Revisiting Plausible Value Estimation Using Bayesian Methods for Large-Scale Assessments

by

Kjorte L. Harra

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

(Educational Psychology)

at the

UNIVERSITY OF WISCONSIN–MADISON

2026

Date of final oral examination: 05/04/2026

The dissertation is approved by the following members of the Final Oral Committee:

David Kaplan, Hilldale Professor and Richard L. Venezky Professor, Educational Psychology

Daniel Bolt, Nancy C. Hoefs Bascom Professor, Educational Psychology

James Wollack, Professor, Director of Testing & Evaluation Services, Educational Psychology

Sameer Deshpande, Assistant Professor, Statistics

To my parents, for believing in the value of my education.

To Evan, for his unwavering support.

And to David Kaplan, for taking a chance on me.

ACKNOWLEDGMENTS

Accomplishing this goal would not have been possible without the help and support of many people in my life, to whom I am sincerely grateful beyond what I can express here.

First, I'd like to thank my advisor, Dr. David Kaplan, whose guidance throughout my graduate career has been invaluable. Through your mentorship, I have grown as a researcher and developed a deep appreciation for the Bayesian perspective.

I am also grateful for my committee, Dr. Daniel Bolt, Dr. James Wollack, and Dr. Sameer Deshpande, for their insightful feedback and continued support of my work.

Next, I am especially thankful to my wonderful parents for their love and constant encouragement, even when I struggled to believe in myself. Your investment in my education has made this possible, and I am deeply indebted to you for the opportunities it has given me.

Last but not least, I'd like to thank my lovely partner, Evan. Your confidence in me, genuine interest in my work, and unwavering support have brought me immense joy. I feel truly fortunate to share this journey with you.

My appreciation extends beyond those I can list here. I would like to thank my friends, family, colleagues, teachers, and professors for their support and contributions along the way.

Use of Artificial Intelligence Tools:

In preparing this dissertation, I used large language model tools (specifically ChatGPT by OpenAI) solely to assist with code troubleshooting and editorial refinement of written prose. No AI-generated content was included without careful review and revision. All substantive ideas, analyses, interpretations, and conclusions are my own. No sensitive or restricted data were shared with these tools. This use is consistent with the University of Wisconsin–Madison's academic integrity policies at the time of submission. I retain full responsibility for the content of this work.

CONTENTS

ABSTRACT

Plausible values (PVs) are widely used in large-scale assessments (LSAs) to estimate student proficiency, yet traditional approaches rely on a single population model and dimensionality reduction techniques such as principal components analysis (PCA). These practices can introduce bias, create statistical uncongeniality, and limit predictive performance in secondary analyses. This dissertation evaluates three Bayesian approaches to PV estimation that increase model flexibility, better account for model uncertainty, and improve predictive accuracy through simulation and empirical studies using PISA 2022 reading data.

The first study investigates Bayesian Model Averaging (BMA) (Leamer, 1978; Raftery, 1995; Raftery et al., 1997) for PV generation, allowing the algorithm to average over the space of candidate models rather than rely on a single imputation model specification. By explicitly incorporating model uncertainty, BMA reduces bias and yields more accurate, predictively optimal PVs. The second study examines sparsity-inducing priors—including the ridge, regularized horseshoe, and R2D2 priors (Hsiang, 1975; Piironen and Vehtari, 2017; Zhang et al., 2022) as alternatives for PV generation. These approaches perform shrinkage and variable selection while preserving the original structure of the conditioning variables, leading to improved recovery of the population distribution and stronger predictive performance. The final study explores Bayesian additive regression trees (BART) (Chipman et al., 2010), a nonparametric ensemble method that flexibly captures model uncertainty and mitigates imputation model misspecification concerns. Simulation results indicate improvements in accuracy, though gains are less consistent in empirical applications.

Overall, the results identify BMA as the most effective approach for PV estimation. Sparsity-inducing priors were more accurate but computationally expensive, whereas BART was computationally cheaper but less consistently effective empirically. These findings highlight key trade-offs and establish BMA as a strong, practical framework for advancing plausible value methodologies in large-scale assessments.

1 INTRODUCTION

Large-scale assessments (LSAs) in education are a valuable tool in evaluating and tracking educational achievement trends and student outcomes in dozens of countries and jurisdictions. These assessments, such as the Programme for International Student Assessment (PISA) (OECD, 2023), aim to understand student populations of interest in terms of academic achievement and relevant social contexts. Given the complexity of these assessments, estimating student proficiency requires careful consideration of the sampling design, relevant background characteristics, secondary analyses, and more. To construct estimates of student latent ability or proficiency in a given academic domain of interest, plausible values (PVs) are implemented.

Student latent ability represents an individual's underlying academic proficiency in a domain such as reading or math. This is referred to as a "latent" ability because although a student's test performance provides an indication of their true ability, that ability cannot be directly observed due to random error and other situational factors such as test-day conditions. LSAs also collect extensive background information, ranging from student demographics, students' home environment, and school characteristics to support this estimation. Accurately specifying the relationship between these characteristics and latent ability is critical, as misspecification can bias estimates and downstream inferences.

Despite their importance, key aspects of PV estimation remain underexplored, particularly the specification of the population model linking ability to background characteristics. Misspecification in this element can produce unreliable results for both reporting agencies and secondary analysts. This work explores potential a potential alternative to PV methodologies to better address sources of uncertainty in PV estimation while balancing statistical accuracy and practical implementation.

Large-Scale Assessments

Large-scale assessments (LSAs) such as the OECD's PISA aim to compile reports on student proficiency in a variety of topics, primarily numeracy and literacy. These assessments focus on group-level trends, as opposed to scoring individual students, to allow for comparisons over time and between relevant groups and jurisdictions (Jewsbury et al., 2024; Wu, 2005; Khorramdel et al., 2020). It is important to note that especially for the purposes of this work, the goal of LSAs like PISA are not to provide accurate and reliable *individual* student scores; the focus is on *population-level* estimates. Reported statistics of interest may be population means, variance, percentages meeting proficiency, and standard errors (Wu, 2005; Johnson and Jenkins, 2004). To secure accurate and reliable estimates across relevant content domains, LSAs employ multi-stage sampling designs, collect key background information, and apply complex statistical methods to obtain such estimates.

These assessments administer cognitive tests in subjects such as mathematics and reading but also collect background information on students, schools, and family life. This contextual information is often used to group students for reporting results, such as by gender or socioeconomic status. The results of these assessments can be used to inform education policy reform, track which education systems are performing well or not, and what student groups or schools may need more support. To best advise stakeholders, it is crucial that ability estimates and their relationship to relevant context indicators are accurate (Monseur and Adams, 2009). These data also allow applied researchers to answer their research questions through secondary analyses (Jewsbury et al., 2024). Secondary analysts use LSAs to answer their own research questions about the state of education and related topics.

Sampling Design

To obtain a representative sample of the student population in a given jurisdiction, LSAs like PISA use multi-stage sampling. This involves sampling some geographic area (e.g., counties), then schools, and then randomly sampling students nested within the selected

schools (Monseur and Adams, 2009; Jewsbury et al., 2024; Mislevy et al., 1992; Johnson and Jenkins, 2004). Sampling weights are implemented as well to more accurately represent population characteristics (Mislevy et al., 1992). To achieve sufficient coverage of topics and student subgroups of interest, the following item sampling designs incorporating planned missingness are adopted.

In short, each participating student in an assessment is given a subset of questions from a larger question bank, known as a cluster or block (Braun and von Davier, 2017; von Davier et al., 2009; Johnson and Jenkins, 2004). Students do not receive a large portion of the possible questions, as an exam that sufficiently covers all the important cognitive material for each participating student would be quite time-consuming and unnecessary when the goal is population-level performance. This design is called matrix item sampling (Monseur and Adams, 2009; Jewsbury et al., 2024). Shorter assessments, featuring a broader range of questions and administered to a larger number of students, provide more accurate population estimates than longer assessments. Additionally, shorter exams are easier to administer logistically in schools (Braun and von Davier, 2017; Mislevy et al., 1992; von Davier et al., 2009).

While statistically accounting for a complex sampling design is more challenging, these designs allow for broader coverage of all relevant cognitive domains of interest without forcing students to take prolonged exams. Scaling procedures are implemented to estimate student proficiency given their cognitive test performance and relevant background information (Grund et al., 2021). This type of sampling design coupled with the group-level focus goals of these assessments require more complex estimation models for student-level proficiency, paving the way for plausible value imputation.

1.1 Conventional Plausible Value Estimation

The proper estimation methods and uses of plausible values are critically important, as they directly inform LSA outcomes such as country rankings, proficiency classifications, and average student achievement scores in assessments like PISA (Özer, 2020; Martens and Niemann, 2013; Araujo et al., 2017). Consequently, any modifications to PV estimation methods discussed in this work could have substantial downstream effects, potentially influencing education policy decisions, resource allocation, and international comparisons of student achievement.

Plausible values are imputed estimates of individual student ability in a given domain, conditioned on both student test performance and relevant background information, where random draws are taken from the resulting posterior distribution of ability. PVs play a central role in LSAs, providing more accurate population-level estimates and accounting for sources of measurement uncertainty. They are widely used to report educational trends within and between subgroups, compare performance across jurisdictions, and support applied research linking student achievement to contextual factors. In practice, primary analysts generate and report PVs, while secondary analysts use them in subsequent research. Together, their work informs stakeholders and policymakers to support evidence-based decision-making in education.

Statistical Framework

Mathematically, plausible values are random variables intended to approximate the conditional distribution of latent ability given a set of a student's item responses and background characteristics (Marsman et al., 2016). The aim of the PV imputation model is to accurately reconstruct the population distribution of latent ability rather than produce precise single point estimates for individuals. We can view plausible value estimation as an extension of missing data imputation, where proficiency estimates are missing for all students

(Rubin, 1987; von Davier, 2013).

To estimate latent ability in LSAs using plausible values, we must specify the posterior distribution from which proficiency draws are obtained for each student. Plausible values are then defined as random draws from an individual's posterior distribution of proficiency, and is denoted as $h(\theta|\mathbf{x})$:

$$h(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta)g(\theta)}{\int f(\mathbf{x}|\theta)g(\theta)d\theta}, \tag{1.1}$$

where a student's item response pattern is denoted as $\mathbf{x}$, latent ability as θ, an IRT model as $f(\mathbf{x}|\theta)$, and a population model that follows a normal distribution, $g(\theta) \sim \mathcal{N}(\mu, \sigma^2)$ (Wu, 2005).

Briefly, the process of generating plausible values, as formulated in Equation (1.1), can be viewed as an imputation model for θ (von Davier, 2013). First, principal components analysis (PCA) is typically run as a preprocessing step to reduce the variance and collinearity with the large set of background variables for computational ease (OECD, 2023). These background variables are often called conditioning variables, as they are used to condition the estimation of student proficiency. Next, the posterior distribution of proficiency for each student is generated, following Equation (1.1). The posterior distribution, $h(\theta|\mathbf{x})$, is obtained under an IRT model, $f(\mathbf{x}|\theta)$, that relates student responses to θ, and a population model, $g(\theta)$, specifying the prior distribution of θ as a function of the background characteristics. Plausible values are then multiple draws from the posterior distribution of latent ability for each student (Jewsbury et al., 2024; von Davier, 2013; OECD, 2023). The generated plausible values can then be used for official reporting and for secondary analyses. A more detailed examination of the posterior distribution formulation is provided below.

Plausible values have been shown to outperform other available methods in estimating a student's latent ability in LSAs such maximum or weighted likelihood estimates (MLE/WLE) (Warm, 1989; von Davier et al., 2009; Wu, 2005; Monseur and Adams, 2009). By using multiple random draws from each student's posterior distribution of latent ability, PVs explicitly

account for uncertainty in individual-level estimates, leading to more accurate population-level inferences (Braun and von Davier, 2017; Khorramdel et al., 2020; von Davier et al., 2009; Wu, 2005; Monseur and Adams, 2009). This bias-reduction effect is most pronounced when multiple PVs are used in analyses, as relying on a single PV to fully represent student proficiency can lead to overconfidence in results.

Another key advantage of PVs is their ability to accurately capture subgroup variability, which is essential for subgroup reporting and for secondary analyses. This holds true even when the conditioning model may be missing relevant variables or the test has few items (Marsman et al., 2016; Wu, 2005). In contrast, maximum likelihood estimates (MLEs) only provide single point estimates, failing to capture the full distribution of proficiency estimates (Wu, 2005). These benefits make PVs particularly effective for accurate subgroup reporting and secondary analyses.

Item Response Theory

Individual proficiency on cognitive tests is commonly estimated using item response theory (IRT), a form of latent variable modeling that relates item responses to underlying ability θ (Khorramdel et al., 2020; Marsman et al., 2016). In this framework, the distribution of observed item responses is obtained by integrating over the latent ability distribution. PISA uses the two-parameter logistic model (2PLM) for dichotomous items and the generalized partial credit model (GPCM) for polytomous items (OECD, 2024). IRT enables common reporting scales and supports comparisons across domains and populations.

Constructing scale scores from cognitive assessments using IRT is widely considered best practice in LSAs and applied research more broadly (Braun and von Davier, 2017). IRT models used in this context involve two key components: item calibration and latent trait estimation. Item calibration refers to the estimation of item parameters, while latent trait estimation refers to the estimation of examinee ability. In LSAs, item parameters are typically estimated using conditional maximum likelihood, whereas ability parameters are

estimated using either maximum likelihood or Bayesian methods.

Due to the item sampling design of LSAs and other constraints, students do not receive enough test items to reliably estimate the population distribution of proficiency using IRT alone. MLEs of student ability obtained from IRT models are not consistently accurate enough to serve as latent ability estimates for reporting purposes. LSAs are often more interested in population and subgroup reporting, which MLEs fail to recover correct population parameters regardless of sample size (Mislevy et al., 1992). Convergence to the true population parameters would require students to receive many more test items. This logistical burden would necessitate substantially longer tests and larger item pools.

The Population Model

While IRT models can generate estimates of latent ability θ, these estimates do not account for the population characteristics, and are therefore insufficient estimates for primary reporting and secondary analyses. In LSAs, θ is conditioned on all the given background characteristics obtained from context questionnaires, allowing for more accurate subgroup analysis. To impute PVs, we can specify a population prior distribution that accounts for subgroup structures and mitigates misrepresentation of the true relationship between the background variables and achievement (Wu, 2005; Marsman et al., 2016). We can define this population model as a prior distribution, $g(\theta)$, as:

$$g(\theta) \sim \mathcal{N}(\mu + x_1 b_1 + x_2 b_2 + x_3 b_3 + \cdots, \sigma^2). \tag{1.2}$$

which follows a normal distribution conditioned on background variables $\boldsymbol{b}$, where proficiency is modeled as a function of the background variables. Including these conditioning variables in the population model that defines the posterior distribution of θ enables secondary analysts to more accurately estimate relationships between background characteristics and group-level achievement (Wu, 2005). Because LSAs typically include hundreds of background variables, principal components analysis (PCA) is used to simplify the variable set while still accounting

for up to 90% of the variance (Braun and von Davier, 2017).

In theory, one could solely rely on the population model to generate PVs, however, doing so requires correct model specification (Marsman et al., 2016). This is unlikely in practice given the complexities of measuring proficiency in LSAs. Under the Bayesian framework used to generate PVs, the influence of the population model prior diminishes as the number of variables increases. When assessments have a sufficient number of test items, a misspecified population model does not prevent the empirical posterior distribution from approximating the true distribution of plausible values (Marsman et al., 2016). Regardless, a better specified population model (i.e., properly accounting for model uncertainty and congeniality) will more quickly converge to the true PV distribution, stressing the importance of the population model within the PV estimation process.

The number of PVs used from random draws of each student's posterior distribution varies across LSAs, with ten in PISA (OECD, 2023). Multiple random draws from the posterior more accurately depict the population distribution of latent ability and account for individual-level uncertainty in scores (Wu, 2005; Jewsbury et al., 2024). However, current PV methodologies rely on a single, fixed population model linking background variables to ability, leaving additional sources of model uncertainty unaddressed.

Secondary Analyses

PVs are a useful tool for applied researchers conducting secondary analyses. In education, psychology, and the social sciences more broadly, LSAs are used to address a wide range of substantive questions. PVs are used by secondary analysts to estimate achievement gaps, evaluate policy effectiveness, and track trends over time, among other applications. Through their estimation methods, PVs can address uncertainty and reduce bias in secondary analyses, allowing for more accurate statistical inferences in a variety of research contexts.

For the purposes of this work, we can distinguish between two categories of secondary analyses based on the role of the analyst: those conducted by *primary analysts* and those

conducted by *secondary analysts*. For LSAs, primary analysts are typically the organizations responsible for administering the assessment and producing official reports. These analysts generate and disseminate plausible values and use them to report main results and rankings, as in the case of PISA. Secondary analysts may include these same organizations conducting additional analyses beyond official reporting, as well as external researchers (e.g., psychologists, education researchers, and other third-party analysts) who use released PVs for independent or downstream analyses. There are several ways PVs can be used in both types of secondary analyses.

As dependent variables in secondary analyses, utilizing PVs follows the same approach that is recommended for multiple imputation procedures, as PVs can be thought of as a special case of missing data with multiple imputation. It is recommended to repeat a given analysis or model estimation procedure K times, once for each of the k plausible values estimated by the assessment (e.g., 10 times for PISA) (OECD, 2023). Then, the average across each of the k results can be taken for the final result of the analysis. This method helps minimize bias due to the randomness of PV draws (Mislevy, 1991; Braun and von Davier, 2017). However, if variables not included in the PV imputation model are used, it's possible some bias could be introduced. In practice, this may not be a large concern as the set of conditioning variables used by LSAs is so large, an outside variable may very well be correlated to an included variable (Braun and von Davier, 2017; Mislevy, 1991; Meng, 1994). When properly applied by pooling results across PV draws, PVs enable secondary analysts to build accurate models to answer their research questions effectively.

As independent variables, PVs can also be used with some consideration. If there is uncongeniality between the process used to generate the PVs and the secondary analyses, bias can be introduced, making the final results less accurate (Braun and von Davier, 2017; Meng, 1994). More research in this area may be needed to ensure PVs are best optimized for all forms of secondary analyses. Issues surrounding congeniality in PV estimation are discussed in more detail in a later section of this work.

To mathematically represent how secondary analyses can be carried out with PVs, we can adopt Rubin's (1987) analysis of multiple imputed datasets. We also want to quantify the uncertainty or measurement error across K plausible values of our estimate of a chosen population statistic S. The measurement or imputation variance is

$$B_K = \frac{1}{K-1}\sum_{k=1}^{K}(s_k - S)^2, \tag{1.3}$$

where B_K represents the uncertainty from the unobserved θ (Mislevy, 1991; OECD, 2009). Lastly, we then can combine the measurement variance and sampling variance (U) to obtain the total variance V as:

$$V = U + (1 + \frac{1}{K})B_k, \tag{1.4}$$

which estimates the variance of our sample statistic of interest s around S across repeated testing (Mislevy, 1991; OECD, 2009).

If we were to take a Bayesian approach, the posterior variance of s is approximated by V, conditional on the observed data (Mislevy, 1991). However, PISA takes a more frequentist approach by assuming plugin estimates as fixed. Given these variance and point estimate formulas, secondary researchers can properly handle PVs and account for estimate uncertainty in their own analyses.

Current Issues and Proposed Alternative

While previous work surrounding PVs has demonstrated their clear efficacy in the LSA world, several unanswered questions or concerns remain apparent that the present work seeks to explore.

Congeniality

One issue that particularly motivates this work surrounds congeniality between the imputer and the secondary analyst. At every stage of the analysis process, from sampling

design to secondary analysis, methodological decisions impact downstream analysts and researchers and can introduce issues such as bias into the analyses (Meng, 1994; Kaplan and Su, 2018; He et al., 2010). *Congeniality* is achieved when both the imputation and a secondary analysis model are compatible, meaning the assumptions underlying the imputation procedure do not conflict with those used in subsequent analyses. The responsibility of the imputer is therefore to specify a reasonably well-specified model with as many relevant covariates as possible, allowing for a broad range of valid and congenial secondary analyses (Meng, 1994). In the context of LSAs and PVs, congeniality is likely to fail, as the researcher who imputes the PVs is often a different individual than the secondary analyst who may be relying on a different set of assumptions (Kaplan and Yavuz, 2020). As Murray (2018) notes, the goal in practice is not necessarily to avoid congeniality failure entirely, but rather to "fail gracefully" (p. 146). Regardless, this statistical disjointedness can lead to bias in secondary analyses. Addressing uncongeniality is crucial, as failing to meet these assumptions undermines the validity of researchers' findings.

Current PV estimation procedures condition on hundreds of background variables, which likely mitigates concerns about uncongeniality since excluded variables are often correlated with those included. Nevertheless, uncongeniality remains a source of uncertainty that warrants attention. Although some evidence suggests that congeniality is less problematic when the PV estimation model is correctly specified with a sufficiently large sample (Marsman et al., 2016), these mixed findings highlight the need for further research to address uncongeniality and better maximize the validity and reliability of LSA results. It is therefore crucial that the imputation model properly accounts for uncertainty and utilizes all relevant information, including considerations around the use of PCA.

Principal Components Analysis

Mislevy et al. (1992) note that as the number of background variables used to estimate PVs has increased over time, researchers adopted PCA on the background variables

for computational ease. This suggests that the implementation of PCA is driven more by pragmatic considerations than by theoretical or statistical advantages. While using PCA to reduce the dimensionality of the background variables is practically reasonable, it raises several methodological issues. First, employing PCA in the imputation model introduces uncongeniality. Although PCA generates orthogonal linear combinations of the original variables, most secondary analyses (e.g., regression models) assume correlations among variables or a specific structural form, such as hierarchical. Because these assumptions are incompatible with the orthogonalized principal components, PVs derived from PCA can lead to bias or inaccurate results in secondary analyses (Benton, 2017). While current PV methods have been used for decades (Mislevy, 1991; Mislevy et al., 1992), the accuracy of the conditioning model, particularly regarding PCA-based dimensionality reduction, remains an active area of research (see also Benton, 2017; Jewsbury et al., 2025; Jewsbury and Johnson, 2025).

Bayesian Implementation

Kaplan (2023) states that taking a Bayesian approach to statistical inference enables additional flexibility and adaptability to the data at hand, better quantifies sources of uncertainty, and supports more straightforward probabilistic interpretations of results, among other benefits of such a framework. More recently, advances in computing power availability, estimation methods, and software have made Bayesian approaches increasingly accessible and practical to implement. Currently, PV estimation incorporates Bayesian elements, but is not a fully Bayesian procedure.

Fully Bayesian approaches to PV estimation are not new. Prior work has introduced Markov chain Monte Carlo (MCMC) methods within hierarchical models that account for complex sampling designs, as an alternative to methods used in NAEP (Johnson and Jenkins, 2004). These approaches offer advantages over traditional methods, including single-step estimation that captures additional sources of uncertainty and improved parameter recovery

in simulations. MCMC-based Bayesian modeling is far more computationally intensive than ML methods, making it challenging for large-scale assessments like PISA, which involve massive samples and hundreds of variables. Fully Bayesian IRT is therefore not implemented in the present work, as it would be computationally demanding while likely producing similar results given the large sample sizes.

1.2 Overview of the Three Studies

This dissertation seeks to explore the above discussed issues in current plausible value estimation methodologies. The alternatives proposed in this work will focus on modifications to the population model prior specification, as this literature review suggests that this aspect of PV estimation is crucial in ensuring PV accuracy and usefulness. It's worth considering these alternatives and others in order to implement PV procedures in a way that allows for relatively straightforward estimation while producing accurate and easily usable results for secondary researchers. More sources of uncertainty can be accounted for, more prior flexibility can be introduced, and efficient estimation processes should be explored to ensure PV procedures are optimized for accuracy and streamlining.

Across all three studies, both simulation and empirical analyses using PISA 2022 reading data are conducted. Simulation studies are designed to establish a baseline, or ground truth, given that plausible values represent imputed estimates of student proficiency rather than directly observable scores. Because true proficiency cannot be observed in practice, simulated data allow for direct comparison of plausible value estimation methods against the known population distribution of proficiency.

The simulation results are used to inform the empirical analyses, which constitute the primary focus of this dissertation, as the overarching goal is to evaluate methods intended for application in real-world settings. The empirical studies draw on PISA 2022 reading data from 76 participating jurisdictions.

Study 1: Bayesian Model Averaging

A common problem with statistical model building is that the final model the end-user is presented with, such as the secondary analyst with LSAs, is treated as if that was the model specified from the outset. In practice, models are often tweaked and re-specified to produce stronger results. This process introduces an additional component of uncertainty, as reliance on a single selected model ignores the uncertainty associated with the model choice itself (Hoeting et al., 1999).

To account for this uncertainty, we can employ model ensembling strategies that do not require the user to specify any given model, but combine the best components of several models to optimize prediction. One such method is Bayesian Model Averaging (Raftery, 1995; Raftery et al., 1997). Essentially, BMA takes a weighted average of models within a set of selected possible models.

In the context of plausible value estimation, BMA offers a way to mitigate bias from population model misspecification. Specifically, we can regress a frequentist IRT-based estimate of latent ability on the full set of background variables while allowing BMA to determine the optimal combination and weighting of these variables. This proposition forms the basis of my first study. BMA appropriately accounts for uncertainty of model and variable selection, as the user does not even need to select a single pre-specified model and instead provides the set of variables to be explored in the model space. For LSAs, this means that the entire set of conditioning variables can be explored, producing PVs that better quantifies the uncertainty surrounding said estimates.

Newer advances in BMA have made it even more appealing and adaptable to the plausible value framework. Specifically, Kaplan and Yavuz (2020) introduce using BMA in multiple imputation for context questionnaire items (i.e., background characteristics) in LSAs. Given the demonstrated effectiveness of these methods in other imputation contexts, it's possible that integrating them into estimating the plausible values themselves may be beneficial in terms of addressing sources of uncertainty and the downstream impact for

secondary analyses. Additionally, adopting a Bayesian approach to PV estimation present advantages not seen with a frequentist approach, as noted earlier.

