CRESST REPORT 842

STUDENT GROWTH PERCENTILES BASED ON MIRT:
IMPLICATIONS OF CALIBRATED PROJECTION

SEPTEMBER 2014

Scott Monroe
Li Cai
Kilchan Choi

National Center for Research
on Evaluation, Standards, & Student Testing
UCLA | Graduate School of Education & Information Studies

Student Growth Percentiles Based on MIRT:

Implications of Calibrated Projection

CRESST Report 842

Scott Monroe, Li Cai, and Kilchan Choi
CRESST/University of California, Los Angeles

September 2014

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Center for the Study of Evaluation (CSE)
Graduate School of Education & Information Studies
University of California, Los Angeles
300 Charles E. Young Drive North
GSE&IS Bldg., Box 951522
Los Angeles, CA 90095-1522
(310) 206-1532

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305D140046 to the National Center for Research on Evaluation, Standards, and Student Testing (CRESST). Li Cai's research was further supported by a grant from the Bill and Melinda Gates Foundation (OPP1088937).

To cite from this report, please use the following as your APA reference: Monroe, S., Cai, L., & Choi, K. (2014). *Student growth percentiles based on MIRT: Implications of calibrated projection.* (CRESST Report 842). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

TABLE OF CONTENTS

STUDENT GROWTH PERCENTILES BASED ON MIRT: IMPLICATIONS OF CALIBRATED PROJECTION

Scott Monroe, Li Cai, and Kilchan Choi

CRESST/University of California, Los Angeles

ABSTRACT

This research concerns a new proposal for calculating student growth percentiles (SGP, Betebenner, 2009). In Betebenner (2009), quantile regression (QR) is used to estimate the SGPs. However, measurement error in the score estimates, which always exists in practice, leads to bias in the QR-based estimates (Shang, 2012). One way to address this issue is to estimate the SGPs using a modeling framework that can directly account for the measurement error. Multidimensional IRT (MIRT) is one such framework, and the one utilized here. To maximize the generality of the approach, the SNP-MIRT model (Monroe, 2014), which estimates the shape of the latent variable density, is used to obtain model parameter estimates. These estimates are then used with the calibrated projection linking methodology (Thissen, Varni, et al., 2011, Thissen, Liu, Magnus, & Quinn, 2014, Cai, in press-a, Cai, in-press-b) to produce SGP estimates. The methods are compared using simulated and empirical data.

Keywords: item response theory, multidimensional, large-scale assessment

1 Introduction

The Student Growth Percentile (SGP, Betebenner, 2009) methodology is used to locate a student's current score in a conditional distribution based on the student's past scores. Instead of focusing solely on current achievement, SGPs provide context for that achievement. For example, suppose a student's current achievement is categorized as "below basic." By itself, this evaluation may be considered disappointing. However, if the accompanying SGP is 90, there is reason for encouragement: the interpretation is that this student's current achievement is higher than 90% of students who share the same score history. In this way, SGPs can add to our understanding of how well students are doing, and how they are progressing. Consequently, the SGP methodology has grown in popularity, and is used in numerous states to describe student performance. Moreover, the measure can be aggregated in an effort to describe teacher performance. In this latter case, the desired inference is that teachers with higher aggregate SGPs are more effective.

The original methodology uses quantile regression (QR) to calculate the SGPs, with past scores serving as covariates and the current score serving as the dependent variable (Betebenner, 2009). Another regression-based methodology was presented in Castellano and Ho (2013), which used OLS regression to calculate the SGPs. With either regression approach, however, measurement error in the scores introduces bias into the regression parameter estimates. As a consequence, the SGP estimates may also be biased. Within the context of QR, Shang (2012) applied simulation extrapolation (SIMEX) in an effort to correct the bias caused by measurement error in the covariates (i.e., past years' scores). An alternative strategy is to adopt a different modeling framework, one that accounts for measurement error more directly. Instead of using QR, the proposed approach is based on multidimensional IRT (MIRT). Notably, Lockwood and Castellano (in press) similarly proposes a MIRT framework for calculating SGPs. Specifically, our work is motivated by recent work on calibrated projection linking (Thissen et al., 2011; Thissen et al. 2014; Cai, in press-a, in press-b). Additionally, we consider a generalization of the MIRT approach based on SNP-MIRT (Monroe, 2014), where the shape of the multidimensional latent variable density is estimated along with the other model parameters.

The primary goals of the research are to validate the proposed methods and compare their performance to that of the QR-based method. To focus on these goals, the scope of the research is limited in numerous ways. For instance, within the MIRT

framework, we show how uncertainty in the latent variable estimates (i.e., achievement) is directly related to uncertainty in the SGP estimates. However, this issue is not further explored, as the primary goals concern only the SGP point estimates. As another example of the limited scope, we do not explore SGP aggregation to the teacher level. We hope to explore these and other topics in future research.

The remainder of this report is organized as follows. Section 2 outlines the proposed MIRT-based approach for estimating SGPs. The main steps in the procedure are presented and relevant supporting methods (e.g., calibrated projection) are reviewed. In Section 3, the SNP-MIRT model is reviewed, as well as its implications for estimating SGPs within a MIRT framework. Section 4 presents a simulation study and results, focusing on the proposed methods and a comparison to the QR-based approach. Then, in Section 5, state achievement data are analyzed with the different approaches and the results are compared. Finally, Section 6 provides a discussion of the research and potential future directions.

2 The Proposed Method

In this Section, the proposed method for calculating MIRT-based SGPs is presented. The method may be applied to latent scores based on response patterns. However, the use of full response patterns is inconvenient for introducing the method, since the number of possible response patterns is exponential in the number of test items. Instead, we use summed score based calculations, as this facilitates the presentation. Further, this choice is not inappropriate, as numerous large-scale assessment programs currently utilize summed score based scaled scores in reporting. Additionally, we limit the number of prior years to 1, again to facilitate the presentation and to make graphical illustrations more straightforward. The proposed method, though, may accommodate multiple prior years.

At this point, it is convenient to introduce some notation. Let there be N students who take tests in years 1 (last year) and 2 (current year). Let there be n_1 and n_2 items for the two years, where the subscript indicates the year. For any student, let y_1 and y_2 be vectors of observed response patterns. Similarly, let $s_1 \in \{0, \dots, S_1\}$ and $\mathrm{s}_2 \in \{0, \dots, \mathrm{S}_2\}$ be the analogously defined summed scores by adding elements of y_1 and y_2, where S_1 and S_2 are the respective maximum summed scores observable in years 1 and years 2. Finally, let θ_1 and θ_2 be the latent achievement scores for years 1 and 2. For convenience, we assume all students take both tests, and that there are no missing data, though both assumptions may be relaxed.

Before presenting the proposed method, we review the QR approach as implemented in this research, which serves as a point of comparison. First, two separate unidimensional IRT models are calibrated, one within each year. The item responses are modeled using the three-parameter logistic (3PL) model, which can be written as:

$$T(y|\theta) = g + (1-g)\frac{1}{1+\exp[-(\alpha+\beta\theta)]}, \tag{1}$$

where y is the item response (keyed 0 for incorrect and 1 for correct), g is the guessing parameter, α is the intercept, and β is the item slope. Note that θ in Equation (1) generally refers to either θ_1 or θ_2, depending on the test. Next, the respective IRT scaled scores based on summed scores are calculated, e.g., using the standard Lord-Wingersky algorithm (see Thissen & Wainer, 2001). For the QR, the estimated scaled scores from year 2 are the dependent variable, and the scores from year 1 are the covariate. Finally, the "SGP" R package (Betebenner, Van Iwaarden, Domingue, & Shang, 2014) can be used to compute the QR-based SGPs. For all packaged functions, we used the default settings.

Next, we turn to the proposed method. Together, the prior and current years imply a two-dimensional MIRT model, where each dimension is measured by one year's items. A multidimensional version of the 3PL is used:

$$P(y|\theta_1,\theta_2) = g + (1-g)\frac{1}{1+\exp[-(\alpha+\beta_1\theta_1+\beta_2\theta_2)]}, \tag{2}$$

where β_1 and β_2 are the slopes for dimensions 1 and 2, respectively. For all items measuring θ_1, β_2 is fixed to 0. Similarly, for all items measuring θ_2, β_1 is fixed to 0. This specification ensures that each latent dimension is only measured by the corresponding year's items, leading to an independent clusters factor pattern. Note, though, that all free parameters are jointly estimated.

To present the method, we introduce various conditional distributions with the generic notation $p(\cdot\,|\,\cdot)$. Unconditional distributions are indicated by the generic notation $p(\cdot)$. An example of the former type is $p(\theta_2|s_1,s_2)$, the posterior distribution of θ_2, given both years' summed scores. An example of the latter type is $p(\theta_1,\theta_2)$, the unconditional distribution (i.e., prior or population distribution) for the latent achievement variables. Let the correlation of $p(\theta_1,\theta_2)$ be ρ. Also, to illustrate the

calculation of various quantities, we define a model with $n_1 = 40$, $n_2 = 40$, so that $s_1, s_2 = \{0, \ldots, 40\}$. For the illustrations, $\rho = 0.85$.

Calculation of MIRT-based SGPs begins with an estimation of all MIRT item parameters, as well as the latent correlation ρ. Then, for any combination of s_1 and s_2, MIRT-based SGPs are calculated using the following steps:

Step 1. Latent Score Estimation. Estimate the current latent score, based on s_1 and s_2. This estimate (e.g., an EAP score) is denoted $\hat{\theta}_2|s_1, s_2$.

Step 2. Calibrated Projection. Using calibrated projection, find the reference conditional distribution, $p(\theta_2|s_1)$. This distribution is based on ρ and only s_1.

Step 3. SGP Estimation. Calculate the SGP using $\hat{\theta}_2|s_1, s_2$ and the cumulative distribution function of $p(\theta_2|s_1)$. The location of the current score estimate within this conditional distribution gives the MIRT-based SGP.

These steps are explained in greater detail in the three subsections to follow.

2.1 Latent Score Estimation

For Step 1, $\hat{\theta}_2|s_1, s_2$ is based on $p(\theta_2|s_1, s_2) = \int p(\theta_1, \theta_2|s_1, s_2)d\theta_1$. Recognizing that the MIRT model used here is a special case of the two-tier item factor model (Cai, 2010c) $p(\theta_1, \theta_2|s_1, s_2)$ may be found using a modified version of the Lord-Wingersky algorithm (Cai, in press-a). This algorithm makes it possible to calculate $p(\theta_1, \theta_2|s_1, s_2)$ without first calculating $p(\theta_1, \theta_2|y_1, y_2)$ for all y_1 and y_2. Again, the numbers of possible response patterns for y_1 and y_2 grow exponentially with n_1 and n_2. Thus, any method requiring calculations for all possible y_1 and y_2 will have practical limitations. On the other hand, the Lord-Wingersky algorithm makes calculation of $p(\theta_1, \theta_2|s_1, s_2)$ feasible for very large n_1 and n_2. As an aside, in the case of response pattern scoring, Step 1 is even more straightforward, as it is then only necessary to calculate $p(\theta_1, \theta_2|y_1, y_2)$ for all *observed* y_1 and y_2.

Before proceeding, we note that the posterior distribution $p(\theta_1, \theta_2|s_1, s_2)$ reflects the correlated measurement errors for the latent dimensions. While estimation of separate unidimensional IRT models, as in the QR approach to SGPs, may be used to estimate $p(\theta_1|s_1)$ and $p(\theta_2|s_2)$, such an approach does not account for measurement errors in the joint distribution of s_1 and s_2. Additionally, $\hat{\theta}_2|s_1, s_2$, obtained via the MIRT framework, will be a more efficient estimate than $\hat{\theta}_2|s_2$, obtained via unidimensional IRT (see Cai, 2010c).

2.2 Calibrated Projection

The development of calibrated projection (Thissen et al., 2011) was motivated by the need to link two highly similar, though not identical, constructs for the purposes of producing a scoring cross-walk. Utilizing a MIRT framework, calibrated projection provides a means to use item responses from one instrument to produce scores on the scale of a second instrument. These scores are summaries (e.g., EAPs) of the posterior distribution $p(\theta_2|s_1)$. In the original application of calibrated projection, $p(\theta_2|s_1)$ is needed because estimates of θ_2 are desired for people who have not taken test 2.

In contrast, in the current application of calibrated projection, we assume s_2 exists for all students. Still, $p(\theta_2|s_1)$ is a key quantity for MIRT-based SGPs: it represents the conditional distribution of the current latent achievement for all students with identical score histories. In other words, it is the reference conditional distribution from which an SGP may be estimated. We now provide an example of how calibrated projection may be used to find this reference distribution.

Consider Figure 1, which is akin to figures in Thissen et al. (2011) and Cai (2013). Given a specified s_1 (e.g., 20), the MIRT model implies a distribution on θ_1, shown on the *y*-axis. This distribution is then projected through the relationship between θ_1 and θ_2 to imply a distribution on θ_2, shown on the *x*-axis. This latter distribution, $p(\theta_2|s_1)$, is the model-implied achievement distribution for all students with identical score histories (here, $s_1 = 20$). It is used in Step 3 to estimate an SGP.

<u>Insert Figure 1 about here</u>

Other features of Figure 1 are worth mentioning. First, the light gray central ellipses represent $p(\theta_1, \theta_2)$, the prior distribution of the latent achievement variables. In Figure 1, $p(\theta_1, \theta_2)$ is bivariate normal with $\rho = 0.85$. Second, the dark gray central ellipses represent $p(\theta_1, \theta_2|s_1)$, which does not condition on s_2. As a result, $p(\theta_1, \theta_2|s_1)$ is more variable for θ_2. Finally, this relative uncertainty is projected onto the *x*-axis in $p(\theta_2|s_1)$. Given the estimated parameters, the location and scale of $p(\theta_2|s_1)$ are completely determined by s_1 and $p(\theta_1, \theta_2)$.

