METHODOLOGY
European Journal of Research Methods for the Behavioral and Social Sciences

PsychOpen GOLD

Original Article

Analyzing Group Differences and Measurement Fairness in Process Data: A Sequential Response Model With Covariates

Yuting Han [1,2,3], Feng Ji [4], Yunxiao Chen [5], Kaiyu Gan [6], Hongyun Liu [6]

[1] *Cognitive Science and Allied Health School, Beijing Language and Culture University, Beijing, China.* **[2]** *Institute of Life and Health Sciences, Beijing Language and Culture University, Beijing, China.* **[3]** *Key Laboratory of Language and Cognitive Science (Ministry of Education), Beijing Language and Culture University, Beijing, China.* **[4]** *Department of Applied Psychology and Human Development, University of Toronto, Toronto, ON, Canada.* **[5]** *Department of Statistics, London School of Economics and Political Science, London, United Kingdom.* **[6]** *Beijing Key Laboratory of Applied Experimental Psychology, National Demonstration Center for Experimental Psychology Education (Beijing Normal University), Faculty of Psychology, Beijing Normal University, Beijing, China.*

Methodology, 2026, Vol. 22(1), 1–26, https://doi.org/10.5964/meth.16999

Received: 2025-02-11 • **Accepted:** 2025-11-04 • **Published (VoR):** 2026-03-27

Handling Editor: Pablo Nájera Álvarez, Universidad Pontificia Comillas, Madrid, Spain

Corresponding Author: Hongyun Liu, Faculty of Psychology, Beijing Normal University, No. 19, XinJieKouWai St., HaiDian District, Beijing, People's Republic of China. E-mail: hyliu@bnu.edu.cn

Supplementary Materials: Code, Data, Materials [see Index of Supplementary Materials]

Abstract

This article introduces the sequential response model with covariates (SRM-C) for analyzing process data, with emphasis on three key capabilities: detecting potential measurement bias in response processes, evaluating group differences in ability distributions and improving parameter estimation precision. The SRM-C combines measurement and structural components, with the measurement component modeling response sequences conditional on abilities and covariates, and the structural component characterizing group-specific ability distributions. Sparsity assumptions implemented through horseshoe prior distributions address identification issues within the Bayesian framework. Monte Carlo simulations demonstrated robust parameter recovery and effective differential item functioning (DIF) detection. An empirical analysis of PISA problem-solving data illustrated the model's utility in distinguishing ability differences from potential measurement bias. The SRM-C offers a comprehensive framework for understanding group differences in process data while ensuring measurement fairness.

Keywords

computer-based assessment, process data, differential item functioning, measurement invariance, Bayesian regularization

Computer-based assessments (CBAs) have emerged as a transformative approach in psychological and educational measurement, particularly for evaluating higher-order cognitive skills (Liu et al., 2018; Shute & Moore, 2017). The adoption of CBAs has been driven by their capacity to provide more authentic assessment contexts while reducing test anxiety through enhanced engagement (Banfield & Wilkerson, 2014; Li et al., 2015). This shift toward computer-based evaluation is evidenced by the implementation of interactive assessment systems in major international educational initiatives, including the Programme for International Student Assessment (PISA) in 2012, 2015, and 2018 (OECD, 2014, 2017), the Programme for International Assessment of Adult Competencies (PIAAC) in 2012 (Goodman et al., 2013; OECD, 2016; Schleicher, 2008), the Assessment and Teaching of 21st Century Skills project (ATC21S) (Griffin et al., 2012), and the National Assessment of Education Progress (NAEP) (National Center for Education Statistics, 2014).

CBAs generate rich process data that enables comprehensive assessment validation, enhanced measurement precision, and detailed analysis of response patterns, group differences, and behavioral patterns (Ercikan & Pellegrino, 2017; Mislevy et al., 2014). Researchers have leveraged this process data to evaluate problem-solving abilities through various methodological approaches (Hesse et al., 2015; Siddiq et al., 2017; Xiao et al., 2022). In tasks with finite state spaces, response sequences can be modeled as stochastic processes, where psychometric models incorporate both frequency and correctness of state transitions to estimate underlying abilities (Chen, 2020; Fu et al., 2023; Han et al., 2022; Han & Wilson, 2022; LaMar, 2018; Shu et al., 2017; Xiao & Liu, 2024; Zhan & Qiao, 2022). These approaches offer interpretable ability estimates while utilizing the complete response process.

Test fairness, particularly differential item functioning (DIF), is a critical consideration in educational and psychological assessment. DIF occurs when examinees of comparable ability levels from different demographic groups exhibit systematic differences in their item responses (American Educational Research Association et al., 2014). The methodological framework for DIF detection encompasses diverse approaches, including the Mantel-Haenszel procedure (Holland & Thayer, 1988), item response theory-based methods (Lord, 1980; Raju, 1988), logistic regression techniques (Swaminathan & Rogers, 1990), likelihood ratio tests (Thissen et al., 1993), and graphical procedures (Magis et al., 2010; Yuan et al., 2021). More recently, LASSO-type regularized estimation procedures have been developed to address model selection and parameter estimation simultaneously (Bauer et al., 2020; Belzak & Bauer, 2020; Huang, 2018; Magis et al., 2015; Schauberger & Mair, 2020; Tutz & Schauberger, 2015). Within the Bayesian paradigm, regularization effects comparable to LASSO have been achieved through specialized prior distributions,

including the Laplace prior (Casella et al., 2010; Park & Casella, 2008), spike-and-slab prior (Mitchell & Beauchamp, 1988), and horseshoe prior (Carvalho et al., 2009, 2010; Piironen & Vehtari, 2017; Polson & Scott, 2011), enabling anchor-free DIF analysis.

While traditional DIF detection methods analyze item responses, CBAs present unique challenges for fairness evaluation due to their complex process data structure. Unlike conventional test responses, process data typically consists of variable-length action sequences, necessitating more sophisticated analytical approaches for identifying potential DIF. Currently, the DIF detecting methods that can be used with process data are still limited to our knowledge, particularly regarding the simultaneous estimation of latent abilities and detection of DIF while controlling for confounding. To address these methodological challenges, we propose the sequential response model with covariates (SRM-C), which makes several methodological contributions that distinguish it from existing approaches. First, unlike traditional DIF methods that analyze discrete item responses, the SRM-C detects differential functioning in sequential state transitions, accommodating the variable-length action sequences characteristic of process data. Second, the model simultaneously estimates group-specific ability distributions while controlling for potential measurement bias in transition processes, enabling separation of genuine ability differences from DIF effects. Third, the SRM-C implements anchor-free DIF detection through horseshoe priors adapted for transition matrices, reflecting the assumption that most state transitions are DIF-free. These innovations enable three key analytical capabilities: detecting potential measurement bias in response processes, evaluating group differences in ability distributions, and improving parameter estimation precision.

