PSYCHOMETRIKA—VOL. 89, NO. 4, 1119–1147
DECEMBER 2024
https://doi.org/10.1007/s11336-024-09978-1

OPTIMIZING LARGE-SCALE EDUCATIONAL ASSESSMENT WITH A "DIVIDE-AND-CONQUER" STRATEGY: FAST AND EFFICIENT DISTRIBUTED BAYESIAN INFERENCE IN IRT MODELS

SAINAN XU AND JING LU

NORTHEAST NORMAL UNIVERSITY

JIWEI ZHANG

NORTHEAST NORMAL UNIVERSITY

CHUN WANG

UNIVERSITY OF WASHINGTON

GONGJUN XU

UNIVERSITY OF MICHIGAN

With the growing attention on large-scale educational testing and assessment, the ability to process substantial volumes of response data becomes crucial. Current estimation methods within item response theory (IRT), despite their high precision, often pose considerable computational burdens with large-scale data, leading to reduced computational speed. This study introduces a novel "divide- and-conquer" parallel algorithm built on the Wasserstein posterior approximation concept, aiming to enhance computational speed while maintaining accurate parameter estimation. This algorithm enables drawing parameters from segmented data subsets in parallel, followed by an amalgamation of these parameters via Wasserstein posterior approximation. Theoretical support for the algorithm is established through asymptotic optimality under certain regularity assumptions. Practical validation is demonstrated using real-world data from the Programme for International Student Assessment. Ultimately, this research proposes a transformative approach to managing educational big data, offering a scalable, efficient, and precise alternative that promises to redefine traditional practices in educational assessments.

Key words: large-scale testing, item response theory, divide-and-conquer strategy, distributed Bayesian inference, Wasserstein posterior.

Large-scale educational testing and assessments have gained global attention, engaging educators, researchers, and policymakers worldwide. This interest arises from significant changes in organized assessments over the last fifty years, resulting in rich and extensive educational big datasets. One significant example is the Programme for International Student Assessment (PISA), conducted by OECD. PISA collects data every three years from 15-year-olds worldwide to measure students' abilities in applying their knowledge and skills in reading, mathematics, and science to solve real-world problems. PISA 2018 collected data from 606,627 students who answered 82 mathematics cognitive items (OECD, 2021) . While not every student responds to each question, after eliminating the responses of students who did not answer any of the items, the

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11336-024-09978-1.

Correspondence should be made to Jing Lu, Key Laboratory of Applied Statistics of MOE, School of Mathematics and Statistics, Northeast Normal University, Changchun, Jilin, China. Email: luj282@nenu.edu.cn

Correspondence should be made to Jiwei Zhang, Faculty of Education, Key Laboratory of Applied Statistics of MOE, Northeast Normal University, Changchun, Jilin, China. Email: zhangjw713@nenu.edu.cn

sample size still remains as high as 267,889. Furthermore, large-scale educational assessments conducted by individuals or schools are also notable. For example, to study the "critical period" in Second Language Acquisition, Hartshorne et al. (2018) administered a grammar test through social media to English-native and non-native speakers (Wu et al., 2020) . The test consists of 95 binary items and was completed by 669,498 participants without missing data. Other large-scale testing like the National Assessment of Educational Progress (NAEP; van Rijn et al. (2016)) and the Trends in International Mathematics and Science Study (TIMSS; Martin and Kelly (1996)) belong to this landscape. Clearly, these extensive educational response datasets share a common characteristic: big data and large sample sizes. Note that "big data" and "large sample size" in this paper refer to high number of individuals rather than a large number of items. Therefore, conducting a sound analysis of these large response data holds significant value for educational evaluation and item quality assessment.

Leveraging these big datasets presents a unique opportunity and a challenge. It enables researchers and educators to gain insights into student performance on an unprecedented scale, thereby allowing more informed decision-making regarding educational policies and teaching strategies. Moreover, it facilitates a deeper grasp of student learning processes, capturing not just final answers but the methods used to reach them. However, analyzing such datasets is complex. It requires robust statistical methodologies and the computational capacity to process the educational big data. Thus, ongoing investment in innovative methods and computational tools is crucial to harness the full potential of these large-scale educational datasets.

Item response theory (IRT; van der Linden and Hambleton (1997)) is a powerful tool for test design, analysis, and scoring. It employs a probabilistic model to explain the relationship between an individual's latent ability and the probability of correctly answering a particular item on a test. This theory is particularly relevant in the context of large-scale educational assessments. For example, IRT models are applied to analyze the collected data for PISA test (Embretson & Reise, 2000) . The analysis offers insights into the mathematical abilities, reading comprehension, and scientific literacy of students from various countries.

Accurate parameter estimation is a fundamental requirement for applying IRT models. Commonly employed methods for estimating parameters in IRT include marginal maximum likelihood estimation via expectation maximization algorithm (MMLE-EM; Bock and Aitkin (1981); Baker and Kim (2004); Schilling and Bock (2005)) and Bayesian Markov chain Monte Carlo (MCMC) methods, such as the Metropolis–Hastings (MH; Metropolis et al. (1953); Hastings (1970)) and Gibbs algorithms (Béguin & Glas, 2001; Fox, 2010; Culpepper, 2016) . The MMLE-EM algorithm, despite being frequently utilized in various IRT models, possesses significant limitations. One primary concern is the requirement of these algorithms to perform high-dimensional numerical integration for parameter estimation. This is an inherently complex process that is computationally intensive. The issue is further compounded when the number of quadrature points needed for integration grows exponentially as the number of latent variables increases linearly (Jiang & Templin, 2019) . To circumvent these limitations, MCMC algorithms, based on posterior sampling, have been widely applied to estimate parameters in various complex IRT models. This approach significantly reduces the complexity of parameter estimation and improves computational efficiency.

However, parameter estimation in IRT models using MCMC methods is not without limitations either. MCMC methods often have difficulty with big data because of the following two reasons. Firstly, there are challenges related to data storage; when the dataset is too large, it becomes impractical for a single computer to store and process it. Secondly, computational time is a concern; applying MCMC methods in settings with big data can be highly time-consuming, as it typically involves numerous iterations, each of which requires several passes over the entire dataset (Xue & Liang, 2019; Srivastava & Xu, 2021; Shyamalkumar & Srivastava, 2022) .

Researchers have been actively addressing the challenges of using MCMC methods with big data. One approach involves approximating posterior analyses, such as variational Bayesian (Hoffman et al., 2013; Tan & Nott, 2014; Lee & Wand, 2016) , Laplace/Gaussian approximations (Rue et al., 2009) , and expectation propagation techniques (Vehtari et al., 2020). Although these methods can generate useful posterior mean estimates for big data, most of them often underestimate the posterior variance and lack theoretical guarantees for quantifying posterior uncertainty (Giordano, 2018) . The second approach focuses on approximating MCMC transition kernels with subsampling methods or easier-to-sample alternatives, without needing to pass through full a dataset (Korattikara et al., 2014; Alquier et al., 2016; Quiroz et al., 2019) . This approach, though promising, requires careful parameter tuning; only with correctly tuned parameters can the algorithm produce reliable posterior uncertainty estimates.

The third approach involves the divide-and-conquer (D&C) strategy, which comprises three stages: initially, the full dataset is partitioned into multiple subsets; subsequently, the sampling algorithm is implemented in parallel on all these subsets to derive posterior samples; and finally, the posterior samples collected from each subset are combined in a specific manner to conduct accurate posterior inference on the full dataset. Several techniques exist within stage 3, with the main difference being how the posterior draws from different subsets are merged (Minsker et al., 2014; Neiswanger et al., 2014; Scott et al., 2016) . Currently, the most advanced merging method is the Wasserstein posterior (WASP) method. Srivastava et al. (2015, 2018) first introduced WASP method, which amalgamates the posterior distribution of subsets by leveraging the Wasserstein barycenter—a concept representing the geometric center for probability measures (Agueh & Carlier, 2011) . Although the method offers extensive applicability and theoretical guarantees, its implementation, which requires computing the Wasserstein barycenter, can be highly complex and resource-intensive. In response, Li et al. (2017) proposed a computationally more straightforward posterior interval estimation algorithm (PIE) that estimates the WASP quantile by averaging the quantiles derived from each subset draw. However, PIE has been criticized for its effectiveness only with one-dimensional parameters.

More recently, the "double-parallel" Monte Carlo (DPMC) method was suggested by Xue and Liang (2019). This approach mainly relies on the mixture distribution derived from the common center of the subset draws to approximate the full data posterior. However, the assumption of asymptotic normality in DPMC can be challenging to validate in practical scenarios as it implies that the covariance matrices of the subset posterior distributions are identical. Álvarez et al. (2016) argued that the computation of the Wasserstein barycenter could be simplified by solving a fixed-point equation on a positive definite matrix space if the subsets posterior distribution pertains to the same location-scatter (LS) family of distributions. Building off this idea, Srivastava and Xu (2021) proposed the LS-WASP algorithm for distributed Bayesian inference in linear mixed-effects models. The LS-WASP algorithm not only preserves the computational simplicity of the PIE and DPMC algorithms, but also delivers asymptotic Monte Carlo and statistical guarantees. Subsequently, Shyamalkumar and Srivastava (2022) as well as Wang and Srivastava (2023) extended this algorithm to generalized linear models and hidden Markov models, respectively.

Despite extensive research into addressing the challenges of applying MCMC methods to mid-sized data, the difficulty of estimating the parameters of the IRT models using the MCMC algorithm remains in the presence of big data. This study presents the location-scatter Wasserstein posterior (LS-WASP) algorithm, a divide-and-conquer-based strategy, to tackle the big data challenge. This divide-and-conquer algorithm, anchored in the LS-WASP concept, transforms parameters drawn from subset posteriors into draws from WASP via a straightforward recentering and rescaling operation. Not only is this algorithm cutting-edge for scalable Bayesian inference applications, but it also presents a simple computation with asymptotic theoretical guarantees. This study applied the LS-WASP algorithm to the two-parameter logistic (2PL; Birnbaum, (1957)) model and the multidimensional two-parameter logistic (M2PL; Reckase (1972, 2009)) model,

confirming its utility and feasibility within the educational and psychometrics field. When dealing with large-scale response data, the LS-WASP algorithm not only accurately estimates each parameter of the IRT models but also substantially accelerates computational speed, supported by asymptotic Monte Carlo and statistical guarantees.

The organizations of this paper are as follows. Section 1 provides a brief introduction to the 2PL and M2PL models. Section 2 provides a detailed description of the preliminary preparations and the specific implementation process of the LS-WASP algorithm within the IRT framework. Section 3 presents the specific theoretical assumptions and the asymptotic statistical properties, and Monte Carlo guarantees under these assumptions. Section 4 presents the simulation studies of the LS-WASP algorithm in the 2PL and M2PL models. In Sect. 5, two examples are provided to show the application of the LS-WASP algorithm in IRT models. Finally, Sect. 6 concludes the paper with a brief summary and discussion.

1. Models

1.1. Unidimensional Two-Parameter Logistic Model

The unidimensional IRT model establishes a basic structure for illustrating the interaction between individuals and items by postulating a single latent trait dimension. Presume a test is composed of J binary response items, with each assessing a unidimensional latent trait, $\boldsymbol{\theta}$. Let $\boldsymbol{y} = \left[y_{ij}\right]_{n\times J}$ be an $n \times J$ matrix representing the responses of n examinees to J items, where $y_{ij} = 1\left(y_{ij} = 0\right)$ if the ith examinee answers the jth item correctly (wrong), for $i = 1, ..., n$ and $j = 1, ..., J$. The correct response probability of the two-parameter logistic (2PL) model is expressed as follows:

$$P\left(y_{ij} = 1 \left| \theta_i, a_j, b_j \right.\right) = \frac{\exp\left\{a_j(\theta_i - b_j)\right\}}{1 + \exp\left\{a_j(\theta_i - b_j)\right\}}, \tag{1}$$

where θ_i is the ability parameter of the examinee, for $i = 1, \cdots, n$; a_j is the discrimination parameter for the item; and b_j is the difficulty parameter for the item, for $j = 1, \cdots, J$.

1.2. Multidimensional Two-Parameter Logistic Model

Multidimensional item response theory (MIRT) models (Ackerman, 1996; Béguin & Glas, 2001) , as an extension of unidimensional IRT models, are extensively applied to illustrate the relationships between test items and multiple latent traits in psychological and educational assessments (Reckase, 2009) . Subsequently, we present the multidimensional two-parameter logistic (M2PL) model, a prevalent model within the domain of MIRT. Consider a test comprising of J items, which is designed to assess Q latant traits among n examinees. Denote the observed responses to the J items for all n examinees as $\boldsymbol{y} = \left[y_{ij}\right]_{n\times J}$. Here, $y_{ij} = 1$ indicates that examinee i answers item j correctly, while $y_{ij} = 0$ indicates a wrong response. For each examinee i $(i = 1, ..., n)$, let $\boldsymbol{\theta}_i = \left(\theta_{i1}, ..., \theta_{iQ}\right)^T$ denote the latent traits being measured, with Q dimensions. The correct response probability of the M2PL model is expressed as follows:

$$P(y_{ij} = 1 \left| \boldsymbol{\theta}_i, \boldsymbol{a}_j, b_j \right.) = \frac{\exp\left(\boldsymbol{a}_j^T\boldsymbol{\theta}_i - b_j\right)}{1 + \exp\left(\boldsymbol{a}_j^T\boldsymbol{\theta}_i - b_j\right)}, \tag{2}$$

where the discrimination and difficulty parameters are $\boldsymbol{a}_j = \left(a_{j1}, ..., a_{jQ}\right)^T$ and b_j, respectively. $a_{jq} \neq 0$ indicates that item j is associated with latent trait q.

Due to the overparametrization that allows for the rotation of discrimination parameters $(a_{j1}, ..., a_{jQ})$, it is necessary to address the model identification in advance. Usually, two approaches are employed to ensure model identification. The first approach constrains all ability parameters to follow a multivariate normal distribution with the mean of zero vector and covariance matrix of an identity matrix, respectively, and introducing the constrains $a_{jq} = 0$, for $j = 1, \cdots, Q-1, q = j+1, \cdots, Q$. The second approach constrains the item parameters by setting Q item parameters b_j to zero. Furthermore, for each $j = 1, \cdots, Q$ and $q = 1, \cdots, Q$, constraints are imposed on a_{jq} such that it equals to 1 if $j = q$, and it is set to zero if $j \neq q$. Meanwhile, the ability parameters are freely estimated. Both approaches effectively guarantee a unique solution for parameter estimation, as emphasized by Béguin & Glas (2001). Similarly, factor analysis models often constrain factor loadings to achieve model identifiability. Please see chapter 5 of Skrondal and Rabe-Hesketh (2004) for details. In this paper, we adopt the second approach, constraining item parameters.

