Two-Step Imputation Combining IRT and Deep-Learning Methods for Large-Scale Survey Assessments

Peter W. van Rijn

Cito, The Netherlands

Usama S. Ali

ETS Research Institute, USA

Qena University, Egypt

Priyadarshini Dwivedi

Indian Institute of Technology Kanpur, India

Abstract

Large-scale assessments (LSAs) rely on matrix-sampling designs that generate substantial planned missing item-response data, necessitating robust imputation methods to support valid population inference. In this study, we report on a two-step imputation method that combines item-response-theory (IRT) and deep-learning techniques to generate a complete item-response matrix that can be used for reporting such as market-basket approach. This method consist of a variation autoencoder (VAE) initiated with IRT-based imputations. Comparisons with standalone IRT and VAE imputation were made using three countries from PIRLS 2021 and TIMSS 2023. The VAE and IRT-VAE can outperform the IRT-based imputation when evaluated on RMSE and accuracy using cross-validation, but results varied somewhat across the data sets. Reliabilities of imputed total scores were generally higher with the VAE and IRT-VAE methods than with IRT, but correlations with official plausible values showed a mixed pattern. Finally, results related to the recovery of the distribution of scores on item pairs (local independence) indicated that VAE performed better than the other two methods on 5 out of 6 data sets. The discussed imputation approach provides a starting point for further research.

Keywords: missing data, multiple imputation, item response theory, variational autoencoders

Introduction

Many large-scale educational survey assessments (LSAs) employ test designs that lead to planned missing data. Plausible-value (PV) methodology is then commonly used which results in multiple imputations of latent proficiency variables (e.g., reading, mathematics). In LSAs, the goal generally is to estimate proficiency means for groups of respondents based on incomplete item-response data using item response theory (IRT) and background information using latent regression modeling (von Davier, Gonzalez, & Mislevy, 2009). PV methodology was developed from more general missing-data techniques (Rubin, 1987) and was first introduced in the United States' National Assessment of Educational Progress (NAEP; Mislevy, 1991). Over the past decades, PV methodology has become a key component of many well-known LSAs (Rutkowski, von Davier, & Rutkowski, 2013), including international studies such as PIRLS, TIMSS, PISA, and PIAAC, and national assessments such as the United States' NAEP and NALS, the Australian NAPLAN, and the German NEPS.

In this paper, however, we aim to develop an alternative to current PV methodology based on modern deep-learning techniques focusing on imputing the planned missing item-response data. The main reasoning is that the assumptions of the IRT and latent regression models used to generate PVs can be considered quite strong (von Davier, Sinharay, Oranje, & Beaton, 2006). That is, issues related to test design (e.g., adaptive testing), dimensionality, local dependence, and position effects can impact the validity of the IRT model (e.g., Robitzsch, 2022), while problems linked with congeniality, heteroskedasticity, and overfitting may affect the latent regression model.

Our research aims to overcome these issues by developing an alternative method that supplements the current PV methodology by employing modern deep-learning techniques. One existing alternative for PVs is the market-basket reporting, explored for NAEP (Mislevy, 1998) and PISA (Zwitser, Glaser, & Maris, 2017), where the market basket is defined as a specific collection of items considered to be representative for a domain (Mazzeo, Kulick, Tay-Lim, & Perie, 2006). If the complete item pool is taken as the market basket, the approach described by Zwitser et al. (2017) entails imputing item responses missing by design using IRT and then reporting the total score on observed and imputed responses.

Although this approach has benefits, its validity still relies on the assumptions of the IRT model. We propose to investigate whether recent advancements in deep-learning methods for imputing missing data (e.g., J. Yoon, Jordon, & van der Schaar, 2018; Lall & Robinson, 2022) offer the potential to reduce reliance on strong model assumptions and extensive background information, while still supporting the reporting of meaningful score distributions for student proficiency.

The research question is whether deep-imputation methods can serve as viable alternatives to current PV methodology for reporting LSA results. Recent studies suggest promising potential. Ali, van Rijn, and Dwivedi (2024) conducted simulations comparing IRT-based imputations with imputations by chained equations (Van Buuren & Groothuis-Oudshoorn, 2011), autoencoders (AE; Lall & Robinson, 2022), and generative adversarial networks (GAN; J. Yoon et al., 2018). Their findings showed that autoencoders produced results most similar to the IRT-based data-generating model. In contrast, GAN-based imputation seemed to have problems with the combination of discrete data (i.e., item responses) and high rates of missingness. Although several GAN-based imputation variants exist (e.g., S. Yoon & Sull, 2020; Wang, Chai, & Li, 2022), our present work focuses on AE-based imputation.

Our choice is also motivated by the fact that autoencoders have been used successfully as an estimation method for multidimensional IRT (MIRT) models (Converse, Curi, Oliveira, & Templin, 2021; Urban & Bauer, 2021). More recently, Veldkamp, Grasman, and Molenaar (2025) employed autoencoders for both imputing missing data and estimating MIRT models (see also, e.g., Freire & Curi, 2024). However, a key difference is that our goal is not to estimate MIRT models, but to enhance missing-data imputation to apply market-basket reporting. That is, we aim to investigate whether the AE-based imputation can be improved by developing a two-step approach. In the first step, starting values are generated using IRT-based imputation. In the second step, the AE-based method is applied starting from the results of the first IRT-based step. This new method is applied to recent LSA data in which substantial rates of missing item responses by design have to be accounted for. For example, in the TIMSS 2023 group-adaptive design (Mullis, Martin, & von Davier, 2021), students took 4 out of 28 blocks, which means that about 86% of the data

is missing by design. In PIRLS 2021 (Mullis & Martin, 2019), students were administered 2 out of 18 blocks, leading to roughly 89% planned missing data. In addition, due to the low-stakes nature of LSAs for students, there is often additional missing data resulting from students omitting and/or not reaching test items. Although the incidence of this latter kind of missing data is less frequent than the high percentages of planned missingness, it may be that these cannot be treated as missing completely at random (Rubin, 1976).

The outline of our paper is as follows. In the Method section, we start with a description of different types of missing data following Rubin's typology, followed by a brief description of imputation methods. Then, we describe the relevant IRT-based and deep-learning-based imputation methods, after which we develop our two-step imputation method. Then, we detail how the method is implemented and evaluated using real data from PIRLS 2021. In the Results section, the findings from this evaluation are presented. The paper concludes with a discussion, while additional materials can be found in the appendix.

Method

Missing Data

Imputation methods are essential for handling missing data in LSA to ensure the integrity of statistical inferences of scores. Although our main focus is on imputation of missing item responses, three types of missing LSA data can generally be differentiated Weirich et al. (2014): 1) Inherently missing data, 2) missing cognitive item responses, 3) and missing background information.

Latent variables are commonly used in modeling LSA data and are, by their nature, unobserved. For this reason, plausible values can be employed to represent student proficiency (Wu, 2005; von Davier et al., 2009). Plausible values are typically drawn from posterior distributions of latent proficiency variables given cognitive item responses and background information. In addition, latent construct variables are often used for certain scales in the student background questionnaire, such as attitudes in relation to the assessed proficiency or behaviors like bullying.

Missing cognitive item responses are also common in LSA as many more items are generally developed to cover the assessed domain(s) than can be administered in a single

test administration. For this reason, matrix sampling designs have often been employed in the past (Gonzalez & Rutkowski, 2010). More recently, complex variations of such designs have been developed to better match student proficiency with test difficulty, such as the group-adaptive designs used in PIRLS and TIMSS and the multistage adaptive testing (MSAT) designs used in PISA. Generally, due to the ratio of the size of the item pool (e.g., 240 items) and the feasible test length (e.g., 30 items), these designs lead to high levels of missing data by design (e.g., 80%). Furthermore, as noted, additional missing data occurs from respondents who omit items or do not finish. For computer-based assessments, the process data can inform decision-making on how to deal with such unplanned missing data (see, e.g., Ulitzsch, 2020).

The third type of missing LSA data concerns background information. For example, demographic information about students can be missing (e.g., gender or variables related to socioeconomic status) or questionnaire items related to attitudes and behaviors are skipped. Furthermore, certain background questions may not be administered by design. For example, a within-construct matrix sampling design was used in PISA 2022 to allow certain constructs to be covered by more items while limiting student burden (OECD, 2024, p. 93).

In addition to distinguishing the above types of missing data in LSA, it is crucial to understand the mechanism behind missing data to select appropriate imputation methods. Following the typology of Rubin (1976), there are three primary types of missing-data mechanisms, which can be described as follows: Missing Completely at Random (MCAR), the probability of missing data is unrelated to both observed and unobserved data. Missing at Random (MAR), the probability of missing data is related to observed data but not to the missing data itself. Missing Not at Random (MNAR), the probability of missing data is related to the missing data itself. Note that MCAR implies MAR.

Rubin (1976) specified sufficient conditions under which the missing-data mechanism can be ignored for likelihood or Bayesian inference. In addition to MCAR and MAR, the parameter of the missing-data process is distinct from the parameter of the data if their joint parameter space is the Cartesian product of their individual parameter spaces. Then, the ignorability principle holds if both MAR and distinctness hold, which generally simplifies

statistical inference.

In general, with matrix-sampling designs, the planned missing item-response data can be considered MCAR (Mislevy & Wu, 1996; Eggen & Verhelst, 2011). However, with designs such as the aforementioned group-adaptive and MSAT designs, this is not necessarily the case and it may mean that additional requirements are needed to satisfy MCAR or that the missing data are MAR (see, e.g., Jewsbury & van Rijn, 2020). As our main focus in this paper is on the imputation of planned missing cognitive item responses as done in Zwitser and Maris (2015), we mostly consider missing data as MCAR. However, we do investigate unplanned missing data in our application to PIRLS and TIMSS data (see the Results Section).

Multiple Imputation

Generally, the choice of imputation method depends on the type of missing data, the missing-data mechanism, and the specific goals of the analysis. Multiple imputations such as PVs require special treatment in subsequent statistical analyses to properly take into account the variance of the sampling error and the variance of the imputation error (Rubin, 1987). That is, statistical summaries such as means and standard deviations need to be calculated for each imputation and then averaged across the number of imputations. In addition, sampling and imputation error variances are estimated to determine the uncertainty in the reported statistic. Before we develop our imputation method using deep-learning techniques, we first describe more classical imputation methods.

Multiple Imputation by Chained Equations. Multiple imputation by chained equations (MICE) is a flexible method for handling missing data, especially when the data is MAR (Van Buuren & Groothuis-Oudshoorn, 2011; White, Royston, & Wood, 2011). It works by creating multiple complete datasets through iterative imputations using regression models. Each variable with missing values is imputed based on the others, considering the data type (e.g., linear regression for continuous, logistic regression for binary, etc.). MICE's key strength lies in its ability to handle mixed data types and complex missingness patterns, offering accurate statistical inferences by generating multiple datasets and pooling the results. This approach reduces bias, particularly in standard errors and confidence intervals.