Implementing BMA to PV estimation presents several potential benefits. First, BMA avoids the need for a single imputation model choice by the imputer, thereby accounting for uncertainty in the selection of a model itself. Next, model-averaged predictions consistently outperform any individual model in the model space (Raftery, 1995; Kaplan and Yavuz, 2020), making BMA a strong choice for maximizing accuracy in model specification. BMA also reduces collinearity between predictors, a key benefit of PCA that is currently used by LSAs. Lastly, BMA is straightforward to execute with a large set of variables as the imputer is allowing the algorithm to decide which effects or variables are most important in order to maximize prediction. However, BMA programs in `R` such as `BMS` (Feldkircher and Zeugner, 2015) do not easily account for potential interactions between predictors, a potential drawback.

Purpose

BMA is an appealing alternative to current PV methods, as it addresses imputation model selection uncertainty via averaging over all possible models in the model space. BMA also helps address issues of congeniality in that BMA maintains the linear relationship between the set of predictors and the outcome of interest (student achievement). Because BMA is fully Bayesian, it is well suited for PV imputation by producing a posterior distribution that addresses some of the non-Building on this, the first study investigates whether incorporating BMA into PV estimation improves the representation of the population distribution of student ability and the predictive performance of secondary analyses.

Study 2: Inducing Sparsity

At this point in time, PISA implements PCA to reduce the size of the large set of conditioning variables for the population model to make it more manageable in their PV

generating processes. However, given the literature previously discussed, it remains unclear if that is truly the best practice to capture the variance and the relevant predictors to most effectively estimate plausible values.

Another potential alternative to better handle the set of background variables used in the estimation of PVs is introducing shrinkage via Bayesian sparsity-inducing priors such as the ridge or regularized horseshoe priors (Hsiang, 1975; Piironen and Vehtari, 2017). These priors have previously been well documented to induce sparsity in linear regression or more complex structural equation models, particularly with models of many variables and small effects (Harra and Kaplan, 2024; van Erp et al., 2019). Bayesian regularization techniques have several desirable characteristics, including the ease of implementation via specific prior distributions and the flexibility on prior choice and hyperparameter tweaking (Harra and Kaplan, 2024). These methods may allow researchers particular flexibility that their current procedures do not allow for.

There are numerous potential benefits of such a method (Harra and Kaplan, 2024; van Erp et al., 2019). Bayesian regularization is a single-step procedure used to select relevant variables and reduce collinearity. This way, the set of conditioning variables can be reduced in size without forfeiting that explanation of variance, all the while streamlining the process of PV estimation. Introducing sparsity-inducing priors such as the regularized horseshoe also allow the imputation model to account for important linear interactions between covariates in a single analytic step.

Bayesian regularization penalizes small regression coefficients by attaching a prior distribution to model parameters (Jacobucci and Grimm, 2018). Researchers have many regularization priors to choose from, beginning with the ridge prior (Hsiang, 1975) that seeks to shrink parameters close to zero and minimize collinearity. The Bayesian lasso (Park and Casella, 2008) improves upon the ridge prior as it enables shrinkage of coefficients to zero, allowing for variable selection.

The Bayesian ridge and lasso priors are extensions of frequentist methods to the

Bayesian context. Strictly Bayesian approaches include the horseshoe prior (Carvalho et al., 2009, 2010), which allows for greater shrinkage than the ridge and the lasso while maintaining unregularized large coefficients. Additionally, the regularized horseshoe (Piironen and Vehtari, 2017), which is sometimes referred to as the *Finnish Horseshoe*, prevents large coefficients from escaping shrinkage, allows further flexibility than the original horseshoe prior, and has been shown to further improve model predictive performance (Piironen and Vehtari, 2017; Harra and Kaplan, 2024). A more recent development is the R2D2 prior (Zhang et al., 2022), which directly places a prior on the model's coefficient of determination (R^2), introducing a way of controlling shrinkage through a global explanation of variance parameter.

Purpose

As uncongeniality and bias may be introduced when performing PCA on the set of conditioning variables, I propose that PV estimation via a sparsity-inducing prior may be better suited for PV estimation. This proposed alternative is more straightforward, allows for the data themselves to reduce dimensionality without sacrificing congeniality, generates PVs via a proper posterior distribution, and could improve out-of-sample predictive performance for increased generalizability with secondary analyses.

Study 3: Bayesian additive regression trees

Lastly, PV estimation via machine learning methods such as Bayesian additive regression trees (BART) may be a reasonable alternative to current PV estimation methods.

BART is a nonparametric regression method using a sum-of-trees model, motivated by other model ensembling and machine learning methods in a fully-Bayesian framework (Chipman et al., 2010). By summing many smaller and "weaker" trees, BART avoids overfitting and can conduct variable selection with strong predictive performance (Chipman et al., 2010; Ročková and Saha, 2019). Rather than relying on a single complex model that may be prone to overfitting and misspecification, BART combines many weaker approximations of

the true data-generating process to improve overall performance without making such strong assumptions. By ensembling many smaller trees, BART implicitly can perform variable selection based on which variables the algorithm deems important to include as meaningful node splits.

With BART, regularization priors are placed on all model parameters to minimize disproportionately impacting individual regression trees in the ensemble (Chipman et al., 2010). This prevents from any one tree from dominating the ensemble. Model fitting is done with iterative back-fitting via a Gibbs sampler across a user-specified number of trees, m, updating individual trees conditioned on the residuals of the others. This process yields posterior draws of $h(x)$ and forms a full posterior distribution of predictions for the given outcome, Y.

There are several statistical and practical benefits to using BART for prediction. First, Chipman et al. (2010) find that BART outperforms other machine learning methods such as random forests and boosting in terms of out-of-sample predictive performance and that its fully-Bayesian methods allow for straightforward interpretations and uncertainty quantification surrounding parameter estimates. BART's nonparametric framework imposes fewer assumptions on the data-generating process and functional forms (e.g., linearity) than traditional modeling approaches. By combining this flexibility with its model-ensembling methods, BART reduces the risk of model misspecification while simultaneously addressing model uncertainty and potential uncongeniality between the imputation model and secondary analyses. These properties make BART an appealing potential alternative for PV generation.

Previous work on BART highlights that it's adaptable to many modeling approaches (Chipman et al., 2010; Hill et al., 2020), however the present work will focus on nonparametric regression for generating plausible values. The flexibility and generalizability of BART, combined with the advantages of Bayesian regularization and uncertainty quantification, make it a promising approach for PV estimation. Its streamlined fully-Bayesian approach has shown more accuracy than other ensembling methods in the machine learning world.

While BART has been well studied and used in the context of machine learning, it appears to be underutilized in the educational statistics and large-scale assessment space. The present work seeks to explore how we can bridge between these two fields to combine the best practices of both to yield optimally predictive and unbiased models in more efficient ways.

Purpose

Unlike traditional approaches, BART is a nonparametric method that makes minimal assumptions about the functional form of the relationships in the data. Additionally, BART inherently accounts for model uncertainty through its ensemble of regression trees, reducing the risk of model misspecification and providing more robust estimates of student abilities. Motivated by these clear benefits of BART, the purpose of this study to see what improvements in terms of accuracy and predictive performance implementing estimation of PVs via BART as opposed to current methods can bring.

Evaluating Predictive Performance

Policy decision-making informed by LSA results are rarely based solely on average student outcomes. Instead, they often target specific subpopulations of interest, such as high-risk, underperforming, or high-achieving students, who lie in the tails of the proficiency distribution. Consequently, plausible value imputation models optimized only for mean performance may fail to capture population heterogeneity and can perform poorly for extreme cases. This limitation can lead to incomplete or potentially misleading inferences, reducing the effectiveness of policy decisions. To maximize the utility of LSAs and PV-based analyses, modeling approaches should aim to capture the full distribution of the population of interest, provide reliable predictions for extreme observations, and remain robust across populations.

Accordingly, the results presented in this dissertation, particularly in the secondary analysis examples, move beyond average-based fit metrics and instead prioritize out-of-sample

predictive performance using the leave-one-out information criterion (LOO-IC) (Vehtari et al., 2017), which is derived from the Bayesian leave-one-out cross-validation (LOO-CV).

LOO-CV is a special case of k-fold cross-validation in which $k = n$, so that each observation is treated as its own validation set. In this framework, the model is repeatedly fit to the data with a single observation omitted, and predictive performance is assessed on that held-out observation. This procedure is particularly well-suited for estimating pointwise out-of-sample predictive accuracy (Allen, 1974; Stone, 1974).

The corresponding *expected log pointwise predictive density* (ELPD) for LOO-CV, denoted ELPD_{loo}, is defined as:

$$\text{ELPD}_{loo} = \sum_{i=1}^{n} \log p(y_i \mid y_{-i}), \quad \text{where} \quad p(y_i \mid y_{-i}) = \int p(y_i \mid \theta)\, p(\theta \mid y_{-i})\, d\theta. \tag{1.5}$$

Here, $p(y_i \mid y_{-i})$ is the LOO predictive density given the data with the i^{th} data point left out (Vehtari et al., 2017). The log sum of these predictive densities in Equation (1.5) is the LOO-CV estimate of the ELPD (Gelman et al., 2014; Gronau and Wagenmakers, 2019; Vehtari et al., 2017). An information criterion based on LOO, referred to as the *LOO-IC*, can be derived as

$$\text{LOO-IC} = -2\,\widehat{\text{ELPD}}_{loo} \tag{1.6}$$

which places the LOO-IC on the deviance scale. Among a set of competing models, the one with the smallest LOO-IC is considered best from an out-of-sample point-wise predictive point of view. This approach helps ensure that selected models better reflect the full range of outcomes relevant for policy decision-making.

1.3 Contributions & Discussion

Previous research highlights the benefits of plausible values (PVs), particularly their statistical consistency, improved estimation accuracy, and ability to support subgroup and secondary analyses (Mislevy et al., 1992; Wu, 2005; von Davier et al., 2009). Current methodologies used in PISA are not particularly computationally demanding given the transformation of the background variables into principal components, combined with the use of plugin fixed estimates and maximum likelihood estimation procedures. However, there are several areas of ambiguity or concern that may require future research and estimation refinement to remedy these issues and bridge the gap between the current PV methodologies with newer statistical modeling approaches more readily available.

The present dissertation seeks to address currently unanswered or unaccounted for issues of plausible value estimation techniques. These proposed alternatives include PV estimation methods Bayesian Model Averaging (BMA), incorporating sparsity via Bayesian regularization priors, and Bayesian additive regression trees (BART). With a comprehensive study of these proposed methods, my dissertation seeks to offer more concrete recommendations on which PV method(s) should be used, and if different methods provide unique benefits. These recommendations would help LSAs and stakeholders alike improve the value and impact of such assessments. Building on the findings from my three studies, this dissertation not only considers technical and methodological aspects of PV estimation but also addresses their practical implications. Plausible values are a crucial tool for disseminating LSA findings and supporting secondary research in education and the social sciences. Ensuring that the procedures used to estimate plausible values are fully optimized is key to maximize the benefits for all researchers and policymakers involved in such research.

2 STUDY 1: BAYESIAN MODEL AVERAGING

As discussed in Chapter 1, Bayesian Model Averaging (BMA) provides a principled approach for addressing population model misspecification, model uncertainty, and uncongeniality in plausible value (PV) estimation. Rather than relying on a single conditioning model, BMA averages over a space of candidate models (Raftery, 1995; Raftery et al., 1997), allowing the full set of background variables to inform estimation while explicitly accounting for model uncertainty. This eliminates the need for dimension reduction techniques such as principal components analysis (PCA). In the context of large-scale assessments (LSAs), this allows for a more flexible use of conditioning variables and a more complete quantification of uncertainty in PVs.

Incorporating BMA directly into PV estimation offers several advantages. It removes the need to pre-specify a single imputation model, improves predictive performance through model averaging, and aligns naturally with the Bayesian foundations of PV methodology (Raftery et al., 1997; Kaplan, 2021). Additionally, by incorporating the full set of predictors, BMA can help mitigate multicollinearity without requiring orthogonalization.

This study investigates whether integrating BMA into PV estimation improves recovery of the population distribution of student ability and enhances the predictive performance of secondary analyses. The chapter proceeds as follows: first, a technical overview of BMA and its integration into PV estimation (BMA-PVs) is presented. Next, a simulation study and an empirical application using PISA 2022 reading data evaluate the statistical and practical performance of the proposed approach. The chapter concludes with a discussion of findings and implications.

2.1 BMA Technical Background

A key limitation of existing PV approaches is the reliance on a single specified population model, which is typically treated as correct despite uncertainty in model selection.

In practice, model specification often involves modification and re-specification to produce stronger results. This process introduces an additional aspect of uncertainty, as reliance on a single selected model ignores the uncertainty associated with the model choice itself, resulting in inaccurate inferences and mis-calibrated predictions (Hoeting et al., 1999). To account for this uncertainty, we can employ model ensembling strategies that do not require the user to specify any given model, but combine the best components of several models to optimize prediction.

One such method is Bayesian Model Averaging (Leamer, 1978; Raftery, 1995; Raftery et al., 1997). Essentially, BMA takes a weighted average of models within a set of selected candidate models. We define the average of the posterior distribution for some statistic of interest, s, given our data D, as follows:

$$P(s|D) = \sum_{k=1}^{K} P(s|M_k, D)P(M_k|D), \tag{2.1}$$

where the posterior probability of the model (PMP) M_k is denoted as

$$P(M_k|D) = \frac{P(D|M_k)P(M_k)}{\sum_{\ell=1}^{K} P(D|M_\ell)P(M_\ell)}, \qquad \ell \neq k. \tag{2.2}$$

The term $P(D|M_k)$, which likely differs between models, can be written as a marginal likelihood

$$P(D|M_k) = \int P(D|\theta_k, M_k)P(\theta_k|M_k)d\theta_k, \tag{2.3}$$

where $P(\theta_k|M_k)$ is the prior distribution of θ_k under model M_k (Raftery et al., 1997). BMA thus provides a framework for combining many models, specified by the set of potentially relevant predictors. In the context of plausible value estimation, BMA offers a way to mitigate bias from population model misspecification. Specifically, we can regress a frequentist IRT-based estimate of latent ability on the full set of background variables while allowing BMA to determine the optimal combination and weighting of these variables. This proposition forms the basis of this chapter.

BMA appropriately accounts model and variable selection uncertainty, as the user does not even need to select a single pre-specified model and instead provides the set of variables to be explored in the linear model space. For LSAs, this means that the entire set of conditioning variables can be explored, producing PVs that better quantifies the uncertainty surrounding said estimates of latent ability.

The $\mathcal{M}$-Closed Assumption

An essential assumption in BMA surrounds the belief that the "true" model M_T, that is the data-generating model, is contained within our set of candidate models $\mathcal{M}$. In other words, among all possible combinations of background variables considered, one of these models is assumed to represent the true process generating latent ability in the population.

The assumption that a true model both exists and is contained within the user's specified model space is known as the $\mathcal{M}$-closed framework (Bernardo and Smith, 2000; Raftery, 1995; Raftery et al., 1997; Hoeting et al., 1999; Kaplan, 2021). Under this assumption, we can assign prior probabilities not only to parameters within each model, but also to the models themselves. This reflects our prior belief of the likelihood a given model is M_T. This $\mathcal{M}$-closed framework is of particular relevance of PV estimation in that we are aiming to better quantify uncertainty surrounding imputation model selection, improving accuracy for population-level inferences and secondary analyses.

Prior Settings for BMA

With BMA, it is necessary to specify priors on both individual model parameters and on the models themselves. Parameter priors are assigned to each coefficient in a given model. Model priors are used to assign prior probability that a given model is M_T, or the data-generating model. These prior settings allow users to quantify uncertainty at both the parameter and model selection levels.

Parameter priors use variants of Zellner's g-prior (Zellner, 1986). The g-prior placed on

regression coefficients, β, follows a natural-conjugate Normal-Gamma distribution. Following Kaplan (2021), we can specify this prior for a given model M_k:

$$y_i = x_i'\beta + \varepsilon, \tag{2.4}$$

where Zellner's g-prior can be written as

$$\beta_k|\sigma^2, M_k, g \sim \mathcal{N}\Big(0, \sigma^2 g\big(x_k' x_k\big)^{-1}\Big). \tag{2.5}$$

Variations on the g-prior include fixed priors and flexible priors (Kaplan, 2021). These priors seek to correct instability that can be caused by a poorly chosen g (Feldkircher and Zeugner, 2009).

Fixed parameter priors include several options. First, there is the *unit information prior* (UIP), where $g = N$ (Liang et al., 2008). The *risk inflation criterion prior* (RIC) uses $g = Q^2$ with Q predictors (Fernández et al., 2001). The last fixed parameter prior is the *Hannan and Quinn prior* (HQ), where $g = \log(N)^3$, originating from a model selection criterion (Hannan and Quinn, 1979).

Only one flexible prior is included in this study, as convergence issues arose when implementing others. Specifically, I examine the *local empirical Bayes* (EBL) prior, which estimates the hyperparameter g via maximum likelihood using the observed data for each model k:

$$g_k = \arg\max(0, F_k - 1), \tag{2.6a}$$

$$F_k = \frac{R_k^2(N - Q_k - 1)}{(1 - R_K^2)Q_k}. \tag{2.6b}$$

Other flexible priors are available, but previous work has found minimal differences in performance across common specifications in large-sample settings (Kaplan and Huang, 2021; Fernández et al., 2001). These findings suggest that parameter prior selection is unlikely to meaningfully affect results in the present context.

While parameter priors govern the distribution of regression coefficients within each candidate model, model priors assign probability mass across the model space itself. The `BMS` package used in this work provides three model prior specifications: uniform, fixed, and random (Feldkircher and Zeugner, 2015). The *uniform* prior assigns equal probability to all models in the model space. The *fixed* prior imposes a binomial distribution on model size with the prior inclusion probability, or the probability that a given predictor is included in the true model, fixed at a specified value for each covariate:

$$p(M_k) = \theta^{q_k}(1-\theta)^{Q-q_m}, \tag{2.7}$$

where each predictor q_k has a θ inclusion probability (Zeugner and Feldkircher, 2015). A more flexible alternative is the *random* model prior, which uses a beta-binomial formulation in which the inclusion probability θ follows a Beta distribution, $\theta \sim \text{Beta}(a, b)$ (Ley and Steel, 2009). The present work evaluates both model and parameter prior selection impact on PV estimation.

2.2 Present Study

Newer applications of BMA within the multiple imputation literature make it particularly well suited for adaptation to the plausible value imputation framework, which can be viewed as a special case of missing data imputation. Specifically, Kaplan and Yavuz (2020) introduced using BMA in multiple imputation for context questionnaire items (i.e., background characteristics) in LSAs and found that BMA better accounts for the model uncertainty present in existing methods of missing data imputation and more accurately recovers the true population distribution. Given the demonstrated effectiveness of these methods in other imputation contexts, it's possible that integrating them into estimating the plausible values themselves may be beneficial in terms of addressing sources of uncertainty and the downstream impact for secondary analyses.

As previously discussed, BMA is an appealing alternative to current PV methods, as it addresses imputation model selection uncertainty via averaging over all possible models in the model space. BMA also helps address congeniality by preserving the correlational relationship between the predictors and the outcome of interest (student achievement), an assumption underlying many secondary analyses. Also, BMA is a fully-Bayesian procedure, making it well-aligned with a Bayes-like posterior distribution framework, addressing the non-Bayesian components of the current PV methods. Multiple random draws from the resulting posterior of each student can be used to form PV population distributions.

For this study, I aim to investigate whether incorporating BMA into PV estimation can better capture the population distribution of student ability, as well as aid in predictive performance of secondary analyses, via both a simulation study and an empirical example using PISA 2022 reading data.

2.3 Methods

Simulation Study

To study how BMA-estimated PVs (BMA-PVs) compare to current PV methods, I propose a comprehensive Monte Carlo simulation study. Data are simulated using the `lsasim` package (Matta et al., 2018) in `R` to best mimic a U.S. sample of PISA for reading achievement ($N = 5000$). I investigate PV accuracy and performance using BMA via `BMS` (Feldkircher and Zeugner, 2015). Model-averaged predictions provide estimates of student ability based on IRT-based scores. The entire set of background variables ($k = 500$) are used as the predictors. Random draws from the resulting posterior distributions are used to directly form the PVs for each student.

To compare BMA-PVs with those generated using current methods, I use the `TAM` package (Robitzsch et al., 2025) to generate 10 PVs for each simulated student based on their item responses, conditioned on the principal components that explain 90% of the variance of

the background variables, following LSA procedures (Braun and von Davier, 2017).

To examine the impact of prior selection on BMA-PVs, four parameter priors were crossed with three model priors, yielding a total of 12 conditions. These include the parameter priors of: EBL, HQ, RIC, and UIP. The model priors are: fixed, random, and uniform. In total, 1000 replications for each condition were done, and 10 plausible values were drawn for each student and aggregated for the results.

Secondary analyses use a Bayesian linear regression model, with the 10 PVs for each student as the outcome regressed on a randomly chosen set of predictors ($p = 9$). PV performance is measured in the form of the leave-one-out cross-validation information criterion (LOO-IC), which is arguably the best out-of-sample predictive performance scoring rule (Vehtari et al., 2017; Harra and Kaplan, 2024), described in the previous chapter.

Empirical Study

For the empirical study, I use PISA 2022 reading data from the 76 countries that had publicly available item response and background items collected. The full list of countries in this analytic sample are contained in Table 2.1. While four other countries (Cambodia, Guatemala, Paraguay, and Vietnam) were PISA 2022 reading participants, their item response data were not publicly available for download. These nations administered PISA 2022 on paper instead of digitally (OECD, 2023), making item response coding more challenging. The following analyses thus exclude these countries.

To prepare data for analysis, missing background questionnaire items were imputed using predictive mean matching via the `mice` package (Van Buuren and Groothuis-Oudshoorn, 2011). Predictive mean matching produces unbiased estimates under the assumption of ignorable missing data (Little and Rubin, 2019). In the present analysis, I assume the data are missing at random. Predictive mean matching imputes missing values by matching the predicted values from the observed data using a predictive mean metric to the predicted values with regression imputation. Next, the observed value is used for the imputation. So

Table 2.1: Countries and economies included in the analytic sample for this paper ($N = 76$)

Albania	Estonia*	Latvia*	Portugal*
Argentina	Finland*	Lithuania*	Qatar
Australia*	France*	Macao (China)	Romania
Austria*	Georgia	Malaysia	Saudi Arabia
Azerbaijan	Germany*	Malta	Serbia
Belgium*	Greece*	Mexico*	Singapore
Brazil	Hong Kong (China)	Moldova	Slovak Republic*
Brunei	Hungary*	Mongolia	Slovenia*
Bulgaria	Iceland*	Montenegro	Spain*
Canada*	Indonesia	Morocco	Sweden*
Chile*	Ireland*	Netherlands*	Switzerland*
Chinese Taipei	Israel*	New Zealand*	Thailand
Colombia*	Italy*	North Macedonia	Türkiye*
Costa Rica*	Jamaica	Norway*	United Arab Emirates
Croatia	Japan*	Palestinian Authority	Ukraine
Czechia*	Jordan	Panama	United Kingdom*
Denmark*	Kazakhstan	Peru	Uruguay
Dominican Republic	Korea*	Philippines	United States*
El Salvador	Kosovo	Poland*	Uzbekistan

An asterisk (*) denotes an OECD member country.

for each regression model, there is a predicted value for both observed and missing data. The predicted value for the observed data is then matched to the predicted missing data value, for example, using a nearest neighbor metric. After a match is made, the missing value is replaced by the observed value (as opposed to the predicted value). Random matches are chosen when multiple matches are found. Multiple draws for missing values should be made to best account for uncertainty surrounding the imputation. However, I only conducted this process for a single imputed data point to ensure computational ease. While using multiple imputations is considered best practice for producing the most reliable results, I argue that using a single imputation via predictive mean matching is a significant improvement over listwise deletion or using a much smaller set of complete predictors.

Item parameters were estimated via a 2PL model in `TAM` (Robitzsch et al., 2025), with a random subsample of 500 students from all participating countries, for a total of 38,000 students across 76 countries and economies. Item parameters were then held fixed for the subsequent analyses. This strategy mimics how PISA calibrates item parameters (Okubo, 2022). The resulting item parameters were found to strongly correlate with PISA's

publicly reported item parameters for both item difficulty and discrimination in a pilot study ($\rho \geq 0.8$), and are used for the subsequent analyses.

It is important to note that the results do not perfectly align with the official PISA reports in terms of country rankings and mean scores, however I found my replicate mean scores had an acceptable amount of bias ($\leq 10\%$) for all countries. This discrepancy arises from details omitted in the PISA technical reports in how PISA generates plausible values and the resulting country rankings (OECD, 2024). To best approximate the original procedure, I used the `TAM` (Robitzsch et al., 2025) package. Even when employing the PISA-reported item parameters, an exact replication was not possible but a strong rank-order correlation ($\rho = .98$) was achieved between PISA and `TAM`-replicate country rankings. This, combined with the fact that four countries were missing complete item and background questionnaire responses (see above), an exact replication of PISA 2022 reading results was not achieved. Nevertheless, these replicated PVs will serve as the baseline, or "ground truth," for all subsequent analyses.

2.4 Results

Simulation Study Results

First I examine the effects, if any, parameter and model priors had on the accuracy of the resulting BMA-PVs. I focus on the mean and standard deviation of the resulting population distribution for each PV, as these are widely reported for LSAs.

Figure 2.1 shows the distribution of BMA-estimated PVs across various prior settings. In terms of distribution means, we see virtually no differences across prior options. There is a bit more variation for the standard deviation of the population, however, all options slightly underestimate the true variability. The HQ parameter prior performed slightly better than the rest. The EBL parameter prior performed the worst, however these differences are quite small. For the remainder of this study and the following empirical study, an HQ parameter prior with a uniform model prior are used.