2.3 SGP Estimation

Step 3 is conceptually the most straightforward. Let $q_{\theta_2|s_1}(\theta_2)$ be the cumulative distribution function of $p(\theta_2|s_1)$, the reference conditional distribution from Step 2. Then, the MIRT-based SGP estimate is $q_{\theta_2|s_1}(\hat{\theta}_2|s_1, s_2)$, where $\hat{\theta}_2|s_1, s_2$ is the score estimate from Step 1. As an example, consider Figure 2, which shows $p(\theta_2|s_1 = 20)$ as

the large light gray distribution, and $p(\theta_2|s_1 = 20, s_2 = 30)$ as the small black distribution. The solid black vertical line segment marks the EAP of $p(\theta_2|s_1 = 20, s_2 = 30)$, $\hat{\theta}_2|s_1, s_2$. Its position within $p(\theta_2|s_1 = 20)$ is marked by the solid light gray vertical line segment, and corresponds to an SGP estimate of 88 for this score combination. That is, $q_{\theta_2|s_1}\left(\hat{\theta}_2\middle|s_1 = 20, s_2 = 30\right) = 88$.

Insert Figure 2 about here

Figure 2 also shows how uncertainty in the SGP estimate is directly related to uncertainty in $\hat{\theta}_2|s_1, s_2$. The dashed vertical lines in Figure 2 correspond to ± 1 and ± 2 standard errors of measurement for $\hat{\theta}_2|s_1, s_2$. Like $\hat{\theta}_2|s_1, s_2$, these values of θ_2 correspond to percentiles of $p(\theta_2|s_1)$, which are displayed in Figure 2. For any given s_1, the uncertainty in the SGP estimate will vary as a function of s_2. This phenomenon is presented graphically in Figure 3, for $s_1 = 20$. For each s_2, the boxplot demarcates the SGP estimates corresponding to $\hat{\theta}_2|s_1, s_2$, and ± 1 and ± 2 standard errors of measurement for $\hat{\theta}_2|s_1, s_2$. Note that the boxplots for s_2 values near $s_1 = 20$ are relatively large. This is because for these values, $p(\theta_2|s_1 = 20, s_2)$ is centrally located in relation to $p(\theta_2|s_1 = 20)$, where small changes in θ_2 lead to large changes in the SGP estimate. At least for this example, there is considerable uncertainty for most of the SGP estimates.

Insert Figure 3 about here

3 Generalizing the Method with SNP-MIRT

Recall the calibrated projection example given in Figure 1. Given the specified MIRT model and estimated parameters, the reference conditional distribution, $p(\theta_2|s_1)$ (shown on the *x*-axis) depends on two quantities: s_1 and $p(\theta_1, \theta_2)$. Consequently, for a given s_1 and s_2, different specifications of $p(\theta_1, \theta_2)$ may lead to different SGP estimates. In other words, the MIRT-based SGP estimates may be sensitive to the specification of the prior density for the latent variables.

As an example of this potential sensitivity we consider two different specifications for $p(\theta_1, \theta_2)$, holding all other aspects of the model constant. The first is the bivariate normal from the examples in Section 2, with null mean vector, unit variances, and $\rho = 0.85$. The second is a nonnormal specification, created using a mixture of normals, shown in Figure 4. This distribution likewise has null mean vector, unit variances, and $\rho = 0.85$. For a given s_1, each of these distributions may lead to a different reference conditional distribution.

Insert Figure 4 about here

Figure 5 shows the reference conditional distributions formed by using the normal (dashed gray curves) and nonnormal (solid black curves) priors for $s_1 = 20$ (left plot) and $s_2 = 35$ (right plot). For $s_1 = 20$, the two reference conditional distributions are quite similar, and the corresponding SGP estimates would likely be comparable. On the other hand, for $s_2 = 35$, the two reference conditional distributions are clearly different, suggesting that the specification of $p(\theta_1, \theta_2)$ can impact the MIRT-based SGP estimates. While $p(\theta_1, \theta_2)$ is typically assumed to be normal for MIRT modeling, there are alternatives, one of which is presented below.

Insert Figure 5 about here

3.1 Estimating the Latent Variable Density in IRT

Numerous efforts have been made to estimate the latent variable density within the framework of maximum marginal likelihood for unidimensional IRT models (e.g., Bock & Aitkin, 1981; Woods & Thissen, 2006; Woods & Lin, 2009; Monroe & Cai; 2014). In comparison, estimating the latent variable density for multidimensional IRT models has received less attention. A notable exception is Monroe (2014), which proposed and evaluated the use of a semi-nonparametric (SNP) density for the (possibly multidimensional) latent variables. The SNP density is a reparameterization of the density proposed by Gallant and Nychka (1987), and implemented for unidimensional IRT in Woods and Lin (2009). The resulting model, called SNP-MIRT, may be used in lieu of a standard MIRT model to estimate SGPs, following the same steps presented in Section 2.

3.2 Key Features of SNP-MIRT

It is beyond the scope of this report to present all of the technical details of the SNP-MIRT model. Instead, we present some key features of the research, with an emphasis on those most relevant to estimating SGPs.

First, the SNP density (Gallant & Nychka, 1987) is quite flexible, and can approximate a wide range of densities, including those with multiple modes. The flexibility of the SNP density is controlled by a tuning constant, with greater values of the tuning constant leading to a greater number of SNP parameters and greater flexibility of the density. For example, for two-dimensional densities, a tuning constant of 2 implies 5 "shape" parameters for the SNP density.

In much of the research utilizing the SNP density (e.g., Zhang & Davidian, 2001; Woods & Lin, 2009), the density is parameterized in such a way that its mean and variance are complex functions of all SNP parameters. Such a parameterization, however, makes it relatively difficult to place constraints on the mean and/or variance. To address this issue, Monroe (2014) introduced a new parameterization for the density with 3 types of parameters. Like a normal density, the new parameterization has mean and variance parameters. Unlike a normal, there are also shape parameters. When these shape parameters are each constrained to 0, the SNP density reproduces a normal. Consequently, the newly parameterized SNP density can be considered a generalization of the multivariate normal. Similarly, SNP-MIRT may be considered a generalization of standard MIRT (where a normal prior is specified).

Before leaving this section, we note that in empirical applications, the data do not always suggest nonnormality of the latent density. The decision on whether to use an SNP density in lieu of a normal can be informed by a likelihood ratio test comparing nested models, or by standard information criteria (e.g., AIC, HQIC).

4 Simulation Study

A simulation study was conducted to evaluate the proposed methods and to compare SGP approaches. In Simulation 1, latent variable scores were generated from a bivariate normal. Then, SGPs were estimated using the QR and MIRT approaches. In Simulation 2, latent variable scores were generated from the normal mixture shown in Figure 4. In this latter simulation, SGPs were estimated using the QR, MIRT, and SNP-MIRT approaches. Generally, data conditions were chosen to be representative of large-scale state assessments.

4.1 Data Generation and SGP Estimation

For both Simulations, the generating density had a null mean vector, variances equal to 1, and a correlation of $\rho = 0.85$. For Simulation 2, the latent variable density was a mixture of 2 bivariate normals with parameters: $\mu_1 = (-0.47, -0.47)'$, $\mu_2 = (1.09, 1.09)'$ $\sigma_1 = (0.48, 0.29, 0.48)'$, $\sigma_2 = (0.48, 0.44, 0.48)'$, $\mathrm{mp}_1 = 0.7$, and $\mathrm{mp}_2 = 0.3$, where $\sigma = \mathrm{vech}(\Sigma)$ stacks the unique elements of the covariance matrix Σ, and "mp" stands for mixing proportion.

Each of the two dimensions was measured by 40 items satisfying the three-parameter logistic (3PL) IRT model. Thus, $s_1, s_2 \in \{0, \ldots, 40\}$. Slopes were drawn from a truncated normal, with mean = 1.5 and standard deviation = 0.5, truncated at 0.5 and 3.

Intercepts were drawn from a normal with mean = 0 and standard deviation = 1. Finally, guessing parameters were drawn from a normal with mean = 0.25 and standard deviation = 0.05, truncated at 0.1 and 0.35. The logits of the guessing parameters, used in estimation, are denoted γ. All of the data-generating item parameters are presented in Table 1.

Insert Table 1 about here

Let π_{s_1,s_2} be the true model-implied probability for the summed-score combination given s_1 and s_2. For all combinations of s_1 and s_2, π_{s_1,s_2} may be calculated using a modified version of the Lord-Wingersky algorithm (Cai, in press-a). Figure 6 presents bubble plots of these probabilities for Simulation 1 (left plot) and Simulation 2 (right plot), with larger bubbles corresponding to greater probabilities. Although the overall patterns are similar, differences can be detected, in particular for score combinations where both s_1 and s_2 are high.

Insert Figure 6 about here

For each replication, SGPs based on QR, MIRT, and SNP-MIRT were estimated as described in Section 2. All SGP estimates were compared to "true" SGPs, calculated using the data-generating parameter values and MIRT model.

4.2 Collected Statistics

Several measures of accuracy were used to evaluate the SGP estimates. Momentarily suppressing reference to s_1 and s_2, let ψ be a true SGP and $\widehat{\psi}$ its corresponding estimate. Bias is defined as $M^{-1}\sum_{m=1}^{M}(\psi - \widehat{\psi}_m)$, where M is the number of Monte Carlo replications (here, 100). The absolute bias is defined as $\delta = M^{-1}\sum_{m=1}^{M}\left|\psi - \widehat{\psi}_m\right|$. Root mean square error (RMSE) is defined as $\sqrt{M^{-1}\sum_{m=1}^{M}\left(\psi - \widehat{\psi}_m\right)^2}$.

For a given s_1, the integrated absolute bias is

$$\bar{\delta}_{s_1} = \sum_{s_2=0}^{S_2} \delta_{s_1,s_2} W_{s_1,s_2}, \tag{3}$$

where $W_{s_1,s_2} = \pi_{s_1,s_2} / \sum_{s_2=0}^{S_2} \pi_{s_1,s_2}$. In words, this measure gives the expected absolute bias for a given s_1, averaged across all possible s_2 values.

The SGP estimates were also evaluated using correct classification rates (CCR). Given a set of cut-percentiles, such as (0, 35, 65, 100), the rate is defined as:

$$\mathrm{CCR} = M^{-1}N^{-1}\sum_{m=1}^{M}\sum_{i=1}^{N}\mathbf{1}_{\kappa(\psi_{im})}(\kappa(\hat{\psi}_{im})), \tag{4}$$

where $\kappa(\cdot)$ maps an SGP to a classification and $1_{\kappa(\psi_{im})}(\kappa(\hat{\psi}_{im})$ is an indicator function that returns a 1 if and only if $\kappa(\hat{\psi}_{im})$ is equal to $\kappa(\psi_{im})$, and 0 otherwise. The CCR is simply the proportion of estimated classifications that agree with true classifications.

Bias and RMSE statistics were also collected for the item parameter estimates. For Simulation 1, since the models are correctly specified, the estimates should be approximately unbiased. However, for Simulation 2, all fitted models are misspecified, since the true latent variable density is a normal mixture. Thus, the bias and RMSE statistics can help to measure the sensitivity of the models to this misspecification. Additionally, given the flexibility of the SNP-MIRT model, the bias and RMSE statistics can shed light on whether SNP-MIRT outperforms standard MIRT in terms of parameter recovery. Finally, the log-likelihoods and HQIC values of the MIRT and SNP-MIRT models were collected to make comparisons of overall fit.

4.3 Results for Simulation 1: Normal Latent Density

Table 2 presents some parameter recovery results for Simulation 1. As expected, the estimates for all parameter types are approximately unbiased. The RMSE values from the MIRT model are slightly smaller than those from the two separate unidimensional IRT models used for the QR approach. This is to be expected since the correlation between dimensions in the MIRT model leads to an increase in efficiency of the parameter estimates. Table 2 also serves as a point of reference for parameter recovery results for Simulation 2, when the fitted models are misspecified.

Insert Table 2 about here

Figures 7 and 8 present the bias in SGP estimates for all cross-classifications of s_1 and s_2 for the QR and MIRT-based approaches, respectively. The MIRT-based estimates are nearly unbiased, and much less biased than the QR-based estimates. Further, the magnitude and direction of bias for the QR-based estimates clearly depend on s_1 and s_2. One notable trend is that there is relatively little bias in SGP estimates for $s_1 \approx s_2$, at least when each summed score is around 15 or greater. However, for score combinations near this diagonal where $s_1 > s_2$, there is a clear pattern of positive bias. In contrast, near this diagonal, when $s_1 < s_2$, there is a clear pattern of negative bias.

Another trend is that for score combinations where both s_1 and s_2 are relatively small, the QR-based estimates are consistently negatively biased.

Insert Figure 7 about here

Insert Figure 8 about here

A shortcoming of the bivariate plots in Figures 7 and 8 is that they do not incorporate the model-implied probabilities for each summed score combination (see Figure 6). For example, the bias corresponding to the combination $s_1 = 0$ and $s_2 = 0$ may not be particularly important, since π_{s_1,s_2} is extremely small in that case. One way to focus our attention is to identify the most probable summed score combinations. Here, the 99% Highest Density Region (HDR, Rosa et al., 2001) of combinations is identified, which comprises the minimum number of most probable combinations sufficient to account for 99% of the probability mass. In other words, the least probable combinations are ignored. Figure 9 plots the mean SGP estimates against the true SGP values for the 99% HDR score combinations, for both the QR (left plot) and MIRT (right plot) approaches. Again, the MIRT-based estimates are nearly unbiased, while the QR-based estimates are systematically biased. Generally, the QR-based SGP estimates are negatively biased for smaller values and positively biased for greater values. The implication is that the QR approach tends to "exaggerate" SGPs.

Insert Figure 9 about here

Another way to focus our attention is to look at the integrated absolute bias, given s_1. Figure 10 shows the integrated absolute bias for both QR and MIRT approaches, with the MIRT approach again outperforming the QR approach. Two other features of Figure 10 are worth mentioning. First, for this condition, the QR approach performs better near the middle of the s_1 range and worse near the extremes. Second, while the integrated absolute bias for low values of s_1 is relatively great, those values of s_1 have low model-implied probabilities (see Figure 6).

Insert Figure 10 about here

Finally, Table 3 presents CCRs for the two approaches for several sets of cut-percentiles. For the set with 3 classes, both approaches are quite accurate, with rates of 0.95 and 0.99 for the QR and MIRT approaches, respectively. As expected, as the number of classes increases, the accuracies for both approaches decrease. However, the accuracy for the MIRT approach decreases relatively slowly with an accuracy rate of 0.97 for 10 classes.

