIMEX Method of calculating Student Growth Percentiles

This brief methodological review is meant to outline the SIMEX process in general and highlight the areas in which additional efficiency might be gained. For a detailed review of the SIMEX measurement error correction method applied to SGP analyses, see Shang, VanIwaarden and Betebenner (2015), and for an in-depth treatment of SIMEX in general see Carroll, Ruppert, Stefanski, & Crainiceanu (2006).

2.1 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 (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 are calculated 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, 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 here, but can be found in Shang, VanIwaarden and Betebenner (2015).

In short, the process by which SIMEX corrected SGP estimates are obtained can be outlined as follows:

1. Calculate the "naive" fitted values from the quantile regression coefficient matrices produced from the unperturbed data.
2. For each value of λ (0.5, 1, 1.5, 2), repeat the following steps:
    a. Create B alternate data sets that adds measurement error to the observed data using λ and the CSEM values.
    b. For each of the B alternate data sets, create a SIMEX coefficient matrix.
    c. Use the B alternate coefficient matrices to produce B predicted score lookup-tables (a predicted score is produced for each percentile, so that when all tables are stacked on top of each other the result is a single table of predicted scores is produced with, generally, a dimension of N x B rows and 100 percentile columns)
    d. Average the predicted scores for each percentile over the B to create a N rows by 100 percentile columns lookup-table for each value of λ.
3. Use a least squares regression model to extrapolate over the increasingly error prone estimates (increasing values of λ) back to the point at which $ME = 0$ (i.e. the predicted value for the extrapolation model when $\lambda = -1$).
4. Use the SIMEX corrected predicted score table to establish the percentile estimation for each students' observed score. For example, a student may have an observed score of 750, which falls between their 50^{th} and 51^{st} percentile predicted scores based on her prior scores, and so her SIMEX SGP is estimated as 50. This step applies the same process used to establish the "naive" SGP, and in this example the students' observed score would likely be situated slightly differently in the uncorrected predicted score table giving her a different SGP (say 47 for this example).

Steps 2 through 3 are repeated for each number of prior scores available. For each student, the SIMEX SGP with the maximum number of priors is selected as the final SIMEX SGP estimate.

A random sample of students may be selected to create the matrices in step 2b (a unique sample for each simulation iteration). This is particularly useful when analyzing large student cohorts. Although this adds in and additional source of sampling variation and uncertainty, large states/consortia have shown the use of the entire cohort data to be computationally prohibitive. In exploring the trade off between cohort sample size and other SIMEX parameters, it was found that increasing the number of simulation iterations, B, provides a good balance between reducing the additional sampling error and computational feasibility. The current defaults in the `SGP` package reflect this, as discussed in the next section.

2.2 SIMEX implementation in the SGP package

The `SGP` package (Betebenner et al., 2017) allows the user to specify any 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`, where assessment specific meta-data is housed in the `SGP` package and identifies which variables in the data to use in the SIMEX process. Alternatively, one can use the following elements to identify the necessary components:
  - `variable` - the variable in the data to be perturbed (typically test scores, e.g. "SCALE_SCORE" in `SGP` vernacular).
  - `csem.data.vnames` - the CSEM variable ("SCALE_SCORE_CSEM" in `SGP` vernacular)
  - `csem.loss.hoss` - a list of the Lowest/Highest Obtainable Scall Scores (L/HOSS) for each grade/content area to be included in the analysis. This will be computed internally if not supplied here.
- `lambda` identifies the desired values of λ.
- `simulation.iterations` identifies the desired number of iterations, B.
- `extrapolation` allows the user to select a "linear" or "quadratic" extrapolant function.

The user may also request optional functionality, including

- `simex.sample.size` allows one to use a sample subet of the data in the production of the coefficient matrices through quantile regression[2],
- `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.

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

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 can be used to calculate baseline referenced SGPs.

Internally, the `studentGrowthPercentiles` function first uses an uncorrected coefficient matrix 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.[3] The function then uses each coefficient matrix to calculate fitted value predictions at each percentile value. Thus at this stage we have a table of predicted values that is $B \times N$ columns (where N is number of students) and 100 columns wide for each percentile.

These predicted values are then averaged for each student 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 values form a lookup table that is N rows and 100 percenile columns wide. The original observed scores are then used to produce SGP values in the typical manner. That is, each students observed score is compared to all 100 of their 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.

3 Ranking SIMEX SGPs in the SGP Package

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 most situations.