However, it can be computationally demanding for large datasets, and its effectiveness depends on correctly specifying the underlying models. Despite challenges, MICE remains a widely used technique in a variety of fields. In LSA, it has been used not only to impute missing variables in student background data from TIMSS (Bouhlila & Sellaouti, 2013), but also in combination with such missing data and PVs (Weirich et al., 2014; Grund, Lüdtke, & Robitzsch, 2021).

Multiple Imputation with Item Response Theory. IRT is a powerful approach for handling missing data through imputation. The fundamental principle of IRT-based imputation is grounded in the probabilistic relationship between latent proficiency variables and item responses, allowing for imputation of latent variables and imputation of missing responses based on the estimated IRT model from observed item-response data. Unlike imputation techniques that rely on basic statistical assumptions (e.g., mean substitution or regression-based methods), IRT makes use of a mathematical model that describes the relation between the difficulty of items and the proficiency of students.

The probability of a correct item response under the two-parameter logistic (2PL; Birnbaum, 1968) model can be given by

$$p(X_j = 1|\theta) = \frac{\exp(\alpha_j\theta + \beta_j))}{1 + \exp(\alpha_j\theta + \beta_j))}, \tag{1}$$

where α_j is an item slope parameter, θ is the latent proficiency, and β_j is an item intercept parameter. Under local independence, the joint probability of a vector $\mathbf{x}$ containing m item responses is

$$p(\mathbf{x}|\theta) = \prod_{j=1}^{m} P(X_j|\theta). \tag{2}$$

If a density $f(\theta)$ is assumed for θ (e.g., the normal density), the marginal (or manifest) probability follows

$$p(\mathbf{x}) = \int p(\mathbf{x}|\theta)f(\theta)d\theta. \tag{3}$$

The posterior of θ is then

$$f(\theta|\mathbf{x}) = \frac{p(\mathbf{x}|\theta)f(\theta)}{\int p(\mathbf{x}|\theta)f(\theta)d\theta}, \tag{4}$$

which can be used to generate PVs for θ. First, the model parameters need to be estimated, where the most common method for the 2PL model is marginal maximum likelihood (Bock

& Aitkin, 1981). However, estimation of IRT models is relatively straightforward nowadays and can, for example, be performed by the mirt package in R (Chalmers, 2012). This package includes the capability of imputing missing item responses, which proceeds as follows. A sample is obtained from the posterior $f(\theta|\mathbf{x})$ (a plausible value) and the probability correct is calculated using Equation 1 with the sampled value of θ and the estimated item parameters. A draw from the binomial distribution with the calculated probability correct then results in an imputed item response. The process is repeated to obtain multiple imputations. A benefit of imputation with IRT is that marginal distributions of the imputed item data follow that of the observed data due to the IRT model parameters, provided that the model has acceptable fit.

In case of multidimensional proficiency $\boldsymbol{\theta}$, the integral in Equation 3 can become difficult to evaluate, especially at higher dimensions. In this case, variational inference can be used in which the posterior density $f(\boldsymbol{\theta}|\mathbf{x})$ is approximated by another density $g(\boldsymbol{\theta}|\mathbf{x})$ that is easier to work with while relying on an analytic approximation (see, e.g., Rijmen, Jeon, & Rabe-Hesketh, 2016; Cho, Wang, Zhang, & Xu, 2021). A similar approach is taken in using variational autoencoders to estimate MIRT models (Converse et al., 2021; Urban & Bauer, 2021).

In the context of reporting LSA results, Bock (1996) described domain-referenced scoring in which a large number of items is considered sufficient for the operational definition of an assessed domain. Expected scores on the collection of items can then be used for reporting. Mislevy (1998) considered summarizing LSA results in terms of a market basket of items with domain-referenced scoring as one possible form market-basket reporting. Zwitser et al. (2017) used this approach with PISA 2006 data and imputed item responses for planned missing data, thus creating a complete item-response matrix and the possibility to report total scores. They used the imputation approach described above using an IRT model. However, in principle, any imputation method can be employed to complete the item-response matrix, and this is why we investigate recent deep-learning-based imputation techniques focusing on autoencoders.

Multiple Imputation with Autoencoders. In general, autoencoders are artificial neural networks that can be used to efficiently represent unlabeled data. In the context

of multiple imputation, autoencoders offer a flexible approach compared to traditional statistical imputation methods. For example, Lall and Robinson (2022) describe a framework for multiple imputation by denoising autoencoders (MIDAS) to model complex relationships between variables and impute missing values. The key idea behind MIDAS is to learn a low-dimensional representation of the data through an autoencoder network, which consists of an encoder that maps the input data into a latent space, and a decoder that reconstructs the data back from this latent space. By training the autoencoder on incomplete data, the network learns to reconstruct the missing values during the imputation process, effectively capturing intricate dependencies among variables.

Autoencoders can be specified as follows. The model for a "forward pass" — or computation of output values given input data — through layer h of a neural network is:

$$\mathbf{y}^{(h)} = \sigma(\mathbf{W}^h \mathbf{y}^{(h-1)} + \mathbf{b}^{(h)}), \tag{5}$$

where $\mathbf{y}^{(h)}$ is a vector of outputs from hidden layer h ($\mathbf{y}^{(0)} = \mathbf{x}$ is the input), $\mathbf{W}^h$ is a matrix of weights connecting the nodes in layer $h-1$ with the nodes in layer h, $\mathbf{b}^{(h)}$ is a vector of biases for layer h, and σ is a nonlinear activation function.

The encoder maps the input $\mathbf{x}$ to a lower-dimensional latent representation $\boldsymbol{\theta}$ with B as the bottleneck layer (i.e., $\boldsymbol{\theta} = \mathbf{y}^{(B)}$) as follows:

$$\boldsymbol{\theta} = \sigma\left(\mathbf{W}^{(B)}\left[\ldots\left[\sigma\left(\mathbf{W}^{(2)}\left[\sigma\left(\mathbf{W}^{(1)}\mathbf{x} + \mathbf{b}^{(1)}\right)\right] + \mathbf{b}^{(2)}\right)\right]\ldots\right] + \mathbf{b}^{(B)}\right]. \tag{6}$$

It is important to note that the latent representation $\boldsymbol{\theta}$ in the autoencoder framework does not necessarily have an interpretation similar to the interpretation of $\boldsymbol{\theta}$ as a latent proficiency variable in the IRT framework. The decoder reconstructs $\mathbf{x}$ from $\mathbf{y}$, with the reconstructed vector $\mathbf{z}$:

$$\mathbf{z} = \Phi\left(\mathbf{W}^{(H)'}\left[\ldots\left[\sigma\left(\mathbf{W}^{(B+2)'}\left[\sigma\left(\mathbf{W}^{(B+1)'}\boldsymbol{\theta} + \mathbf{b}^{(B+1)'}\right)\right] + \mathbf{b}^{(B+2)'}\right)\right]\ldots\right] + \mathbf{b}^{(H)'}\right). \tag{7}$$

The network's parameters consist of weights and biases, which are trained by minimizing a loss function $L(\mathbf{x}, \mathbf{z})$ that quantifies the difference between actual and reconstructed outputs. Training proceeds through four steps, collectively referred to as an epoch (Lall & Robinson, 2022): (1) perform a forward pass through the network with the current parameters; (2) calculate the loss function L; (3) apply the chain rule to compute error gradients

for the weights in each layer (backpropagation); and (4) update the weights in the direction of the negative gradient for the subsequent forward pass. *Hyperparameters*, such as the network architecture and the number of training cycles per epoch, have to be set in advance.

One of the primary advantages of autoencoders is that it can handle large datasets with high-dimensional and non-linear relationships. Unlike traditional methods such as MICE or statistical approaches like KNN, autoencoders can capture non-linear interactions between features without requiring the specification of an explicit imputation model for each variable.

Variational autoencoders (VAEs) are a class of deep generative models that combine autoencoder architectures with variational inference (Kingma & Welling, 2013) and can address missing-data imputation (Lall & Robinson, 2022). VAEs learn a probabilistic latent representation of the data, enabling them to model uncertainty and generate plausible imputations for missing values.

In general, the input data for the i-th sample is denoted as $\mathbf{x}_i \in \mathbb{R}^d$, where d is the dimensionality. The VAE imputation framework consists of the following components:

- *Probabilistic Encoder*: The encoder maps the item responses $\mathbf{x}_i$ to parameters of the latent distribution:

$$\boldsymbol{\mu}_i = f_{\boldsymbol{\phi}}(\mathbf{x}_i), \tag{8}$$

$$\log \boldsymbol{\sigma}_i^2 = g_{\boldsymbol{\phi}}(\mathbf{x}_i), \tag{9}$$

  where $f_{\boldsymbol{\phi}} : \mathbb{R}^d \to \mathbb{R}^k$ and $g_{\boldsymbol{\phi}} : \mathbb{R}^d \to \mathbb{R}^k$ are neural networks parameterized by $\boldsymbol{\phi}$, producing the mean $\boldsymbol{\mu}_i$ and log-variance $\log \boldsymbol{\sigma}_i^2$ of the latent variable $\boldsymbol{\theta}_i \in \mathbb{R}^k$ ($k \ll d$).

- *Latent Sampling via Reparameterization*: The latent variable $\boldsymbol{\theta}_i$ is sampled using:

$$\boldsymbol{\theta}_i = \boldsymbol{\mu}_i + \boldsymbol{\sigma}_i \odot \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{10}$$

  where $\odot$ denotes element-wise multiplication.

- *Stochastic Decoder*: The decoder reconstructs response probabilities through:

$$\boldsymbol{\pi}_i = h_{\boldsymbol{\psi}}(\boldsymbol{\theta}_i), \tag{11}$$

where $h_{\boldsymbol{\psi}} : \mathbb{R}^k \to \mathbb{R}^d$ is a neural network parameterized by ψ. The function $h_{\boldsymbol{\psi}}$ outputs parameters of a specified distribution (e.g., mean and variance of normal distribution for continuous variables, success probability for binary variables).

