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

Online Supplement

S1. Gibbs Sampler via the Pólya-Gamma Distribution for the full data

First, we introduce the Pólya-Gamma distribution. A random variable, denoted as X, follows a Pólya-Gamma distribution with parameters $b > 0$ and $c \in R$ (denoted as $X \sim PG(b, c)$) if it satisfies the following condition:

$$X \sim \sum_{h=1}^{\infty} G(b,1)/(2\pi^2(h-0.5)^2 + \frac{c^2}{2}). \tag{1}$$

Here, $G(b,1)$ represents a Gamma distribution with parameters b and 1. Let $\{(y_{ij}, \omega_{ij})\}$ be the independent random pairs, where $y_{ij} \sim \text{Binom}(1, p_{ij})$ and augmented variable $\omega_{ij} \sim PG(1,0)$. Leveraging Theorem 1 from Polson et al. (2013), we can represent the likelihood contribution of the ith examinee's answer to the jth item as follows:

$$\begin{aligned} L_{ij}(\boldsymbol{\theta}, \boldsymbol{a}, \boldsymbol{b}) &= \frac{\exp\{a_j(\theta_i - b_i)\}^{y_{ij}}}{1 + \exp\{a_j(\theta_i - b_i)\}} \\ &\propto \exp\{\kappa_{ij} a_j(\theta_i - b_i)\} \int_0^{\infty} \exp\{-\frac{\omega_{ij}(a_j(\theta_i - b_i))^2}{2}\} p(\omega_{ij}|1,0) d\omega_{ij}, \end{aligned} \tag{2}$$

where $\kappa_{ij} = y_{ij} - \frac{1}{2}$.

The details of the MCMC sampling process for 2PL model are as follows:

Step 1: Given the parameters θ_i, a_j, b_j, and the observed data y_{ij}, we sample the auxiliary variable ω_{ij}. The full-conditional posterior distribution of ω_{ij} is given by:

$$\omega_{ij}|\theta_i, a_j, b_j \sim PG(1, a_j|\theta_i - b_j|). \tag{3}$$

Step 2: Sampling the ability parameter θ_i for each examinee i. The prior distribution of θ_i is assumed to follow a normal distribution, denoted by $\theta_i \sim N(\mu_1, \sigma_1^2)$. Given $\boldsymbol{a}$, $\boldsymbol{b}$, $\boldsymbol{\omega}$, and $\boldsymbol{y}$, the full-conditional posterior distribution of θ_i can be expressed as follows:

$$f(\theta_i|\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\omega}, \boldsymbol{y}) \propto f(\theta_i) \exp\left\{-\frac{1}{2}\left(\boldsymbol{z}_\theta - \boldsymbol{a}\theta_i\right)^T \boldsymbol{\Omega}_\theta \left(\boldsymbol{z}_\theta - \boldsymbol{a}\theta_i\right)\right\}, \tag{4}$$

where $\boldsymbol{z}_\theta = \left(\frac{a_1 b_1 \omega_{i1} + \kappa_{i1}}{\omega_{i1}}, \cdots, \frac{a_J b_J \omega_{iJ} + \kappa_{iJ}}{\omega_{iJ}}\right)^T$ and $\boldsymbol{\Omega}_\theta = \text{diag}(\omega_{i1}, \cdots, \omega_{iJ})$. Thus, the full-conditional posterior distribution of θ_i follows a normal distribution with a mean of $m_{\theta_i} = V_{\theta_i}(\boldsymbol{a}^T \boldsymbol{\Omega}_\theta \boldsymbol{z}_\theta + \frac{\mu_1}{\sigma_1^2})$ and a variance of $V_{\theta_i} = (\boldsymbol{a}^T \boldsymbol{\Omega}_\theta \boldsymbol{a} + \frac{1}{\sigma_1^2})^{-1}$.

Step 3: Sampling the discrimination parameter a_j for each item j. The prior distribution of a_j is assumed to follow a truncated normal distribution, such that $a_j \sim TN_{(0,+\infty)}(\mu_2, \sigma_2^2)$. Given $\boldsymbol{\theta}$, $\boldsymbol{b}$, $\boldsymbol{\omega}$, and $\boldsymbol{y}$, the full-conditional posterior distribution of a_j is given by:

$$f(a_j|\boldsymbol{\theta}, \boldsymbol{b}, \boldsymbol{\omega}, \boldsymbol{y}) \propto f(a_j) \exp\left\{-\frac{1}{2}\left[\boldsymbol{z}_a - (\boldsymbol{\theta} - \mathbf{1}b_j)a_j\right]^T \boldsymbol{\Omega}_{ab} \left[\boldsymbol{z}_a - (\boldsymbol{\theta} - \mathbf{1}b_j)a_j\right]\right\}, \tag{5}$$

where $\boldsymbol{z}_a = \left(\frac{\kappa_{1j}}{\omega_{1j}}, \cdots, \frac{\kappa_{nj}}{\omega_{nj}}\right)^T$, $\mathbf{1} = (1, \cdots, 1)^T_{n\times 1}$ and $\boldsymbol{\Omega}_{ab} = \text{diag}(\omega_{1j}, \cdots, \omega_{nj})$. Therefore, the full-conditional posterior distribution of a_j follows a truncated normal distribution at zero, with a mean of $m_{a_j} = V_{a_j}\left[(\boldsymbol{\theta} - \mathbf{1}b_j)^T \boldsymbol{\Omega}_{ab} \boldsymbol{z}_a + \frac{\mu_2}{\sigma_2^2}\right]$ and variance $V_{a_j} = \left[(\boldsymbol{\theta} - \mathbf{1}b_j)^T \boldsymbol{\Omega}_{ab}(\boldsymbol{\theta} - \mathbf{1}b_j) + \frac{1}{\sigma_2^2}\right]^{-1}$.

Step 4: Sampling the difficulty parameter b_j for each item j. The prior distribution of b_j is assumed to follow a normal distribution, i.e., $b_j \sim N(\mu_3, \sigma_3^2)$. Given $\boldsymbol{\theta}$, $\boldsymbol{a}$, $\boldsymbol{\omega}$, and $\boldsymbol{y}$, the full-

conditional posterior distribution of b_j is given by:

$$f(b_j|\boldsymbol{\theta},\boldsymbol{a},\boldsymbol{\omega},\boldsymbol{y}) \propto f(b_j)\exp\left\{-\frac{1}{2}\big(\boldsymbol{z}_b+\mathbf{1}a_jb_j\big)^T\boldsymbol{\Omega}_{ab}\big(\boldsymbol{z}_b+\mathbf{1}a_jb_j\big)\right\}, \tag{6}$$

where $\boldsymbol{z}_b = \left(\frac{\kappa_{1j}-a_j\theta_1\omega_{1j}}{\omega_{1j}},\cdots,\frac{\kappa_{nj}-a_j\theta_n\omega_{nj}}{\omega_{nj}}\right)^T$ and $\boldsymbol{\Omega}_{ab} = \text{diag}\,(\omega_{1j},\cdots,\omega_{nj})$. Hence, the full-conditional posterior distribution of b_j follows a normal distribution with a mean $m_{b_j} = V_{b_j}\left[-(\mathbf{1}a_j)^T\boldsymbol{\Omega}_{ab}\boldsymbol{z}_b+\frac{\mu_3}{\sigma_3^2}\right]$ and a variance $V_{b_j} = \left[(\mathbf{1}a_j)^T\boldsymbol{\Omega}_{ab}\mathbf{1}a_j+\frac{1}{\sigma_3^2}\right]^{-1}$.

Similarly, based on the Pólya-Gamma distribution, the full-conditional posterior distributions of the parameters for M2PL model can be expressed as:

$$\begin{aligned}
&\omega_{ij}|\boldsymbol{\theta}_i,\boldsymbol{a}_j,b_j \sim PG\left(1,|\boldsymbol{a}_j^T\boldsymbol{\theta}_i-b_j|\right),\\
&f(\boldsymbol{\theta}_i|\boldsymbol{a},\boldsymbol{b},\boldsymbol{\omega},\boldsymbol{y}) \propto f(\boldsymbol{\theta}_i)\exp\left\{-\frac{1}{2}(\boldsymbol{z}_\theta-\boldsymbol{a}\boldsymbol{\theta}_i)^T\boldsymbol{\Omega}_\theta(\boldsymbol{z}_\theta-\boldsymbol{a}\boldsymbol{\theta}_i)\right\},\\
&f(a_{jq}|\boldsymbol{\theta},\boldsymbol{b},\boldsymbol{\omega},\boldsymbol{y}) \propto f(a_{jq})\exp\left\{-\frac{1}{2}\big(\boldsymbol{z}_{aq}-\boldsymbol{\theta}_{\cdot q}a_{jq}\big)^T\boldsymbol{\Omega}_{ab}\big(\boldsymbol{z}_{aq}-\boldsymbol{\theta}_{\cdot q}a_{jq}\big)\right\},\\
&f(b_j|\boldsymbol{\theta},\boldsymbol{a},\boldsymbol{\omega},\boldsymbol{y}) \propto f(b_j)\exp\left\{\frac{1}{2}\big(\boldsymbol{z}_b-\mathbf{1}b_j\big)^T\boldsymbol{\Omega}_{ab}\big(\boldsymbol{z}_b-\mathbf{1}b_j\big)\right\},
\end{aligned} \tag{7}$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_1,\cdots,\boldsymbol{\theta}_n)^T$ is the $n\times Q$ matrix of ability parameters, $\boldsymbol{\theta}_i = (\theta_{i1},\cdots,\theta_{iQ})^T$ and $\boldsymbol{\theta}_{\cdot q} = (\theta_{1q},\cdots,\theta_{nq})^T$ denote the vectors consisting of the elements of the ith row and qth column of $\boldsymbol{\theta}$, respectively. Also, $\boldsymbol{a} = (\boldsymbol{a}_1,\cdots,\boldsymbol{a}_J)^T$ is the $J\times Q$ matrix of discrimination parameters with $\boldsymbol{a}_j = (a_{j1},\cdots,a_{jQ})^T$. $\boldsymbol{\Omega}_\theta$ and $\boldsymbol{\Omega}_{ab}$ are $\text{diag}(\omega_{i1},\cdots,\omega_{iJ})$ and $\text{diag}(\omega_{1j},\cdots,\omega_{nj})$, respectively. In the M2PL model, the vectors denoted by $\boldsymbol{z}_\theta$, $\boldsymbol{z}_b$ and $\boldsymbol{z}_{aq}$ are as follows:

$$\begin{aligned}
&\boldsymbol{z}_\theta = (\frac{b_1\omega_{i1}+\kappa_{i1}}{\omega_{i1}},\cdots,\frac{b_J\omega_{iJ}+\kappa_{iJ}}{\omega_{iJ}})^T,\quad \boldsymbol{z}_b = (\frac{\boldsymbol{a}_j^T\boldsymbol{\theta}_1\omega_{1j}-\kappa_{1j}}{\omega_{1j}},\cdots,\frac{\boldsymbol{a}_j^T\boldsymbol{\theta}_n\omega_{nj}-\kappa_{nj}}{\omega_{nj}})^T,\\
&\boldsymbol{z}_{aq} = (\frac{\kappa_{1j}+\omega_{1j}b_j-\omega_{1j}\boldsymbol{\theta}_{1,-q}^T\boldsymbol{a}_{j,-q}}{\omega_{1j}},\cdots,\frac{\kappa_{nj}+\omega_{nj}b_j-\omega_{nj}\boldsymbol{\theta}_{n,-q}^T\boldsymbol{a}_{j,-q}}{\omega_{nj}})^T,
\end{aligned}$$

where $\kappa_{ij} = y_{ij}-\frac{1}{2}$, $\boldsymbol{\theta}_{i,-q}$ denotes the vector of $\boldsymbol{\theta}_i$ excluding the element $\boldsymbol{\theta}_{iq}$, and $a_{j,-q}$ denotes the vector of $\boldsymbol{a}_j$ excluding the element a_{jq}.