2. The Location-Scatter Wasserstein Posterior (LS-WASP) Algorithm with a Divide-and-Conquer-Based Strategy

2.1. Preliminaries

This section introduces fundamental concepts, definitions, and theorems used in the LS-WASP algorithm, which serves as the foundation for the theoretical properties of the LS-WASP algorithm in Sect. 3. Next, several concepts and definitions related to the Wasserstein space are introduced. The Wasserstein space of order 2 on $\mathbb{R}^p$ is defined as:

$$\mathcal{P}_2(\mathbb{R}^p) = \left\{ \nu \in \mathcal{P}(\mathbb{R}^p) : \int \|x\|_2^2 \nu\,(dx) < \infty \right\},$$

where $\mathcal{P}(\mathbb{R}^p)$ is the set of Borel probability measures on $\mathbb{R}^p$ and $\|\cdot\|_2$ represents the Euclidean metric. The L_2-Wasserstein distance given by $(\mu, \nu) \in \mathcal{P}_2(\mathbb{R}^p)$ can be defined as:

$$W_2(\mu, \nu) = \left(\inf \left\{ \int \|x - y\|_2^2 d\pi(x, y),\ \pi \in \mathcal{P}_2(\mathbb{R}^p \times \mathbb{R}^p) \text{ with marginals } \mu, \nu \right\} \right)^{\frac{1}{2}}.$$

Definition 1. (Wasserstein barycenter) Given $\nu_1, \cdots, \nu_K$ in $\mathcal{P}_2(\mathbb{R}^p)$ and denoting w_k as the weight tied to ν_k, their Wasserstein barycenter can be expressed as:

$$\widetilde{\nu} = \arg \min_{\nu \in \mathcal{P}_2(\mathbb{R}^p)} \sum_{k=1}^{K} \frac{w_k}{2} W_2^2(\nu, \nu_k). \tag{3}$$

Note that $\sum_{k=1}^{K} w_k = 1$ and all $w_1, \cdots, w_K > 0$.

The Wasserstein barycenter exhibits important properties that render it a useful tool across different domains. First, the barycenter provides an effective measure of central tendency, representing a compromise between the multiple probability measures. Second, according to Agueh and Carlier (2011), the Wasserstein barycenter exists uniquely and optimally, as it minimizes the

sum of the squared Wasserstein distances to the individual measures. This implies that it provides the most accurate approximation under the Wasserstein metric. In practice, these properties allow the Wasserstein barycenter to aggregate complex, multi-modal distributions, thereby playing a critical role in applications such as optimal transport, machine learning, and computer vision, to name a few. When applying this idea to IRT model estimation, ν_k is the kth subset posterior, and $\widetilde{\nu}$ is the Wasserstein posterior (WASP) which replaces the full data posterior for inference. Nonetheless, the computation of the Wasserstein barycenter poses a significant challenge, prompting ongoing research into effective algorithms and computational strategies. Álvarez-Esteban et al. (2016) argued that when $\nu_1, \cdots, \nu_K$ belong to the same location-scatter family, the computation of the Wasserstein barycenter can be simplified to solving a fixed-point equation in the space of positive definite matrices. The definition of the location-scatter family is as follows.

Definition 2. (Location-scatter family) Consider a random vector $\boldsymbol{U}$ that conforms to the probability law $G \in \mathcal{P}_2(\mathbb{R}^p)$, such that its expectation $E(\boldsymbol{U})$ is zero and its variance $\mathrm{Var}(\boldsymbol{U})$ equals the identity matrix of dimensions $p \times p$. Let $\mathcal{L}$ denote the probability distribution of a random variable W, and $\mathcal{M}_+^{p\times p}$ represent the collection of $p \times p$ positive definite matrices. We define the family $\mathcal{F}(G) = \{\mathcal{L}(\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{U} + \boldsymbol{\mu}) : \boldsymbol{\Sigma} \in \mathcal{M}_+^{p\times p}, \boldsymbol{\mu} \in \mathbb{R}^p\}$, composed of probability laws formed by positive definite affine transformations derived from G, as a location-scatter family. Here, $\boldsymbol{\Sigma}^{\frac{1}{2}}$ denotes the symmetric square root of $\boldsymbol{\Sigma}$.

The location-scatter family offers a flexible and efficient method to generate a plethora of probability measures from a base measure. By adjusting the location parameter $\boldsymbol{\mu}$ and the positive definite matrix $\boldsymbol{\Sigma}$, we can effectively shift and scale the original measure, generating a comprehensive family of measures. This flexibility finds widespread use in statistical modeling and analysis. For instance, in multivariate analysis, the location-scatter family allows for the modeling of diverse and complex data distributions. Moreover, in machine learning, location-scatter families can help construct flexible probabilistic models that can adapt to the specific characteristics of the data. For additional details, please refer to the study by Álvarez-Esteban et al. (2016). Next, Theorem 1 provides a simplified process of the Wasserstein barycenter under the assumption of location-scatter family.

Theorem 1. *Suppose $\nu_1, \cdots, \nu_K \in \mathcal{F}(G)$ for a certain G, where $\boldsymbol{\mu}_k$ and $\boldsymbol{B}_k$ denote the mean vector and covariance matrix of $\nu_k (k = 1, \cdots, K)$, respectively. Under general conditions, the Wasserstein barycenter of $\nu_1, \cdots, \nu_K$ with weights $w_1, \cdots, w_K$, indicated as $\widetilde{\nu}$, is also an element of $\mathcal{F}(G)$. Its mean vector $\widetilde{\boldsymbol{\mu}}$ is the weighted average of $\boldsymbol{\mu}_k$ values, i.e., $\widetilde{\boldsymbol{\mu}} = \sum_{k=1}^K w_k \boldsymbol{\mu}_k$, and the covariance matrix $\widetilde{\boldsymbol{B}}$ corresponds to the fixed-point of the sequence $\{\boldsymbol{\Delta}_t\}_{t=0}^{\infty}$, which can be expressed as follows:*

$$\boldsymbol{\Delta}_{t+1} = \boldsymbol{\Delta}_t^{-\frac{1}{2}} \left\{ \sum_{k=1}^K w_k \left(\boldsymbol{\Delta}_t^{\frac{1}{2}} \boldsymbol{B}_k \boldsymbol{\Delta}_t^{\frac{1}{2}} \right)^{\frac{1}{2}} \right\}^2 \boldsymbol{\Delta}_t^{-\frac{1}{2}}, t = 0, 1, 2, \ldots, +\infty, \tag{4}$$

where $\boldsymbol{\Delta}_0$ is set as the unit array.

Theorem 1 reveals a fundamental and appealing property of the location-scatter family. When the probability measures $\nu_1, \cdots, \nu_K$ belong to the same location-scatter family $\mathcal{F}(G)$, their Wasserstein barycenter, denoted by $\widetilde{\nu}$, also belongs to the same family under general assumptions. This powerful result provides a direct and efficient computational route to determine the Wasserstein barycenter of a collection of measures within a location-scatter family, bypassing the need for solving the potentially complex Wasserstein barycenter problem directly. The theorem

also provides explicit analytical forms for the mean vector $\widetilde{\boldsymbol{\mu}}$ and covariance matrix $\widetilde{\boldsymbol{B}}$ of the barycenter $\widetilde{\nu}$. Specifically, the mean vector $\widetilde{\boldsymbol{\mu}}$ is the weighted average of the mean vectors $\boldsymbol{\mu}_k$ of the individual measures ν_k, while the covariance matrix $\widetilde{\boldsymbol{B}}$ is given by the fixed-point of the iteratively defined sequence $\{\boldsymbol{\Delta}_t\}_{t=0}^{\infty}$. Please refer to Álvarez-Esteban et al. (2016) for the proof of Theorem 1.

2.2. *Implementation of the LS-WASP Algorithm within the IRT Framework*

Consider an item response dataset comprising n examinees and J items from a large-scale testing, the detailed stages of the LS-WASP algorithm to approximate the posterior of the 2PL model are as follows:

Stage 1: Partitioning of the Full Data

Divide the full data set $\boldsymbol{y}$ into K disjoint subsets $\boldsymbol{y}^{(1)}, \cdots, \boldsymbol{y}^{(K)}$ through random subsampling. Specifically, let s_k be the number of examinees in subset k, ensuring that $\sum_{k=1}^{K} s_k = n$, where n is the total number of examinees. The 2PL model for subset k is given by:

$$p_{ij}^{(k)} = P\left(y_{ij}^{(k)} = 1 \left| \theta_i^{(k)}, a_j^{(k)}, b_j^{(k)} \right.\right) = \frac{\exp\left\{a_j^{(k)}\left(\theta_i^{(k)} - b_j^{(k)}\right)\right\}}{1 + \exp\left\{a_j^{(k)}\left(\theta_i^{(k)} - b_j^{(k)}\right)\right\}}. \tag{5}$$

Here, $y_{ij}^{(k)}$ represents the response data for subset k. $p_{ij}^{(k)}$ is the probability that the ith examinee in subset k answers the jth item correctly. $\theta_i^{(k)}$ represents the ability parameter of the ith examinee in subset k. The parameters $a_j^{(k)}$ and $b_j^{(k)}$ refer to the discrimination and difficulty parameters for the jth item in subset k, respectively ($i = 1, \cdots, s_k$, $j = 1, \cdots, J$, $k = 1, \cdots, K$).

Note that we partition the examinees into different subsets, but the items remain the same across all subsets. This means that even though each subset has distinct individuals, they respond to the same items. In this case, our divide-and-conquer approach primarily focuses on the item parameters: discrimination and difficulty. The identifiability constraints in each subset are set to be consistent with those of the full dataset, so that the estimated item parameters from different subsets are put on the same scale.

Stage 2: Parallel Sampling of Each Subset using Modified Likelihood method

Denote the parameter of item j as $\boldsymbol{\eta}_j = (a_j, b_j)$. $p(\boldsymbol{\eta}_j)$ is the prior distribution of $\boldsymbol{\eta}_j$, and $\ell_k(\boldsymbol{\eta}_j)$ is the likelihood of $\boldsymbol{\eta}_j$ on subset k. Consequently, the conditional posterior of $\boldsymbol{\eta}_j$ on subset k is given by:

$$\pi(\boldsymbol{\eta}_j \left| \boldsymbol{y}^{(k)} \right.) = \frac{\left\{\ell_k(\boldsymbol{\eta}_j)\right\}^{n/s_k} p(\boldsymbol{\eta}_j)}{\int \{\ell_k(\boldsymbol{\eta}_j)\}^{n/s_k} p(\boldsymbol{\eta}_j) d\boldsymbol{\eta}_j}, \; k = 1, \cdots, K, \; j = 1, \cdots, J. \tag{6}$$

Here, the likelihood has been raised to the power of n/s_k. This adjustment, known as the stochastic approximation strategy, was proposed by Minsker et al. (2014; 2017) to ensure that the variance of the subset posterior roughly matches the variance of the full data posterior.

While the LS-WASP algorithm can be applied to any sampling technique in MCMC, this paper chooses the Gibbs sampling algorithm based on the Pólya-Gamma distribution (PG-Gibbs) (Jiang & Templin, 2019; Balamuta & Culpepper, 2022; Jimenez et al., 2023) for subset posterior sampling. The reasons for this choice are as follows. First, the M-H algorithm, though a common choice for Markov chain generation, introduces potential inefficiency because the acceptance rate of each MC draw is only between 20% and 40%. Additionally, applying the traditional Gibbs

sampling algorithm to logistic regression also presents considerable challenges. Specifically, this function, which forms the basis for logistic regression, leads to non-standard and highly intricate posterior distributions, which often lacks straightforward analytical forms. However, the PG-Gibbs sampler not only adeptly addresses this challenge by providing a full conditional posterior distribution that allows for comprehensive analytical manipulation, but also provides essential theoretical support such as ergodicity (Choi & Hobert, 2013; Polson et al., 2013) . Finally, Balamuta and Culpepper (2022) showcased that the application of the Pólya-gamma formulation to the logit link function outperforms the traditional probit link in computational speed. Notably, they also highlighted its exceptional proficiency when dealing with large-scale test data, making it a compelling choice for research involving substantial datasets. As a result, these combined factors substantially motivate our decision to utilize the PG-Gibbs algorithm.

Therefore, when the discrimination parameters and difficulty parameters are independent, the modified posterior density of subset k ($k = 1, \cdots, K$) based on PG-Gibbs algorithm for the discrimination parameter a_j and difficulty parameter b_j is given by:

$$
\begin{aligned}
&f\left(a_j^{(k)}|\boldsymbol{\theta}^{(k)}, \boldsymbol{b}^{(k)}, \boldsymbol{\omega}^{(k)}, \boldsymbol{y}^{(k)}\right) \\
&\propto f\left(a_j^{(k)}\right) \exp\left\{-\frac{n}{s_k} \cdot \frac{1}{2}\left[z_a^{(k)} - \left(\boldsymbol{\theta}^{(k)} - \mathbf{1}b_j^{(k)}\right) a_j^{(k)}\right]^T \boldsymbol{\Omega}_{ab}^{(k)} \left[z_a^{(k)} - \left(\boldsymbol{\theta}^{(k)} - \mathbf{1}b_j^{(k)}\right) a_j^{(k)}\right]\right\},
\end{aligned} \tag{7}
$$

$$
\begin{aligned}
&f\left(b_j^{(k)}|\boldsymbol{\theta}^{(k)}, \boldsymbol{a}^{(k)}, \boldsymbol{\omega}^{(k)}, \boldsymbol{y}^{(k)}\right) \\
&\propto f\left(b_j^{(k)}\right) \exp\left\{-\frac{n}{s_k} \cdot \frac{1}{2}\left(z_b^{(k)} + \mathbf{1}a_j^{(k)} b_j^{(k)}\right)^T \boldsymbol{\Omega}_{ab}^{(k)} \left(z_b^{(k)} + \mathbf{1}a_j^{(k)} b_j^{(k)}\right)\right\}.
\end{aligned} \tag{8}
$$

The prior distributions of $\theta_i^{(k)}, a_j^{(k)}$ and $b_j^{(k)}$ are assumed to follow $N(\mu_1^{(k)}, \sigma_1^{2(k)})$, $TN_{(0,+\infty)}(\mu_2^{(k)}, \sigma_2^{2(k)})$, and $N(\mu_3^{(k)}, \sigma_3^{2(k)})$, respectively. Subsequently, the basic steps for sampling a subset are as follows:

(1) Given $\boldsymbol{\theta}^{(k)}$, $\boldsymbol{a}^{(k)}$, and $\boldsymbol{b}^{(k)}$, draw augmented variable ω_{ij} from the $PG(1, a_j^{(k)}|\theta_i^{(k)} - b_j^{(k)}|)$ distribution;

(2) Given $\boldsymbol{\omega}^{(k)}$, $\boldsymbol{a}^{(k)}$, and $\boldsymbol{b}^{(k)}$, draw $\theta_i^{(k)}$ from a normal distribution $N(m_{\theta_i}^{(k)}, V_{\theta_i}^{(k)})$, where $m_{\theta_i}^{(k)} = V_{\theta_i}^{(k)}\left((\boldsymbol{a}^{(k)})^T \boldsymbol{\Omega}_\theta \boldsymbol{z}_\theta + \frac{\mu_1^{(k)}}{\sigma_1^{2(k)}}\right)$, $V_{\theta_i}^{(k)} = \left((\boldsymbol{a}^{(k)})^T \boldsymbol{\Omega}_\theta \boldsymbol{a}^{(k)} + \frac{1}{\sigma_1^{2(k)}}\right)^{-1}$,
$\boldsymbol{z}_\theta = \left(\frac{a_1^{(k)} b_1^{(k)} \omega_{i1} + \kappa_{i1}}{\omega_{i1}}, \cdots, \frac{a_J^{(k)} b_J^{(k)} \omega_{iJ} + \kappa_{iJ}}{\omega_{iJ}}\right)^T$, $\boldsymbol{\Omega}_\theta = \text{diag}(\omega_{i1}, \cdots, \omega_{iJ})$, and $\kappa_{ij} = y_{ij}^{(k)} - \frac{1}{2}$.