Without a continuous value, the percentile ranking[4] 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 allow for a more continuous SIMEX SGP to be established. The authors suggested that more granular SIMEX SGPs be established in the simulations (personal communication), however this would require the already computationally and time intensive process to take **at least** 10 times longer. Furthermore, previously calculated (unranked) SIMEX values would no longer be reproducible.

A simpler solution was used that allows the estimated SIMEX values to be placed on a $1/8^{th}$

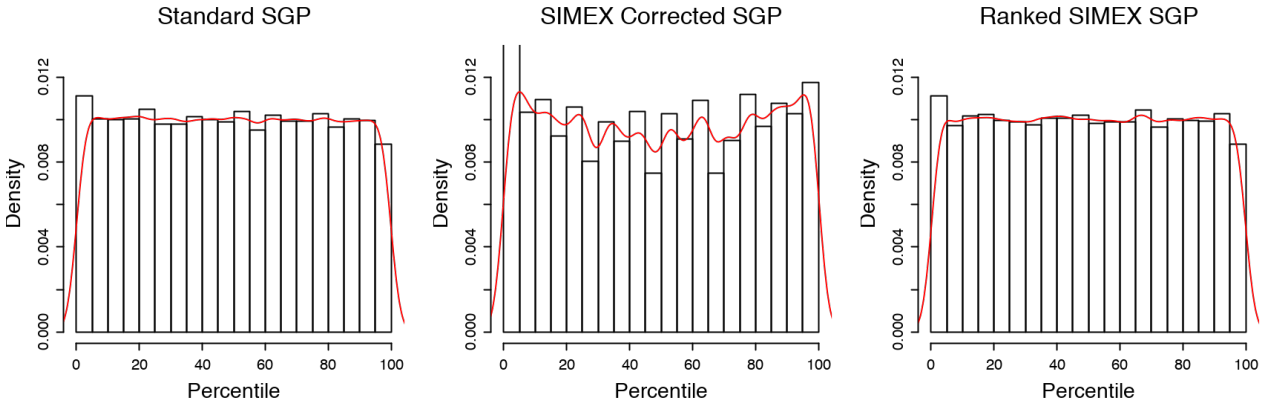
[3] This includes producing the knots and boundaries used in the quantile regressions.

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

interval by calculating arithmetic midpoints between each percentile's predicted score values[5]. This resulted in RS-SGP values that were more uniformly distributed in initial tests with real and simulated data.

Figure 1 shows the distribution of the three types of SGP estimates for 2017 8^{th} Grade Mathematics SGP analyses for an example state: uncorrected ("standard"), SIMEX corrected and Ranked SIMEX. Note that these results are from analyses that use two years of data (a single prior and the current year). Adding additional prior years data will also greatly reduce ME bias.

Figure 1: 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. 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).

[5]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.

These plots display typical distributions of each SGP variant from the same 2017 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 2: Goodness of Fit Plot for 2017 ***Standard*** 8^{th} Grade Mathematics SGPs.

