Title:
Examining the Reliability of Student Growth Percentiles using Multidimensional IRT

Authors:
Scott Monroe
Li Cai

Journal publication date:
2015

IES grant information:
Grant number R305D140046
Funded by National Center for Education Research (NCER)

Examining the Reliability of Student Growth Percentiles using Multidimensional IRT

Scott Monroe
University of Massachusetts, Amherst

Li Cai
University of California, Los Angeles

Part of this research is supported by an Institute of Education Sciences statistical methodology grant (R305D140046). The views expressed here belong to the authors and do not reflect views or policies of the funding agency.

Address all correspondence to: Scott Monroe, 111 Thatcher Rd., UMass, Amherst, MA, USA 01003. Email: smonroe@educ.umass.edu. Phone: 413.545.8909.

Examining the Reliability of Student Growth Percentiles using Multidimensional IRT

Abstract

Student Growth Percentiles (SGP, Betebenner, 2009) are used to locate a student's current score in a conditional distribution based on the student's past scores. Currently, following Betebenner (2009), quantile regression is most often used operationally to estimate the SGPs. Alternatively, multidimensional item response theory (MIRT) may also be used to estimate SGPs, as proposed by Lockwood and Castellano (2015). A benefit of using MIRT to estimate SGPs is that techniques and methods already developed for MIRT may readily be applied to the specific context of SGP estimation and inference. This research adopts a MIRT framework to explore the reliability of SGPs. More specifically, we propose a straightforward method for estimating SGP reliability. Additionally, we use this measure to study how SGP reliability is affected by two key factors: the correlation between prior and current latent achievement scores, and the number of prior years included in the SGP analysis. These issues are primarily explored via simulated data. Additionally, the quantile regression and MIRT approaches are compared in an empirical application.

Keywords: Student Growth Percentiles, Item Response Theory, High-Stakes Testing, Teacher Evaluation

1 Introduction

Numerous states use the Student Growth Percentile (SGP, Betebenner, 2009) methodology to make inferences about student academic progress. An SGP locates a student's current achievement score in a conditional distribution dependent on the student's prior achievement scores. In this way, an SGP provides context for the current achievement. Some states also aggregate SGPs (e.g., using a mean) for the purposes of teacher evaluation. The original methodological framework for SGPs is quantile regression (QR), and an R package (Betebenner, VanIwaarden, Domingue, & Shang, 2014) has been developed in support of the methodology. Within this framework, which has been the focus of several recent research efforts (e.g., Castellano & Ho, 2013; Shang, VanIwaarden, & Betebenner, 2015; McCaffrey, Castellano, & Lockwood, 2015), SGPs are calculated in multiple steps. First, student scores are generated for each year's test. Second, based on the observed scores, QR is used to obtain conditional quantiles. Optionally, a bias correction is applied to the conditional quantile estimates (Shang et al., 2015). Finally, the quantiles and observed scores are used to estimate SGPs.

Given that SGPs may be used for high-stakes decisions, such as teacher evaluation, it is important that the statistical properties of the estimates are well understood. The present research focuses primarily on the reliability of SGP estimates. Generally, research has shown that SGP estimates have low levels of reliability at the student level. Examples of research reaching this conclusion include Wells, Sireci, and Bahry (2014), Shang et al. (2015), and McCaffrey et al. (2015). In all of the cited research, true SGPs are used to determine that estimates produced via the QR framework have large amounts of random error. However, some questions remain. For instance, what is responsible for the low reliability? And, are there realistic conditions, as yet unconsidered, where the reliability attains an acceptable level? Finally, can reliability be estimated without true SGPs, available only in a simulation study? Answers to these questions will not only offer methodological insights but are also relevant to policy discussions.

In this research, we explore these questions. First, we propose a straightforward method for estimating marginal reliability that does not depend on true SGPs. An advantage of this measure is that it is familiar and easily interpretable. Then, using simulated data examples, we study how reliability is affected by two key features of the SGP analysis: the correlations among the latent achievement scores, and the number of prior years included in the analysis.

Instead of adopting the QR framework for SGPs, we use a multidimensional item response theory (MIRT) framework, as advocated by Lockwood and Castellano (2015), among others. This latter approach is appealing because MIRT is a relatively flexible modeling framework and the focus of much ongoing research. Consequently, techniques and tools already developed for MIRT may readily be applied to the specific context of SGP estimation and inference.

For instance, in the present research, a standard error of the SGP estimate is needed for the proposed reliability measure. In the MIRT framework, standard errors for SGPs are readily available, as established methods used to estimate latent traits and their standard errors may easily be extended to SGPs (Lockwood & Castellano, 2015). In contrast, in the QR framework, defining and computing a standard error appear more involved. Ideally, the standard error should account for the uncertainty in each of the multiple steps (i.e., calibration and linking of the instruments, scaled score computations, quantile regression, etc.). We will demonstrate that integrating SGP estimation into MIRT provides straightforward methods for studying the uncertainty of the resulting estimates by leveraging existing knowledge in latent variable modeling.

Since, in practice, SGPs are mostly calculated using QR, an important question concerns the relevance of results obtained from studying MIRT-based SGPs. Generally speaking, due to the fact that MIRT-based SGPs (in particular, the EAP estimator to be discussed later) make use of all available information from the item response data set across multiple years, the MIRT-based reliability may be considered a best-case sce-

nario, to which alternative methods tend to approximate. Thus, the proposed method does provide useful information about the statistical properties of QR-based SGPs. Also, we believe and show that patterns found in the reliability results hold across modeling frameworks, thus contributing to the understanding of when individual SGPs will be most and least useful.

The remainder of this article is organized as follows. First, SGPs are defined, and their calculation illustrated with graphical examples. Next, the QR and MIRT frameworks for estimating SGPs are introduced and compared, and the proposed method for calculating marginal reliability for SGPs is presented. Then, using the MIRT framework, simulated data examples are provided to explore what factors drive the SGP reliability. This is followed by an empirical data example where both the QR and MIRT approaches are used. Finally, there is a discussion and potential directions for future research are presented.

2 Details and Definitions

2.1 Student Growth Percentiles

As observed test scores contain measurement error, observed growth likewise contains measurement error. Thus, we assume that instead of observed growth, the proper focus of inquiry is latent, or "true", growth. Let θ_c be the current latent achievement of a student, and $\boldsymbol{\theta}_p$ be an $m \times 1$ vector of past latent achievement scores, where m is the number of prior years to be included. Then, let $\boldsymbol{\theta} = (\theta_c, \boldsymbol{\theta}_p')'$, and let $g(\boldsymbol{\theta})$ be the distribution for the latent achievement scores. To simplify the presentation, we assume that each latent variable has a mean of zero and variance of one, so that $\text{Var}(\boldsymbol{\theta}) = \boldsymbol{\Sigma}$ is a correlation matrix. Following Lockwood and Castellano (2015), the SGP is defined as

$$S(\theta_c, \boldsymbol{\theta}_p) = \int_{-\infty}^{\theta_c} p(t|\boldsymbol{\theta}_p)dt, \tag{1}$$

where the integrand $p(\theta_c|\boldsymbol{\theta}_p)$ is the conditional distribution of the current score given prior scores. That is, $S(\theta_c, \boldsymbol{\theta}_p)$ is a conditional cumulative distribution function (CDF).[1] Generally, $p(\theta_c|\boldsymbol{\theta}_p)$ depends on the form of $g(\boldsymbol{\theta})$, as well as the latent (i.e., unobserved) $\boldsymbol{\theta}$. But, regardless of the form of $g(\boldsymbol{\theta})$, S will be uniformly distributed over random samples of $\boldsymbol{\theta}$. Note that as defined in Equation (1), S is on a scale of 0 to 1. For reporting, this value would be multiplied by 100.