Insert Table 3 about here

Based on the measures considered, the MIRT-based approach performed extremely well in Simulation 1. These results, however, represent the unrealistic situation where the model, including the model for the latent variable density, is exactly correctly specified. This correct specification, in conjunction with the ML estimator, leads to asymptotically unbiased IRT parameter estimates. Further, the large sample sizes in the replications (N = 10,000) resulted in highly efficient parameter estimates. And, since the MIRT-based SGPs are a function of the latent variable density and parameter estimates, it should not be surprising that the MIRT approach performed so strongly. Simulation 2, however, presents the more challenging condition where the latent variable density is misspecified.

4.4 Results for Simulation 2: Nonnormal Latent Density

The plots in Figure 11 show the true generating latent variable density (left column) and the SNP-MIRT estimated density for Simulation 2. The top row displays the true bivariate contour and the median of the estimated SNP densities. The resemblance between the two plots indicates the SNP-MIRT model was, to some degree, effective in estimating the shape of the latent density. The middle and bottom rows show the univariate marginal densities for θ_1 and θ_2, respectively. In the left column for these plots, the true generating density is represented by the black curve, while the gray curve is a standard normal, provided as a reference. In the right column for these plots, the median of the estimated SNP densities is shown in black, while the dashed gray curves provide an empirical 90% confidence interval. Again, the right column (SNP estimate) resembles the left column (true generating). These results of density recovery are similar to those reported in Monroe (2014), and suggest that the SNP-MIRT model can be effective in capturing nonnormality in the latent variable density.

Insert Figure 11 about here

Table 4 presents parameter recovery results disaggregated by parameter type for the different estimation approaches, including SNP-MIRT. For all parameter types, the SNP-MIRT model yields estimates with the least bias and the lowest RMSE values. And in contrast to the results from Simulation 1 (see Table 2), Table 4 suggests that the QR and standard MIRT approaches lead to biased parameter estimates. These results are consistent with other research on nonnormal latent variables in IRT (e.g., Woods & Thissen, 2006; Monroe, 2014) that has found that failing to account for the nonnormality can lead to biased parameter estimates. Also, for every replication, the SNP-MIRT

model was preferred over the standard MIRT model based on both $-2\times$ log-likelihood and HQIC values.

Insert Table 4 about here

To summarize the results for Simulation 2 thus far, the SNP-MIRT model was fairly successful in estimating the shape of the true nonnormal latent density, and performed the best among the methods in terms of parameter recovery. Since the SGP estimation approach presented in this research depends on the latent density and item parameter estimates, we should expect the SGP estimates based on the SNP-MIRT model to be more accurate than those based on the standard MIRT model. We now turn to those results.

Figures 12-14 present the bias in SGP estimates for all summed score cross-classifications for the QR, MIRT, and SNP-MIRT approaches, respectively. The pattern of bias for the QR-based estimates in Figure 12 is similar to the corresponding pattern from Simulation 1 in Figure 7. Specifically, the pattern of positive and negative bias is similar. One apparent difference, though, is the magnitude of the bias, in particular for the highest summed score combinations. The bias in the QR-based estimates for these combinations is greater in Simulation 2 than the corresponding bias in Simulation 1.

Insert Figure 12 about here

Recall Figure 8 from Simulation 1, which showed that the MIRT-based SGP estimates were approximately unbiased when the latent trait density was correctly specified as normal. Figure 13 presents a sharp contrast, as there is considerable bias in the MIRT-based SGP estimates for many cross-classifications. Also, the pattern of bias appears to be mostly a function of s_2. For lower and higher values of s_2, the bias tends to be positive, whereas for more central values of s_2, the bias tends to be negative. We do not offer an explanation for this pattern, beyond that it likely depends on the shape of $p(\theta_1, \theta_2)$. In any event, Figure 8 makes clear that the proposed method of estimating SGPs based on standard MIRT is clearly sensitive to the misspecification of the latent trait density.

Insert Figure 13 about here

Finally, Figure 14 presents the bias by cross-classification for the SGP estimates based on the SNP-MIRT model. Overall, the SNP-MIRT approach results in much less bias than the QR and MIRT approaches (see Figures 12 and 13). At the same time, the approach results in more bias than the MIRT approach in Simulation 1 (see Figure 8).

This can be explained by the small amount of bias in the item parameter and density estimation for the SNP-MIRT approach.

Insert Figure 14 about here

As in Simulation 1, we also present plots of mean SGP estimates against true SGP values for the 99% HDR score combinations. Figure 15 shows these plots for the QR (left plot), MIRT (center plot), and SNP-MIRT (right plot) approaches. As in Simulation 1, the QR approach seems to exaggerate the SGP estimates at the high and low ends. The MIRT and SNP-MIRT estimates, in comparison, are better aligned with the 45° line, and it is more difficult to discern any pattern in the plot. Comparing just the MIRT and SNP-MIRT estimates, the latter shows a tighter correspondence with the true SGP values.

Insert Figure 15 about here

The final plot for the results of Simulation 2 is Figure 16, which displays the integrated absolute bias, given s_1, for all three methods. Comparing the QR and SNP-MIRT methods is straightforward: both have higher values at the extremes of s_1, but the bias for the SNP-MIRT approach is always smaller. The results corresponding to the MIRT method, however, are not easily summarized. Overall, the bias values for the MIRT method tend to fall between the SNP-MIRT approach and QR approach, but there are exceptions. For instance, for $s_1 = 38$, the MIRT method produces the smallest bias, while for $s_1 = 25$, it produces the largest bias. This variability across s_1 is likely due to the shape of the latent trait density.

Insert Figure 16 about here

As with Simulation 1, we can also examine CCRs for the 3 methods, presented in Table 5. The SNP-MIRT approach is the most accurate, regardless of which set of cut-percentiles is used. The other two methods, QR and MIRT, are comparable to one another, but clearly less accurate than the SNP-MIRT approach.

Insert Table 5 about here

5 Empirical Application

To illustrate the proposed SGP estimation methodology, we use longitudinally-matched student achievement data from the 2011-2012 and 2012-2013 academic years. The data are from a state's mathematics assessments, but due to confidentiality agreements, the state is not identified. For each year, 44 dichotomous items were

analyzed. These items do not constitute a vertical scale. A random sample of 10,000 complete cases was drawn.

SGPs based on the QR, MIRT, and SNP-MIRT approach were calculated as described and illustrated in Sections 3 and 4. As before, a tuning constant of 2 is used for the SNP density, leading to 5 shape parameters in the SNP-MIRT model. Also, as before, the 3PL model, or its multidimensional version, was used for all items.

Figure 17 shows the contour plots of the latent trait density for the standard MIRT (left plot) and SNP-MIRT (right plot) models. The estimated SNP density appears approximately normal, although the estimated correlation of 0.86 is slightly smaller than the 0.88 estimate for the standard MIRT model. Figure 18 shows the estimated univariate marginals for the SNP-density (solid black curves) along with normal densities (dashed gray curves) provided for reference. For θ_1, the estimated SNP density is slightly peaked and left-skewed in relation to the normal. A practical interpretation of this is that a greater proportion of students in this sample had lower latent achievement levels in 2011-2012 than would be expected using a normal distribution. On the other hand, for θ_2, the estimated SNP density cannot be distinguished from a normal.

Insert Figure 17 about here

Insert Figure 18 about here

Turning to model comparison, Table 6 provides $-2\times$ log-likelihood and HQIC values for the different models. Focusing on the multidimensional models, the SNP-MIRT model is preferred by both of these criteria. Also, since the multidimensional models are nested, a likelihood ratio test can be used to judge whether the additional constraints placed on the shape parameters by the standard MIRT model lead to a significant decrement in model fit. Since the test statistic is highly significant ($\chi^2_5 = 452.5, p < 0.001$), we conclude that the standard MIRT model does not fit the data as well as the SNP-MIRT model.

Insert Table 6 about here

Next, we can compare the SGP estimates from the 3 methods. Figure 19 displays bivariate plots of SGP estimates for a random subsample of 1,000 students. The estimates based on the MIRT and SNP-MIRT approaches are highly similar, as evidenced by the correspondence of estimates in the lower-right plot. As for the QR-based estimates, they tend to be more extreme than the estimates based on the other

two methods. Interestingly, this is the same pattern exhibited by the QR-based estimates in Simulation 1 (see Figure 9) and Simulation 2 (see Figure 15). However, here, the QR-based estimates are being plotted against the MIRT and SNP-MIRT-based approaches, as opposed to the true values.

Insert Figure 19 about here

Table 7 also measures the similarity in the 3 sets of estimates by looking at the pairwise classification agreement rates. For the set of cut-percentiles with 3 classes, all methods produce similar results, with agreement rates of 0.94 and higher. However, for the set of cut-percentiles with 10 classes, the results depend substantially on the method. In particular, the QR-based method has low classification agreement rates (< 0.68) with the MIRT and SNP-MIRT methods, while the latter two methods have a relatively high agreement rate (0.92).

Insert Table 7 about here

6 Conclusion and Future Directions

In this research, a new method was presented to calculate SGPs within a MIRT framework, capitalizing on recent research on calibrated projection. The calibrated projection technique can be used to find the reference conditional distribution, which is necessary for SGP estimation. This research also presented a generalization or variation of the MIRT approach, based on SNP-MIRT (Monroe, 2014). The new methods were compared to the QR-based method using both simulated and empirical data.

The proposed methods performed well in the simulation study and are worthy of further investigation. In Simulation 1, when the true latent trait density was specified as normal, the MIRT-based approach produced nearly unbiased SGP estimates, while the QR-based approach led to "exaggerated" SGP estimates. In Simulation 2, when the true latent trait density was nonnormal, the results were more mixed. For this condition, the SNP-MIRT approach was effective in estimating the latent trait density shape, and led to the most accurate SGP estimates.

The empirical example provokes several questions and ideas. The SNP-MIRT method yielded a density estimate that was nonnormal, but only slightly so. Consequently, SGP estimates based on the SNP-MIRT and standard MIRT approaches were highly similar. This suggests that the SNP-MIRT approach may serve as a type of sensitivity analysis for the standard MIRT approach. It is unclear, however, how different the SGP estimates need to be to justify the use of the more complex SNP-MIRT

model. Further, the empirical latent trait density suggests that the specified density in Simulation 2 may have been too extreme in its nonnormality. Additional empirical datasets should be analyzed to develop a better understanding of "typical" density shapes for large-scale longitudinal achievement data.

Finally, there are numerous topics for future research. Some of these topics involve the generality of the proposed methods. Theoretically, the framework accommodates scale scores based on response patterns (as opposed to summed scores) as well as multiple prior years of achievement data. The MH-RM algorithm (Cai, 2010a, 2010b) may be used to obtain maximum marginal likelihood estimates for high-dimensional MIRT models with arbitrary factor structures. Additionally, Thissen et al. (2014) demonstrated that calibrated projection could be generalized to more than two dimensions. However, these generalizations have not been applied to the MIRT-based approach for SGP estimation. Further, the framework should accommodate residual dependencies among items across years (Cai, in press-a) that may result from, for instance, the use of the same item in consecutive year. This too has yet to be demonstrated. Another interesting direction concerns the uncertainty in the individual SGP estimates. The preliminary research here suggests a great deal of uncertainty at the level of the student. On a related note, future research should focus on SGP aggregation. For instance, how does uncertainty at the student level affect uncertainty in teacher level estimates? Research on these last two topics, in particular, would be of great interest to policymakers.

7 References

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

Betebenner, D. W., Van Iwaarden, A., Domingue, B. D., & Shang, Y. (2014). *An R package for the calculation and visualization of student growth percentiles and percentile growth trajectories*. Retrieved on June 25, 2014, from http://cran.r-project.org/web/packages/SGP/

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*, 443–459.

Cai, L. (in press-a). Lord-Wingersky algorithm version 2.0 for hierarchical item factor models with applications in test scoring, scale alignment, and model fit testing. *Psychometrika*.

Cai, L. (in press-b). Two-tier item factor analysis modeling. In W. J. van der Linden, & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (2nd edition). Boca Raton, FL: Chapman & Hall/CRC.

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

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.

Castellano, K. E., & Ho, A. D. (2013). Contrasting OLS and quantile regression approaches to student "growth" percentiles. *Journal of Educational and Behavioral Statistics*, *38*, 190-215.

Gallant, A. R., & Nychka, D. W. (1987). Semi-nonparametric maximum likelihood estimation. *Econometrica*, *55*(2), 363-390.

Lockwood, J. R., & Castellano, K. E. (in press). Alternative statistical frameworks for student growth percentile estimation. *Statistics and Public Policy*.

Monroe, S. (2014). *Multdimensional item factor analysis with semi-nonparametric latent densities*. Unpublished doctoral dissertation, Education Department, University of California, Los Angeles.

Monroe, S., & Cai, L. (2014). Estimation of a Ramsay-curve IRT model by the Metropolis-Hastings Robbins-Monro algorithm. *Educational and Psychological Measurement*, *74*(2), 343-369.

Rosa, K., Swygert, K. A., Nelson, L., & Thissen, D. (2001). Item response theory applied to combinations of multiple-choice and constructed-response items—scale scores for patterns of summed scores. In D. Thissen & H. Wainer (Eds.), *Test Scoring* (pp. 253-292). Mahwah, NJ: Lawrence Erlbaum Associates.

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

Thissen, D., & Wainer H. (2001). *Test Scoring*. Mahwah, NJ: Lawrence Erlbaum Associates.

Thissen, D., Liu, Y., Magnus, B., & Quinn, H. (2014). E*xtending the use of multidimensional IRT calibration as projection: Many-to-one linking and linear computation of projected scores.* Paper presented at the International Meeting of the Psychometric Society. Madison, Wisconsin.

Thissen, D., Varni, J. W., Stucky, B. D., Liu, Y., Irwin, D. E., & DeWalt, D. A. (2011). Using the PedsQL™ 3.0 asthma module to obtain scores comparable with those of the PROMIS pediatric asthma impact scale (PAIS). *Quality of Life Research*, *20*, 1497-1505.

Woods, C. M., & Lin, N. (2009). Item response theory with estimation of the latent density using Davidian curves. *Applied Psychological Measurement*, *33*(2), 102-117.

Woods, C. M., & Thissen, D. (2006). Item response theory with estimation of the latent population distribution using spline-based densities. *Psychometrika*, *71*, 281–301.

Zhang, D., & Davidian, M. (2001). Linear mixed models with flexible distributions of random effects for longitudinal data. *Biometrics*, *57*(3), 795-802.