- *Imputation-Specific Loss Function*: The model optimizes a modified Evidence Lower Bound (ELBO):

$$\mathcal{L}_{\text{ELBO}} = \underbrace{\mathbb{E}_{q_\phi(\boldsymbol{\theta}_i|\mathbf{x}_i)}\left[\sum_{j\in\mathcal{O}_i} \log p_\psi(x_{ij}|\boldsymbol{\theta}_i)\right]}_{\text{Reconstruction term}} - \beta \cdot \underbrace{D_{\text{KL}}\left(q_\phi(\boldsymbol{\theta}_i|\mathbf{x}_i)\|p(\boldsymbol{\theta}_i)\right)}_{\text{Regularization term}} \tag{12}$$

  where:

  - $\mathcal{O}_i$ denotes observed feature indices for sample i
  - $p_\psi(x_{ij}|\boldsymbol{\theta}_i)$ is the likelihood function (e.g., Bernoulli for binary data)
  - $p(\boldsymbol{\theta}_i) = \mathcal{N}(\mathbf{0}, \mathbf{I})$ is the latent prior
  - β controls the trade-off between reconstruction and regularization

- *Training Procedure*: The training comprises the following steps:

  1. **Masking**: Apply binary mask $\mathbf{m}_i \in \{0,1\}^d$ to exclude missing entries:

$$\mathbf{x}_i^{\text{obs}} = \mathbf{x}_i \odot \mathbf{m}_i \tag{13}$$

  2. **Forward Pass**:
     - Compute $\boldsymbol{\mu}_i$, $\boldsymbol{\sigma}_i^2$ via the encoder
     - Sample $\boldsymbol{\theta}_i$ using reparameterization
     - Reconstruct $\boldsymbol{\pi}_i$ via the decoder
  3. **Backward Pass**: Update parameters ϕ, ψ by maximizing the loss function $\mathcal{L}_{\text{ELBO}}$ via gradient ascent.

  During inference, missing entries x_{ij} $(j \notin \mathcal{O}_i)$ are estimated as:

$$\hat{x}_{ij} = \mathbb{E}_{q_\phi(\boldsymbol{\theta}_i|\mathbf{x}_i^{\text{obs}})}[p_\psi(x_{ij}|\boldsymbol{\theta}_i)] \tag{14}$$

Although Veldkamp et al. (2025) describe VAE-based methods for estimating MIRT models with incomplete data, our approach is different in the sense that our goal is not to

estimate a MIRT model. We merely aim to complete the item-response matrix and use a market-basket approach for reporting. Veldkamp et al. (2025) distinguish four methods for dealing with missing data in their application of VAE-methods for MIRT modeling. The first is referred to as input dropout, where missing values are replaced by zero during training. Note that this is not the same as scoring them as zero, since the loss function is calculated on observed and reconstructed data. The second method is to add the missing-data mask to the input, which doubles the number of input variables. The third method is a partial VAE. The fourth method is an imputation VAE where the missing values are replaced with expected responses given the current MIRT model parameters. Since IRT-based imputation performed reasonably well and our final goal is not to estimate a complex MIRT model, we created a two-step imputation approach where the initial IRT-based imputations are feeded into the VAE-based method (see the details in the setup of the Application Section).

Multiple Imputation with Generative Adversarial Networks. Another framework for multiple imputation using deep learning methods is based on generative adversarial networks (J. Yoon et al., 2018). It provides an efficient approach for dealing with missing data by leveraging an adversarial learning framework to generate imputations for missing data. The method consists of two neural networks: a generator G and a discriminator D. The generator is tasked with imputing the missing values, while the discriminator aims to distinguish between the truly observed values and the imputed values. The generator learns to improve its imputation quality by trying to fool the discriminator into believing that the imputed values are real, thus encouraging more accurate and realistic imputations. This adversarial training process continues until the discriminator can no longer differentiate between the real and imputed values, at which point the generator has learned to produce high-quality imputations.

One of the major strengths of this method is its flexibility and ability to handle complex, high-dimensional datasets with a variety of data types, including continuous and categorical variables. Unlike traditional imputation techniques, such as mean or regression-based imputations, GAN-based approach does not rely on strong assumptions about the underlying data distribution, making it highly adaptable to datasets with complex, non-linear relationships. Additionally, GAN-based method can handle various patterns of missingness,

including MCAR, MAR, and even MNAR, which are often challenging for other imputation methods. However, GANs can be difficult to train, especially on small or tabular datasets (Zhang, Zhang, & Zhao, 2023). Problems such as collapsing modes, vanishing gradients, or failing to learn the true data distribution can occur. More specifically for the LSA case, binary or categorical data with high missingness can exacerbate these issues, leading to cases where the discriminator overpowers the generator, resulting in poor imputations.

GAN-based imputation method operates by imputing missing data based on both the observed values and the hidden structure learned from the data. The imputation process is guided by a "hint" mechanism, where part of the observed data is provided as a hint to the discriminator, helping it learn more effectively. This mechanism improves the quality of the imputations and stabilizes the training process. Compared to other machine learning-based imputation techniques, such as deep autoencoders (used in methods like denoising autoencoders), GAN-based imputation framework allows it to generate more realistic and coherent imputations, particularly in datasets with complex inter-variable dependencies. Despite the high-quality imputations it produces, it requires careful tuning of hyperparameters, and training can be computationally intensive, especially for very large datasets or complex architectures.

Several important variants of GAN-based imputation have been developed to address specific imputation challenges:

- GAIN (Generative Adversarial Imputation Nets; J. Yoon et al., 2018): The foundational model introduces a key innovation: a hint mechanism. This mechanism provides the discriminator with additional information about which parts of the data were originally observed and which were imputed, preventing it from being trivially fooled and stabilizing training. This guides the generator to produce more accurate imputations that are coherent with the observed data.

- GAMIN (Generative Adversarial Multiple Imputation Nets; S. Yoon & Sull, 2020; Kim, Jeong, & Yoon, 2020): This extension modifies the GAIN architecture to formally support multiple imputation, creating several plausible imputed datasets to account for the uncertainty of the missing values. This is a critical advancement for valid statistical inference, as it allows analysts to pool results across datasets and ob-

tain accurate standard errors, a requirement often lacking in single-imputation deep learning methods.

- GAGIN (Generative Adversarial Gaussian Imputation Nets; Wang et al., 2022; Seo, Korbak, Zhou, Bhooshan, & Gweon, 2023): Designed to improve performance on mixed-type data (continuous and categorical), GAGIN uses a Gaussian mixture model in the generator's output layer. This provides a probabilistic approach to imputation that can better handle the different distributions of continuous variables and the discrete nature of categorical variables, a common challenge in survey data like LSAs.

Two-Step Imputation Combining IRT and Deep Learning

In our two-step approach, the first step consists of IRT-based imputation as described above using a basic IRT model such as the 2PL model. The aim here is to provide a reasonable starting point for the deep-learning based method.

The second step consisted of a VAE similar to the ones described by Veldkamp et al. (2025), but with the missing responses replaced by the IRT-based imputations from the first step. Initially, we also planned to include GAN-based methods. However, since the above GAN-based methods did not perform as expected (most likely due to combination of high missingness and discrete data), we limit our focus to the VAE method. The goal of this second step is to further enhance the IRT-based imputations by providing a flexible, data-driven approach for imputing missing data.

Research Quesions

Our main research objectives are

- To compare the performance of the two-step approach using IRT and deep-learning imputation with each method individually using real LSA data.
- To investigate how the resulting scores compare to current official PVs of student proficiency.
- To explore how the scores can be used to evaluate model fit.

Application

Setup

To study the performance of the imputation methods, we make use of real LSA data. Specifically, we used the data from PIRLS 2021 from three countries: Brazil, England, and Türkiye. In addition, data from the same three countries was taken from TIMSS 2023. These countries were selected as their performance in these studies has substantial variation. To ease the analysis, polytomous items were dichotomized. In addition, to ensure stable cross-validation results, we selected students who responded to at least 20 items. Finally, sampling weights were not used as these are not needed for our purpose of evaluation the imputation methods.

For each imputation method, ten-fold cross-validation was applied to calculate the root mean squared error (RMSE) and accuracy statistics. For each fold k, we obtain for 10% of the data both observed and imputed values. If ten imputations are created, this means that ten data sets are obtained across all folds and imputations for which we have both observed and imputed values. The RMSE for the k-th fold and l-th imputation can be defined as

$$\text{RMSE}(k,l) = \sqrt{\frac{\sum_{i,j}\left[(1-M_{ij}^{(k)})X_{ij} - (1-M_{ij}^{(k)})X_{ij}^{(k,l,\text{imp})}\right]^2}{\sum_{i,j}(1-M_{ij}^{(k)})}}, \tag{15}$$

where $M_{ij}^{(k)}$ is the indicator for fold k, X_{ij} is the original data, and $X_{ij}^{(k,l,\text{imp})}$ is the imputed data. Note here that the original missing data would already be left out. The overall RMSE is then obtained by $\sqrt{\frac{\sum_{k,l}\text{RMSE}_{(k,l)}}{KL}}$. The accuracy statistic is merely the percentage matching between observed and imputed item responses. The cross-validation ensures that the results are not affected by a specific train-test split and provides more robust performance estimates for each model. The cross-validation is not straightforward, as students answer different numbers of items, and thus different splits are required. As additional check, we calculate the correlations of the total score using the imputed complete data across the imputations and with the official reported PVs. For this last evaluation, a single training run had to be performed only instead of using ten folds to calculate the RMSE.

Table 1

Descriptives for selected PIRLS 2021 and TIMSS 2023 data.

				Percent Missing		Official Score		
Study	Country	*N*	Number of Items	Planned	Unplanned	Reading	Math	Science
PIRLS 2021	Brazil	3,972	296	88.9	2.1	419	-	-
	England	4,036	296	88.9	0.5	558	-	-
	Türkiye	5,914	296	88.9	0.5	496	-	-
TIMSS 2023	Brazil	21,924	587	85.7	2.7	-	400	425
	England	4,083	602	85.7	0.7	-	552	556
	Türkiye	4,541	603	85.7	0.3	-	553	570

All models were implemented in Python. The 2PL model was estimated with the MML algorithm implemented in IRTorch (Wallmark, 2024) using 61 Gauss-Hermite quadrature points and the VAE models were implemented in PyTorch (Paszke et al., 2019). The details of the VAE implementation are as follows. The dimensionality of the hidden layer was fixed at 128. The dimensionality of the bottleneck layer was fixed at 20. The batch size was set at 128. Additional dropout was created in the hidden layer of the encoder to create more robust features. Models were trained using the Adam optimizer with a learning rate of .001 (the default). Training was stopped when the loss function did not decrease more than 10 epochs.

Results

Table 1 shows the sample sizes, number of items, percentages of planned and unplanned missing data, and mean official scale scores of the six selected data sets from PIRLS 2021 (Grade 4) and TIMSS 2023 (Grade 4). It can be seen that the Brazil data have higher proportions of unplanned missing data than the data from England and Türkiye. Also, many more students in Brazil participated in TIMSS 2023.

Table 2 presents the mean and standard deviation (SD) of the RMSE and accuracy statistics for the three imputation methods applied to the PIRLS 2021 and TIMSS 2021 data. For the PIRLS 2021 data, the IRT-based method outperformed the VAE-based approach, achieving a lower RMSE and higher accuracy, although the differences varied between the three countries. Furthermore, the two-step IRT-VAE method yielded better

Table 2

Mean and SD of RMSE and accuracy of different imputation methods for selected PIRLS 2021 and TIMSS 2023 data.