The prior distributions of $\boldsymbol{\theta}_i$, a_{jq} and b_j are assumed to follow $N(\boldsymbol{\mu}_1, \Sigma_1)$, $TN_{(0,+\infty)}(\mu_2, \sigma_2^2)$ and $N(\mu_3, \sigma_3^2)$, respectively. Subsequently, the sampling steps for M2PL model are as follows:

(1) Given $\boldsymbol{\theta}$, $\boldsymbol{a}$, and $\boldsymbol{b}$, draw ω_{ij} from the $PG(1, |\boldsymbol{a}_j^T\boldsymbol{\theta}_i - b_j|)$ distribution;

(2) Given $\boldsymbol{\omega}$, $\boldsymbol{a}$, and $\boldsymbol{b}$, draw $\boldsymbol{\theta}_i$ from $N(\boldsymbol{m}_{\theta_i}, \boldsymbol{V}_{\theta_i})$, where $\boldsymbol{m}_{\theta_i} = \boldsymbol{V}_{\theta_i}\left(\boldsymbol{a}^T\boldsymbol{\Omega}_\theta\boldsymbol{z}_\theta + \Sigma_1^{-1}\boldsymbol{\mu}_1\right)$, $\boldsymbol{V}_{\theta_i} = \left(\boldsymbol{a}^T\boldsymbol{\Omega}_\theta\boldsymbol{a} + \Sigma_1^{-1}\right)^{-1}$;

(3) Given $\boldsymbol{\omega}$, $\boldsymbol{\theta}$, and $\boldsymbol{b}$, draw a_{jq} from the $TN_{(0,+\infty)}\left(m_{a_{jq}}, V_{a_{jq}}\right)$, where $m_{a_{jq}} = V_{a_{jq}}\left(\boldsymbol{\theta}_{\cdot q}^T\boldsymbol{\Omega}_{ab}\boldsymbol{z}_{aq} + \frac{\mu_2}{\sigma_2^2}\right)$, $V_{a_{jq}} = \left(\boldsymbol{\theta}_{\cdot q}^T\boldsymbol{\Omega}_{ab}\boldsymbol{\theta}_{\cdot q} + \frac{1}{\sigma_2^2}\right)^{-1}$;

(4) Given $\boldsymbol{\omega}$, $\boldsymbol{\theta}$, and $\boldsymbol{a}$, draw b_j from $N(m_{b_j}, V_{b_j})$, where $m_{b_j} = V_{b_j}\left(\mathbf{1}^T\boldsymbol{\Omega}_{ab}\boldsymbol{z}_b + \frac{\mu_3}{\sigma_3^2}\right)$ and variance $V_{b_j} = \left(\mathbf{1}^T\boldsymbol{\Omega}_{ab}\mathbf{1} + \frac{1}{\sigma_3^2}\right)^{-1}$.

S2. Additional results of the simulation studies

The average Bias and RMSE in the discrimination parameters as a function of the number of subsets K in different sample sizes and test lengths in simulation study 1 are shown in Figure S-1 and S-2.

The bias and RMSE of each item for $J = 40$ in simulation study 1 are shown in Figure S-3.

The bias and RMSE of each item for $J = 40$ in simulation study 2 are shown in Figure S-4.

S3. Results of ability parameter estimates in section 6

Figure S-5a depicts the differences in EAP estimates of ability parameters between different subsets and full data in empirical example 1, and Figure S-5b displays the square of these differences. The differences in EAP estimates of ability parameters between different subsets and full data tend

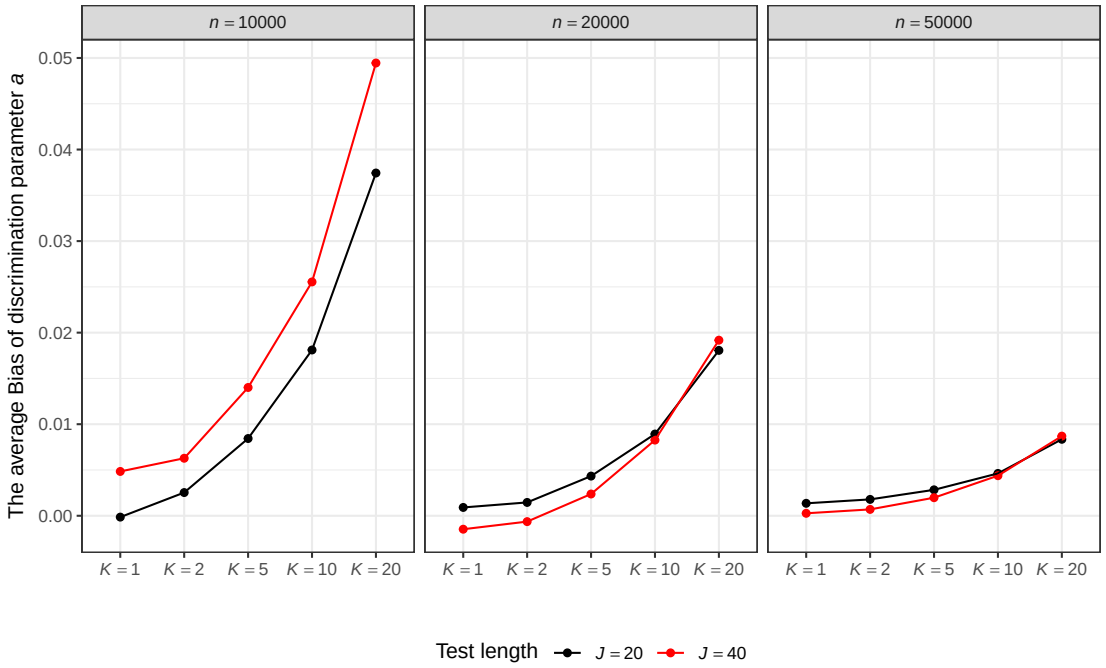

Figure S-1: The average Bias in the discrimination parameters as a function of the number of subsets K in different sample sizes and test lengths.

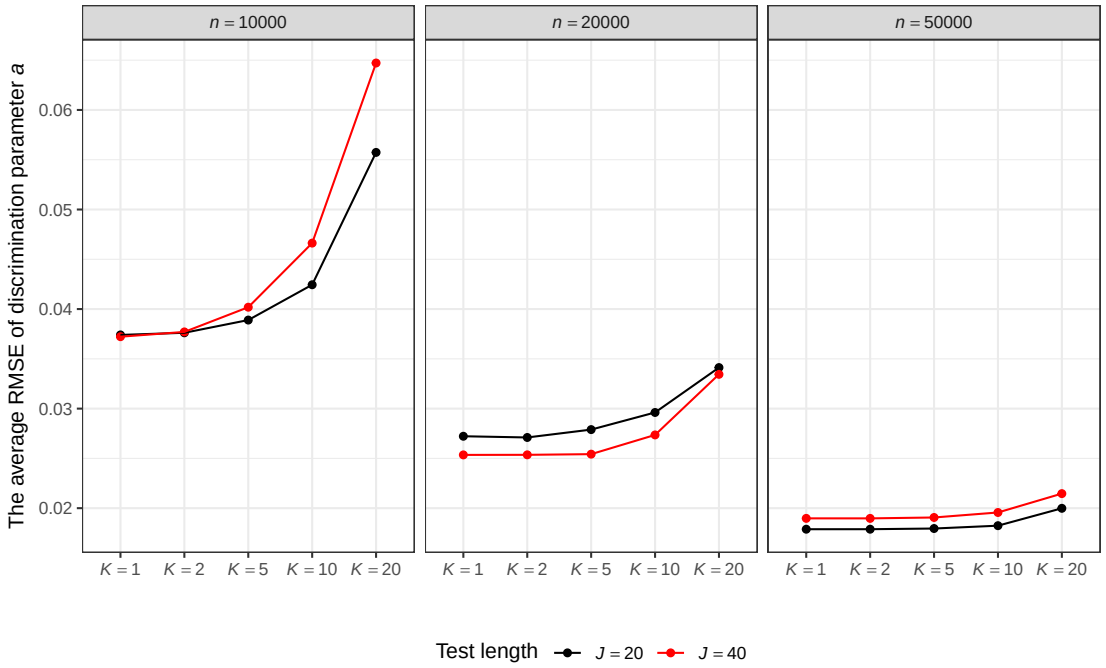

Figure S-2: The average RMSE in the discrimination parameters as a function of the number of subsets K in different sample sizes and test lengths.

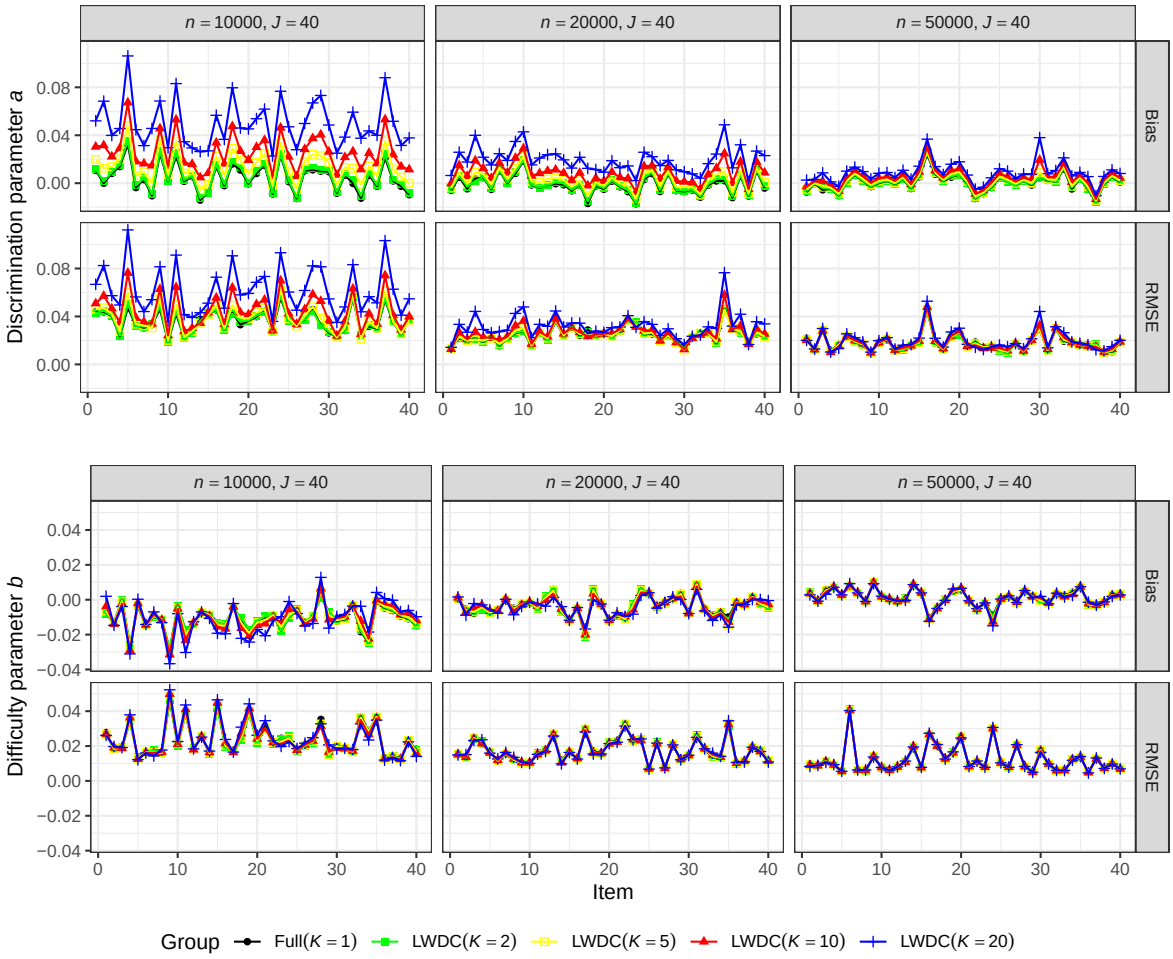

Figure S-3: The bias and RMSE of each item parameter estimate across various sample sizes with a fixed test length $J = 40$ in simulation study 1. Note that 'Group' indicates the number of subsets.