This article proceeds as follows: We first introduce the SRM-C and detail its Bayesian estimation framework. We then evaluate the model's parameter recovery capabilities through simulation studies. The model's practical utility is demonstrated through an empirical analysis of problem-solving process data, with particular attention to DIF detection. We conclude by discussing implications and future directions for process data analysis in educational measurement.

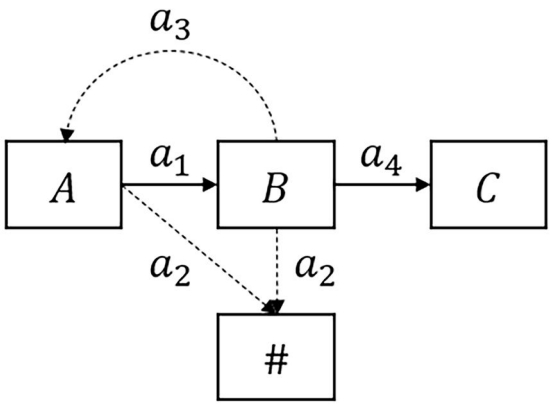
The Sequential Response Model With Covariates

Mayer and Wittrock (2006) define problem solving as "cognitive processing directed at transforming a given situation into a goal situation when no obvious method of solution is available to the problem solver". This process is inherently personalized (depending on the solver's existing knowledge and abilities), cognitive (occurring within the solver's cognitive system), goal-directed (aimed at achieving specific objectives), and process-based (involving sequential mental computations and representations) (Mayer, 1992; Mayer & Wittrock, 2006). The comprehensive recording of external behavioral processes — represented by transitions between problem states from initial to goal states — provides essential information for measuring latent problem-solving ability.

In CBA, a task state is defined as the cumulative sum of system changes resulting from actions taken from the beginning to a given moment. This includes all factors associated with decisions made within the task, which together constitute complex performance (LaMar, 2018). Consequently, each action chosen is equivalent to selecting the subsequent state achievable within the current task state. In well-designed tasks, the reachable states in the next step are a finite set that depends on the current state (Han et al., 2022; Xiao & Liu, 2024). Figure 1 illustrates a simplified state transition diagram where A denotes the initial state, B an intermediate state, C the target state, and # a premature termination state. Transitions between states occur through actions (a_1-a_4), where solid arrows indicate correct transitions (A⟶B, B⟶C) that progress toward the target state, while dotted arrows represent incorrect transitions that either deviate from or fail to advance toward the solution. The optimal solution path is represented by the state sequence A⟶B⟶C.

Figure 1

State Transition Diagram for Task Response Process



Note. A, B, C, and # represent initial, intermediate, target, and termination states, respectively. Actions a_1-a_4 facilitate state transitions, with solid arrows indicating correct transitions and dotted arrows indicating incorrect ones.

The Sequential Response Model

The response process can be conceptualized as a sequence of task states following a temporal stochastic process with the conditional Markov property, grounded in problem-solving theory where an individual's next action depends on both the current problem state and their latent problem-solving ability (Shu et al., 2017). For a task with finite states $\boldsymbol{s} = \{s_1, s_2, s_3, \ldots, s_z\}$, the sequential response model (SRM; Han et al., 2022) models

the probability of transitioning to the next state $S_{i,t+1}$ conditional on the current state $S_{i,t}$ and latent ability θ_i:

$$P(S_{i,t+1} = s_k | S_{i,t} = s_j, \theta_i, a, \mathscr{R}) = \frac{\exp\left(a_{j,k} + I^{+}_{j,k} \cdot \theta_i\right)}{\sum_{s_h \in M_j} \exp\left(a_{j,h} + I^{+}_{j,h} \cdot \theta_i\right)}, s_j \in s\ , s_k \in M_j, \tag{1}$$

where M_j denotes the set of reachable states from s_j. $I^{+}_{j,k}$ is a binary indicator (1 for correct, -1 for incorrect transitions) specified a priori in $\mathscr{R}$. The transition tendency parameter $a_{j,k}$ captures task-specific characteristics — the inherent propensity for transitioning from state s_j to state s_k. It functions analogously to item difficulty parameters in traditional item response theory (IRT) models. Positive values indicate transitions that are relatively easy to occur, while negative values indicate more difficult transitions. Empirically, these parameters typically range from -3 to 3, consistent with standard IRT parameterization. When combined with the directional indicator $I^{+}_{j,k}$, the model ensures that higher-ability individuals are more likely to make correct transitions ($a_{j,k} + \theta_i$ for correct transitions) and less likely to make errors ($a_{j,k} - \theta_i$ for incorrect transitions), consistent with established psychometric principles. Note that the SRM focuses on transition directionality (correct vs. incorrect) rather than discrimination, as the directional indicator $I^{+}_{j,k}$ already captures this information. The vector a contains all transition tendency parameters. As a multinomial logit model, the SRM requires $\sum_{s_h \in M_j} a_{j,h} = 0$ for identification (McFadden, 1974; Thissen & Steinberg, 1986).

The Sequential Response Model With Covariates

For process data, we define differential item functioning (DIF) as systematic differences in state transition probabilities between groups of examinees with equal ability levels, attributable to construct-irrelevant variables. Building upon the foundational SRM framework, the SRM-C introduces three key extensions to address group differences and measurement fairness in process data. The measurement component incorporates covariate effects on state transition probabilities, the structural component allows for group-specific ability distributions, and the identification strategy employs horseshoe priors to enable anchor-free DIF detection without requiring pre-specified invariant transitions. For illustration, we present the model for a two-group scenario where the covariate is binary, with values 0 and 1 representing membership in the reference and focal groups, respectively.