		IRT		VAE		IRT-VAE	
Study	Country	RMSE (SD)	Accuracy (SD)	RMSE (SD)	Accuracy (SD)	RMSE (SD)	Accuracy (SD)
PIRLS 2021	Brazil	.5448 (.0028)	.7032 (.0029)	.5559 (.0032)	.6910 (.0035)	.5440 (.0023)	.7040 (.0025)
	England	.5037 (.0022)	.7462 (.0022)	.5497 (.0027)	.6978 (.0029)	.4932 (.0020)	.7568 (.0029)
	Türkiye	.5418 (.0013)	.7065 (.0015)	.5688 (.0040)	.6764 (.0046)	.5319 (.0024)	.7171 (.0026)
TIMSS 2023	Brazil	.5557 (.0005)	.6912 (.0005)	.5235 (.0010)	.7260 (.0010)	.5487 (.0007)	.6989 (.0008)
	England	.5332 (.0011)	.7157 (.0012)	.5072 (.0018)	.7428 (.0018)	.5174 (.0030)	.7322 (.0031)
	Türkiye	.5225 (.0012)	.7270 (.0013)	.4972 (.0021)	.7527 (.0021)	.5118 (.0038)	.7381 (.0039)

results than the IRT-based imputation alone. For the TIMSS 2023 data, however, the VAE-based method outperformed the IRT-based approach, while the IRT-VAE method performed in between these methods. Notably, RMSE and accuracy values for the Brazil data seem to be slightly worse than those for the other countries. This may be due to its lower overall performance, its larger proportion of unplanned missing data (Table 1), or a combination of both.

Table 3 displays the reliability of the imputed total score and the mean correlation with the official PVs for the selected PIRLS 2021 and TIMSS 2023 data sets for each of the three imputation methods. For the TIMSS 2023 data, we show the results for the official mathematics PVs only, as the results for the science PVs are not essentially different. The reliability is calculated as the mean of the correlations between all unique pairs of imputed total scores. That is, for 10 imputations, 5 unique pairs are defined (i.e., 1-2, 3-4, etc.). The reliability of the scores is the highest for the VAE method followed by the IRT-VAE and IRT methods. A similar pattern is seen for the mean correlation with the official PVs, although the values and the differences are somewhat smaller. One noticeable difference is seen between the IRT and VAE methods for the England PIRLS 2021 data: .827 vs. .854. It is interesting to see that VAE and IRT-VAE methods can outperform the IRT method on these statistics, even though the official PVs are also based on IRT modeling (although using international data from many countries).

Results related to the recovery of person-level and item-level statistics can be found in Appendix A, which includes an indicatation of position effects. As a final exploration of

Table 3

Reliability of imputed total score and mean correlation with official PV for the three imputation methods and selected PIRLS 2021 and TIMSS 2023 data.

		IRT		VAE		IRT-VAE	
Study	Country	Reliability	Correlation	Reliability	Correlation	Reliability	Correlation
PIRLS 2021	Brazil	.951	.895	.983	.902	.976	.896
	England	.926	.827	.978	.854	.973	.844
	Türkiye	.939	.878	.978	.888	.976	.886
TIMSS 2023	Brazil	.953	.818	.973	.821	.972	.820
	England	.968	.878	.978	.863	.979	.876
	Türkiye	.974	.904	.980	.896	.979	.905

model fit, we compared observed and expected frequencies for each score combination on a pair of items. In the 2PL IRT model, local independence is assumed and violations of this assumption can be indicated by discrepancies between these frequencies. Using the imputed values for which observed data is also available, expected frequencies are easily calculated so that a Pearson X^2 statistic can be computed on the distribution for each item pair. Table 4 shows the results of this exploration, where the mean is calculated across imputation. The X^2 was computed only when both the observed and imputed frequency was at least five for each of the four score combinations. This is indicated by the mean number of valid item pairs. The SD shown in the table is that of the mean X^2. It can be seen that the VAE method, except for the TIMSS 2023 Brazil data, returns substantially lower values than the IRT and IRT-VAE method. Although we do not claim that the statistic follows a χ^2 distribution, the result can be seen as an indication that the VAE method does a better job in capturing conditional dependence.

Discussion

In this study, we developed and evaluated a two-step imputation method for handling planned missing item-response data in LSAs. The approach integrates the aspects of IRT and VAE, first leveraging an IRT model to generate initial imputations, which are then refined by a VAE to capture potentially more complex, non-linear patterns in the data.

Our method was tested on real-world data from the PIRLS 2021 and TIMSS 2023

Table 4

Mean number of valid item pairs (SD) and mean of Pearson χ^2 statistic (SD) for observed and imputed item pair score frequencies of three imputation methods for selected PIRLS 2021 and TIMSS 2023 data.

Study	Country	Mean # Item Pairs (SD)	IRT Mean X^2 (SD)	VAE Mean X^2 (SD)	IRTVAE Mean X^2 (SD)
PIRLS 2021	Brazil	5,452 (30)	4.38 (0.14)	3.45 (0.07)	4.75 (0.18)
	England	4,454 (18)	4.33 (0.13)	2.01 (0.07)	4.11 (0.17)
	Türkiye	5,872 (23)	4.61 (0.16)	2.74 (0.07)	4.52 (0.14)
TIMSS 2023	Brazil	34,705 (36)	11.34 (0.49)	15.95 (0.99)	11.46 (0.54)
	England	29,118 (85)	4.93 (0.13)	3.84 (0.09)	4.66 (0.11)
	Türkiye	29,395 (64)	4.80 (0.08)	4.28 (0.09)	5.13 (0.10)

assessments and compared against standalone IRT and VAE imputation techniques using tenfold cross-validation. For the PIRLS 2021 data, it was found that the two-step IRT-VAE method achieved superior performance when evaluated on RMSE and imputation accuracy than either method used in isolation. This suggests that the sequential application of these models can successfully compensate for their individual limitations, with IRT providing a psychometrically sound foundation and VAE enhancing the plausibility of the imputed values. However, for the TIMSS 2023 data, where the item pool was roughly twice as large, the results were best in terms of RMSE and imputation accuracy with the standalone VAE method, although the IRT-VAE method did do better than the IRT method. These combined findings suggest that the VAE and IRT-VAE methods can perform well but that comparisons are required for a proper evaluation. Furthermore, further enhancing the architecture of these methods (e.g., adding additional layers) likely leads to improved performance on RMSE and accuracy statistics.

The additional results on reliability of imputed total scores and correlations with official PVS indicate that the VAE and IRT-VAE methods generally produce higher reliabilities than the IRT method, while the pattern of the correlations with the official PVs is more mixed. The exploratory results on local dependence and position effects seen in the person means at the booklet level (see appendix A) can be taken as a starting point for further investigation.

Our findings must be interpreted within the context of the complex sampling de-

sign inherent to LSAs like PIRLS and TIMSS. These studies employ a stratified multistage cluster sampling design, resulting in a hierarchical data structure and unequal selection probabilities. While our current implementation focused on the imputation model itself, future work should explicitly account for this design in the imputation process to ensure the validity of inferences drawn from the completed datasets. A promising direction, for example, as suggested by Weirich et al. (2014), is to explore a nested multiple imputation framework. This would involve first imputing any missing background data at the student level before generating PVs for the cognitive items, thereby more fully respecting the structure of the data and the intended population inferences.

Furthermore, the international nature of these assessments raises critical questions about measurement invariance, or differential item functioning (DIF), across countries. The presence of DIF indicates that an item measures abilities differently for members of different groups, potentially biasing cross-national comparisons (Zwitser et al., 2017). The observed variation in imputation performance, such as the slightly worse results for the Brazil data, could be related to country-specific factors like the proportion of unplanned missingness or underlying DIF. A significant avenue for future research is to extend the proposed imputation framework to explicitly test for and account for DIF, perhaps by implementing group-specific models or incorporating DIF parameters directly into the imputation algorithm. This would strengthen the methodological foundation for producing valid and comparable cross-national scores.

In conclusion, both the VAE and the proposed two-step IRT-CVAE method offer promising and flexible approaches for imputing planned missing data in LSAs. By combining parametric and machine-learning techniques, it provides a path to more accurate and efficient handling of complex missing data patterns, ultimately supporting more reliable educational measurement.

References

Ali, U. S., van Rijn, P. W., & Dwivedi, P. (2024, July 17). *Deep-learning methods for multiple imputation in large-scale survey assessments [Paper presentation].* The International Meeting of the Psychometric Society (IMPS). Prague, Czech Republic.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. Lord & M. Novick (Eds.), *Statistical theories of mental test scores* (p. 397-472). Reading, MA: Addison-Wesley.

Bock, R. D. (1996). *Domain referenced reporting in large scale educational assessments* (Tech. Rep.). Paper commissioned by the National Academy of Education for the Capstone Report of the NAE Technical Review Panel on State/NAEP Assessment, Washington, DC.

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

Bouhlila, D. S., & Sellaouti, F. (2013). Multiple imputation using chained equations for missing data in timss: a case study. *Large-scale Assessments in Education*, *1*(1), 4.

Chalmers, R. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1-29.

Cho, A. E., Wang, C., Zhang, X., & Xu, G. (2021). Gaussian variational estimation for multidimensional item response theory. *British Journal of Mathematical and Statistical Psychology*, *74*, 52–85.

Converse, G., Curi, M., Oliveira, S., & Templin, J. (2021). Estimation of multidimensional item response theory models with correlated latent variables using variational autoencoders. *Machine learning*, *110*(6), 1463–1480.

Eggen, T. J. H. M., & Verhelst, N. D. (2011). Item calibration in incomplete testing designs. *Psicológica*, *32*(1), 107–132.

Freire, G., & Curi, M. (2024). Masked autoencoder transformer for missing data imputation of pisa. In *International conference on artificial intelligence in education* (pp. 364–372).

Gonzalez, E., & Rutkowski, L. (2010). *Principles of multiple matrix booklet designs and parameter recovery in large-scale assessments* (Vol. 3; Tech. Rep.). IEA-ETS Research Institute Monograph.

Grund, S., Lüdtke, O., & Robitzsch, A. (2021). On the treatment of missing data in background questionnaires in educational large-scale assessments: An evaluation of different procedures. *Journal of Educational and Behavioral Statistics*, *46*(4), 430–

465.

Jewsbury, P. A., & van Rijn, P. W. (2020). IRT and MIRT models for item parameter estimation with multidimensional multistage tests. *Journal of Educational and Behavioral Statistics*, *45*, 383–402.

Kim, J., Jeong, H., & Yoon, S. (2020). Gamin: Generative adversarial multiple imputation network for highly missing data. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8456–8464.

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Lall, R., & Robinson, T. (2022). The midas touch: Accurate and scalable missing-data imputation with deep learning. *Political Analysis*, *30*(2), 179–196.