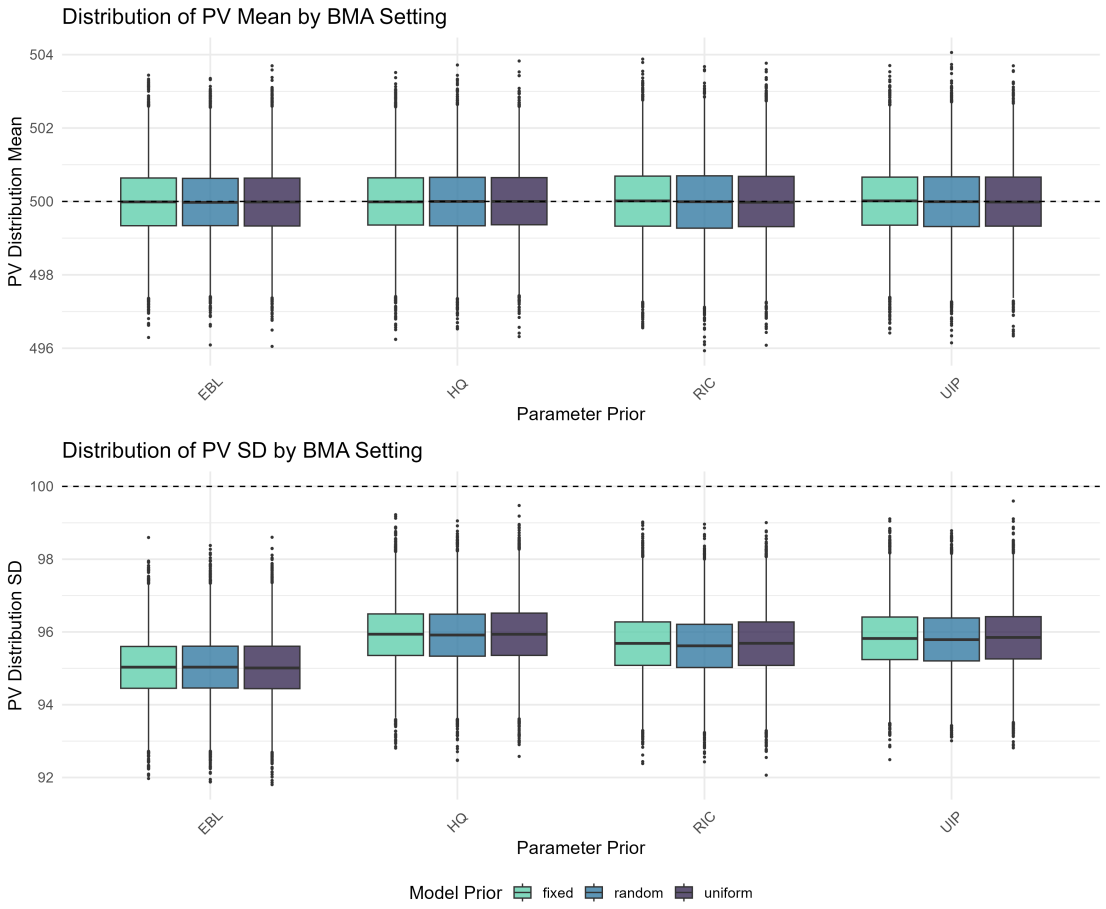


Figure 2.1: Distributions of PV means and standard deviations by BMA settings in the simulation study.

Next, BMA-PVs were compared to the current PV estimation methods and to the simulated distribution of latent ability. Figure 2.2 demonstrates that while both BMA and current methods sufficiently capture the true latent ability mean of the population, the current methods actually overestimate variability in the population, whereas BMA slightly underestimates it. We see the standard deviation of the replicated methods is about 15% higher than the true variability, and BMA is about 5% lower. These differences are small, whereas the most clear benefits of BMA-PVs are seen in a secondary analysis example shown in Figure 2.3.

Figure 2.3 shows that out-of-sample predictive performance is strongest with BMA-PVs, where the PISA-replicate's performance is notably worse. This finding supports the

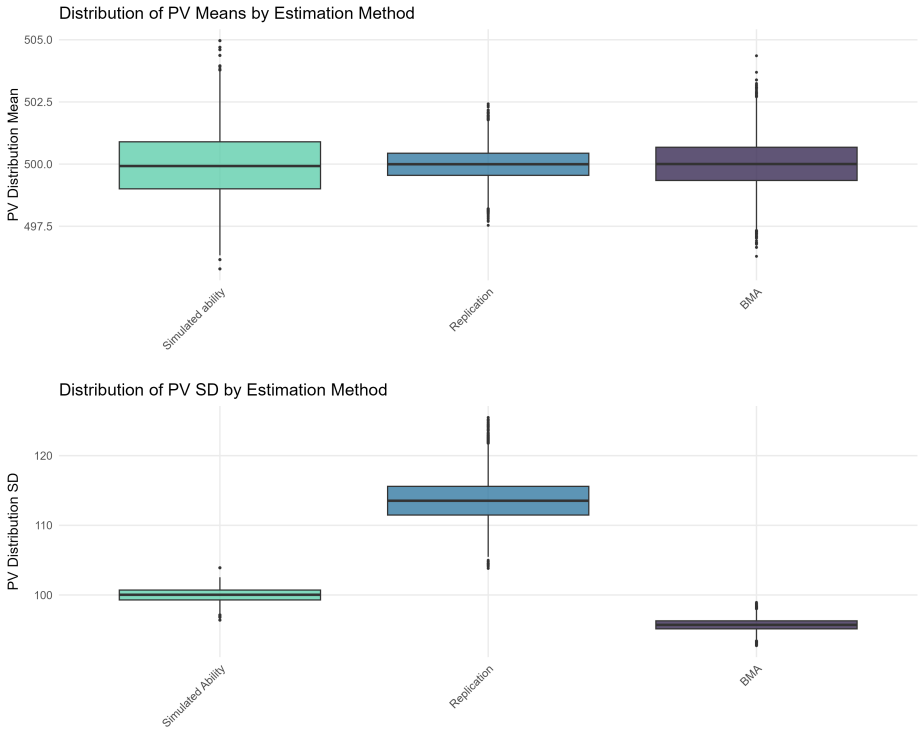


Figure 2.2: Distributions of PV means and standard deviations across methods in the simulation study.

notion that using BMA-generated PVs can induce stronger out-of-sample predictive performance in a secondary analysis predicting reading achievement, increasing the confidence and generalizability of findings for secondary researchers.

These simulation study results show that BMA-PVs can effectively capture the true population distribution of latent ability in a simulated population similar to a U.S. PISA 2022 sample. BMA-PVs are also clearly more effective in optimizing out-of-sample predictive performance in comparison to current methods, and interestingly enough, better than using the simulated latent ability scores themselves. The following section takes these methods shown to be effective in a simulated data setting into a real-world application with PISA 2022.

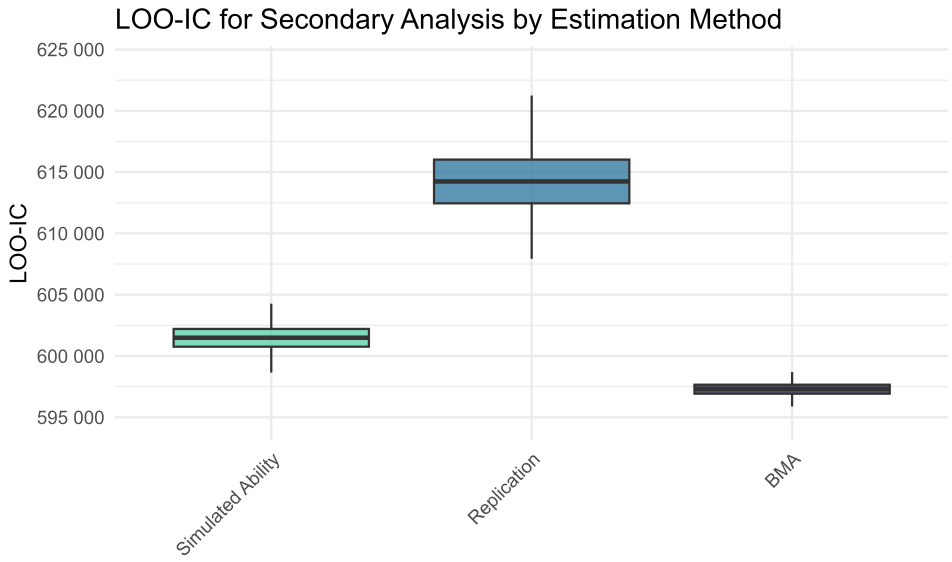


Figure 2.3: Distributions of LOO-IC's for a secondary analysis model by method in the simulation study.

Empirical Study Results

For the empirical study, 76 countries and jurisdictions with reading data were used (see Table 2.1). Country rankings and secondary analysis examples are compared to assess the performance of BMA-PVs in a real-data setting. An example of the resulting PV distributions for the United States between PISA-reported PVs, my PISA replicates using `TAM`, and BMA-PVs is contained in Figure 2.4. We can see virtually no difference between these distributions except for small amounts of random noise, highlighting BMA-PVs' sufficient population reconstruction abilities.

BMA Diagnostics

To directly quantify model uncertainty, we can look at the posterior model probability, which is the probability that any given model in our model space is the true data-generating model after seeing the data, assuming the $\mathcal{M}$-closed assumption holds. Figure 2.5 shows that the cumulative posterior probability of the top 100 models varies slightly by country, averaging around 45%. In other words, these top 100 models together account for roughly

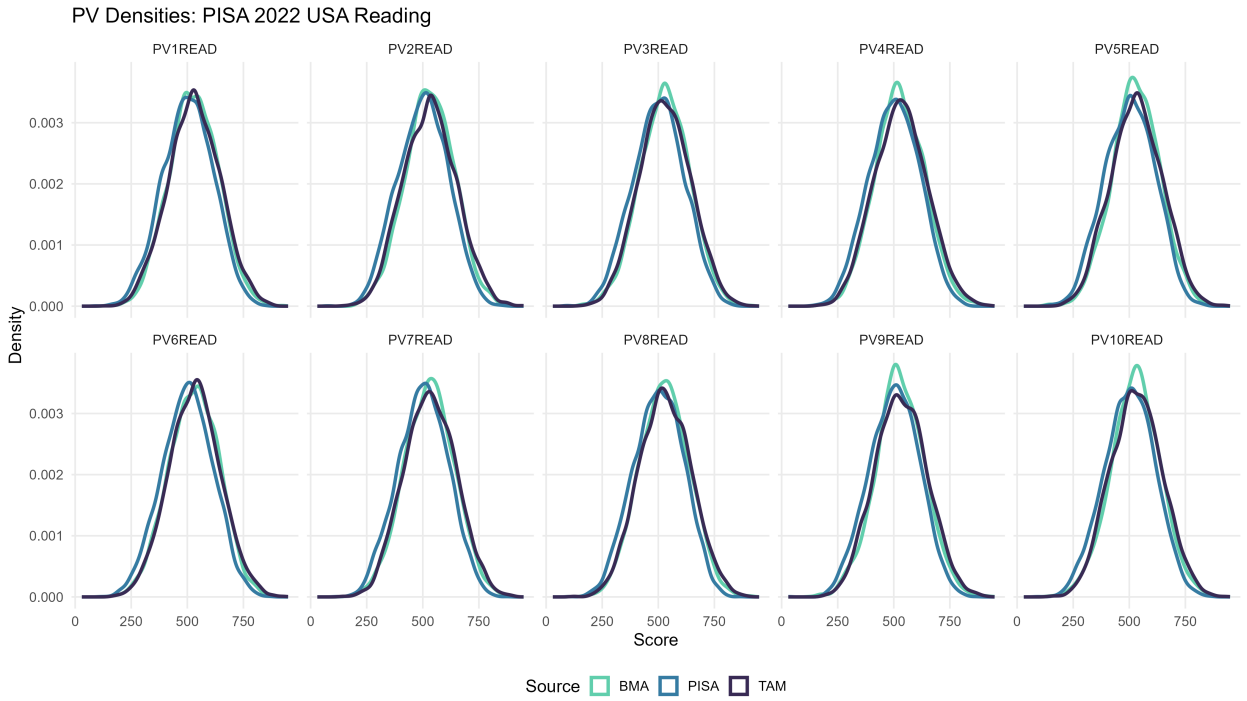


Figure 2.4: PV densities by estimation method using PISA 2022 United States (USA) reading data.

45% of the posterior probability mass, while each individual model contributes only a small fraction. Since this sum is not close to 100%, this suggests considerable model uncertainty that is not captured by conventional PV methods. Across countries, Greece had the highest PMP, at approximately 52%, and Brazil with the lowest at approximately 37%. We see that BMA-PVs can be effectively applied across countries of varying sample size and population characteristics.

Another result of interest extends beyond posterior model probabilities to examine which background characteristics most consistently emerged as influential predictors of student reading achievement. Within the BMA framework, predictors were evaluated over a model space of approximately 2^{672} candidate models for each country[1]. Because evaluating every possible model is computationally infeasible, the `BMS` package uses stochastic search algorithms to sample models with high posterior probability rather than exhaustively evaluating all

[1]For comparison, the estimated number of atoms in the observable universe is roughly $2^{266} \approx 10^{80}$ (Kiernan, 2024)

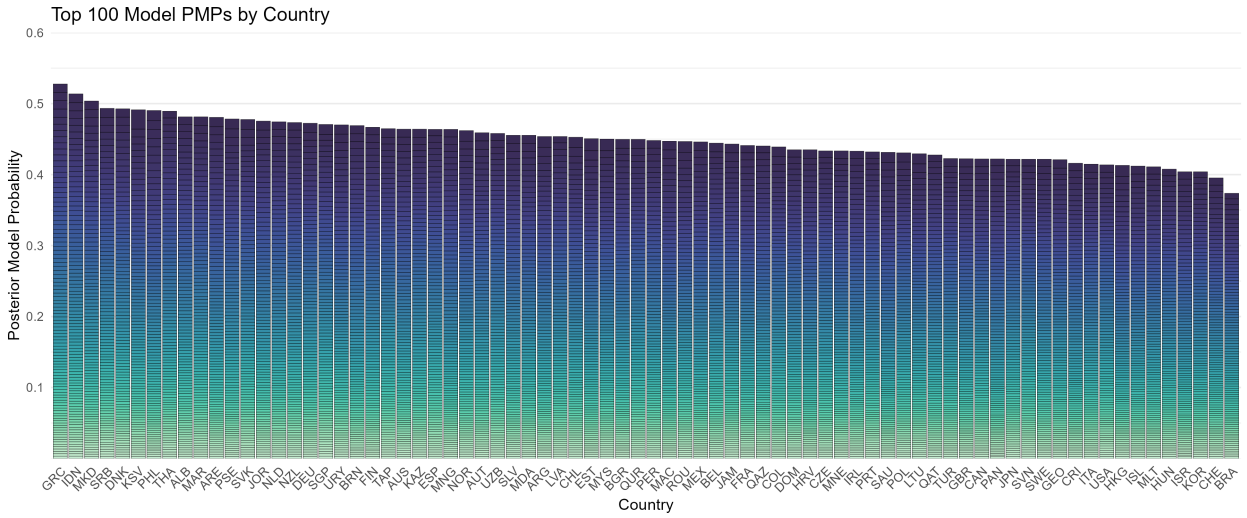


Figure 2.5: Cumulative posterior probability of the top 100 models (out of approximately 1000 visited by the algorithm) from a model space of about 2^{672} possible models.

models (Feldkircher and Zeugner, 2015). Approximately 1,000 models per country were therefore explored. Predictors that appeared frequently in the posterior-weighted model space exhibited high posterior inclusion probabilities (PIPs), defined as the estimated proportion of posterior probability-weighted models visited by the algorithm that include the variable. Table 2.2 reports, across the 76 countries, the frequency with which each predictor ranked among the most important, as determined by having the highest PIPs within each country.

Table 2.2: Predictors most frequently identified as top predictors across countries based on PIPs. The *Frequency* column reports the number of countries (out of 76) in which each variable ranks among the top 10 predictors by PIP.

Predictor	Frequency	Predictor Description
ST253Q01JA	18	Number of digital devices with screens at home.
ST001D01T	8	Student international grade.
ST255Q01JA	8	Number of books in home.
ST297Q09JA	8	Participation in additional math instruction.
EFFORT2	7	Post-assessment intended effort.
PROGN	7	Unique national program study code.
ST059Q02JA	7	Number of total class periods per week (all subjects).
ST256Q03JA	7	Number of contemporary literature books at home.
ST022Q01TA	6	Language most often spoken at home (language of test or not).
ST294Q05JA	6	Number of days a week before school exercising/practicing sport.

Across countries, reading proficiency appears to be commonly associated with both a student's home resources and classroom engagement. For example, the most frequently selected top predictor, appearing among the highest-ranked variables via posterior inclusion probabilities in 18 countries, is the student-reported number of digital devices with screens at home. Other factors, while not consistently among the top predictors based on the PIPs, still contribute meaningfully to predicting reading achievement in PISA. This distinction is important both substantively and statistically, as including the same set of predictors in both the imputation models and secondary analysis models helps ensure congeniality and supports the validity of the resulting estimates.

Country Rankings

As stated earlier, a key aspect of PISA reporting surrounds country rankings of reading achievement (Özer, 2020; Araujo et al., 2017; Martens and Niemann, 2013). These rankings rely on PVs to compare country and jurisdiction performance. Figure 2.6 depicts changes in country rankings for PISA 2022 reading mean scores. To understand what changes we see, note that filled-in circles represent the PISA-replicate rank, and the empty circles denote BMA-PV ranks. Line segments between circles depict what difference, if any, arises between the two methods of PV generation.

Changes in country rankings with BMA-PVs are generally small, and are primarily concentrated among top-performing countries. In the upper half of the distribution, rank changes are likely more frequent due to closer clustering of mean estimates; even minor adjustments to estimation procedures can reorder countries. For example, Czechia (CZE) drops from 4th place in the PISA-replicate ranks to 7th with BMA-PVs, whereas Denmark (DNK) rises from 25th to 21st. In contrast, countries in the lower half of the distribution exhibit fewer ranking changes, likely due to wider gaps between countries decreasing sensitivities to methodology changes. For instance, Albania (ALB) moves from 68th to 67th. These results suggest that incorporating model uncertainty through BMA-PVs improves the statistical

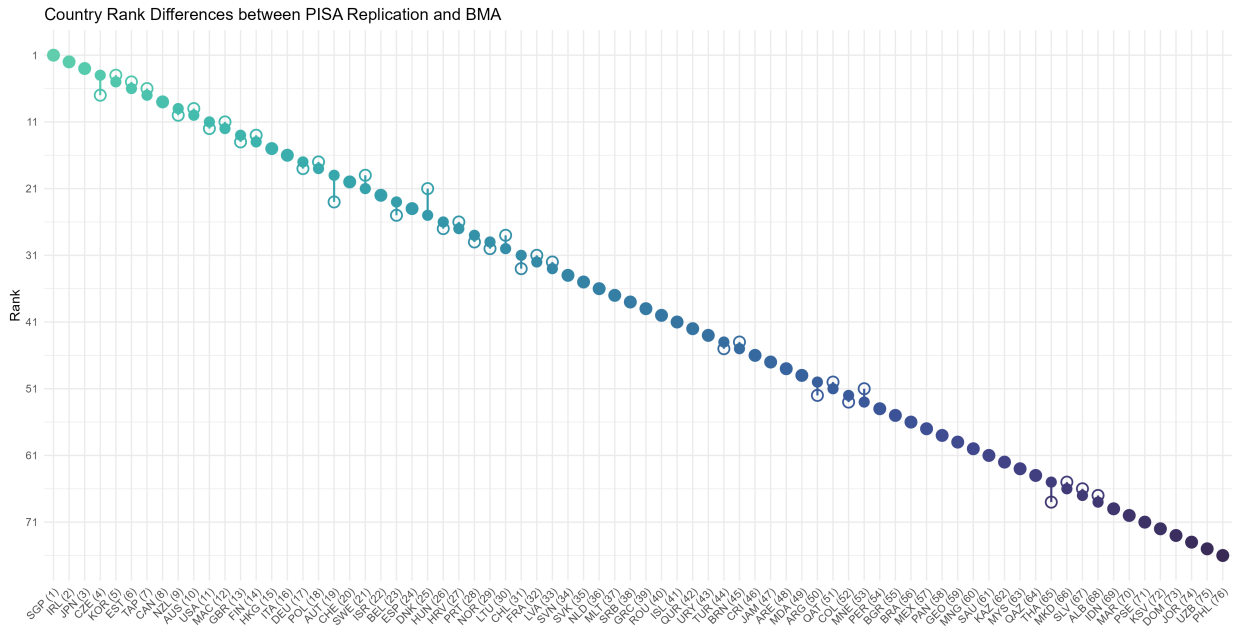


Figure 2.6: Country rankings of PISA-replicates using `TAM` and BMA-PVs for reading in 2022. Note the PISA-replicate ranks in parentheses on the bottom labels. Filled-in circles denote the PISA-replicate rank, and the empty circles denote BMA-PV ranks. Line segments between circles depict the magnitude of disagreement.

properties of PV estimation without substantially altering overall country rankings. That said, even small shifts in country rankings can have meaningful policy and political consequences, particularly for countries that place substantial weight on PISA results in educational planning and international comparisons. Changes of even a few positions could influence public perception, resource allocation, and policy priorities. While the changes shown in the present study are relatively modest, the advantages of using BMA-PVs become more evident in secondary analysis, where accounting for model uncertainty by incorporating information across multiple possible models can lead to more robust and reliable inferences.

Secondary Analysis

As discussed previously, an important use of PVs is in applied secondary analyses in a variety of fields. Reading achievement plausible values are commonly used as an outcome of interest, as predicted by background variables. To examine how well BMA-PVs can be

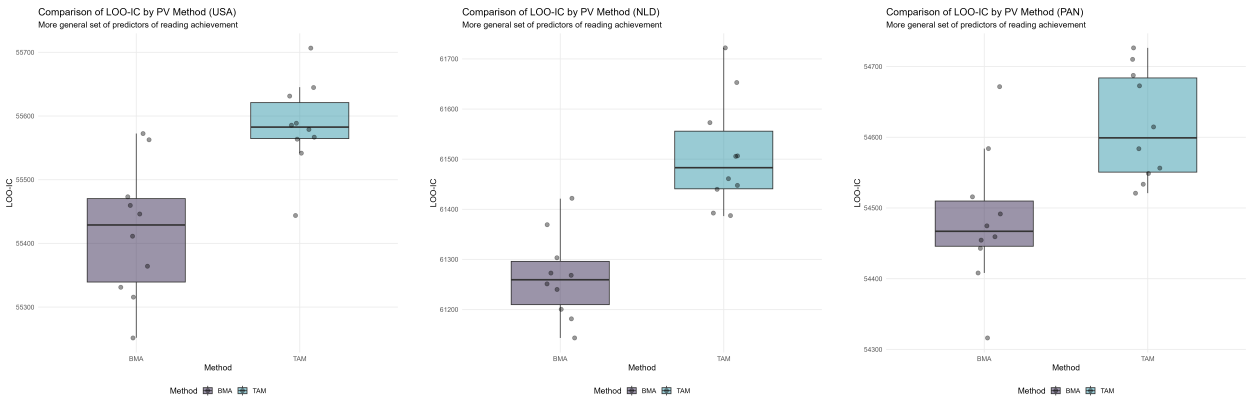
used in these types of analyses, I use an example Bayesian linear regression model to study predictive performance. To illustrate performance across varying positions in the achievement distribution, I examine three countries: the United States, the Netherlands, and Panama, representing different points in overall country rankings. This analysis was conducted 10 times for each country and method, one for each PV as the outcome variable. Results are pooled and described below.

The model of interest to predict reading achievement within a given country for each student i can be specified as:

$$\begin{aligned} \text{READ}_i = \beta_0 + \beta_1 \text{ESCS}_i + \beta_2 \text{HISCED}_i + \beta_3 \text{IMMIG}_i + \beta_4 \text{LANGN}_i \\ + \beta_5 \text{SEX}_i + \beta_6 \text{AGE}_i + \beta_7 \text{HISEI}_i + \beta_8 \text{HOMEPOS}_i \\ + \beta_9 \text{GROSAGR}_i + \beta_{10} \text{PROBSELF}_i + \beta_{11} \text{DISCLIM}_i \\ + \beta_{12} \text{RELATST}_i + \beta_{13} \text{BELONG}_i + \beta_{14} \text{EFFORT1}_i + \epsilon_i \end{aligned} \tag{2.8}$$

where reading achievement is predicted by a student's socioeconomic status (ESCS_i), the highest education attainment by a parent (HISCED_i), student's immigration status (IMMIG_i), language spoken most at home (LANGN_i), sex, age, a student's parental occupational status index (HISEI_i), a home possessions index reflecting family wealth and resources (HOMEPOS_i), growth mindset index (GROSAGR_i), problems with self-directed learning index (PROBSELF_i), disciplinary climate index (DISCLIM_i), student–teacher relationship quality index (RELATST_i), sense of belonging (BELONG_i), and self-reported effort on the cognitive exam (EFFORT1_i).

Figure 2.7 displays the distribution of LOO-ICs in the secondary analysis example. Emphasis should be placed on the median line in the box plots, as a regression using just a single PV (here, represented as the points), does not constitute a sufficient estimate of model performance. Across all three countries, BMA-PVs yield lower average LOO-ICs, indicating improved out-of-sample predictive performance relative to current approaches. These findings suggest that implementing BMA-PVs can improve generalizability of secondary analyses to



United States Netherlands Panama

Figure 2.7: Secondary analysis LOO-ICs for a Bayesian linear regression model predicting reading achievement with 14 predictors across three countries and two methods.

the target population, boosting analysts' confidence that their inferences may more accurately reflect the underlying population structure.