As an example of the SGP definition, assume that $g(\boldsymbol{\theta})$ is multivariate normal. Then,

$$\begin{pmatrix} \theta_c \\ \boldsymbol{\theta}_p \end{pmatrix} \sim \mathcal{N}_{1+m} \left(\begin{pmatrix} 0 \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} 1 & \boldsymbol{\sigma}'_{pc} \\ \boldsymbol{\sigma}_{pc} & \boldsymbol{\Sigma}_{pp} \end{pmatrix} \right), \tag{2}$$

where $\boldsymbol{\Sigma}_{pp}$ is an $m \times m$ matrix and $\boldsymbol{\sigma}_{pc}$ is an $m \times 1$ vector. By standard normal distribution theory, the conditional density $p(\theta_c|\boldsymbol{\theta}_p)$ is that of univariate normal. The mean of the conditional distribution is

$$E(\theta_c|\boldsymbol{\theta}_p = \mathbf{x}) = \boldsymbol{\sigma}'_{pc}\boldsymbol{\Sigma}_{pp}^{-1}\mathbf{x}, \tag{3}$$

and the conditional variance is

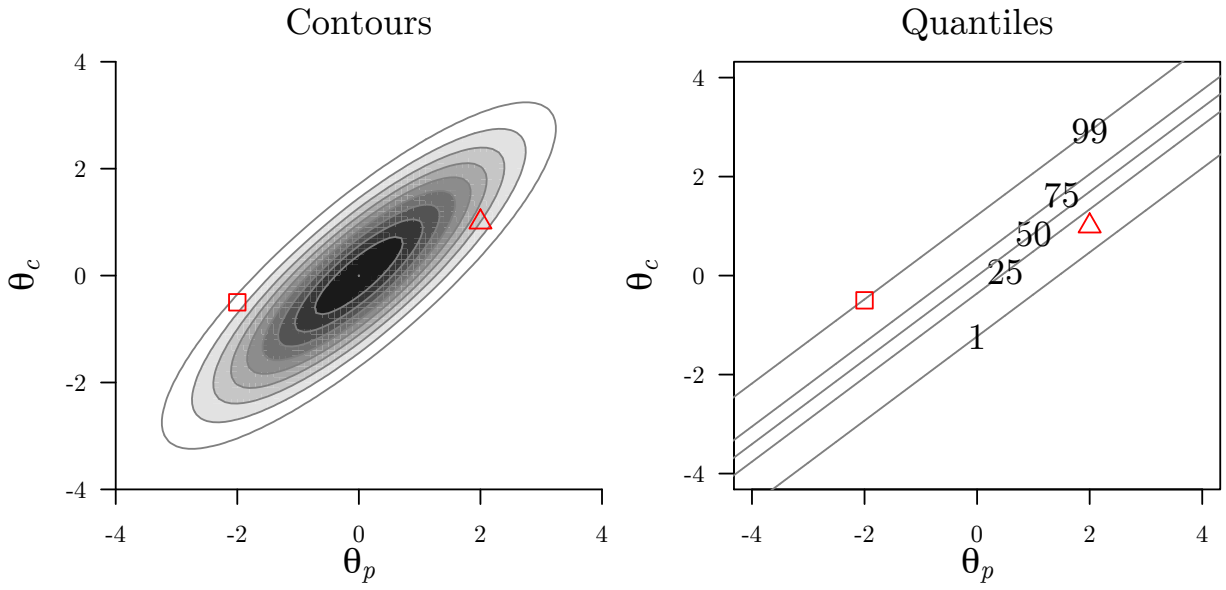
$$\text{Var}(\theta_c|\boldsymbol{\theta}_p = \mathbf{x}) = \text{Var}(\theta_c|\boldsymbol{\theta}_p) = 1 - \boldsymbol{\sigma}'_{pc}\boldsymbol{\Sigma}_{pp}^{-1}\boldsymbol{\sigma}_{pc}, \tag{4}$$

where $\mathbf{x}$ is a realized value of the random variable $\boldsymbol{\theta}_p$. So, by Equation (1), S would be found by evaluating the normal CDF defined by Equations (3) and (4). Continuing with this example, let $m = 1$ (i.e., one prior year is included) so that $g(\boldsymbol{\theta})$ is bivariate normal. Further, let the correlation r be equal to 0.85, a value that is representative of correlations we have observed in analyses of operational state summative assessments. Then, $g(\boldsymbol{\theta})$ is fully defined and shown as a contour plot on the left of Figure 1. On the right of Figure 1, corresponding conditional percentiles are shown. In both plots, the triangle and

[1] As noted in the literature (e.g., Betebenner, 2009; Castellano & Ho, 2013), and made clear by Equation (1), SGPs are conditional status measures, and, as such, should not be interpreted as magnitudes of growth.

square symbols represent true latent achievement values, with corresponding $S \approx .10$ and $S \approx .99$, respectively.

Figure 1: Graphic Representation of SGP Definition: Normal Distribution



Note. Distribution is a standard bivariate normal with correlation $r = .85$. The triangle and square symbols represent two points in the bivariate space with $S \approx .10$ and $S \approx .99$, respectively.

2.2 Quantile Regression Framework

Before introducing the MIRT framework, we briefly describe the QR approach to estimating SGPs. The QR approach has many implementation details that we will not review. Interested readers are referred to Betebenner (2009) and Shang et al. (2015). Let $\hat{\theta}_c$ and $\hat{\boldsymbol{\theta}}_p$ be estimates of the current and prior achievements, respectively. Often, each year's achievement estimate is obtained via independent unidimensional IRT analyses. Then, QR is used to estimate a large number of conditional quantiles via $\hat{\boldsymbol{\beta}}$, a vector of QR coefficient estimates. For example, 100 quantiles, ranging from .005 to .995, may be estimated. For each student, conditional on $\hat{\boldsymbol{\theta}}_p$, the observed current score, $\hat{\theta}_c$, is compared to these quantiles. $\hat{\theta}_c$ will be positioned between two estimated quantiles (e.g., .665 and .675). The mean of these two quantiles (.670) is the estimate of S.

Measurement error in the observed scores, however, implies that $\hat{\boldsymbol{\beta}}$ will be biased (Shang, 2012). That is, the quantiles estimated from the observed scores, without further adjustment, will be biased estimates of the quantiles of $g(\boldsymbol{\theta})$. This bias then spills over into estimates of S. To address this issue, the SIMEX algorithm (Shang, 2012; Shang et al., 2015) has been proposed as an additional step for the QR approach. Application of the SIMEX method can be viewed as an effort to obtain unbiased estimates of the quantiles of $g(\boldsymbol{\theta})$ and, by extension, S as defined in Equation (1).

Ideally, standard errors for QR-based estimates of S should account for the uncertainty in all steps of the overall procedure. That is, the standard errors should represent the uncertainty in the estimates of the observed achievement scores, the QR coefficients, and, if applicable, the estimated parameters of the SIMEX method. However, fully accounting for the uncertainty may be computationally demanding, particularly if the SIMEX step is included (see Carroll, Küchenhoff, Lombard, & Stefanski, 1996). Further, each of the steps has a number of implementation details, which makes it challenging to find a sufficiently general approach.

2.3 MIRT Framework

Provided item-level data are available to the researcher, the MIRT framework can naturally accommodate SGP estimation (Lockwood & Castellano, 2015). This is because MIRT presupposes a latent distribution, which in the current context is $g(\boldsymbol{\theta})$, the distribution of the latent achievement scores. In this section, a MIRT model that facilitates the estimation of S is outlined. Then scoring for $\boldsymbol{\theta}$ and estimation of S are both presented, as the latter may be considered an extension of the former (Lockwood & Castellano, 2015). Some equations supporting the presentation in this section may be found in the Appendix.