(3) Given $\boldsymbol{\omega}^{(k)}$, $\boldsymbol{\theta}^{(k)}$, and $\boldsymbol{b}^{(k)}$, draw $a_j^{(k)}$ from a truncated normal distribution $TN_{(0,+\infty)}\left(m_{a_j}^{(k)}, V_{a_j}^{(k)}\right)$, where $m_{a_j}^{(k)} = V_{a_j}^{(k)}\left(\frac{n}{s_k} \cdot (\boldsymbol{\theta}^{(k)} - \mathbf{1}b_j^{(k)})^T \boldsymbol{\Omega}_{ab}^{(k)} \boldsymbol{z}_a^{(k)} + \frac{\mu_2^{(k)}}{\sigma_2^{2(k)}}\right)$, $V_{a_j}^{(k)} = \left(\frac{n}{s_k} \cdot (\boldsymbol{\theta}^{(k)} - \mathbf{1}b_j^{(k)})^T \boldsymbol{\Omega}_{ab}^{(k)} (\boldsymbol{\theta}^{(k)} - \mathbf{1}b_j^{(k)}) + \frac{1}{\sigma_2^{2(k)}}\right)^{-1}$, $z_a^{(k)} = \left(\frac{\kappa_{1j}}{\omega_{1j}}, \cdots, \frac{\kappa_{s_k j}}{\omega_{s_k j}}\right)^T$, $\mathbf{1} = (1, \cdots, 1)^T_{s_k \times 1}$, and $\boldsymbol{\Omega}_{ab}^{(k)} = \text{diag}(\omega_{1j}, \cdots, \omega_{s_k j})$;

(4) Given $\boldsymbol{\omega}^{(k)}, \boldsymbol{\theta}^{(k)}$, and $\boldsymbol{a}^{(k)}$, draw $b_j^{(k)}$ from a normal distribution $N(m_{b_j}^{(k)}, V_{b_j}^{(k)})$, where $m_{b_j}^{(k)} = V_{b_j}^{(k)}\left(-\frac{n}{s_k} \cdot (\mathbf{1}a_j^{(k)})^T \boldsymbol{\Omega}_{ab}^{(k)} \boldsymbol{z}_b^{(k)} + \frac{\mu_3^{(k)}}{\sigma_3^{2(k)}}\right)$, $V_{b_j}^{(k)} = \left(\frac{n}{s_k} \cdot (\mathbf{1}a_j^{(k)})^T \boldsymbol{\Omega}_{ab}^{(k)} \mathbf{1}a_j^{(k)} + \frac{1}{\sigma_3^{2(k)}}\right)^{-1}$,
$\boldsymbol{z}_b^{(k)} = \left(\frac{\kappa_{1j} - a_j^{(k)} \theta_1^{(k)} \omega_{1j}}{\omega_{1j}}, \cdots, \frac{\kappa_{s_k j} - a_j^{(k)} \theta_{s_k}^{(k)} \omega_{s_k j}}{\omega_{s_k j}}\right)^T$, and $\boldsymbol{\Omega}_{ab}^{(k)} = \text{diag}\left(\omega_{1j}, \cdots, \omega_{s_k j}\right)$.

For more details of the PG-Gibbs algorithm, please see Jiang and Templine (2019). The online supplementary S1 provides details of the PG-Gibbs algorithm for full data.

Stage 3: Assembling and Integrating Sampled Parameters from Each Subset

Denote the item parameters of the j-th item be $\boldsymbol{\eta}_j = (a_j, b_j)$. Assume that $\pi_{(1)}, \cdots, \pi_{(K)}$ represent the posterior distributions of the K subsets of $\boldsymbol{\eta}_j$, respectively. Let $\boldsymbol{\eta}_j^{(k,1)}, \cdots, \boldsymbol{\eta}_j^{(k,M)}$ denote the samples from the kth subset posterior of $\boldsymbol{\eta}_j^{(k)}$, where M is the total number of post-burn-in iterations. In distributed Bayesian applications for IRT model, ν_k of Definition 1 is the kth subset posterior distribution $\pi_{(k)}$. Therefore, the Wasserstein barycenter of $\pi_{(1)}, \cdots, \pi_{(K)}$ is $\widetilde{\pi}$, which is also known as the Wasserstein posterior (WASP). We use WASP instead of the full data posterior for parameter inference. Let $\boldsymbol{\mu}_{\eta_j}^{(k)}$ and $\boldsymbol{\Sigma}_{\eta_j}^{(k)}$ represent the mean vector and covariance matrix of the kth subset posterior, respectively. From stage 2, we can obtain the Monte Carlo estimates of $\boldsymbol{\mu}_{\eta_j}^{(k)}$ and $\boldsymbol{\Sigma}_{\eta_j}^{(k)}$ as follows:

$$\widehat{\boldsymbol{\mu}}_{\eta_j}^{(k)} = \frac{1}{M}\sum_{m=1}^{M} \boldsymbol{\eta}_j^{(k,m)}, \tag{9}$$

$$\widehat{\boldsymbol{\Sigma}}_{\eta_j}^{(k)} = \frac{1}{M}\sum_{m=1}^{M} \left(\boldsymbol{\eta}_j^{(k,m)} - \widehat{\boldsymbol{\mu}}_{\eta_j}^{(k)}\right)\left(\boldsymbol{\eta}_j^{(k,m)} - \widehat{\boldsymbol{\mu}}_{\eta_j}^{(k)}\right)'. \tag{10}$$

Therefore, assuming $\pi_{(1)}, \cdots, \pi_{(K)}$ belong to the same location-scatter family, the estimates of the mean vector $\widetilde{\boldsymbol{\mu}}_{\eta_j}$ and covariance matrix $\widetilde{\boldsymbol{\Sigma}}_{\eta_j}$ of the WASP are obtained from Theorem 1 as follows:

$$\widehat{\widetilde{\boldsymbol{\mu}}}_{\eta_j} = \sum_{k=1}^{K} w_k \widehat{\boldsymbol{\mu}}_{\eta_j}^{(k)}, \tag{11}$$

$$\widehat{\widetilde{\boldsymbol{\Sigma}}}_{\eta_j} = \widehat{\boldsymbol{\Delta}}_{\infty}, \tag{12}$$

where

$$\widehat{\boldsymbol{\Delta}}_{t+1} = \widehat{\boldsymbol{\Delta}}_t^{-\frac{1}{2}} \left\{\sum_{k=1}^{K} w_k \left(\widehat{\boldsymbol{\Delta}}_t \widehat{\boldsymbol{\Sigma}}_{\eta_j}^{(k)}\right)^{\frac{1}{2}}\right\} \left\{\sum_{k=1}^{K} w_k \left(\widehat{\boldsymbol{\Delta}}_t \widehat{\boldsymbol{\Sigma}}_{\eta_j}^{(k)}\right)^{\frac{1}{2}}\right\}' \widehat{\boldsymbol{\Delta}}_t^{-\frac{1}{2}}, t = 0, 1, 2, \ldots, +\infty, \tag{13}$$

and we set $\widehat{\boldsymbol{\Delta}}_0 = \boldsymbol{I}$, where $\boldsymbol{I}$ represents the identity matrix. Note that Eq. (13) is the numerically stabilized version proposed by Srivastava and Xu (2021) to solve the rank deficiency problem of $\widehat{\boldsymbol{\Delta}}_t$ in Eq. (4). Thus, we use the fixed-point iterations in Eq. (13) instead of Eq. (4) for computing the covariance matrix $\widehat{\widetilde{\boldsymbol{\Sigma}}}_{\eta_j}$ and use the following convergence criterion: $|\text{tr}(\widehat{\boldsymbol{\Delta}}_{t+1} - \widehat{\boldsymbol{\Delta}}_t)| < 10^{-6}$. In this study, given the similarities in the sizes of the K subsets $s_1, \cdots, s_K$, we set the weights of $\pi_{(k)}$ to $w_k = \frac{1}{K}$. However, when the sizes $s_1, \cdots, s_K$ differ substantially, it would be more appropriate to set the weights as $w_k = s_k/(s_1 + \cdots + s_K)$. This approach assigns greater importance to subsets that contain more samples, thereby reflecting their greater contribution to the overall data set. Furthermore, adjusting the weights in this manner can ensure a more accurate representation of the data and potentially improve the robustness of the subsequent analysis.

According to the definition of the location-scatter family, we can obtain the WASP draws through the following steps for $k = 1, \cdots, K, m = 1, \cdots, M, j = 1, \cdots, J$:

1. Centralize and standardize the kth subset of posterior draws for item j to obtain $\widehat{\boldsymbol{u}}_j^{(k,m)} = (\widehat{\boldsymbol{\Sigma}}_{\eta_j}^{(k)})^{-\frac{1}{2}}(\boldsymbol{\eta}_j^{(k,m)} - \widehat{\boldsymbol{\mu}}_{\eta_j}^{(k)})$. Consequently, $\widehat{\boldsymbol{u}}_j^{(k,m)}$ has a mean vector of zero and a covariance of the identity matrix.

2. Recentralize and restandardize $\widehat{\boldsymbol{u}}_j^{(k,m)}$ to derive the WASP draws as $\widehat{\widehat{\boldsymbol{\eta}}}_j^{(k,m)} = \widehat{\widehat{\boldsymbol{\mu}}}_{\eta_j} + (\widehat{\widehat{\boldsymbol{\Sigma}}}_{\eta_j})^{\frac{1}{2}}\widehat{\boldsymbol{u}}_j^{(k,m)}$.

In this study, we assume that the prior distributions of the discrimination parameters and difficulty parameters are independent. Assuming independence simplifies the model estimation process, making the estimation and inference of parameters more direct and computationally manageable. This is particularly important in cases of large parameter spaces or data volumes to ensure the feasibility and efficiency of model estimation. The independent prior is also supported by many literature (e.g., Wang et al. 2013, 2018). Furthermore, there is often a lack of prior knowledge about the specific relationship between discrimination and difficulty parameters in practical applications, with no theoretical or empirical basis to support significant correlation between these parameters (van der Linden 2007, please see Table 1 in their paper), making the use of independent prior distributions a prudent choice. Therefore, in this case, we can compute the mean and variance of each parameter separately. Take the discrimination parameter a_j as an example. The Monte Carlo estimates of $\mu_{a_j}^{(k)}$ and $\sigma_{a_j}^{(k)}$, and the estimates of the mean and variance of the WASP are shown below:

$$\widehat{\mu}_{a_j}^{(k)} = \frac{1}{M}\sum_{m=1}^{M} a_j^{(k,m)}, \quad \widehat{\sigma}_{a_j}^{(k)} = \frac{1}{M}\sum_{m=1}^{M}(a_j^{(k,m)} - \widehat{\mu}_{a_j}^{(k)})^2. \tag{14}$$

$$\widehat{\widehat{\mu}}_{a_j} = \frac{1}{K}\sum_{k=1}^{K}\widehat{\mu}_{a_j}^{(k)}, \quad \widehat{\widehat{\sigma}}_{a_j} = \widehat{\Delta}_{\infty} = \left(\frac{1}{K}\sum_{k=1}^{K}\left(\widehat{\sigma}_{a_j}^{(k)}\right)^{\frac{1}{2}}\right)^2. \tag{15}$$

Note that since the variance $\widehat{\widehat{\sigma}}_{a_j}$ is a unidimensional variable (as is the case for the difficulty parameter), the sequences $\widehat{\Delta}_t$ actually reach the fixed point at $t = 1$, which simplifies the model estimation process and improves computational efficiency. The process of deriving WASP draws for the difficulty parameter $\boldsymbol{b}$ follows a similar approach as the discrimination parameter $\boldsymbol{a}$. The details of stage 3 of the LS-WASP algorithm for the independent item parameters are given in Algorithm 1.

Remark 1. By the WASP draws $\widehat{\widehat{a}}_j^{(k,m)}$ of parameter a_j, we can derive the WASP estimator of a_j as follows:

$$\begin{aligned}\widehat{\widehat{a}}_j &= \frac{1}{KM}\sum_{k=1}^{K}\sum_{m=1}^{M}\widehat{\widehat{a}}_j^{(k,m)} = \frac{1}{KM}\sum_{k=1}^{K}\sum_{m=1}^{M}\left(\widehat{\widehat{\mu}}_{a_j} + (\widehat{\widehat{\sigma}}_{a_j})^{\frac{1}{2}}\widehat{u}_j^{(k,m)}\right)\\ &= \widehat{\widehat{\mu}}_{a_j} + (\widehat{\widehat{\sigma}}_{a_j})^{\frac{1}{2}}\frac{1}{KM}\sum_{k=1}^{K}\sum_{m=1}^{M}(\widehat{\sigma}_{a_j}^{(k)})^{-\frac{1}{2}}(a_j^{(k,m)} - \widehat{\mu}_{a_j}^{(k)}) = \widehat{\widehat{\mu}}_{a_j}.\end{aligned} \tag{16}$$

According to Eq. (16), we note that the mean of WASP is essentially the average of the posterior means across all subsets. Note that our algorithm offers theoretical guarantee (please refer to Sect. 3 for details). In fact, after partitioning the data, we can only make posterior inferences on parameters of subsets, rather than on the full dataset's posterior distribution. However, the proposed LS-WASP algorithm enables us to combine all posterior distributions from all subsets to form a new distribution, i.e., the WASP distribution. This WASP effectively approximates the