2.5 Discussion

This paper presents a workflow and justification for a novel method of PV estimation using Bayesian Model Averaging (BMA), designed to directly address statistical uncongeniality and incorporate imputation model uncertainty into PV estimation and secondary analyses. Rather than relying on PCA to reduce the dimensionality of background variables, I include them directly in the imputation model space using BMA. This preserves the correlated covariate structure typically observed in secondary analyses, improving alignment between the imputation and analysis models and enhancing congeniality. Additionally, BMA achieves these benefits without the computational burden of full MCMC over the total model space, making it a fast and practical approach for large-scale assessments. In the simulation study, BMA-PVs effectively captured the true population distribution of student ability, albeit slightly underestimating variability. Their primary advantages emerged in secondary analyses, where BMA-PVs as an outcome variable of interest improved out-of-sample predictive performance compared to PISA-replicate PVs. The empirical application using PISA 2022 reading data

from 76 countries and jurisdictions corroborated these findings. Although country rankings changed only modestly, predictive performance improved across multiple countries. Together, these results indicate that BMA-PVs are a theoretically justified and statistically viable alternative to existing plausible value methodologies.

Plausible values serve multiple purposes in LSAs, extending beyond simply measuring population and subgroup proficiency. While this paper focuses on their uses for rankings and applied analyses, PVs could also be re-envisioned to be generated with specific uses in mind, tailored to different analytical needs. For example, one could create distinct sets of PVs optimized for evaluating subgroup differences or for producing more optimally predictive applied secondary analyses. The present work found that BMA-PVs did not substantially alter country rankings, but their value was apparent in improving out-of-sample predictive performance in a secondary analysis example. Developing detailed guidelines or recommendations for generating PVs for these alternative purposes is beyond the scope of this paper; nevertheless, recognizing these possibilities highlights the flexibility of plausible values and their potential role in improving the reliability and interpretability of LSAs.

These findings expand upon the extensive previous literature surrounding plausible values and their uses in LSAs. While previous work has been clear in the benefits of PVs in the context of LSAs (Mislevy, 1991; Mislevy et al., 1992; Wu, 2005; von Davier et al., 2009), given the relative ease in computational implementation relying on PCA and plug-in fixed estimates. However, as detailed in this work, several pressing issues have previously been underexplored. Specifically, properly addressing model uncertainty and statistical agreement or congeniality between imputation and secondary analyst models. The addition of using BMA-generated PVs helps address these gaps in the literature, serving as a reasonable addition to the plausible value and large-scale assessment world. This approach provides a justified, practical, and computationally accessible alternative to PV methodology in LSAs.

Limitations and Future Directions

The present study has several limitations. First, despite a very strong rank-order correlation and minimal bias, the inability to exactly replicate PISA's reported reading scores may limit the generalizability of these findings. Additionally, the use of only Bayesian linear regression models to compare predictive performance across methods and countries provides a limited demonstration of the benefits of BMA-PVs. It's possible that these findings may not generalize to all modeling contexts, specifically more complex approaches or latent variable modeling. More work in this area would be needed to ensure the robustness of these methods. Another promising direction of research could be incoporating Kaplan and Yavuz's (2020) approach for context questionnaire item imputation with the BMA methods for PV estimation put forth in this study. Despite these limitations, the methods presented here offer identification of present issues and clear improvements over the traditional approaches that rely on orthogonalizing predictors in the conditioning model used to generate PVs.

Another option instead of the `BMS` package used in this work is the `BAS` package (Clyde, 2025), which offers additional flexibility, including support for interaction terms and explicit variable inclusion. This allows, for example, key indicators of substantive importance to be forced into the imputation model to ensure their selection and mitigate congeniality. However, implementing these alternative approaches may require additional fine-tuning to achieve optimal efficiency and predictive accuracy.

Alternatives to BMA-based approaches for PV estimation, such as Bayesian stacking or Bayesian additive regression trees, may further improve predictive performance by relaxing assumptions surrounding the existence and functional form of a single data-generating model (Yao et al., 2018; Chipman et al., 2010; Harra and Kaplan, 2025; Kaplan et al., 2025). Another avenue of PV estimation may be incorporating sparsity into a single imputation model as a single-model alternative that automatically performs variable selection via regularization priors such as the regularized horseshoe (Piironen and Vehtari, 2017). Building on prior work demonstrating the value of plausible values in large-scale assessments, this study introduces

an alternative estimation method via BMA designed to overcome key limitations and better harness the potential of these assessments. This approach aims to provide more accurate and informative results in educational measurement.

3 STUDY 2: INDUCING SPARSITY

Large-scale assessments such as PISA rely on principal components analysis (PCA) to reduce the dimensionality of the conditioning variables used in plausible value (PV) estimation (OECD, 2024). While this approach is motivated by computational convenience (Mislevy et al., 1992), it remains unclear whether PCA preserves the most relevant information for accurately specifying the population model. In particular, orthogonalizing conditioning variables through PCA may introduce uncongeniality and bias into PV estimation and secondary analyses.

An alternative strategy I propose is to incorporate Bayesian regularization through sparsity-inducing priors, such as the ridge, regularized horseshoe, and R2D2 priors (Hsiang, 1975; Piironen and Vehtari, 2017; Zhang et al., 2022). These methods are well suited for high-dimensional data settings with many small effects, providing a method for simultaneous shrinkage and variable selection. More flexible priors, such as the regularized horseshoe (Piironen and Vehtari, 2017) and R2D2 (Zhang et al., 2022), improve upon earlier approaches by more aggressively shrinking unimportant noise while preserving meaningful effects, leading to improvements in predictive performance (Harra and Kaplan, 2024; van Erp et al., 2019).

Within the context of PV estimation, Bayesian regularization offers a straightforward procedure for handling large sets of conditioning variables. By retaining the original scale and structure of the predictors via a linear regression framework, this approaches diminishes the uncongeniality introduced by PCA and reduces the risk of bias resulting from the imputation model. Consequently, this approach has the potential to improve both recovery of the population distribution and the predictive utility of PVs in secondary analyses.

Motivated by these potential advantages, this study investigates whether PV estimation based on sparsity-inducing priors provides a viable alternative to current PCA-based approaches. Specifically, it evaluates whether this procedure improves population recovery, reduces bias, and enhances predictive performance in secondary analyses.

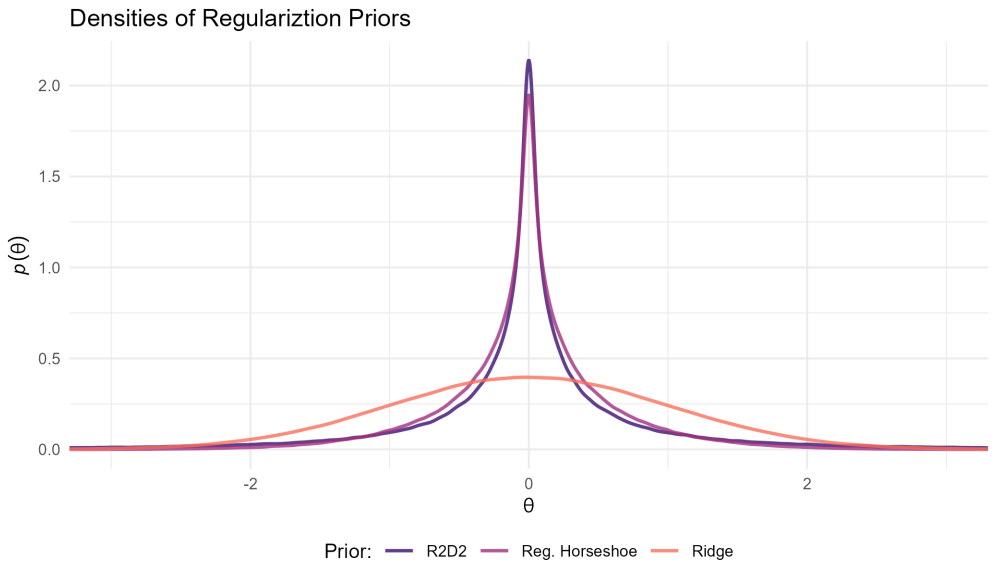


Figure 3.1: Regularization priors discussed in this work: The ridge normal prior, the R2D2 prior, and the regularized horseshoe prior.

The remainder of this chapter is organized as follows. First, a technical overview of Bayesian regularization and sparsity-inducing priors is presented, along with their implementation in PV estimation. Next, a simulation study assesses the proposed methods with respect to latent ability estimation accuracy and predictive performance. This is followed by an empirical application using PISA data to assess its practical implications. The chapter concludes with a discussion of findings, limitations, and directions for future research.

3.1 Regularization Priors to be Investigated

Figure 3.1 displays the density plots for the three regularization priors examined in this chapter: the Bayesian ridge prior, the regularized horseshoe, and the R2D2 prior. Each is discussed below.

The Ridge Prior

Frequentist ridge regression (Hoerl and Kennard, 1970; Hoerl, 1985) aims to yield a parsimonious regularized regression model in the presence of highly correlated variables. The Bayesian specification of ridge regression was suggested by Hsiang (1975), who showed that if the ridge estimator, $\boldsymbol{\beta}$, has a mean of zero and covariance matrix $\boldsymbol{\Sigma} = (\sigma^2/\lambda)\mathbf{I}$, and if $\epsilon \sim N(0, \sigma_\epsilon^2\mathbf{I})$, then the posterior mean of $\boldsymbol{\beta}$ is $(\boldsymbol{x'x} + \lambda\mathbf{I})^{-1}\boldsymbol{x'y}$, which is an alternative specification of the ridge estimator. The penalty term (λ) is captured through normally distributed independent priors placed on the regression slope parameters. These normal priors have mean hyperparameter values fixed at zero in order to control shrinkage toward zero. The variance hyperparameter is typically rescaled to be in standard deviation form and is set to define the degree of spread that the distribution exhibits. A representation of the ridge prior, specified as a $\mathcal{N}(0, 1)$ prior, is given in Figure 3.1.

The Regularized Horseshoe Prior

A major drawback of ridge regression is that it does not improve parsimony in that all of the variables still remain in the model after penalization (Zou and Hastie, 2005). To perform variable selection, other sparsity priors need to be used.

One such prior is the regularized horseshoe, a variant of the original horseshoe prior (Carvalho et al., 2009, 2010). It can be expressed as a scale mixture of normals with half-Cauchy tails, enabling a form of shrinkage that differs from that of traditional regularization priors. More specifically, the tails of its $\mathcal{C}^+$ distribution permit large parameters to remain unregularized, while the global shrinkage parameter τ severely shrinks parameters that are small. The regularized horseshoe corrects issues seen with other regularization priors such as the ridge, lasso, and original horseshoe.

A limitation of the original horseshoe prior relates to cases where large coefficients can transcend the global scale set by τ_0 with the impact being that the posteriors of these large coefficients can become quite diffused, particularly in the case of weakly-identified coefficients

(Betancourt, 2018; Piironen and Vehtari, 2017; Kaplan, 2023). To remedy this issue, Piironen and Vehtari (2017) proposed a *regularized* version of the horseshoe prior. Following the notation used in Betancourt (2018) the regularized horseshoe prior takes the form of the following, for $j = 1, ..., p$, where p are the number of predictors,

$$\beta_j \sim \mathcal{N}(0, \tau^2 \tilde{\lambda}_j^2), \tag{3.1a}$$

$$\tilde{\lambda}_j = \frac{c\lambda_j}{\sqrt{c^2 + \tau^2 \lambda_j^2}}, \tag{3.1b}$$

$$\lambda_j \sim \mathcal{C}^+(0, 1), \tag{3.1c}$$

$$c^2 \sim \mathcal{IG}\left(\frac{\nu}{2}, \frac{\nu}{2} s^2\right), \tag{3.1d}$$

$$\tau \sim \mathcal{C}^+(0, \tau_0), \tag{3.1e}$$

where $c > 0$ and s^2 is the variance for each of the p predictor variables. Those variables that have large variances would be considered more relevant a priori, and while it is possible to provide predictor-specific values for s^2, generally we scale the variables ahead of time so that $s^2 = 1$. Finally, c^2 is the slab width, which controls the size of the large regression coefficients (Piironen and Vehtari, 2017). The density plot for the regularized horseshoe is given in Figure 3.1. The regularized horseshoe prior can be viewed as a continuous equivalent to the "gold standard" spike-and-slab prior, which is often more difficult to implement in many `R` programs (Mitchell and Beauchamp, 1988; George and McCulloch, 1993). As a result, the regularized horseshoe provides an appealing alternative for inducing sparsity within a Bayesian framework.

Another way to understand the regularized horseshoe prior is through its shrinkage profile following Harra and Kaplan (2024), shown below. The distribution is asymmetric, with more mass at the left mode, because the regularized horseshoe is shifted 0.05 units to the right of 0.0 (see also Piironen and Vehtari, 2017). As a result, even large parameters experience some shrinkage ($\kappa_j = 0.05$), while small parameters are subject to total shrinkage ($\kappa_j =$

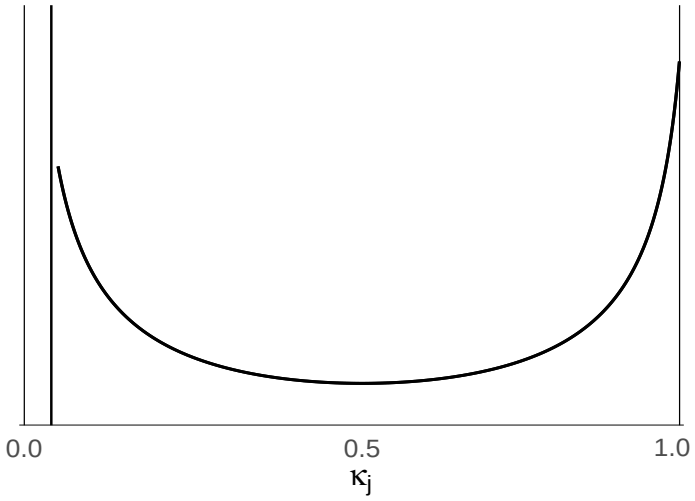


Figure 3.2: Density of shrinkage weight for the regularized horseshoe prior (Harra and Kaplan, 2024).

1.0). Essentially, the key difference between the original horseshoe prior and the regularized horseshoe is this feature of preventing large coefficients from entirely escaping shrinkage and remaining too large.

Eliciting Hyperparameters for τ

Inspection of Equation (3.1e) shows that a value of τ_0 must be specified, as it determines the amount of global shrinkage induced by τ in Equation (3.1a).Piironen and Vehtari (2017) relate the choice of τ_0 to the effective number of non-zero coefficients in the model. Specifically, they define this effective number of non-zero coefficients in terms of the shrinkage parameter, κ_j, as:

$$m_{eff} = \sum_{j=1}^{p} (1 - \kappa_j), \tag{3.2}$$

where we see that if κ_j is close to zero or one, the result is effectively the number of non-zero coefficients in the model. Next, Piironen and Vehtari (2017) show that, under the assumption that $\mathbb{V}(x_j) = 1$,

$$\mathbb{E}(m_{eff}|\tau,\sigma) = \frac{\tau\sigma^{-1}\sqrt{n}}{1+\tau\sigma^{-1}\sqrt{n}} \times p. \tag{3.3}$$

Let $p_0 \equiv \mathbb{E}(m_{\text{eff}} \mid \tau, \sigma)$. Then, τ_0 can be expressed as

$$\tau_0 = \frac{p_0}{p - p_0} * \frac{\sigma}{\sqrt{n}}, \tag{3.4}$$

where again, p is the total number of coefficients in the model, and p_0 ($\neq p$) is the number of coefficients assumed to be effectively large. Piironen and Vehtari (2017) note that using a standard $\mathcal{C}^+(0,1)$ prior for τ tends to place substantial mass on large coefficients, particularly when p is large, resulting in many coefficients escaping shrinkage. A more effective strategy is to incorporate prior knowledge to select a reasonable value for p_0. This consideration is particularly relevant for the present study, where the total number of model parameters is substantial ($p \geq 500$), motivating the use of the regularized horseshoe. However, Harra and Kaplan (2024) suggest that the impact of this choice is less consequential in large sample settings, such as those considered here, where posterior inference is less sensitive to the prior.

The R2D2 Prior

Another global-local shrinkage prior is the R2D2 prior, or the R^2-induced Dirichlet decomposition (R2D2) prior (Zhang et al., 2022). This prior is designed to improve shrinkage by inducing heavier tails and stronger concentration near zero compared to earlier priors such as the Bayesian Lasso (Park and Casella, 2008) or the horseshoe (Carvalho et al., 2009, 2010). A key feature of the R2D2 prior is that, rather than directly specifying a joint prior on $\boldsymbol{\beta}$, it first places a Beta prior on the model R^2 and then derives the implied prior on the regression coefficients (Zhang et al., 2022).

Zhang et al. (2022) show that this prior can be expressed as a hierarchical scale mixture of normals:

$$\beta_j \mid \sigma^2, \lambda_j \sim \mathrm{DE}\left(\sigma\sqrt{\frac{\lambda_j}{2}}\right), \tag{3.5a}$$

$$\lambda_j \sim \mathrm{BP}(a_\pi, b), \tag{3.5b}$$

where λ_j is a local shrinkage parameter and $\mathrm{BP}(a_\pi, b)$ denotes a Beta Prime distribution (Johnson et al., 1994). The Beta Prime prior, which differs from a Beta prior in that it is defined on $(0, \infty)$ rather than $(0, 1)$, arises from the variance decomposition implied by the R^2 prior and governs the induced distribution over the local variance components. Each β_j follows a Laplace (double-exponential) distribution with scale parameter proportional to $\sigma\sqrt{\lambda_j/2}$.

In this framework, the R2D2 prior first places a Beta prior on the model R^2, which determines the total proportion of variance explained by the model. This total variance is then allocated across predictors via a Dirichlet-based decomposition. When all concentration parameters are set equally (e.g., equal to 1), this corresponds to a uniform prior over variance allocation, implying equal a priori importance of predictors, as in the present work. As the model is updated with the data, the posterior reallocates variance, allowing relevant predictors to receive a larger share while less relevant ones are increasingly shrunk toward zero.

This variance decomposition induces a Beta Prime prior on the local shrinkage parameter λ_j, yielding a Laplace prior on the regression coefficients. The Laplace component concentrates mass near zero, encouraging sparsity, while the Beta Prime component induces heavy tails, mitigating over-shrinkage. Although the R2D2 prior incorporates elements of the Bayesian Lasso, as both share a double-exponential distribution, its heavier tails and explicit control over total explained variance make it particularly well suited to high-dimensional settings with many predictors.

Thus, the R2D2 prior can be viewed as a global-local shrinkage prior that is more

directly tied to overall model fit rather than solely coefficient magnitude. It also performs particularly well in high-dimensional data spaces (Zhang et al., 2022), which is relevant for the present study in PV estimation. Figure 3.1 illustrates its combination of strong shrinkage near zero and heavy tails relative to the regularized horseshoe prior.

3.2 Present Study

Building on the discussion of global-local shrinkage priors, it becomes clear that priors capable of both strong shrinkage of small effects with heavy tails to regularize large effects are particularly useful in high-dimensional settings with many predictors in the model. These priors allow the model to concentrate on the most relevant predictors while still accounting for uncertainty across the full set.

Given these properties, sparsity-inducing priors provide a framework for PV estimation, especially in contexts where traditional dimension reduction techniques like PCA may introduce bias. As uncongeniality and bias may be introduced when performing PCA on the large set of conditioning variables, I propose that PV estimation via a sparsity-inducing prior may be better suited. This proposed alternative is more straightforward, allows for the data themselves to reduce dimensionality without compromising congeniality, generates PVs with a proper posterior distribution, and could improve out-of-sample predictive performance for increased generalizability with secondary analyses.

3.3 Methods

To ensure the results of this work are broadly applicable and easily reproducible across countries and modeling contexts, I adopt the default prior specifications from `brms` without further modification (Bürkner, 2017) in both the present simulation and empirical studies. Models using a ridge prior were fit with the `cmdstanr` (Gabry et al., 2025) backend to improve computational efficiency in Stan, using two chains with 1,000 warmup iterations

and 2,000 post-warmup iterations per chain. For the regularized horseshoe and R2D2 prior specifications, more iterations were required due to increased posterior complexity. These models were also fit using `cmdstanr`, but with four chains and 2,000 warmup iterations followed by 4,000 iterations per chain. To improve sampling stability and reduce divergent transitions, the target acceptance rate was increased to 0.99. Table 3.1 presents the full specifications of the priors examined.

The following sections describe the simulation study and empirical analyses used to assess prior performance in the context of PV accuracy and predictive performance for secondary analyses.

Table 3.1: Default shrinkage prior specifications in `brms` used in both the simulation and empirical studies.

Ridge	**Regularized Horseshoe**	**R2D2**
$\beta_j \sim \mathcal{N}(0, 1)$	$\beta_j \sim \mathcal{N}(0, \tau^2 \tilde{\lambda}_j^2)$ $\tilde{\lambda}_j = \frac{c\lambda_j}{\sqrt{c^2 + \tau^2 \lambda_j^2}}$ $\lambda_j \sim \mathcal{C}^+(0, 1)$ $c^2 \sim \text{Inv-Gamma}(2, 8)$ $\tau \sim \mathcal{C}^+(0, 1)$	$\beta_j \sim \mathcal{N}(0, \sigma^2 \phi_j \tau^2)$ $R^2 \sim \text{Beta}(1, 1)$ $\phi \sim \text{Dirichlet}(0.5, \ldots, 0.5)$ $\tau^2 = \frac{R^2}{1 - R^2}$

Simulation Study

This study investigates whether sparsity-inducing priors used in the estimation of plausible values improve the recovery of the population distribution of latent ability, as well as predictive performance in secondary analyses. To this end, a comprehensive Monte Carlo simulation study is conducted, with data generated following the design of Study 1.

PV accuracy and performance are evaluated using Bayesian linear regression models estimated via `brms` (Bürkner, 2017). Each model predicts student ability using an IRT-based score as the outcome, with all background variables ($k = 500$) included as predictors. Within each prior condition, a common regularization prior is applied to all regression

coefficients. Posterior draws from the fitted models are then used to construct 10 plausible values per simulated student. To compare the sparsity-based PVs against current practices, PISA-replicate PVs are generated using the `TAM` package (Robitzsch et al., 2025), following procedures described in Study 1.

Three prior specifications are considered: the ridge normal prior, the regularized horseshoe prior, and the R2D2 prior. Prior settings using `brms` defaults are summarized in Table 3.1. The Bayesian Lasso (Park and Casella, 2008) is not considered, despite its wide use, as it has been deprecated in `brms` in favor of the regularized horseshoe and R2D2 priors. Each condition is replicated 1000 times, with 10 PVs generated per student in each replication and aggregated for analysis.

Secondary analyses are conducted using a Bayesian linear regression model, where the aggregated PVs serve as the outcome and are regressed on a randomly selected subset of predictors ($p = 9$). PV performance is evaluated using the leave-one-out cross-validation information criterion (LOO-IC) (Vehtari et al., 2017), as discussed previously.

Empirical Study

For the empirical study, I use PISA 2022 reading data from the 76 countries that have publicly available item response and background items collected. The full list of countries in this analytic sample are contained in Table 2.1. For each PISA country included, 10 sparsity-estimated PVs per student were generated under each of three priors: the ridge, regularized horseshoe, and R2D2 (see Table 3.1).

Data preparation for analysis followed the same procedure as in Study 1, including imputation of missing background items and estimation of item parameters.

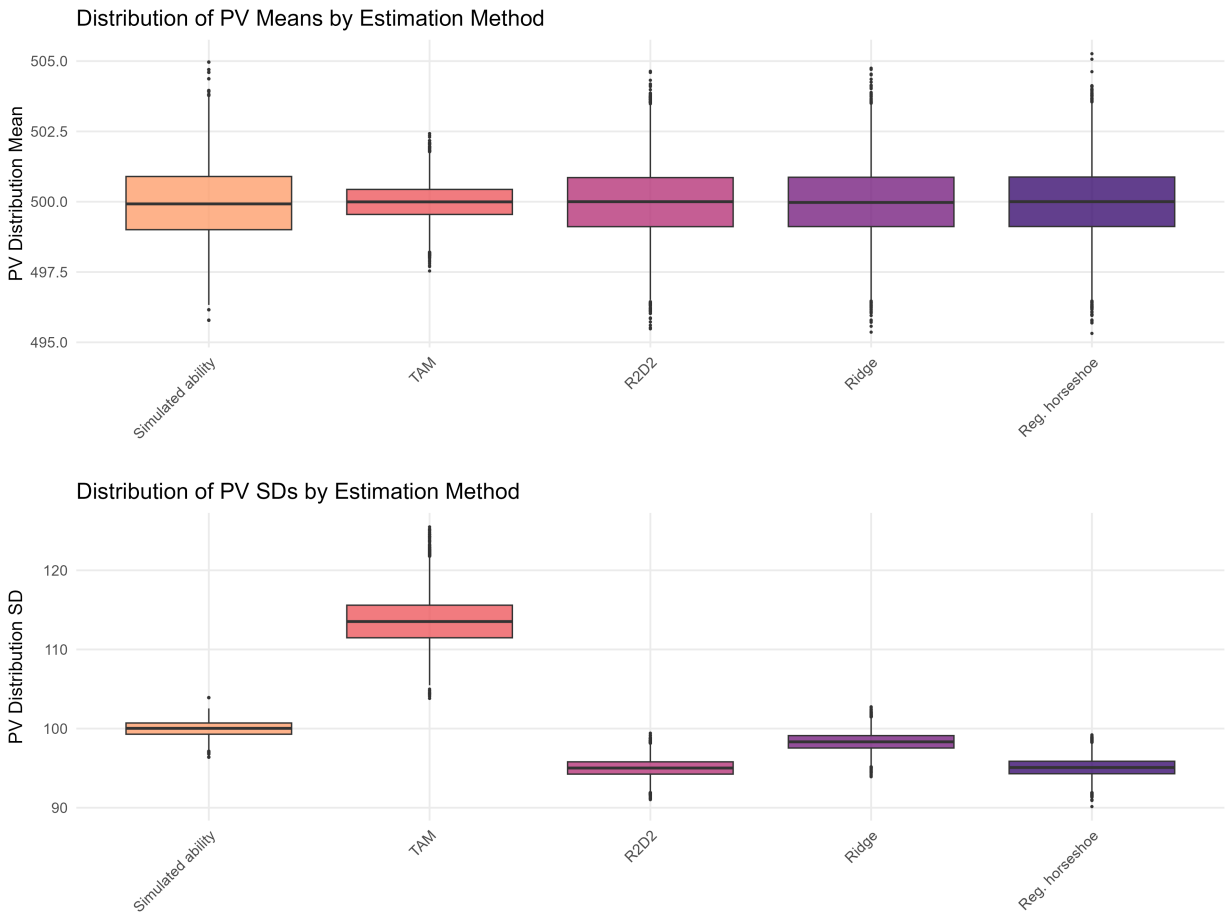


Figure 3.3: Distributions of PV means and standard deviations across methods in the simulation study.