Mazzeo, J., Kulick, E., Tay-Lim, B., & Perie, M. (2006). *Technical report for the 2000 market-basket study in mathematics* (Tech. Rep. No. ETS-NAEP 06-T01). Princeton, NJ: Educational Testing Service.

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika*, *56*, 177–196.

Mislevy, R. J. (1998). Implications for market-basket reporting for achievement-level setting. *Applied Measurement in Education*, *11*, 49-63.

Mislevy, R. J., & Wu, P.-K. (1996). *Missing responses and IRT ability estimation: Omits, choice, time limits, and adaptive testing* (ETS Research Report RR-96-30). Princeton, NJ: Educational Testing Service.

Mullis, I. V. S., & Martin, M. O. (2019). *PIRLS 2021 assessment frameworks* (Tech. Rep.). International Association for the Evaluation of Educational Achievement.

Mullis, I. V. S., Martin, M. O., & von Davier, M. (2021). *TIMSS 2023 assessment frameworks* (Tech. Rep.). International Association for the Evaluation of Educational Achievement.

OECD. (2024). *Pisa 2022 technical report* (Tech. Rep.). OECD Publishing, Paris. doi: http://doi.org/https://doi.org/10.1787/01820d6d-en

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... others (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in*

neural information processing systems, *32*.

Rijmen, F., Jeon, M., & Rabe-Hesketh, S. (2016). Variational approximation methods. In W. J. van der Linden (Ed.), *Handbook of item response theory, volume two: Statistical tools* (pp. 259–270). CRC Press Boca Raton.

Robitzsch, A. (2022). On the choice of the item response model for scaling pisa data: Model selection based on information criteria and quantifying model uncertainty. *Entropy*, *24*(6), 760.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.

Rutkowski, L., von Davier, M., & Rutkowski, D. (2013). *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis*. CRC Press.

Seo, M., Korbak, T., Zhou, Y., Bhooshan, S., & Gweon, H. (2023). Gagin: Generative adversarial gaussian imputation network for missing data. *Neural Processing Letters*, *55*(2), 1179–1205.

Ulitzsch, E. (2020). *Using response times for modeling missing responses in large-scale assessments*. Freie Universitaet Berlin (Germany).

Urban, C. J., & Bauer, D. J. (2021). A deep learning algorithm for high-dimensional exploratory item factor analysis. *Psychometrika*, *86*(1), 1–29.

Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, *45*, 1–67.

Veldkamp, K., Grasman, R., & Molenaar, D. (2025). Handling missing data in variational autoencoder based item response theory. *British Journal of Mathematical and Statistical Psychology*, *78*(1), 378–397.

von Davier, M., Gonzalez, E., & Mislevy, R. J. (2009). What are plausible values and why are they useful. *IERI Monograph Series*, *2*(1), 9–36.

von Davier, M., Sinharay, S., Oranje, A., & Beaton, A. (2006). The statistical procedures used in national assessment of educational progress: Recent developments and future directions. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics, vol. 26* (pp. 1039–1055). Elsevier.

Wallmark, J. (2024). Irtorch: An item response theory python package.

Wang, W., Chai, Y., & Li, Y. (2022). Gagin: generative adversarial guider imputation network for missing data. *Neural Computing and Applications*, *34*(10), 7597–7610.

Weirich, S., Haag, N., Hecht, M., Böhme, K., Siegle, T., & Lüdtke, O. (2014). Nested multiple imputation in large-scale assessments. *Large-scale assessments in education*, *2*(1), 9.

White, I. R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: issues and guidance for practice. *Statistics in Medicine*, *30*(4), 377–399.

Wu, M. (2005). The role of plausible values in large-scale surveys. *Studies in educational Evaluation*, *31*(2-3), 114–128.

Yoon, J., Jordon, J., & van der Schaar, M. (2018). GAIN: Missing data imputation using generative adversarial nets. In *International conference on machine learning* (pp. 5689–5698).

Yoon, S., & Sull, S. (2020). Gamin: Generative adversarial multiple imputation network for highly missing data. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 8456–8464).

Zhang, Y., Zhang, R., & Zhao, B. (2023). A systematic review of generative adversarial imputation network in missing data imputation. *Neural Computing and Applications*, *35*(27), 19685–19705.

Zwitser, R. J., Glaser, S. F., & Maris, G. (2017). Monitoring countries in a changing world: A new look at dif in international surveys. *Psychometrika*, *82*(1), 210–232.

Zwitser, R. J., & Maris, G. (2015). Conditional statistical inference with multistage testing designs. *Psychometrika*, *80*(1), 65–84.

Appendix A.

Figures 1 - 3 show scatterplots of the person and item means of observed and imputed data for the three imputation methods for three PIRLS 2021 data sets. In addition, Figures 4 - 6 display similar scatterplots for the three TIMSS 2023 data sets. It can be seen that the item means are recovered better than the person means, which is not surprising given that more data is generally available for the former. However, an interesting pattern is seen when the person means are shown at the booklet level, which is illustrated for the TIMSS 2023 Türkiye data in Figures 7 and 8. Figure 7 shows the scatterplots for the seven more difficult booklets and Figure 8 displays the person means for the seven less difficult booklets (see Mullis et al., 2021, Chapter 4). It is clearly seen that the person means for booklets 6 and 7 are higher than the observed means (across the three imputation methods), while for booklets 9 and 11 (and to a lesser extent for 8 and 12) the imputed means are lower than the observed means.

This last finding indicates that there are position effects. Since item position was neither taken into account in the data structure (i.e., the data from the same item in different booklets appears in a single column of the data matrix) nor in the modeling (e.g., by means of a booklet or position indicator), the issue is seen in all three imputation methods. Since the booklets are randomly assigned, these effects do not have to be problematic per se. However, they can give issues when there are interactions between position effects and other variables of interest, a topic we leave for further investigation.

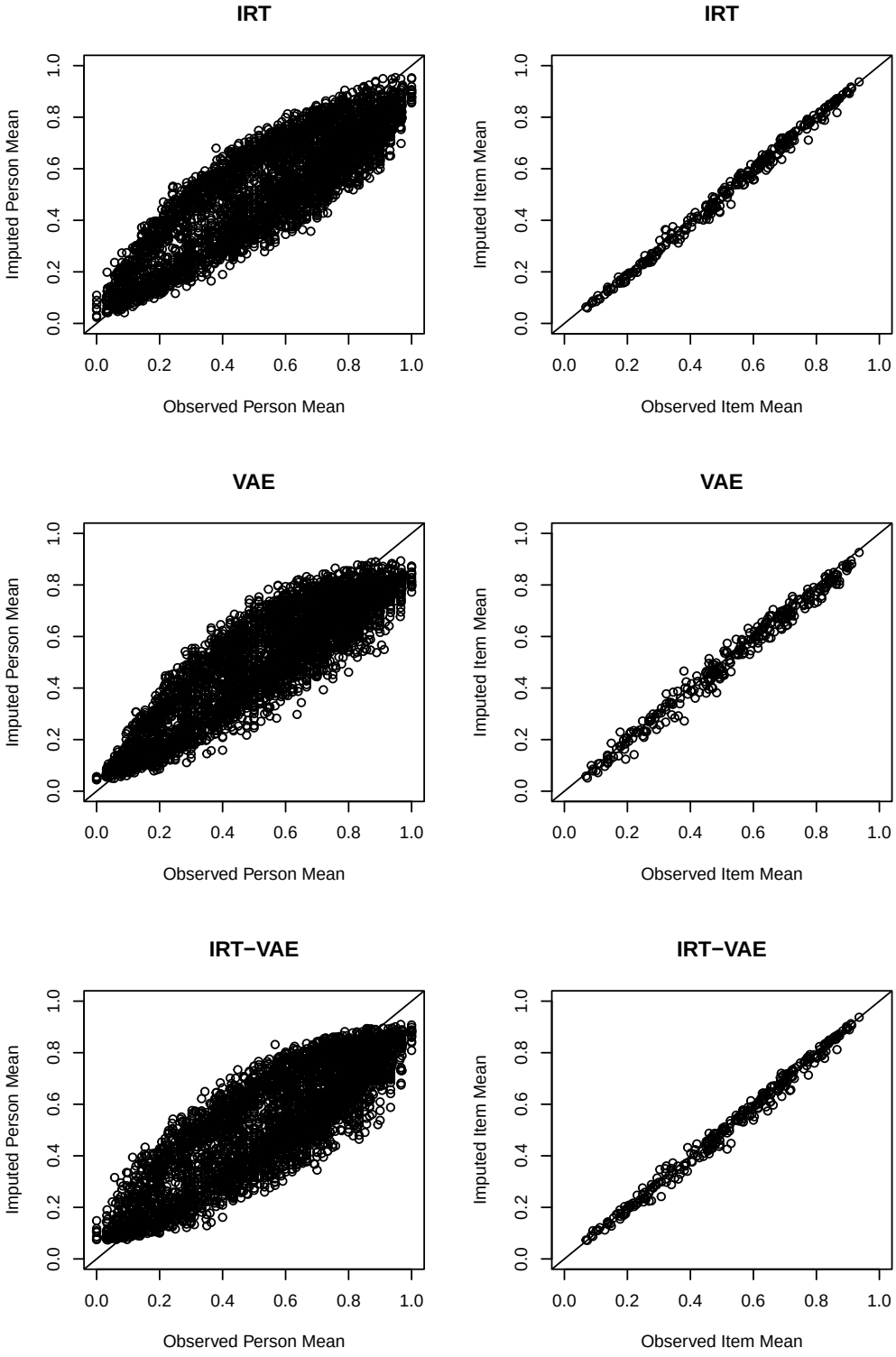

Figure 1. Person and Item Means for IRT, VAE, and IRT-VAE imputation methods for Brazil PIRLS 2021 data.

