Rethinking Student Growth Percentile Estimation

J.R. Lockwood, Katherine E. Castellano, Daniel F. McCaffrey

SREE Fall Research Conference
September 5, 2014

This material is based on work supported by Educational Testing Service and the U.S. Department of Education Institute of Education Sciences through Grant R305D140032. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of these organizations.

Corresponding author: jrlockwood@ets.org
Overview

- Student Growth Percentiles (SGP) are being used in about half of US states to monitor student progress and evaluate educators.

- Statistical methods for SGP computation mostly have been taken at face value.
  - R package “SGP” for implementing computations being used by states and researchers.
  - We call this the “standard SGP approach”.

- We suggest two alternative statistical frameworks that may be better suited to the SGP estimation problem.
  - Talk summarizes paper to appear among group of articles on SGP forthcoming in *Statistics and Public Policy*. 
Student Growth Percentiles: The Concept

- SGP is the percentile rank of a student’s current score in the conditional distribution of current scores among students sharing the same past test score(s).
  - E.g., SGP of “70” is intended to mean student scored higher than 70% of students in a reference population who had scored similarly in the past.

- Ideally equal to the conditional CDF of the current score given past score(s), evaluated at the current score.

- But ties in the scores complicate formal definitions and estimation of the SGP.
Some Advantages of SGP

- Intuitively appealing
- Easy to explain to non-technical audiences
- Minimal reliance on test scaling assumptions
  - Requires only ordinal scale rather than interval scale
- Transparent aggregation of student-level SGP to teacher, school or other hierarchical unit by taking mean or median, called “MGP”
  - Being used in some states for formal teacher or school leader evaluation
- Available to implement in R package “SGP”
The Standard SGP Estimation Approach

1. Choose percentile points $\tau$, typically from 0.005 to 0.995 in increments of 0.01 (100 total)

2. For each $\tau$, separately fit quantile regression of current score as smooth function of past score(s)
   - E.g. $\tau = 0.105$ models the 10.5th percentile of current score
   - Uses B-splines for regression on past scores to avoid linearity and implicit interval scaling assumptions

3. Pointwise re-sort 100 estimated quantile functions so that they are monotone non-decreasing for each student

4. Assign SGP by figuring out which pair of percentile cuts the outcome score lies between
Student Growth Percentile Model Demonstration

Grade 4 Score vs. Grade 5 Score

Copyright ©2014 by Educational Testing Service. (6)
Student Growth Percentile Model Demonstration

![Graph showing the relationship between Grade 4 and Grade 5 scores.](image)

Copyright ©2014 by Educational Testing Service.
Student Growth Percentile Model Demonstration

![Graph showing the relationship between Grade 4 and Grade 5 scores. The x-axis represents Grade 4 score ranging from 200 to 450, and the y-axis represents Grade 5 score ranging from 200 to 500. The graph has several data points at (200, 200), (250, 250), (300, 300), (350, 350), (400, 400), and (450, 450).]
Student Growth Percentile Model Demonstration

Grade 4 Score vs. Grade 5 Score

Copyright ©2014 by Educational Testing Service.
Student Growth Percentile Model Demonstration
Student Growth Percentile Model Demonstration

[Graph showing the relationship between Grade 4 Score and Grade 5 Score.]
Student Growth Percentile Model Demonstration
Student Growth Percentile Model Demonstration
Student Growth Percentile Model Demonstration
Student Growth Percentile Model Demonstration
Some Issues with Standard SGP Approach

1. SGP are basically conditional CDFs but standard approach models quantile functions (inverse CDFs) instead
   - Requires ad hocery to get monotonicity of percentile cuts
   - Requires manual inversion of estimated functions to create estimated SGP
   - Discreteness of observed scores causes theoretical problems for quantile function estimation and difficulties in treating ties consistently across students
   - Standard SGP estimator has algorithmic definition, so hard to assess statistical properties (e.g. standard errors)

2. Modeling with observed scores and their distributions causes bias in estimated SGP from measurement error in both current and prior scores
Two Suggestions for Other Ways to Estimate SGP

1. Because conditional CDFs are the target, model conditional CDFs directly rather than indirectly through collection of quantile functions

2. Multidimensional Item Response Theory (MIRT) models using longitudinal item-level data can in principle get around most of the problems of standard approach, notably bias due to test measurement error
Conditional CDF Modeling

- Model $P(Y \leq y | X)$ directly
- Long history in statistics (e.g. survival analysis and Cox regression)
- Standard proportional hazards assumptions may not hold for test scores because test score distributions do not look anything like exponential-type distributions
- Whole line of more recent work on estimating conditional CDFs that require minimal distributional or functional form assumptions (e.g. Hall et al. 1999, JASA)
  - Most of these methods use nonparametric kernel smoothing estimators of the conditional CDF and result in estimated functions that satisfy the required monotonicity constraints
  - Standard errors or MSE are available for some estimators
Conditional CDF Modeling Deals Better With Discreteness

- Observed test scores have discrete support
- Quantile regression will generally be modeling some quantiles that are not well defined
  - E.g. if $Y$ has support $y_1, \ldots, y_K$ then $F_{Y|X}(y|x)$ is a step function with steps at $y_1, \ldots, y_K$ for each $x$
  - The step sizes will generally depend on $x$
  - Some quantiles will be unresolveable for some $x$ because it will not align with the steps
- Modeling conditional CDF for discrete ordinal outcome gets around this problem
  - E.g. ordered logit models, loglinear models
An Empirical Example to Demonstrate Possible Benefits of Conditional CDF Modeling