For SGP estimation, the MIRT model is specified as a correlated-traits between-item MIRT model (e.g., Reckase, 2009). In this model, each item loads on one and only one latent variable, and the latent dimensions may be correlated. As a simple example,

again consider $m = 1$ prior year. In this case, the item responses from the current year depend only on θ_c, and the responses from the prior year depend only on θ_p. In the factor analysis literature, this pattern of loadings is referred to as the *independent cluster* pattern.

After the MIRT model is specified, the unknown parameters of the model are estimated in a step known as calibration. Let $\boldsymbol{\gamma}$ collect together the free parameters of the MIRT model. So, $\boldsymbol{\gamma}$ typically includes free parameters of the item response models (e.g., intercepts and slopes), but may also include free parameters of the model for $g(\boldsymbol{\theta})$. Importantly, for SGP estimation, the correlations of $g(\boldsymbol{\theta})$ are free parameters to be estimated. Typically $g(\boldsymbol{\theta})$ is specified as multivariate normal. However, recent research has also investigated specifying $g(\boldsymbol{\theta})$ as a more flexible distribution (Monroe, 2014). Use of this distribution for MIRT-based SGPs has been explored in Monroe, Cai, and Choi (2014). In this latter case, additional free parameters of $g(\boldsymbol{\theta})$ are estimated from the data.

Calibration of the MIRT model results in estimates $\hat{\boldsymbol{\gamma}}$ of $\boldsymbol{\gamma}$ based on a calibration sample. Then, estimates of individual achievement scores $\boldsymbol{\theta}$ may be produced using various estimators, such as Maximum Likelihood (ML) or *Expected A Posteriori* (EAP) scoring. In this research, we consider EAP scoring, as it is the minimum mean squared error estimator of the latent variables $\boldsymbol{\theta}$ (Bock & Mislevy, 1982). For the ith examinee, let the EAP estimates be $\text{EAP}(\boldsymbol{\theta}_i)$ and the corresponding standard errors be $\text{SE}(\boldsymbol{\theta}_i)$. Both $\text{EAP}(\boldsymbol{\theta}_i)$ and $\text{SE}(\boldsymbol{\theta}_i)$ are functions of the posterior distribution of $\boldsymbol{\theta}_i$, given the examinee's observed item responses. Conceptually, $\text{EAP}(\boldsymbol{\theta}_i)$ averages the latent achievement over the uncertainty in estimating $\boldsymbol{\theta}_i$ as characterized by the posterior distribution.

As with $\boldsymbol{\theta}$, estimates of S may be produced using various estimators, and we again opt for the EAP estimator in this research. For the ith examinee, let the EAP estimate be $\text{EAP}(S_i)$ and the corresponding standard error be $\text{SE}(S_i)$. As with $\text{EAP}(\boldsymbol{\theta}_i)$, $\text{EAP}(S_i)$ is found by averaging over the posterior distribution. However, while $\text{EAP}(\boldsymbol{\theta}_i)$ averages the latent achievement over the posterior distribution, $\text{EAP}(S_i)$ averages the (latent)

definition of S in Equation (1) over the posterior distribution. Expressions for the EAP estimators and standard errors are provided in the Appendix.

2.4 Proposed Reliability Index

The reliability of the SGP estimate, for either MIRT or QR modeling frameworks, can be calculated in a straightforward manner. The proposed index is analogous to the *marginal reliability* index used to describe test precision in a unidimensional IRT framework, suggested by Green, Bock, Humphreys, Linn, and Reckase (1984). Thus, it is instructive to review marginal reliability for IRT before presenting the proposed measure for SGP estimates.

For IRT, the reliability coefficient may be written as

$$\rho_\theta = 1 - \frac{\sigma_e^2(\theta)}{\sigma_\theta^2}, \tag{5}$$

where σ_θ^2 is the prior value of the variance of θ, and $\sigma_e^2(\theta)$ is the marginal or average error variance of θ.[2] Often, σ_θ^2 is fixed to one for purposes of model identification, and the right-hand side of Equation (5) simplifies to $1 - \sigma_e^2(\theta)$. Thus, all that is required to compute the marginal reliability of the test is an estimate of the average error variance.

In IRT, the magnitude of error variance depends on the level of the latent trait. The conditional error variance may be averaged, however, using one of two methods. The first approach uses expected error variance. This expectation may be calculated by integrating the conditional standard error of measurement function over the latent variable distribution of θ. The conditional standard errors, in turn, depend solely on the expected test information function (see, e.g., Thissen & Orlando, 2001). Alternatively, $\sigma_e^2(\theta)$ may be calculated as an average over a random sample of individuals from the population

[2]Marginal reliability for IRT is comparable to, but distinct from, reliability as defined in classical test theory (e.g., Lord & Novick, 1968). For comparisons between the two measures, interested readers are referred to Green et al. (1984), Sireci, Thissen, and Wainer (1991), and the references therein.

distribution,

$$\overline{\mathrm{SE}}^2(\theta) = \frac{1}{N}\sum_{i=1}^{N}\mathrm{SE}^2(\theta_i), \tag{6}$$

where $\mathrm{SE}^2(\theta_i)$ is the squared standard error for the ith examinee. In other words, given a large random sample from the examinee population, and the availability of standard errors of individual θ estimates, the empirical average in Equation (6) provides a consistent estimate of $\sigma_e^2(\theta)$ by law of large numbers.

The proposed method of calculating reliability for estimates of S is completely analogous. Let the SGP reliability coefficient be

$$\rho_S = 1 - \frac{\sigma_e^2(S)}{\sigma_S^2}, \tag{7}$$

which parallels the construction of Equation (5). By definition, S is distributed as a uniform(0,1) random variable. Since the variance of a standard uniform is $1/12$, this value is used for σ_S^2, and the right-hand side of Equation (7) simplifies to $1 - 12\sigma_e^2(S)$. For $\sigma_e^2(S)$, we may use the empirical average

$$\overline{\mathrm{SE}}^2(S) = \frac{1}{N}\sum_{i=1}^{N}\mathrm{SE}^2(S_i), \tag{8}$$

which is analogous to Equation (6). With this approach, individual standard errors for SGP estimates are needed to calculate ρ_S. Within a MIRT framework, using the EAP estimator, these standard errors may are described in Section 2.3 (and defined in the Appendix). This approach is equally applicable to the QR framework, assuming the availability of reasonably accurate standard errors for the individual SGP estimates (see Section 2.2).

2.5 Factors That Affect SGP Reliability

With the foregoing development, we can explore why the reliability of individual SGPs tends to be low in practice. From another perspective, made clear by Equation

(7), we are interested in why $\sigma_e^2(S)$ tends to be large. One possibility is that estimation uncertainty for $\boldsymbol{\theta}$ tends to be too large. As estimation uncertainty for $\boldsymbol{\theta}$ is reduced, $\sigma_e^2(S)$ will decrease, and ρ_S will increase. In the most extreme case, the tests are perfectly reliable, $\sigma_e^2(S) = 0$, and $\rho_S = 1$. While this relationship is true, it is unsatisfying as an explanation of empirically observed low reliabilities. After all, the marginal reliabilities of the annual summative tests measuring achievement are typically high, usually in the 0.9 range. Hence, the estimation uncertainty for $\boldsymbol{\theta}$ is typically small.

A second possibility is that larger correlations among the latent achievement variables lead to lower reliability of SGPs. Though this idea has been previously presented (McCaffrey et al., 2015), it has not been examined with respect to the SGP definition (i.e., Equation 1). To facilitate the presentation of why the correlations might be so important, we make a few simplifying assumptions. We temporarily assume that the analysis is based on the current and immediate past year's achievement data, so that $m = 1$ and there is a single correlation, r. Also, we assume that $g(\boldsymbol{\theta})$ is bivariate normal.