Measurement Model — The measurement component of the SRM-C models the relationship between response sequences and latent abilities while accounting for task features and covariates. This component extends Equation (1) by incorporating covariate effects:

$$P(S_{i,t+1} = s_k | S_{i,t} = s_j, \theta_i, a, b, x_i, \mathcal{R}) = \frac{\exp\left[a_{j,k} + I^{+}_{j,k} \cdot \theta_i + b_{j,k} \cdot x_i\right]}{\sum_{s_h \in M_j} \exp\left[a_{j,h} + I^{+}_{j,h} \cdot \theta_i + b_{j,h} \cdot x_i\right]}, \; s_j \in s \,, s_k \in M_j, \quad (2)$$

where x_i denotes the covariate indicating group membership, and the covariate effect parameter $b_{j,k}$ quantifies differential transition probabilities between groups at equal ability levels. When $b_{j,k} > 0$, the focal group ($x_i = 1$) shows higher propensity for transition $s_j \longrightarrow s_k$ compared to the reference group ($x_i = 0$) at the same ability level. This constitutes uniform DIF detection in the context of process data, where group differences manifest as horizontal shifts in transition probability functions. The vector b contains all covariate effects, and other notation follows Equation (1).

Structural Model — The structural component characterizes group-specific ability distributions:

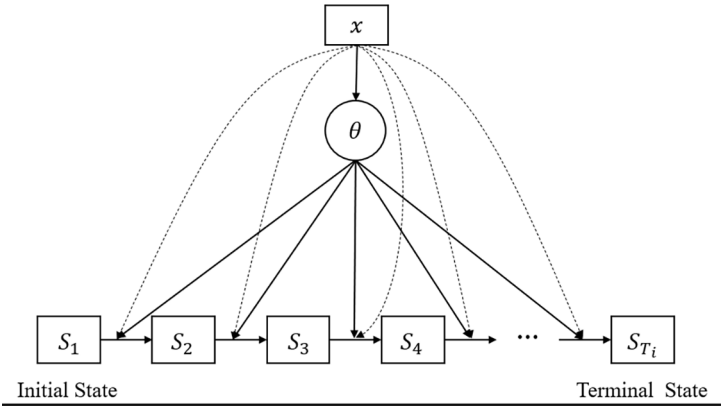
$$\theta_i | x_i \sim N\left(\mu x_i, \sigma^{2x_i}\right). \quad (3)$$

Abilities in the reference group ($x_i = 0$) follow a standard normal distribution to identify the location and scale of the latent trait, while abilities in the focal group ($x_i = 1$) follow $N(\mu, \sigma^2)$, where μ and σ are group-specific parameters. While the normality assumption represents a simplification of potentially more complex ability distributions, empirical research consistently demonstrates that cognitive abilities approximate normal distributions in educational contexts, and this assumption aligns with standard practice in psychometric modeling, including IRT and structural equation modeling approaches. In addition, allowing different location and scale parameters across groups accommodates realistic scenarios where groups differ in both mean ability and variability — a common finding in cross-cultural and demographic studies — thereby maintaining sufficient flexibility while avoiding overfitting risks associated with more complex distributional assumptions.

For an examinee with sequence length T_i, the complete SRM-C combines Equations (2) and (3). Figure 2 presents the path diagram of the model. The SRM-C framework is specifically designed for tasks with well-defined state spaces and clear transition correctness criteria, such as computer-based problem-solving assessments. The model's strength lies in its ability to simultaneously account for process complexity and group heterogeneity while maintaining interpretability.

Figure 2

Path Diagram of the SRM-C



Note. Subscript i is omitted for simplicity. Dashed lines from x to S_i represent DIF effects.

Model Identification — Similar to the SRM, the SRM-C requires constraints $\sum_{s_h \in M_j} a_{j,h} = 0$ and $\sum_{s_h \in M_j} b_{j,h} = 0$ for identification. The marginal likelihood function is:

$$L = \prod_{i=1}^{N} \int \left(\prod_{t=1}^{T_i - 1} \frac{\exp\left(a_{j,k} + I^{+}_{j,k} \cdot \theta_i + b_{j,k} \cdot x_i\right)}{\sum_{s_h \in M_j} \exp\left(a_{j,h} + I^{+}_{j,h} \cdot \theta_i + b_{j,h} \cdot x_i\right)} \right) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(\theta - \mu x_i)^2}{2\sigma^{2x_i}} \right) d\theta \,. \tag{4}$$

The model exhibits a location shift invariance: for any constant c, replacing μ with $\mu + c$ and $b_{j,h}$ with $b_{j,h} - c \cdot I^{+}_{j,h}$ yields an equivalent model. To resolve this non-identification issue (San Martín, 2016), we adopt a sparsity assumption that most state transitions are DIF-free, implemented through a horseshoe prior distribution. This approach aligns with common DIF detection methods that do not require anchor items (Chen et al., 2023) and offers favorable theoretical properties with straightforward implementation (Carvalho et al., 2009, 2010).

Bayesian Estimation

Under the conditional first-order Markov assumption, the joint probability of a state sequence S_i is:

$$p(S_i | \theta_i, a, b, x_i, \mathscr{R}) = \prod_{t=1}^{T_i - 1} p(S_{i,t+1} | S_{i,t}, \theta_i, a, b, x_i, \mathscr{R}) \,. \tag{5}$$

The joint posterior distribution is:

$$p(\theta, a, b | S, x, \mathscr{R}) \propto p(S | \theta, a, b, x, \mathscr{R}) p(\theta, a, b)$$

$$= \prod_{i=1}^{n} \prod_{t=1}^{T_i - 1} p(S_{i,t+1} | S_{i,t}, \theta_i, a, b, x_i, \mathscr{R}) p(\theta_i) p(a) p(b)$$

$$= \prod_{i=1}^{n} \prod_{t=1}^{T_i - 1} p(S_{i,t+1} | S_{i,t}, \theta_i, a, b, x_i, \mathscr{R}) p(\theta_i | x_i, \mu, \sigma) p(\mu) p(\sigma) p(a) p(b | \lambda, \tau) p(\lambda) p(\tau). \quad (6)$$