Algorithm 1
The details for the stage 3 of LS-WASP Algorithm

1. Input:
   - Samples for $\theta_i^{(k,m)}, a_j^{(k,m)}$, and $b_j^{(k,m)}$ are drawn from the kth subset posterior at the stage 2, where $i = 1, \cdots, s_k$, $j = 1, \cdots, J$, $k = 1, \cdots, K$ and $m = 1, \cdots, M$.
   - Calculate Monte Carlo estimates of mean vectors and covariance matrices for the subset posterior distributions and the WASPs of $\widehat{\boldsymbol{\mu}}_a^{(k)} = (\widehat{\mu}_{a_1}^{(k)}, \cdots, \widehat{\mu}_{a_J}^{(k)})^T$, $\widehat{\boldsymbol{\Sigma}}_a^{(k)} = \text{diag}(\widehat{\sigma}_{a_1}^{(k)}, \cdots, \widehat{\sigma}_{a_J}^{(k)})$, $\widehat{\widehat{\boldsymbol{\mu}}}_a = (\widehat{\widehat{\mu}}_{a_1}, \cdots, \widehat{\widehat{\mu}}_{a_J})^T$, $\widehat{\widehat{\boldsymbol{\Sigma}}}_a = \text{diag}(\widehat{\widehat{\sigma}}_{a_1}, \cdots, \widehat{\widehat{\sigma}}_{a_J})$, $\widehat{\boldsymbol{\mu}}_b^{(k)} = (\widehat{\mu}_{b_1}^{(k)}, \cdots, \widehat{\mu}_{b_J}^{(k)})^T$, $\widehat{\boldsymbol{\Sigma}}_b^{(k)} = \text{diag}(\widehat{\sigma}_{b_1}^{(k)}, \cdots, \widehat{\sigma}_{b_J}^{(k)})$, $\widehat{\widehat{\boldsymbol{\mu}}}_b = (\widehat{\widehat{\mu}}_{b_1}, \cdots, \widehat{\widehat{\mu}}_{b_J})^T$ and $\widehat{\widehat{\boldsymbol{\Sigma}}}_b = \text{diag}(\widehat{\widehat{\sigma}}_{b_1}, \cdots, \widehat{\widehat{\sigma}}_{b_J})$ according to Eqs. (14), (15).
2. Do:
   - Centralize and standardize the kth subset posterior draws of item parameters as

$$\widehat{\boldsymbol{u}}^{(k,m)} = (\widehat{\boldsymbol{\Sigma}}_a^{(k)})^{-\frac{1}{2}}\left(\boldsymbol{a}^{(k,m)} - \widehat{\boldsymbol{\mu}}_a^{(k)}\right),$$
$$\widehat{\boldsymbol{q}}^{(k,m)} = (\widehat{\boldsymbol{\Sigma}}_b^{(k)})^{-\frac{1}{2}}\left(\boldsymbol{b}^{(k,m)} - \widehat{\boldsymbol{\mu}}_b^{(k)}\right),$$

   where $\widehat{\boldsymbol{u}}^{(k,m)} = (\widehat{u}_1^{(k,m)}, \cdots, \widehat{u}_J^{(k,m)})^T$, $\boldsymbol{a}^{(k,m)} = (a_1^{(k,m)}, \cdots, a_J^{(k,m)})^T$, $\widehat{\boldsymbol{q}}^{(k,m)} = (\widehat{q}_1^{(k,m)}, \cdots, \widehat{q}_J^{(k,m)})^T$ and $\boldsymbol{b}^{(k,m)} = (b_1^{(k,m)}, \cdots, b_J^{(k,m)})^T$.
   - Recentralize and restandardize $\widehat{\boldsymbol{u}}^{(k,m)}$ and $\widehat{\boldsymbol{q}}^{(k,m)}$ to obtain the WASP draws as

$$\widehat{\widehat{\boldsymbol{a}}}^{(k,m)} = \widehat{\widehat{\boldsymbol{\mu}}}_a + (\widehat{\widehat{\boldsymbol{\Sigma}}}_a)^{\frac{1}{2}}\widehat{\boldsymbol{u}}^{(k,m)},$$
$$\widehat{\widehat{\boldsymbol{b}}}^{(k,m)} = \widehat{\widehat{\boldsymbol{\mu}}}_b + (\widehat{\widehat{\boldsymbol{\Sigma}}}_b)^{\frac{1}{2}}\widehat{\boldsymbol{q}}^{(k,m)},$$

   where $\widehat{\widehat{\boldsymbol{a}}}^{(k,m)} = (\widehat{\widehat{a}}_1^{(k,m)}, \cdots, \widehat{\widehat{a}}_J^{(k,m)})^T$ and $\widehat{\widehat{\boldsymbol{b}}}^{(k,m)} = (\widehat{\widehat{b}}_1^{(k,m)}, \cdots, \widehat{\widehat{b}}_J^{(k,m)})^T$.
3. Return:
   - Assemble and integrate sampled parameters from each subset, $\widehat{\widehat{\boldsymbol{a}}}^{(1,1)}, \cdots, \widehat{\widehat{\boldsymbol{a}}}^{(1,M)}, \cdots, \widehat{\widehat{\boldsymbol{a}}}^{(K,M)}$ and $\widehat{\widehat{\boldsymbol{b}}}^{(1,1)}, \cdots, \widehat{\widehat{\boldsymbol{b}}}^{(1,M)}, \cdots, \widehat{\widehat{\boldsymbol{b}}}^{(K,M)}$, as an approximation of the WASP.

full data posterior distribution, making it a suitable alternative for diverse analyses and posterior evaluations instead of the full data posterior.

Remark 2. Note that our algorithm exhibits high computational efficiency. It adopts a "divide-and-conquer" strategy to split larger datasets into smaller, more manageable subsets, thereby effectively reducing the computational and storage burden. Specifically, when the dataset is divided into K subsets, the effective sample size for each MCMC operation becomes $\frac{1}{K}$ of the total sample size; thus, the computation time for each subset approximately becomes $\frac{1}{K}$ of the total runtime. Leveraging the capabilities of parallel processing, as well as the simple and efficient merging operation in the third step of the algorithm, the overall runtime of the algorithm is essentially consistent with the computation time of a single subset, which is significantly less than traditional MCMC method that processes the entire dataset at once. Intuitively, the runtime of our algorithm is with a fraction of the runtime of the full dataset under MCMC method; the more subsets the data is divided into, the shorter the runtime. This point is validated in the subsequent simulation studies.

3. Theoretical Properties

The Bayesian inference using the LS-WASP algorithm involves two types of errors. The first is the statistical error, originating from using the WASP approximation posterior instead of the full data posterior. This deviation from the original full data posterior introduces discrepancies, consequently generating statistical errors. The second type is the Monte Carlo error, which arises from using Monte Carlo estimates. The inherent randomness in Monte Carlo simulations can introduce variations and errors, leading to the overall error in the LS-WASP algorithm.

In this section, we discuss the theoretical properties of the LS-WASP algorithm. First, we present five essential assumptions forming the foundation of our theoretical framework. Each assumption is outlined and justified for its significance to the overall methodology and its role in quantifying the errors previously mentioned. Second, we further explore the two sources of errors, gaining insights into their origins and impacts on the LS-WASP algorithm's results. Moreover, it opens up opportunities for potential improvements to reduce these errors and enhance the algorithm's precision and robustness.

3.1. Assumptions

For the notation simplicity in the following text, we define $\boldsymbol{\eta}$ as either the $\boldsymbol{a}$ parameter or the $\boldsymbol{b}$ parameter, even though $\boldsymbol{\eta}$ can also represent the set of item parameters, i.e., $\boldsymbol{\eta} = \{\boldsymbol{a}, \boldsymbol{b}\}$. Below are the key assumptions for establishing the theoretical foundation of the LS-WASP algorithm.

Assumption 1. (Independent and identically distributed data): We assume that the observations $\boldsymbol{y}_1, \cdots, \boldsymbol{y}_n$ are independent and identically distributed (i.i.d), where $\boldsymbol{y}_i = (y_{i1}, \cdots, y_{iJ})^T$ $(i = 1, \cdots, n)$.

Assumption 2. (Location-scatter family): There exist probability distributions G specifying location-scatter families $\mathcal{F}(G)$. Both the full data posterior distribution π of $\boldsymbol{\eta}$ and subset posterior distributions $\pi_{(k)}(k = 1, \cdots, K)$ of $\boldsymbol{\eta}$ belong to $\mathcal{F}(G)$ with $P^n_{\boldsymbol{\eta}^*}$-probability 1, where $P^n_{\boldsymbol{\eta}^*}$ is the joint probability distribution of the full data, $\boldsymbol{\eta}^* \in \Theta$ represents the true value of parameters $\boldsymbol{\eta}$, and $\Theta \subset \mathbb{R}^J$ is the parameter space.

Assumption 3. (Regularity conditions for Laplace approximation): We denote an open ball of radius δ centered at $\boldsymbol{\eta}$ as $B_\delta(\boldsymbol{\eta})$. Suppose $h_n(\boldsymbol{\eta}) = -\frac{1}{n}\sum_{i=1}^n \log f(\boldsymbol{y}_i|\boldsymbol{\eta})$ is a six times continuously differentiable real function on Θ (Kass et al., 1990) . Let $\widehat{\boldsymbol{\eta}}_n$ be the maximum likelihood estimate (MLE) of $\boldsymbol{\eta}$, and $D^2 h_n(\boldsymbol{\eta})$ be the Hessian matrix of $h_n(\boldsymbol{\eta})$ at $\boldsymbol{\eta}$. There exist positive numbers ϵ, N, and ξ and an integer n_0 such that for all $n \geq n_0$:

(a) For every $\boldsymbol{\eta} \in B_\epsilon(\widehat{\boldsymbol{\eta}}_n)$ and all $1 \leq j_1, \cdots, j_d \leq J$ with $1 \leq d \leq 6$, the absolute value of the dth derivative of the log likelihood $|\partial_{j_1,\cdots,j_d} h_n(\boldsymbol{\eta})| < N$;
(b) The determinant of $D^2 h_n(\widehat{\boldsymbol{\eta}}_n)$ is greater than ξ, that is, $|D^2 h_n(\widehat{\boldsymbol{\eta}}_n)| > \xi$;
(c) For every δ satisfying $0 < \delta < \epsilon$, $B_\delta(\widehat{\boldsymbol{\eta}}_n) \subseteq \Theta$ and the upper limit as n goes to infinity of the supremum of $h_n(\widehat{\boldsymbol{\eta}}_n) - h_n(\boldsymbol{\eta})$ over $\boldsymbol{\eta} \in \Theta - B_\delta(\widehat{\boldsymbol{\eta}}_n)$ is less than zero, that is,

$$\limsup_{n\to\infty} \sup_{\boldsymbol{\eta}\in\Theta - B_\delta(\widehat{\boldsymbol{\eta}}_n)} \{h_n(\widehat{\boldsymbol{\eta}}_n) - h_n(\boldsymbol{\eta})\} < 0.$$

Assumption 4. (Disjoint subsets of equal size): The subsets are disjoint, and the number of subsets K and the sample size of subsets s satisfy the conditions $K = o(n^{\frac{1}{2}})$, and $Ks = n$, where $s = s_1 = \ldots = s_K$.

Assumption 5. (Number of iterations): Let M be the number of iterations. M satisfies the conditions $n = o(\sqrt{M})$ and $\widehat{\boldsymbol{\mu}}_{(k)} - \boldsymbol{\mu}_{(k)} = O_p(M^{-\frac{1}{2}})$ and $\widehat{\boldsymbol{\Sigma}}_{(k)} - \boldsymbol{\Sigma}_{(k)} = O_p(M^{-\frac{1}{2}})$ in P_k-probability$(k = 1, \cdots, K)$, where $\widehat{\boldsymbol{\mu}}_{(k)} = (\widehat{\mu}_1^{(k)}, \cdots, \widehat{\mu}_J^{(k)})^T$ with $\widehat{\mu}_j^{(k)} = \frac{1}{M}\sum_{m=1}^{M} \eta_j^{(k,m)}$ represents the sample mean of the sequence of $\boldsymbol{\eta}_{(k)}$ values and $\widehat{\boldsymbol{\Sigma}}_{(k)} = \text{diag}(\widehat{\sigma}_1^{(k)}, \cdots, \widehat{\sigma}_J^{(k)})$ with $\widehat{\sigma}_j^{(k)} = \frac{1}{M}\sum_{m=1}^{M}(\eta_j^{(k,m)} - \widehat{\mu}_j^{(k)})^2$ denotes the sample covariance of the same sequence. P_k is the probability measure of the posterior draw $\{\boldsymbol{\eta}_{(k)}^{(m)}, m = 1, \cdots, M\}$ on subset k.

Assumptions 1–4 are standard in the divide-and-conquer strategy for Bayesian inference, which is typically applied to the exponential family (Xue & Liang, 2019; Shyamalkumar & Srivastava, 2022) . In other words, if P_{η^*} belongs to the exponential family, Assumptions 1–4 are valid. Assumption 2 can be used to derive the mean and covariance matrix of the WASP posterior of item parameters. In fact, in IRT models, given the data $\boldsymbol{y}$, the full data and subset posterior distributions of item parameters do not belong to the same location-scatter family. However, in many divide-and-conquer studies, it is common to approximate the subset posterior distribution using the same location-scatter family (Srivastava & Xu, 2021) , making Assumption 2 is reasonable. By combining the LS-WASP algorithm with subset parameter sampling, we can obtain a true WASP approximation under Assumption 2, using the location-scatter family-based subset posterior distribution. Moreover, when the sample size of each subset is large, we can demonstrate that the Bernstein–von Mises theorem holds, indicating that substituting the full data posterior distribution with a true WASP approximation is justifiable. Assumption 3 is a standard requirement for the posterior expansion based on the Laplace method in the quantization of the approximation (Kass et al., 1990) . Although Assumption 3 is commonly applied in various statistical models (e.g., Xue and Liang 2019; Shyamalkumar and Srivastava 2022), our study exclusively focuses on IRT models. Therefore, we introduce a proposition to verify Assumption 3 (a) and (b) holds specifically for 2PL and M2PL models. This proposition necessitates that the discrimination parameter $\boldsymbol{a}$, the difficulty parameter $\boldsymbol{b}$, and the ability parameter $\boldsymbol{\theta}$ are all bounded. The proposition and its detailed proof are provided in S5.1 of the online supplement. Assumption 3 (c) is a reasonable assumption to ensure the uniqueness of MLE $\widehat{\boldsymbol{\eta}}_n$ and has been studied in the IRT literature (e.g., San Martín 2016). Assumption 4 requires a uniform subset sample size for simplifying the analysis, although the LS-WASP algorithm can still be applied when subset sizes vary. In this study, to simplify the calculations, we assume $K = o(n^{\frac{1}{2}})$, which follows the assumptions made by Xue and Liang (2019). Assumption 5 is valid when the subset sampling scheme exhibits geometric ergodicity. With geometric ergodicity, the Markov chain (which the subsets are based on) mixes quickly, reducing the autocorrelation and, consequently, the Monte Carlo error. This, in turn, allows for an accurate approximation of the expectation, contributing to the overall effectiveness and precision of the LS-WASP algorithm. In this study, we adopt the PG-Gibbs algorithm as the subset sampling scheme. Choi and Hobert (2013) have demonstrated that the Pólya-Gamma data augmentation strategy exhibits uniform ergodicity.

3.2. Statistical Error

Note that the LS-WASP algorithm introduces a source of error for posterior inference on parameter $\boldsymbol{\eta}$ when it uses $\widetilde{\pi}$ to infer parameter $\boldsymbol{\eta}$, rather than the true π. The reason it is called statistical error is due to the absence of any Monte Carlo approximation. Another source of error is the Monte Carlo error, which is described in detail in subsection 3.3.

To accurately quantify the statistical error, the following specific corollary is necessary. This corollary provides a mathematical framework for precise computation of the statistical error. The magnitude and impact of the statistical error on the results may vary depending on factors such

as the complexity of the model, the nature of the data, and the specifics of the algorithmic implementation. Therefore, proper management of the statistical error is crucial for the performance and reliability of the LS-WASP algorithm.