As r increases, the prior achievement θ_p holds greater predictive power for θ_c. This will be directly reflected in a decrease of $\text{Var}(\theta_c|\theta_p)$, the variance of the conditional distribution $p(\theta_c|\theta_p)$ in Equation (1). When $g(\boldsymbol{\theta})$ is bivariate normal, the variance of the conditional distribution, stated in Equation (4), is completely determined by r: $\text{Var}(\theta_c|\theta_p) = 1 - r^2$.

And, as $\text{Var}(\theta_c|\theta_p)$ decreases, ρ_S will almost certainly decrease. Recall that, in operational settings, the estimation uncertainty for $\boldsymbol{\theta}$ is typically small. As $\text{Var}(\theta_c|\theta_p)$ approaches zero, the small (though non-negligible) estimation uncertainty for $\boldsymbol{\theta}$ will lead to greater and greater estimation uncertainty for S. On the other hand, as $\text{Var}(\theta_c|\theta_p)$ increases, the small estimation uncertainty for $\boldsymbol{\theta}$ should lead to relatively smaller estimation uncertainty for S. Thus, we argue that large correlations across years contribute to small ρ_S, due to the relationship between $\text{Var}(\theta_c|\boldsymbol{\theta}_p)$ and the estimation uncertainty for $\boldsymbol{\theta}$.

There are likely other systematic causes of low SGP reliability. However, given the complex interplay between the definition of S, the form of $g(\boldsymbol{\theta})$, and uncertainty in the estimate of $\boldsymbol{\theta}$, it is challenging to identify these causes. Additionally, given our observations regarding $\text{Var}(\theta_c|\boldsymbol{\theta}_p)$, it is unclear whether including more prior years in the analysis will increase ρ_S.

3 Simulated Data Examples

3.1 Generating Conditions

We use simulated data examples to further study the reliability of SGP estimators, focusing on the correlation of the latent dimensions r, as well as the number of prior years m included in the analysis. To generate latent true scores $\boldsymbol{\theta}$, $N = 10,000$ random vectors were sampled from a 4-dimensional normal distribution (i.e., $m = 3$) with zero means and covariance matrix $\boldsymbol{\Sigma}$.

Two generating covariance structures were considered, an auto-regressive structure (AR), and a compound-symmetric (CS) structure. These structures are defined as

$$\boldsymbol{\Sigma}_{AR} = \begin{pmatrix} 1 & r & r^2 & r^3 \\ r & 1 & r & r^2 \\ r^2 & r & 1 & r \\ r^3 & r^2 & r & 1 \end{pmatrix}, \text{ and } \boldsymbol{\Sigma}_{CS} = \begin{pmatrix} 1 & r & r & r \\ r & 1 & r & r \\ r & r & 1 & r \\ r & r & r & 1 \end{pmatrix},$$

respectively, for four time points. The AR and CS covariance matrices were chosen because of their importance in longitudinal data analysis. Each represents a plausible interpretation of the underlying structural relationship among the latent achievement variables. The AR structure models the first-order dependence (Markovian) nature of individual longitudinal data. The CS structure can be understood as arising from modeling the dependence over time with the inclusion of a common random effect on the intercept. Values of r ranged from 0.05 to 0.95 in increments of 0.05.

For each simulee, three true S values were computed, corresponding to the inclusion

of $m = 1$, 2, or 3 prior years. Generally, for each simulee, the values of S will be different depending on m, and the extent of the difference will depend on the form of $g(\boldsymbol{\theta})$ and its covariance structure. An exception to this rule is discussed in the results section.

Item response data were simulated according to the three-parameter logistic model (Birnbaum, 1968). As described earlier, each item loaded on one and only one latent dimension, corresponding to the testing year. For each year, 50 items were used. The true item parameters were chosen to resemble values encountered in typical large-scale assessment programs. Further, the item slopes were chosen in such a manner that the marginal reliability for each dimension (calculated using test information for the corresponding 50 items) was 0.9. We consider this value to be high, but not unrealistic.

To summarize the conditions, two covariance structures were crossed with 19 different correlations to create 38 datasets. Each of these 38 datasets, however, could be analyzed using $m = 1$, 2, or 3 prior years.

3.2 Estimation and Evaluation Statistics

The true generating models were fit to the data by maximum likelihood, using flexMIRT® (Cai, 2013). For $g(\boldsymbol{\theta})$, all variances were fixed to one, and all correlations were estimated as free parameters. Thus, the data-generating covariance structures were not imposed on the estimated covariance matrix, but rather an unstructured covariance matrix was estimated. No misspecification is introduced (albeit there may be a slight loss in statistical efficiency), but an unstructured covariance matrix is operationally more feasible and realistic, given software capabilities. For $m = 2$ and 3 (i.e., the 3 and 4-dimensional models), the MH-RM algorithm (Cai, 2010a, 2010b) was used in order to improve the computational speed of model fitting. Adopting the MIRT framework, EAP estimates and error variances were produced for both $\boldsymbol{\theta}$ and S.

The list of collected evaluation statistics from the simulation is quite brief. First, the SGP marginal reliability ρ_S was computed. Second, the conditional variance of current year achievement $\text{Var}(\theta_c|\boldsymbol{\theta}_p)$ was collected, to examine its relation to ρ_S. Finally, the mea-

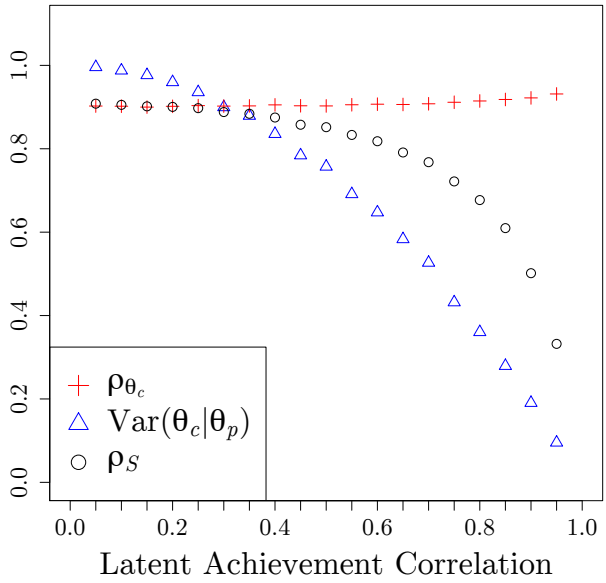
surement error variance for the current year achievement $\sigma_e^2(\theta_c)$ was collected, using the empirical average in Equation (6). Since the true marginal reliability for each dimension is the same, the scalar-valued $\sigma_e^2(\theta_c)$ is representative of the estimation uncertainty for $\boldsymbol{\theta}$. Also, because the total variance of θ_c is fixed to one as part of model identification, $1 - \sigma_e^2(\theta_c)$ provides an empirical estimate of ρ_{θ_c}, the marginal reliability of the current year's test. Given the data-generating process, if only the current year's test items are used to estimate θ_c, then $\sigma_e^2(\theta_c)$ should be around 0.10, and the empirical value of ρ_{θ_c} should be around 0.90. But, if additional years' items are taken into account, as with EAP scoring, then $\sigma_e^2(\theta_c)$ may be less than 0.10, and ρ_{θ_c} may exceed 0.90. This is because the EAP estimator can "borrow strength" from other parts of the model to produce more efficient estimates (Cai, 2010c) by utilizing the latent variable correlations in the MIRT model. The magnitude of the decrease depends on the specific structure of the MIRT model, but generally, it will be greater for larger r and m.

