Title: A Framework for Improving Student Growth Percentiles by Accounting for Test Score Measurement Error

Authors and Affiliations: J.R. Lockwood, Katherine E. Castellano & Daniel F. McCaffrey, Educational Testing Service
**Background**

Student growth percentiles (SGPs) are currently being considered or used in more than half of the states in the country to gauge the academic progress of both individual students and groups of students (Betebenner, 2008; Castellano & Ho, 2013; Shang, 2012). The widespread interest in SGPs is in part due to their conceptual simplicity: each student’s current test score is expressed as a percentile rank in the distribution of current test scores among students who had the same past test scores (typically from the previous school year). For example, a student with SGP of 70 is meant to convey that the student scored higher on the current test than 70% of his or her peers who had scored similarly in the past. SGPs are commonly interpreted as a measure of relative growth or conditional status (Castellano & Ho, 2013) in achievement, and because of the percentile rank scaling, SGPs are applicable even to test score scales that are not vertically aligned.

SGPs are currently calculated in terms of observed test scores in a target population of students (e.g., all students at a given grade level in a state). However, test scores are error-prone measures of latent achievement (Lord, 1980). Although SGPs are formally only a descriptive statistic, in practice, they are interpreted as relative growth in achievement and thus it is likely that inference desired by consumers of SGPs is more closely aligned with latent achievement than with observed test scores.

Due to the nature of the complex calculations and nonlinear functions involved to translate observed test scores into SGPs (Shang, 2012), it is generally the case that a student’s observed SGP is a biased estimator of his/her SGP defined in terms of latent achievement, even if each test score used in the calculation is an unbiased estimate of the latent achievement construct it is intended to measure. In fact, even if the SGP could be calculated using the latent values of past achievement, measurement error in the current year score alone would be sufficient to cause bias in the SGP due to the nonlinear functions used to calculate SGPs. For example, if true achievement and measurement errors were normally distributed, SGPs on observed scores would be biased toward 50.

The potential for bias in inferences from SGPs stemming from test measurement error may be exacerbated when median or mean SGPs are calculated for schools or individual teachers. Because students in different schools and different teachers’ classes are likely to differ with respect to latent achievement, errors in student SGPs caused by test score measurement error that are systematically related to latent achievement will not cancel when aggregated. This can result in systematic errors in mean or median SGPs for teachers or schools that are correlated with the students’ background characteristics. Because the aggregate SGPs are sometimes interpreted as indicators of relative effectiveness of schools or teachers, this can lead to certain teachers or schools being chronically advantaged or disadvantaged by the evaluation metrics depending on the types of students they serve. The potential for such systematic error has been raised and studied in the broader literature on teacher value-added for more than a decade (Harris, 2011; Kane, McCaffrey, Miller, & Staiger, 2013; McCaffrey, Lockwood, Koretz, & Hamilton, 2003; Rothstein, 2010) and is equally applicable to inferences based on SGPs.

Shang (2012) provides an initial attempt to address test score measurement error in SGPs by using Simulation-Extrapolation (SIMEX), a method for providing approximately consistent parameter estimates in nonlinear models (Carroll, Ruppert, Stefanski, & Crainiceanu, 2006; Cook & Stefanski, 1994). However, the methods in that study, which apply SIMEX to the parameters of quantile regression functions that are only an intermediate step in the calculation of SGPs, are insufficient to fully correct for bias induced by measurement error.

**Objective**

In this paper and presentation, we develop a theoretically coherent approach to handling test score measurement error in SGPs. We establish an isomorphism between using item-level data from a single
test to make inferences about latent student achievement and using item-level data from multiple tests to make inferences about latent SGPs. We discuss the main practical problems that need to be solved in order for this approach to be implemented. We recommend solutions to these problems and will demonstrate their potential with test score data from a large school system.

**Relevance to Conference Theme**

Our work is a direct result of a collaboration with practitioners in a state department of education. The state found that school-level median SGPs were correlated greater than 0.4 with school-level average prior achievement, a value that seemed unreasonably high and likely to cause mistrust of the SGP metric among educators. Inflated correlations between the aggregate SGP metric and prior student achievement is a predictable consequence of test score measurement error (Fuller, 2006). We have been working with the state to understand how much of this correlation might due to test measurement error, and to develop strategies for mitigating it. The framework developed here is intended to help the state move forward its use of SGPs, and will help other states and practitioners think about how to best use the available data to improve SGP estimation and reporting.