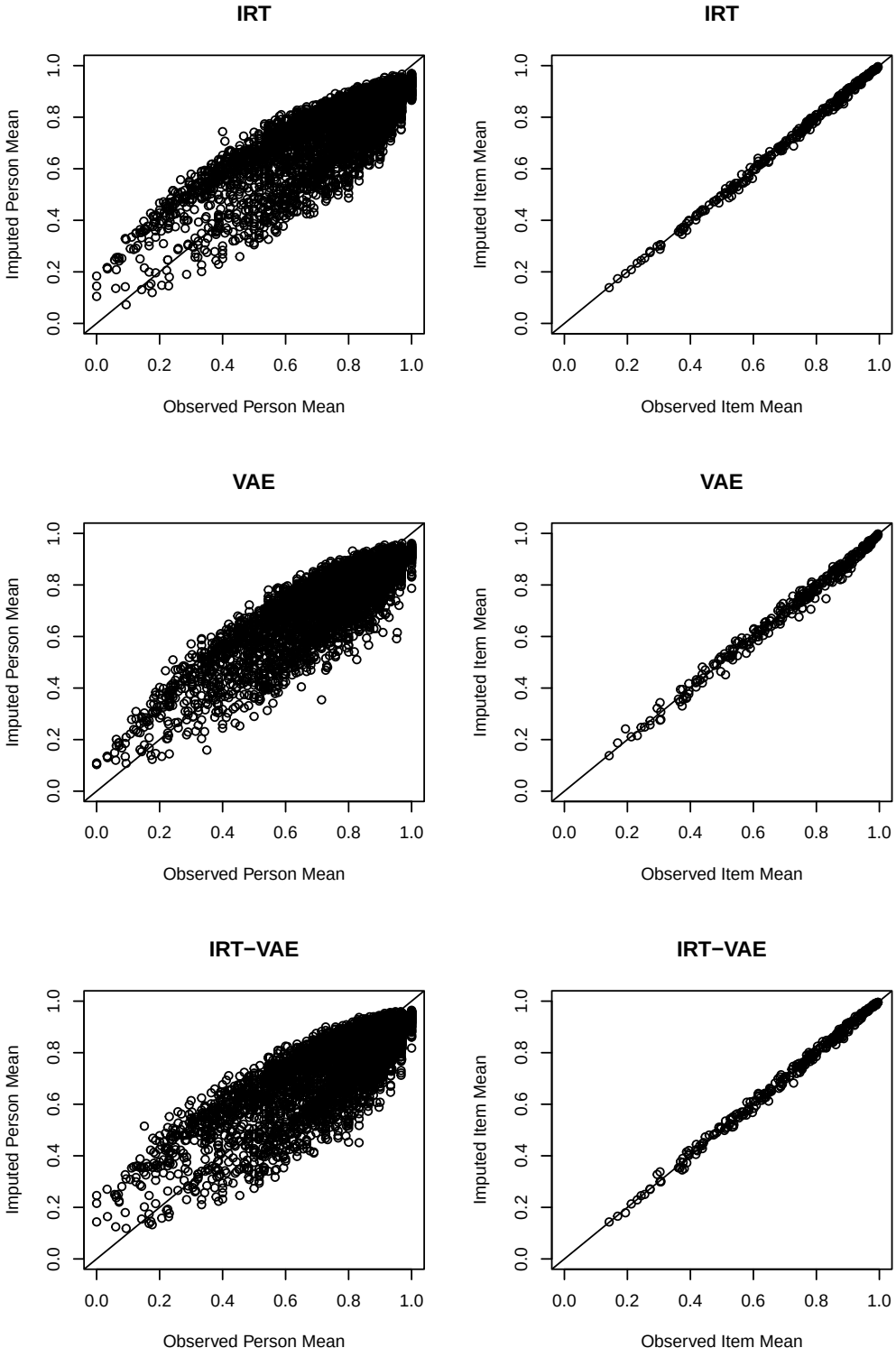

Figure 2. Person and Item Means for IRT, VAE, and IRT-VAE imputation methods for England PIRLS 2021 data.

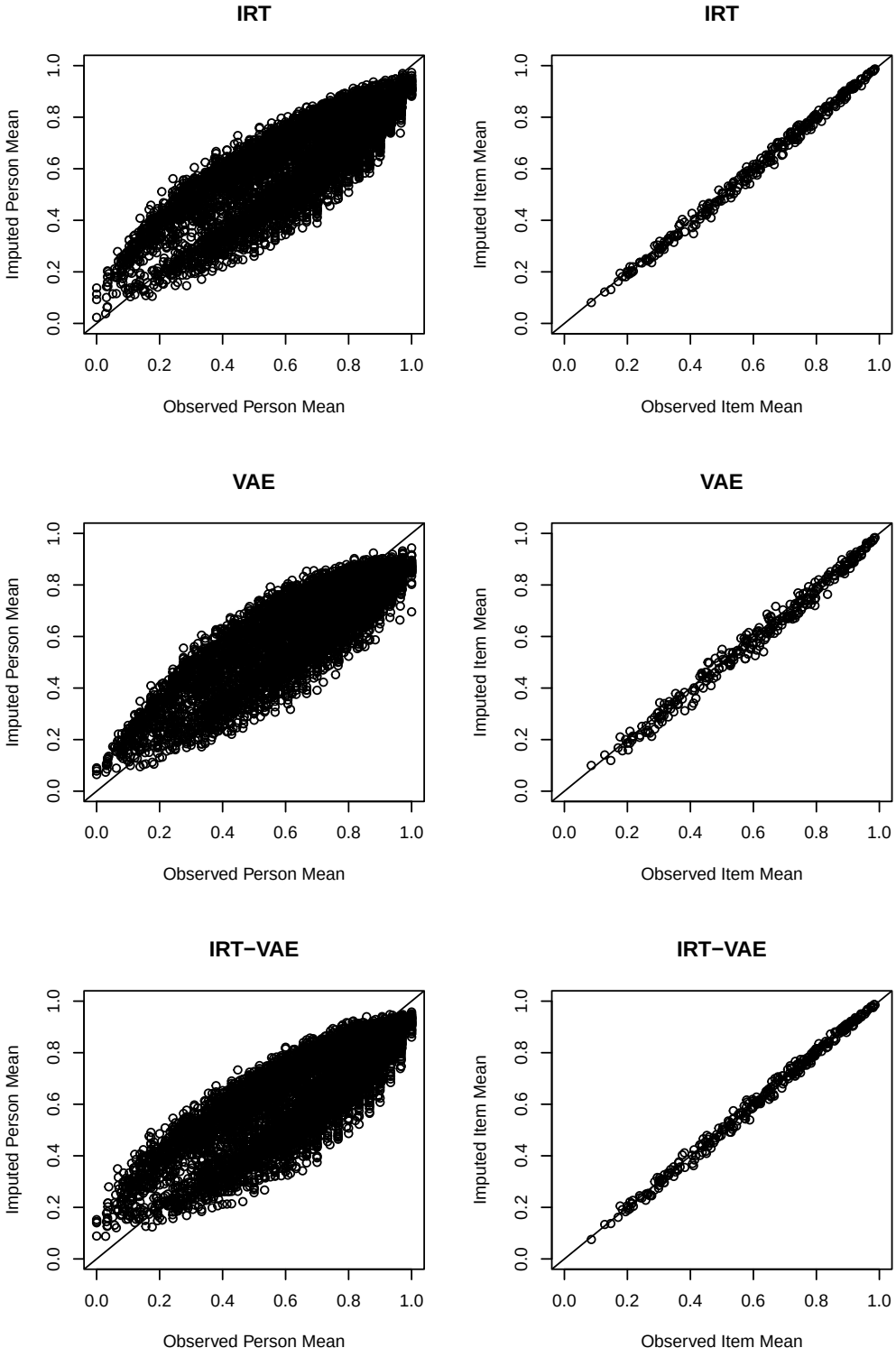

Figure 3. Person and Item Means for IRT, VAE, and IRT-VAE imputation methods for Türkiye PIRLS 2021 data.

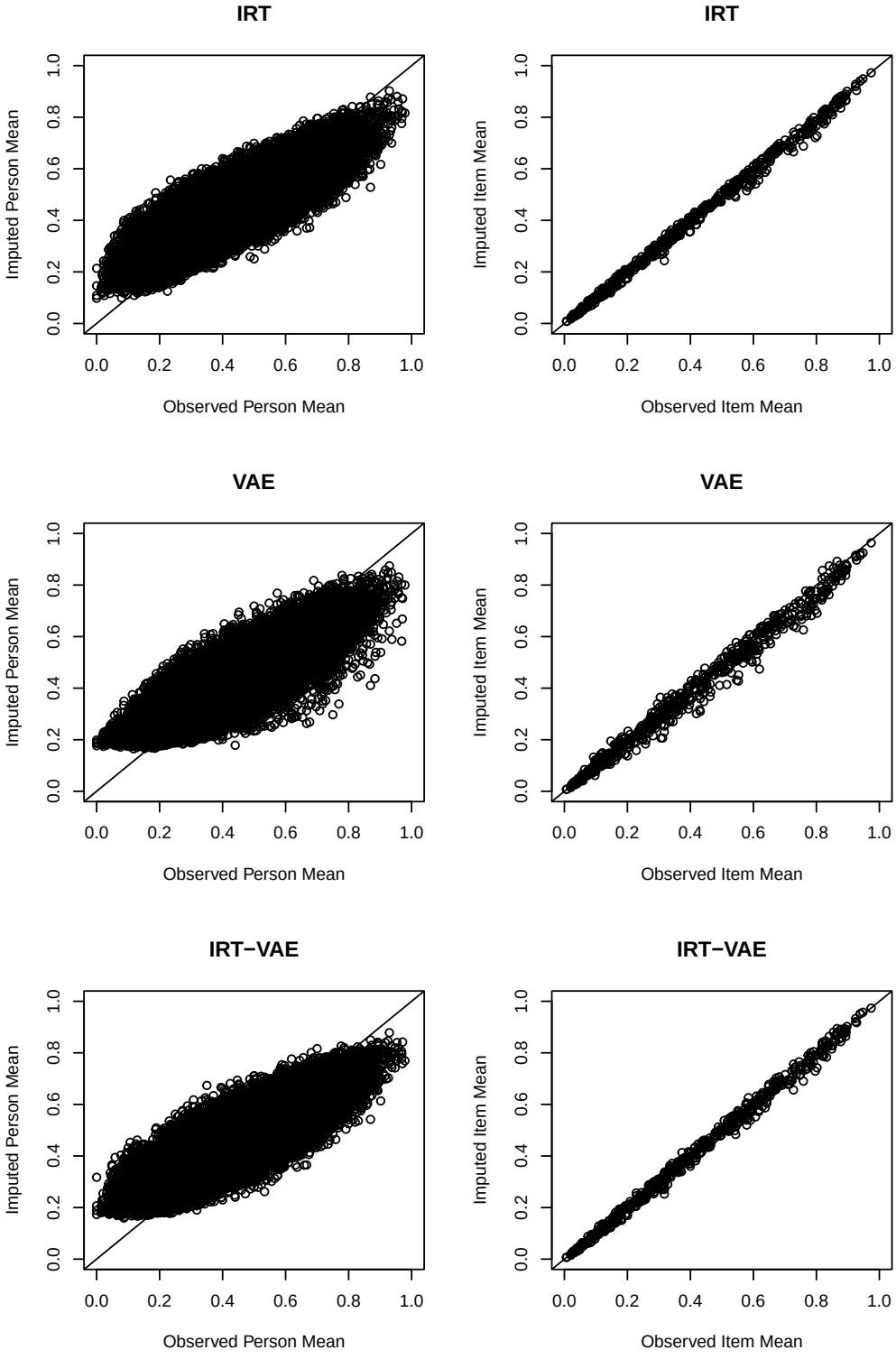

Figure 4. Person and Item Means for IRT, VAE, and IRT-VAE imputation methods for Brazil TIMSS 2023 data.