8 Tables 1-7

Table 1
Generating Parameters for Simulation Studies 1 & 2

item	1	2	3	4	5	6	7	8	9	10
γ	-1.41	-0.79	-1.19	-1.34	-1.16	-1.15	-0.82	-1.08	-0.91	-1.24
α	-0.71	0.26	-0.25	-0.35	-0.95	-0.05	-0.78	-1.67	-0.38	0.92
β_1	1.22	1.38	2.28	1.54	1.56	2.36	1.73	0.87	1.16	1.28
item	11	12	13	14	15	16	17	18	19	20
γ	-1.04	-1.19	-1.07	-1.35	-1.49	-0.62	-0.94	-1.47	-1.27	-1.45
α	-0.58	0.61	-1.62	-0.06	0.52	0.3	0.11	-0.64	-0.85	-1.02
β_1	2.11	1.68	1.7	1.56	1.22	2.39	1.75	0.52	1.85	1.26
item	21	22	23	24	25	26	27	28	29	30
γ	-0.77	-1.17	-0.96	-1.21	-1.23	-1.32	-1.26	-0.7	-1.11	-1.07
α	0.12	-0.95	-0.49	-0.26	1.84	-0.65	0.24	0.08	-0.96	-0.07
β_1	0.97	1.39	0.99	1.14	1.19	0.66	1.92	1.58	0.93	2.13
item	31	32	33	34	35	36	37	38	39	40
γ	-1.03	-0.79	-1.24	-1.38	-0.69	-1.22	-1.3	-1.46	-1.48	-1.26
α	1.44	0.45	0.04	-0.42	-2.05	1.13	-1.46	0.74	1.91	-1.44
β_1	1.71	1.35	1.95	1.94	1.91	1.84	1.78	1.47	1.35	1.31
item	41	42	43	44	45	46	47	48	49	50
γ	-0.94	-0.82	-0.92	-1.2	-1.08	-1.3	-1.3	-0.87	-1.39	-0.63
α	0.7	-0.26	-1.57	-1.51	-1.6	-0.53	-1.46	0.69	2.1	-1.29
β_2	1.15	1.4	0.87	2.58	2.1	0.94	1.3	1.27	1.89	1.46
item	51	52	53	54	55	56	57	58	59	60
γ	-1.12	-1.04	-1.31	-1.26	-1.49	-1.15	-0.99	-1.01	-1.32	-1.32
α	0.79	0.77	0.33	-1.01	-0.12	-0.28	0.56	-0.37	0.98	-0.37
β_2	1.63	1.49	1.48	2.18	1.39	2.26	0.73	1.79	1.56	1.61
item	61	62	63	64	65	66	67	68	69	70
γ	-1.24	-0.73	-1.43	-1.15	-0.64	-1.13	-1.5	-1.28	-0.97	-1.2
α	1.05	-1.05	-1.26	3.24	-0.42	0.3	0.64	-0.48	0.52	0.37
β_2	1.69	1.25	1.33	0.99	0.96	1.65	1.72	1.53	1.96	2.53
item	71	72	73	74	75	76	77	78	79	80
γ	-1.25	-1.19	-1.07	-0.71	-1.12	-0.83	-0.94	-1.13	-1.56	-1.24
α	-0.22	0.07	-0.03	2.13	-0.74	-1.1	0.04	0.31	0.44	-0.46
β_2	1.25	2	1.15	1.16	2.01	1.36	0.89	1.59	1.43	1.5

Table 2
Simulation 1 Results: Parameter Recovery

	Bias			RMSE		
	Parameter Type			Parameter Type		
Method	γ	α	β	γ	α	β
QR	0.01	-0.01	0.01	0.15	0.12	0.09
MIRT	0.01	-0.01	0.01	0.14	0.11	0.09

Note. "QR" refers to calibration via two separate unidimensional IRT models, one for year/test 1, one for year/test 2. "γ" is logit-guessing parameter; "α" is intercept; "β" is slope.

Table 3
Simulation 1 Results: Correct Classification Rates for SGPs

		Correct Classification Rate	
Classes	Cut-Percentiles	QR	MIRT
3	(0, 35, 65, 100)	0.949	0.989
4	(0, 25, 50,…, 100)	0.916	0.988
5	(0, 20, 40,…, 100)	0.893	0.985
10	(0, 10, 20,…, 100)	0.749	0.966

Note. Figures based on simulation study with *N*=10,000 and 100 replications.

Table 4
Simulation 2 Results: Parameter Recovery

	Bias			RMSE		
	Parameter Recovery			Parameter Recovery		
Method	γ	α	β	γ	α	β
QR	0.22	-0.25	0.23	0.26	0.28	0.26
MIRT	0.20	-0.19	0.18	0.24	0.23	0.21
SNP-MIRT	-0.04	0.05	-0.00	0.17	0.13	0.10

Note. "QR" refers to calibration via two separate unidimensional IRT models, one for year/test 1, one for year/test 2. "γ" is logit-guessing parameter; "α" is intercept; "β" is slope.

Table 5

Simulation 2 Results: Correct Classification Rates for SGPs

		Correct Classification Rate		
Classes	*Cut-Percentiles*	*QR*	*MIRT*	*SNP-MIRT*
3	(0, 35, 65, 100)	0.922	0.917	0.967
4	(0, 25, 50,…, 100)	0.882	0.893	0.947
5	(0, 20, 40,…, 100)	0.835	0.848	0.932
10	(0, 10, 20,…, 100)	0.677	0.705	0.854

Note. Figures based on simulation study with N=10,000 and 100 replications.

Table 6

Empirical Application Results: Model Comparisons

Model	*N*	*Parameters*	$\hat{\rho}$	$-2\times$ *LogL*	*HQIC*
IRT ($\times 2$)	10000	264	0	987218.00	988390.30
MIRT	10000	265	0.881	978034.28	978545.35
SNP-MIRT	10000	270	0.860	977581.80	978102.51

Note. "IRT ($\times 2$)" refers to two unidimensional IRT models, the first for data from year 1, the second for data from year 2. Together, the fitted models are formally equivalent to a two-dimensional MIRT model where the latent variable correlation is fixed to 0.

Table 7

Empirical Application Results: SGP Classification Agreement Rates

		Classification Agreement Rate		
Classes	*Cut-Percentiles*	*QR/MIRT*	*QR/SNP-MIRT*	*MIRT/SNP-MIRT*
3	(0, 35, 65, 100)	0.953	0.946	0.982
4	(0, 25, 50,…, 100)	0.907	0.890	0.959
5	(0, 20, 40,…, 100)	0.846	0.837	0.953
10	(0, 10, 20,…, 100)	0.676	0.660	0.915

Note. Figures based on sample of N=10,000. Classification Agreement Rates are pairwise.

9 Figure Captions

Figure 1. *Calibrated projection linking. Given a specified* s_1 *(here,* $s_1 = 20$*), the MIRT model implies a distribution on* θ_1*, shown on the* y*-axis. This distribution,* $p(\theta_1|s_1)$*, is then projected through the relationship between* θ_1 *and* θ_2 *to imply a distribution on* θ_2*, shown on the* x*-axis. The dark gray central ellipses approximate* $p(\theta_1, \theta_2|s_1)$*. The light gray central ellipses represent the prior distribution of latent scores,* $p(\theta_1, \theta_2)$*.*

Figure 2. *Illustration of MIRT-based SGP calculation. The dominating light gray curve is* $p(\theta_2|s_1)$ *(here,* $s_1 = 20$*). The smaller dark gray curve is* $p(\theta_2|s_1, s_2)$ *(here,* $s_2 = 30$*). The 5 vertical line segments demark the expectation of* $p(\theta_2|s_1, s_2)$*, as well as* ± 1 *and* ± 2 *standard errors of measurement. The extended line segments (light gray) correspond to percentile values for* $\mathrm{p}(\theta_2|\mathrm{s}_1)$*. Here,* $\mathrm{s}_1 = 20$ *and* $\mathrm{s}_2 = 30$ *yields an SGP point estimate of* 88*.*

Figure 3. *Boxplots of MIRT-based SGPs corresponding to EAP scores and* ± 1 *and* ± 2 *standard errors of measurement for* $\mathrm{s}_1 = 20$ *and all possible* s_2*. The horizontal dotted lines correspond to possible SGP cut-values of 35 and 65. Many boxplots span all 3 "classifications." The boxplot above* $\mathrm{s}_2 = 30$ *corresponds to Figure 2.*

Figure 4. *Example of bivariate non-normal density created as a mixture of normals. The variance for each dimension is* 1*, and the correlation between dimensions is* $\rho = 0.85$*. Each marginal distribution is standardized, but skewed right.*

Figure 5. *Examples of* $\mathrm{p}(\theta_2|\mathrm{s}_1)$ *for* $\mathrm{s}_1 = 20$ *(left plot) and* $\mathrm{s}_1 = 35$ *(right plot) using a normal prior distribution (light gray, dashed curve) and the nonnormal distribution from Figure 4 (black solid curve). For both prior distributions,* $\rho = 0.85$*.*

Figure 6. *Bubble plots of model-implied probabilities of summed score combinations for Simulations. Larger bubbles correspond to greater model-implied probabilities.*

Figure 7. *Simulation 1 results: bias for QR-based SGP estimates for all cross-classifications of s_1 and s_2. Bias is defined as the average estimate across all 100 replications minus the true value.*

Figure 8. *Simulation 1 results: bias for MIRT-based SGP estimates for all cross-classifications of s_1 and s_2. Bias is defined as the average estimate across all 100 replications minus the true value.*

Figure 9. *Simulation 1 results: integrated absolute bias for QR- and MIRT-based SGP estimates for all s_1. For each replication, for each s_1, the absolute bias is integrated over the true model-implied probabilities for all s_2. These values are then averaged over all 100 replications.*

Figure 10. *Simulation 1 results: plots of SGP estimates (*y*-axis) against true SGP values (*x*-axis) for QR (left plot) and MIRT (right plot). Estimates are averages over all 100 replications.*

Figure 11. *Simulation 2 results: plots of true generating (left column) and SNP-estimated (right column) prior latent trait densities. The top row shows the bivariate distributions, $p(\theta_1, \theta_2)$. The middle row shows the univariate marginal, $p(\theta_1)$. On the left, the light gray distribution is a standard normal, shown for reference. On the right, the dashed light gray curves give a 90% empirical confidence interval. The bottom row provides the same information as the middle row, but for θ_2.*

Figure 12. *Simulation 2 results: bias for QR-based SGP estimates for all cross-classifications of s_1 and s_2. Bias is defined as the average estimate across all 100 replications minus the true value.*

Figure 13. *Simulation 2 results: bias for MIRT-based SGP estimates for all cross-classifications of s_1 and s_2. Bias is defined as the average estimate across all 100 replications minus the true value.*

Figure 14. *Simulation 2 results: bias for SNP-MIRT-based SGP estimates for all cross-classifications of* s_1 *and* s_2*. Bias is defined as the average estimate across all 100 replications minus the true value.*

Figure 15. *Simulation 2 results: integrated absolute bias for SGP estimates for all* s_1*. For each replication, for each* s_1*, the absolute bias is integrated over the true model-implied probabilities for all* s_2*. These values are then averaged over all 100 replications.*

Figure 16. *Simulation 2 results: plots of SGP estimates (y-axis) against true SGP values (x-axis) for QR (left plot) and MIRT (middle plot) and SNP-MIRT (right plot). Estimates are averages over all 100 replications.*

Figure 17. *Empirical application results: estimates of bivariate normal (left plot) and SNP density (right plot) for the prior distribution of latent trait scores. The estimated correlations are* $\rho = 0.88$ *for the normal and* $\rho = 0.86$ *for the SNP density.*

Figure 18. *Empirical application results: estimates of univariate marginal prior distributions of the latent trait scores for* θ_1 *(left) and* θ_2 *(right). The light gray dashed curves are normal from the standard MIRT estimation; the black solid curves are from the SNP-MIRT estimation.*

Figure 19. *Empirical application results: plots of estimated SGPs for a random sub-sample of 1,000 from the full sample of 10,000. The same random sub-sample is used in all 3 plots.*