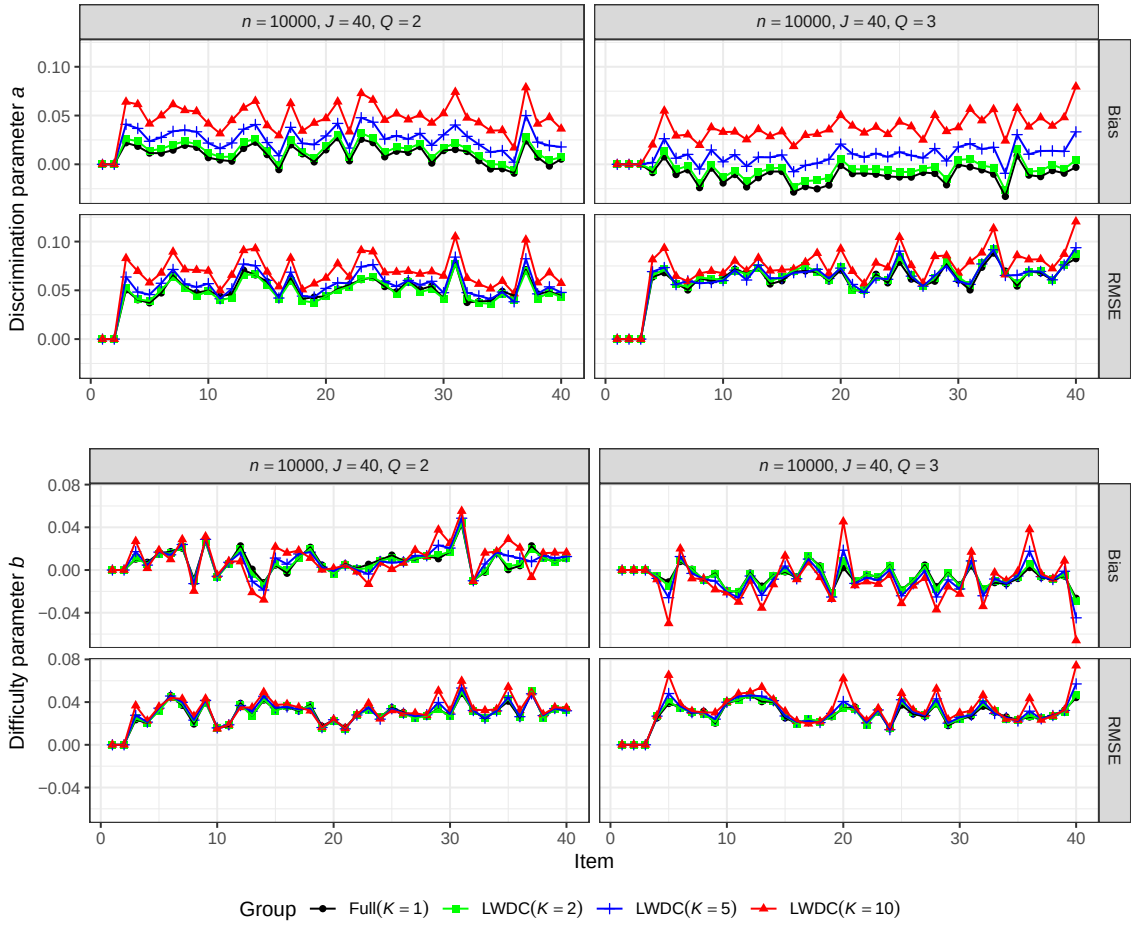

Figure S-4: The bias and RMSE of each item parameter estimate across various latent trait dimensions with a fixed test length $J = 40$ in simulation study 2. Note that 'Group' indicates the number of subsets.

to increase as the number of subsets grows. However, most differences in Figure S-5a fall between –0.025 and 0.025, and most squares of these differences in Figure S-5b are below 0.002. Therefore, compared to the full data, our algorithm does not significantly deviate in its estimation of ability parameters.

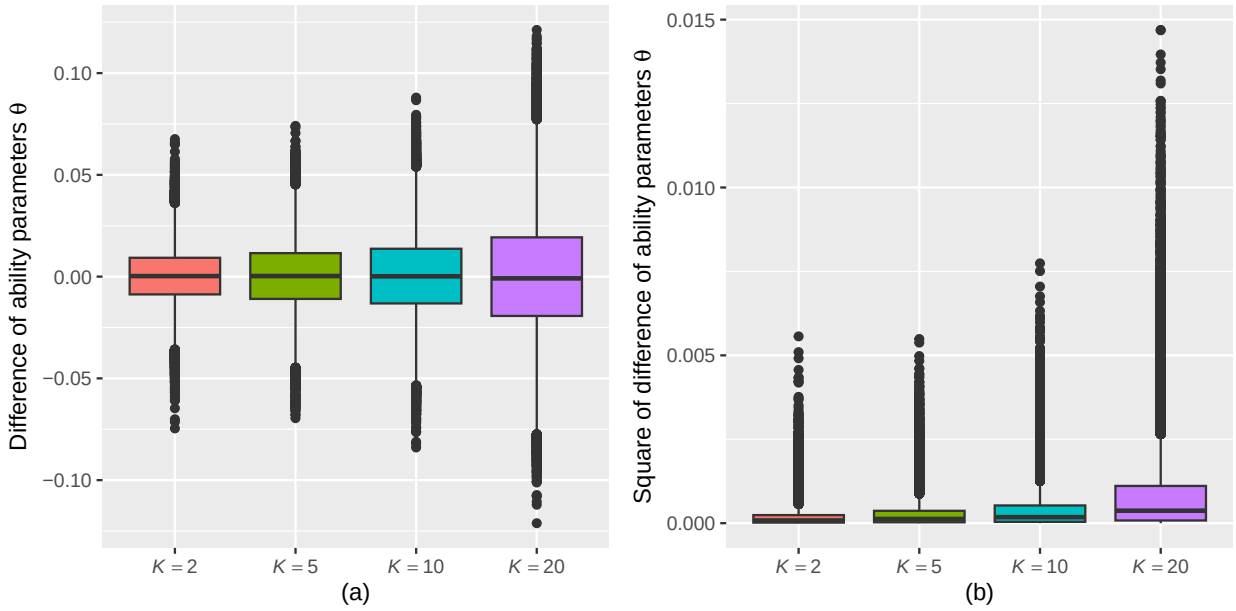

Figure S-5: Boxplots of the differences in EAP estimates of ability parameters between different subsets and full data in empirical example 1.

S4. Real Data Example 2

S4.1. Data description

Here, we focused on the PISA 2015 computer-based mathematics data (OECD, 2018) and specifically considered 11 items that were previously analyzed by Man et al. (2019). The item IDs include CM474Q01S, CM155Q01S, CM411Q01S, CM411Q02S, CM442Q02S, CM305Q01S,

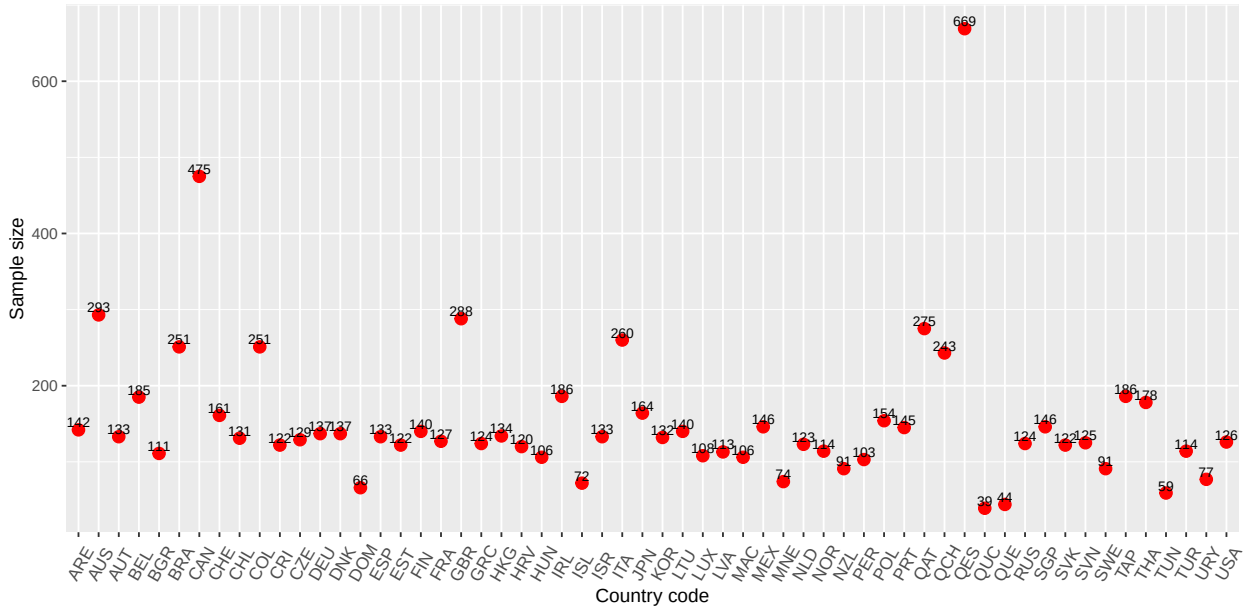

Figure S-6: 58 countries participating in the 2015 PISA mathematics cognitive test and the number of students from each country.

CM496Q01S, CM496Q02S, CM603Q01S, CM564Q01S, and CM564Q02S. According to the PISA 2015 mathematics framework codebook, these 11 items assess two dimensions: (a) employing mathematical concepts, facts procedures, and reasoning and (b) context societal knowledge. After excluding non-response data as well as "Not Reached", "Not Applicable", and "Invalid" data, a total of 9000 students from 58 countries responded to these 11 binary-scored items. As shown in Figure S-6, the country with the largest sample size of 669 is "QES" (i.e., Spain (Regions)), while "QUC" (i.e., Massachusetts, USA)) has the smallest sample size of 39. The correct response rates for the 11 items are 65.10%, 70.60%, 50.00%, 48.03%, 31.60%, 43.60%, 48.33%, 66.18%, 36.88%, 48.86%, and 46.38%. The item "CM155Q01S" has the highest correct response rate at 70.60%, while "CM442Q02S" has the lowest at 31.60%.

Table S-1: The EAPs and SD values for item parameters for PISA 2015 mathematics cognitive test.