The model extends the standard SRM assumptions to accommodate group differences. For ability parameters, the SRM-C allows group-specific distributions $\theta_i | x_i \sim N(\mu x_i, \sigma^{2x_i})$ and maintains normal priors for state transition parameters $a_{j,k} \sim N(0, 1)$, with $\mu \sim N(0, 1)$ and $\sigma^2 \sim$Inv-Gamma$(1, 1)$. The key methodological innovation lies in the DIF parameter specification, where we adopt the horseshoe prior for the covariate coefficients: $b_{j,k} \sim N\left(0, \lambda_{j,k}^2 \cdot \tau^2\right)$, $\lambda_{j,k} \sim$half-Cauchy$(0, \quad 1)$, $s_j, s_k \in s$. Following Piironen and Vehtari (2017), the effective number of non-zero coefficients m_{eff} can be approximated as

$$E(m_{\text{eff}} | \tau, \sigma) = \frac{\tau \sigma^{-1} \sqrt{n}}{1 + \tau \sigma^{-1} \sqrt{n}} \cdot D, \quad (7)$$

where τ represents the global shrinkage parameter, σ denotes the noise standard deviation, n is the total sample size, and D indicates the number of state transitions. For our simulation conditions with $\tau = 0.2, \quad \sigma = 2$, and $D = 18$, this corresponds to expected DIF proportions of approximately 76% ($n = 1000$) and 82% ($n = 2000$). This choice reflects a conservative approach prioritizing the detection of measurement bias over model parsimony. In DIF detection, false negatives pose greater threats to educational equity than false positives, as undetected bias can perpetuate unfair assessment practices. Overly aggressive sparsity assumptions risk shrinking genuine DIF effects toward zero, potentially masking critical fairness violations. In addition, the horseshoe prior's hierarchical structure mitigates these concerns through its dual-layer shrinkage mechanism: while the global parameter τ establishes a lenient framework, the local parameters $\lambda_{j,k}$ provide adaptive regularization based on empirical evidence. This design preserves the prior's ability to distinguish signal from noise while reducing the risk of overlooking substantive bias effects.

Parameters are estimated using Markov chain Monte Carlo with a Gibbs sampler incorporating Metropolis-Hastings steps (Hastings, 1970). Detailed estimation procedures are provided in Supplemental Materials Section A (see Han et al., 2026).

Simulation Study

A Monte Carlo simulation study was conducted to evaluate three aspects of the SRM-C: parameter recovery capabilities, relative performance compared to the SRM, and efficacy in detecting DIF in response processes.

Design

The simulation design followed a state transition framework with eight states (Han et al., 2022; see Supplemental Materials Section B for more details, in Han et al., 2026).

The study employed a factorial design with four manipulated factors:

1. *Sample size* (n): 1000, 2000, with equal group sizes (i.e., 500 and 1000 per group, respectively).
2. *Sequence length*: Short (≤ 10 transitions), Medium (~20 transitions) and Long (~40 transitions). Sequence length was controlled by transition parameters, with larger values for transitions returning to states distant from the target producing longer sequences (see Table B2 in the Supplemental Materials in Han et al., 2026).
3. *Group difference*: Abilities were drawn from $N(0, 1)$ for both groups in the no-difference condition, and from $N(0, 1)$ for the reference group and $N(1, 1.5^2)$ for the focal group in the difference condition.
4. *DIF Pattern*: Three conditions were examined: DIF-free where all covariate coefficients ($b_{j,k}$) were set to 0, balanced DIF where the focal group showed higher probability for the correct transition E⟶F but lower probability for correct transition C⟶D, and unbalanced DIF where the focal group showed lower probabilities for both correct transitions (C⟶D and E⟶F). The specific coefficient values are shown in Table 1.

Table 1

True Values of Covariate Coefficients by DIF Pattern

DIF Pattern	b_{CA}	b_{CB}	b_{CD}	b_{EC}	b_{EF}	$b_{E\#}$
Balanced	0.4	0.6	-1.0	-1.0	1.0	0
Unbalanced	0.4	0.6	-1.0	0	-1.0	1.0

Note. For state C, only C⟶D is correct. For state E, only E⟶F is correct.

The simulation employed a 2(sample size) × 3(sequence length) × 2(group difference) × 3(DIF pattern) factorial design, yielding 36 conditions. Each condition was replicated 100 times with constant transition parameters and coefficients. Groups were balanced with equal numbers of examinees. Abilities were randomly generated per replication based on the group difference condition.

Parameter Estimation

Parameters were estimated using a custom Bayesian sampler in R (R Core Team, 2018), following the procedure described in Section A of the Supplemental Materials (see Han et al., 2026). Each of two chains ran for 15,000 iterations, with 10,000 burn-in and thinning by 10. Random initial values were used, and estimation was restarted with new initial

values if the potential scale reduction factor (PSRF) exceeded 1.2 (Gelman & Rubin, 1992). The sampler code is available at Han and Ji (2025).

To assess robustness to prior specification, we conducted sensitivity analyses using $\tau \in \{0.05, 0.1, 0.2\}$ for a representative subset of conditions ($n = 2000$, medium sequence, with group difference and unbalanced DIF). Results remained consistent across this range, supporting our choice of $\tau = 0.2$ for balancing DIF detection sensitivity with error control. Complete sensitivity analyses are reported in Supplemental Materials Section D (see Han et al., 2026).

Results of the Simulation Study

Convergence and Model Fit Comparison — All parameters converged with PSRF less than 1.2 (Gelman & Rubin, 1992). Trace plots (Figure B1 in the Supplemental Materials in Han et al., 2026) from an exemplar condition demonstrated convergence to the same posterior distributions despite different initial values. Similar convergence patterns were observed across all conditions.

Model fit was compared using the deviance information criterion (DIC; Spiegelhalter et al., 2002) and double log of pseudo Bayes factor ($2\log(\text{PsBF}_{21})$; Levy & Mislevy, 2016). A lower DIC indicates better fit, and $2\log(\text{PsBF}_{21})$ greater than 0 favors the SRM-C over the SRM, with values exceeding 10 suggesting strong evidence (Kass & Raftery, 1995).

Figure B3 in Supplemental Materials (see Han et al., 2026) presents detailed fit indices across conditions from 100 replications. The SRM-C showed consistently better fit (lower DIC and positive 2log(PsBF21)) in conditions with DIF or ability differences between groups. This advantage increased with larger sample sizes and longer sequences. The two models showed comparable fit only in conditions with no DIF and no ability differences.