10 Figures 1-19

Figure 1

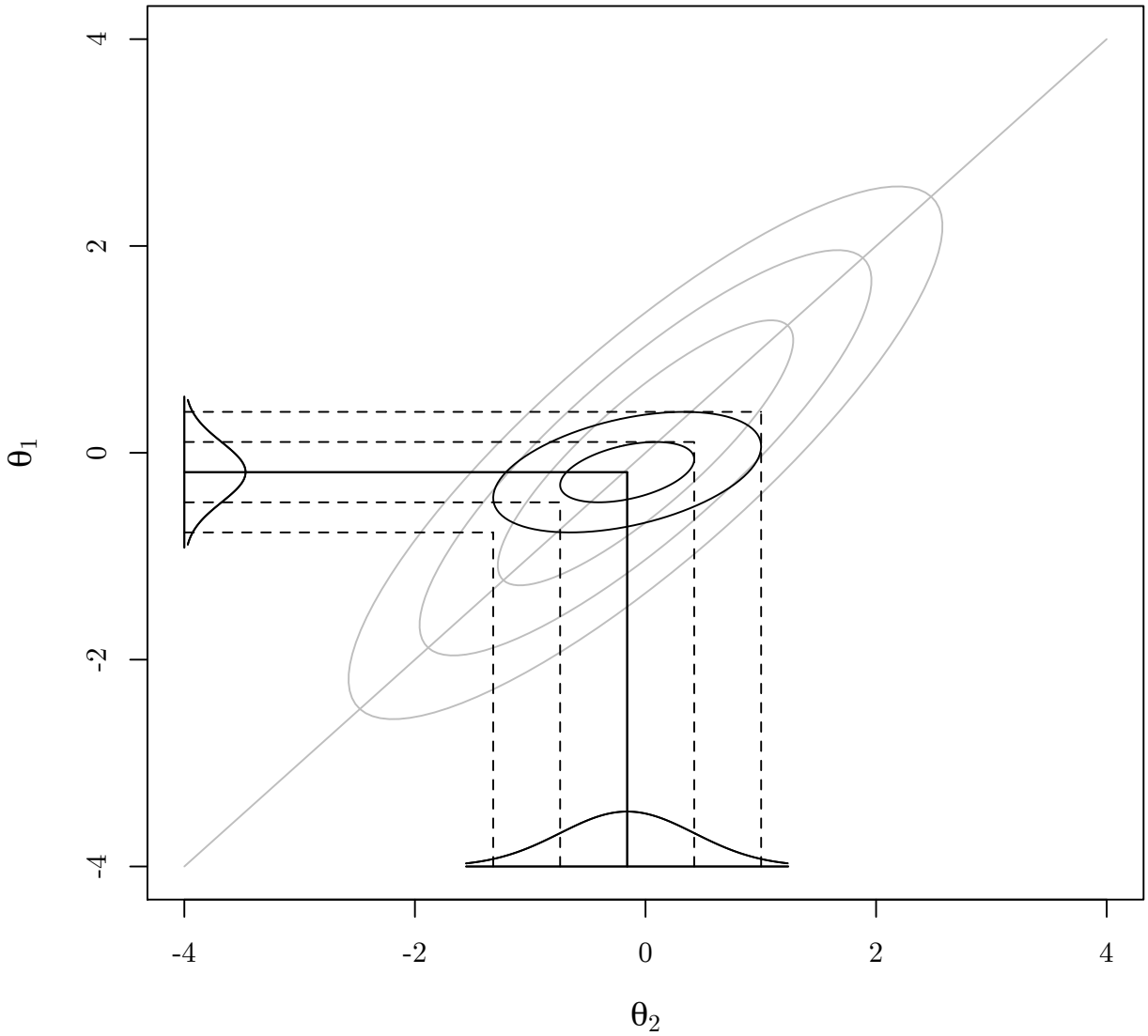

Figure 2

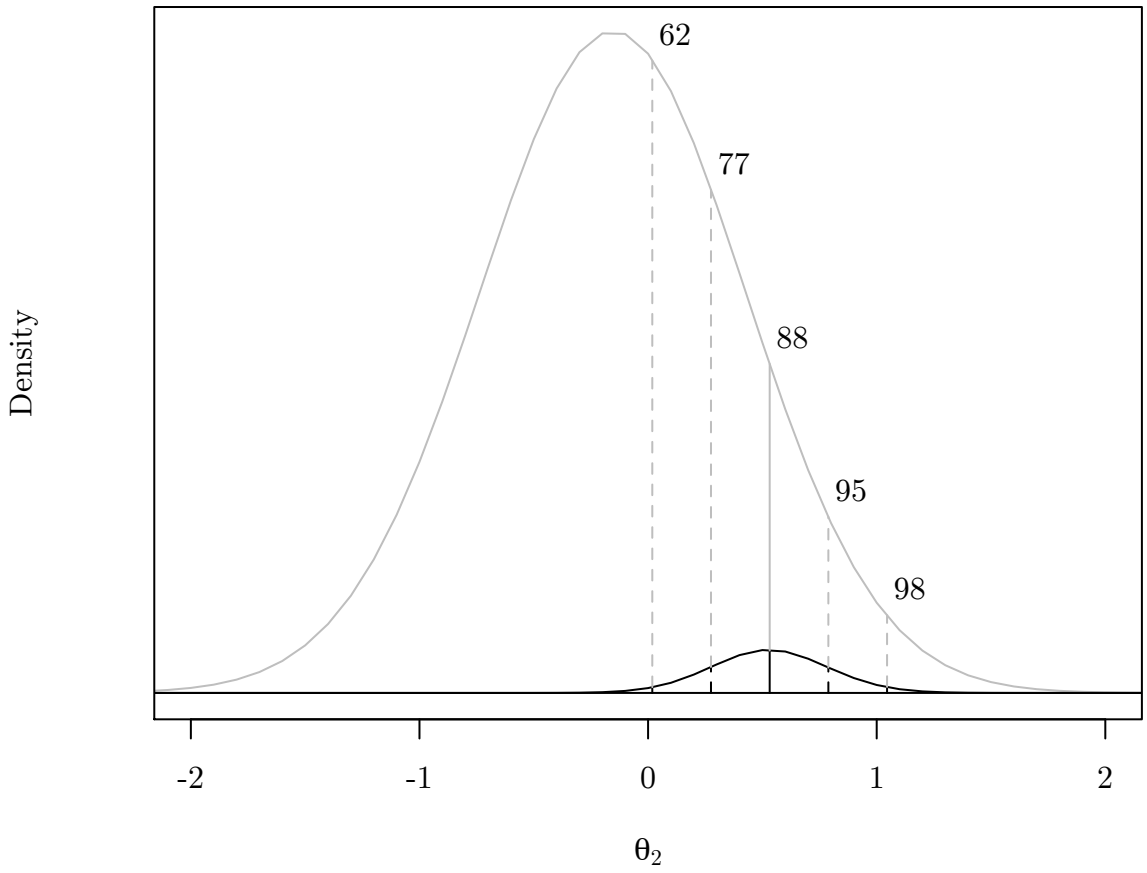

Figure 3

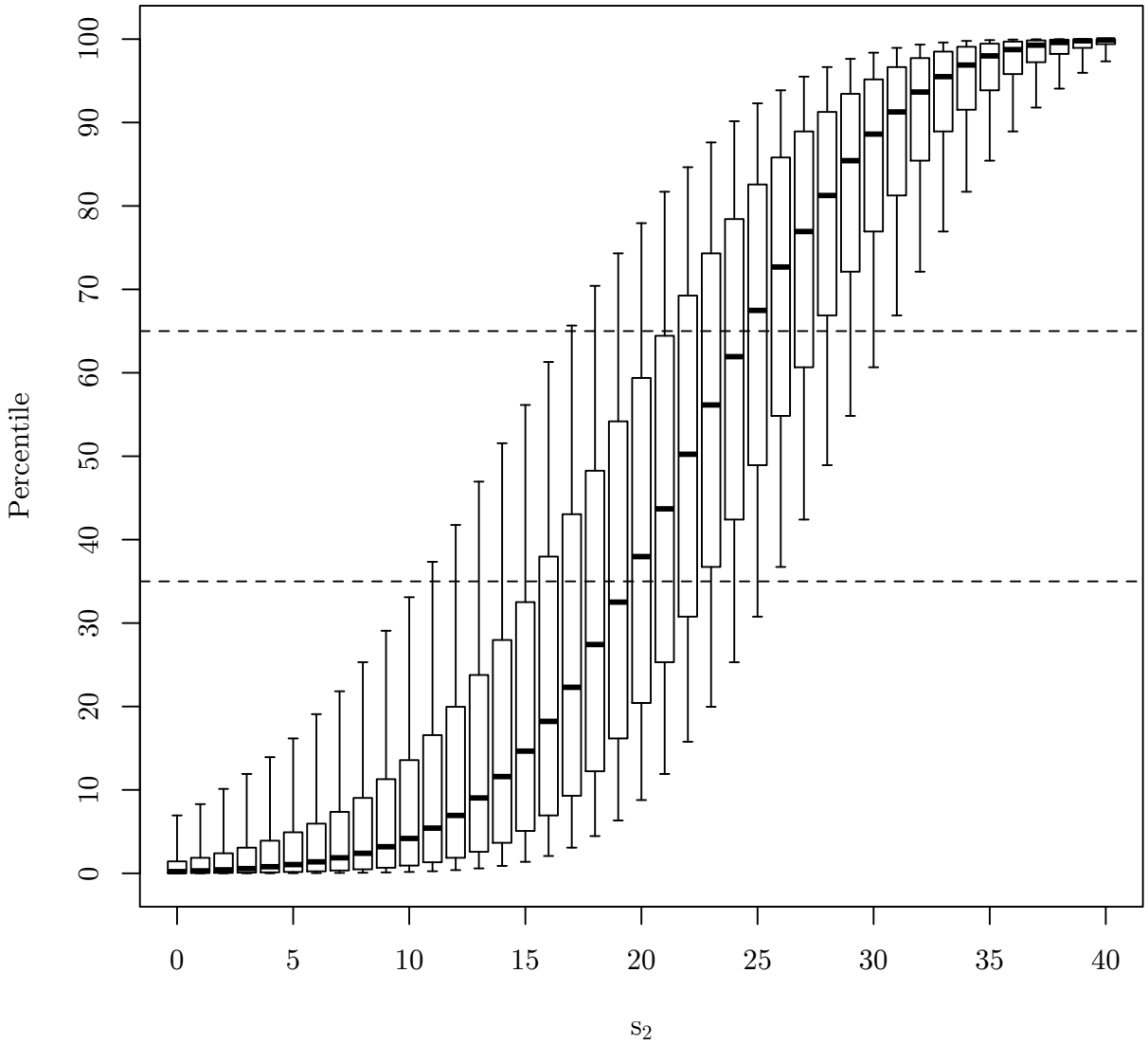
Percentile
0
10
20
30
40
50
60
70
80
90
100
0
5
10
15
20
25
30
35
40
s_2