3.3 Results: One Prior Year

For both the $\boldsymbol{\Sigma}_{AR}$ and $\boldsymbol{\Sigma}_{CS}$ conditions, the covariance structure for $m = 1$ is identical. That is, $\boldsymbol{\Sigma}$ reduces to the correlation matrix of a standard bivariate normal variable, with a single correlation to be estimated. Therefore, for $m = 1$, differences between the $\boldsymbol{\Sigma}_{AR}$ and $\boldsymbol{\Sigma}_{CS}$ conditions are solely due to sampling variability (as they are independent conditions under the simulation design). As the sample size is quite large ($N = 10,000$), we observe little effect of sampling variability, so only the results from one condition, $\boldsymbol{\Sigma}_{AR}$, are presented.

Figure 2 displays empirical estimates of ρ_{θ_c} (plus signs), $\text{Var}(\theta_c|\boldsymbol{\theta}_p)$ (triangles), and ρ_S (circles) as r increases from 0.05 to 0.95. While the values of ρ_{θ_c}, the marginal reliability for the current year's test, are relatively stable, they do increase slightly from around 0.90 to 0.93 as r increases. This is simply an example of how the EAP estimator is able to "borrow strength." Solely focusing on ρ_{θ_c}, we would predict that ρ_S would increase with r.

Figure 2: Marginal Reliability of SGP for One Prior Year



Note. ρ_{θ_c} = marginal reliability for current year's test; $\text{Var}(\theta_c|\boldsymbol{\theta}_p)$ = conditional variance of current year achievement; ρ_S = SGP marginal reliability.

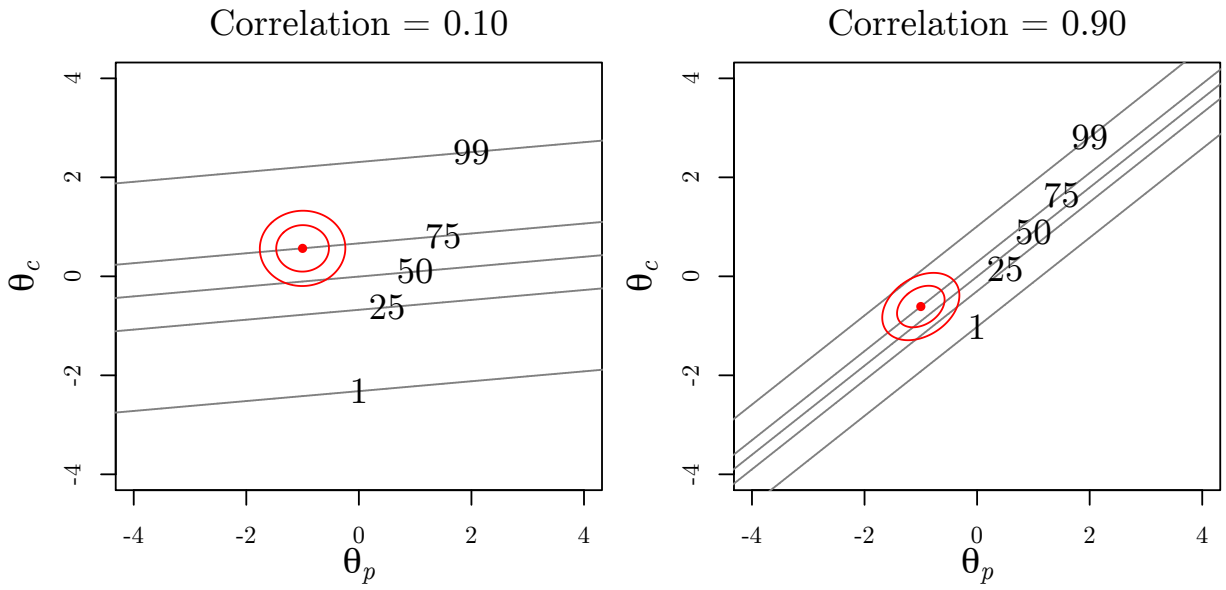
Examining the role of $\text{Var}(\theta_c|\boldsymbol{\theta}_p)$, we see that it decreases rapidly as r increases. In fact, as $g(\boldsymbol{\theta})$ is specified as multivariate normal, the points in Figure 2 correspond exactly to $1 - \hat{r}^2$, where $\hat{r}$ is the maximum likelihood estimate of r. This functional form is responsible for the accelerating change in $\text{Var}(\theta_c|\boldsymbol{\theta}_p)$ (being quadratic in r).

Finally, Figure 2 displays ρ_S, the SGP marginal reliability. Interestingly, the values vary greatly, from around $\rho_S = 0.9$ for small values of r, to $\rho_S = 0.33$ at $r = 0.95$. For small correlations, ρ_S is nearly indistinguishable from ρ_{θ_c}. For moderate correlations, such as 0.5, ρ_S is still quite high (0.85). But, for values of r typical for state achievement tests, such as 0.7 to 0.9, ρ_S decreases quickly, dropping from moderate to low values. It appears that ρ_S is more influenced by $\text{Var}(\theta_c|\boldsymbol{\theta}_p)$ than by ρ_{θ_c}.

To further elucidate the relationship between r and ρ_S, Figure 3 presents conditional percentile plots for $r = 0.10$ (left panel) and $r = 0.90$ (right panel). Each plot also displays representations of average posterior distributions of latent achievement given observed item responses from the simulation. More specifically, the ellipses are based on bivariate

normal approximations of the average posterior covariance matrix across simulees, and demarcate 68% and 95% of the central density. The ellipses are arbitrarily centered on the 75th conditional percentile in both plots with no loss of generality in interpretation.

Figure 3: Effect of Correlation on SGP Uncertainty



Note. For each plot, the ellipses demarcate 68% and 95% of the central density of an average posterior distribution. The locations (i.e., centers) of the ellipses are arbitrary.

For $r = 0.10$, the ellipses are intersected by two of the displayed conditional percentile lines. The corresponding ρ_S value is 0.91. For $r = 0.90$, the ellipses are somewhat more compact than those for $r = 0.10$, reflecting the smaller value of $\sigma_e^2(\theta_c)$. Nevertheless, the ellipses for $r = 0.90$ intersect more of the displayed conditional percentiles. The fact that the conditional percentiles in the $r = 0.90$ plot are closer to one another is a direct consequence of the smaller value of $\text{Var}(\theta_c|\boldsymbol{\theta}_p)$. For $r = 0.90$, the corresponding ρ_S value is 0.50. To summarize the interpretation of Figure 3, the extent to which the ellipses intersect the conditional percentiles provides information about the uncertainty in determining the SGP (as represented by $\sigma_e^2(S)$), and by extension, the marginal reliability ρ_S. In general, more intersections imply lower values of ρ_S.

3.4 Results: Multiple Prior Years

Before presenting results for ρ_S, we briefly discuss how true S values change with the inclusion of additional years in the analysis. In general, for any simulee, including an additional prior year will result in a change in true SGP. That is, for any simulee, true S will be different depending on m. However, it can be shown that for Σ_{AR}, assuming a multivariate normal distribution, S does not depend on m. In this case, S is the same no matter how many prior years are included. This is because of the particular structure of Σ_{AR}. The underlying structural interpretation of a first-order AR model stipulates that dependence is solely modeled by the immediately preceding data point. Therefore, both the mean and variance of the conditional distribution $p(\theta_c|\boldsymbol{\theta}_p)$ remain unchanged as additional prior years are included.

Table 1 presents correlations among various true S values for the AR and CS structures, for typical values of r. For $\boldsymbol{\Sigma}_{AR}$, since S does not vary with m, all of the corresponding correlations are one. On the other hand, for $\boldsymbol{\Sigma}_{CS}$, the correlations between the true S values decrease with increases in m and r. At the most extreme, for $r = 0.9$, the correlation between S based on $m = 1$ and $m = 3$ years is 0.82. Information regarding whether S varies with m could be used to determine the most appropriate number of years to include in an analysis. Also relevant to this determination would be how m affects ρ_S.