PARM	K=1		K=2		K=4		K=6	
	EAP	SD	EAP	SD	EAP	SD	EAP	SD
$a_{1.3}$	1.0871	0.0629	1.0785	0.0755	1.0156	0.0984	1.0498	0.1321
$a_{1.4}$	0.5868	0.0483	0.5714	0.0573	0.4951	0.0713	0.4812	0.1074
$a_{1.5}$	1.2259	0.0740	1.2019	0.0908	1.0855	0.1078	1.1283	0.1552
$a_{1.6}$	0.2356	0.0416	0.2277	0.0483	0.2245	0.0609	0.2155	0.0705
$a_{1.7}$	0.8685	0.0701	0.8680	0.0890	1.0694	0.1307	1.0636	0.1759
$a_{1.8}$	0.6427	0.0644	0.6526	0.0844	0.8760	0.1162	0.8592	0.1838
$a_{1.9}$	0.6572	0.0500	0.6466	0.0622	0.5896	0.0766	0.5942	0.0968
$a_{1.10}$	0.7663	0.0565	0.7727	0.0668	0.6483	0.0837	0.7119	0.1185
$a_{1.11}$	0.6668	0.0523	0.6680	0.0623	0.5875	0.0795	0.6297	0.1038
$a_{2.3}$	1.0688	0.0555	1.1003	0.0688	1.1680	0.0884	1.1512	0.1242
$a_{2.4}$	0.4793	0.0424	0.4928	0.0505	0.5652	0.0648	0.5880	0.1000
$a_{2.5}$	1.1851	0.0625	1.2149	0.0791	1.3214	0.1064	1.3719	0.1679
$a_{2.6}$	0.2056	0.0382	0.2142	0.0432	0.2163	0.0541	0.2246	0.0625
$a_{2.7}$	1.9171	0.0971	1.9105	0.1416	1.8091	0.1697	1.8181	0.2348
$a_{2.8}$	1.7280	0.0868	1.6860	0.1144	1.5123	0.1326	1.5623	0.1981
$a_{2.9}$	0.4843	0.0424	0.4967	0.0545	0.5420	0.0657	0.5491	0.0863
$a_{2.10}$	0.3215	0.0427	0.3190	0.0521	0.4066	0.0671	0.3670	0.0947
$a_{2.11}$	0.3981	0.0429	0.3966	0.0503	0.4641	0.0648	0.4387	0.0855
b_3	0.3239	0.0312	0.3304	0.0356	0.3333	0.0435	0.3371	0.0496
b_4	0.2453	0.0251	0.2461	0.0266	0.2484	0.0299	0.2531	0.0334
b_5	1.5275	0.0488	1.5318	0.0550	1.5438	0.0688	1.5913	0.0925
b_6	0.3281	0.0229	0.3288	0.0234	0.3294	0.0246	0.3314	0.0260
b_7	0.5651	0.0406	0.5647	0.0499	0.5761	0.0612	0.5713	0.0739
b_8	−0.6708	0.0348	−0.6657	0.0405	−0.6652	0.0484	−0.6832	0.0599
b_9	0.7826	0.0278	0.7844	0.0308	0.7860	0.0355	0.7921	0.0403
b_{10}	0.2045	0.0251	0.2046	0.0275	0.2033	0.0310	0.2083	0.0343
b_{11}	0.3169	0.0252	0.3176	0.0270	0.3192	0.0308	0.3217	0.0336

Note: 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.

S4.2. Designs

As the data set sample size is limited to 9,000, it is concluded in Section 5 that, when dealing with a relatively small sample size, the number of subsets should not exceed 10 for optimal performance. Therefore, we partitioned the data into $K = 2$, 4, and 6 subsets. As these 11 items measure

two dimensions, we set the number of latent traits to $Q = 2$. These two latent dimensions are compensatory each other, because if students possess only societal knowledge but lack mathematical concepts and reasoning skills, they cannot properly answer the items. Therefore, the M2PL model is utilized to fit this dataset. We conducted 10,000 MCMC iterations, discarding the initial 5,000 as burn-in. Subsequently, we analyzed the running time of the LS-WASP algorithm under different subsets.

S4.3. Results

Table S-1 presents the item parameter estimates under different subsets using the LS-WASP algorithm. The most accurate estimates are obtained when data is partitioned into two subsets. When the number of subsets is increased to 4 or 6, there is a slight deviation in estimates compared to using the full data. However, the difference for the majority of parameters remains within 0.1, with only a few exceeding 0.2. This might result from the reduction in sample size per subset as the number of subsets increases.

Table S-1 displays the SDs of item parameter estimates under different conditions, Figure S-7 shows the bar plots of the differences in the SD values of EAP estimates of item parameters between different subsets and full data in empirical example 2. The SD values progressively increases with the number of subsets. However, Figure S-7 illustrates that the maximum difference in SD for parameter $\boldsymbol{a}$ is less than 0.02, while for parameter $\boldsymbol{b}$, it mainly remains within 0.04. Figures S-8 presents the difference in the ability parameter $\boldsymbol{\theta}$ in empirical example 2. As the number of subsets increases, the difference between the two latent traits of $\boldsymbol{\theta}$ maintains an increasing trend. However, most of the differences between the two latent traits range from -0.1 to 0.1. The maximum

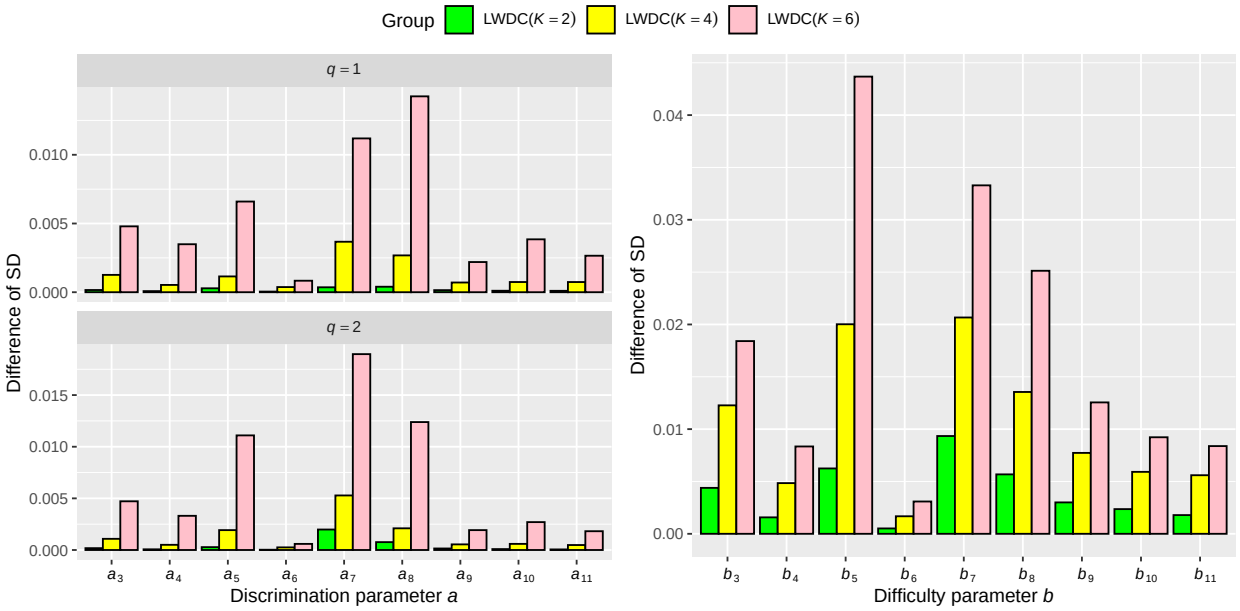

Figure S-7: Differences between the SD values of item parameter estimates under different subsets and the SD values of item parameter estimates under the full data in empirical example 2.

squared difference presented in Figure S-8b is 0.02. This concludes that the LS-WASP algorithm can accurately estimate the ability parameters of the M2PL model.

Figure S-9 shows the running time of the LS-WASP algorithm, which still exhibits a multiplicative relationship with the full data running time. In summary, the LS-WASP algorithm exhibits applicability to the M2PL model. With proper data partitioning, it continues to accurately and efficiently estimate the M2PL model's parameters.

S5. Verification of Assumption 3 (a) and (b) and proof of the theorems

S5.1. Verification of Assumption 3 (a) and (b)

Proposition 1. Given that the discrimination parameter $\boldsymbol{a}$, the difficulty parameter $\boldsymbol{b}$, and the ability parameter $\boldsymbol{\theta}$ are all bounded, then Assumption 3 (a) and (b) is satisfied for the 2PL model and M2PL model.

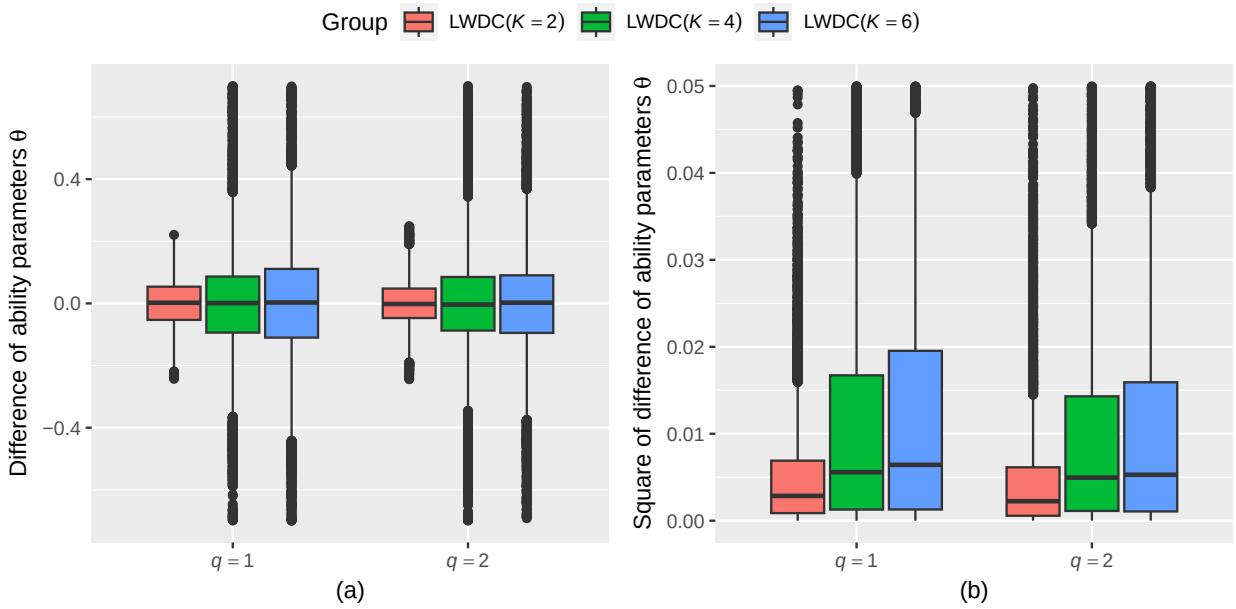

Figure S-8: Boxplots of the differences between the EAPs of ability parameters under different subsets and the EAP estimates of ability parameters under the full data in empirical example 2.