Figure 4

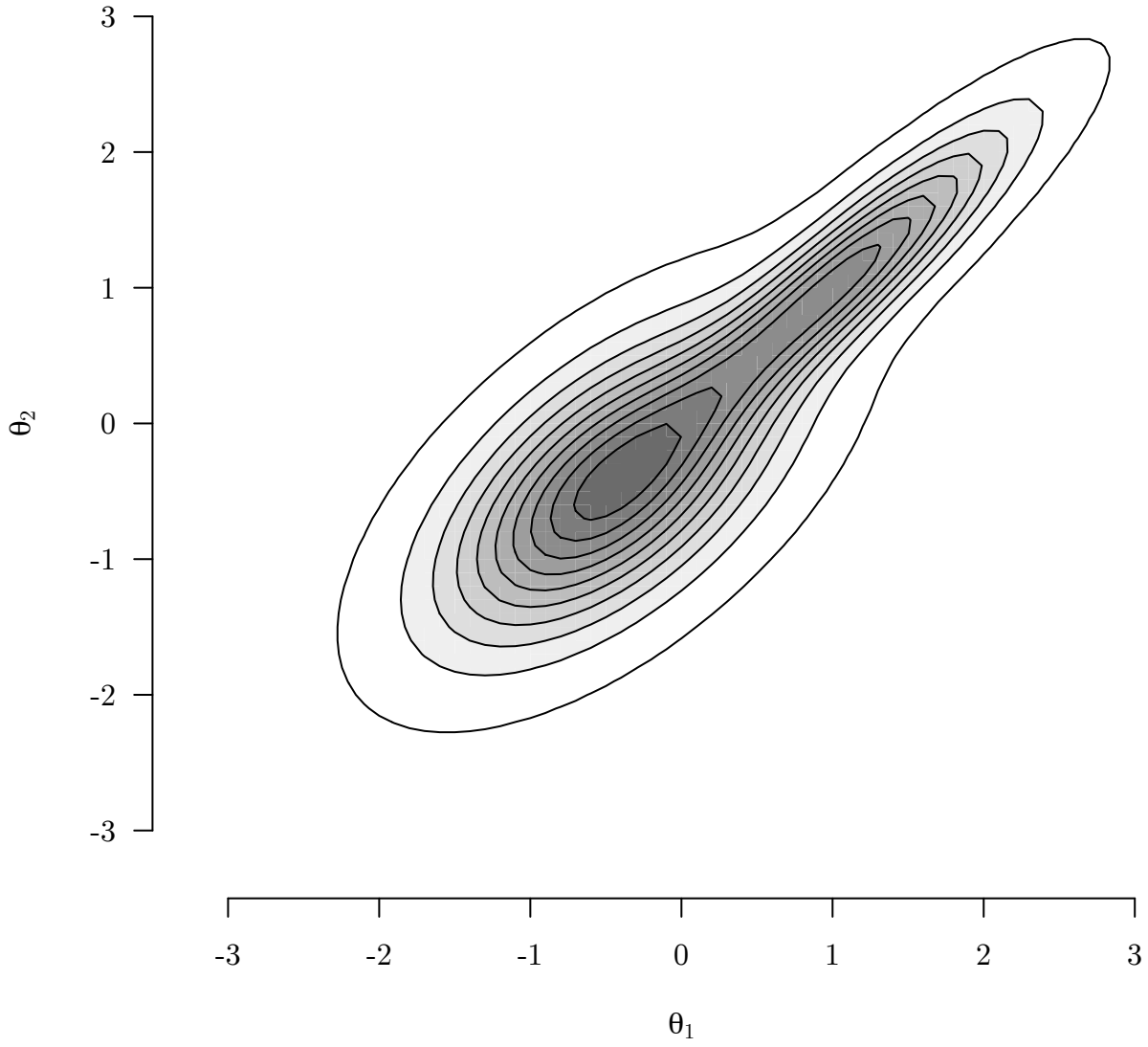

Figure 5

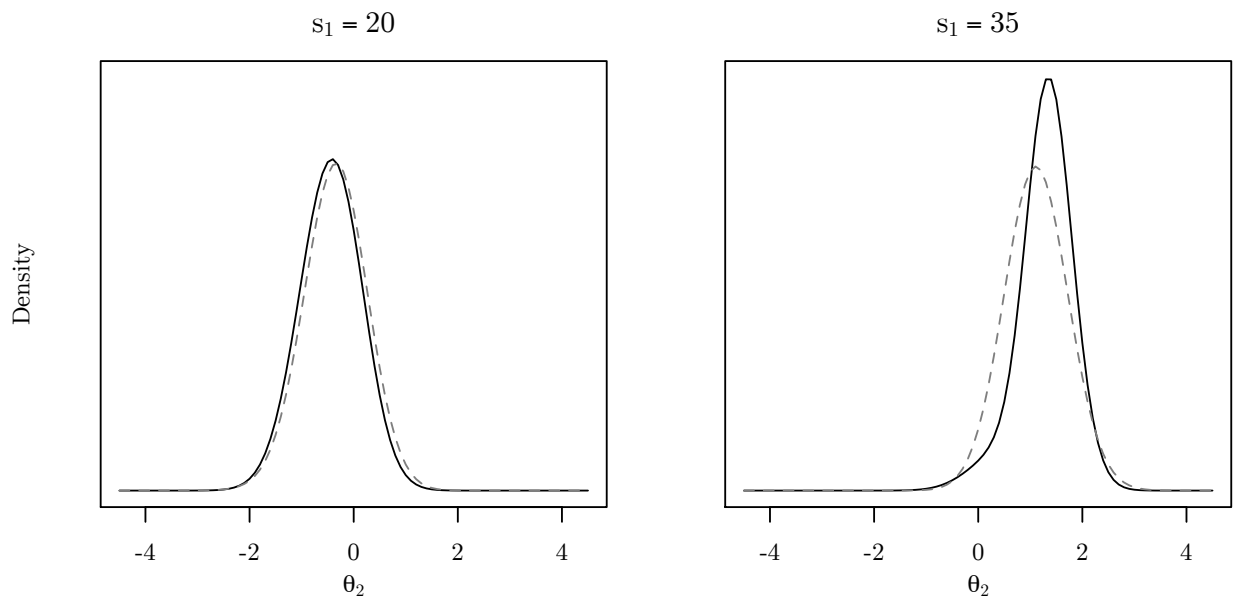

Figure 6

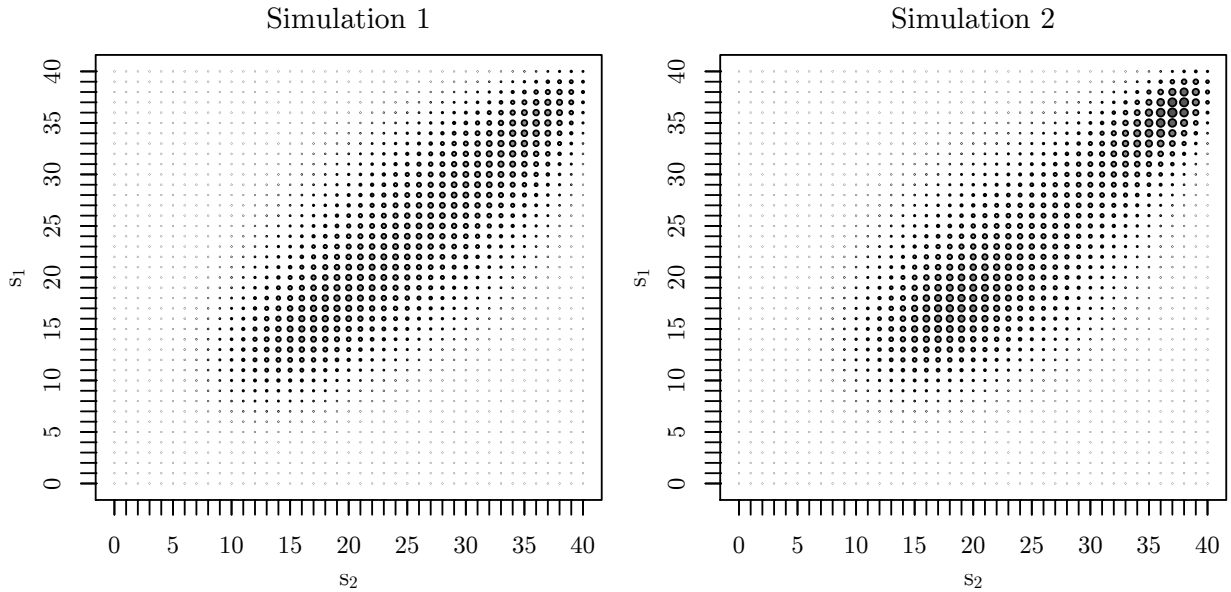

Figure 7

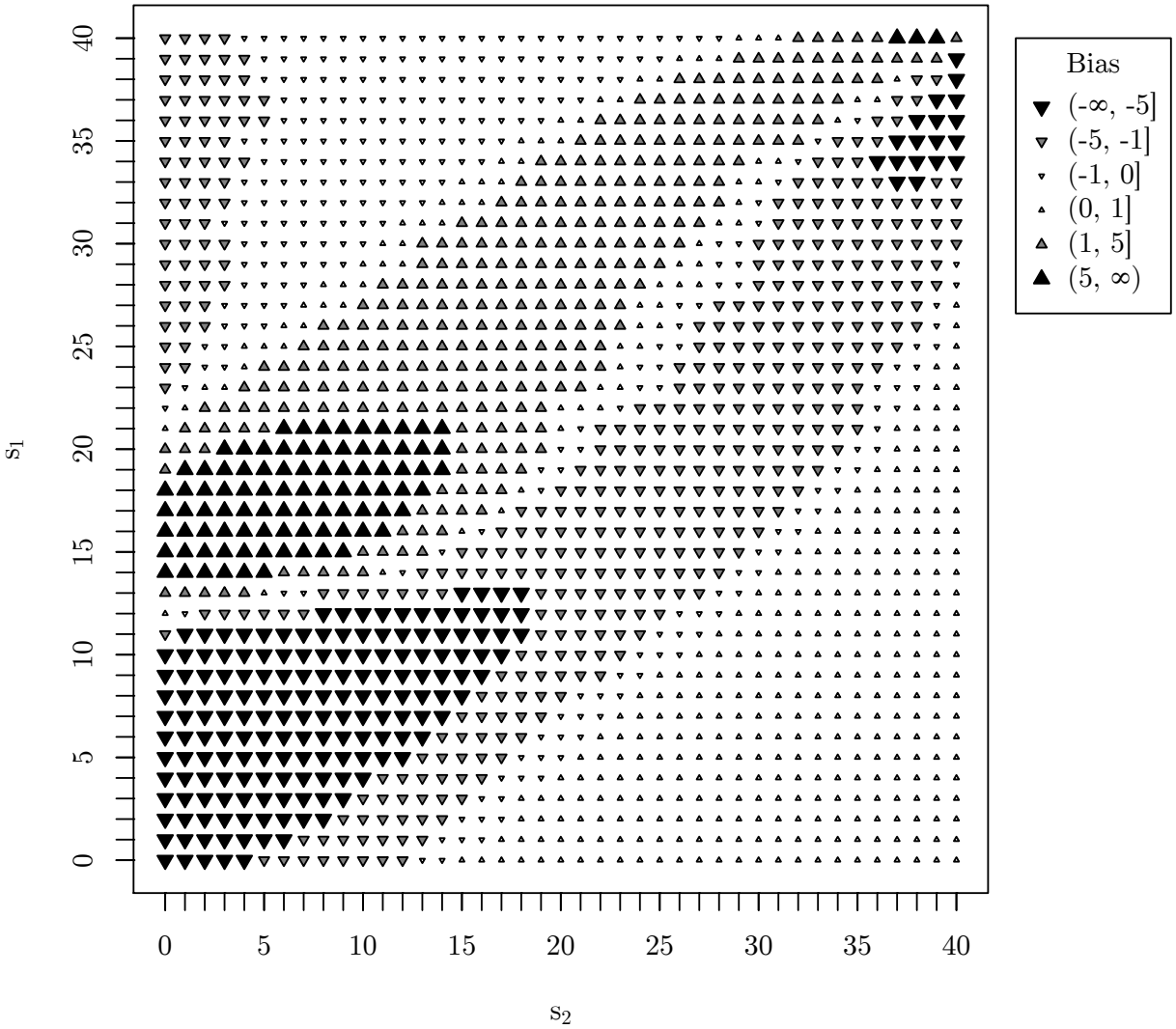

Figure 8

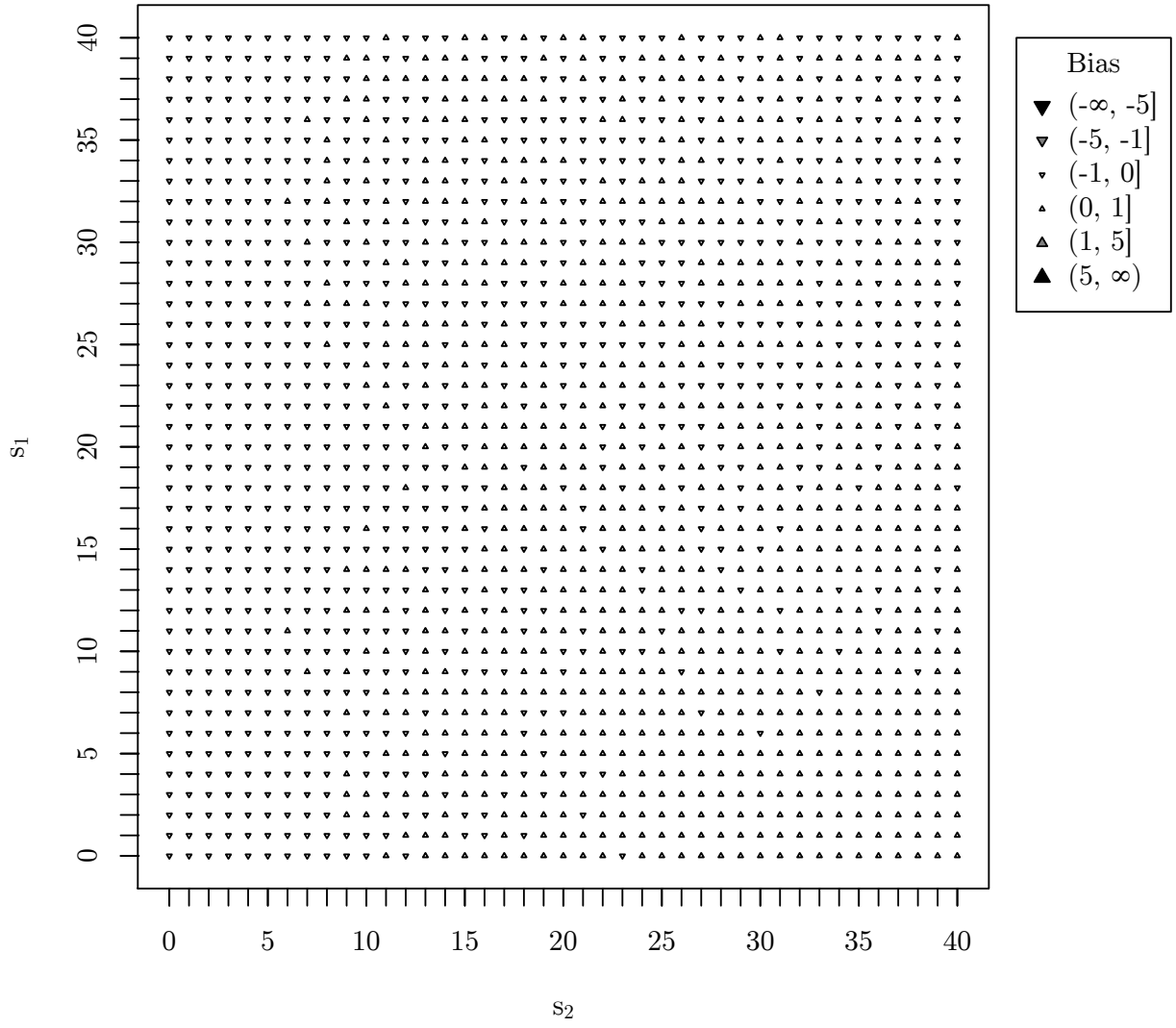

Figure 9

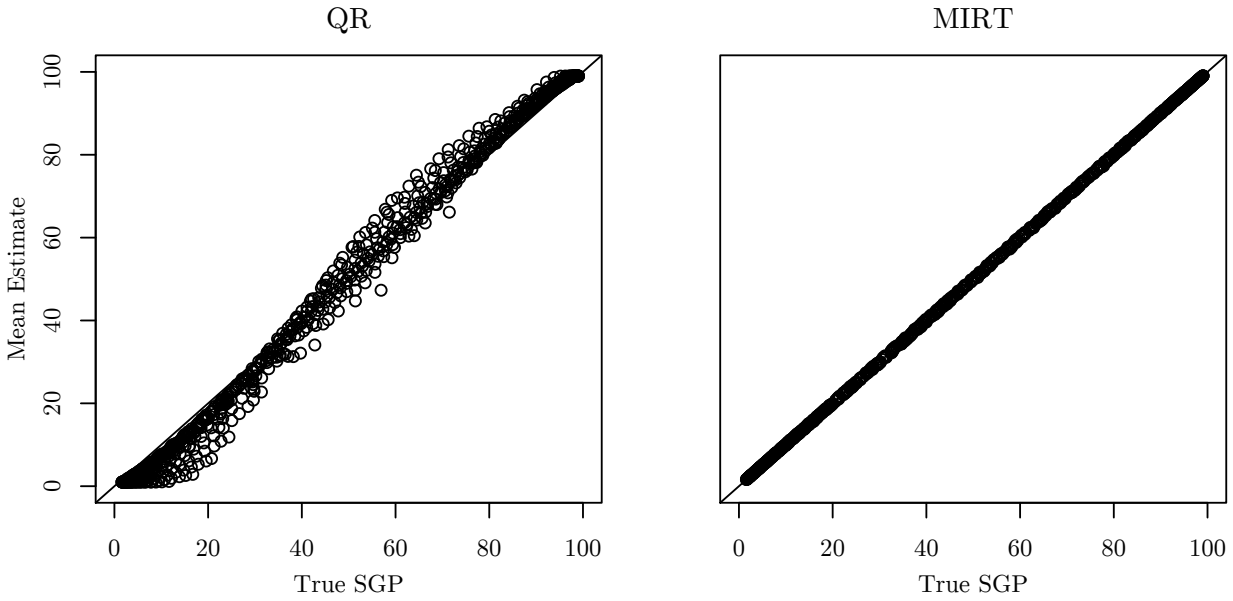

Figure 10

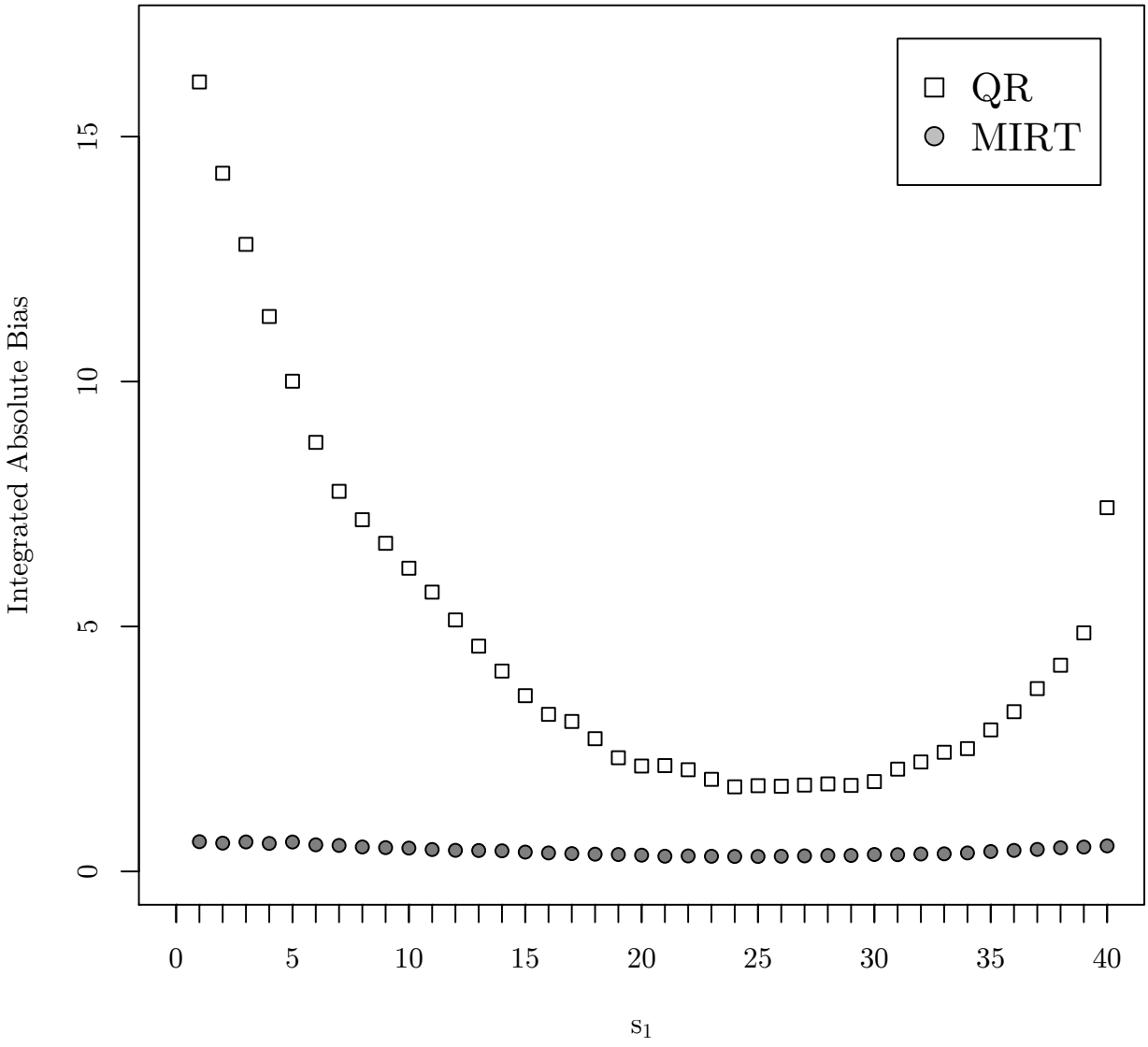
QR
MIRT
Integrated Absolute Bias
0
5
10
15
0
5
10
15
20
25
30
35
40
s_1