Corollary 1. *Consider two probability measures A and B in* $\mathcal{P}_2(\mathbb{R}^p)$. *Let* $\boldsymbol{m}_A$ *and* $\boldsymbol{m}_B$ *represent the means of A and B, and let* $\boldsymbol{\Sigma}_A$ *and* $\boldsymbol{\Sigma}_B$ *denote the covariance matrices of A and B, respectively. Under the assumption that* $\boldsymbol{\Sigma}_A$ *is nonsingular, the following inequality holds*

$$W_2^2(A, B) \geq \|\boldsymbol{m}_A - \boldsymbol{m}_B\|_2^2 + tr(\boldsymbol{\Sigma}_A + \boldsymbol{\Sigma}_B - 2(\boldsymbol{\Sigma}_A^{\frac{1}{2}}\boldsymbol{\Sigma}_B\boldsymbol{\Sigma}_A^{\frac{1}{2}})^{\frac{1}{2}}), \tag{17}$$

where $\|\cdot\|_2$ *denotes the Euclidean distance and* $tr(\cdot)$ *represents the trace of a matrix. The equality holds when the map* $T(x) = (\boldsymbol{m}_B - \boldsymbol{m}_A) + \boldsymbol{D}x$ *transports A to B, where* $\boldsymbol{D} = \boldsymbol{\Sigma}_A^{-\frac{1}{2}}(\boldsymbol{\Sigma}_A^{\frac{1}{2}}\boldsymbol{\Sigma}_B\boldsymbol{\Sigma}_A^{\frac{1}{2}})^{\frac{1}{2}}\boldsymbol{\Sigma}_A^{-\frac{1}{2}}$. *Here,* $\boldsymbol{D}$ *is a positive semi-definite matrix.*

For the detailed proof of this corollary, please refer to the proof of Theorem 2.3 in Álvarez-Esteban et al. (2016). Next, we introduce the notational conventions used to state the theoretical results. We define the full data posterior, the kth subset posterior, and the WASP approximation of the interested parameter $\boldsymbol{\eta}$, as π, $\pi_{(k)}$ and $\widetilde{\pi}$, for $k = 1, \cdots, K$, respectively. Let $\boldsymbol{\mu}$, $\boldsymbol{\mu}_{(k)}$, $\widetilde{\boldsymbol{\mu}}$, and $\boldsymbol{\Sigma}$, $\boldsymbol{\Sigma}_{(k)}$, $\widetilde{\boldsymbol{\Sigma}}$ represent the means and covariance matrices of π, $\pi_{(k)}$, and $\widetilde{\pi}$, respectively. Therefore, we define the statistical error as $W_2^2(\pi, \widetilde{\pi})$. Under Corollary 1, we have

$$W_2^2(\pi, \widetilde{\pi}) = \|\boldsymbol{\mu} - \widetilde{\boldsymbol{\mu}}\|_2^2 + \mathrm{tr}(\boldsymbol{\Sigma} + \widetilde{\boldsymbol{\Sigma}} - 2(\widetilde{\boldsymbol{\Sigma}}^{\frac{1}{2}}\boldsymbol{\Sigma}\widetilde{\boldsymbol{\Sigma}}^{\frac{1}{2}})^{\frac{1}{2}}). \tag{18}$$

Subsequently, we establish asymptotic statistical guarantees for the LS-WASP algorithm, provided that Assumptions 1–4 are satisfied, as detailed below.

Theorem 2. *If* Assumptions 1–4 *hold, as* $n, s \to \infty$,

$$W_2^2(\pi, \widetilde{\pi}) = O_p(s^{-2}) + O_p(n^{-\frac{3}{2}}), \tag{19}$$

where π *denotes the full data posterior;* $\widetilde{\pi}$ *denotes the WASP approximate posterior. And* n *and* s *represent the sample sizes of the full data and subset, respectively.*

The detailed proof of Theorem 2 is provided in online supplement S5. According to Theorem 2, as the sample size, n, of the full data increases, the sample size, s, of each subset will also increase under a specific number of partitions, leading to a smaller statistical error that tends toward zero. Therefore, to minimize the statistical error, it is essential to maximize the sample size whenever possible. In cases where the sample size of the full data is limited, one can effectively reduce the statistical error by controlling the number of samples in each subset through appropriate partitioning.

3.3. Monte Carlo Error

The Monte Carlo error is also introduced due to the approximation of the Wasserstein barycenter of subset posterior distributions, $\widetilde{\pi}$, by an empirical measure $\widehat{\widetilde{\pi}}$. This empirical measure, while being a necessary approximation for computational tractability, is not the true Wasserstein barycenter of subset posterior distributions. Therefore, the distance between these two measures can be seen as the 'Monte Carlo error.' However, the true Wasserstein barycenter and its empirical approximation are both random measures, which presents additional challenges in the analysis of

this type of error. To overcome this, we consider coupled versions of these measures, which evolve from the same randomness and are therefore dependent. The Wasserstein distance between these coupled measures provides a controlled environment in which we can investigate the magnitude of the Monte Carlo error. Next, we will provide a detailed introduction of this coupling procedure and outline the key steps to establish the asymptotic order of the Monte Carlo error.

Let M denote the number of iterations after burn-in. We assume $\boldsymbol{\eta}$, $\boldsymbol{\eta}_{(k)}$, and $\widetilde{\boldsymbol{\eta}}$ to follow the distributions of π, $\pi_{(k)}$, and $\widetilde{\pi}$, respectively. Provided that Assumption 2 is satisfied, the mth iterative draw of parameter $\boldsymbol{\eta}_{(k)}$ in subset k can be obtained as $\boldsymbol{\eta}_{(k)}^{(m)} = \boldsymbol{\mu}_{(k)} + \boldsymbol{\Sigma}_{(k)}^{\frac{1}{2}} \boldsymbol{\xi}_{(k)}^{(m)}$ for $m = 1, \cdots, M$, as stated in Definition 2. According to this definition, $\boldsymbol{\xi}_{(k)}^{(m)}$ $(k = 1, \cdots, K, m = 1, \cdots, M)$ represent KM independent draws from distribution G, which has a zero mean and a covariance matrix of identity matrix. Moreover, Definition 2 allows us to derive the posterior draw of the WASP approximation posterior $\widetilde{\pi}$ as follows:

$$\widetilde{\boldsymbol{\eta}} = \widetilde{\boldsymbol{\mu}} + \widetilde{\boldsymbol{\Sigma}}^{\frac{1}{2}} \boldsymbol{\xi}_{(k)}^{(m)} = \widetilde{\boldsymbol{\mu}} + \widetilde{\boldsymbol{\Sigma}}^{\frac{1}{2}} \boldsymbol{\Sigma}_{(k)}^{-\frac{1}{2}} \left(\boldsymbol{\eta}_{(k)}^{(m)} - \boldsymbol{\mu}_{(k)} \right), \quad k = 1, \cdots, K, m = 1, \cdots, M, \tag{20}$$

where $\widetilde{\boldsymbol{\mu}} = \frac{1}{K} \sum_{k=1}^{K} \boldsymbol{\mu}_{(k)}$ and $\widetilde{\boldsymbol{\Sigma}} = \boldsymbol{\Delta}_{\infty}$. As $\widetilde{\boldsymbol{\mu}}$, $\boldsymbol{\mu}_{(k)}$, $\widetilde{\boldsymbol{\Sigma}}$, and $\boldsymbol{\Sigma}_{(k)}$ are unknown, they are replaced by their Monte Carlo estimates $\widehat{\widetilde{\boldsymbol{\mu}}}$, $\widehat{\boldsymbol{\mu}}_{(k)}$, $\widehat{\widetilde{\boldsymbol{\Sigma}}}$, and $\widehat{\boldsymbol{\Sigma}}_{(k)}$ in the LS-WASP algorithm, yielding $\widehat{\widetilde{\boldsymbol{\eta}}}$, the Monte Carlo estimate of $\widetilde{\boldsymbol{\eta}}$. Let $\widehat{\widetilde{\pi}}$ represent the empirical measure of the WASP approximate posterior acquired from Monte Carlo estimation in the LS-WASP algorithm. Consequently, $\widehat{\widetilde{\boldsymbol{\eta}}}$ follows the distribution of $\widehat{\widetilde{\pi}}$, and the empirical measure $\widehat{\widetilde{\pi}}$ is comprised of observations

$$\widehat{\widetilde{\boldsymbol{\eta}}} = \widehat{\widetilde{\boldsymbol{\mu}}} + \widehat{\widetilde{\boldsymbol{\Sigma}}}^{\frac{1}{2}} \widehat{\boldsymbol{\Sigma}}_{(k)}^{-\frac{1}{2}} \left(\boldsymbol{\mu}_{(k)} - \widehat{\boldsymbol{\mu}}_{(k)} \right) + \widehat{\widetilde{\boldsymbol{\Sigma}}}^{\frac{1}{2}} \widehat{\boldsymbol{\Sigma}}_{(k)}^{-\frac{1}{2}} \boldsymbol{\Sigma}_{(k)}^{\frac{1}{2}} \boldsymbol{\xi}_{(k)}^{(m)}, \quad k = 1, \cdots, K, m = 1, \cdots, M, \tag{21}$$

where $\widehat{\widetilde{\boldsymbol{\mu}}} = \frac{1}{K} \sum_{k=1}^{K} \widehat{\boldsymbol{\mu}}_{(k)}$ and $\widehat{\widetilde{\boldsymbol{\Sigma}}} = \widehat{\boldsymbol{\Delta}}_{\infty}$. In light of Corollary 1, we characterize the Monte Carlo error as $W_2(\widetilde{\pi}, \widehat{\widetilde{\pi}})$. Consequently, the following theorem is presented.

Theorem 3. *Under* Assumptions 1–5*, when* $n, M \to \infty$*,*

$$W_2^2(\widetilde{\pi}, \widehat{\widetilde{\pi}}) = O_p(M^{-1}) + o_p(n^{-1}). \tag{22}$$

where n represents the sample size of the full data and M denotes the number of iterations after burn-in.

The detailed proof of Theorem 3 is provided in online supplement S5.

4. Simulation Studies

Two simulation studies were conducted to assess the effectiveness and feasibility of the proposed LS-WASP algorithm. Simulation studies 1 and 2 performed the LS-WASP algorithm on the 2PL model and M2PL model, respectively. The sample sizes, test lengths, and the number of subsets under different partitions were varied to evaluate the robustness of the LS-WASP algorithm across diverse scenarios. The MC chain length was set to be 10,000, with the first 5000 iterations

as burn-in. The bias and RMSE of item parameters and ability parameters were computed to assess the accuracy of the parameter estimates:

$$\text{Bias}(\boldsymbol{\eta}) = \frac{1}{R}\sum_{r=1}^{R}(\hat{\boldsymbol{\eta}}_r - \boldsymbol{\eta}^*), \quad \text{RMSE}(\boldsymbol{\eta}) = \sqrt{\frac{1}{R}\sum_{r=1}^{R}(\hat{\boldsymbol{\eta}}_r - \boldsymbol{\eta}^*)^2}, \tag{23}$$

where R represents the number of replications, $\hat{\boldsymbol{\eta}}_r$ denotes the parameter estimate of the rth replication, and $\boldsymbol{\eta}^*$ indicates the true value of the parameter. Each simulation condition was conducted $R = 25$ replications. Additionally, we calculated the running time for each replication to assess the computational efficiency of our LS-WASP algorithm. The maximum value of each replication running time was displayed as the final running time. The computation implementation was executed on an AMD EPYC 7542 32-Core Processor (2.90 GHz) with 1.5 TB RAM on a Windows Server 2019 Standard operating system. Note that the parallel sampling of subsets is conducted using different cores of the computer. Specifically, after partitioning the full data into K subsets, the K cores of the computer are used for parallel sampling, with one core used for sampling one subset.

4.1. Study 1: Performance of LS-WASP Algorithm for the 2PL Model

4.1.1. Design The purpose of study 1 is to examine the performance of the LS-WASP algorithm on the 2PL model. Three factors were manipulated to vary different simulation conditions: (1) The number of examinees, i.e., n=10,000, 20,000, 50,000; (2) The number of items, i.e., J=20, 40; (3) The number of subsets, i.e., K=1, 2, 5, 10, 20, which is different number of partitions of the full data. Varying different levels of these three factors produced 30 simulation conditions. Each simulation condition was replicated 25 times.

Note that $K = 1$ corresponds to the scenario where no partitioning is performed, and the full data set is used for analysis. Thus, we employed the PG-Gibbs Sampler for the full data set hereafter, rather than using the LS-WASP algorithm which requires data partitioning.

4.1.2. Data Generation The response data of 2PL model are generated from Eq. (1). The ability parameters were generated from a normal distribution $N(0, 1)$. The discrimination parameters and difficulty parameters were sampled from a lognormal distribution $\log N(0.3, 0.2)$ and a normal distribution $N(0, 1)$, respectively. The manner of data generation is consistent with that of Jiang and Templin Jiang and Templin (2019). To ensure the model identification, the priors for θ_i, a_j, and b_j were set to a normal distribution $N(0, 1)$, a truncated normal distribution $TN_{(0,+\infty)}(0, 10)$, and a normal distribution $N(0, 10)$, respectively. Our selection of these distributions and priors is based on their theoretical properties and widespread use in the existing literature. However, it is important to note that the LS-WASP algorithm can work with other distributions and priors as well.

4.1.3. Results Table 1 shows the bias and RMSE of parameter estimates and running time across different conditions. We observe a slight increase in the bias and RMSE for the ability parameters $\boldsymbol{\theta}$ and difficulty parameters $\boldsymbol{b}$ as the number of subsets increases. However, this increase is minimal, indicating that the number of subsets has a negligible effect on $\boldsymbol{\theta}$ and $\boldsymbol{b}$. Furthermore, Figures S-1 and S-2 in online supplement S2 display the average bias and RMSE of the discrimination parameter $\boldsymbol{a}$ across different number of subsets K. From Table 1 and Figures S-1 and S-2, it can be observed that both the average bias and RMSE for parameter $\boldsymbol{a}$ increase with K, but this increasing trend becomes more slower as the sample size n increases. This is because an increase in K leads to a smaller subset sample size, resulting in larger bias and RMSE. However, as the

TABLE 1.
Bias and RMSE of parameter estimates and running time across different conditions in simulation study 1.

n	J	K	Bias			RMSE			Time (h)
			$\boldsymbol{\theta}$	$\boldsymbol{a}$	$\boldsymbol{b}$	$\boldsymbol{\theta}$	$\boldsymbol{a}$	$\boldsymbol{b}$	
10000	20	1	−0.0034	−0.0001	−0.0023	0.3821	0.0374	0.0233	0.393
		2	−0.0031	0.0025	−0.0019	0.3822	0.0376	0.0233	0.170
		5	−0.0030	0.0084	−0.0020	0.3828	0.0389	0.0233	0.071
		10	−0.0024	0.0181	−0.0014	0.3837	0.0424	0.0234	0.036
		20	−0.0016	0.0374	−0.0005	0.3864	0.0557	0.0231	0.019
20000	20	1	−0.0054	0.0009	−0.0021	0.3886	0.0272	0.0192	0.795
		2	−0.0051	0.0015	−0.0019	0.3886	0.0271	0.0191	0.384
		5	−0.0051	0.0043	−0.0018	0.3893	0.0279	0.0190	0.137
		10	−0.0054	0.0089	−0.0020	0.3897	0.0296	0.0192	0.070
		20	−0.0058	0.0181	−0.0026	0.3909	0.0341	0.0192	0.037
50000	20	1	0.0027	0.0014	0.0024	0.3905	0.0179	0.0120	1.985
		2	0.0029	0.0018	0.0026	0.3905	0.0179	0.0120	0.973
		5	0.0028	0.0028	0.0026	0.3908	0.0180	0.0121	0.392
		10	0.0028	0.0046	0.0026	0.3911	0.0182	0.0120	0.174
		20	0.0029	0.0084	0.0029	0.3918	0.0200	0.0121	0.089
10000	40	1	−0.0111	0.0048	−0.0110	0.2808	0.0372	0.0252	0.974
		2	−0.0108	0.0063	−0.0107	0.2813	0.0377	0.0249	0.507
		5	−0.0108	0.0140	−0.0109	0.2821	0.0402	0.0252	0.184
		10	−0.0110	0.0255	−0.0109	0.2847	0.0466	0.0254	0.093
		20	−0.0110	0.0495	−0.0109	0.2885	0.0647	0.0261	0.049
20000	40	1	−0.0053	−0.0015	−0.0036	0.2899	0.0254	0.0185	1.587
		2	−0.0053	−0.0006	−0.0037	0.2900	0.0254	0.0186	0.771
		5	−0.0053	0.0024	−0.0038	0.2909	0.0254	0.0185	0.316
		10	−0.0051	0.0083	−0.0038	0.2913	0.0274	0.0184	0.138
		20	−0.0055	0.0192	−0.0042	0.2927	0.0334	0.0185	0.071
50000	40	1	0.0029	0.0003	0.0015	0.2912	0.0190	0.0144	4.003
		2	0.0027	0.0007	0.0013	0.2912	0.0190	0.0143	1.938
		5	0.0027	0.0020	0.0011	0.2916	0.0191	0.0144	0.790
		10	0.0028	0.0044	0.0012	0.2920	0.0196	0.0144	0.399
		20	0.0027	0.0087	0.0011	0.2928	0.0215	0.0144	0.176