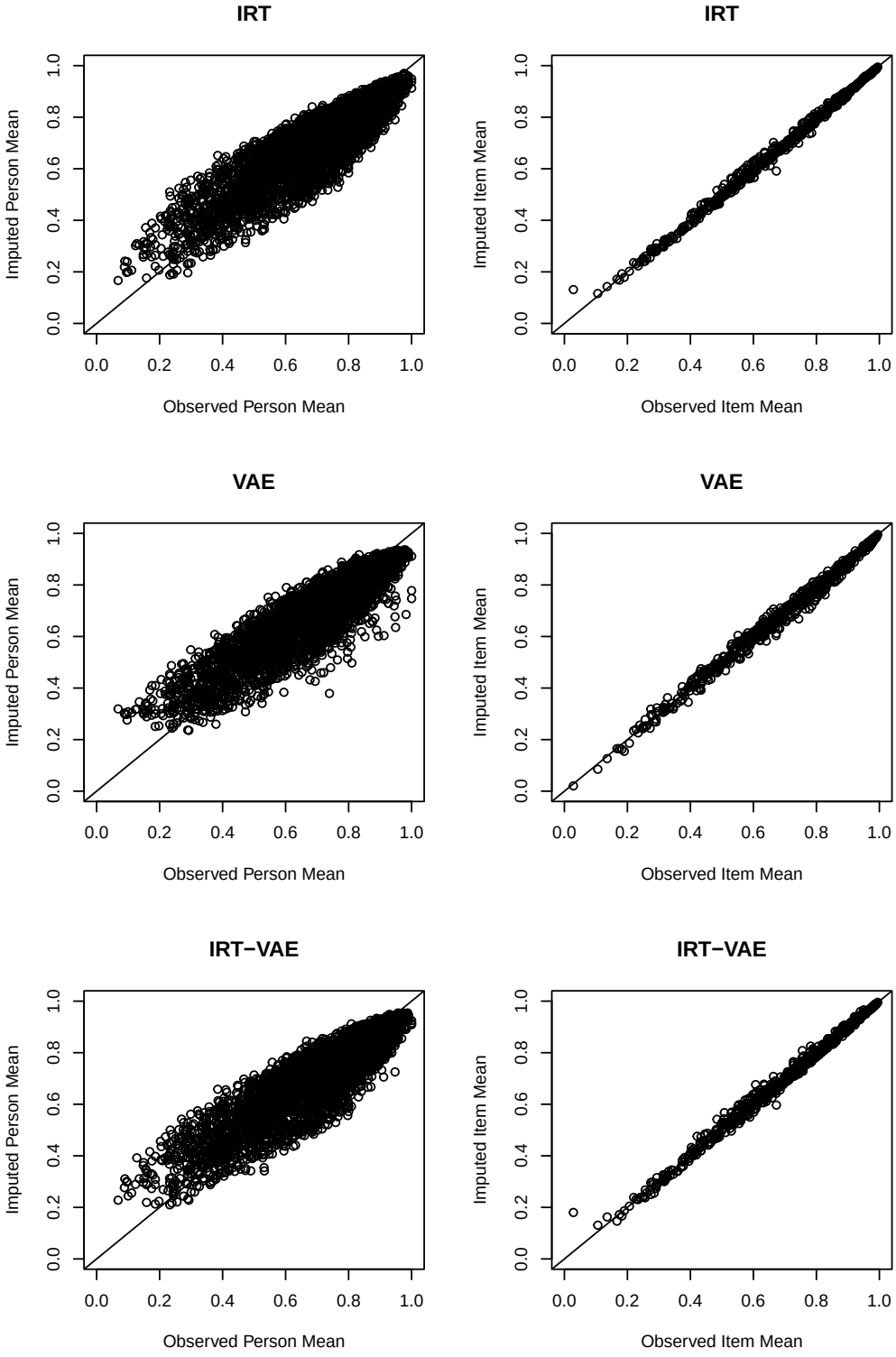

Figure 5. Person and Item Means for IRT, VAE, and IRT-VAE imputation methods for England TIMSS 2023 data.

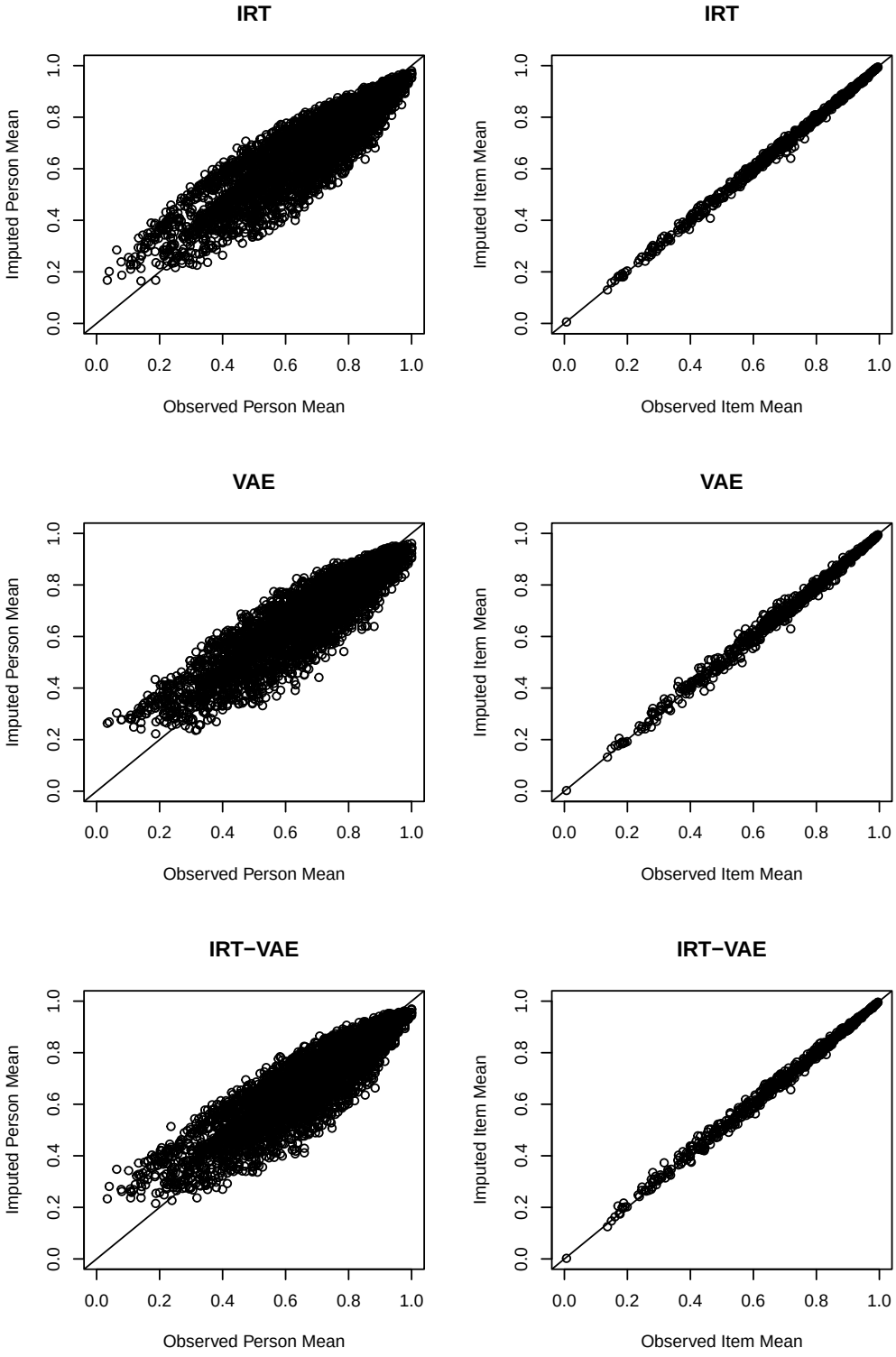

Figure 6. Person and Item Means for IRT, VAE, and IRT-VAE imputation methods for Türkiye TIMSS 2023 data.

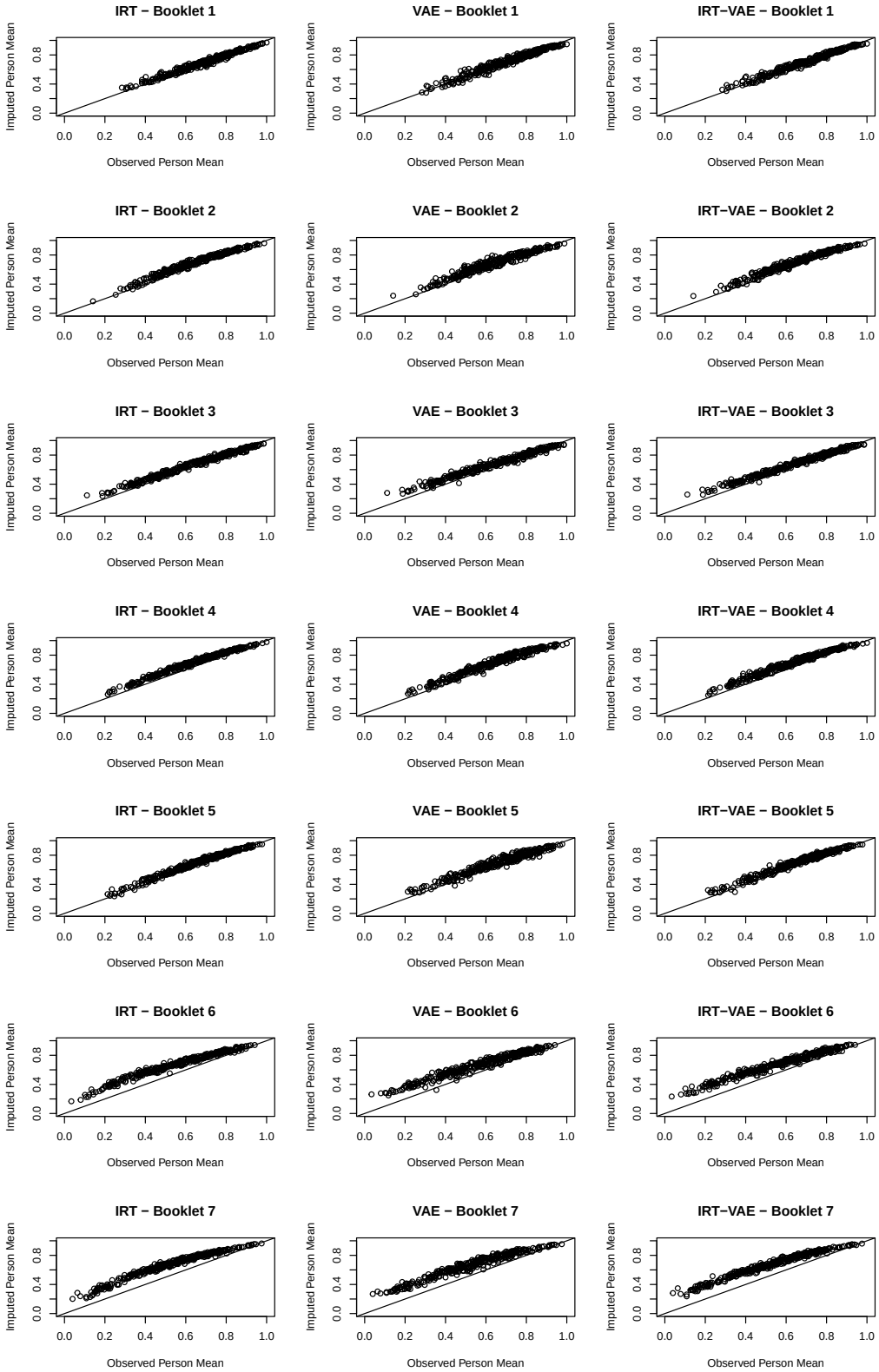

Figure 7. Person Means for IRT, VAE, and IRT-VAE imputation methods for Türkiye TIMSS 2023 data, more difficult booklets.