Figure 11

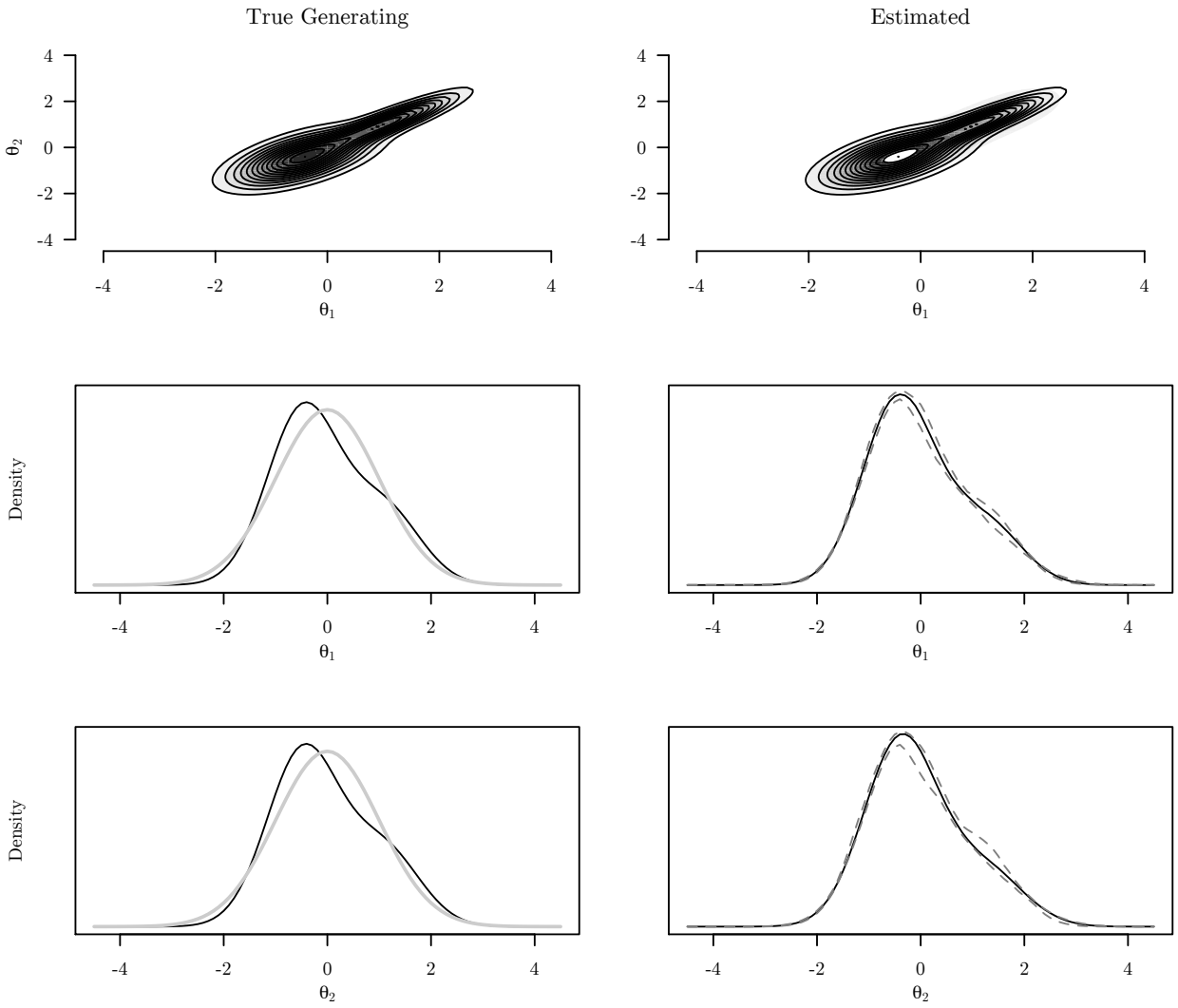

Figure 12

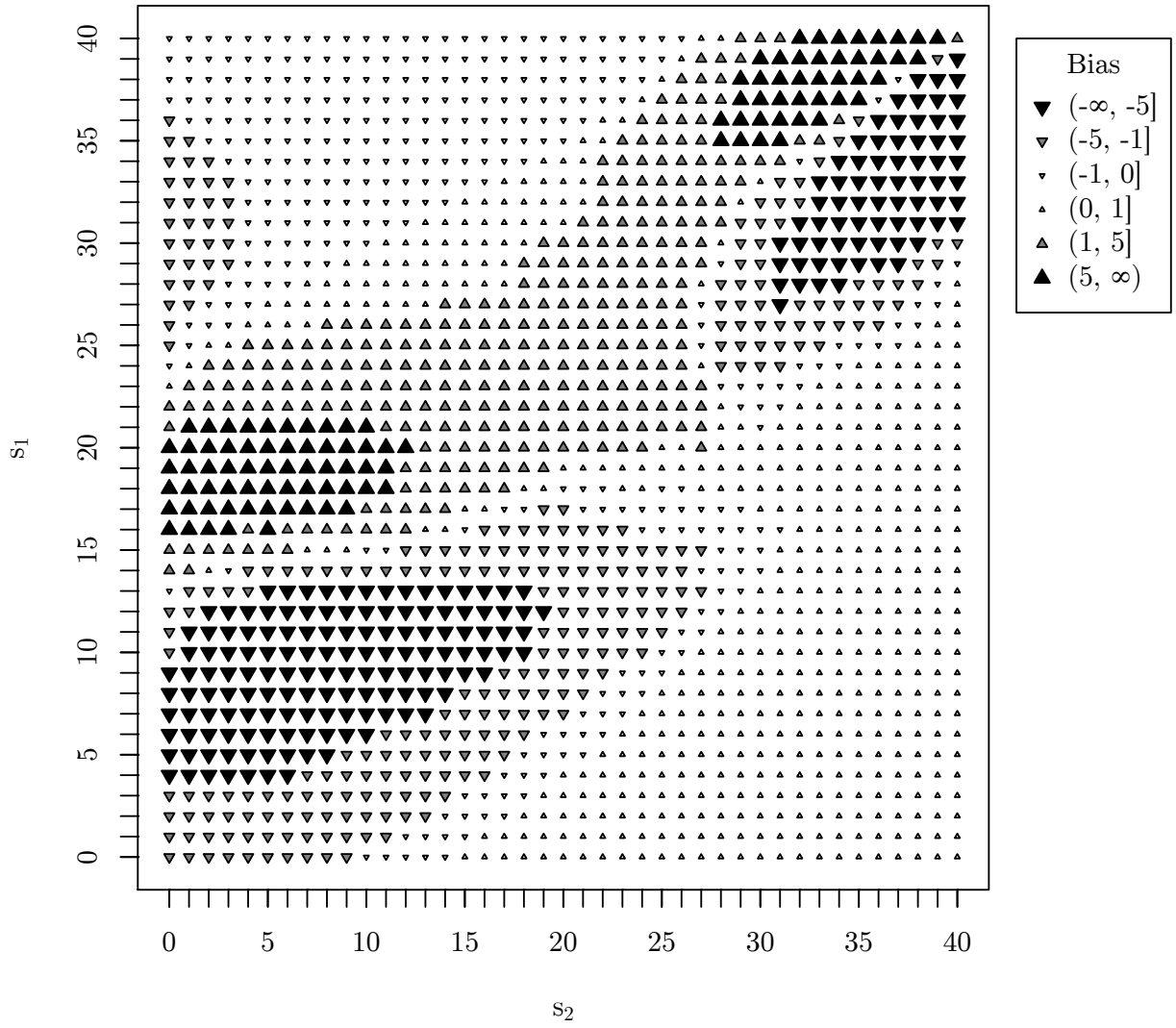

Figure 13

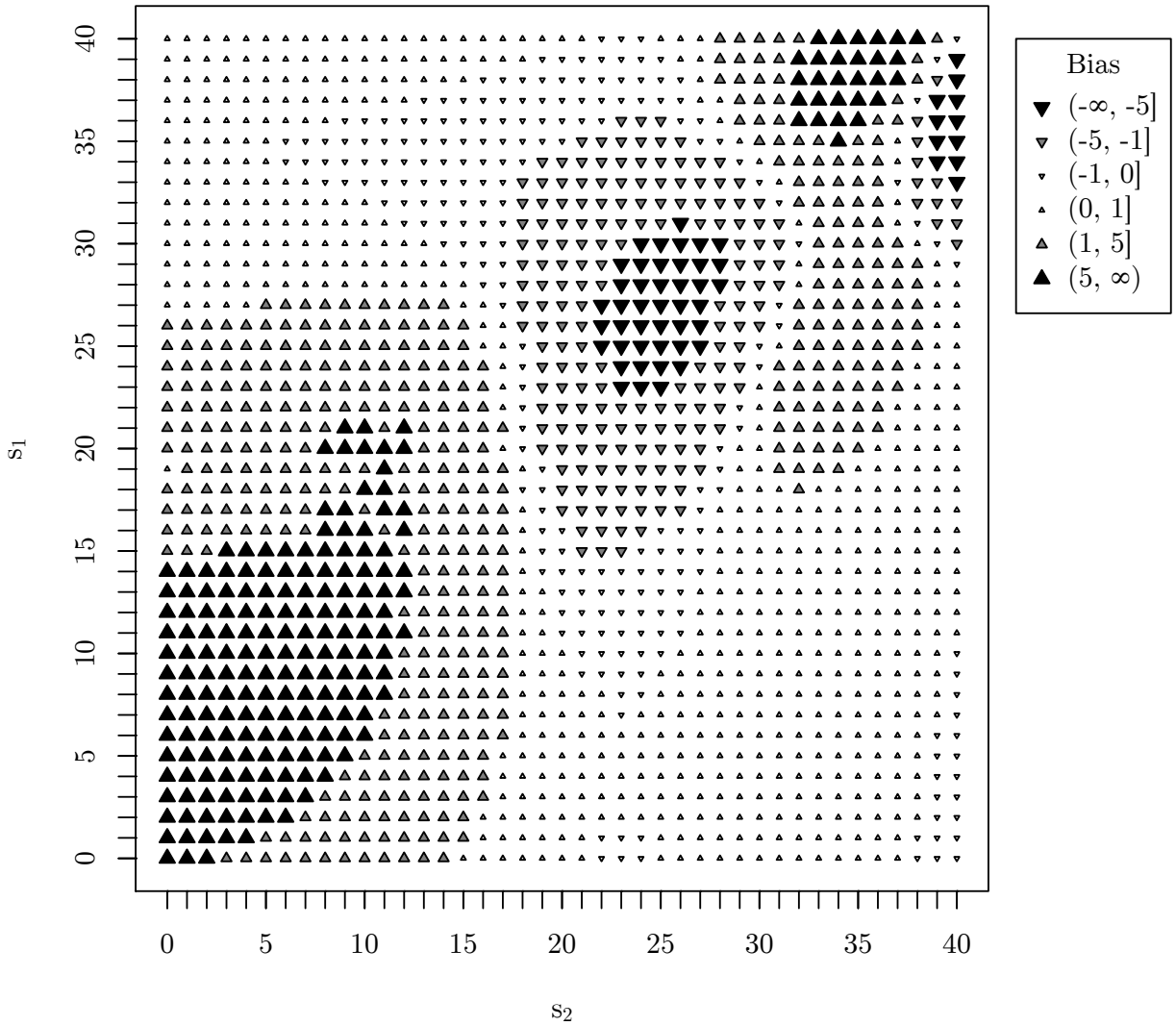

Figure 14

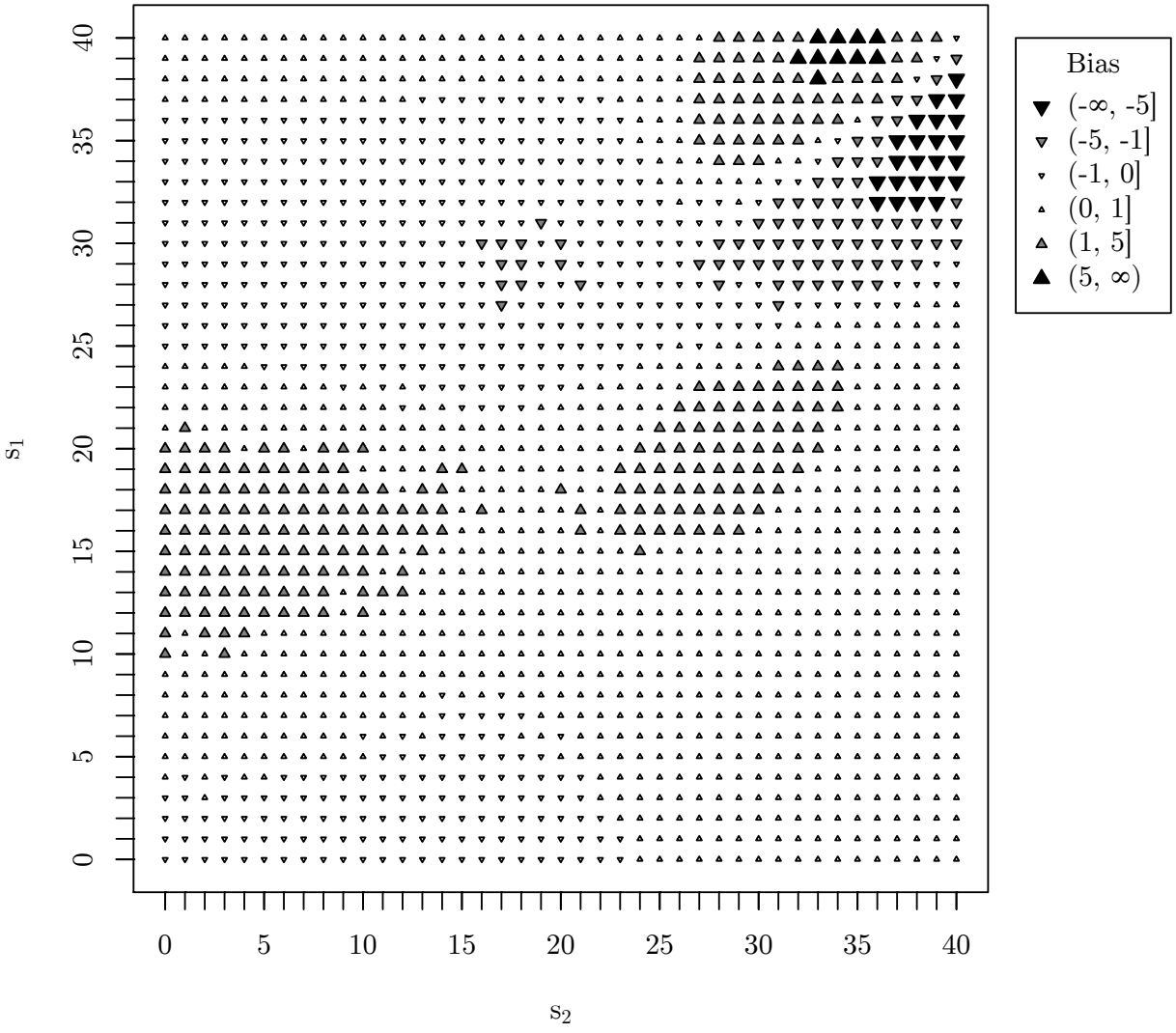

Figure 15

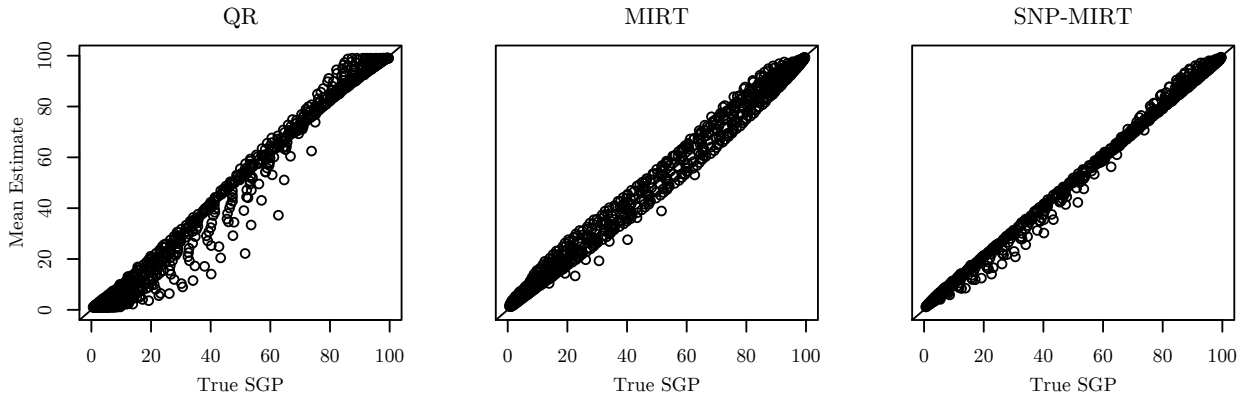
QR
MIRT
SNP-MIRT
Mean Estimate