Bias and RMSE denote the average bias and RMSE for the parameter estimates. $\boldsymbol{a}$ represents all discrimination parameters, $\boldsymbol{b}$ represents all difficulty parameters, and $\boldsymbol{\theta}$ denotes all ability parameters.

full dataset sample size increases, the subset sample size also increases for the same K, thus slowing down the increase in estimation error. Additionally, we observe that as the sample size increases, the overall bias and RMSE decrease, indicating that the larger the sample size, the more accurate the parameter estimates become. It is noteworthy that when $K = 20$, the bias of $\boldsymbol{a}$ is twice as high as that when $K = 10$. Nevertheless, even with small sample sizes when $K = 20$, the RMSE results remain satisfactory. Note that the selection of the optimal subset sample size will be influenced by the complexity of the model. Specifically, more complex model generally requires larger sample size per subset to ensure adequate accuracy of parameter estimation. Therefore, it is suggested that the sample size within each subset should at least meet the minimum requirement necessary for accurate model parameter estimation. For instance, an accurate estimation of a 2PL model requires at least 500 individuals (König et al., 2020) , a 3PL model needs at least 1000 (de la Torre & Hong, 2010; De Ayala, 2013) , and a 4PL model necessitates at least 4000

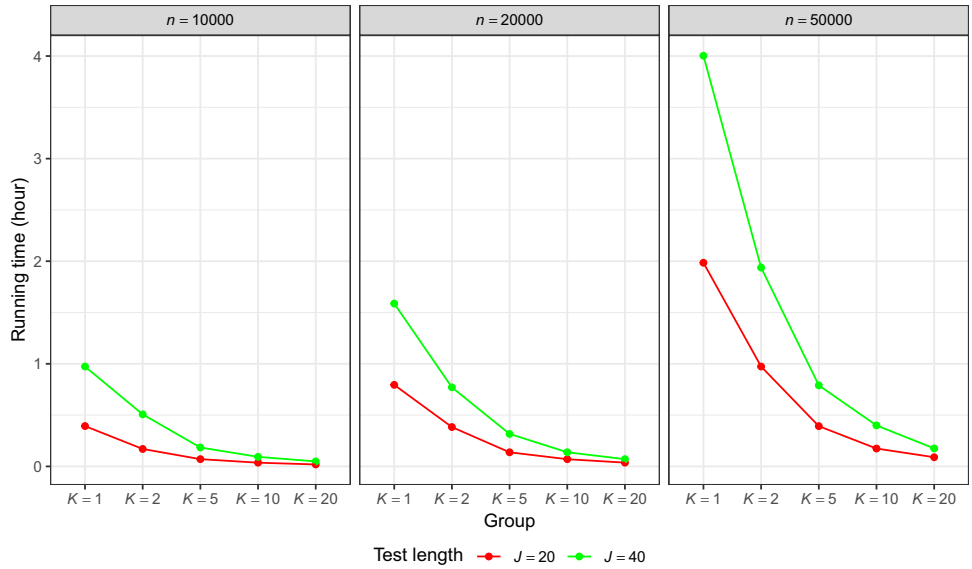


FIGURE 1.
Running times under different subset conditions in simulation study 1. Note that 'Group' indicates the number of subsets.

individuals (Cuhadar, 2022) . This strategy ensures that the accuracy of estimates within each subset is maintained, thereby ensuring the overall accuracy of the parameter estimation.

In addition, from the running time in Table 1, the running time for the full data is significantly larger than that of subset using the LS-WASP algorithm. For example, the running time of the LS-WASP algorithm is halved when partitioning the full data into two subsets, and it becomes a tenth when partitioned into ten subsets, and so forth. To provide a detailed analysis of running time, Fig. 1 presents a graphical comparison of the running time across conditions. We observe that the running time increases with more examinees and items, while decreasing with a larger number of subsets. When the full data are partitioned into $K = 10$ and $K = 20$ subsets, the running time remains below 0.5 h for each condition. Thus, our algorithm holds a distinct advantage in computational speed.

To delve deeper into these differences, we compared the bias and RMSE of each item. Figure 2 shows the results of $J = 20$. Please see online supplement S2 for results of $J = 40$. Clearly, as the number of examinees increases, the bias and RMSE lines of item parameters converge and overlap more. In essence, as examinee numbers grow, LS-WASP algorithm's item parameter estimation aligns more closely with full data, diminishing partitioning-related discrepancies. This is due to a larger sample size in each subset with a constant partition number, reducing partitioning effects. Moreover, overall estimation accuracy improves with more examinees, highlighting our algorithm's advantages as examinee numbers increase.

In cases of small samples split into numerous subsets, the LS-WASP algorithm may slightly compromise discriminant parameter accuracy. However, it still produces accurate difficulty parameter estimates and offers a clear running time advantage. For larger samples, regardless of subset numbers, our method ensures accurate parameter estimation and maintains fast computation.

4.2. Study 2: Performance of LS-WASP Algorithm for the M2PL Model

4.2.1. Design The aim of study 2 is to investigate the performance of the LS-WASP algorithm for the M2PL model. 10,000 examinees were considered to be consistent with the sample size

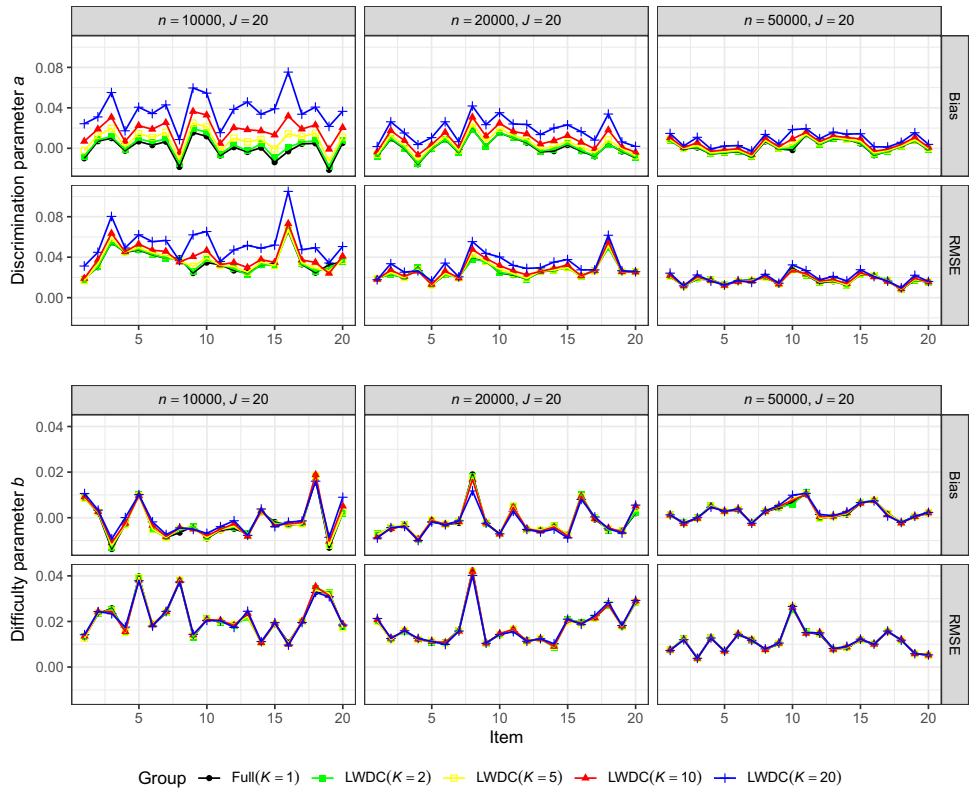


FIGURE 2.
Bias and RMSE of each item parameter estimate across various sample sizes with $J = 20$ in simulation study 1. Note that 'Group' indicates the number of subsets.

in empirical example 2 for the multidimensional case. Three factors were manipulated to vary different simulation conditions: (1) The number of items, i.e., J=20, 40; (2) The number of subsets, i.e., K=1, 2, 5, 10. In this case, we excluded $K = 20$ because the M2PL model required a minimum of 1000 examinees per subset to ensure precise parameter estimation when $n = 10,000$; (3) The number of dimensions of latent traits, i.e., Q=2, 3. In total, 16 simulation conditions were conducted, and each condition was performed 25 replications.

4.2.2. Data Generation We generated item response data for the M2PL model from Eq. (2). The ability parameters $\boldsymbol{\theta}_i$ for each examinee i were sampled from a multivariate normal distribution $N(\mathbf{0}, \boldsymbol{I}_Q)$, where $\mathbf{0}$ is a Q-dimensional vector with all elements being 0 and $\boldsymbol{I}_Q$ is a $Q \times Q$ dimensional unit matrix. The discrimination parameter a_{jq} and difficulty parameter b_j for each item j were generated from $\log N(0.3, 0.2)$ and $N(0, 1)$ (for $q = 1, \cdots, Q$), respectively. The model identification condition of the M2PL model follows the constrains given in subsection 1.2 (Béguin & Glas, 2001) . The prior settings for the M2PL are same with those for the 2PL model, i.e., the prior for $\boldsymbol{\theta}_i$, a_{jq}, and b_j follows $N(\mathbf{0}, \boldsymbol{I}_Q)$, $TN_{(0,+\infty)}(0, 10)$, and $N(0, 10)$, respectively.

4.2.3. Results Table 2 presents bias and RMSE of parameter estimates and running time across different conditions for the M2PL model. The trend of parameter estimates in the M2PL model mirrors that of the 2PL model; that is, as the number of subsets increases, the bias and RMSE for each parameter estimate slightly grow, but the increase remains minimal. Furthermore, the bias

TABLE 2.
Bias and RMSE of parameter estimates and running time across different conditions in simulation study 2.

Q	J	K	Bias			RMSE			Time (h)
			$\boldsymbol{\theta}$	$\boldsymbol{a}$	$\boldsymbol{b}$	$\boldsymbol{\theta}$	$\boldsymbol{a}$	$\boldsymbol{b}$	
2	20	1	0.0007	−0.0059	0.0021	0.6437	0.0536	0.0289	2.761
		2	0.0005	−0.0024	0.0014	0.6448	0.0533	0.0287	1.680
		5	0.0007	0.0073	0.0009	0.6489	0.0557	0.0308	0.701
		10	0.0005	0.0242	−0.0018	0.6559	0.0637	0.0353	0.345
	40	1	0.0039	0.0104	0.0084	0.5868	0.0509	0.0310	3.860
		2	0.0038	0.0143	0.0080	0.5887	0.0495	0.0310	2.088
		5	0.0040	0.0262	0.0089	0.5944	0.0562	0.0320	0.905
		10	0.0041	0.0466	0.0099	0.6042	0.0695	0.0345	0.463
3	20	1	−0.0006	0.0125	−0.0010	0.7381	0.0725	0.0302	3.693
		2	−0.0003	0.0167	−0.0005	0.7401	0.0719	0.0306	1.903
		5	−0.0005	0.0283	−0.0004	0.7462	0.0785	0.0334	0.852
		10	−0.0003	0.0500	−0.0005	0.7567	0.0943	0.0409	0.448
	40	1	−0.0015	−0.0104	−0.0065	0.7016	0.0630	0.0299	4.156
		2	−0.0014	−0.0054	−0.0063	0.7041	0.0647	0.0304	2.181
		5	−0.0016	0.0092	−0.0084	0.7121	0.0649	0.0321	0.982
		10	−0.0017	0.0353	−0.0099	0.7261	0.0767	0.0368	0.494

Bias and RMSE denote the average bias and RMSE for the parameters. $\boldsymbol{a}$ represents all discrimination parameters across Q dimensions, $\boldsymbol{b}$ represents all difficulty parameters, and $\boldsymbol{\theta}$ denotes all ability parameters across Q dimensions.

and RMSE for each parameter expand as latent trait dimensions increase. However, the difference in results between the LS-WASP algorithm and the full data set is not significant, indicating its applicability to the M2PL model. Figure 3 depicts the running time under different subsets for the M2PL model. With more subsets, LS-WASP algorithm speeds up significantly. Additionally, computation time slightly increases as latent trait dimensions grow. Figure 4 shows the bias and RMSE of each item parameter estimates for $J = 20$. Please see online supplement S2 for results of $J = 40$. As the number of subsets and latent trait dimensions increase, there is a noticeable increase in bias and RMSE for each item. The bias and RMSE for the discrimination parameter $\boldsymbol{a}$ stay mostly within 0.1, given that true values of $\boldsymbol{a}$ are generally between 1 and 2.5. For the difficulty parameter $\boldsymbol{b}$, where true values are primarily around 0 with a few exceeding 1, the bias and RMSE are kept within 0.08. Therefore, the LS-WASP algorithm ensures accurate parameter estimation and notably improves computational efficiency in the M2PL model.

5. Empirical Examples

Two examples from PISA computer-based mathematics data were analyzed using our proposed LS-WASP algorithm. These two examples showcased the application of the LS-WASP algorithm for the 2PL model and M2PL model, respectively. Due to page limit, please see online supplement S4 for details on the empirical example of M2PL model.

5.1. Data Description

The first data set is from PISA 2018 computer-based cognitive mathematics test. We selected 9 scored items, i.e., CM447Q01S, CM273Q01S, CM408Q01S, CM420Q01S, CM446Q01S, CM559Q01S, CM828Q03S, CM464Q01S, and CM800Q01S. After excluding students with missing responses, the remaining sample size was 80,352. Further, 20,412 students with Not Reached

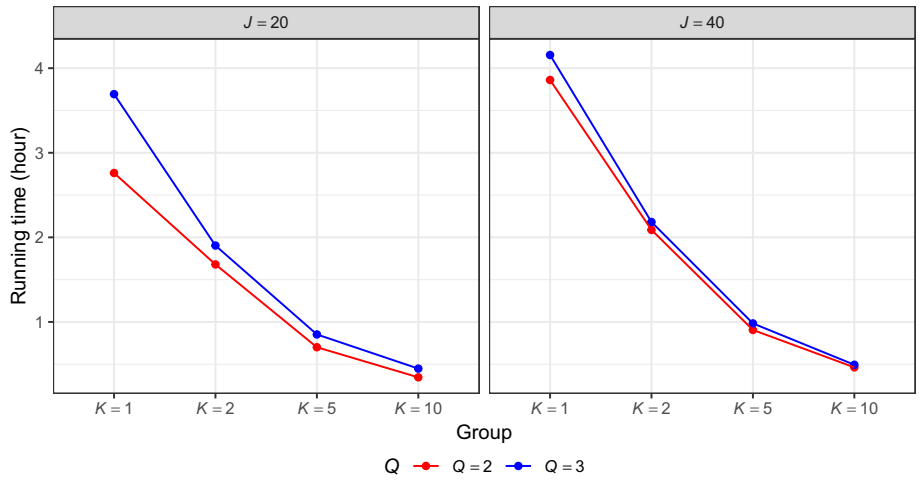


FIGURE 3.
Running times under different subset conditions in simulation study 2. Note that 'Group' indicates the number of subsets.