Estimation Accuracy — Parameter recovery was evaluated using Root Mean Squared Error (RMSE). For ability parameters, posterior estimates were averaged across identical response patterns to account for MCMC sampling variability. Figure 3 presents RMSE values for both theta and item parameters, while Figure 4 displays the estimated ability distributions for the target group. Detailed numerical results, including bias statistics, correlations between estimated and true values, as well as RMSE and average occurrence frequency for individual covariate coefficients across conditions, are provided in Supplemental Materials B (see Han et al., 2026).

Figure 3

RMSE Values for Ability Estimates (θ), Transition Parameters (a), and Covariate Coefficients (b)

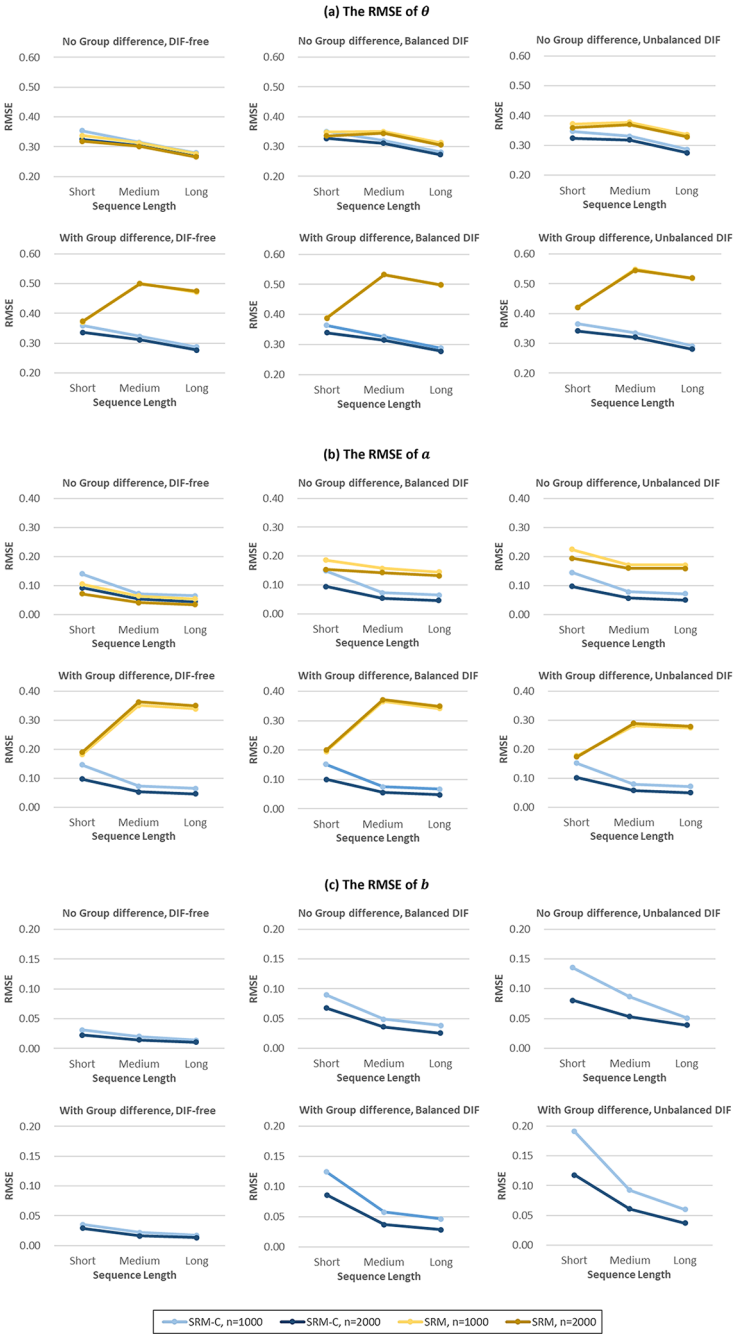
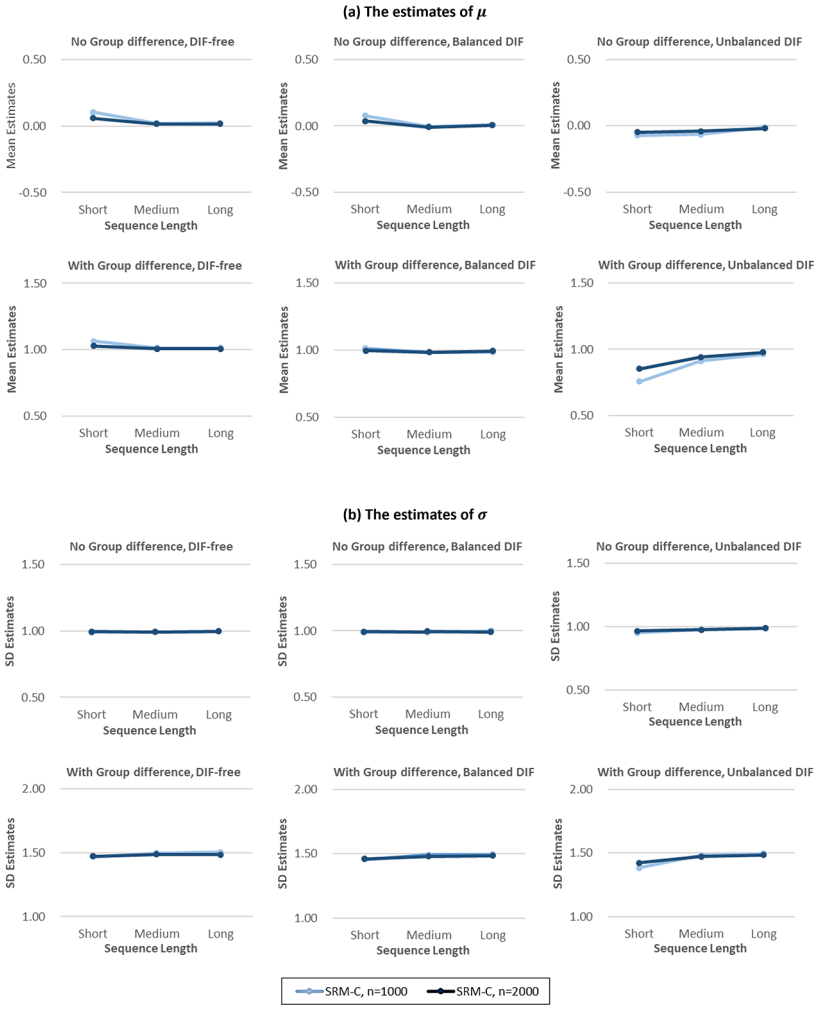