Table 2 presents results for ρ_S for $m = 2$ and $m = 3$ prior years. To focus attention on the results most relevant to realistic testing scenarios, results are only reported for r between 0.7 and 0.9. The first set of entries in Table 2, for $m = 1$ prior year, corresponds to points from Figure 2.

Overall, the inclusion of additional prior years has little effect on ρ_S, regardless of the structure of $\boldsymbol{\Sigma}$. For the $\boldsymbol{\Sigma}_{AR}$ conditions, this is entirely predictable given that increasing m has no effect on $\text{Var}(\theta_c|\boldsymbol{\theta}_p)$ at the population level. The differences in $\text{Var}(\theta_c|\boldsymbol{\theta}_p)$ for the $\boldsymbol{\Sigma}_{AR}$ conditions are entirely due to sampling variability. There is a small decrease in

Table 1: Correlations Among Various True SGPs

		Correlation (r)				
Cov. Structure		0.70	0.75	0.80	0.85	0.90
$\boldsymbol{\Sigma}_{AR}$	$\text{cor}(S^{(1)}, S^{(2)})$	1.0	1.0	1.0	1.0	1.0
	$\text{cor}(S^{(1)}, S^{(3)})$	1.0	1.0	1.0	1.0	1.0
$\boldsymbol{\Sigma}_{CS}$	$\text{cor}(S^{(1)}, S^{(2)})$	.90	.89	.89	.88	.87
	$\text{cor}(S^{(1)}, S^{(3)})$	.86	.85	.84	.83	.82

Note. $\boldsymbol{\Sigma}_{AR}$ = auto-regressive covariance structure for $g(\boldsymbol{\theta})$. $\boldsymbol{\Sigma}_{CS}$ = compound-symmetric covariance structure for $g(\boldsymbol{\theta})$. $S^{(m)} = S$ based on m prior years.

Table 2: Marginal Reliability of SGP for Multiple Prior Years

			Correlation (r)				
Cov. Structure	Prior Years (m)		0.70	0.75	0.80	0.85	0.90
$\boldsymbol{\Sigma}$	1	ρ_S	.768	.722	.677	.610	.502
		$\text{Var}(\theta_c \vert \boldsymbol{\theta}_p)$	.527	.432	.360	.279	.191
		$\sigma_e^2(\theta_c)$	.092	.088	.085	.082	.078
$\boldsymbol{\Sigma}_{AR}$	2	ρ_S	.774	.729	.689	.631	.542
		$\text{Var}(\theta_c \vert \boldsymbol{\theta}_p)$	.508	.441	.362	.289	.196
		$\sigma_e^2(\theta_c)$	.091	.088	.085	.081	.077
	3	ρ_S	.774	.729	.692	.634	.549
		$\text{Var}(\theta_c \vert \boldsymbol{\theta}_p)$	.517	.450	.371	.267	.198
		$\sigma_e^2(\theta_c)$	.091	.087	.086	.080	.076
$\boldsymbol{\Sigma}_{CS}$	2	ρ_S	.750	.708	.663	.596	.496
		$\text{Var}(\theta_c \vert \boldsymbol{\theta}_p)$	.420	.365	.295	.225	.148
		$\sigma_e^2(\theta_c)$	.087	.084	.081	.075	.070
	3	ρ_S	.742	.703	.655	.596	.490
		$\text{Var}(\theta_c \vert \boldsymbol{\theta}_p)$	.398	.334	.265	.194	.134
		$\sigma_e^2(\theta_c)$	.085	.082	.078	.073	.066

Note. ρ_S = marginal reliability of SGP. $\text{Var}(\theta_c|\boldsymbol{\theta}_p)$ = conditional variance of θ_c given $\boldsymbol{\theta}_p$. $\sigma_e^2(\theta_c)$ = average error variance for θ_c.

$\sigma_e^2(\theta_c)$ as m increases, again demonstrating how the EAP estimator borrows strength to estimate θ_c more efficiently. And, there is a small corresponding increase in ρ_S.

For the $\boldsymbol{\Sigma}_{CS}$ conditions, Table 2 shows that $\text{Var}(\theta_c|\boldsymbol{\theta}_p)$ actually decreases as m increases. As demonstrated in Section 3.3, a decrease in $\text{Var}(\theta_c|\boldsymbol{\theta}_p)$ can be expected to correspond to a decrease in ρ_S, which is what we observe for the $\boldsymbol{\Sigma}_{CS}$ conditions. So, in the case of the $\boldsymbol{\Sigma}_{CS}$ covariance structure, increasing the number of years in the analysis actually leads to a slight decrease in ρ_S.

4 Empirical Application

As an illustration of the methods discussed in this research, we used MIRT to analyze longitudinal assessment data in order to estimate S and characterize its reliability. Additionally, we used the QR-based approach to produce another set of estimates as a comparison. The item-level data come from a mathematics assessment with 4th and 5th grade data for $N = 10,000$ students in a mid-western state. To be consistent with our notation, we consider the 4th and 5th grade years to be the prior and current years, respectively. The state is not identified for legal reasons.

For each year, 44 dichotomous items were modeled using the three-parameter logistic model. The MIRT approach followed the methods presented in Section 2.3, and EAP estimates were produced for S. For the QR approach, observed scores for the QR were produced in the following way. A unidimensional IRT model was fit to the data for each grade separately, to mimic how the state usually produces the test scores. Within each year, a set of EAP scores and standard errors was produced. These EAP estimates were used as the observed scores for the QR analysis. Additionally, these unidimensional analyses provided marginal reliability values for the 4th and 5th grade tests. These test reliabilities were 0.87 and 0.89, respectively.

Next, the "SGP" package (with default settings) was used to obtain QR-based estimates of S. This analysis also yielded estimates of the conditional quantiles of $g(\boldsymbol{\theta})$. As this research focuses on variability for the SGP estimate, we decided it was unnecessary

to apply the SIMEX method to correct for any bias in the quantile estimates. To obtain standard errors for the QR-based SGP estimates, we used the following imputation-based scheme.

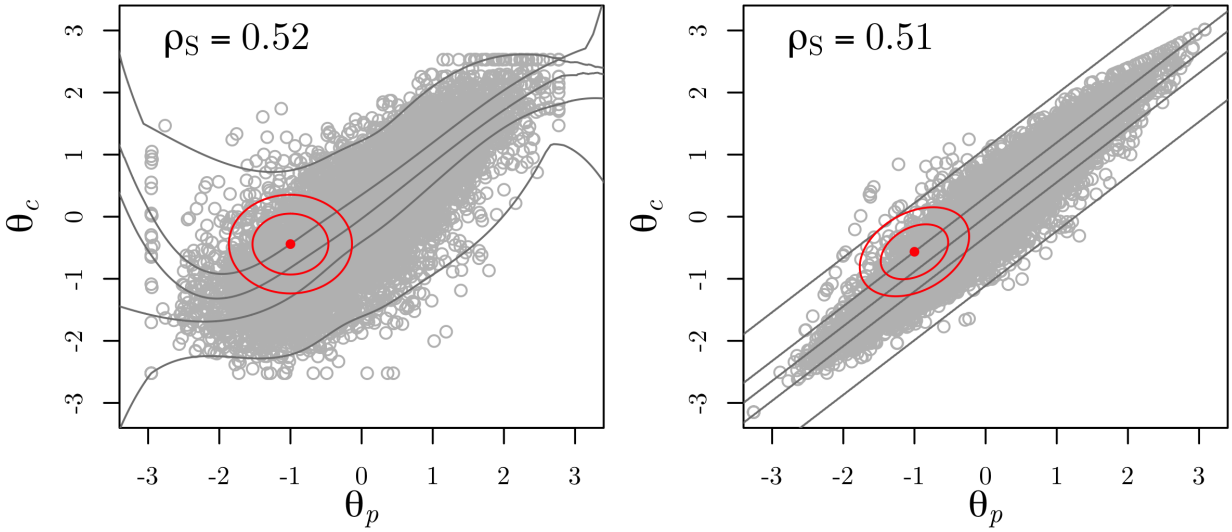
To explain the scheme, we focus on one student. A normal distribution was defined using the student's 4th grade EAP score (i.e., mean) and standard error (i.e., standard deviation). 400 imputations were drawn from this distribution. This process was repeated with the 5th grade EAP score and standard error. Thus, 400 pairs of imputed scores were created. Then, the pairs of imputed scores, along with the original estimated quantiles, were used to create a distribution of SGP estimates. The standard deviation of this distribution was used as a standard error for the QR-based SGP estimate. In turn, the SGP standard errors for all students could be used to calculate the average error variance, as in Equation (8).