- Simple example of prior and current year mathematics scores from three large urban school districts
- Modeled current score as ordered categorical using logit model with flexible spline parameterization in prior score
  - Has order of magnitude fewer parameters than family of quantile regression models, but was better able to recover the empirical conditional CDFs
- Demonstrated that ties can matter for SGP and MGP
  - Modeling conditional CDFs makes it easy to define SGP using standard percentile rank estimator which gives students exactly half credit for students they tie
  - Standard approach less consistent in its treatment of ties
Prior Score = 890 (n = 354)

Cumulative Probability

Current Score

Prior Score = 890 (n = 354)

Standard KS = 0.076

Logit KS = 0.021
MGP Differences By Prior Achievement

Teacher Average Prior Score

- MGP Difference
- Standard MGP – Empirical MGP
- Logit MGP – Empirical MGP
Conditional CDF Modeling Does Not Solve Everything

- Standard SGP approach, and direct CDF modeling, both still treat the observed test scores and their distributions as the quantities of interest.

- But presumably stakeholders care about relative growth or conditional status in achievement, not test scores.

- Measurement error in both current and past test scores bias SGP estimators for “latent SGP” defined in terms of either true scores or latent proficiencies from Item Response Theory (IRT) models.
Bias Caused by Test Score Measurement Error

- Prior scores lead to bias because of standard errors-in-variables logic
  - If this was the sole source of bias, it would cause students with high true prior achievement to have positively biased SGP

- Current scores lead to bias because of nonlinearity of conditional CDFs or quantile functions
  - Compression toward median

- Actual bias depends on combination of both sources and can go in either direction

- Bias in SGP will not cancel when aggregating to MGP for teachers, schools, etc.
SGP Estimation from Item-Level Data Can Solve These Problems

- Define latent SGP in terms of latent achievement attributes and their distributions in a reference population
- Assume current test and single prior test for simplicity
- Prior test measures $\theta_0$ and current test measures $\theta_1$

\[
\pi(\theta_1, \theta_0) := F_{\Theta_1|\Theta_0}(\theta_1|\theta_0) = \int_{-\infty}^{\theta_1} p_{\Theta_1|\Theta_0}(u|\theta_0)du
\]

- Longitudinal item-level data can be used to estimate $p_{\Theta_1,\Theta_0}$ in the population and individual students’ item-level data can be used to estimate their value of $\pi(\theta_1, \theta_0)$
Practical Challenges: Latent Distribution

- Spirit of SGP is to make minimal distributional assumptions, but parametric assumptions are probably required to estimate $p_{\Theta_1, \Theta_0}$ using MIRT.
- Multivariate normal assumption might be sufficient in practice.
- Relaxations to try include mixtures of normals, multivariate skew normal, or latent class specifications.
Practical Challenges: Student SGP Estimator

- Isomorphism between using items from one assessment to make inferences about student achievement, and using items from multiple assessments to make inferences about student SGP

- Different estimators, different tradeoffs
MLE SGP Estimator

- MLE conditional on estimated item parameters and population distribution, themselves estimated from marginal maximum likelihood (MML)

- Approximately unbiased for large number of students and items

- Standard error available

- Existence problems for extreme response patterns (e.g. all correct or all incorrect)
  - Adaptation of Warm’s Weighted Likelihood may be useful here
Empirical Bayes SGP Estimators

- EAP (posterior mean) or MAP (posterior mode) available from MML
- Always exist, intuitive interpretation
- Empirical Bayes posterior standard deviation serves as measure of uncertainty
- But biased from shrinkage which might be viewed unfavorably for policy or accountability purposes
- Shrinkage bias will carry over to MGP and compress perceived variability among teachers
Fully Bayesian SGP Estimators

- Fully Bayesian model of item-level data, item parameters, latent achievement distribution
- Posterior distribution of each student’s SGP accounts for uncertainty due to limited item responses, unknown item parameters, unknown latent distribution
- May be overkill when estimating for an entire state
Summary

- Stakeholders evidently really like SGP and MGP
- It may be possible to improve upon the standard approach
- Conditional CDF modeling easier to try first because it uses the same scale score data as the current approach
- But item-level modeling seems to make the most efficient, coherent use of the available data and tackles test measurement error head-on
  - Practical issues include latent achievement distribution specification and choice of SGP estimator
Limited Reliability of SGP Probably Transcends Estimation Methods

- Latent achievement attributes tend to be strongly correlated within student over time
- Even if we observed $\theta_0$ without error, the conditional reliability of $S_1$ for $\theta_1$ given $\theta_0$ will tend to be very low
  - In bivariate normal case, the conditional reliability is the same as the reliability would be if we were measuring $\theta_1$ with error variance $\sigma^2/(1 - r^2)$ rather than $\sigma^2$, where $r$ is $\text{cor}(\theta_0, \theta_1)$
- SGP for individual students likely to be dominated by measurement error no matter what we do
- MGP will be more reliable, but if the sole goal is measures of relative educator performance, there are probably better approaches (e.g. value-added methods)