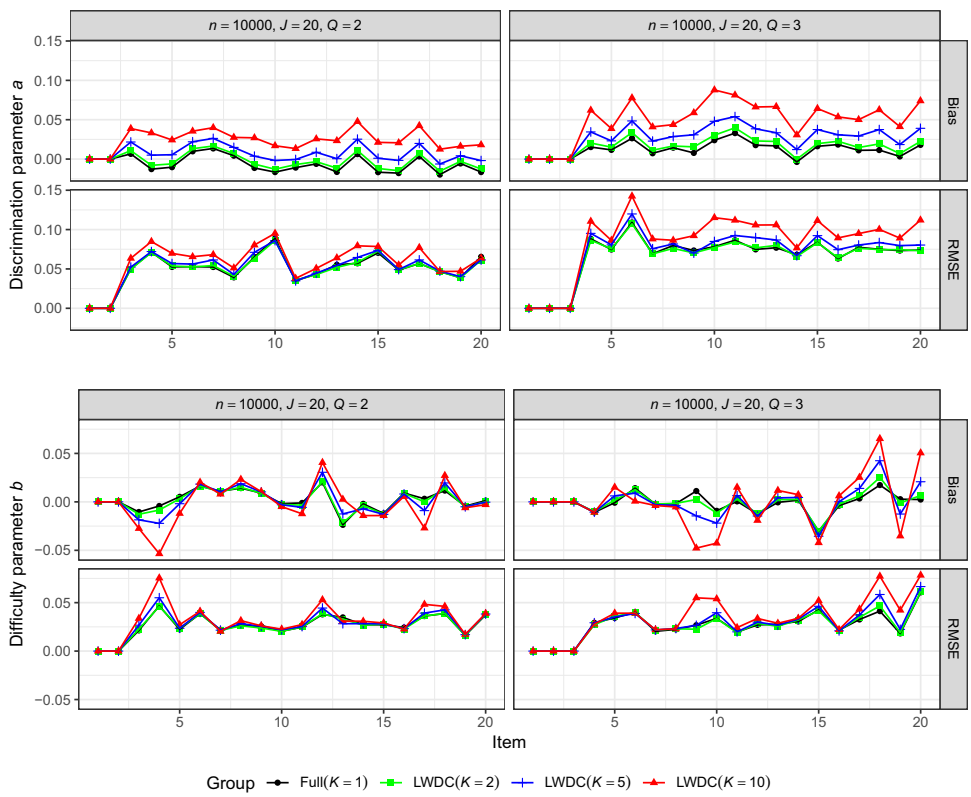


FIGURE 4.
Bias and RMSE of each item parameter estimate across various latent trait dimensions with $J = 20$ in simulation study 2. Note that 'Group' indicates the number of subsets.

(original code 6), Not Applicable (original code 7), Invalid (original code 8), or No Response (original code 9) were removed. Thus, the final sample size consisted of 59,940 students.

5.2. Purposes and Designs

We calculated the standard deviations (SD) and 95% highest posterior density (HPD) intervals for each item parameter estimates. Our proposed LS-WASP algorithm primarily addresses the significant challenges in parameter estimation arising from large sample sizes in large-scale testing. The primary idea of our algorithm is as follows. At stage 1 of our algorithm, the full dataset is partitioned, i.e., the "persons" are grouped. At stage 2, the same set of items are estimated across different data subsets. At stage 3, the item parameters obtained from the stage 2 are integrated to yield the final item parameter estimates. Since the LS-WASP algorithm is grounded in the full Bayesian framework, the ability parameters are sampled from PG-Gibbs algorithm of stage 2. Here, we also present the results of ability parameter estimation. The running time was given to illustrate the time efficiency of the LS-WASP algorithm.

With the sample size exceeding 50,000, we divided the data into subsets of $K = 2, 5, 10, 20$ to align with simulation studies. We employed the 2PL model to fit the data to adhere to the PISA operational analysis plan. We set the MCMC iterations to 10,000, with the initial 5000 iterations as burn-in. Unlike simulation studies, real data do not have predefined "true" parameter values. To verify the effectiveness of the LS-WASP algorithm, we computed parameter estimates from the full data using the Pólya-Gamma Gibbs sampler for comparison.

5.3. Results

Table 3 shows the expected a posteriori (EAP) values and SD values of item parameters under different subsets. As the number of subsets increases, sample sizes within each subset decrease, potentially impacting parameter accuracy for inference. Conversely, fewer subsets yield larger sample sizes, theoretically obtaining more accurate parameter estimates. From Table 3, with a maximum of 20 subsets, the EAP of $\boldsymbol{a}$ increases by about 0.01 compared to full data. Similarly, for parameter $\boldsymbol{b}$, EAP differences primarily stay within 0.0007, with only a few exceeding 0.001. Therefore, for these two parameters, estimations remain consistent between the full data and conditions with 20 subsets.

Figure 5 shows the SD differences of item parameter estimates between different subsets and full data in empirical example 1. Clearly, as the number of subsets increases, the differences in item parameter SDs are larger. However, the largest difference in SD of parameter $\boldsymbol{b}$ is around 0.025. Notably, the SD differences for parameter $\boldsymbol{a}$ are even smaller, staying within 0.004. This indicates that the LS-WASP algorithm can accurately estimate the item parameters. In addition, the differences in EAP estimates of ability parameters between different subsets and full data are presented in online supplement S3. The results show that our algorithm's estimation of ability parameters remains closely aligned with the full data, without any significant deviation.

Figure 6 shows the 95% HPDIs of nine item parameters under different subsets. While the HPDI ranges of parameters $\boldsymbol{a}$ and $\boldsymbol{b}$ appear to increase with more subsets, the extent of this increase remains minimal. The running time of full data is approximately 1.3 h. When the full data are partitioned into two subsets, the time required is reduced by approximately half compared to the full data. When the full data are partitioned into 20 subsets, the computation time of the LS-WASP algorithm takes approximately three minutes. Therefore, the LS-WASP algorithm substantially enhances computational speed for big data and accurately estimates parameters in the 2PL model.

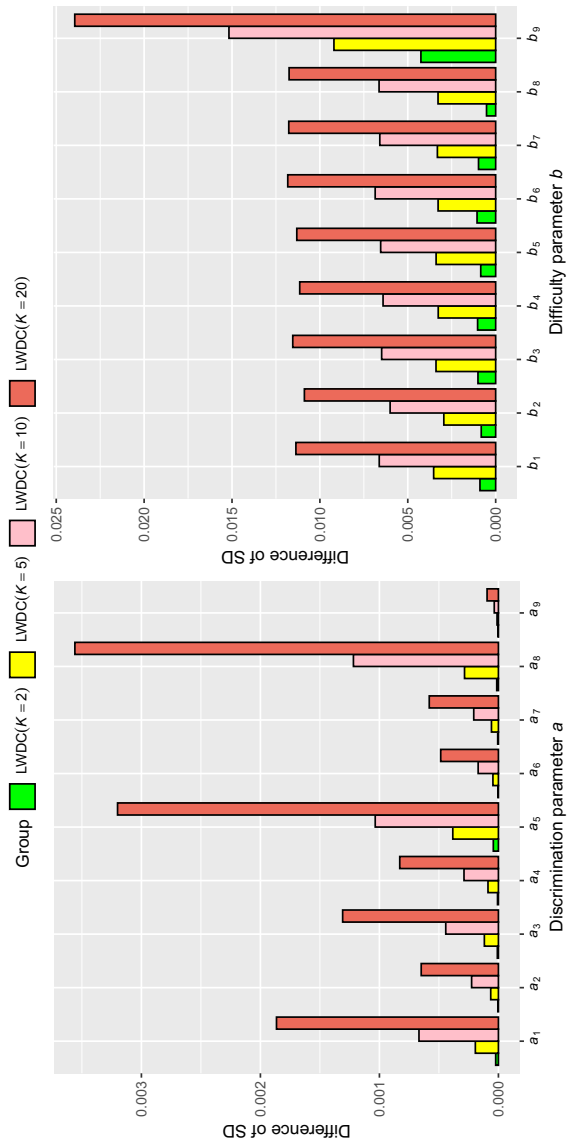


FIGURE 5.
Differences in SD values of item parameter estimates between different subsets and full data in empirical example 1.

TABLE 3.
EAPs and SD item parameters for PISA 2018 mathematics cognitive test.

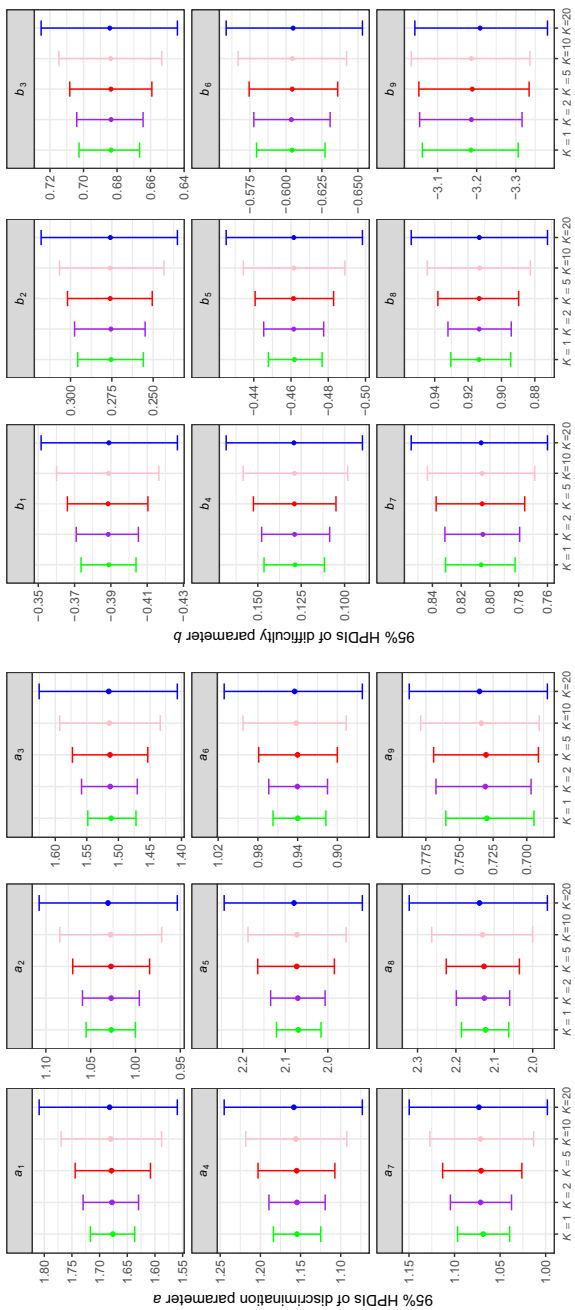
PARM	K=1		K=2		K=5		K=10		K=20	
	EAP	SD	EAP	SD	EAP	SD	EAP	SD	EAP	SD
a_1	1.6760	0.0206	1.6776	0.0254	1.6785	0.0345	1.6803	0.0465	1.6819	0.0638
a_2	1.0270	0.0140	1.0271	0.0162	1.0274	0.0220	1.0279	0.0290	1.0307	0.0395
a_3	1.5114	0.0197	1.5129	0.0227	1.5133	0.0306	1.5141	0.0408	1.5152	0.0559
a_4	1.1544	0.0151	1.1545	0.0179	1.1548	0.0245	1.1559	0.0321	1.1583	0.0439
a_5	2.0692	0.0266	2.0704	0.0332	2.0730	0.0462	2.0724	0.0588	2.0794	0.0832
a_6	0.9399	0.0135	0.9402	0.0153	0.9401	0.0203	0.9414	0.0266	0.9430	0.0356
a_7	1.0685	0.0146	1.0715	0.0170	1.0709	0.0222	1.0715	0.0290	1.0732	0.0387
a_8	2.1228	0.0319	2.1260	0.0357	2.1269	0.0488	2.1305	0.0668	2.1385	0.0916
a_9	0.7297	0.0166	0.7308	0.0181	0.7303	0.0200	0.7337	0.0225	0.7352	0.0264
b_1	−0.3888	0.0078	−0.3885	0.0087	−0.3885	0.0113	−0.3887	0.0144	−0.3889	0.0192
b_2	0.2754	0.0101	0.2756	0.0110	0.2760	0.0131	0.2760	0.0161	0.2759	0.0210
b_3	0.6836	0.0092	0.6836	0.0102	0.6836	0.0126	0.6838	0.0157	0.6843	0.0207
b_4	0.1285	0.0089	0.1288	0.0099	0.1289	0.0122	0.1289	0.0153	0.1292	0.0201
b_5	−0.4618	0.0074	−0.4615	0.0083	−0.4613	0.0108	−0.4616	0.0140	−0.4615	0.0187
b_6	−0.6043	0.0123	−0.6038	0.0134	−0.6044	0.0156	−0.6045	0.0192	−0.6050	0.0242
b_7	0.8061	0.0124	0.8049	0.0134	0.8054	0.0158	0.8055	0.0190	0.8061	0.0242
b_8	0.9134	0.0091	0.9134	0.0097	0.9134	0.0124	0.9131	0.0158	0.9133	0.0209
b_9	−3.1850	0.0630	−3.1865	0.0672	−3.1888	0.0722	−3.1864	0.0781	−3.2088	0.0869

PARM represents parameter, EAP denotes the expected a posteriori estimation, and SD is the standard deviation. $K = 1$ indicates the full data set, i.e., no data partitioning.

6. Discussion

For large-scale educational assessment data, current MCMC methods are significantly time-consuming when estimating the IRT models. To address this issue, we propose a divide-and-conquer algorithm named LS-WASP for distributed Bayesian inference in IRT models. This algorithm partitions the data into several subsets and then conducts parallel sampling of parameters for these subsets. Under the assumption of a location-scatter family, we propose an approximate Wasserstein posterior method as a substitute for the full data posterior sampling of parameters.