One way to compare the QR and MIRT approaches is to evaluate model fit for the respective IRT models. The QR approach, with two unidimensional IRT models, is formally equivalent to a 2-dimensional IRT model where the latent achievement correlation is constrained to zero. In other words, the QR and MIRT approaches specify the same model, except they differ in whether r is constrained to zero or estimated. (Note, however, that with the QR approach, the empirical correlation of scores across years need not be zero.) The MIRT model (with $\hat{r} = 0.88$) is preferred by both $-2\times$log-likelihood and Bayesian information criteria values.[3]

Another obvious point of comparison for the two approaches is ρ_S. For the QR approach, $\rho_S = 0.52$, while for the MIRT approach, $\rho_S = 0.51$. Given $\hat{r} = 0.88$, it is unsurprising that the values are low. What is, perhaps, surprising is that the two approaches produce such similar values. Figure 4, inspired by a similar figure in Betebenner (2009), illustrates some of the differences in the approaches.

[3]The two models were also compared using a likelihood-ratio test. The test statistic is highly significant ($\chi^2_1 = 9183.72, p < 0.001$), suggesting that estimation of the correlation parameter yields a better-fitting model.

Figure 4: Plots of Score Estimates for Longitudinal Test Data



Note. Gray circles show EAP scores using a 2-dimensional IRT model (MIRT) and two unidimensional models (QR). The lines mark, from left to right, the 1st, 25th, 50th, 75th, and 99th percentiles. For each plot, the ellipses demarcate 68% and 95% of the central density of an average posterior distribution. The locations (i.e., centers) of the ellipses are arbitrary.

First, consider the distributions of EAP scores for the two approaches, given by the gray circles. A byproduct of the QR approach, using two unidimensional models, is that the score estimates are not as highly correlated as in the MIRT case. Compared to the model-based $\hat{r} = 0.88$ in the MIRT case, the empirical correlation for the scores is 0.77 in the QR case. Also, with the QR approach, we observe certain unlikely score combinations, such as those where $\theta_p \approx -3$ and $\theta_c > 0$.

A second obvious difference between the two plots is the shape and location of the conditional quantiles. In the MIRT case, the conditional quantiles are linear, and entirely determined by the multivariate normal assumption for $g(\boldsymbol{\theta})$ and $\hat{r} = 0.88$. On the other hand, for the QR approach, the conditional quantiles are curvilinear, and depend on the empirical distribution of score estimates. Interestingly, at the left of the plot, the conditional quantiles curve upwards, to better fit the "unlikely" score combinations mentioned above. Additionally, compared to the QR conditional quantiles, the MIRT conditional quantiles are relatively close to one another, which reflects the higher correlation for the latent achievement dimensions.

A third clear difference is the size and shape of the ellipses, which are representative of an average uncertainty in estimates of $\boldsymbol{\theta}$. Since the QR approach utilizes two unidimensional IRT models, the standard errors for estimates of θ_p and θ_c are uncorrelated. For the MIRT case, the standard errors for θ_p and θ_c are correlated, since their calculation depends in part on $g(\boldsymbol{\theta})$. Also, for the MIRT approach, the EAP estimator "borrows strength" which leads to smaller average standard errors, and smaller ellipses.

Despite these differences in the QR and MIRT approaches, the results for the estimates of S are surprisingly similar. In addition to the similar values for ρ_S, the correlation between the two sets of SGP estimates is 0.98. This is not to say that the estimates are interchangeable, as they may differ in important ways with respect to bias (see Shang et al., 2015; McCaffrey et al., 2015).

5 Discussion & Conclusion

In this research, we proposed a measure to characterize SGP reliability. It is straightforward to calculate and has the advantage that it is easily interpretable. The measure was demonstrated using a MIRT approach with simulated data, and also calculated using the conventional QR approach in an empirical data analysis. This research also identified a major contributing factor to low SGP reliability: high correlations between latent achievement variables. The high correlations mean that most of the variation in the current achievement score can be explained by past achievement scores. Thus, the variance of the conditional distribution of current achievement given past achievement is typically small. Yet, the uncertainty of the latent achievement estimate is sizable, relative to the conditional distribution's variance. In this scenario, the reliability for SGP will tend to be low. Finally, this research demonstrated that including additional years of prior test scores should not be expected to increase ρ_S much, if at all. In fact, via simulation, it was shown that under certain circumstances, ρ_S will actually *decrease* when additional years are included.

While SGP estimates at the student level will typically have low reliability, aggregate estimates currently used in numerous states may have higher reliability. Nevertheless, given the high-stakes nature surrounding the use of aggregate SGPs, it is important to assess the reliability of these aggregate measures, particularly when a formal multilevel measurement model may be specified. The methods presented in this research may be applicable to this aggregate setting. This is one direction for future research.

Another direction for future research concerns the generalization of the MIRT approach along the lines of the semi-nonparametric MIRT (SNP-MIRT) model used in Monroe et al. (2014) for SGPs. There are numerous questions to explore, such as whether the SNP-MIRT approach can accommodate multiple prior years. Another question, directly related to the present research, is how the reliability of SGP estimates using SNP-MIRT would compare to the reliabilities using either the QR or MIRT approaches.

A final set of questions raised by this research stems from the findings regarding the inclusion of additional prior years. Generally speaking, the motivation for including additional years is to more fully contextualize current student achievement. Our research shows, however, that the effects of including additional years are hard to predict. The true S values sometimes, but not always, change as more years are included. And, ρ_S may (slightly) increase or decrease. The outcomes depend on subtleties in the specifications of the models that are used to summarize the data. Given these findings, when should additional years be included in the analysis? While this question is more policy-oriented, the longitudinal structure of the data also provokes methodological questions. For instance, can modeling techniques popularized in other fields, such as econometrics (e.g., Baltagi, 2008), be fruitfully applied to longitudinal student achievement data?

Due to interesting measurement issues and policy questions surrounding SGPs, the related methodologies deserve further attention, particularly by research psychometricians and assessment policy experts. This research was an attempt to explore and clarify aspects of the methodology, but much work remains.