**Statistical Framework**

For clarity we focus on the bivariate case where there is a current test score and a single prior test score, although the framework we discuss is equally applicable to conditioning on multiple past test scores. We follow the notation of Carroll et al. (2006) by using \( X \) to denote latent variables that are measured with error by \( W \). Specifically, we use \( X_0 \) to denote the latent prior achievement and \( X_1 \) to denote the latent current achievement. We are defining \((X_0, X_1)\) as hypothetical scores that would be realized by letting the number of test items get large for the same tests administered at the same time to the same students as the actual tests. In this sense the only source of measurement error we are addressing is that caused by the finite number of test items and the imperfect information about latent achievement provided by each item.

We let \((W_0, W_1)\) be the observed test scores providing error-prone measures of \((X_0, X_1)\). Some of the methods we discuss make reference to individual item-level data which we denote \((I_0, I_1)\) where \(I_t\) is the vector of item responses to questions on the test from time \(t\). We also define our percentile ranks on the \((0, 1)\) scale rather than the \((0, 100)\) scale.

**The Standard SGP Definition**

The standard SGP is defined as

\[
\rho(w_1, w_0) := F_{W_1|W_0}(w_1|w_0) = \int_{-\infty}^{w_1} p_{W_1|W_0}(u|w_0)du
\]

(1)

where \( p_{W_1|W_0} \) is the conditional density of \( W_1 \) given \( W_0 \) and \( F_{W_1|W_0} \) is the corresponding CDF. These distributions and the resulting SGPs are defined with respect to the error-prone scores in some target population of students. In practice \( \rho(w_1, w_0) \) is estimated indirectly through estimated quantile functions for \( W_1 \) given \( W_0 \) which are parameterized using B-splines to allow for flexibility in the conditional distributions (Betebenner, 2008).

**The Latent SGP Definition**

We begin by assuming that the target SGP estimand involves the latent scores and their distributions rather than the error-prone scores and their distributions. We thus define the latent SGP as

\[
\pi(x_1, x_0) := F_{X_1|X_0}(x_1|x_0) = \int_{-\infty}^{x_1} p_{X_1|X_0}(u|x_0)du.
\]

(2)
Comparing Equation 2 to 1 shows that the definitions are parallel but one is defined in terms of latent scores and distributions and the other in terms of observed scores and distributions. As noted previously, it is generally the case that \( E[\rho(W_1, W_0)|X_1 = x_1, X_0 = x_0] \neq \pi(x_1, x_0) \) so that the standard SGP is a biased estimator of the latent SGP.

Under the assumption that \( \pi(x_1, x_0) \) is the desired inference from SGPs, the challenge is to develop an estimator for it given the observed item response data \((I_0, I_1)\) in a target population of students. This estimation problem parallels that faced when having item-level data on a single test and wanting to use those data to provide an estimate of each student’s latent achievement. However, the SGP case presents two specific challenges: 1) the target parameter \( \pi(x_1, x_0) \) involves two latent achievement values; and 2) the target parameter is a complex nonlinear function of those latent achievement values that is defined only in reference to the distribution of latent achievement in some target population, which itself must be estimated from the observed data. These problems have been studied in the literature in contexts other than SGPs, and we next discuss how those solutions could be applied to the SGP problem.

**Solving Estimation Challenges**

1. **Estimating Latent Score Distributions**

   The first challenge to estimating \( \pi(x_1, x_0) \) is that it requires estimating the conditional distributions of the latent scores in the target population from the observed item response data. These conditional distributions would follow directly from an estimate of the joint distribution \( p_{X_0,X_1}(x_0, x_1) \) in the target population. This is the well-studied problem of deconvolution of a latent distribution given error-prone measures (Carroll et al., 2006; Laird, 1978; Lockwood & McCaffrey, 2014; Mislevy, 1984; Rabe-Hesketh, Pickles, & Skrondal, 2003; Roeder, Carroll, & Lindsay, 1996). Deconvolution methods use the fact that the item response theory model (IRT; van der Linden & Hambleton, 1996) defines the probability structure of the observed item responses conditional on the latent scores, and therefore the observed item response distributions can be used to infer the latent score distributions.

   We propose that states could use the longitudinal item response data on individual students to estimate bivariate IRT models (Reckase, 2009) that would provide an estimate of \( p_{X_0,X_1}(x_0, x_1) \) in the target population. Standard software available for MIRT (e.g. flexMIRT, Cai, 2012) could be used to perform the estimation under different assumptions about \( p_{X_0,X_1}(x_0, x_1) \). For deconvolution, these assumptions range from highly parametric to nonparametric and the key question for the SGP application is how much parametric structure is required to estimate \( p_{X_0,X_1}(x_0, x_1) \) in such a way that the conditional distributions are reasonably smooth and well-identified from the available data. Fully nonparametric approaches to deconvolution (Rabe-Hesketh et al., 2003; Roeder et al., 1996) tend to result in distributions that are not smooth and therefore it is likely that parametric assumptions will be required in practice to achieve the desired smoothness. For example, it is possible that the standard MIRT assumption of bivariate normality of \((X_0, X_1)\) would fit the data sufficiently well. Parametric extensions of the multivariate normal family that allow for skewness are also available (Azzalini & Dalla Valle, 1996) and could be tested against the standard normality assumption, as could mixtures of normal distributions.