Proof. The log marginal likelihood function for the 2PL model is as follows:

$$\ell(\boldsymbol{\eta}) = \sum_{i=1}^{n} \log f(\boldsymbol{y}_i \mid \boldsymbol{\eta}) = \sum_{i=1}^{n} \log \int p(\boldsymbol{y}_i \mid \boldsymbol{\eta}, \theta_i)\, \phi(\theta_i) d\theta_i, \tag{8}$$

where $p(\boldsymbol{y}_i \mid \boldsymbol{\eta}, \theta_i) = \prod_{j=1}^{J} P_j(\theta_i)^{y_{ij}} Q_j(\theta_i)^{1-y_{ij}}$, $P_j(\theta_i) = \frac{\exp\{a_j(\theta_i - b_j)\}}{1+\exp\{a_j(\theta_i - b_j)\}}$, $Q_j(\theta_i) = 1 - P_j(\theta_i)$, and $\phi(\theta_i)$ is the population distribution of θ_i. Note that the log-likelihood function involved an integral which can be challenging to evaluate on a digital computer. Consequently, Hermite-Gauss quadrature is frequently utilized to approximate these integrals in IRT studies (Bock & Aitkin, 1981; Harwell et al., 1988; Baker & Kim, 2004). Specifically, if $\phi(\theta_i)$ is a continuous distribution with finite moments, it can be approximated to any desired degree of accuracy by a discrete distribution over a finite number of points (Baker & Kim, 2004). Using the Hermite-Gauss quadrature approximation to approximate this integral, we derive the first and second order derivatives of $h_n(\boldsymbol{\eta}) = -\frac{1}{n}\ell(\boldsymbol{\eta})$

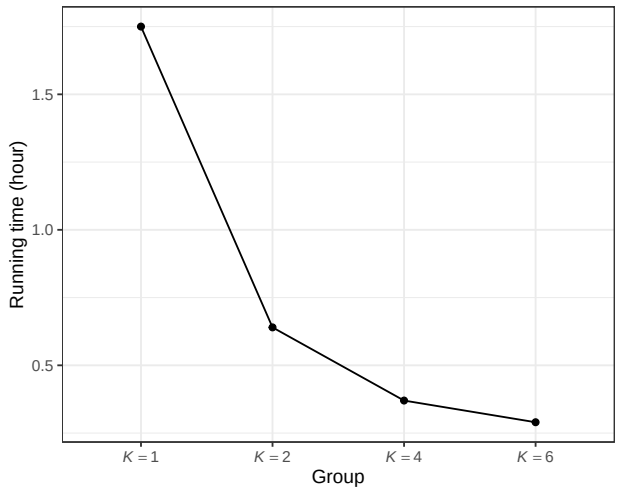

Figure S-9: Running times under different subset conditions in empirical example 2.

with respect to the item parameters as follows. For the calculation procedure and derivatives of higher orders, please refer to Baker and Kim (2004).

$$\frac{\partial h_n(\boldsymbol{\eta})}{\partial a_j} = -\sum_{q=1}^{Q} \big(\overline{r}_{jq} - \overline{f}_{jq} P_j(X_q)\big)(X_q - b_j), \tag{9}$$

$$\frac{\partial^2 h_n(\boldsymbol{\eta})}{\partial a_j^2} = \sum_{q=1}^{Q} (X_q - b_j)^2 \overline{f}_{jq} P_j(X_q) Q_j(X_q), \tag{10}$$

$$\frac{\partial h_n(\boldsymbol{\eta})}{\partial b_j} = \sum_{q=1}^{Q} a_j \big(\overline{r}_{jq} - \overline{f}_{jq} P_j(X_q)\big), \tag{11}$$

$$\frac{\partial^2 h_n(\boldsymbol{\eta})}{\partial b_j^2} = \sum_{q=1}^{Q} a_j^2 \overline{f}_{jq} P_j(X_q) Q_j(X_q), \tag{12}$$

$$\frac{\partial^2 h_n(\boldsymbol{\eta})}{\partial a_j \partial b_j} = \sum_{q=1}^{Q} \big(\overline{r}_{jq} - \overline{f}_{jq} P_j(X_q)\big) - \sum_{q=1}^{Q} (X_q - b_j) a_j \overline{f}_{jq} P_j(X_q) Q_j(X_q), \tag{13}$$

where $\overline{r}_{jq} = \frac{1}{n}\sum_{i=1}^{n} y_{ij} p\left(X_q \mid \boldsymbol{y}_i, \boldsymbol{\eta}\right)$, $\overline{f}_{jq} = \frac{1}{n}\sum_{i=1}^{n} p\left(X_q \mid \boldsymbol{y}_i, \boldsymbol{\eta}\right)$, $P_j(X_q) = \frac{\exp\{a_j(X_q - b_j)\}}{1+\exp\{a_j(X_q - b_j)\}}$, and $Q_j(X_q) = 1 - P_j(X_q)$. Additionally, X_q $(q = 1, \cdots, Q)$ is referred to as a Hermite-Gauss quadrature "node", and the number of nodes, Q, is finite. Since y_{ij} can only take values of 0 or 1, and the probability values satisfy $0 < P_j\left(X_q\right) < 1$, $0 < p\left(X_q \mid \boldsymbol{y}_i, \boldsymbol{\eta}\right) < 1$, it is assured that both

$\overline{r}_{jq} = \frac{1}{n}\sum_{i=1}^{n} y_{ij} p\left(X_q \mid \boldsymbol{y}_i, \boldsymbol{\eta}\right)$ and $\overline{f}_{jq} = \frac{1}{n}\sum_{i=1}^{n} p\left(X_q \mid \boldsymbol{y}_i, \boldsymbol{\eta}\right)$ are bounded. Additionally, since the MLE $\widehat{\boldsymbol{\eta}}_n$ remains bounded when the parameters $\boldsymbol{a}$, $\boldsymbol{b}$, $\boldsymbol{\theta}$ are bounded, Assumption 3 (a) is satisfied for the 2PL model. It is obvious that the parameters $\boldsymbol{a}$, $\boldsymbol{b}$, $\boldsymbol{\theta}$ are all bounded. For instance, although theoretically the ability could range from $-\infty$ to $+\infty$, in practical applications, it typically varies between -4 and 4.

We then proceed to verify Assumption 3 (b). Denote $\overline{f}_{jq} P_j(X_q) Q_j(X_q) \triangleq w_{jq}$ and $\sum_{q=1}^{Q}\left(\overline{r}_{jq} - \overline{f}_{jq} P_j(X_q)\right) \triangleq d_j$. According to the definition of maximum likelihood estimation, we have

$$\frac{\partial h_n(\boldsymbol{\eta})}{\partial a_j} \mid_{a_j=\widehat{a}_j} = -\sum_{q=1}^{Q}\left(\widehat{\overline{r}}_{jq} - \widehat{\overline{f}}_{jq}\widehat{P}_j(X_q)\right)(X_q - \widehat{b}_j) = 0,$$
$$\frac{\partial h_n(\boldsymbol{\eta})}{\partial b_j} \mid_{b_j=\widehat{b}_j} = \sum_{q=1}^{Q}\widehat{a}_j\left(\widehat{\overline{r}}_{jq} - \widehat{\overline{f}}_{jq}\widehat{P}_j(X_q)\right) = 0.$$

Then, we have $\widehat{d}_j = \sum_{q=1}^{Q}\left(\widehat{\overline{r}}_{jq} - \widehat{\overline{f}}_{jq}\widehat{P}_j(X_q)\right) = 0$. Hence, the Hessian matrix of $h_n(\boldsymbol{\eta})$ at $\widehat{\boldsymbol{\eta}}_n$ is

$$\begin{aligned} D^2(h_n(\widehat{\boldsymbol{\eta}}_n)) &= \begin{pmatrix} \frac{\partial^2 h_n(\boldsymbol{\eta})}{\partial a_j^2} & \frac{\partial^2 h_n(\boldsymbol{\eta})}{\partial a_j \partial b_j} \\ \frac{\partial^2 h_n(\boldsymbol{\eta})}{\partial b_j \partial a_j} & \frac{\partial^2 h_n(\boldsymbol{\eta})}{\partial b_j^2} \end{pmatrix} \mid_{\boldsymbol{\eta}=\widehat{\boldsymbol{\eta}}_n} \\ &= \begin{pmatrix} \sum_{q=1}^{Q}(X_q - \widehat{b}_j)^2 \widehat{w}_{jq} & -\sum_{q=1}^{Q}(X_q - \widehat{b}_j)\widehat{a}_j \widehat{w}_{jq} \\ -\sum_{q=1}^{Q}(X_q - \widehat{b}_j)\widehat{a}_j \widehat{w}_{jq} & \sum_{q=1}^{Q}\widehat{a}_j^2 \widehat{w}_{jq} \end{pmatrix} = \sum_{q=1}^{Q}\widehat{w}_{jq}\widehat{\boldsymbol{z}}_q \widehat{\boldsymbol{z}}_q^T. \end{aligned} \tag{14}$$

where $\widehat{\boldsymbol{z}}_q = \left(X_q - \widehat{b}_j, -\widehat{a}_j\right)^T$. Let $\widehat{\boldsymbol{Z}} = (\widehat{\boldsymbol{z}}_1, \cdots, \widehat{\boldsymbol{z}}_Q)^T$, $\widehat{\boldsymbol{W}}_j = \mathrm{diag}(\widehat{w}_{j1}, \cdots, \widehat{w}_{jQ})$, we obtain $D^2\left(h_n(\widehat{\boldsymbol{\eta}}_n)\right) = \widehat{\boldsymbol{Z}}^T \widehat{\boldsymbol{W}}_j \widehat{\boldsymbol{Z}} \triangleq \boldsymbol{H}$. For any nonzero vector $\boldsymbol{x} = (x_1, x_2)^T$, we have

$$\boldsymbol{x}^T \boldsymbol{H} \boldsymbol{x} = \left(\widehat{\boldsymbol{Z}}\boldsymbol{x}\right)^T \widehat{\boldsymbol{W}}_j \widehat{\boldsymbol{Z}}\boldsymbol{x} = \boldsymbol{P}^T \widehat{\boldsymbol{W}}_j \boldsymbol{P} = \sum_{q=1}^{Q}\widehat{w}_{jq} p_q^2, \tag{15}$$

where $\boldsymbol{P} = \widehat{\boldsymbol{Z}}\boldsymbol{x} = (p_1, \cdots, p_Q)^T$. Moreover, for any q, it is evident that $\widehat{w}_{jq} = \widehat{\overline{f}}_{jq}\widehat{P}_j(X_q)\widehat{Q}_j(X_q) > 0$, i.e., the diagonal matrix $\widehat{\boldsymbol{W}}_j > 0$, and thus $\boldsymbol{x}^T \boldsymbol{H} \boldsymbol{x} = \boldsymbol{P}^T \boldsymbol{W}_j \boldsymbol{P} > 0$. Then, by the sufficient and

necessary condition for a matrix to be positive definite, the matrix $\boldsymbol{H} = D^2(h_n(\widehat{\boldsymbol{\eta}}_n))$ is confirmed to be a positive definite. Since the determinant of a positive definite matrix is always positive, it can be concluded that Assumption 3 (b) is satisfied for the 2PL model.

Here,we use the 2PL model as an example for verification, and the verification process for the M2PL model follows a similar approach. Thus, Assumption 3 (a) and (b) are also satisfied for the M2PL model.

S5.2. Proof of the theorems

Consider the function $d(\boldsymbol{A}, \boldsymbol{B})$, which is referred to as the Wasserstein metric in the fields of statistics and optimal transport theory. It is defined as follows:

$$d(\boldsymbol{A}, \boldsymbol{B}) = \sqrt{\mathrm{tr}(\boldsymbol{A} + \boldsymbol{B} - 2(\boldsymbol{A}^{\frac{1}{2}}\boldsymbol{B}\boldsymbol{A}^{\frac{1}{2}})^{\frac{1}{2}})}, \tag{16}$$

where both $\boldsymbol{A}$ and $\boldsymbol{B}$ are positive semi-definite matrices of the size $p \times p$. Bhatia et al. (2019, please see page 3) demonstrated that the function $d(\cdot, \cdot)$ establishes a metric in the space of positive semi-definite matrices. The term "Wasserstein mean" of K positive semi-definite matrices $\boldsymbol{A}_k$ (where $k = 1, \cdots, K$) denotes the Wasserstein barycenter corresponding to the variance-covariance matrix of the normal distributions $N(\boldsymbol{0}, \boldsymbol{A}_k)$, for $k = 1, \cdots, K$.

Lemma 1. Suppose $\boldsymbol{A}_k$, $k = 1, \cdots, K$ represents a sequence of $p \times p$ positive definite matrices, and their Wasserstein mean is signified by $\bar{\boldsymbol{A}}$. When considering another positive definite matrix A_0, we have

$$d(\bar{\boldsymbol{A}}, \boldsymbol{A}_0) \leq 2\sqrt{\frac{p}{K}\sum_{k=1}^{K}\|\boldsymbol{A}_k - \boldsymbol{A}_0\|} \leq 2\sqrt{\frac{p}{K}\sum_{k=1}^{K}\|\boldsymbol{A}_k - \boldsymbol{A}_0\|_F}. \tag{17}$$

where $\|\cdot\|$ signifies the operator norm, while $\|\cdot\|_F$ represents the Frobenius norm.