Figure 4

Mean Estimated Values of Target Group Ability Distribution Parameters: Mean (μ) and Standard Deviation (σ)



The SRM-C demonstrated robust parameter recovery across all conditions. RMSE values were below .35 for ability estimates, below .15 for transition parameters, and below .20 for covariate coefficients. Focal group distribution parameters were accurately recovered in most conditions, with mean and standard deviation estimates close to true values ($\mu = 0, \sigma = 1$ for no difference; $\mu = 1, \sigma = 1.5$ for with difference). The only exception occurred in unbalanced DIF conditions with short sequences (average length = 7) and

group differences, where estimates were slightly biased (n = 1000: $\widehat{\overline{\mu}}$ = 0.76, $\widehat{\overline{\sigma}}$ = 1.38; n = 2000: $\widehat{\overline{\mu}}$ = 0.85, $\widehat{\overline{\sigma}}$ = 1.42).

As expected, estimation accuracy improved with larger sample sizes and longer sequences. With n = 1000 (500 per group), parameter recovery was adequate under most conditions, but showed some deterioration with short sequences, particularly in complex scenarios involving both group differences and unbalanced DIF. For instance, RMSE values for ability estimates approached 0.35 under these challenging conditions, suggesting that smaller sample sizes may compromise estimation accuracy when combined with limited process information. When n > 1000 and sequences were not short, RMSE remained below .1 for transition parameters a and below .12 for covariate coefficients b. Additionally, group differences had minimal impact on estimation accuracy, particularly with longer sequences and larger samples.

The SRM showed comparable performance to the SRM-C only under DIF-free conditions without group differences. With DIF or group differences present, the SRM's accuracy deteriorated substantially, with the performance gap widening as sequence length and sample size increased. These results suggest the SRM-C should be preferred when sample heterogeneity exists, while serving as a viable alternative to the SRM when group abilities are homogeneous.

Type I Error Rate and Statistical Power Using the SRM-C to Assess DIF — DIF detection was based on 95% highest posterior density (HPD) intervals for $b_{j,k}$ parameters, with significance determined by whether intervals contained zero. Type I error rate was defined as the proportion of truly zero coefficients incorrectly identified as significant, while power was the proportion of truly non-zero coefficients correctly identified as significant. Results across replications are summarized in Tables 2 and Table 3.

Type I error rates remained well-controlled (below 0.5%) across all conditions. In DIF-free conditions, error rates were predominantly 0.00%, with only a slight increase to 0.06% in conditions with n = 2000 and long sequences, regardless of group differences. Under DIF conditions, error rates remained low but showed slight sensitivity to sequence length and sample size, with balanced DIF conditions showing better control than unbalanced conditions. The highest error rate (0.31%) was observed in the unbalanced DIF conditions with n = 1000 and short sequences, regardless of group differences. This rate remained well below the conventional 5% level.

Table 2

Type I Error Rates (%) for DIF Detection

Group Difference	Sample Size	Sequence Length	DIF Pattern		
			DIF-Free	Balanced	Unbalanced
No	1000	Short	0	0.00	0.31
		Medium	0	0.00	0.08
		Long	0	0.00	0
	2000	Short	0	0.00	0
		Medium	0	0.00	0.08
		Long	0.06	0.00	0
Yes	1000	Short	0	0.08	0.31
		Medium	0	0.00	0.08
		Long	0	0.00	0.08
	2000	Short	0	0.08	0.15
		Medium	0	0.08	0.23
		Long	0.06	0.00	0

Table 3

Statistical Power (%) for DIF Detection

DIF Pattern	Group Difference	Sample Size	Sequence Length	Average Power (%)	Coefficients				
					b_{CA}	b_{CB}	b_{CD}	b_{EF}	$b_{EC}/b_{E\#}$
Balanced	No	1000	Short	85.2	49	97	100	95	85
			Medium	100	100	100	100	100	100
			Long	100	100	100	100	100	100
		2000	Short	97.6	90	100	100	100	98
			Medium	100	100	100	100	100	100
			Long	100	100	100	100	100	100
	Yes	1000	Short	65.8	29	87	98	74	41
			Medium	100	100	100	100	100	100
			Long	100	100	100	100	100	100
		2000	Short	93.6	80	99	100	99	90
			Medium	100	100	100	100	100	100
			Long	100	100	100	100	100	100

DIF Pattern	Group Difference	Sample Size	Sequence Length	Average Power (%)	Coefficients				
					b_{CA}	b_{CB}	b_{CD}	b_{EF}	$b_{EC}/b_{E\#}$
Unbalanced	No	1000	Short	75.6	19	81	95	88	95
			Medium	98.8	97	100	100	97	100
			Long	100	100	100	100	100	100
		2000	Short	95.8	79	100	100	100	100
			Medium	100	100	100	100	100	100
			Long	100	100	100	100	100	100
	Yes	1000	Short	49.6	8	53	72	43	72
			Medium	98.8	96	100	100	98	100
			Long	100	100	100	100	100	100
		2000	Short	87.6	46	95	98	99	100
			Medium	100	100	100	100	100	100
			Long	100	100	100	100	100	100

As shown in Table 3, power exceeded 85% in most conditions except those with short sequences and n = 1000. Statistical power was higher for balanced than unbalanced DIF conditions, and showed strong sensitivity to sequence length, exceeding 98% for medium and long sequences. Sample size and group differences also affected power, particularly with short sequences. For example, under unbalanced DIF with group differences, power increased from 49.6% (n = 1000) to 87.6% (n = 2000) with short sequences, suggesting that large samples can compensate for short sequence lengths (< 10 transitions). These results highlight that while the method maintains excellent Type I error control even with moderate sample sizes (500 per group), sufficient statistical power for DIF detection requires either larger samples or longer process sequences, with the combination of small samples and short sequences presenting the most challenging scenario for reliable DIF detection.

Power varied across coefficients, influenced by both effect size and transition frequency. Coefficients with smaller absolute values (e.g., b_{CA} = 0.4) showed lower detection rates than those with larger values. The transition frequency also impacted power, as illustrated by b_{EF}: in balanced DIF conditions, its positive value (1) increased the tendency and average frequency of this transition for the focal group, while in unbalanced conditions, its negative value (-1) decreased this tendency, resulting in higher power for balanced conditions.