True SGP
True SGP
True SGP
0
20
40
60
80
100

Figure 16

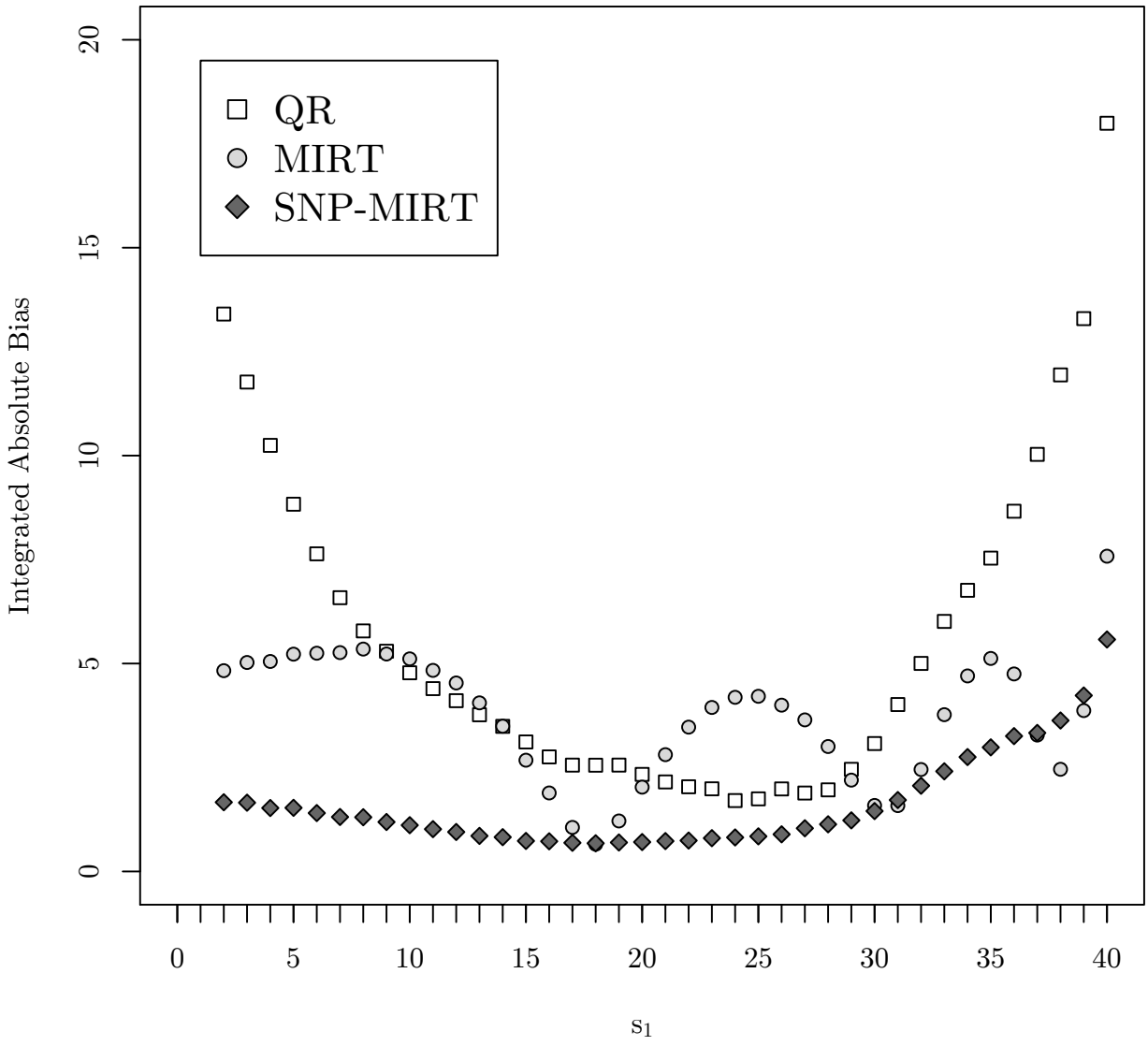
QR
MIRT
SNP-MIRT
Integrated Absolute Bias
0
5
10
15
20
0
5
10
15
20
25
30
35
40
s_1

Figure 17

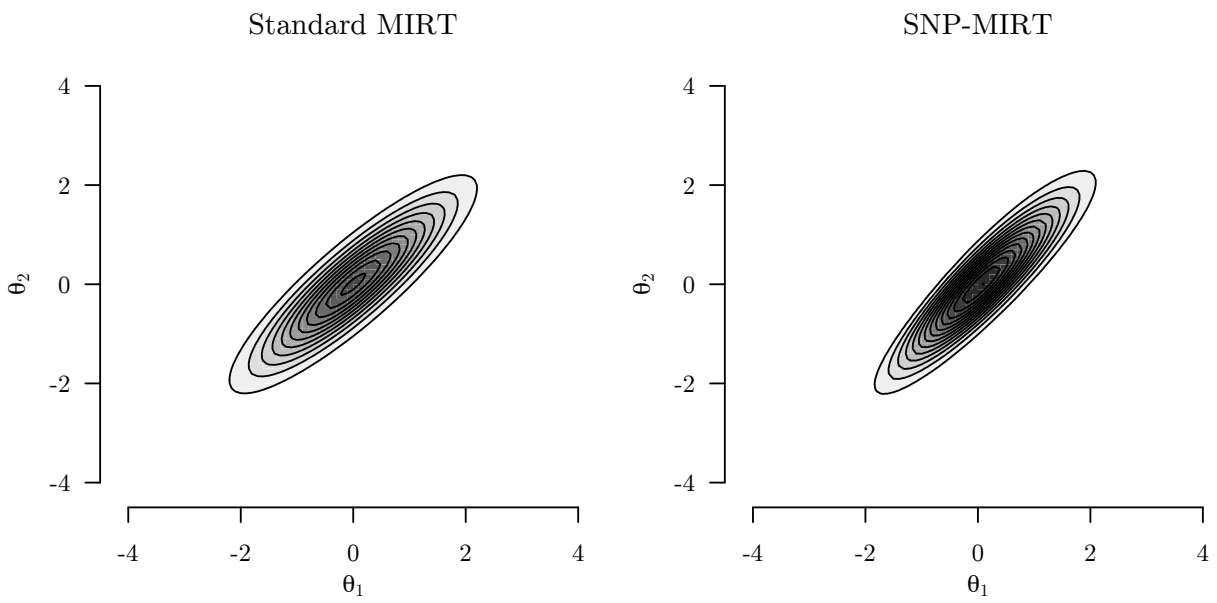

Figure 18

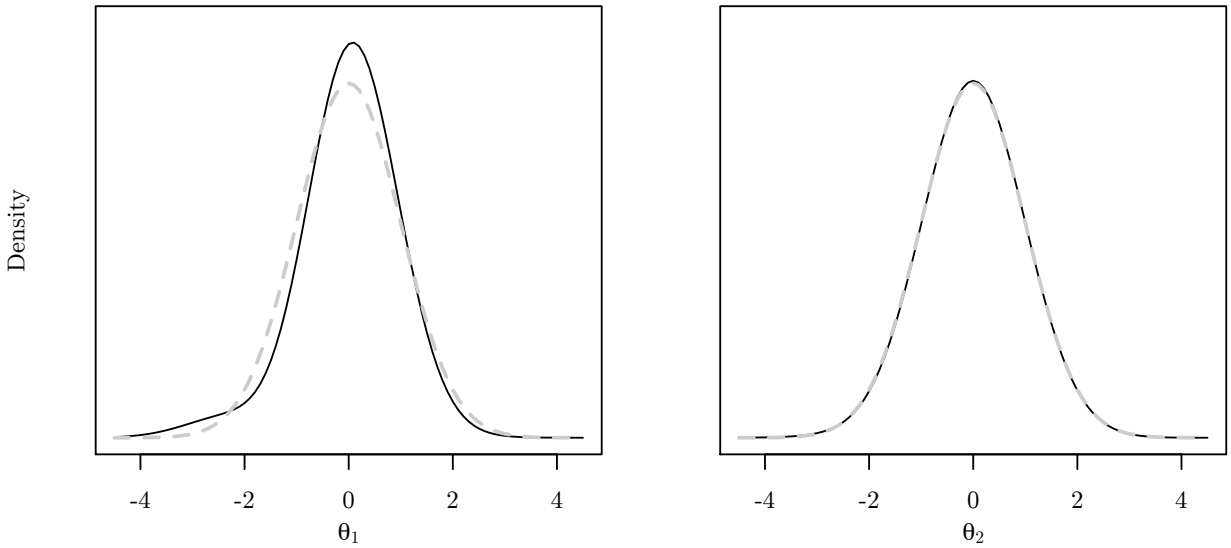

Figure 19