Proof. As per the definition of $\bar{\boldsymbol{A}}$, it is established that

$$\bar{\boldsymbol{A}} := \arg\min_{\boldsymbol{X}>0} \sum_{k=1}^{K} w_k d^2(\boldsymbol{X}, \boldsymbol{A}_k). \tag{18}$$

This implies that

$$\frac{1}{K}\sum_{k=1}^{K} d^2(\bar{\boldsymbol{A}}, \boldsymbol{A}_k) \leq \frac{1}{K}\sum_{k=1}^{K} d^2(\boldsymbol{A}_0, \boldsymbol{A}_k). \tag{19}$$

Applying the triangle inequality and the arithmetic mean-geometric mean (AM-GM) inequality, we derive that

$$\begin{aligned} d^2(\bar{\boldsymbol{A}}, \boldsymbol{A}_0) &\leq 2\left[\frac{1}{K}\sum_{k=1}^{K} d^2(\bar{\boldsymbol{A}}, \boldsymbol{A}_k) + \frac{1}{K}\sum_{k=1}^{K} d^2(\boldsymbol{A}_0, \boldsymbol{A}_k)\right] \\ &\leq \frac{4}{K}\sum_{k=1}^{K} d^2(\boldsymbol{A}_0, \boldsymbol{A}_k) \\ &\leq \frac{4}{K}\sum_{k=1}^{K} \|\boldsymbol{A}_0^{\frac{1}{2}} - \boldsymbol{A}_k^{\frac{1}{2}}\|_F^2 \quad \text{(by Theorem 1 of Bhatia et al. (2019))} \\ &\leq \frac{4p}{K}\sum_{k=1}^{K} \|\boldsymbol{A}_0^{\frac{1}{2}} - \boldsymbol{A}_k^{\frac{1}{2}}\|^2. \end{aligned} \tag{20}$$

Since the square root function demonstrates operator monotonicity, Theorem X.1.1 from Bhatia (2013) can consequently be implemented to the prior inequality, yielding

$$d(\bar{\boldsymbol{A}}, \boldsymbol{A}_0) \leq 2\sqrt{\frac{p}{K}\sum_{k=1}^{K} \|\boldsymbol{A}_k - \boldsymbol{A}_0\|}. \tag{21}$$

The final inequality presented in Equation (17) is established upon recognizing that the Frobenius norm serves as an upper bound for the operator norm. The details of this proof can also be found in Lemma 2 in Shyamalkumar and Srivastava (2022). □

Lemma 2. Consider two $p \times p$ positive semi-definite matrices, denoted as $\boldsymbol{A}$ and $\boldsymbol{B}$, we have

$$\|\boldsymbol{A} - \boldsymbol{B}\|_F \leq d(\boldsymbol{A}, \boldsymbol{B})\left(\sqrt{\mathrm{tr}(\boldsymbol{A})} + \sqrt{\mathrm{tr}(\boldsymbol{B})}\right). \tag{22}$$

Proof. Considering the $p \times p$ positive semi-definite matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, we can represent $\mathcal{F}(\boldsymbol{A})$ and $\mathcal{F}(\boldsymbol{B})$ as

$$\begin{aligned} \mathcal{F}(\boldsymbol{A}) &:= \{\boldsymbol{M}_{p\times p} : \boldsymbol{A} = \boldsymbol{M}\boldsymbol{M}^T\}, \\ \mathcal{F}(\boldsymbol{B}) &:= \{\boldsymbol{N}_{p\times p} : \boldsymbol{B} = \boldsymbol{N}\boldsymbol{N}^T\}. \end{aligned} \tag{23}$$

Subsequently, we find that

$$\begin{aligned} \|\boldsymbol{A} - \boldsymbol{B}\|_F &= \|\boldsymbol{M}\boldsymbol{M}^T - \boldsymbol{N}\boldsymbol{N}^T\|_F \\ &= \|\boldsymbol{M}\boldsymbol{M}^T - \boldsymbol{M}\boldsymbol{N}^T + \boldsymbol{M}\boldsymbol{N}^T - \boldsymbol{N}\boldsymbol{N}^T\|_F \\ &\leq \|\boldsymbol{M}\|_F\|\boldsymbol{M}^T - \boldsymbol{N}^T\|_F + \|\boldsymbol{N}^T\|_F\|\boldsymbol{M} - \boldsymbol{N}\|_F \\ &= \|\boldsymbol{M} - \boldsymbol{N}\|_F(\|\boldsymbol{M}\|_F + \|\boldsymbol{N}\|_F) \\ &= \|\boldsymbol{M} - \boldsymbol{N}\|_F \left(\sqrt{\mathrm{tr}(\boldsymbol{A})} + \sqrt{\mathrm{tr}(\boldsymbol{B})}\right). \end{aligned} \tag{24}$$

Based on Theorem 1 of Bhatia et al (2019), we have

$$d(\boldsymbol{A}, \boldsymbol{B}) = \min_{\boldsymbol{M}\in\mathcal{F}(\boldsymbol{A});\boldsymbol{N}\in\mathcal{F}(\boldsymbol{B})} \|\boldsymbol{M} - \boldsymbol{N}\|_F. \tag{25}$$

Hence, we concluded

$$\begin{aligned} \|\boldsymbol{A} - \boldsymbol{B}\|_F &\leq \left(\sqrt{\mathrm{tr}(\boldsymbol{A})} + \sqrt{\mathrm{tr}(\boldsymbol{B})}\right) \min_{\boldsymbol{M}\in\mathcal{F}(\boldsymbol{A});\boldsymbol{N}\in\mathcal{F}(\boldsymbol{B})} \|\boldsymbol{M} - \boldsymbol{N}\|_F \\ &= d(\boldsymbol{A}, \boldsymbol{B}) \left(\sqrt{\mathrm{tr}(\boldsymbol{A})} + \sqrt{\mathrm{tr}(\boldsymbol{B})}\right). \end{aligned} \tag{26}$$

The details of this proof can also be found in Lemma 3 in Shyamalkumar and Srivastava (2022).

□

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

$$W_2^2(\pi, \widetilde{\pi}) = O_p(s^{-2}) + O_p(n^{-\frac{3}{2}}), \tag{27}$$

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

Proof. We first define some necessary notations for the proof. We denote n and s as the sample sizes of the full data and subset, respectively. Let π, $\pi_{(k)}$ and $\widetilde{\pi}$ represent the full data posterior, the kth subset posterior, and the WASP posterior, respectively. Given Assumption 2 and Corollary 1, we establish that

$$\begin{aligned} W_2^2(\pi, \widetilde{\pi}) &= \|\boldsymbol{\mu} - \widetilde{\boldsymbol{\mu}}\|_2^2 + \mathrm{tr}(\boldsymbol{\Sigma} + \widetilde{\boldsymbol{\Sigma}} - 2(\widetilde{\boldsymbol{\Sigma}}^{\frac{1}{2}}\boldsymbol{\Sigma}\widetilde{\boldsymbol{\Sigma}}^{\frac{1}{2}})^{\frac{1}{2}}) \\ &= \|\boldsymbol{\mu} - \widetilde{\boldsymbol{\mu}}\|_2^2 + d^2(\boldsymbol{\Sigma}, \widetilde{\boldsymbol{\Sigma}}), \end{aligned} \tag{28}$$

where $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ and $\widetilde{\boldsymbol{\mu}}$, $\widetilde{\boldsymbol{\Sigma}}$ denote the means and covariance matrices of π and $\widetilde{\pi}$ respectively. We denote $\boldsymbol{\mu}_{(k)}$ and $\boldsymbol{\Sigma}_{(k)}$ as the mean and covariance matrix of the kth subset posterior $\pi_{(k)}$, respectively. Assume $\boldsymbol{I}_{\eta^*}$ be the Fisher information matrix under the truth value of the parameter $\boldsymbol{\eta}^*$. The maximum likelihood estimators (MLE) for the full data and the kth subset are defined as $\widehat{\boldsymbol{\eta}}_{MLE}$ and $\widehat{\boldsymbol{\eta}}_{MLE}^{(k)}$, respectively. The Fisher information matrices for the full data and subset, under $\widehat{\boldsymbol{\eta}}_{MLE}$ and $\widehat{\boldsymbol{\eta}}_{MLE}^{(k)}$, can be written as

$$\widehat{\boldsymbol{I}} = -\frac{1}{n}\frac{\partial^2 \log f(\boldsymbol{y}|\boldsymbol{\eta})}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}^T}\bigg|_{\boldsymbol{\eta}=\widehat{\boldsymbol{\eta}}_{MLE}}, \quad \widehat{\boldsymbol{I}}_{(k)} = -\frac{1}{s}\frac{\partial^2 \log f(\boldsymbol{y^{(k)}}|\boldsymbol{\eta})}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}^T}\bigg|_{\boldsymbol{\eta}=\widehat{\boldsymbol{\eta}}_{MLE}^{(k)}}. \tag{29}$$

First, we focus on the first term of Equation (28) on the right side. Based on Kass et al.

(1990), $\mu_{(k)}$ and $\widehat{\boldsymbol{\eta}}^{(k)}_{MLE}$ would unfold in the following manner under Assumption 3

$$\begin{aligned}\boldsymbol{\mu}_{(k)} &= \widehat{\boldsymbol{\eta}}^{(k)}_{MLE} + \frac{\boldsymbol{\gamma}_{(k)}}{n} + O(n^{-2}),\\ \widehat{\boldsymbol{\eta}}^{(k)}_{MLE} &= \boldsymbol{\eta}^* + \frac{\boldsymbol{\delta}_{(k)}}{\sqrt{s}} + O_p(s^{-1}),\end{aligned} \tag{30}$$

where

$$\begin{aligned}\boldsymbol{\gamma}_{(k)} &= \widehat{\boldsymbol{I}}^{-1}_{(k)}\left[\frac{\partial \log g(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}}\bigg|_{\boldsymbol{\eta}=\widehat{\boldsymbol{\eta}}^{(k)}_{MLE}} - \frac{1}{2}\left(-\frac{1}{s}\frac{\partial^3 \log f(\boldsymbol{y^{(k)}}|\boldsymbol{\eta})}{\partial \boldsymbol{\eta}\partial \boldsymbol{\eta}^T \partial \boldsymbol{\eta}}\bigg|_{\boldsymbol{\eta}=\widehat{\boldsymbol{\eta}}^{(k)}_{MLE}}\right)\widehat{\boldsymbol{I}}^{-1}_{(k)}\right],\\ \boldsymbol{\delta}_{(k)} &= \frac{1}{\sqrt{s}}\boldsymbol{I}^{-1}_{\eta^*}\sum_{i=1}^{s}\frac{\partial \log f(\boldsymbol{y}^{(k)}_i|\boldsymbol{\eta}^*)}{\partial \boldsymbol{\eta}}.\end{aligned} \tag{31}$$

Consequently, we represent the mean of the WASP posterior $\widetilde{\pi}$ as follows

$$\begin{aligned}\widetilde{\boldsymbol{\mu}} &= \frac{1}{K}\sum_{k=1}^{K}\boldsymbol{\mu}_{(k)}\\ &= \boldsymbol{\eta}^* + \frac{1}{K}\sum_{k=1}^{K}\frac{\boldsymbol{\gamma}_{(k)}}{n} + \frac{1}{K}\sum_{k=1}^{K}\frac{\boldsymbol{\delta}_{(k)}}{\sqrt{s}} + O(n^{-2}) + O_p(s^{-1}).\end{aligned} \tag{32}$$