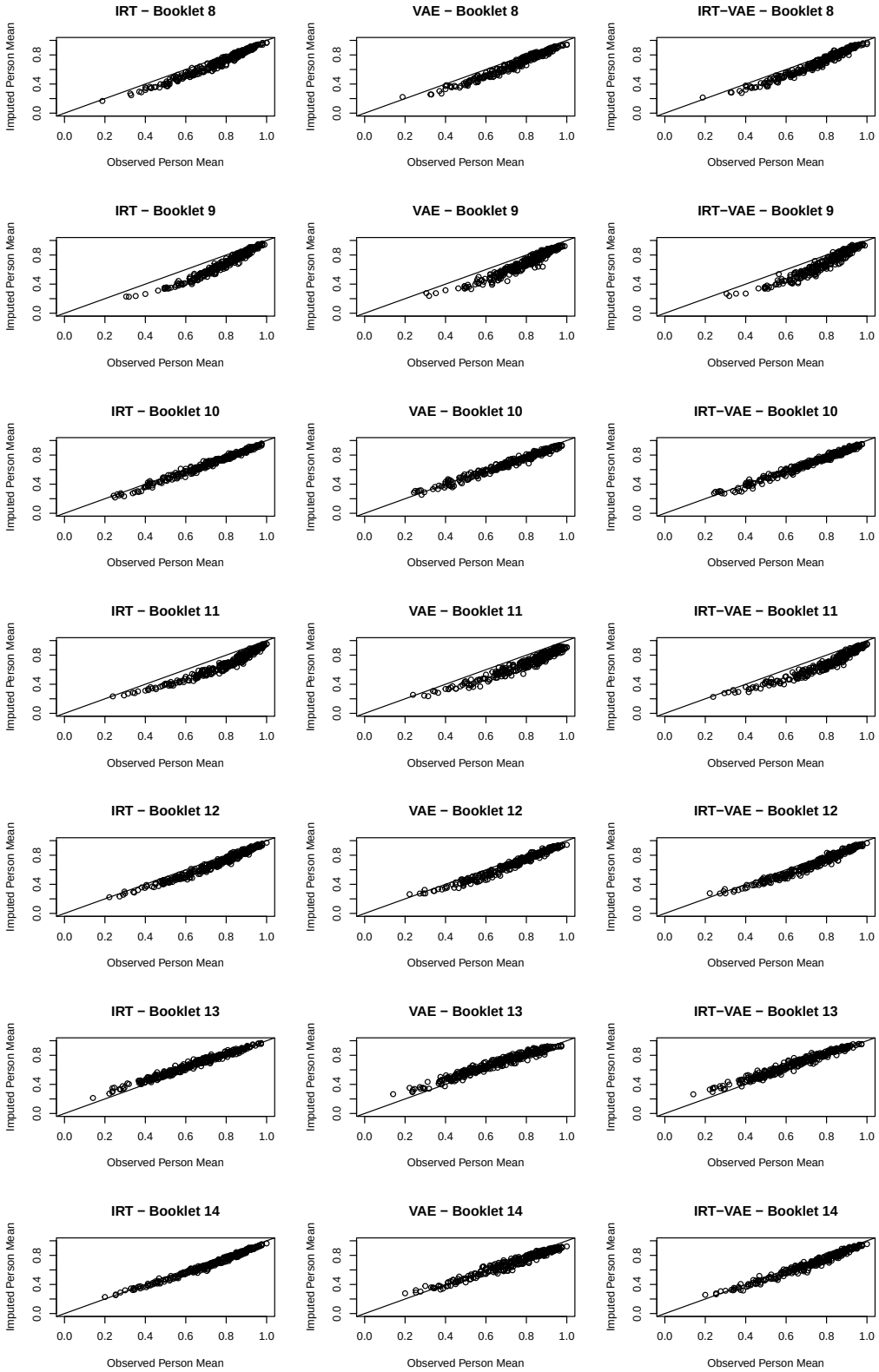

Figure 8. Person Means for IRT, VAE, and IRT-VAE imputation methods for Türkiye TIMSS 2023 data, less difficult booklets.

Appendix B. Response to Reviewers' Comments

Reviewer #1

All edits in response to reviewer's comments are in text with track changes.

Reviewer #2

This project compared several imputation methods based on deep neural networks. I am glad to see the authors to test modern machine learning approaches for analysing data from large-scale assessments. However, I think it is not entirely clear how such results can be used in the analysis and reporting of ILSAs. My major comments are listed below. Maybe the authors could consider adding some discussions.

1. The report says the research aims to overcome the issues with the PV method by developing an alternative method that supplements the current PV methodology.

   (a) In what sense does the proposed approach supplement the current PV approach? Do you suggest replacing the PVs by some other sub-score measures? If so, what measures do you use? With the imputation approach, you can estimate some total scores (as now you have each student's (imputed) responses to all the items). Are you going to use that? (Do the "resulting scores" in your research questions and in the analysis refer to this kind of total scores? If so, please clarify in the report.)

   Response: The final goal is to have an alternative method that can operationally replace traditional PV-based reporting under specific analytic scenarios. At the current stage, the project explores whether deep-learning-based imputations can generate complete response matrices that may support new types of score derivations (including total-score–like aggregates) for more flexible reporting. These "resulting scores" refer to the imputed, complete-response–based outcomes that allow comparison with standard PV-based proficiency estimates in later stages.

   (b) If you answer to the above question is no, then how does the proposed imputation methods supplement the PV method? In other words, how do you integrate the imputation methods with the IRT-based PV approach?

Response: The intention is not to integrate neural-network imputation inside the PV pipeline, but rather to create a methodology that uses both the cognitive and contextual information to generate such a complete data matrix and develop new indicators for checking e.g., comparability and item invariance.

(c) If your answer to the above question is yes, then how does this approach address the issues you mentioned about the PVs? For example, one issue you mentioned is the fairness issue with the PVs (which I do not think is a big problem given that ILSAs are low-stake to students). However, the imputation procedure does not solve the fairness issue, in the sense that the summary scores you use depends on the imputations, and the imputations are based on the entire dataset, including the background variables. Therefore, bias still exists with your approach. In fact, the fairness issue will be even more severe due to the use of black-box machine learning algorithms (as with the IRT framework, you can control DIF, and also at least know where bias may appear in the latent regression, which has a clear form). I am not against using machine learning methods, but I do not want to give practitioners the hallucination that deep learning or AI can solve all the problems. Sometimes, the issues are just covered but not addressed.

Response: See the above response. The fairness issue and other psychometric aspects will also be studied within the potential deep learning approach.

2. Another issue with the deep-learning-based imputation procedures is that performing statistical inference based on the multiple imputations is problematic. Even if we do not talk about the potential high variance brought by the large number of parameters in these deep neural nets (which can be a huge problem if you do this country by country as the current PV procedure), the standard inference procedures, like Rubin's formula, no longer apply due to the large number of parameters in these neural nets. In other words, when people use the multiple imputation datasets for subsequent analysis, the p-values they obtain using Rubin's formula are no longer valid (note that Rubin's formula only applies under the parametric setting). As a result, it is unclear how and whether the multiple imputation dataset can be used by practitioners. The current research only evaluates the accuracy of the imputation but did not consider

the statistical inference based on the imputed data.

Response: We agree; this is a critical aspect to consider and now how to address it. Again, the current phase focuses on imputation quality and robustness, not on producing inference-ready imputed datasets. From another perspective, Rubin's formula does not only apply under parametric settings. For example, predictive mean matching is a nonparametric approach, which involves replacing missing values with observed ones from similar rows. Furthermore, Lall and Robinson (2022) apply Rubin's rules as well in the context of missing-data imputation with autoencoders

3. Finally, I want to share my view on the fully deep-neural-network-based approach to the scaling of ILSAs. While I agree that some assumptions of the IRT latent regression model is oversimplified and can be extended for better measurement validity, the deep learning approach is a bit too much for several reasons.

   (a) Fairness. The IRT framework allows us to adjust for DIF. The bias that may arise through the latent regression model is also clear. The deep learning models implicitly bring bias into the scaling (as long as you use background variables). The fairness concepts in machine learning do not directly apply to the educational testing setting (see e.g., recent papers by Mathew Johnson for a discussion). Therefore, it is much harder to be fair when we use the purely machine learning approach. In particular, if you do the analysis using the whole dataset, rather than country by country, I suspect that a big bias will appear against certain countries.

   Response: We fully agree that fairness is a central methodological consideration. However, addressing this issue is beyond the scope of the current project. As noted earlier, DIF analyses would need to be defined differently under deep-learning–based imputation frameworks, analogous to how DIF methodologies evolved when assessment practice transitioned from classical test theory (CTT) to IRT.

   (b) Bias-variance trade-off. Data from ILSAs differ quite substantially from other fields where these deep learning models are found successful. In typical appli-

cations of deep learning models (e.g., large language models), you have a very big sample size, highly heterogeneous and complex data, and the information in each variable may be quite low (but when you aggregate over many variables using deep net, you can still achieve quite good accuracy). However, ILSAs are different. The signal is quite strong in both our cognitive and noncognitive variables (thanks to the careful design of these surveys), but the sample size for each country is not super big (as in the successful applications of deep neural nets). In that case, deep neural net models do not perform well due to the high variance (I suspect that some nonlinear factor models can achieve better accuracy than the deep learning algorithms in the current report). Overall, I think the performance of the current IRT latent regression model can be improved by some parametric extensions, but I think deep neural network models are too flexible for this problem.

Response: This is an important point and aligns with our empirical observations.