Simulation results and real data analysis validate the effectiveness of the LS-WASP algorithm in estimating IRT models. First, the LS-WASP algorithm exhibits a significant advantage in computational time. Specifically, the computational time of the LS-WASP algorithm is roughly inversely proportional to the number of subsets K. Second, the LS-WASP algorithm accurately estimates IRT model parameters. Simulation studies and real data analyses show that when sample sizes are large, regardless of the number of subsets, the estimates for item parameters from the LS-WASP algorithm closely align with those derived from the full data. However, when the sample size is small but the number of subsets is large, our method exhibits a slight difference in estimating discrimination parameters compared to the results based on full data, yet the estimation of difficulty parameters remains precise. The advantages of our proposed method become more prominent with larger sample sizes, and in the case of smaller sample sizes, we suggest to select an appropriate number of subsets to obtain the precise parameter estimates. The optimal subset sample size depends on the model's complexity; as complexity increases, so does the needed sample size per subset. Therefore, each subset's sample size should meet the minimum requirement for precise estimation, ensuring accuracy both within subsets and across the entire dataset. Furthermore, although our method primarily focuses on item parameters based on the data partitioning (i.e.,



(a) Discrimination parameter a (b) Difficulty parameter b

FIGURE 6.
95% HPDIs of the nine item parameters in empirical example 1.

grouping the persons), simulation results show that our method can still estimate ability parameters accurately.

The LS-WASP algorithm is not only conceptually simple and easy to implement, but also facilitates fast and accurate parameter estimation of IRT models as well as provides theoretical guarantees. We use this algorithm in conjunction with subsampling of parameters, and then based on the assumption that the posterior distributions of each subset belong to the same location-scatter family, an approximation of the true WASP can be derived. Along with other regularity assumptions, we derived the asymptotic Monte Carlo and statistical theoretical guarantees for the theoretical foundation of this algorithm.

In addition, we would like to reiterate the purpose of this study. As known, the EM algorithm has been the de facto method in operationally processing large educational datasets. However, it is crucial to acknowledge the fundamental differences between the two approaches: EM being an optimization algorithm requires relatively fewer iterations for convergence, and Bayesian estimation being a sampling-based approach requires a larger number of iterations for parameter convergence. This fundamental difference suggests that comparing them directly may not be entirely fair. Therefore, our aim is not to replace the EM algorithm but to offer an alternative that leverages the strengths of Bayesian inference in standard IRT settings, and provide a strong theoretical and practical framework for managing complex datasets. Our research seeks to explore the potential and applicability of the Bayesian "divide-and-conquer" method in scenarios where the Bayesian paradigm offers distinct advantages, particularly in handling the complexities inherent in large-scale educational data.

Despite its advantages, the LS-WASP algorithm has several limitations. First, the algorithm requires large sample size and relies on the appropriate selection of the number of subsets. The larger the sample size, the better the performance of the algorithm. When the sample size is small, choosing a large number of subsets may deteriorate the estimation performance, making the selection of an appropriate subset number crucial. It is recommended that each subset's sample size should meet the minimum requirement for precise parameter estimation. Second, the sampling algorithm used in this paper is the Pólya-Gamma Gibbs algorithm, but the LS-WASP algorithm is actually applicable to any MCMC sampling algorithms, such as the M-H algorithm, the slice algorithm (Neal, 2003; Lu et al., 2018) , and so on. Thus, additional sampling methods combined with our proposed LS-WASP algorithm can be explored in the future. Third, our results can also be generalized to cases with different subset sample sizes, but a common subset sample size still needs to be assumed to simplify the analysis. Fourth, the complexity of the LS-WASP algorithm may increase with complex IRT models; the further investigation of our proposed algorithm can be explored. Finally, the accuracy of parameter estimation using the "divide-and-conquer" approach can also be influenced by external factors, such as the unbalanced subset response data due to data partitioning or improper handling of missing data. Inappropriate data partitioning can easily lead to highly unbalanced response data within subsets, for instance, when the data allocated to a subset consist entirely of 0 s or 1 s. In such cases, some subsets may not contain enough responses from the minority groups, potentially leading to inaccurate parameter estimates. Typically, we need to be cautious with data partitioning to ensure that each subset contains enough minority class samples. After data partitioning, we should check each subset as thoroughly as possible to avoid unbalanced data splitting. Additionally, handling missing data improperly can lead to biased estimates. Imputing missing responses without proper consideration could deviate from our research goals and potentially distort the results (e.g., Robitzsch and Rupp 2009; Pohl et al. 2014; Sportisse et al. 2020; Du et al. 2022), especially when the missing response data are incorrectly imputed as complete data. Therefore, in this study, we believe that strictly managing missing data and relying solely on complete cases are pivotal for the reliability and validity of the study outcomes. Thus, while the "missing data" approach may be beneficial in certain contexts,

given its potential problems and complexity, we recommend further exploration and evaluation of the feasibility and effectiveness of distributed Bayesian estimation in future research.

Acknowledgments

This research was supported by the general projects of National Social Science Fund of China on Statistics (Grant No. 23BTJ067).

Declarations

Conflict of interest Each author signed a form for conflict of interest of potential conflict of interest. No authors reported any financial or other conflict of interest in relation to the work described. The author(s) declared no potential conflict of interest with respect to the research, authorship, and/or publication of this article

Ethical approval The authors affirm having followed professional ethical guidelines in preparing this work. These guidelines include obtaining informed consent from human participants, maintaining ethical treatment and respect for the rights of human or animal participants, and ensuring the privacy of participants and their data, such as ensuring that individual participants cannot be identified in reported results or from publicly available original or archival data.

Data availability The real data set was derived from the following resources available in the public domain: http://www.oecd.org/pisa.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

Ackerman, T. (1996). Graphical representation of multidimensional item response theory analyses. *Applied Psychological Measurement, 20*(4), 311–329.

Agueh, M., & Carlier, G. (2011). Barycenters in the Wasserstein space. *SIAM Journal on Mathematical Analysis, 43*(2), 904–924.

Alquier, P., Friel, N., Everitt, R., & Boland, A. (2016). Noisy Monte Carlo: Convergence of Markov chains with approximate transition kernels. *Statistics and Computing, 26*(1–2), 29–47.

Álvarez-Esteban, P. C., Del Barrio, E., Cuesta-Albertos, J. A., & Matrán, C. (2016). A fixed-point approach to barycenters in Wasserstein space. *Journal of Mathematical Analysis and Applications, 441*(2), 744–762.

Baker, F. B., & Kim, S. H. (2004). *Item response theory: Parameter estimation techniques*. New York: Dekker.

Balamuta, J. J., & Culpepper, S. A. (2022). Exploratory restricted latent class models with monotonicity requirements under Pólya-Gamma data augmentation. *Psychometrika, 87*(3), 903–945.

Béguin, A. A., & Glas, C. A. (2001). MCMC estimation and some model-fit analysis of multidimensional IRT models. *Psychometrika, 66*, 541–561.

Birnbaum, A. (1957). Efficient design and use of tests of a mental ability for various decision-making problems. *Series Report No. 58–16. Randolph Air Force Base*. USAF School of Aviation Medicine.

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

Choi, H. M., & Hobert, J. P. (2013). The Pólya-Gamma Gibbs sampler for Bayesian logistic regression is uniformly ergodic. *Electronic Journal of Statistics, 7*, 2054–2064.

Cuhadar, I. (2022). Sample size requirements for parameter recovery in the 4-Parameter logistic model. *Measurement: Interdisciplinary Research and Perspectives, 20*(2), 57–72.

Culpepper, S. A. (2016). Revisiting the 4-parameter item response model: Bayesian estimation and application. *Psychometrika, 81*(4), 1142–1163.

De Ayala, R. J. (2013). *Theory and practice of item response theory*. Cham: Guilford Publications.
de la Torre, J., & Hong, Y. (2010). Parameter estimation with small sample size a higher-order IRT model approach. *Applied Psychological Measurement, 34*, 267–285.
Du, H., Enders, C., Keller, B. T., Bradbury, T. N., & Karney, B. R. (2022). A Bayesian latent variable selection model for nonignorable missingness. *Multivariate Behavioral Research, 57*(2–3), 478–512.
Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.
Fox, J. P. (2010). *Bayesian item response modeling: Theory and applications*. New York: Springer.
Giordano, R., Broderick, T., & Jordan, M. I. (2018). Covariances, robustness and variational bayes. *Journal of Machine Learning Research, 19*(51), 1–49.
Hartshorne, J. K., Tenenbaum, J. B., & Pinker, S. (2018). A critical period for second language acquisition: Evidence from 2/3 million English speakers. *Cognition, 177*, 263–277.
Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika, 57*(1), 97–109.
Hoffman, M. D., Blei, D. M., Wang, C., & Paisley, J. (2013). Stochastic variational inference. *Journal of Machine Learning Research., 14*(5), 1303–1347.
Jiang, Z., & Templin, J. (2019). Gibbs samplers for logistic item response models via the Pólya-Gamma distribution: A computationally efficient data-augmentation strategy. *Psychometrika, 84*(2), 358–374.
Jimenez, A., Balamuta, J. J., & Culpepper, S. A. (2023). A sequential exploratory diagnostic model using a Pólya-gamma data augmentation strategy. *British Journal of Mathematical and Statistical Psychology, 76*(3), 513–538.
Kass, R. E., Tierney, L., & Kadane, J. B. (1990). The validity of posterior expansions based on Laplace's method. *Bayesian and Likelihood Methods in Statistics and Econometrics, 7*, 473–487.
König, C., Spoden, C., & Frey, A. (2020). An optimized Bayesian hierarchical two-parameter logistic model for small-sample item calibration. *Applied Psychological Measurement, 44*(4), 311–326.
Korattikara, A., Chen, Y., & Welling, M. (2014). Austerity in MCMC land: Cutting the Metropolis-Hastings budget. In *International Conference on Machine Learning*, pp. 181–189.
Lee, C. Y. Y., & Wand, M. P. (2016). Streamlined mean field variational Bayes for longitudinal and multilevel data analysis. *Biometrical Journal, 58*(4), 868–895.
Li, C., Srivastava, S., & Dunson, D. B. (2017). Simple, scalable and accurate posterior interval estimation. *Biometrika, 104*(3), 665–680.
Lu, J., Zhang, J. W., & Tao, J. (2018). Slice-Gibbs sampling algorithm for estimating the parameters of a multilevel item response model. *Journal of Mathematical Psychology, 82*, 12–25.
Martin, M. O., & Kelly, D. L. (1996). *Third international mathematics and science study technical report volume 1: Design and development*. Chestnut Hill: Boston College.
Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equations of state space calculations by fast computing machines. *The Journal of Chemical Physics, 21*(6), 1087–1092.
Minsker, S., Srivastava, S., Lin, L., & Dunson, D. B. (2017). Robust and scalable Bayes via a median of subset posterior measures. *The Journal of Machine Learning Research, 18*(1), 4488–4527.
Minsker, S., Srivastava, S., Lin, L., & Dunson, D. (2014). Scalable and robust Bayesian inference via the median posterior. In *International Conference on Machine Learning*, pp. 1656–1664.
Neal, R. (2003). Slice sampling. *The Annals of Statistics, 31*(3), 705–767.
Neiswanger, W., Wang, C., & Xing, E. (2014). Asymptotically exact, embarrassingly parallel MCMC. In *Proceedings of the 30th International Conference on Uncertainty in Artificial Intelligence*, pp. 623–632.
OECD. (2021). *PISA 2018 technical report*. Paris: OECD Publishing.
Pohl, S., Gräfe, L., & Rose, N. (2014). Dealing with omitted and not-reached items in competence tests: Evaluating approaches accounting for missing responses in item response theory models. *Educational and Psychological Measurement, 74*(3), 423–452.
Polson, N. G., Scott, J. G., & Windle, J. (2013). Bayesian inference for logistic models using Pólya-Gamma latent variables. *Journal of the American Statistical Association, 108*(504), 1339–1349.
Quiroz, M., Kohn, R., Villani, M., & Tran, M. N. (2019). Speeding up MCMC by efficient data subsampling. *Journal of the American Statistical Association, 114*(526), 831–843.
Reckase, M. D. (1972). *Development and application of a multivariate logistic latent trait model*. Syracuse University.
Reckase, M. D. (2009). *Multidimensional item response theory*. New York, NY: Springer.
Robitzsch, A., & Rupp, A. A. (2009). Impact of missing data on the detection of differential item functioning: The case of Mantel–Haenszel and logistic regression analysis. *Educational and Psychological Measurement, 69*(1), 18–34.
Rue, H., Martino, S., & Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society Series B: Statistical Methodology, 71*(2), 319–392.
San Martín, E. (2016). Identification of item response theory models. *Handbook of item response theory, 2*, 127–150.
Schilling, S., & Bock, R. D. (2005). High-dimensional maximum marginal likelihood item factor analysis by adaptive quadrature. *Psychometrika, 70*, 533–555.
Scott, S. L., Blocker, A. W., Bonassi, F. V., Chipman, H. A., George, E. I., & McCulloch, R. E. (2016). Bayes and big data: The consensus Monte Carlo algorithm. *International Journal of Management Science and Engineering Management, 11*(2), 78–88.
Shyamalkumar, N. D., & Srivastava, S. (2022). An algorithm for distributed Bayesian inference. *Stat, 11*(1), e432.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Cham: Crc Press.
Sportisse, A., Boyer, C., & Josse, J. (2020). Imputation and low-rank estimation with missing not at random data. *Statistics and Computing, 30*(6), 1629–1643.
Srivastava, S., Cevher, V., Dinh, Q., & Dunson, D. (2015). WASP: Scalable Bayes via barycenters of subset posteriors. In *Artificial Intelligence and Statistics*, pp. 912–920.
Srivastava, S., Li, C., & Dunson, D. B. (2018). Scalable Bayes via barycenter in Wasserstein space. *The Journal of Machine Learning Research, 19*(1), 312–346.
Srivastava, S., & Xu, Y. (2021). Distributed Bayesian inference in linear mixed-effects models. *Journal of Computational and Graphical Statistics, 30*(3), 594–611.
Tan, L. S., & Nott, D. J. (2014). A stochastic variational framework for fitting and diagnosing generalized linear mixed models. *Bayesian Analysis, 9*(4), 963–1004.
van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika, 72*(3), 287–308.
van der Linden, W. J., & Hambleton, R. K. (Eds.). (1997). *Handbook of modern item response theory*. New York: Springer-Verlag.
van Rijn, P. W., Sinharay, S., Haberman, S. J., & Johnson, M. S. (2016). Assessment of fit of item response theory models used in large-scale educational survey assessments. *Large-Scale Assessments in Education, 4*, 1–23.
Vehtari, A., Gelman, A., Sivula, T., Jylänki, P., Tran, D., Sahai, S., & Robert, C. P. (2020). Expectation propagation as a way of life: A framework for Bayesian inference on partitioned data. *The Journal of Machine Learning Research, 21*(1), 577–629.
Wang, C., Fan, Z., Chang, H. H., & Douglas, J. (2013). A semiparametric model for jointly analyzing response times and accuracy in computerized testing. *Journal of Educational and Behavioral Statistics, 38*(4), 381–417.
Wang, C., & Srivastava, S. (2023). Divide-and-conquer Bayesian inference in hidden Markov models. *Electronic Journal of Statistics, 17*(1), 895–947.
Wang, C., Xu, G., & Shang, Z. (2018). A two-stage approach to differentiating normal and aberrant behavior in computer based testing. *Psychometrika, 83*, 223–254.
Wu, M., Davis, R. L., Domingue, B. W., Piech, C., & Goodman, N. (2020). Variational item response theory: Fast, accurate, and expressive. ArXiv:2002.00276
Xue, J., & Liang, F. (2019). Double-parallel Monte Carlo for Bayesian analysis of big data. *Statistics and Computing, 29*(1), 23–32.

Manuscript Received: 25 OCT 2023
Accepted: 2 MAY 2024
Published Online Date: 30 MAY 2024