Similarly, we derive the mean of the full data posterior in the following manner

$$\boldsymbol{\mu} = \boldsymbol{\eta}^* + \frac{\boldsymbol{\gamma}}{n} + \frac{\boldsymbol{\delta}}{\sqrt{n}} + O(n^{-2}) + O_p(n^{-1}), \tag{33}$$

where

$$\begin{aligned}\boldsymbol{\gamma} &= \widehat{\boldsymbol{I}}^{-1}\left[\frac{\partial \log g(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}}\bigg|_{\boldsymbol{\eta}=\widehat{\boldsymbol{\eta}}_{MLE}} - \frac{1}{2}\left(-\frac{1}{n}\frac{\partial^3 \log f(\boldsymbol{y}|\boldsymbol{\eta})}{\partial \boldsymbol{\eta}\partial \boldsymbol{\eta}^T \partial \boldsymbol{\eta}}\bigg|_{\boldsymbol{\eta}=\widehat{\boldsymbol{\eta}}_{MLE}}\right)\widehat{\boldsymbol{I}}^{-1}\right],\\ \boldsymbol{\delta} &= \frac{1}{\sqrt{n}}\boldsymbol{I}^{-1}_{\eta^*}\sum_{i=1}^{n}\frac{\partial \log f(\boldsymbol{y}_i|\boldsymbol{\eta}^*)}{\partial \boldsymbol{\eta}}.\end{aligned} \tag{34}$$

Taking into account that $\frac{\boldsymbol{\delta}}{\sqrt{n}} = \frac{1}{K}\sum_{k=1}^{K}\frac{\boldsymbol{\delta}_{(k)}}{\sqrt{s}}$, $\boldsymbol{\gamma}_{(k)}$ and $\boldsymbol{\gamma}$ are $O(1)$, along with the condition $s < n$,

we have

$$\begin{aligned}\widetilde{\boldsymbol{\mu}} - \boldsymbol{\mu} &= \frac{1}{K}\sum_{k=1}^{K}\frac{\boldsymbol{\gamma}_{(k)}}{n} - \frac{\boldsymbol{\gamma}}{n} + O_p(s^{-1}) + O(n^{-2}) + O_p(n^{-1}) \\ &= O_p(s^{-1}) + O(n^{-2}) + O_p(n^{-1}) \\ &= O_p(s^{-1}). \end{aligned} \tag{35}$$

Therefore

$$\|\widetilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_2^2 = O_p(s^{-2}) \text{ in } P_{\boldsymbol{\eta}^*}^n\text{-probability.} \tag{36}$$

Next, we focus on the second term on the right within $W_2^2(\pi, \widetilde{\pi})$. In accordance with Assumption 3 presented in this paper and Theorem 4 from Kass et al. (1990), we derive the Laplace approximation for the covariance matrix related to the full data posterior and the subset posterior as follows

$$\begin{aligned}\boldsymbol{\Sigma} &= \frac{1}{n}\widehat{\boldsymbol{I}}^{-1} + O(n^{-2}), \\ \boldsymbol{\Sigma}_{(k)} &= \frac{1}{n}\widehat{\boldsymbol{I}}_{(k)}^{-1} + O(n^{-2}). \end{aligned} \tag{37}$$

According to the consistency in the maximum likelihood estimation, we have

$$\widehat{\boldsymbol{I}} = -\frac{1}{n}\frac{\partial^2 \log f(\boldsymbol{y}|\boldsymbol{\eta})}{\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}^T}\bigg|_{\boldsymbol{\eta}=\widehat{\boldsymbol{\eta}}_{MLE}} = -\frac{1}{n}\frac{\partial^2 \log f(\boldsymbol{y}|\boldsymbol{\eta})}{\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}^T}\bigg|_{\boldsymbol{\eta}=\boldsymbol{\eta}^*} + O_p(n^{-\frac{1}{2}}) = \boldsymbol{I}_{\eta^*} + O_p(n^{-\frac{1}{2}}), \tag{38}$$

where $\boldsymbol{\eta}^*$ indicates the true value of the parameter $\boldsymbol{\eta}$. Similarly, we have $\widehat{\boldsymbol{I}}_{(k)} = \boldsymbol{I}_{\eta^*} + O_p(n^{-\frac{1}{2}})$. Since the matrix inversion acts as a continuous operator within the subspace of reversible matrices, we have

$$\begin{aligned}\widehat{\boldsymbol{I}}^{-1} &= \boldsymbol{I}_{\eta^*}^{-1} + O_p(n^{-\frac{1}{2}}), \\ \widehat{\boldsymbol{I}}_{(k)}^{-1} &= \boldsymbol{I}_{\eta^*}^{-1} + O_p(n^{-\frac{1}{2}}). \end{aligned} \tag{39}$$

By integrating Equation (39) into Equation (37), we have

$$
\begin{aligned}
n\boldsymbol{\Sigma} - \boldsymbol{I}_{\eta^*}^{-1} &= O_p(n^{-\frac{1}{2}}), \\
n\boldsymbol{\Sigma}_{(k)} - \boldsymbol{I}_{\eta^*}^{-1} &= O_p(n^{-\frac{1}{2}}).
\end{aligned} \tag{40}
$$

Additionally, we find that $\boldsymbol{\Sigma}_{(k)} - \boldsymbol{\Sigma} = O_p(n^{-\frac{3}{2}})$. Finally, by invoking Lemma 1, we have

$$
d^2(\widetilde{\boldsymbol{\Sigma}}, \boldsymbol{\Sigma}) \leq \frac{4J}{K} \sum_{k=1}^{K} \|\boldsymbol{\Sigma}_{(k)} - \boldsymbol{\Sigma}\|_F = O_p(n^{-\frac{3}{2}}). \tag{41}
$$

Therefore, when $n, s \to \infty$, it follows that $W_2^2(\pi, \widetilde{\pi}) = O_p(s^{-2}) + O_p(n^{-\frac{3}{2}})$ in $P_{\boldsymbol{\eta}^*}^n$-probability. This completes the proof of the Theorem 2. □

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

$$
W_2^2(\widetilde{\pi}, \widehat{\widetilde{\pi}}) = O_p(M^{-1}) + o_p(n^{-1}). \tag{42}
$$

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

Proof. As stated in section 4.3 of this paper, the KM posterior drawing pertaining to $\widetilde{\pi}$ and the empirical measure, $\widehat{\widetilde{\pi}}$, for $k = 1, \cdots, K$ and $m = 1, \cdots, M$, can be expressed as

$$
\begin{aligned}
\widetilde{\boldsymbol{\eta}} &= \widetilde{\boldsymbol{\mu}} + \widetilde{\boldsymbol{\Sigma}}^{\frac{1}{2}} \boldsymbol{\xi}_{(k)}^{(m)}, \\
\widehat{\widetilde{\boldsymbol{\eta}}} &= \widehat{\widetilde{\boldsymbol{\mu}}} + \widehat{\widetilde{\boldsymbol{\Sigma}}}^{\frac{1}{2}} \widehat{\boldsymbol{\Sigma}}_{(k)}^{-\frac{1}{2}} \big(\boldsymbol{\mu}_{(k)} - \widehat{\boldsymbol{\mu}}_{(k)}\big) + \widehat{\widetilde{\boldsymbol{\Sigma}}}^{\frac{1}{2}} \widehat{\boldsymbol{\Sigma}}_{(k)}^{-\frac{1}{2}} \boldsymbol{\Sigma}_{(k)}^{\frac{1}{2}} \boldsymbol{\xi}_{(k)}^{(m)}.
\end{aligned} \tag{43}
$$

Consequently, given the definition of the Wasserstein distance, it follows that

$$
\begin{aligned}
W_2^2(\widetilde{\pi}, \widehat{\widetilde{\pi}}) &\leq \frac{1}{KM} \sum_{k=1}^{K} \sum_{m=1}^{M} \|\widehat{\widetilde{\boldsymbol{\eta}}} - \widetilde{\boldsymbol{\eta}}\|_2^2 \\
&\leq \frac{1}{KM} \sum_{k=1}^{K} \sum_{m=1}^{M} \|\widehat{\widetilde{\boldsymbol{\mu}}} - \widetilde{\boldsymbol{\mu}} + \widehat{\widetilde{\boldsymbol{\Sigma}}}^{\frac{1}{2}} \widehat{\boldsymbol{\Sigma}}_{(k)}^{-\frac{1}{2}} \big(\boldsymbol{\mu}_{(k)} - \widehat{\boldsymbol{\mu}}_{(k)}\big) + \widehat{\widetilde{\boldsymbol{\Sigma}}}^{\frac{1}{2}} \widehat{\boldsymbol{\Sigma}}_{(k)}^{-\frac{1}{2}} \boldsymbol{\Sigma}_{(k)}^{\frac{1}{2}} \boldsymbol{\xi}_{(k)}^{(m)} - \widetilde{\boldsymbol{\Sigma}}^{\frac{1}{2}} \boldsymbol{\xi}_{(k)}^{(m)}\|_2^2 \\
&\leq \frac{2}{K} \sum_{k=1}^{K} \|\widehat{\widetilde{\boldsymbol{\mu}}} - \widetilde{\boldsymbol{\mu}} + \widehat{\widetilde{\boldsymbol{\Sigma}}}^{\frac{1}{2}} \widehat{\boldsymbol{\Sigma}}_{(k)}^{-\frac{1}{2}} \big(\boldsymbol{\mu}_{(k)} - \widehat{\boldsymbol{\mu}}_{(k)}\big)\|_2^2 + \frac{2}{KM} \sum_{k,m} \|\big(\widehat{\widetilde{\boldsymbol{\Sigma}}}^{\frac{1}{2}} \widehat{\boldsymbol{\Sigma}}_{(k)}^{-\frac{1}{2}} \boldsymbol{\Sigma}_{(k)}^{\frac{1}{2}} - \widetilde{\boldsymbol{\Sigma}}^{\frac{1}{2}}\big) \boldsymbol{\xi}_{(k)}^{(m)}\|_2^2 \\
&\leq 4\|\widehat{\widetilde{\boldsymbol{\mu}}} - \widetilde{\boldsymbol{\mu}}\|_2^2 + \frac{4}{K} \sum_{k=1}^{K} \|\widehat{\widetilde{\boldsymbol{\Sigma}}}^{\frac{1}{2}} \widehat{\boldsymbol{\Sigma}}_{(k)}^{-\frac{1}{2}} \big(\boldsymbol{\mu}_{(k)} - \widehat{\boldsymbol{\mu}}_{(k)}\big)\|_2^2 + 2 \max_{1 \leq k \leq K} \|\boldsymbol{D}_k\|_2^2 \frac{1}{KM} \sum_{k,m} \|\boldsymbol{\xi}_{(k)}^{(m)}\|_2^2, \quad (44)
\end{aligned}
$$

where $\boldsymbol{D}_k = \widehat{\widetilde{\boldsymbol{\Sigma}}}^{\frac{1}{2}} \widehat{\boldsymbol{\Sigma}}_{(k)}^{-\frac{1}{2}} \boldsymbol{\Sigma}_{(k)}^{\frac{1}{2}} - \widetilde{\boldsymbol{\Sigma}}^{\frac{1}{2}}$.