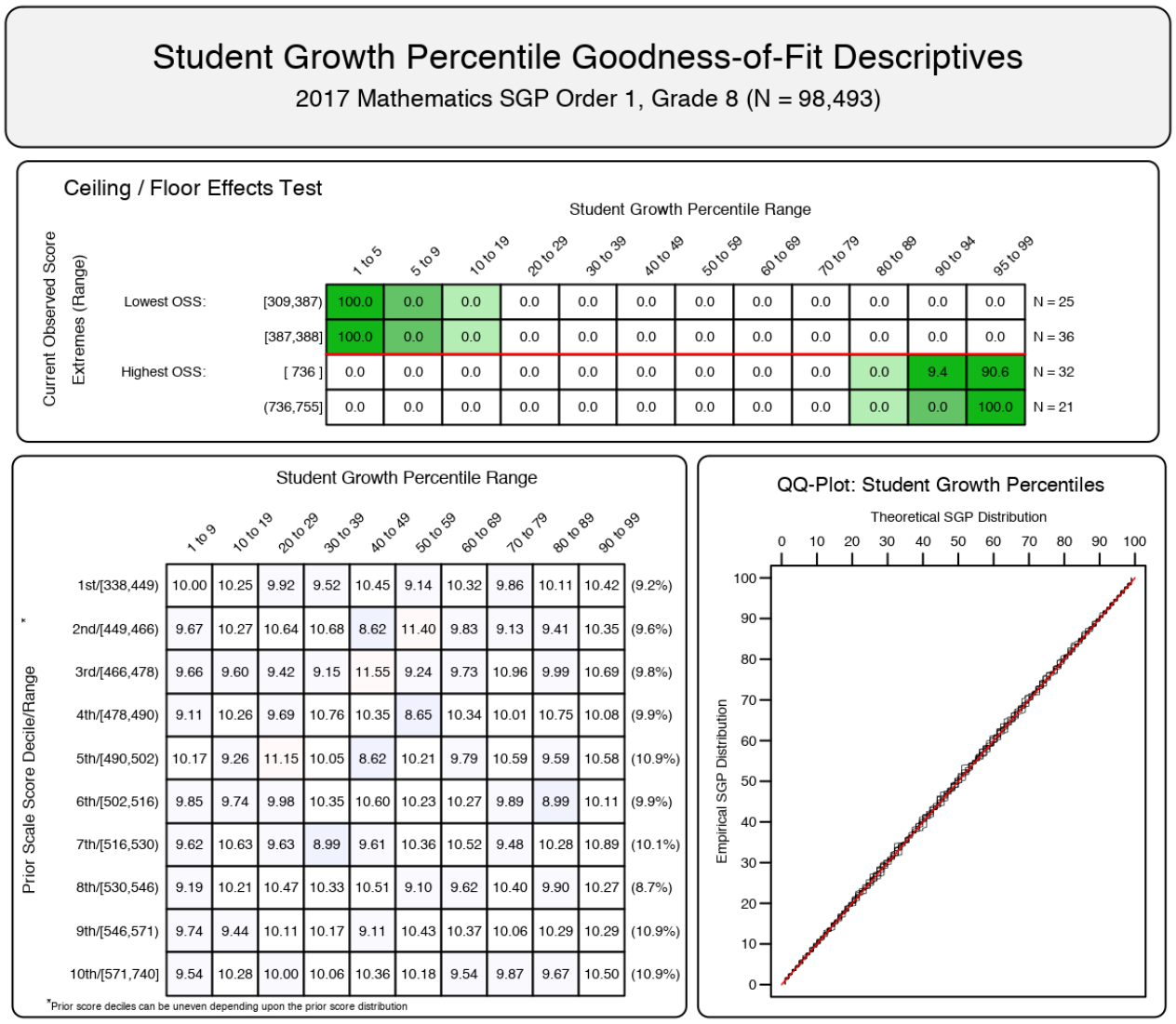


Figure 3: Goodness of Fit Plot for 2017 ***SIMEX*** 8^{th} Grade Mathematics SGPs.