Appendix

This Appendix presents some technical details regarding the MIRT model used for estimating SGPs, as well as the EAP estimators for $\boldsymbol{\theta}$ and S. Let $\mathbf{y}_c$ be observed responses on the current year's test, and $\mathbf{y}_p$ be observed responses for all of the prior tests. More formally, let $\mathbf{y}_p$ be partitioned into m sub-vectors $\mathbf{y}_p = (\mathbf{y}'_{p1}, \ldots, \mathbf{y}'_{pm})'$. Recall also the vector of past latent achievements $\boldsymbol{\theta}_p = (\theta_{p1}, \ldots, \theta_{pm})'$. The likelihood of the response pattern $\mathbf{y} = (\mathbf{y}'_c, \mathbf{y}'_p)'$, conditional on the latent variables is

$$L(\mathbf{y} \mid \boldsymbol{\theta}; \boldsymbol{\gamma}) = L(\mathbf{y}_c, \mathbf{y}_p \mid \theta_c, \boldsymbol{\theta}_p; \boldsymbol{\gamma}) = L(\mathbf{y}_c \mid \theta_c; \boldsymbol{\gamma}) \prod_{j=1}^{m} L(\mathbf{y}_{pj} \mid \theta_{pj}; \boldsymbol{\gamma}), \tag{9}$$

where $\boldsymbol{\gamma}$ collects together the free parameters of the MIRT model. Upon introducing the latent variable distribution $g(\boldsymbol{\theta})$, the marginal likelihood of $\mathbf{y}$ can be found by integrating over the unobserved latent variables $\boldsymbol{\theta}$

$$L(\mathbf{y}; \boldsymbol{\gamma}) = \int L(\mathbf{y}_c \mid \theta_c; \boldsymbol{\gamma}) \prod_{j=1}^{m} L(\mathbf{y}_{pj} \mid \theta_{pj}; \boldsymbol{\gamma}) g(\boldsymbol{\theta}; \boldsymbol{\gamma}) d\boldsymbol{\theta}. \tag{10}$$

Maximizing the marginal likelihood in Equation (10) yields $\hat{\boldsymbol{\gamma}}$, the vector of maximum likelihood estimates.

Whereupon plugging in the estimates of MIRT model parameters $\hat{\boldsymbol{\gamma}}$, the EAP estimator is defined as

$$\text{EAP}(\boldsymbol{\theta}_i) = E(\boldsymbol{\theta} \mid \mathbf{y}_{ic}, \mathbf{y}_{ip}) = \int \boldsymbol{\theta} \; \pi(\theta_c, \boldsymbol{\theta}_p \mid \mathbf{y}_{ic}, \mathbf{y}_{ip}; \hat{\boldsymbol{\gamma}}) d\boldsymbol{\theta}, \tag{11}$$

for a given individual i's item response pattern $\mathbf{y}_i = (\mathbf{y}'_{ic}, \mathbf{y}'_{ip})'$ on the current and all prior years' tests. The posterior distribution $\pi(\theta_c, \boldsymbol{\theta}_p \mid \mathbf{y}_c, \mathbf{y}_p; \boldsymbol{\gamma})$ is defined as

$$\pi(\theta_c, \boldsymbol{\theta}_p \mid \mathbf{y}_c, \mathbf{y}_p; \boldsymbol{\gamma}) = \frac{L(\mathbf{y}_c \mid \theta_c; \boldsymbol{\gamma}) \prod_{j=1}^{m} L(\mathbf{y}_{pj} \mid \theta_{pj}; \boldsymbol{\gamma}) g(\boldsymbol{\theta}; \boldsymbol{\gamma})}{L(\mathbf{y}; \boldsymbol{\gamma})}.$$

The accompanying standard errors, $\mathrm{SE}(\boldsymbol{\theta})$, are taken as the square roots of the diagonal elements of $\mathrm{Var}(\boldsymbol{\theta}|\mathbf{y}_c, \mathbf{y}_p)$, the posterior covariance matrix.

The EAP estimator of S is defined as

$$\mathrm{EAP}(S_i) = E\left[S(\theta_c, \boldsymbol{\theta}_p) \mid \mathbf{y}_{ic}, \mathbf{y}_{ip}\right] = \int S(\theta_c, \boldsymbol{\theta}_p)\pi(\theta_c, \boldsymbol{\theta}_p \mid \mathbf{y}_{ic}, \mathbf{y}_{ip}; \hat{\boldsymbol{\gamma}})d\boldsymbol{\theta}. \tag{12}$$

As with the standard errors for $\mathrm{EAP}(\boldsymbol{\theta}_i)$, the standard error for $\mathrm{EAP}(S_i)$ is the square root of the posterior variance,

$$\mathrm{SE}(S_i) = \sqrt{E\left\{[S(\theta_c, \boldsymbol{\theta}_p)]^2 \mid \mathbf{y}_{ic}, \mathbf{y}_{ip}\right\} - E^2\left[S(\theta_c, \boldsymbol{\theta}_p) \mid \mathbf{y}_{ic}, \mathbf{y}_{ip}\right]}, \tag{13}$$

where the first term on the right-hand side is

$$E\left\{[S(\theta_c, \boldsymbol{\theta}_p)]^2 \mid \mathbf{y}_{ic}, \mathbf{y}_{ip}\right\} = \int [S(\theta_c, \boldsymbol{\theta}_p)]^2 \pi(\theta_c, \boldsymbol{\theta}_p \mid \mathbf{y}_{ic}, \mathbf{y}_{ip}; \hat{\boldsymbol{\gamma}})d\boldsymbol{\theta}.$$

References

Baltagi, B. (2008). *Econometric analysis of panel data* (Vol. 1). John Wiley & Sons.

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. (2014). *SGP: Student growth percentiles & percentile growth trajectories.* Retrieved from `http://centerforassessment.github.com/SGP/` (R package version 1.2-0.0)

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (chap. 17-20). Reading, MA: Addison-Wesley.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, *6*, 431-444.

Cai, L. (2010a). High-dimensional exploratory item factor analysis by a Metropolis-Hastings Robbins-Monro algorithm. *Psychometrika*, *75*, 33–57.

Cai, L. (2010b). Metropolis-Hastings Robbins-Monro algorithm for confirmatory item factor analysis. *Journal of Educational and Behavioral Statistics*, *35*, 307–335.

Cai, L. (2010c). A two-tier full-information item factor analysis model with applications. *Psychometrika*, *75*, 581–612.

Cai, L. (2013). flexMIRT® version 2.0: Flexible multilevel and multidimensional item response theory analysis and test scoring [Computer software]. Chapel Hill, NC: Vector Psychometric Group, LLC.

Carroll, R. J., Küchenhoff, H., Lombard, F., & Stefanski, L. A. (1996). Asymptotics for the SIMEX estimator in nonlinear measurement error models. *Journal of the American Statistical Association*, *91*, 242-250.

Castellano, K. E., & Ho, A. D. (2013). Contrasting OLS and quantile regression approaches to student "growth" percentiles. *Journal of Educational and Behavioral Statis-*

tics, *38*(2), 23-28.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement*, *21*, 347-360.

Lockwood, J. R., & Castellano, K. (2015). Alternative statistical frameworks for Student Growth Percentile estimation. *Statistics and Public Policy*, *2*(1).

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

McCaffrey, D., Castellano, K., & Lockwood, J. R. (2015). The impact of measurement error on the accuracy of individual and aggregate SGP. *Educational Measurement: Issues and Practice*, *34*(1), 4-14.

Monroe, S. (2014). *Multidimensional item factor analysis with semi-nonparametric latent densities.* Graduate School of Education and Information Sciences, University of California, Los Angeles. (Unpublished Doctoral Dissertation)

Monroe, S., Cai, L., & Choi, K. (2014). *Student growth percentiles based on MIRT: Implications of calibrated projection* (Tech. Rep. No. 842). Los Angeles, CA: University of California: National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Reckase, M. D. (2009). *Multidimensional item response theory*. New York, NY: Springer.

Shang, Y. (2012). Measurement error adjustment using the SIMEX method: an application to student growth percentiles. *Journal of Educational Measurement*, *49*(4), 446-465.

Shang, Y., VanIwaarden, A., & Betebenner, D. (2015). Covariate measurement error correction for student growth percentiles using the SIMEX method. *Educational Measurement: Issues and Practice*, *34*(1), 15-21.

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement*, *28*(3), 237-247.

Thissen, D., & Orlando, M. (2001). Item response theory for items scored in two categories. In *Test scoring* (p. 73-140). Mahwah, NJ: Lawrence Erlbaum Associates.

Wells, C., Sireci, S., & Bahry, L. (2014). *The effect of conditioning years on the reliability of SGPs.* (Paper presented at the 2014 annual meeting of the National Council of Measurement in Education, Philadelphia, PA)