3.4 Results

Simulation Study

First, I evaluate the ability of sparsity-inducing priors to recover a simulated population distribution of latent ability. I focus on the mean and standard deviation of each resulting PV distribution, as these are commonly reported in large-scale assessments.

Across methods, there are minimal differences in capturing the mean of the estimated latent ability distribution, both across the sparsity-inducing priors and the current PV methods as replicated in `TAM`. However, the R2D2 and regularized horseshoe priors slightly

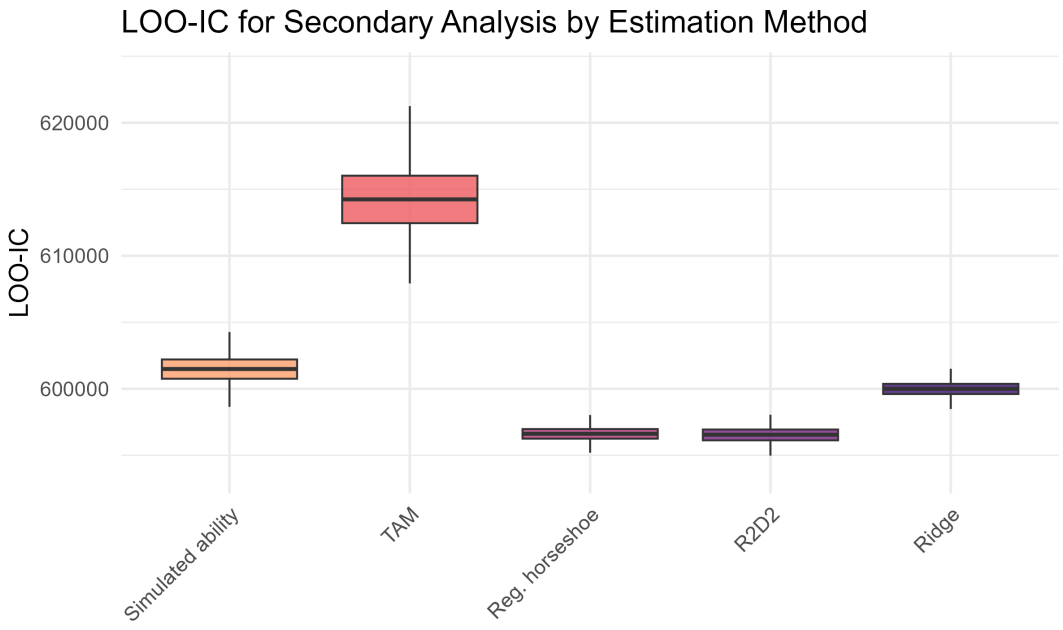


Figure 3.4: Distributions of LOO-IC's for a secondary analysis model by method in the simulation study.

underestimate the true variability, while the ridge prior aligns more closely with the simulated population distribution. Additionally, current PV methods tend to slightly overestimate population variability. These results suggest that sparsity-inducing priors offer a viable alternative for generating PVs and may help address limitations in existing PV methodologies.

However, the benefits of these alternatives becomes more apparent when looking at a secondary analysis example. Figure 3.4 shows the out-of-sample-predictive performance in the simulation study across replications and methods. Here, we see that the regularized horseshoe and the R2D2 prior show a strong improvement in out-of-sample predictive performance in the form of a LOO-IC compared to the ridge prior, and especially compared to existing PV methods via `TAM`. Interestingly, all three sparsity conditions outperformed using the simulated ability estimates themselves, especially the regularized horseshoe and R2D2. These findings demonstrate that the sparsity priors are enacting shrinkage and enhancing predictive performance as expected. To examine the generalizability of these findings to real-world data, the study proceeds with an empirical application.

Empirical Study

Prior to interpreting the results of the empirical study, I conducted diagnostic checks to assess model convergence. All models, estimated using `bmrs` (Bürkner, 2017) with sparsity-inducing priors in a linear regression framework, demonstrated satisfactory convergence. In particular, all $\hat{R}$ values were approximately 1.00, indicating no evidence of non-convergence.

Computation time for the regularized PV generation models varied substantially across countries, primarily reflecting differences in sample size, which ranged from roughly 5,000 to 30,000 students. In this study, runtime per model ranged from about one hour for smaller samples to over ten hours for the largest. Spain, with more than 30,000 participants, required the longest processing time, particularly for the more complex regularized horseshoe and R2D2 priors. Such variation emphasizes the importance of accounting for model runtime when considering applying these methods to LSAs.

Figure 3.5 shows the resulting distribution of estimated latent ability for the U.S. sample of PISA 2022 reading across methods. We see that all three sparsity conditions closely mimicked the `TAM`-replicated PISA estimates. As noted previously, there is a small amount of bias between the reported PISA PVs and the present study's replication. The Bayesian ridge prior-estimated PVs appear closest to the PISA-replicates, whereas the R2D2 and regularized horseshoe-estimated PVs slightly underestimate the variability in the sample, however the bias is minimal.

Country Rankings

As stated earlier, a key aspect of PISA reporting surrounds country rankings of reading achievement (Özer, 2020; Araujo et al., 2017; Martens and Niemann, 2013). These rankings rely on PVs to compare country and jurisdiction performance. Figure 3.6 depicts changes in country rankings for PISA 2022 reading mean scores. To understand what changes we see, note that filled-in circles represent the PISA-replicate rank, and the empty circles denote sparsity-estimated PV ranks. Line segments between circles depict what difference, if any,

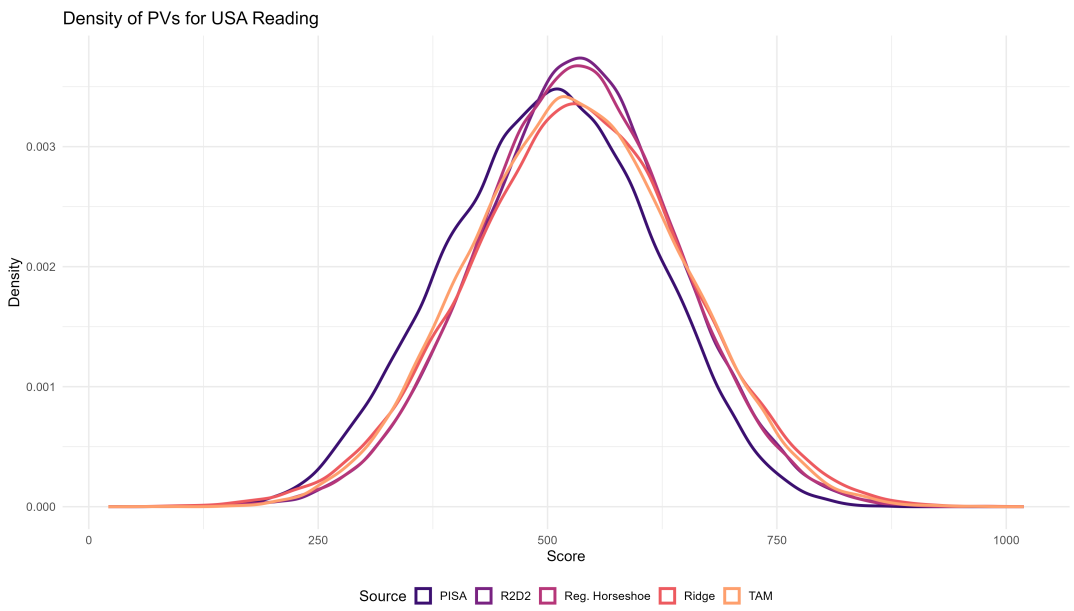


Figure 3.5: PV densities by estimation method using PISA 2022 United States (USA) reading data.

arises between the two methods of PV generation.

Across all three sparsity conditions, country rankings show a high degree of agreement between conventional PV estimation via `TAM` and the proposed sparsity-inducing approaches. In most cases, rank differences are negligible and typically no more than three positions. One notable exception is Denmark, which shifts from 26th up to 19th across all three prior specification conditions. Other changes are more minor, such as South Korea (KOR) moving from 5th to 4th, and positional swaps between countries like Palestine (PSE) and Morocco (MAR) between 70th and 71st ranks. I hypothesize that these minor shifts in country rankings are simply the result of random noise.

Differences in country rankings across the sparsity-inducing priors are minimal. This is unsurprising, as all PVs were generated using the same Bayesian linear regression framework on the same data. Because each prior is applied uniformly across countries within a given condition, shrinkage and regularization effects operate globally rather than differentially by country. With large samples and many parameters, the data themselves dominate the posterior distribution. As a result, while prior choice may influence the overall scale of effect

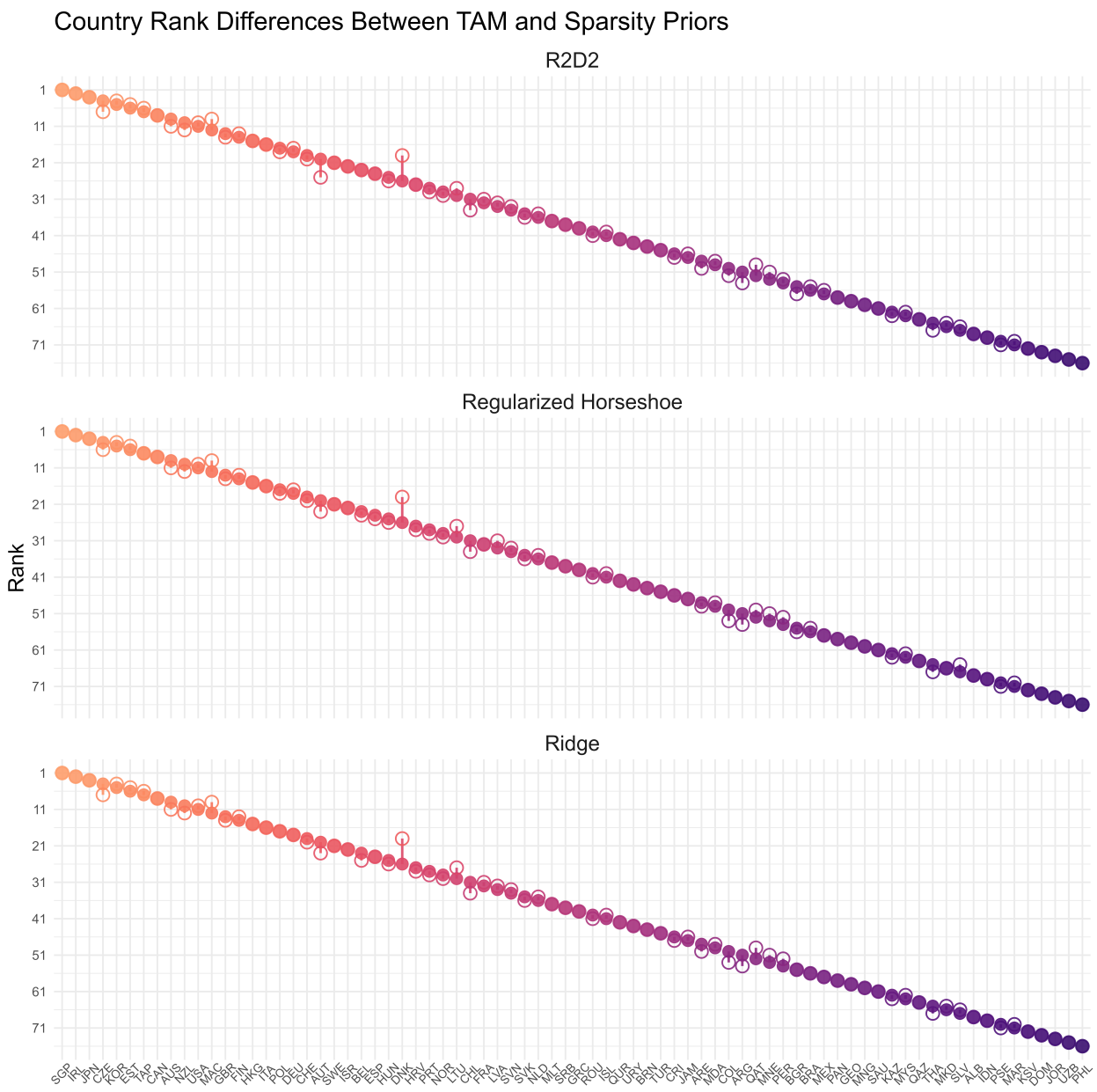


Figure 3.6: Country rankings of PISA-replicates using `TAM` and sparsity-estimated PVs for reading in 2022. Note the PISA-replicate ranks in parentheses on the bottom labels. Filled-in circles denote the PISA-replicate rank, and the empty circles denote sparsity-estimated PV ranks. Line segments between circles depict the magnitude of disagreement.

sizes, it is unlikely to substantially alter relative country rankings.

The strong agreement with conventional PV estimates is unsurprising. Traditional PV methods rely on a single linear imputation model that is largely data-driven, with latent ability estimates conditioned on a normal distribution of principal components of the back-

ground variables. Although introducing sparsity-inducing priors can improve regularization, prediction, and parsimony, these priors do not substantially alter country rankings because the underlying effect sizes in the data remain consistent. Overall, these findings suggest that the proposed methods preserve the relative rankings of countries while providing a more theoretically-sound framework for accounting for uncertainty in the PV imputation model. Across the different sparsity priors, there appear to be no meaningful differences in their effects on country rankings; however, differences in prior selection are more apparent in the secondary analysis example.

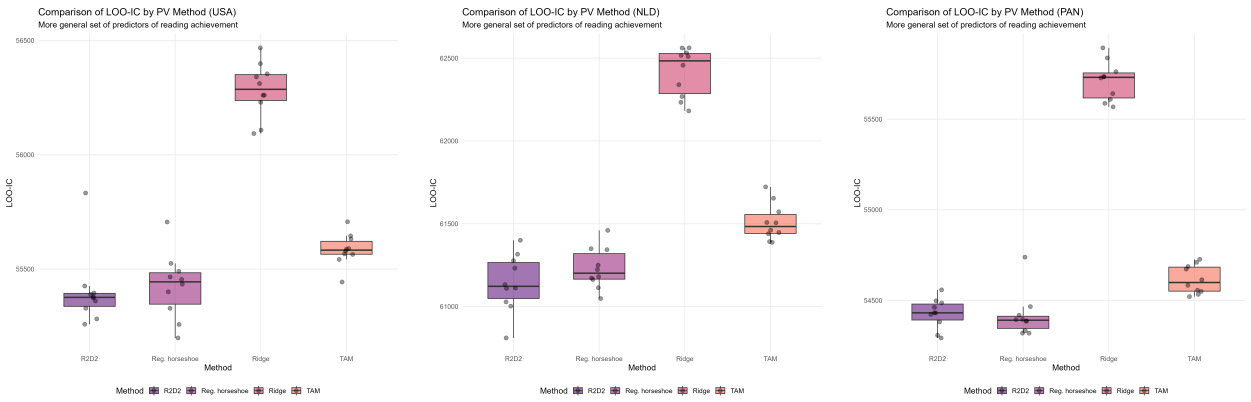
Secondary Analysis

As discussed previously, an important use of PVs is in applied secondary analyses in a variety of fields. Most commonly is using reading achievement plausible values as an outcome of interest, as predicted by background variables. To examine how well sparsity-inducing prior-estimated PVs can be used in these types of analyses, I use an example Bayesian linear regression model to study predictive performance. To illustrate performance across varying positions in the achievement distribution, I examine three countries: the United States, the Netherlands, and Panama, representing different points in overall country rankings. This analysis was conducted 10 times for each country and method, one for each PV as the outcome variable. Results are pooled and described below.

The model of interest to predict reading achievement within a given country for each student i can be specified as:

$$\begin{aligned} \text{READ}_i = \beta_0 &+ \beta_1 \text{ESCS}_i + \beta_2 \text{HISCED}_i + \beta_3 \text{IMMIG}_i + \beta_4 \text{LANGN}_i \\ &+ \beta_5 \text{SEX}_i + \beta_6 \text{AGE}_i + \beta_7 \text{HISEI}_i + \beta_8 \text{HOMEPOS}_i \\ &+ \beta_9 \text{GROSAGR}_i + \beta_{10} \text{PROBSELF}_i + \beta_{11} \text{DISCLIM}_i \\ &+ \beta_{12} \text{RELATST}_i + \beta_{13} \text{BELONG}_i + \beta_{14} \text{EFFORT1}_i + \epsilon_i \end{aligned} \tag{3.6}$$

where reading achievement is predicted by a student's socioeconomic status (ESCS_i), the



United States Netherlands Panama

Figure 3.7: Secondary analysis LOO-ICs for a Bayesian linear regression model predicting reading achievement with 14 predictors across three countries and four methods.

highest education attainment by a parent (HISCED_i), student's immigration status (IMMIG_i), language spoken most at home (LANGN_i), sex, age, a student's parental occupational status index (HISEI_i), a home possessions index reflecting family wealth and resources (HOMEPOS_i), growth mindset index (GROSAGR_i), problems with self-directed learning index (PROBSELF_i), disciplinary climate index (DISCLIM_i), student–teacher relationship quality index (RELATST_i), sense of belonging (BELONG_i), and self-reported effort on the cognitive exam (EFFORT1_i).

Figure 3.7 presents the distribution of LOO-ICs in the secondary analysis example. Emphasis should be placed on the median line in the box plots, as a regression using just a single PV (here, represented as the points), does not constitute a sufficient estimate of model performance. Across all three countries, both the R2D2 and regularized horseshoe priors strongly outperform traditional PV-replicates using `TAM`. In contrast, the ridge prior-estimated PVs performed substantially worse in terms of higher LOO-ICs across all three example countries. Differences between the R2D2 and regularized horseshoe prior are negligible, suggesting both methods of PV estimation are viable from the perspective of a secondary analysis's predictive performance.

These patterns are consistent with the underlying shrinkage properties of the R2D2 and regularized horseshoe priors. The normal ridge prior applies uniform and relatively weak

shrinkage across all imputation model coefficients and does not effectively distinguish between meaningful effects and noise. This is particularly the case with large samples and predictor sets, such as with PV estimation. As a result, it does not substantially improve out-of-sample predictive performance. In contrast, the regularized horseshoe and R2D2 priors induce more aggressive and adaptable shrinkage, where small coefficients are shrunk much more effectively towards zero and preserves true effects. This more targeted regularization leads to more parsimonious and more accurate models, inducing improved predictive accuracy.

3.5 Discussion

This paper presents a workflow and rationale for a novel approach to PV estimation using sparsity-inducing priors in Bayesian linear regression. These methods are designed to directly address statistical uncongeniality and to incorporate population model uncertainty into both PV estimation and secondary analyses. Rather than relying on PCA to reduce the dimensionality of background variables, I include them directly in the imputation model, where regularization priors can autonomously perform variable selection and shrinkage of unimportant predictors of latent ability. This preserves the correlated covariate structure typically observed in secondary analyses, improving alignment between the imputation and analysis models and enhancing congeniality.

In the simulation study, PVs estimated via sparsity-inducing priors effectively captured the true population distribution of student ability, albeit slightly underestimating variability across all three prior conditions that include the R2D2 prior (Zhang et al., 2022), the regularized horseshoe (Piironen and Vehtari, 2017), and the ridge normal prior (Hsiang, 1975). For the stronger regularized priors, the R2D2 and regularized horseshoe, their primary advantages emerged in secondary analyses, where the resulting PVs an outcome variable of interest improved out-of-sample predictive performance compared to PISA-replicate PVs, but the ridge prior performed poorly. The empirical application using PISA 2022 reading data

corroborated these findings. Although country rankings changed only modestly, and were consistent across prior selection, predictive performance improved across multiple countries using the R2D2 and regularized horseshoe priors to estimate PVs. Together, these results indicate that a regularized population model specification with the R2D2 or regularized horseshoe prior is a theoretically justified and statistically viable alternative to existing plausible value methodologies.

Plausible values serve multiple purposes in LSAs, extending beyond simply measuring population and subgroup proficiency. While this chapter focuses on their uses for rankings and applied analyses, PVs could also be re-envisioned to be generated with specific uses in mind, tailored to different analytical needs. For example, one could create distinct sets of PVs optimized for evaluating subgroup differences or for producing more optimally predictive applied secondary analyses across the entire distribution of proficiency. The present work found that sparsity-based PVs did not substantially alter country rankings, but their value was apparent in improving out-of-sample predictive performance in a secondary analysis example when using a strong shrinkage prior like the R2D2 or regularized horseshoe.

These findings expand upon the extensive previous literature surrounding plausible values and their uses in LSAs. While previous work has been clear in the benefits of PVs in the context of LSAs (Mislevy, 1991; Mislevy et al., 1992; Wu, 2005; von Davier et al., 2009), given the relative ease in computational implementation relying on PCA and plug-in fixed estimates. However, as detailed in this work, several pressing issues have previously been underexplored. Specifically addressing statistical agreement or congeniality between imputation and secondary analyst models. The addition of using sparsity-inducing priors in PV estimation helps address these gaps in the literature, serving as a reasonable addition to the plausible value and large-scale assessment world.

Limitations and Future Directions

The present study has several limitations. First, despite a very strong rank-order correlation and minimal bias, the inability to exactly replicate PISA's reported reading scores may limit the generalizability of these findings. Additionally, the use of only Bayesian linear regression models to compare predictive performance across methods and countries provides a limited demonstration of the benefits of sparsity-based PVs. It's possible that these findings may not generalize to all modeling contexts, specifically more complex approaches or latent variable modeling. More work in this area would be needed to ensure the robustness of these methods. Despite these limitations, the methods presented here offer identification of present issues and clear improvements over the traditional approaches that rely on orthogonalizing predictors in the conditioning model used to generate PVs.

Another notable limitation of this work concerns computational feasibility. Runtime for the PV generation models varied across countries, with larger country samples requiring ten hours or more to converge and generate PVs. Scaling this process across dozens of countries, using average computing resources, would impose substantial demands on both time and computing power making PV estimation for LSAs like PISA impractical. While the present study implemented parallel processing on a high-throughput computing system (Center for High Throughput Computing, 2006), organizations such as the OECD may not have access nor wish to devote the necessary time and computing capacity for these proposed methods. Consequently, despite the theoretical advantages of sparsity-induced PVs over existing methods, their practical implementation in LSAs is currently limited without faster estimation techniques, which constrains the feasibility of adopting the proposed workflow in real-world assessment contexts. These methods proposed in this work may be better suited for small sample situations, where regularization has shown to substantially improve accuracy and performance (Harra and Kaplan, 2024; van Erp et al., 2019).

Future work in this topic may also want to explore other approaches for PV estimation, such as Bayesian stacking or Bayesian additive regression trees, that may further improve

predictive performance by relaxing assumptions surrounding the existence and functional form of a single data-generating model (Yao et al., 2018; Chipman et al., 2010; Harra and Kaplan, 2025; Kaplan et al., 2025). Building on prior work demonstrating the value of plausible values in large-scale assessments, this study introduces an alternative estimation method via Bayesian regularization designed to overcome key limitations and better harness the potential of these assessments. This approach aims to provide more accurate and informative results in educational measurement.

4 STUDY 3: BAYESIAN ADDITIVE REGRESSION TREES

Lastly, plausible value estimation via the machine learning method Bayesian additive regression trees (BART) offers a promising alternative to traditional plausible value estimation approaches. BART takes a nonparametric approach, modeling outcomes (i.e., student proficiency) as a sum of many small regression trees. This method is motivated by other ensemble strategies and is implemented within a Bayesian framework (Chipman et al., 2010).

BART's approach provides several advantages worth studying in the context of PV estimation. It imposes fewer assumptions on the data-generating process than traditional regression approaches, addresses model selection uncertainty by ensembling many regression trees, and mitigates uncongeniality by directly incorporating the set of conditioning variables into the ensemble while implicitly performing variable selection. Its Bayesian regularization techniques, described below, also further improves predictive stability and accuracy (Chipman et al., 2010; Brent et al., 2017; Tan and Roy, 2019). Previous research has shown that BART outperforms other machine learning methods, including random forests (Chipman et al., 2010; Hill et al., 2020), motivating its use in the present study.

Although widely applied in machine learning, BART remains underutilized in educational statistics and measurement, specifically within the context of large-scale assessments. This study seeks to bridge this gap by exploring BART as a method for PV generation (BART-PVs),combining the flexibility and predictive power of machine learning with Bayesian uncertainty quantification.

The remainder of this chapter is organized as follows. First, a technical overview of BART is presented. Next, simulation and empirical studies using PISA 2022 data are conducted to evaluate the performance of BART-based plausible values (BART-PVs) in terms of population distribution recovery and predictive accuracy for secondary analyses. The chapter concludes with a discussion of findings, limitations, and directions for future research.

4.1 Overview of Bayesian additive regression trees

BART is a nonparametric regression method using a sum-of-trees model, motivated by other model ensembling and machine learning methods in a fully-Bayesian framework (Chipman et al., 2010). To understand how BART works, first we must define a sum-of-trees model as:

$$Y = f(x) + \epsilon, \quad \epsilon \sim N(0, \sigma^2), \tag{4.1}$$

given an outcome Y predicted by some unknown function $f(x)$. BART attempts to approximate the true data-generating model, where $f(x) = E(Y|x)$, via a fitted sum-of-trees $h(x)$:

$$f(x) \approx h(x) \equiv \sum_{j=1}^{m} g_j(x), \tag{4.2}$$

where each $g_j(x)$ is a regression tree out of m total trees. By summing many smaller and "weaker" trees, BART avoids overfitting and can conduct variable selection with strong predictive performance (Chipman et al., 2010; Ročková and Saha, 2019). Rather than relying on a single complex model that may be prone to overfitting and misspecification, BART models the data-generating process as a sum of many simple components, improving predictive performance without requiring strong assumptions about functional form. By ensembling many smaller trees, BART implicitly can perform variable selection based on which variables the algorithm deems important to include as meaningful node splits.