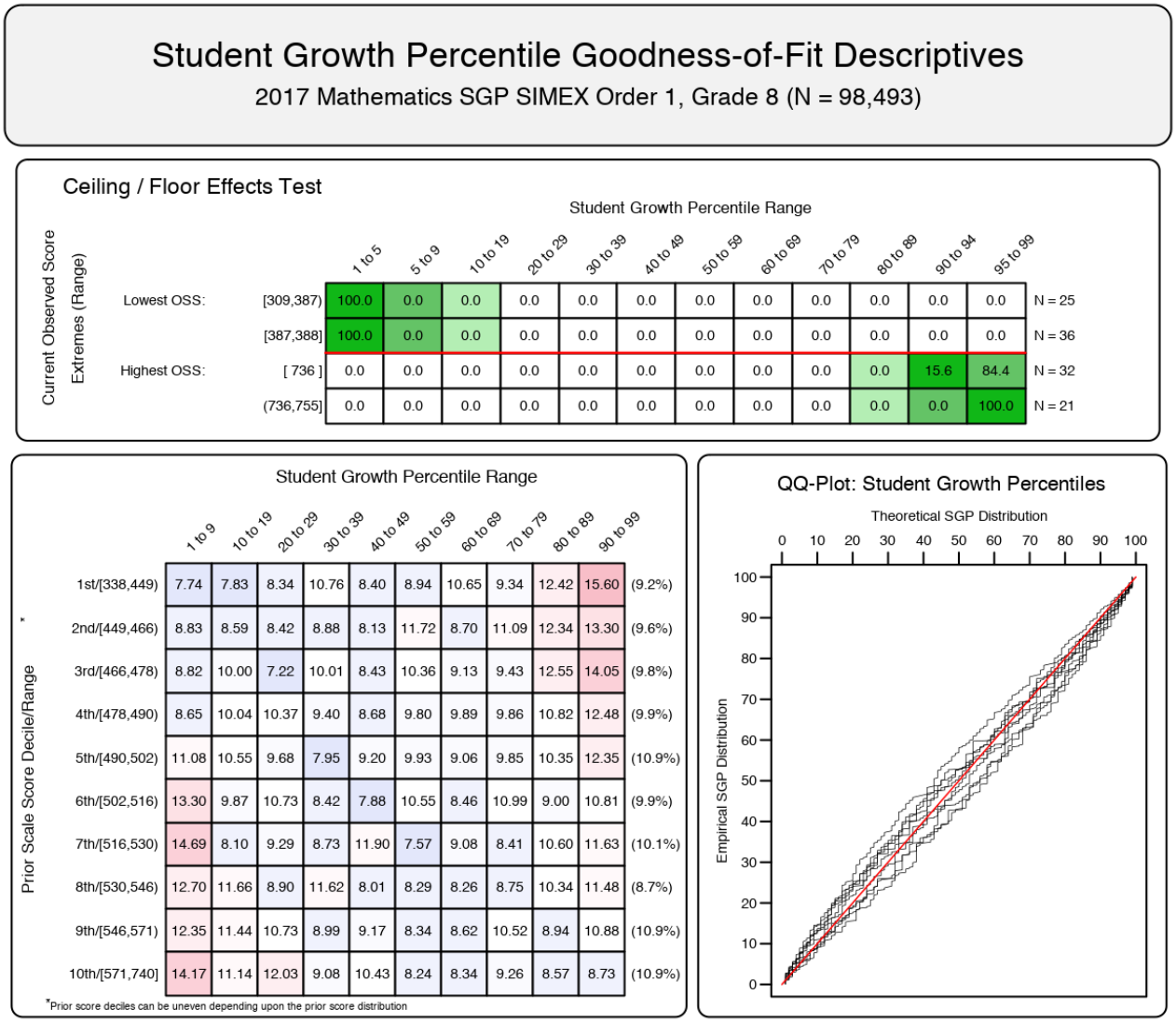
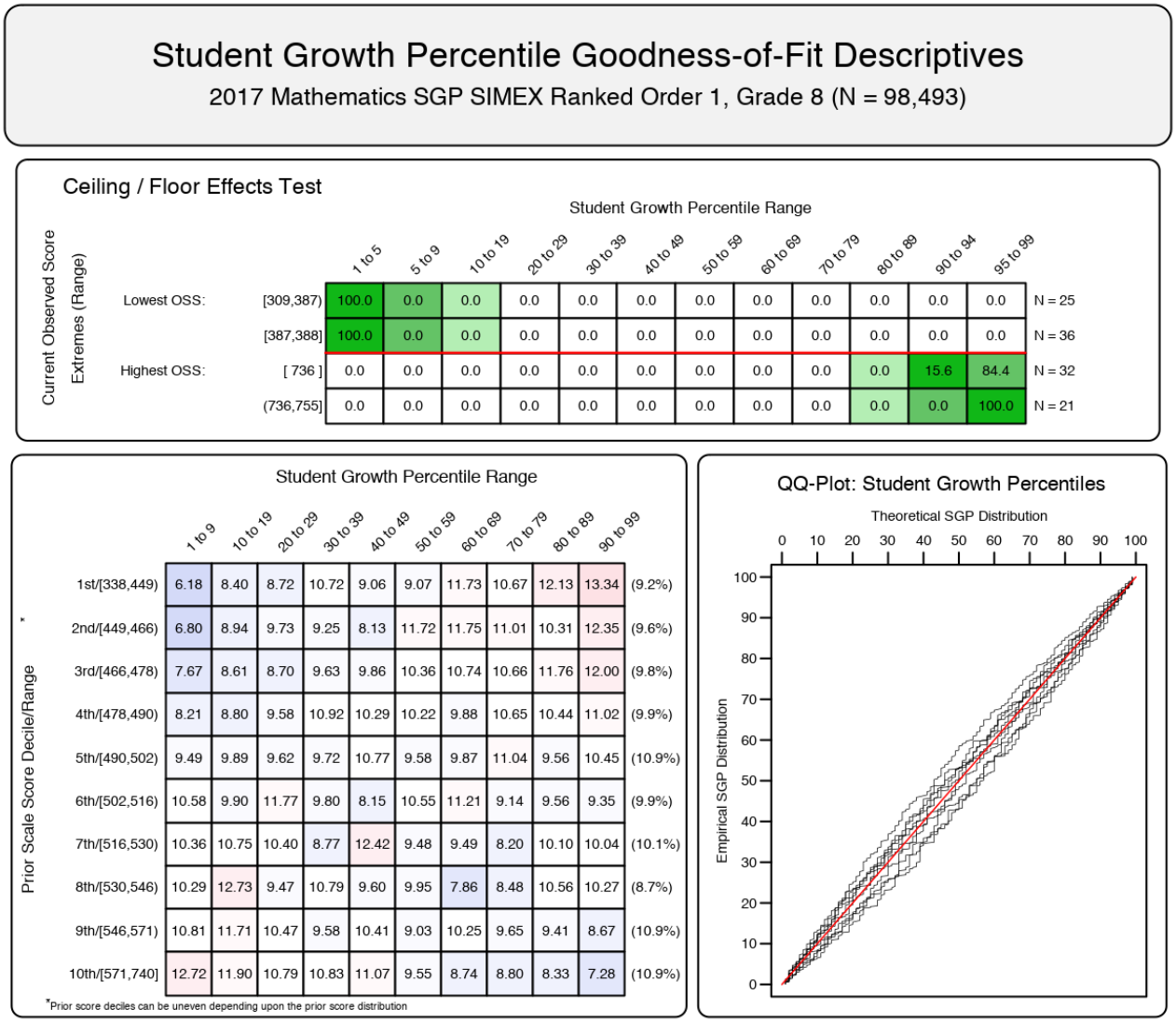