Empirical Study

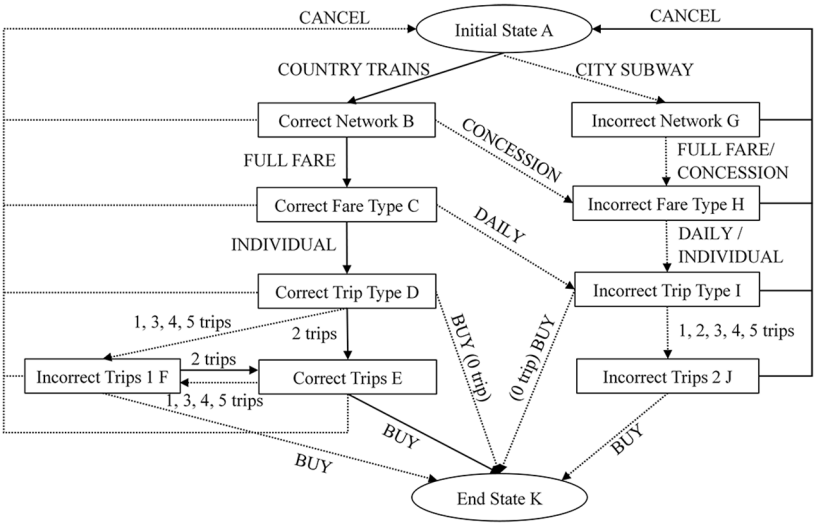
An empirical analysis was conducted using process data from the Tickets task in PISA 2012 to demonstrate the practical application of the SRM-C.

The Tickets Task

The *Tickets* task from PISA 2012 was selected for empirical analysis due to its structured interface design and extensive use in process data modeling literature (Chen, 2020; Fu et al., 2023; Han et al., 2022; Xiao & Liu, 2024). In the focal item (CP038Q02), examinees operated a virtual ticketing machine to purchase a full fare country train ticket with two individual trips (OECD, 2014)[1]. The task's well-defined structure allows for clear state decomposition. Following Han et al. (2022), we decomposed the response process into 11 distinct states with 27 possible transitions between them (Figure 5).

Figure 5

State-Transition Diagram for PISA Tickets Task (CP038Q02)



Note. (Han et al., 2022). Ellipses denote start/end states; rectangles denote intermediate states. Solid arrows indicate correct transitions; dotted arrows indicate incorrect transitions. States represent distinct interface stages; transitions represent examinees' actions between states.

Method

The analysis included response sequences from 1,672 Finnish (reference group) and 1,752 Australian (focal group) examinees (n = 3,424). Sequence lengths ranged from 5 to 43 actions (M = 6.79). Success rates differed notably between Finnish (57.5%) and Australian (70.3%) examinees, necessitating investigation of whether this disparity stems from ability differences or potential differential item functioning.

1) The original process data file in the empirical study reported here is available at OECD website: http://www.oecd.org/pisa/pisaproducts/database-cbapisa2012.htm.

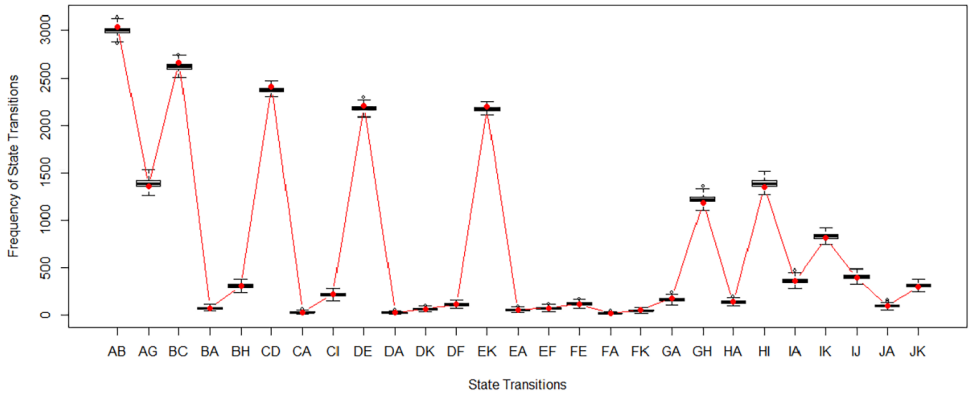
The SRM-C was fitted to the data using Bayesian estimation with the same MCMC specifications as the simulation study. The SRM-C implementation treated nationality as a binary covariate (0 = Finnish, 1 = Australian) to examine its effects on both ability distribution and response processes. All data and analysis code are available at Han and Ji (2025).

Results of the Empirical Study

Convergence and Model Fit — Model convergence was achieved with PSRF < 1.2 for all parameters (Gelman & Rubin, 1992). Model fit was evaluated using posterior predictive checks (PPC) based on 1,000 MCMC iterations. Figure 6 presents the comparison between observed state transition frequencies (points) and their posterior predictive distributions (boxplots). The observed values align well with model predictions, with all observations falling within their 95% prediction intervals. The posterior predictive p-value (*ppp* = 0.139) falls within the acceptable range [0.05, 0.95], indicating adequate model fit (Gelman et al., 2013).

Figure 6

Posterior Predictive Check for State Transition Frequencies



Note. Boxplots show posterior predictive distributions (median, IQR, and 2.5th-97.5th percentiles); points indicate observed frequencies.

Group-Specific Ability Parameters — The Australian group (focal group) demonstrated significantly higher mean ability (posterior *Mean* = 0.289, *SD* = 0.054, 95% Highest Posterior Density (HPD): [0.168, 0.375]) compared to the Finnish group (reference group), and their ability variance (0.609) was lower than the reference group's standardized variance. These differences indicate distinct ability distributions between the two groups in this problem-solving task.

Country Effects on Response Processes — The posterior distributions of country coefficients for all state transitions showed small magnitudes ($|b_{j,k}| < 0.2$) with 95% HPD intervals containing zero (Table 4). This absence of significant country effects on transition probabilities suggests measurement invariance across nations, indicating that performance differences between Finnish and Australian examinees are attributable to ability differences rather than DIF.