With BART, regularization priors are placed on all model parameters to minimize disproportionately impacting individual regression trees in the ensemble (Chipman et al., 2010). This prevents from any one tree from dominating the ensemble. Model fitting is done with iterative back-fitting via a Gibbs sampler across a user-specified m number of trees, updating individual trees conditioned on the residuals of the others. This process yields posterior draws of $h(x)$ and forms a full posterior distribution of predictions for Y.

There are several statistical and practical benefits to using BART for PV estimation

and optimized prediction. First, Chipman et al. (2010) find that BART outperforms other machine learning methods such as random forests and boosting in terms of out-of-sample predictive performance, while its fully-Bayesian methods allow for straightforward interpretations and uncertainty quantification surrounding parameter estimates. BART's nonparametric framework imposes fewer assumptions on the data-generating process and functional forms (e.g., linearity) than traditional modeling approaches. By combining this flexibility with its ensembling methods, BART reduces the risk of model misspecification while simultaneously addressing model uncertainty and potential uncongeniality between the imputation model and secondary analyses. These properties make BART an appealing potential alternative for PV generation.

BART also allows for flexible settings to optimize performance. We must specify a prior on each tree structure T and terminal node parameters μ (Brent et al., 2017). For the tree structure prior, we denote the probability that a current node will further split as:

$$p_{\text{split}}(d) = \frac{\alpha}{(1+d)^{\beta}}, \tag{4.3}$$

where for a given depth d, α is the base parameter, controlling the overall likelihood of a node splitting, and β is the power, which controls how quickly the splitting probability decreases as d increases. Large β values diminish the number of terminal nodes. This means that trees are more shallow with fewer terminal nodes. These parameters control the complexity of trees and mitigate the likelihood of overfitting, effectively inducing regularization (Brent et al., 2017; Tan and Roy, 2019).

For placing priors on terminal node predictions, μ_i, where $\mu_i \sim N(0, \tau^2)$, we specify the parameter k to obtain the variance τ, which controls the degree of prediction shrinkage towards the global mean:

$$\tau = \frac{\max(y) - \min(y)}{2k\sqrt{m}}, \tag{4.4}$$

with m denoting the total number of trees in the ensemble, and k controls shrinkage of

predictions towards the mean. Smaller k values induce greater amounts of shrinkage.

Previous work on BART highlights that it's adaptable to many modeling approaches (Chipman et al., 2010; Hill et al., 2020), however the present work will focus on nonparametric regression for generating plausible values. The flexibility and generalizability of BART, combined with the advantages of Bayesian regularization and uncertainty quantification, make it a promising approach for PV estimation. Its streamlined fully-Bayesian approach has shown more accuracy than other ensembling methods in the machine learning world.

While BART has been well studied and used in the context of machine learning (Chipman et al., 2010; Hill et al., 2020), it appears to be underutilized in the educational statistics and large-scale assessment space. The present work seeks to explore how we can bridge between these two fields to combine the best practices of both to yield optimally predictive and unbiased models in more efficient ways.

4.2 Present Study

Unlike conventional PV estimation approaches that rely on PCA for dimensionality reduction, BART's nonparametric approach makes minimal assumptions about the functional form of the relationships in the data, helping to preserve congeniality between the population model and PV estimation. Additionally, BART inherently accounts for model selection uncertainty through its ensemble of regression trees, reducing the risk of model misspecification and providing more robust estimates of student abilities. Motivated by these advantages, this study examines the potential improvements in accuracy and predictive performance offered by estimating PVs via BART (BART-PVs) compared to conventional methods.

4.3 Methods

Simulation Study

In this study, I examine how using BART to estimate plausible values better captures the population distribution of latent ability and affects the predictive performance of secondary analyses. To do so, I propose a comprehensive Monte Carlo simulation study. The data generation process follows that of Study 1.

I investigate PV accuracy and performance using Bayesian additive regression trees estimated via `BART` (Sparapani et al., 2021). Each ensemble estimates student ability, given an IRT-based estimate as the outcome of interest. The entire set of background variables ($k = 500$) are used as the predictors, with 10 folds for cross-validation. For each posterior draw, PVs were generated by augmenting draw-specific predictions of latent ability with the estimated posterior residual standard deviation, to better reflect population variability.

To compare BART-PVs with those generated using current methods, I use the `TAM` package (Robitzsch et al., 2025) to generate 10 PVs for each simulated student based on their item responses, conditioned on the principal components that explain 90% of the variance of the background variables, following LSA procedures (Braun and von Davier, 2017). In total, 1000 replications were done for each method.

For the simulation study, I investigate what impacts parameter or setting choices make on the resulting PVs. I identify four settings available in the `BART` package to examine. These settings include:

- $ntree$: or m, the number of regression trees to be included in the ensemble.
- α: or the base, is the probability of a node splitting when growing a tree.
- β: the power parameter, governing how quickly the splitting probability decreases as tree depth increases; larger values produce shallower trees with fewer terminal nodes.

- k: the shrinkage parameter controlling the degree to which terminal node predictions are pulled toward the overall mean.

Sparapani et al. (2021) recommend the defaults of setting the parameter $\alpha = 0.95$ and $\beta = 2$, making non-null but small trees likely to reduce overfitting. (Sparapani et al., 2021; Brent et al., 2017; Tan and Roy, 2019). While many other options exist within the `BART` package, the present study identified these four parameters of interest due to their importance in controlling regularization and optimizing predictions.

While the present simulation study does investigate BART hyperparameter selection, the final comparisons between current PV methods and BART-PVs uses `BART` package defaults (Sparapani et al., 2021). Specifically, the model was fit with 200 trees ($ntree = 200$) and a global shrinkage parameter of $k = 2$, inducing regularization. Tree depth was specified by a prior with $\alpha = 0.95$ and $\beta = 2$, which favors shallow trees by assigning a high probability of splitting nodes that decays rapidly with depth. Together, these hyperparameter choices produce an ensemble of weak individual trees, with the aim of supporting accurate predictions while mitigating the risk of overfitting.

Secondary analyses use a Bayesian linear regression model, with the 10 PVs as the outcome, regressed on a randomly chosen set of predictors ($p = 9$). PV performance is measured in the form of the leave-one-out cross-validation information criterion (LOO-IC) (Vehtari et al., 2017), as previously discussed.

Empirical Study

For the empirical study, I use PISA 2022 reading data from the 76 countries that had publicly available item response and background items collected. The full list of countries in this analytic sample are contained in Table 2.1. Data preparation for analysis followed the same procedure as in Study 1, including imputation of missing background items and estimation of item parameters.

For the empirical study, BART-PVs were generated using the default settings of the `BART` package (Sparapani et al., 2021), as described in the previous section. Ten folds were used for cross-validation, where students were randomly assigned to training and test sets. A burn-in of 100 iterations was used, and 200 post burn-in posterior draws were retained per fold. For each posterior draw, PVs were generated by augmenting draw-specific predictions of latent ability with the estimated posterior residual standard deviation, to better reflect population variability.

4.4 Results

Simulation Study

First, I assess the impact of BART settings on the accuracy of the resulting BART-PVs, focusing on the mean and standard deviation of each PV, which are commonly reported in LSAs. Results indicate that BART settings do not make any substantial difference on the recovery of the simulated latent ability distribution mean. All combinations effectively captured the population mean. In contrast, greater variability is observed in the recovery of the population standard deviation across settings. The most accurate representation of the population standard deviation is obtained when using a larger number of trees ($ntree = 500$), in combination with a higher base parameter ($\alpha = 0.9$). A larger number of trees combined with a higher base promotes many small trees included in the ensemble, introducing a greater amount of variability that allows BART to effectively capture the population's variability. However, comparable and promising performance was observed with 200 trees and a smaller shrinkage parameter, k, while maintaining $\alpha = 0.9$. For the remaining analyses, I implement default `BART` settings, as described in the Methods section, to optimize generalizability across populations while maintaining appropriate regularization and mitigating overfitting.

Figure 4.2 shows the accuracy of BART-PVs in reconstructing the population distribution of latent ability as compared to conventional methods and a simulated distribution.

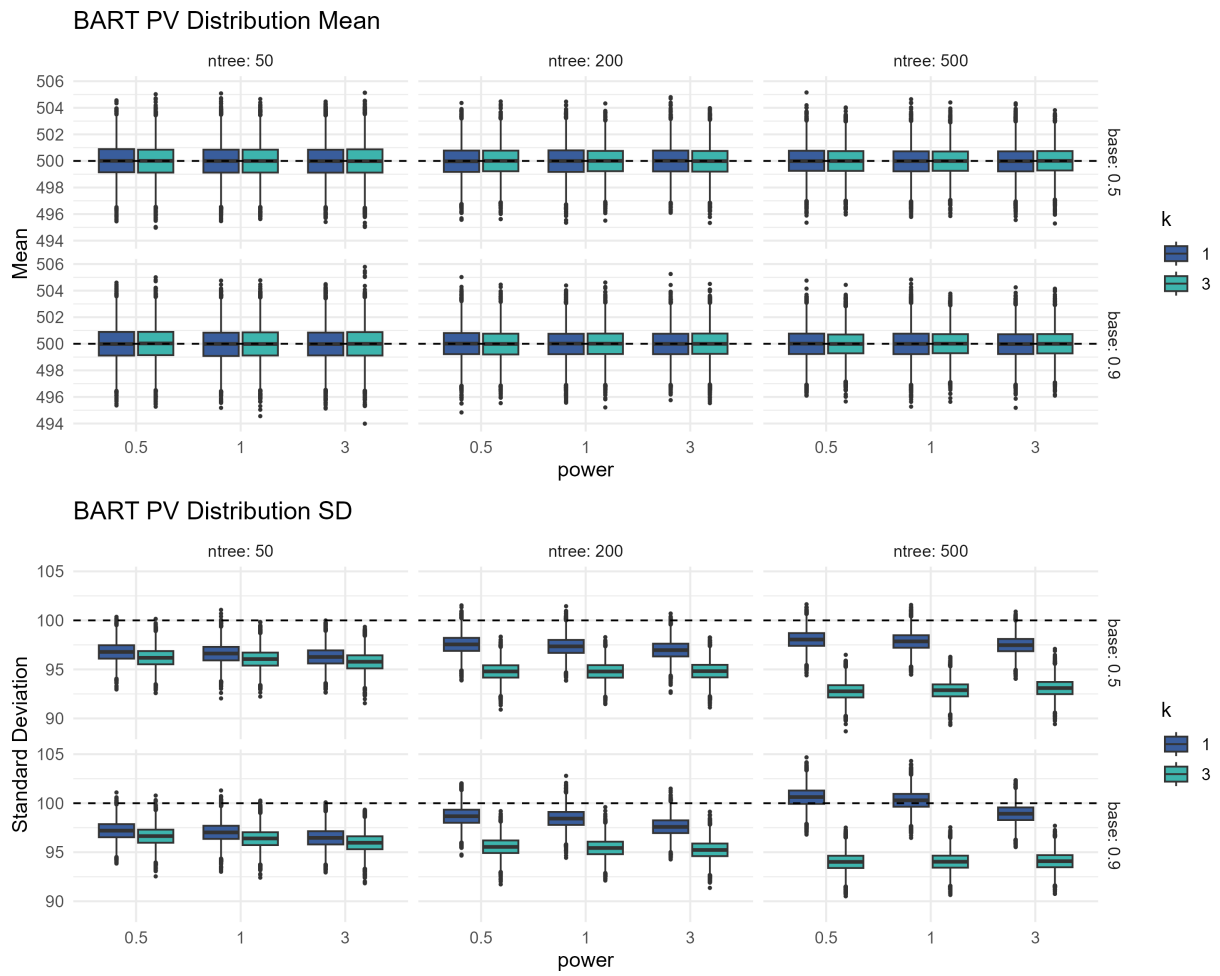


Figure 4.1: Distributions of PV means and standard deviations by BART settings in the simulation study.

Overall, BART accurately recovers both the population mean and variance, indicating that this approach is well-suited for PV generation despite its nonparametric estimation procedures. In contrast, the PISA-replicates obtained via `TAM` show a slight overestimation of population variability, although the population mean remains accurately estimated. Notably, even without specifying a parametric functional form for the latent distribution, BART is able to approximate the underlying structure and recover the population distribution. This result is particularly relevant in the context of addressing uncongeniality, as it suggests that relaxing population model assumptions can improve PV estimation while maintaining compatibility with downstream secondary analyses.

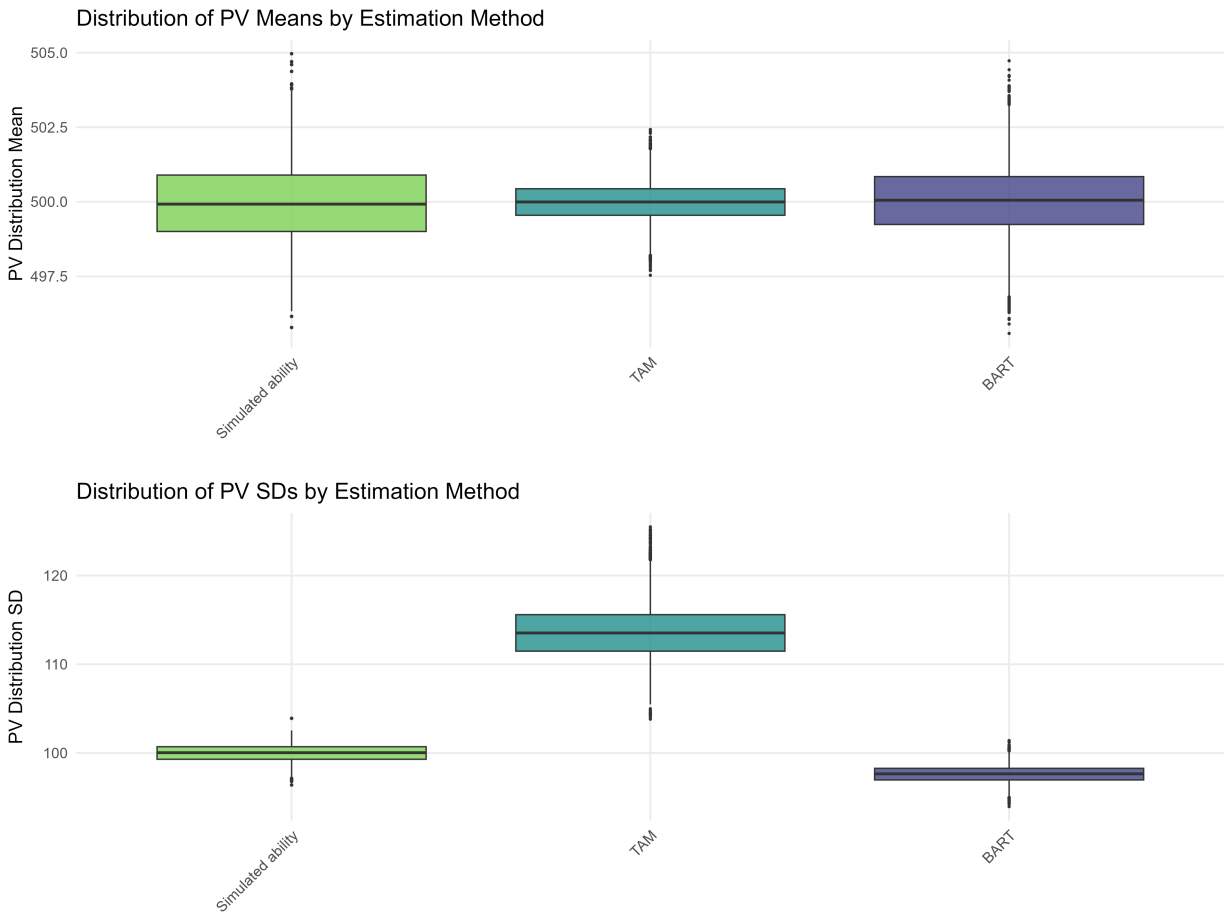


Figure 4.2: Distributions of PV means and standard deviations across methods in the simulation study.

Results from a secondary analysis (Figure 4.3) indicate that using BART-PVs as the outcome improves out-of-sample predictive performance (LOO-IC) relative to current methods and, notably, outperforms models based on the simulated latent ability estimates themselves. These simulated results indicate that BART-PVs may be a reasonable alternative to current methods of PV generation.

By taking a nonparametric, data-driven approach that does not assume a specific functional form for the latent distribution, BART relaxes restrictive model assumptions and flexibly identifies which conditioning variables are most predictive of proficiency, as well as the strength of their relationships. This flexibility not only improves the recovery

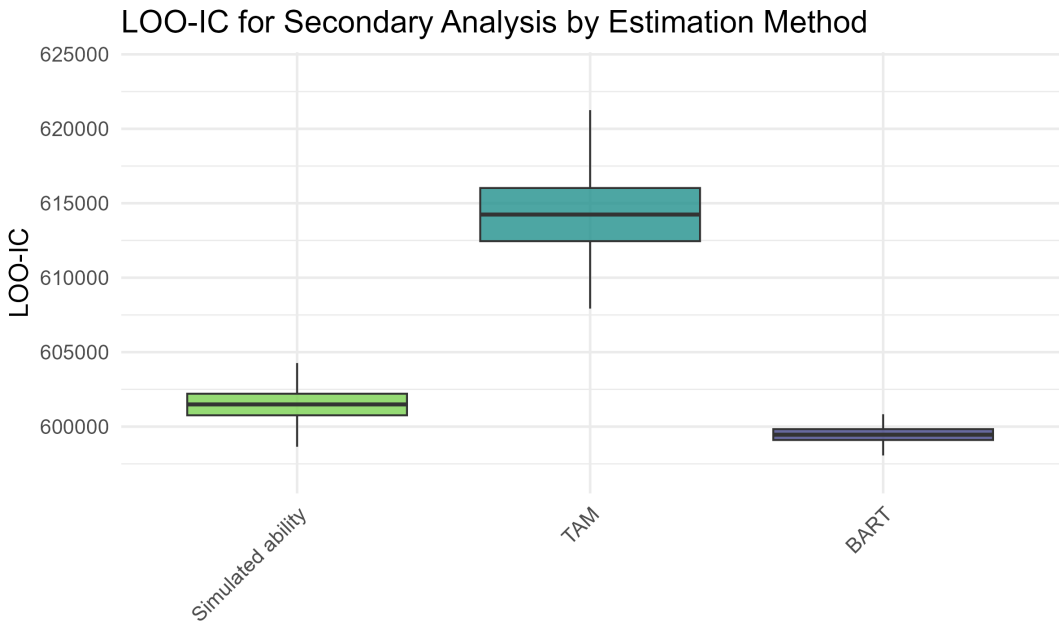


Figure 4.3: Distributions of LOO-IC's for a secondary analysis model by method in the simulation study.

of the population's latent ability distribution. but also improves out-of-sample predictive performance for secondary analysts, highlighting BART-PVs ability to generalize beyond the observed data and provide reliable estimates for secondary analyses in this simulation study.

Empirical Study

For the empirical study, 76 countries and jurisdictions with reading data were used (see Table 2.1). Country rankings and secondary analysis examples are compared to assess the performance of BART-PVs in a real-data setting. An example of the resulting PV distributions for the United States between PISA-reported PVs, my PISA replicates using `TAM`, and BART-PVs is contained in Figure 4.4.A small degree of mean bias is observed between the replication and the PISA-reported PVs. BART also slightly underestimates population variability relative to conventional methods; however, the resulting distribution remains approximately normal, and the differences are minor.

To evaluate the validity and reliability of these results and their applications, we must

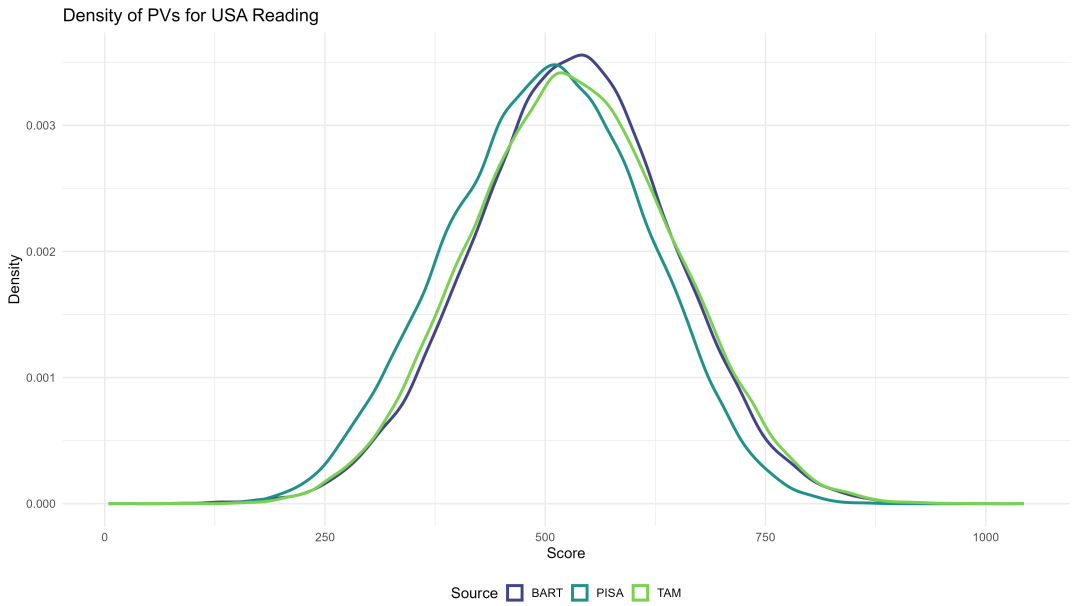


Figure 4.4: PV densities by estimation method using PISA 2022 United States (USA) reading data

first examine some relevant BART diagnostics to ensure model convergence and understand what background characteristics were the most frequently used in BART-PV estimation.

BART Diagnostics

In BART analysis, the residual standard deviation (σ) represents the model's estimate of the variability in student proficiency that is not explained by the background characteristics included as predictors. To assess stability and convergence, students were randomly divided into 10 folds, and the model was fit iteratively across these folds. A trace plot depicting the residual standard deviation across folds and iterations is shown in Figure 4.5. There, each fold is shown in a distinct color, and the dashed vertical line indicates the burn-in period. Across all folds, the chains stabilize quickly after burn-in, with σ converging and fluctuating around 95, slightly below the expected population variance of 100. This pattern demonstrates that the BART algorithm converged efficiently across folds without any distinctions between them, making the resulting estimates more reliable.

Another diagnostic of interest is the frequency of which background characteristics

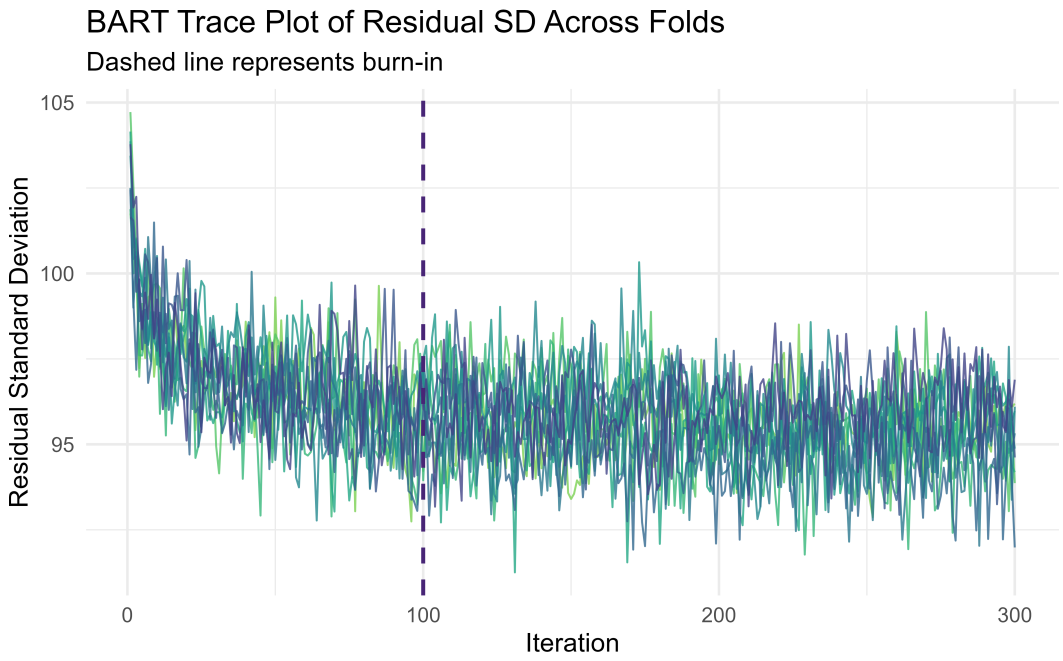


Figure 4.5: Trace plot of residual standard deviations across folds in the BART model used to estimate student proficiency for PISA 2022 United States in reading.

were used in the regression trees comprising the BART ensemble. Figure 4.6 shows the top 50 predictors included in the United States' PISA-PV generating BART model. Variables with higher inclusion proportions were more frequently used in the ensemble to partition the data, indicating they were more informative in predicting student proficiency within the ensemble. However, that does not necessarily indicate that the most frequently used predictors hold the strongest relationship between student proficiency.

For the United States, the most frequently included predictor is "ST038Q11JA" appearing at just below 0.8% of time. This variable corresponds to a student-reported item on how often they gave money to someone who threatened them at school. The next most frequently included predictor, "ST059Q02JA", measures the student-reported number of class periods per week for all subjects (OECD, 2023). Since no one predictor dominated inclusion in the ensemble, this pattern highlights that BART-PVs rely on a combination of many variables to generate the most accurate estimates of student proficiency while improving congeniality and relaxing assumptions on the functional form of the population model.