To start with, we focus on the first term in Equation (44). According to Assumption 5, we have

$$
\|\widehat{\widetilde{\boldsymbol{\mu}}} - \widetilde{\boldsymbol{\mu}}\|_2^2 = \|\frac{1}{K} \sum_{k=1}^{K} \big(\widehat{\boldsymbol{\mu}}_{(k)} - \boldsymbol{\mu}_{(k)}\big)\|_2^2 \leq \frac{1}{K} \sum_{k=1}^{K} \|\widehat{\boldsymbol{\mu}}_{(k)} - \boldsymbol{\mu}_{(k)}\|_2^2 = O_p(M^{-1}). \quad (45)
$$

Then, we focus on the second term in Equation (44). From Theorem 9 of Bhatia et al. (2019), we have $\widehat{\widetilde{\boldsymbol{\Sigma}}} \leq \frac{1}{K} \sum_{k=1}^{K} \widehat{\boldsymbol{\Sigma}}_{(k)}$. Consequently, it follows that

$$
\begin{aligned}
\|\widehat{\widetilde{\boldsymbol{\Sigma}}}^{\frac{1}{2}} \widehat{\boldsymbol{\Sigma}}_{(k)}^{-\frac{1}{2}}\|_2^2 &\leq \|\big(\frac{1}{K} \sum_{k=1}^{K} n\widehat{\boldsymbol{\Sigma}}_{(k)}\big)^{\frac{1}{2}} \big(n\widehat{\boldsymbol{\Sigma}}_{(k)}\big)^{-\frac{1}{2}}\|_2^2 \\
&\leq \sqrt{\|\big(\frac{1}{K} \sum_{k=1}^{K} n\widehat{\boldsymbol{\Sigma}}_{(k)}\big) \big(n\widehat{\boldsymbol{\Sigma}}_{(k)}\big)^{-1}\|_2^2} \\
&\leq \sqrt{\frac{1}{K} \sum_{k=1}^{K} \|n\widehat{\boldsymbol{\Sigma}}_{(k)}\|_2^2 \|\big(n\widehat{\boldsymbol{\Sigma}}_{(k)}\big)^{-1}\|_2^2} \\
&= O_p(1). \quad (46)
\end{aligned}
$$

In conjunction with Assumption 5, we have

$$\|\widehat{\widetilde{\boldsymbol{\Sigma}}}^{\frac{1}{2}}\widehat{\boldsymbol{\Sigma}}_{(k)}^{-\frac{1}{2}}\big(\boldsymbol{\mu}_{(k)}-\widehat{\boldsymbol{\mu}}_{(k)}\big)\|_2^2 \leq \|\widehat{\widetilde{\boldsymbol{\Sigma}}}^{\frac{1}{2}}\widehat{\boldsymbol{\Sigma}}_{(k)}^{-\frac{1}{2}}\|_2^2\|\boldsymbol{\mu}_{(k)}-\widehat{\boldsymbol{\mu}}_{(k)}\|_2^2 = O_p(M^{-1}). \tag{47}$$

For the last term in Equation (44), break down D_k to yield

$$\begin{aligned}
\|D_k\|_2 &= \|\widehat{\widetilde{\boldsymbol{\Sigma}}}^{\frac{1}{2}}\widehat{\boldsymbol{\Sigma}}_{(k)}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{(k)}^{\frac{1}{2}}-\widetilde{\boldsymbol{\Sigma}}^{\frac{1}{2}}\|_2 \\
&= \|n^{-\frac{1}{2}}\big[\big((n\widehat{\widetilde{\boldsymbol{\Sigma}}})^{\frac{1}{2}}-\boldsymbol{I}_{\eta^*}^{-\frac{1}{2}}+\boldsymbol{I}_{\eta^*}^{-\frac{1}{2}}\big)\widehat{\boldsymbol{\Sigma}}_{(k)}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{(k)}^{\frac{1}{2}}-\big((n\widetilde{\boldsymbol{\Sigma}})^{\frac{1}{2}}-\boldsymbol{I}_{\eta^*}^{-\frac{1}{2}}+\boldsymbol{I}_{\eta^*}^{-\frac{1}{2}}\big)\big]\|_2 \\
&\leq n^{-\frac{1}{2}}\big[\|((n\widehat{\widetilde{\boldsymbol{\Sigma}}})^{\frac{1}{2}}-\boldsymbol{I}_{\eta^*}^{-\frac{1}{2}})\widehat{\boldsymbol{\Sigma}}_{(k)}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{(k)}^{\frac{1}{2}}\|_2+\|\boldsymbol{I}_{\eta^*}^{-\frac{1}{2}}(\widehat{\boldsymbol{\Sigma}}_{(k)}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{(k)}^{\frac{1}{2}}-\boldsymbol{I})\|_2+\|((n\widetilde{\boldsymbol{\Sigma}})^{\frac{1}{2}}-\boldsymbol{I}_{\eta^*}^{-\frac{1}{2}})\|_2\big],
\end{aligned} \tag{48}$$

where $\boldsymbol{I}$ represents the identity matrix. For the final term $(n\widetilde{\boldsymbol{\Sigma}})^{\frac{1}{2}}-\boldsymbol{I}_{\eta^*}^{-\frac{1}{2}}$ in Equation (48), using Lemma 1, Lemma 2, and Equation (40), we have

$$\begin{aligned}
\|n\widetilde{\boldsymbol{\Sigma}}-\boldsymbol{I}_{\eta^*}^{-1}\|_2 &\leq \|n\widetilde{\boldsymbol{\Sigma}}-\boldsymbol{I}_{\eta^*}^{-1}\|_F \\
&\leq d(n\widetilde{\boldsymbol{\Sigma}},\boldsymbol{I}_{\eta^*}^{-1})\left[\sqrt{\mathrm{tr}(n\widetilde{\boldsymbol{\Sigma}})}+\sqrt{\mathrm{tr}(\boldsymbol{I}_{\eta^*}^{-1})}\right] \\
&\leq 2\sqrt{\frac{J}{K}\sum_{k=1}^{K}\|n\boldsymbol{\Sigma}_{(k)}-\boldsymbol{I}_{\eta^*}^{-1}\|_F}\left[\sqrt{\frac{1}{K}\sum_{k=1}^{K}\mathrm{tr}(n\boldsymbol{\Sigma}_{(k)}-\boldsymbol{I}_{\eta^*}^{-1}+\boldsymbol{I}_{\eta^*}^{-1})}+\sqrt{\mathrm{tr}(\boldsymbol{I}_{\eta^*}^{-1})}\right] \\
&\leq 2\sqrt{\frac{J}{K}\sum_{k=1}^{K}\|n\boldsymbol{\Sigma}_{(k)}-\boldsymbol{I}_{\eta^*}^{-1}\|_F}\left[\sqrt{\frac{1}{K}\sum_{k=1}^{K}\mathrm{tr}(n\boldsymbol{\Sigma}_{(k)}-\boldsymbol{I}_{\eta^*}^{-1})+\frac{1}{K}\sum_{k=1}^{K}\mathrm{tr}(\boldsymbol{I}_{\eta^*}^{-1})}+\sqrt{\mathrm{tr}(\boldsymbol{I}_{\eta^*}^{-1})}\right] \\
&= o_p(1)\times O_p(1) \\
&= o_p(1),
\end{aligned} \tag{49}$$

where $\|\cdot\|_F$ represents the Frobenius norm. Thus, we have $\|(n\widetilde{\boldsymbol{\Sigma}})^{\frac{1}{2}}-\boldsymbol{I}_{\eta^*}^{-\frac{1}{2}}\|_2 \leq \sqrt{\|n\widetilde{\boldsymbol{\Sigma}}-\boldsymbol{I}_{\eta^*}^{-1}\|_2} = o_p(1)$. Similarly, we establish that $\|(n\widehat{\widetilde{\boldsymbol{\Sigma}}})^{\frac{1}{2}}-\boldsymbol{I}_{\eta^*}^{-\frac{1}{2}}\|_2 = o_p(1)$. Hence, the first term of Equation (48) can be obtained as follows

$$\|\big((n\widehat{\widetilde{\boldsymbol{\Sigma}}})^{\frac{1}{2}}-\boldsymbol{I}_{\eta^*}^{-\frac{1}{2}}\big)\widehat{\boldsymbol{\Sigma}}_{(k)}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{(k)}^{\frac{1}{2}}\|_2 \leq \sqrt{\|(n\widehat{\widetilde{\boldsymbol{\Sigma}}}-\boldsymbol{I}_{\eta^*}^{-1})\|_2}\|\widehat{\boldsymbol{\Sigma}}_{(k)}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{(k)}^{\frac{1}{2}}\|_2 = o_p(1). \tag{50}$$

For the term $\widehat{\boldsymbol{\Sigma}}_{(k)}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{(k)}^{\frac{1}{2}} - \boldsymbol{I}$, based on Assumption 5, we have

$$\begin{aligned}
\|\widehat{\boldsymbol{\Sigma}}_{(k)}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{(k)}^{\frac{1}{2}} - \boldsymbol{I}\|_2 &= \|\widehat{\boldsymbol{\Sigma}}_{(k)}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{(k)}^{\frac{1}{2}} - \widehat{\boldsymbol{\Sigma}}_{(k)}^{-\frac{1}{2}}\widehat{\boldsymbol{\Sigma}}_{(k)}^{\frac{1}{2}}\|_2 \\
&\leq \sqrt{\|\widehat{\boldsymbol{\Sigma}}_{(k)}^{-1}\|_2}\|\boldsymbol{\Sigma}_{(k)}^{\frac{1}{2}} - \widehat{\boldsymbol{\Sigma}}_{(k)}^{\frac{1}{2}}\|_2 \\
&\leq \sqrt{\|(n\widehat{\boldsymbol{\Sigma}}_{(k)})^{-1}\|_2}\sqrt{\|n\boldsymbol{\Sigma}_{(k)} - n\widehat{\boldsymbol{\Sigma}}_{(k)}\|_F} \\
&= o_p(1). \qquad (51)
\end{aligned}$$

Therefore, we derive

$$\begin{aligned}
\|D_k\|_2 &\leq n^{-\frac{1}{2}}\left[\|((n\widehat{\widetilde{\boldsymbol{\Sigma}}})^{\frac{1}{2}} - \boldsymbol{I}_{\eta^*}^{-\frac{1}{2}})\widehat{\boldsymbol{\Sigma}}_{(k)}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{(k)}^{\frac{1}{2}}\|_2 + \|\boldsymbol{I}_{\eta^*}^{-\frac{1}{2}}(\widehat{\boldsymbol{\Sigma}}_{(k)}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{(k)}^{\frac{1}{2}} - \boldsymbol{I})\|_2 + \|((n\widetilde{\boldsymbol{\Sigma}})^{\frac{1}{2}} - \boldsymbol{I}_{\eta^*}^{-\frac{1}{2}})\|_2\right] \\
&= o_p(n^{-\frac{1}{2}}). \qquad (52)
\end{aligned}$$

In other words, $\|D_k\|_2^2 = o_p(n^{-1})$. Finally, integrating Equations (45), (47) and (52), we have

$$W_2^2(\widetilde{\pi}, \widehat{\widetilde{\pi}}) = O_p(M^{-1}) + o_p(n^{-1}) \text{ in } P^K\text{-probability}, \qquad (53)$$

where $P^K = P_1 \otimes \cdots \otimes P_K$ is the probability measure on the Markov chain generated by the posterior distribution of all subsets. This completes the proof of the Theorem 3. □

References

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

Bhatia, R. (2013). *Matrix analysis* (Vol. 169). Cham: Springer Science & Business Media.

Bhatia, R., Jain, T., & Lim, Y. (2019). On the Bures−Wasserstein distance between positive definite matrices. *Expositiones Mathematicae*, *37*(2), 165−191.

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

Harwell, M. R., Baker, F. B., & Zwarts, M. (1988). Item parameter estimation via marginal maximum likelihood and an EM algorithm: A didactic. *Journal of Educational Statistics, 13*(3), 243−271.

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.

Man, K., Harring, J. R., Jiao, H., & Zhan, P. (2019). Joint modeling of compensatory multidimensional item responses and response times. *Applied Psychological Measurement*, *43*(8), 639−654.

OECD (2018). *PISA 2015 technical report*. Paris: OECD Publishing.

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.

Shyamalkumar, N. D., & Srivastava, S. (2022). An algorithm for distributed Bayesian inference. *Stat*, *11*(1), e432.