2. **Estimating Individual SGPs**

   Given an estimate \( \hat{p}_{X_0,X_1}(x_0, x_1) \), there is still the problem of how to estimate the SGP for each student given his or her item responses on the two tests. This problem is isomorphic to how item responses from a single test are used to estimate the latent achievement for that test. For that standard case, there are many different estimators available, including maximum likelihood estimation (MLE),
expected a posteriori (EAP), maximum a posteriori (MAP), among others. Each of these has analogs for SGP, and different advantages and disadvantages that would have to be evaluated in the specific context of SGPs just as they have been studied for estimating latent achievement from a single test. We discuss some alternative estimation approaches here. We present the estimators conditional on \( \hat{p}_{X_0,X_1}(x_0, x_1) \); we will address how uncertainty in this estimated distribution could be handled in practice.

The first estimator of \( \pi(x_1, x_0) \) from Equation 2 uses the fact that if \( W_t \) is a maximum likelihood estimator (MLE) of \( x_t \) for an individual student then \( \hat{\pi}_{MLE}(W_1, W_0) = \pi(W_1, W_0) \) is the MLE for \( \pi(x_1, x_0) \). This is distinctly different from the standard SGP in Equation 1 because the error-prone scores are being plugged into the \( \pi \) function defined in terms of the latent achievement. Another possible estimator is the unbiased estimator \( \hat{\pi}_U(I_1, I_0) \) for a function \( \hat{\pi}_U \) satisfying the conditional expectation constraint \( E[\hat{\pi}_U(I_1, I_0)|X_1 = x_1, X_0 = x_0] = \pi(x_1, x_0) \). This estimator would have the property that students with a given latent SGP \( \pi(x_1, x_0) \) would, on average over the distribution of item responses, receive an estimated SGP \( \hat{\pi}_U(I_1, I_0) \) whose mean was \( \pi(x_1, x_0) \). We can closely approximate the function \( \hat{\pi}_U \) by solving an integral equation, and Monte Carlo and Fourier approaches for this are possible (Carroll et al., 2006; McCaffrey, Lockwood, & Setodji, 2013). Finally, EAP and MAP estimators can be calculated analogously to the single test case because the empirical Bayes posterior distribution of any function of \( (X_0, X_1) \) given \((I_0, I_1)\) is available from Bayes’ rule. For example, the EAP SGP estimator would be \( \hat{\pi}_{EAP}(I_1, I_0) = E[\pi(X_1, X_0)|I_1, I_0] \). This estimator would not be unbiased in the sense of \( \hat{\pi}_U(I_1, I_0) \), but would be calibrated in the sense that students with a given value of \( \hat{\pi}_{EAP}(I_1, I_0) \) would, on average over their corresponding distribution of latent scores, have latent SGP equal to \( \hat{\pi}_{EAP}(I_1, I_0) \).

Each of these estimators has a corresponding measure of uncertainty that we can calculate using standard methods. In addition, each of these could be aggregated to teacher or school levels to obtain estimates of group-level SGPs, and associated standard error calculations could be developed under different assumptions about the sources of uncertainty in the SGP estimators. The impact of uncertainty in the estimated latent population distribution and its conditional distributions, as well as in the item parameters from the MIRT model, would be most straightforwardly addressed in a fully Bayesian analysis where the latent distribution was parameterized or otherwise given a prior distribution along with the item parameters. In this case, the posterior distribution of the SGP would simultaneously account for uncertainty about the population distribution of latent achievement, item parameters, and each student’s latent achievement.

**Conclusions and Practical Significance**

Due to their intuitive appeal and minimal reliance on scale assumptions, SGPs are positioned to grow as part of education research, practice and reporting. Test score measurement error poses particular challenges for SGP estimation that go beyond the problems caused by measurement error in a single test score. We demonstrate a clear parallel between how item response data from a single test is analyzed to estimate latent achievement for each student on that test, and how item response data from multiple tests can be used to estimate latent SGPs for each student in a target population. States have access to all of the required item-level data and standard methods exist for carrying out the required steps to go from longitudinal item-level data to estimators for latent SGPs. Our presentation will describe the framework and report on initial efforts to conduct estimation using longitudinal item-level data from a large school system.
Appendix A: References