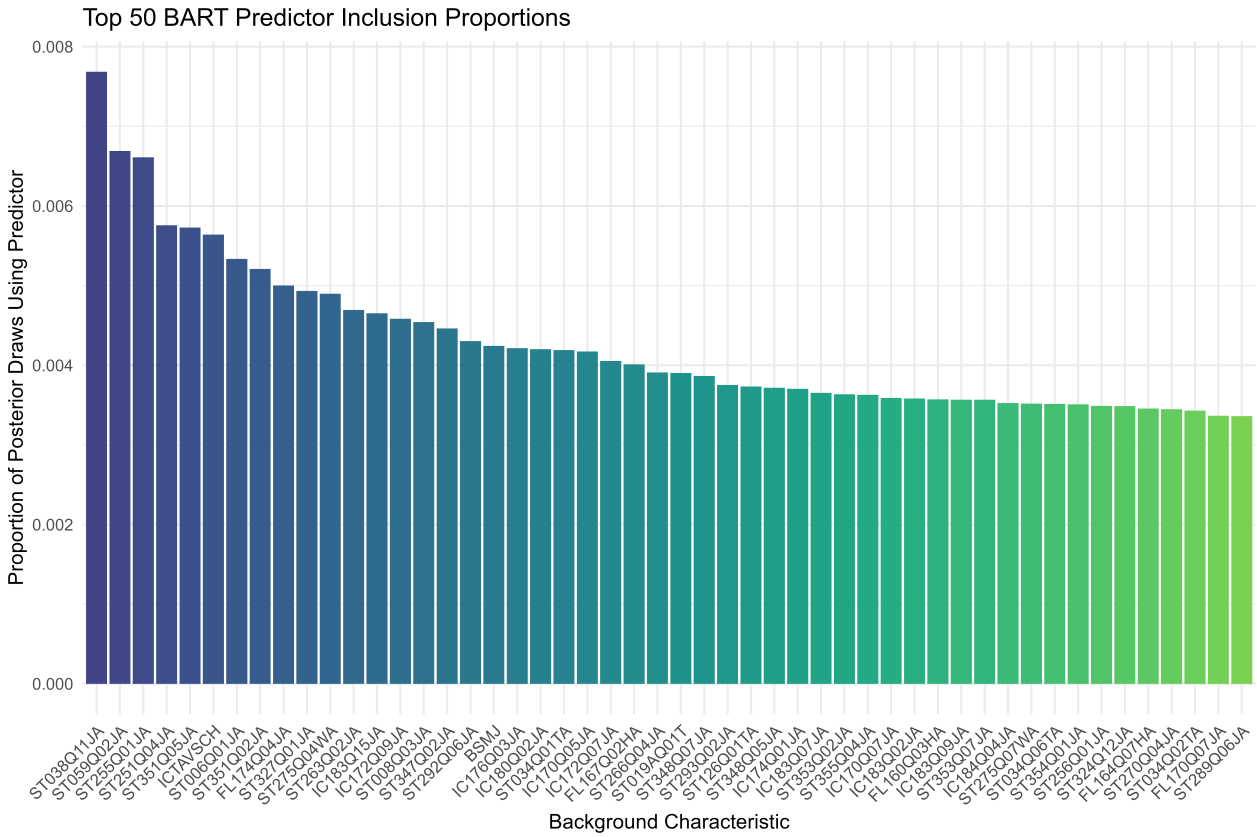


Figure 4.6: Inclusion proportions of the top 50 BART background characteristics predictors of student proficiency for the United States in PISA 2022 reading.

Country Rankings

After establishing that the BART algorithm ran as expected and produced reliable results, we can not further examine country rankings and secondary analysis applications.

Compared to PISA-replicated country rankings, BART-PVs produced only minor changes across the distribution of ranks, with the notable exception of Denmark (DNK), whose ranking increased from 26th to 18th. Aside from this case, rank changes are generally small and roughly uniformly distributed across PISA countries, suggesting BART-PVs do not introduce systematic discrepancies relative to current PV-estimation procedures aside from small amounts of random noise in estimation. Other examples of ranking changes introduced by BART-PVs include Czechia (CZE), which shifts slightly from 4th to 6th. Palestine (PSE) and Morocco (MAR) also swap places between 70th and 71st. Because these changes are

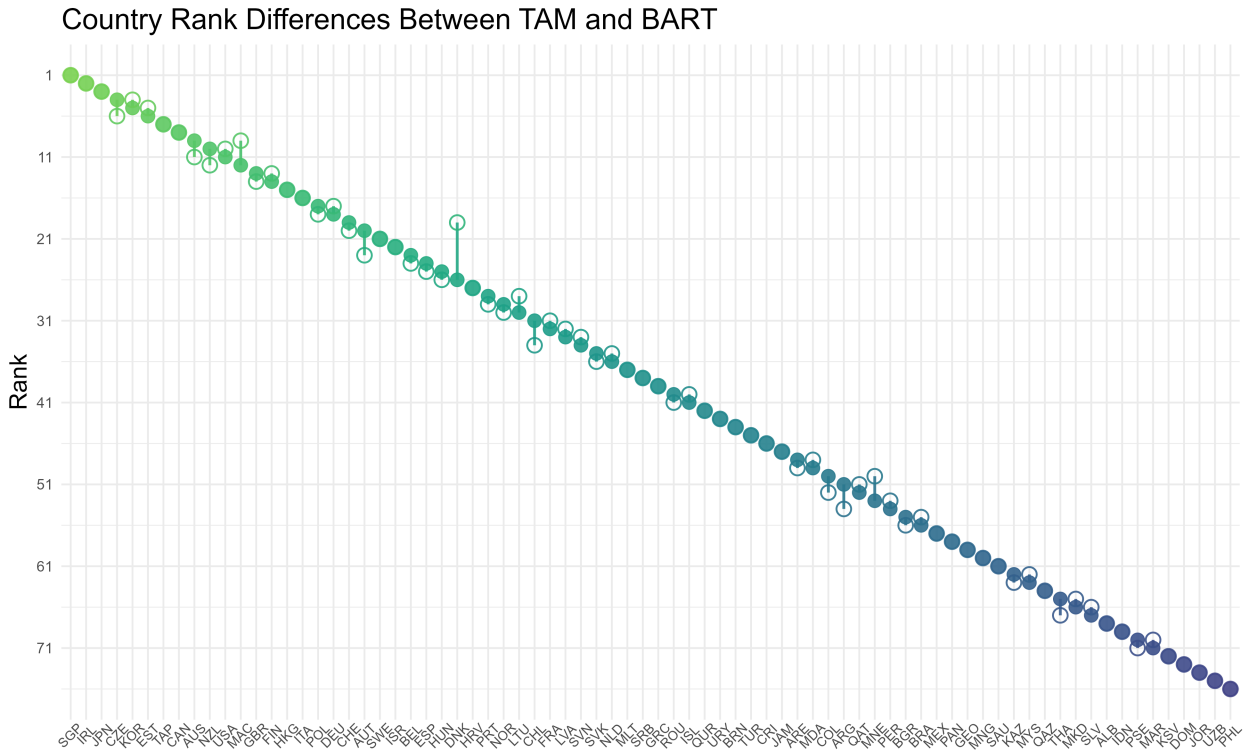


Figure 4.7: Country rankings of PISA-replicates using `TAM` and BART-PVs for reading in 2022. Note the PISA-replicate ranks in parentheses on the bottom labels. Filled-in circles denote the PISA-replicate rank, and the empty circles denote BART-PV ranks. Line segments between circles depict the magnitude of disagreement.

generally small and evenly distributed across countries, BART appears to generate more data-driven estimates of student proficiency while largely preserving the original PISA-replicated rankings. To further investigate the utility of BART-PVs in PISA-like PV estimation, an applied secondary analysis focusing on three countries of interest produces more mixed results.

Secondary Analysis

As discussed previously, an important use of PVs is in applied secondary analyses in a variety of fields to better understand factors associated with student achievement. To examine how well BART-PVs can be used in these types of analyses, I use an example Bayesian linear regression model to study predictive performance. To illustrate performance across varying positions in the achievement distribution, I examine three countries: the United States, the

Netherlands, and Panama, representing different points in overall country rankings. This analysis was conducted 10 times for each country and method, one for each PV as the outcome variable. Results are pooled and described below.

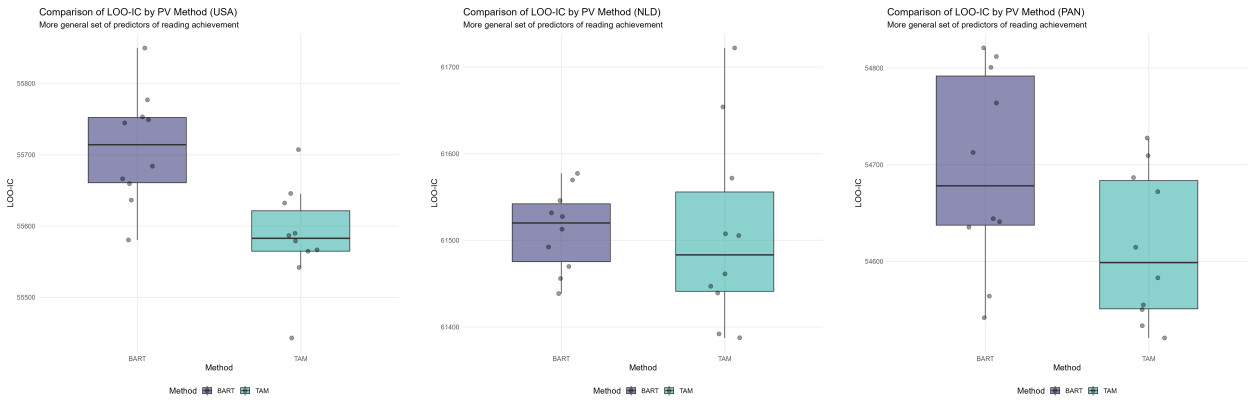
The model of interest to predict reading achievement within a given country for each student i can be specified as:

$$\begin{aligned} \text{READ}_i = \beta_0 + \beta_1\text{ESCS}_i + \beta_2\text{HISCED}_i + \beta_3\text{IMMIG}_i + \beta_4\text{LANGN}_i \\ +\beta_5\text{SEX}_i + \beta_6\text{AGE}_i + \beta_7\text{HISEI}_i + \beta_8\text{HOMEPOS}_i \\ +\beta_9\text{GROSAGR}_i + \beta_{10}\text{PROBSELF}_i + \beta_{11}\text{DISCLIM}_i \\ +\beta_{12}\text{RELATST}_i + \beta_{13}\text{BELONG}_i + \beta_{14}\text{EFFORT1}_i + \epsilon_i \end{aligned} \tag{4.5}$$

where reading achievement is predicted by a student's socioeconomic status (ESCS_i), the highest education attainment by a parent (HISCED_i), student's immigration status (IMMIG_i), language spoken most at home (LANGN_i), sex, age, a student's parental occupational status index (HISEI_i), a home possessions index reflecting family wealth and resources (HOMEPOS_i), growth mindset index (GROSAGR_i), problems with self-directed learning index (PROBSELF_i), disciplinary climate index (DISCLIM_i), student–teacher relationship quality index (RELATST_i), sense of belonging (BELONG_i), and self-reported effort on the cognitive exam (EFFORT1_i).

Figure 4.8 displays the distribution of LOO-ICs in the empirical secondary analysis example. Emphasis should be placed on the median line in the boxplots, as a regression using just a single PV (here, represented as the points), does not constitute a sufficient estimate of model performance. Across all three countries, BART-PVs yield similar and slightly higher average LOO-ICs, indicating they do not improve out-of-sample predictive performance in this empirical contexts. These findings suggest despite BART-PVs' ability to effectively reconstruct latent ability distributions, there is insufficient evidence that they improve the generalizability of secondary analyses to a given target population in PISA.

Interestingly, BART-PVs performed notably better than current approaches in the



United States Netherlands Panama

Figure 4.8: Secondary analysis LOO-ICs for a Bayesian linear regression model predicting reading achievement with 14 predictors across three countries and two methods.

simulation study (see Figure 4.3); however, these improvements did not carry over to the empirical study. This discrepancy may reflect the greater complexity, noise, or unobserved variation present in real-world PISA data, where BART's ensembling approach cannot fully specify. It's possible that BART's preference for smaller and shallower trees may have led to underfitting in this more complex scenario with over 600 predictors and 5,000 to 30,000 students observed per country. While BART-PVs can effectively reconstruct latent ability distributions, achieving consistently strong out-of-sample predictive performance appears more sensitive to noise and additional complexity in empirical settings.

4.5 Discussion

This study explored the application of Bayesian additive regression trees (BART) for generating plausible values in large-scale assessments, using both simulation studies and an empirical example with PISA 2022 reading data for 76 countries. In simulations, BART-PVs accurately recovered the population distribution of latent ability, capturing both the population distribution mean and standard deviation. Compared to traditional PV methods via `TAM` (Robitzsch et al., 2025) in a simulated secondary analysis example, BART-PVs yielded stronger out-of-sample predictive performance as well. These results demonstrate BART's

utility in its flexibility and nonparametric, data-driven approach that ensembles many weaker regression trees to produce overall better results.

However, the empirical study using PISA revealed important limitations of the current BART-PV workflow. Across the secondary analysis examples of three countries, even though BART-PVs' resulting posterior distribution of latent ability closely mimicked PISA-replicate PVs, BART-PVs did not see a boost in predictive performance over traditional PV methods compared to the results found in the simulation study. This may be due to more complex and messy data in a real world setting, where default hyperparameters were not able to adequately fit the data to improve out-of-sample predictive performance. While BART's relaxed assumptions surrounding the functional form of the data-generating population model help address model uncertainty and congeniality issues, more work is still to be done to find how BART can be used to both accurately depict PISA student populations while improving predictive performance over conventional methods.

These findings expand upon the extensive previous literature surrounding plausible values and their uses in LSAs. While previous work has been clear in the benefits of PVs in the context of LSAs (Mislevy, 1991; Mislevy et al., 1992; Wu, 2005; von Davier et al., 2009), given the relative ease in computational implementation relying on PCA and plug-in fixed estimates. However, as detailed in this work, several pressing issues have previously been underexplored. Specifically, properly addressing model uncertainty and statistical agreement or congeniality between imputation and secondary analyst models. The addition of using BART-generated PVs helps address these gaps in the literature, serving as a potentially reasonable addition to the plausible value and large-scale assessment world. If its shortcomings can be properly addressed and corrected, this approach provides a theoretically justified, practical, and computationally accessible alternative to PV methodology in LSAs.

Limitations and Future Directions

In addition to BART-PVs' failure to improve predictive performance in an applied secondary analysis example, the present work has several other limitations. First, despite a very strong rank-order correlation and minimal bias, the inability to exactly replicate PISA's reported reading scores may limit the generalizability of these findings. Additionally, the use of only Bayesian linear regression models to compare predictive performance across methods and countries provides a limited demonstration of BART-PVs. It's possible that these findings may not generalize to all modeling contexts.

Improved BART specifications are likely necessary to enhance predictive performance in real-world data settings for LSAs. In particular, a more comprehensive exploration of hyperparameter settings, such as prior specifications, tree depth, and the number of trees in the ensemble, may allow the ensemble to better capture complex, high-dimensional relationships like those with PISA. Default settings, which favor many shallow trees and strong regularization, may result in poor fit in real-world settings with substantial variation and noise.

Beyond hyperparameter tuning, more advanced extensions of BART may further improve performance. For example, new super learner approaches that combine multiple machine learning algorithms within an ensemble framework (Tyralis et al., 2021) could provide more robust predictions by leveraging complementary strengths across models. Additionally, heteroskedastic BART (HBART) (Pratola et al., 2020) allows for variance to vary across observations, offering a more flexible representation of the data-generating process. Introducing such methods may better accommodate the complex structure of large-scale assessment data, ultimately improving out-of-sample predictive performance. As interest in applying machine learning methods across diverse modeling contexts continues to grow, developing practical and effective approaches to PV estimation may further enhance the utility of large-scale assessment results across a range of applications.

5 DISCUSSION AND CONCLUSIONS

This chapter synthesizes the main findings of this dissertation and reflects on the methodological, practical, and theoretical implications for large-scale assessments. First, results from the studies are compared, evaluating how the alternative PV methods proposed in this dissertation perform relative to each other and to established conventional approaches, with a focus on population-level latent ability distribution accuracy and predictive performance. Next, a more comprehensive evaluation of these methods highlights the strengths and limitations of each alternative. This comparison focuses on their ability to handle model uncertainty and congeniality, the usefulness of the resulting PVs for secondary analyses, and practical considerations such as computation time and feasibility. Based on these factors, implementing Bayesian Model Averaging (BMA) for PV estimation emerges as a strong, practical, and straightforward approach for LSAs to address the previously identified challenges. Finally, these findings are situated within the broader literature, and suggestions for future research in this area are provided.

5.1 Key Findings

Simulation Studies

First, I compared each study's proposed method to both the current PV generation procedures and to the simulated underlying ability distribution, focusing on the mean and standard deviation of the resulting PV distribution for each replication in the simulation studies. Each simulation study used 1000 replications per method. Figure 5.1 presents these results.

Across PV estimation methods, virtually no differences were observed in the recovery of the population mean of the resulting PV distributions. This is likely because both the simulated population and the IRT-estimated ability estimates were centered around a score of

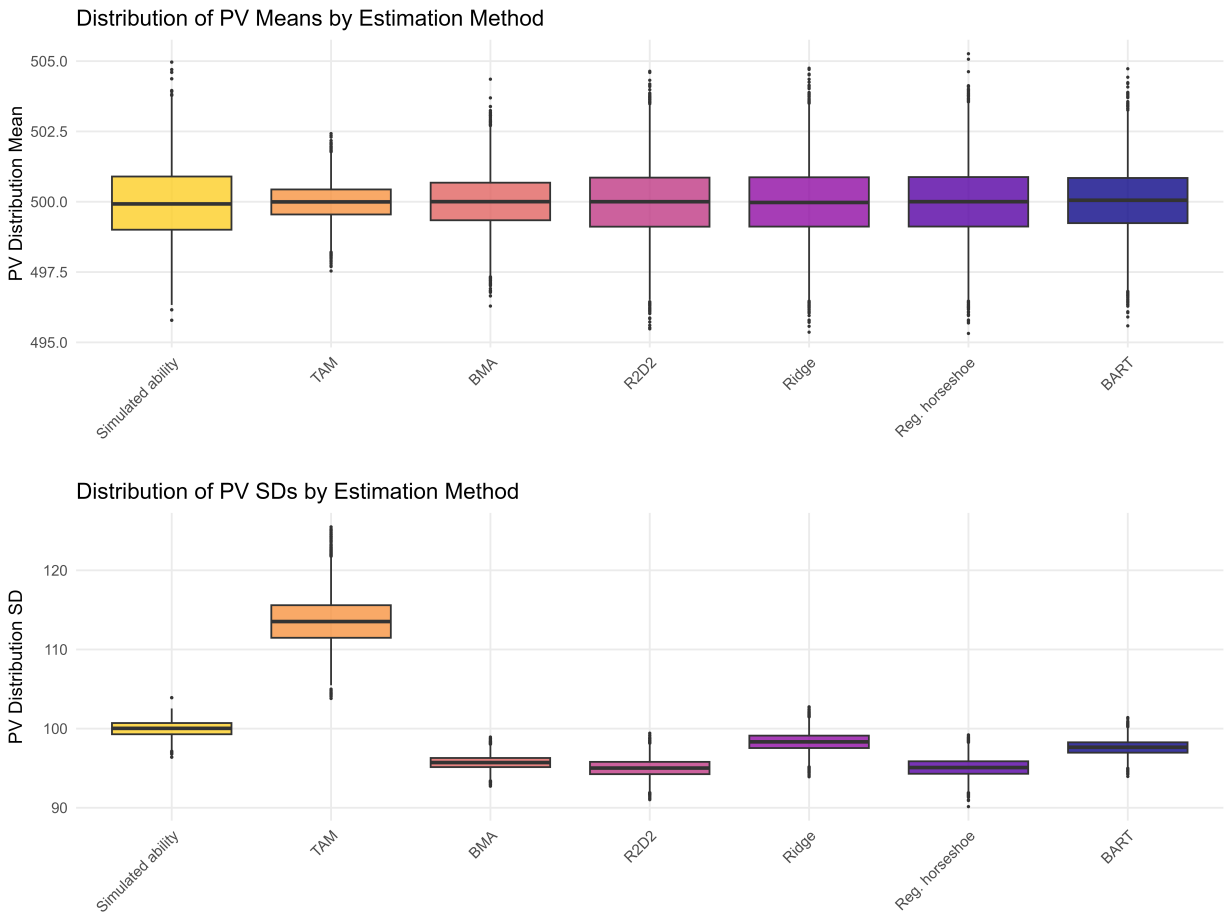


Figure 5.1: Distributions of PV means and standard deviations across methods in the simulation studies.

500, essentially anchoring the resulting PVs. Minor deviations from the mean were negligible, indicating that all methods provide reliable population mean estimates.

Differences in population distribution accuracy became more apparent when examining the resulting standard deviations. Compared to the simulated ability distribution, PVs replicating conventional methods estimated using `TAM` (Robitzsch et al., 2025) tended to slightly overestimate variability by about 10%. PVs generated using a ridge normal prior produced standard deviations that were closest to the simulated distribution. This result is expected because the ridge prior imposes a standard normal prior on all regression coefficients, aligning with the normality assumption underlying PV estimation. Consequently, the ridge

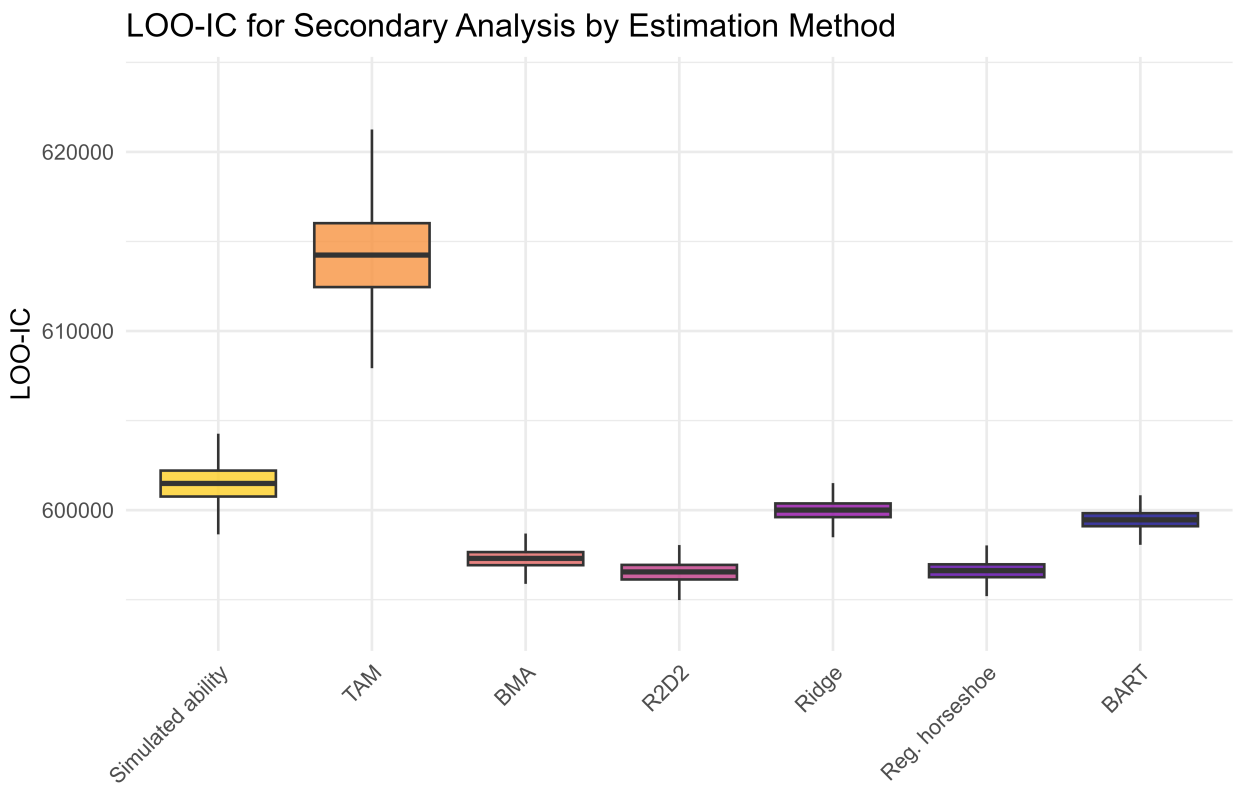


Figure 5.2: Distributions of LOO-IC's for a secondary analysis model by method across simulation studies.

prior also preserves the overall spread of the PV distribution, accurately reflecting the true variability in the simulated abilities. BART also performed well, slightly underestimating population variability but remaining closer to the simulated distribution than most other methods.

In contrast, BMA, the R2D2 prior, and the regularized horseshoe prior consistently slightly underestimated variability by roughly 5%, suggesting that small adjustments to these estimation procedures could potentially improve variance recovery. However, the magnitude of these inaccuracies are small. Overall, these results indicate that while all methods capture the mean accurately, differences in variance recovery may be an important consideration when choosing a PV estimation approach.