Figure 4: Goodness of Fit Plot for 2017 ***Ranked SIMEX*** 8^{th} Grade Mathematics SGPs.



4 Relationship of Ranked SIMEX with Prior Student Achievement

An important consequence of the typical SIMEX and RS-SGP conditional distributions is that a negative correlation is created between them and prior test scores. This is true at the student level and also translates to school and teacher aggregations. These negative relationships are indicative of the reduction in ME induced bias. The following tables show the results for the English/Language Arts and Mathematics 2017 example analyses. As can be seen, the SIMEX and RS-SGP correlations are nearly identical in both settings. This is unsurprising as the maximum differences between the two values are between -3 and 3 with near-zero averages for all grade-by-subject specific analyses (see subsequent sections for more details on the observed differences between SIMEX and RS-SGP).

Table 1: Student Level Correlations between Prior Standardized Scale Score and 1) Current Scale Score, 2) SGP, 3) SIMEX SGP and 4) Ranked SIMEX SGP.

Content Area	Grade	$r_{TestScores}$	r_{SGP}	$r_{SIMEXSGP}$	$r_{RankedSIMEX}$
Language Arts	4	0.830	0.000	-0.121	-0.122
	5	0.843	0.000	-0.090	-0.090
	6	0.839	0.000	-0.086	-0.086
	7	0.839	-0.001	-0.086	-0.086
	8	0.846	0.001	-0.087	-0.088
Mathematics	4	0.843	-0.001	-0.096	-0.096
	5	0.856	-0.001	-0.071	-0.071
	6	0.842	-0.001	-0.068	-0.069
	7	0.866	0.000	-0.073	-0.073
	8	0.819	0.003	-0.064	-0.064

4.1 Schools

It is critical to also consider the impact the SIMEX ranking has on aggregated SGPs since they are used for school and teacher accountability in several states. The following tables looks at the example state's analyses over the past several years at the school level.

Table 2: 2016 and 2017 School Level Correlations between Mean Prior Standardized Scale Score and Aggregate SGPs by Content Area.

Content Area	Year	R Mean SGP	R Mean SIMEX	R Mean Ranked SIMEX
Language Arts	2016	0.583	0.425	0.425
	2017	0.519	0.330	0.330
Mathematics	2016	0.512	0.413	0.413
	2017	0.417	0.303	0.303

Table 3: 2017 School Level Correlations between Mean Prior Standardized Scale Score and Aggregate SGPs by Content Area and Grade.

Content Area	Grade	R Mean SGP	R Mean SIMEX	R Mean Ranked SIMEX
Language Arts	4	0.432	0.236	0.236
	5	0.283	0.119	0.119
	6	0.273	0.117	0.117
	7	0.445	0.314	0.314
	8	0.415	0.272	0.272
Mathematics	4	0.268	0.150	0.150
	5	0.199	0.109	0.109
	6	0.203	0.128	0.128
	7	0.359	0.251	0.251
	8	0.297	0.219	0.219

References

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., VanIwaarden, A., Domingue, B., & Shang, Y. (2017). *SGP: Student growth percentiles & percentile growth trajectories.* Retrieved from sgp.io

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.

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

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

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

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