Next, the simulated secondary analysis example highlights clear advantages of several

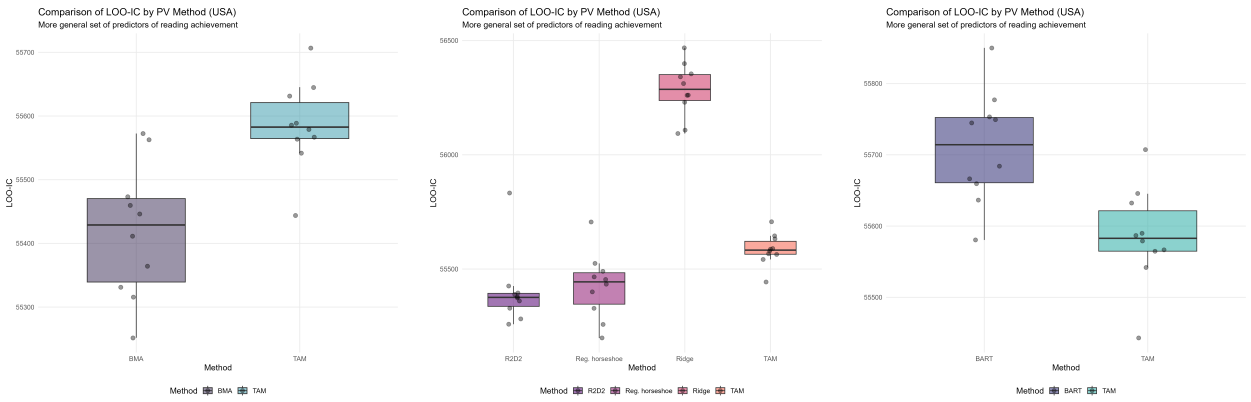
methods proposed in this dissertation relative to conventional PV estimation approaches. In terms of out-of-sample predictive performance in the form of the LOO-IC, we see that PISA-replicates via `TAM` performed substantially worse than the other methods. The BMA, R2D2, and regularized horseshoe conditions held the strongest performance, followed by the ridge and BART conditions.

From the perspective of secondary analysts seeking improved generalizability and predictive accuracy, BMA-PVs or sparsity-estimated PVs via a strong regularization prior like the R2D2 (Zhang et al., 2022) or the regularized horseshoe (Piironen and Vehtari, 2017) may be reasonable alternatives. Conversely, that the ridge-estimated PVs and BART-PVs may be exhibiting signs of underfitting. While they were able to adequately reconstruct the population distribution, they do not achieve optimal predictive performance in this comprehensive simulated study.

Empirical Studies

For the empirical study, I used PISA 2022 reading data from the 76 countries that had publicly available item responses and background items collected. The full list of countries in this analytic sample are contained in Table 2.1.

Across all three studies, country rankings exhibited only minor and seemingly random variation. This stability is expected for two main reasons. First, large sample sizes within each PISA country (5,000 students or more) limit the influence of modeling choices and prior specifications on country-level estimates. Second, PV estimation methods were applied uniformly across countries within each study, ensuring internal consistency. Consequently, any observed ranking differences are most plausibly attributable to random fluctuation rather than systematic bias. These results indicate that, although PV-generation methods may influence other outcomes, cross-country rankings in PISA are relatively robust to the choice of estimation approach. That said, even small shifts in rankings can carry meaningful implications for education policy and public perception.



BMA Sparsity BART

Figure 5.3: Secondary analysis LOO-ICs for a Bayesian linear regression model predicting reading achievement with 14 predictors using the United States' sample across studies. Note the y-axis varies across figures.

However, more substantial differences emerged in secondary analyses that used PISA PVs as outcomes in Bayesian linear regression models. Here, the choice of PV-generation method had a clear impact on out-of-sample predictive performance, as reflected in variation in LOO-IC estimates. Results from all three studies focusing on the United States' sample are summarized in Figure 5.3, highlighting the sensitivity of predictive accuracy to the PV-generation approach in applied analyses.

Figure 5.3 presents out-of-sample predictive performance across studies based on real PISA 2022 reading data for the United States. PVs generated using BMA, as well as those generated using the R2D2 and regularized horseshoe priors, showed the best performance, with the lowest average LOO-IC values compared to conventional PVs estimated via `TAM`. BART-PVs showed somewhat lower predictive performance on average, while PVs produced with a ridge prior performed notably worse than all other methods. These results suggest that again, for secondary analyses, PVs generated using BMA or strong regularization priors such as the R2D2 or regularized horseshoe priors represent reliable and robust options for improving out-of-sample predictive accuracy in a real-world data setting. However, when choosing a PV-estimation method, there are other factors to consider besides population distribution recovery and predictive performance.

5.2 Comparative Evaluation of Methods

Taken together, the findings of this work illustrate not only differences in PV accuracy and predictive performance in downstream analyses, but also the characteristics and trade-offs of each approach. To further summarize these considerations between methods to inform recommendations to LSAs like PISA, Table 5.1 presents a brief comparison of the strengths and limitations of all methods discussed.

Conventual PV estimation methods, such as those used in PISA and similar large-scale assessments, provide stable population-level inferences and appropriately account for measurement uncertainty through multiple draws from the posterior distribution. These methods have been well-studied for decades (Mislevy et al., 1992; Wu, 2005; von Davier et al., 2009; Marsman et al., 2016; Braun and von Davier, 2017; Jewsbury et al., 2024). However, several key limitations remain. Current approaches introduce uncongeniality and fail to fully account for model uncertainty by relying on a single imputation model. As demonstrated in this work, they also exhibit weaker predictive performance in secondary analyses compared to the Bayesian approaches presented here. As illustrated throughout this work, these limitations must be addressed to further refine conventional PV methods used by LSAs like PISA.

Bayesian Model Averaging-based PVs (BMA-PVs) addresses these limitations by explicitly incorporating model uncertainty and directly including background characteristics in the population model, improving congeniality and generating more accurate and predictively optimal secondary analyses. Computationally, BMA is faster than other Bayesian approaches since it uses a Metropolis-Hastings algorithm to explore the model space with a birth-death sampler, as opposed to calculating the marginal likelihood of all possible models (Zeugner and Feldkircher, 2015).

An additional benefit of BMA-PVs is that BMA essentially induces a spike-and-slab prior, a widely recognized and highly regarded approach for regularization and variable selection (Mitchell and Beauchamp, 1988; George and McCulloch, 1993; Ishwaran and Rao,

Table 5.1: Statistical and practical comparisons of methods presented in this work.

Method	Strengths	Limitations
Conventional Methods	• Widely accepted in LSAs. • Accounts for measurement uncertainty. • Stable population-level estimates.	• Not optimized for prediction. • Introduces uncongeniality and bias. • Ignores model specification uncertainty.
Bayesian Model Averaging	• Accounts for model uncertainty. • Strong predictive performance. • Mitigates uncongeniality. • Computationally efficient relative to other Bayesian approaches.	• More complex and less familiar in LSAs. • Assumes the true model is in the candidate set.
Ridge Prior	• Handles multicollinearity well. • Stable and computationally efficient. • Mitigates uncongeniality.	• Weak predictive performance. • Limited shrinkage. • Does not account for model uncertainty.
R2D2 Prior	• Adaptive global and local shrinkage with fewer hyperparameters. • Strong predictive performance. • Mitigates uncongeniality.	• Computationally intensive. • Less established and understood in practice. • Does not account for model uncertainty.
Regularized Horseshoe Prior	• Adaptive global and local shrinkage. • Strong predictive performance. • Mitigates uncongeniality.	• Computationally intensive. • Can be sensitive to hyperparameters. • Does not account for model uncertainty.
Bayesian Additive Regression Trees	• Computationally efficient relative to other Bayesian approaches. • Addresses model uncertainty and uncongeniality. • Accurate distribution recovery.	• Lower predictive performance with PISA. • Limited interpretability. • More complex and less familiar in LSAs.

2005). The model-averaged coefficients from BMA reflect the sparsity effects induced by this prior. This means that BMA simultaneously addresses model selection uncertainty and performs shrinkage, yielding better predictive performance than any single candidate model, assuming the true data-generating model is included in the model space. We can thus view BMA-PVs as a combination of model ensembling and sparsity-inducing. These advantages of BMA come with an increased methodological complexity, requiring familiarity with BMA and potentially institutional and software adjustments to implement in LSAs.

Study 2 introduces Bayesian regularization as an alternative approach to address limitations of conventional PV methods, using ridge, R2D2, and regularized horseshoe priors (Hsiang, 1975; Zhang et al., 2022; Piironen and Vehtari, 2017). The R2D2 and regularized horseshoe priors substantially improve predictive performance in secondary analyses while enhancing congeniality, as they allow the population model to select the most relevant variables directly rather than relying on PCA. The ridge prior, though computationally simpler and faster, does not offer meaningful predictive gains over conventional methods. These regularization approaches, however, do not address imputation model selection uncertainty and can be sensitive to hyperparameter choices, requiring additional expertise to fine-tune and troubleshoot.

Another glaring practical limitation of R2D2 and regularized horseshoe-estimated PVs is computation time. For example, generating PISA 2022 reading PVs for Spain required over 10 hours. This substantial computational cost currently limits the feasibility of applying these methods in LSAs. Future improvements in computational efficiency may mitigate this issue, however, that lies outside the scope of this work.

Finally, Bayesian additive regression trees (BART) offer a flexible, nonparametric machine learning approach that relaxes assumptions about the form of the data-generating model (Chipman et al., 2010). BART addresses both congeniality and model uncertainty while implicitly performing variable selection through its ensemble structure. Its iterative back-fitting algorithm with a Gibbs sampler allows faster computation compared to other

MCMC-based methods, depending on the number of cross-validation folds used (Chipman et al., 2010). In this work, BART was able to successfully recover the population distribution of ability and showed strong predictive performance in a simulated data setting, but these predictive gains were lacking using real PISA 2022 data. This may be due to increased complexity and noise that the default BART settings in the `BART` package could not fully accommodate (Sparapani et al., 2021). Additional fine-tuning of hyperparameters may improve results, but this would require additional research and careful adjustments for each country-specific ensemble to ensure accuracy.

5.3 Recommendations & Implications

Based on the findings of this work, several key takeaways emerge for LSAs moving forward. Traditional PV methods, such as those used in PISA, are reliable and well-established, but they have clear limitations. By relying on a single imputation model that uses PCA, they can introduce bias and reduce predictive accuracy in secondary analyses.

BMA, as demonstrated in Study 1, addresses these issues by considering a much larger possible population model space. This approach not only improves predictive performance but also implicitly induces shrinkage and performs variable selection, allowing the algorithm to identify the most important combinations of variables and take a weighted average. BMA is also computationally practical and feasible to implement compared to other Bayesian methods, making it a strong option for real-world PV generation. For these reasons, BMA appears to be the most promising approach proposed in this work, balancing statistical soundness, predictive performance, and practical considerations. It offers LSAs a flexible and robust method for generating plausible values that are well-suited for secondary analysis.

Alternatively, large-scale assessments could consider implementing different methods of PV estimation tailored to specific purposes. For example, conventional PVs could continue to be used for official country-level reporting and rankings, ensuring continuity and comparability

with previous cycles, while BMA-PVs could be recommended for secondary analyses, where improved predictive accuracy and generalizability are more critical. Both approaches are relatively straightforward to implement and computationally efficient, making such a multi-PV strategy feasible in practice.

However, adopting different PVs for different purposes is not without potential drawbacks. Using distinct PV sets may introduce a new source of uncongeniality, as secondary analyses would rely on posterior distributions of student proficiency that differ from those used for official reporting. This could create confusion for researchers and policymakers interpreting cross-country comparisons, particularly if different PV sets yield subtly different rankings or inferences. Careful documentation and guidance would be necessary to mitigate these risks and ensure that users understand the intended applications of each PV set. Regardless, there may be a substantial gap between the intended use of PVs and how they are actually applied by typical researchers. Introducing additional complexity and nuance into PVs' applications could increase the likelihood of misuse, potentially undermining the reliability of secondary inferences.

Other proposed approaches, like Bayesian regularization via sparsity-inducing priors or BART, offer interesting alternatives that may require more study before their benefits in PV generation can be fully realized. Sparsity-inducing priors to generate PVs can improve predictive performance but are computationally much slower and require more careful attention to convergence and fine-tuning. BART is highly flexible but in practice requires additional fine-tuning to optimize its benefits in real-world settings.

In summary, LSAs should prioritize PV methods that explicitly handle model uncertainty, improve congeniality, and incorporate Bayesian approaches to maximize the usefulness of their results across the entire distribution of latent ability. Among the methods considered here, BMA offers the strongest combination of practical feasibility and predictive performance, making it the clear front-runner.

5.4 Limitations & Future Research

This dissertation holds several limitations that should be acknowledged. First, the simulation studies, while extensive, rely on specific assumptions about the population distribution and structure of the background variables, which may not fully capture all the complexities present in a real-world LSA setting. Additionally, both the simulation study and empirical study's secondary analysis examples only examined a Bayesian linear regression example. While this is a commonly used modeling strategy in many fields that rely on LSAs for applied research, the results of this work may not necessarily generalize to alternative approaches, such as latent variable models or more complex hierarchical or structural models.

Second, the methods proposed in this work rely on the presence of meaningful relationships between student ability and observed background characteristics. In settings where covariates are sparse, poorly measured, or only weakly associated with student ability, the underlying assumptions of the proposed alternative methods may not hold. Under such conditions, approaches that depend on predicting ability from background characteristics, like those considered here, may yield unstable estimates, excessive shrinkage, or effectively arbitrary model weights. This can reduce the predictive accuracy of the resulting plausible values and diminish the practical advantages of these methods. As a result, their utility is limited in contexts where background characteristics are limited or uninformative, and more conventional estimation approaches may be better suited for settings that differ substantially from those examined in this work.

Third, while the proposed methods improve predictive performance and address issues of current PV methods, they may introduce additional complexity that could affect practical implementation and the interpretability of PVs for secondary analysts, as noted in the previous section.

Despite these limitations, there are several promising avenues for future research in plausible value estimation. Further fine-tuning of the methods used here, such as R2D2 and regularized horseshoe-based PVs, or BART-PVs, could address some of the present

work's limitations. Researchers could also explore alternative modeling approaches that further balance accuracy and interpretability, or examine the performance of PV methods in different contexts, including smaller samples or subgroups. Future work could also investigate strategies for managing multiple sets of reported PVs for different purposes while minimizing uncongeniality and potential misuse.

Integrating these methods into applied software tools, such as new `R` packages could increase accessibility for applied researchers and policymakers. Comprehensive documentation that includes clear, non-technical explanations of these methods would be essential.

In conclusion, this dissertation has explored new Bayesian approaches to plausible value estimation, demonstrating both methodological advancements and practical implications for large-scale assessments. By implementing modern Bayesian methods, sparsity-inducing priors, and rigorous evaluation through simulation and empirical studies, this work has highlighted the potential for more accurate and generalizable PVs that better support secondary analyses. At the same time, it has acknowledged limitations and practical considerations, emphasizing the importance of careful implementation and transparency. Ultimately, these contributions aim to bridge the gap between methodological innovation and applied research, providing tools and insights that can elevate the rigor, reliability, and interpretability of future large-scale assessment studies.

REFERENCES

Allen, D. (1974). The relationship between variable selection and data augmentation and a method for prediction. *Technometrics*, 16:125–127.

Araujo, L., Saltelli, A., and Schnepf, S. (2017). Do PISA data justify PISA-based education policy? *International Journal of Comparative Education and Development*, 19.

Benton, T. (2017). The effect of using principal components to create plausible values. In *The Annual Meeting of the Psychometric Society*, volume 265, pages 293–306. Springer International Publishing.

Bernardo, J. and Smith, A. F. (2000). *Bayesian Theory*. Wiley, New York.

Betancourt, M. (2018). Bayes sparse regression.

Braun, H. and von Davier, M. (2017). The use of test scores from large-scale assessment surveys: Psychometric and statistical considerations. *Large-scale Assessments in Education*, 5:1–16.

Brent, L., McCulloch, R., Sparapani, R., and Laud, P. (2017). Introduction to BART: BART::wbart. Presentation, Medical College of Wisconsin, September 30, 2017. Accessed: 2025-10-07.

Bürkner, P. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, 80(1):1–28.

Carvalho, C., Polson, N., and Scott, J. (2009). Handling sparsity via the horseshoe. *Artificial Intelligence and Statistics*, pages 73–80.

Carvalho, C., Polson, N., and Scott, J. (2010). The horseshoe estimator for sparse signals. *Biometrika*, 97:465–480.

Center for High Throughput Computing (2006). Center for high throughput computing.

Chipman, H., George, E., and McCulloch, R. (2010). BART: Bayesian additive regression trees. *Annals of Applied Statistics*, 4:266–298.

Clyde, M. (2025). *BAS: Bayesian Variable Selection and Model Averaging using Bayesian Adaptive Sampling*. R package version 2.0.2.

Feldkircher, M. and Zeugner, S. (2009). Benchmark priors revisited: On adaptive shrinkage and the supermodel effect in Bayesian model averaging. IMF Working Paper 09/202, International Monetary Fund.

Feldkircher, M. and Zeugner, S. (2015). Bayesian model averaging employing fixed and flexible priors: The BMS package for R. *Journal of Statistical Software*, 68.

Fernández, C., Ley, E., and Steel, M. (2001). Benchmark priors for Bayesian model averaging. *Journal of Econometrics*, 100:381–427.

Gabry, J., Češnovar, R., Johnson, A., and Bronder, S. (2025). *cmdstanr: R Interface to 'CmdStan'*. R package version 0.9.0, https://discourse.mc-stan.org.

Gelman, A., Hwang, J., and Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, 24:997–1016.

George, E. and McCulloch, R. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88:881–889.

Gronau, Q. and Wagenmakers, E. (2019). Limitations of Bayesian leave-one-out cross-validation for model selection. *Computational Brain and Behavior*, 2.

Grund, S., Lüdtke, O., and 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*, 46:430–465.

Hannan, E. and Quinn, B. (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(2):190–195.

Harra, K. and Kaplan, D. (2024). On the performance of horseshoe priors for inducing sparsity in structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, 31(4):667–684.

Harra, K. and Kaplan, D. (2025). Incorporating sparsity into Bayesian stacking procedures. In *Proceedings of the 89th Annual International Meeting of the Psychometric Society, Prague, Czech Republic, 2024.*

He, Y., Zaslavsky, A., Landrum, M., Harrington, D., and Catalano, P. (2010). Multiple imputation in a large-scale complex survey: A practical guide. *Statistical Methods in Medical Research*, 19:653–670.

Hill, J., Linero, A., and Murray, J. (2020). Bayesian additive regression trees: A review and look forward. *Annual Review of Statistics and Its Applications*, 7:251–278.

Hoerl, A. and Kennard, R. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Hoerl, R. (1985). Ridge analysis 25 years later. *The American Statistician*, 39(3):186–192.

Hoeting, J., Madigan, D., Raftery, A., and Volinsky, C. (1999). Bayesian model averaging: A tutorial. *Statistical Science*, 14:382–417.

Hsiang, T. (1975). A Bayesian view on ridge regression. *Source: Journal of the Royal Statistical Society. Series D (The Statistician)*, 24:267–268.

Ishwaran, H. and Rao, J. S. (2005). Spike and slab variable selection: Frequentist and bayesian strategies.

Jacobucci, R. and Grimm, K. J. (2018). Comparison of frequentist and Bayesian regularization in structural equation modeling. *Structural Equation Modeling*, 25:639–649.

Jewsbury, P., Jia, Y., and Gonzalez, E. (2024). Considerations for the use of plausible values in large-scale assessments. *Large-Scale Assessments in Education*, 12.

Jewsbury, P. and Johnson, M. (2025). Principal component analysis on the covariance matrix for data reduction in large-scale assessments. *Large-Scale Assessments in Education*, 13.

Jewsbury, P., Lockwood, J., and Johnson, M. (2025). IRT-latent regression with many predictors: limits and solutions. *Large-scale Assessments in Education*, 13.

Johnson, M. and Jenkins, F. (2004). A Bayesian hierarchical model for large-scale educational surveys: An application to the national assessment of educational progress. *ETS Research Report Series*, 2004(2):i–28.

Johnson, N., Kotz, S., and Balakrishnan, N. (1994). *Continuous univariate distributions, volume 1*, volume 1. John wiley & sons.

Kaplan, D. (2021). On the quantification of model uncertainty: A Bayesian perspective. *Psychometrika*, 86:215–238.

Kaplan, D. (2023). *Bayesian statistics for the social sciences*. Guilford Press, New York, 2nd edition.

Kaplan, D., Harra, K., Stampka, J., and Jude, N. (2025). Stacking models of growth: A methodology for predicting the pace of progress to the education sustainable development targets using international large-scale assessments. *Psychometrika*, 90(2):658–686.

Kaplan, D. and Huang, M. (2021). Bayesian probabilistic forecasting with large-scale educational trend data: a case study using NAEP. *Large-Scale Assessments in Education*, 9(1).

Kaplan, D. and Su, D. (2018). On imputation for planned missing data in context questionnaires using plausible values: a comparison of three designs. *Large-Scale Assessments in Education*, 6.

Kaplan, D. and Yavuz, S. (2020). An approach to addressing multiple imputation model uncertainty using Bayesian model averaging. *Multivariate Behavioral Research*, 55:553–567.

Khorramdel, L., von Davier, M., Gonzalez, E., and Yamamoto, K. (2020). *Plausible Values: Principles of Item Response Theory and Multiple Imputations*, pages 27–47. Springer Nature.

Kiernan, P. (2024). Which is greater? the number of atoms in the universe or the number of chess moves? Accessed: 2026-02-26.

Leamer, E. (1978). *Specification Searches: Ad Hoc Inference with Nonexperimental Data*. Wiley, New York.

Ley, E. and Steel, M. F. J. (2009). On the effect of prior assumptions in Bayesian model averaging with applications to growth regression. *Journal of Applied Econometrics*, 24:651–674.

Liang, F., Paulo, R., Molina, G., Clyde, M., and Berger, J. (2008). Mixtures of g-priors for Bayesian variable selection. *Journal of the American Statistical Association*, 103:410–423.

Little, R. and Rubin, D. (2019). *Statistical analysis with missing data*. John Wiley & Sons, 3rd. edition.

Marsman, M., Maris, G., Bechger, T., and Glas, C. (2016). What can we learn from plausible values? *Psychometrika*, 81:274–289.

Martens, K. and Niemann, D. (2013). When do numbers count? the differential impact of the PISA rating and ranking on education policy in germany and the us. *German Politics*, 22:314–332.

Matta, T., Rutkowski, L., Rutkowski, D., and Liaw, Y.-L. (2018). lsasim: An R package for simulating large-scale assessment data. *Large-scale Assessments in Education*, 6(1):1–33.

Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical science*, pages 538–558.

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

Mislevy, R., Beaton, A., Kaplan, B., and Sheehan, K. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement*, 29:133–161.

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83(404):1023–1032.

Monseur, C. and Adams, R. (2009). Plausible values: How to deal with their limitations. *Journal of Applied Measurement*, 10.

Murray, J. (2018). Multiple imputation: a review of practical and theoretical findings. *Statistical Science*, 33(2):142–159.

OECD (2009). Plausible values. In *PISA Data Analysis Manual: SPSS Second Edition*, pages 93–322. OECD, Paris.

OECD (2023). PISA 2022 results (volume I): The state of learning and equity in education.

OECD (2024). *PISA 2022 Technical Report*. OECD Publishing, Paris.

Okubo, T. (2022). Theoretical considerations on scaling methodology in PISA. OECD education working papers, no. 282, OECD Publishing, Paris, France.

Özer, M. (2020). What does PISA tell us about performance of education systems? *Bartın University Journal of Faculty of Education*, 9:217–228.

Park, T. and Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association*, 103:681–686.

Piironen, J. and Vehtari, A. (2017). Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Statistics*, 11:5018–5051.

Pratola, M., Chipman, H., George, E., and McCulloch, R. (2020). Heteroscedastic bart via multiplicative regression trees. *Journal of Computational and Graphical Statistics*, 29(2):405–417.

Raftery, A. (1995). Bayesian model selection in social research (with discussion). In Marsden, P. V., editor, *Sociological Methodology*, volume 25, pages 111–196. Blackwell, New York.

Raftery, A., Madigan, D., and Hoeting, J. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, 92:179–191.

Robitzsch, A., Kiefer, T., and Wu, M. (2025). *TAM: Test Analysis Modules*. R package version 4.3-25.

Ročková, V. and Saha, E. (2019). On theory for BART. In Chaudhuri, K. and Sugiyama, M., editors, *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS 2019)*, volume 89 of *Proceedings of Machine Learning Research*, pages 2839–2848. PMLR.

Rubin, D. (1987). *Multiple imputation for nonresponse in surveys*. John Wiley & Sons.

Sparapani, R., Spanbauer, C., and McCulloch, R. (2021). Nonparametric machine learning and efficient computation with Bayesian additive regression trees: The BART R package. *Journal of Statistical Software*, 97:1–66.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36:111–147.

Tan, Y. and Roy, J. (2019). Bayesian additive regression trees and the general BART model. *Statistics in Medicine*, 38(25):5048–5069.

Tyralis, H., Papacharalampous, G., and Langousis, A. (2021). Super ensemble learning for daily streamflow forecasting: Large-scale demonstration and comparison with multiple machine learning algorithms. *Neural Computing and Applications*, 33(8):3053–3068.

Van Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of statistical software*, 45:1–67.

van Erp, S., Oberski, D., and Mulder, J. (2019). Shrinkage priors for Bayesian penalized regression. *Journal of Mathematical Psychology*, 89:31–50.

Vehtari, A., Gelman, A., and Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and Computing*, 27:1413–1432.

von Davier, M. (2013). Imputing proficiency data under planned missingness in population models. *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis*, 8:175–201.

von Davier, M., Gonzalez, E., and Mislevy, R. (2009). What are plausible values and why are they useful. *IERI monograph series*, 2(1):9–36.

Warm, T. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, 54(3):427–450.

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

Yao, Y., Vehtari, A., Simpson, D., and Gelman, A. (2018). Using stacking to average Bayesian predictive distributions (with discussion). *Bayesian Analysis*, 13:917–1007.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g prior distributions. In Goel, P. and Zellner, A., editors, *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti. Studies in Bayesian Econometrics*, pages 233–243. Elsevier, New York.

Zeugner, S. and Feldkircher, M. (2015). Bayesian model averaging employing fixed and flexible priors: The BMS package for R. *Journal of Statistical Software*, 68(4):1–37.

Zhang, Y., Naughton, B., Bondell, H., and Reich, B. (2022). Bayesian regression using a prior on the model fit: The R2-D2 shrinkage prior. *Journal of the American Statistical Association*, 117:862–874.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67:301 – 320.