Table 4

Posterior Distributions of Country Coefficients for State Transitions

Parameter	Mean	Standard Deviation	Lower Bound of the 95% HPD Interval	Upper Bound of the 95% HPD Interval
b_{AB}	**-0.026**	**0.041**	**-0.123**	**0.012**
b_{AG}	0.026	0.041	-0.012	0.123
b_{BC}	**-0.051**	**0.075**	**-0.222**	**0.062**
b_{BA}	0.034	0.078	-0.097	0.210
b_{BH}	0.018	0.049	-0.067	0.135
b_{CD}	**0.034**	**0.067**	**-0.082**	**0.168**
b_{CA}	-0.021	0.074	-0.183	0.123
b_{CI}	-0.012	0.048	-0.126	0.082
b_{DE}	**0.197**	**0.150**	**-0.052**	**0.468**
b_{DA}	-0.118	0.160	-0.467	0.125
b_{DK}	-0.034	0.115	-0.351	0.146
b_{DF}	-0.044	0.102	-0.304	0.110
b_{EK}	**0.057**	**0.079**	**-0.055**	**0.237**
b_{EA}	-0.007	0.076	-0.183	0.154
b_{EF}	-0.050	0.086	-0.258	0.092
b_{FE}	**0.018**	**0.084**	**-0.154**	**0.183**
b_{FA}	-0.021	0.086	-0.204	0.138
b_{FK}	0.003	0.080	-0.187	0.170
b_{GA}	**0.011**	**0.036**	**-0.051**	**0.086**
b_{GH}	-0.011	0.036	-0.086	0.051
b_{HA}	**0.010**	**0.035**	**-0.046**	**0.103**
b_{HI}	-0.010	0.035	-0.103	0.046
b_{IA}	**0.013**	**0.043**	**-0.073**	**0.109**
b_{IK}	-0.020	0.039	-0.110	0.046
b_{IJ}	0.007	0.036	-0.059	0.098
b_{JA}	**-0.007**	**0.042**	**-0.094**	**0.085**
b_{JK}	0.007	0.042	-0.085	0.094

Note. Correct transitions in bold.

Response Patterns and Task Characteristics — Analysis of ability estimates by response pattern and state transition parameters revealed meaningful patterns in both examinee behavior and task design effects (see Online Supplemental Materials C for detailed results in Han et al., 2026). Ability estimates aligned with theoretical expectations: the optimal sequence (ABCDEK) yielded the highest estimates, while sequences with uncorrected errors showed the lowest estimates. State transition parameters revealed an interface-dependent correction pattern: examinees showed reluctance to correct errors at intermediate stages but demonstrated higher correction probabilities at the final purchase interface where ticket details were explicitly displayed, suggesting the impact of interface design on problem-solving behavior.

Summary and Discussion

Process data in CBAs holds significant potential to uncover and address inequities experienced by examinees, thereby ensuring fairness in the evaluation process. This study introduced the SRM-C to analyze such process data while accounting for group differences. The model demonstrates several key capabilities: accurate estimation of examinees' abilities and task characteristic parameters, evaluation of covariate effects on ability distributions, and detection of DIF in response processes. Our simulation studies revealed that the SRM-C achieves robust parameter recovery and effective DIF detection under various conditions. Notably, adequate model performance can be achieved with either large samples ($n > 2000$) with short sequences or moderate samples with sequence lengths exceeding 10 transitions. The empirical analysis of PISA data demonstrated the model's practical utility in distinguishing whether group performance differences stem from genuine ability variations or potential measurement bias. These findings suggest that the SRM-C can serve as a valuable tool for multiple educational applications: enabling researchers to understand group differences in ability distributions, helping test developers identify and minimize potential biases in assessments, assisting practitioners in analyzing problem-solving strategies across different populations, and supporting fair evaluation in high-stakes assessment contexts.

While the SRM-C demonstrates theoretical soundness and favorable performance under simulation conditions, several practical considerations warrant discussion regarding its real-world applicability. Model identification in the SRM-C parallels that of the multiple indicators, multiple causes (MIMIC) model (Muthén, 1985; Muthén et al., 1991; Muthén & Lehman, 1985), where both measurement and structural components require careful consideration of identifiability constraints. Like the MIMIC model for DIF detection, the SRM-C faces an identification challenge when no anchor items (or state transitions) are pre-specified. Following recent developments in DIF detection methodology (Chen et al., 2023), we addressed this issue through a sparsity assumption — the premise that most state transitions are DIF-free. This assumption, which is reasonable for

most educational contexts where well-designed assessments should exhibit measurement invariance across groups, was implemented within a Bayesian framework using a horseshoe prior with the global shrinkage parameter fixed at 0.2, following Betancourt's (2021) one-horse-town approach. The method is most appropriate for applications where the majority of process elements are expected to function equivalently across groups, consistent with best practices in fair assessment design. However, the sparsity assumption may be violated when systematic measurement bias is pervasive across state transitions. In such cases, alternative identification strategies warrant future investigation, including: (1) anchor-based approaches where domain experts could pre-specify invariant transitions based on substantive knowledge, (2) alternative modeling frameworks that do not rely on sparsity assumptions, such as constrained multi-group extensions with different identification strategies. When the sparsity assumption is uncertain, we currently recommend conducting preliminary analyses to assess the plausibility of measurement invariance before model fitting, and when possible, triangulating findings with external validity evidence or alternative analytical approaches.

Within the adopted sparsity framework, we fixed τ at 0.2, reflecting a relatively liberal approach that prioritizes sensitivity for detecting measurement bias over strict sparsity enforcement. The horseshoe prior's local parameters $\lambda_{j,k}$ are estimated from the data, providing adaptive regularization for individual coefficients. Both simulation and empirical studies demonstrated satisfactory performance under this specification. Sensitivity analyses across $\tau \in \{0.05, 0.1, 0.2\}$ revealed that the model demonstrated robust performance across this range with stable parameter recovery and effective DIF detection. For practical applications, we recommend conducting sensitivity analyses when uncertain about the chosen value. Future research could explore full Bayesian inference for τ through hierarchical modeling with appropriate hyperpriors, enabling data-driven estimation of the global shrinkage parameter.

Process data in CBAs may contain systematic bias where interface elements, cultural content, or technological familiarity systematically advantage or disadvantage certain groups across all ability levels, creating consistent rather than ability-dependent group differences. This study therefore focused on uniform DIF with a binary categorical covariate to address such bias patterns. Beyond its current capabilities, the SRM-C framework demonstrates strong extensibility potential. A particularly promising extension involves developing a two-parameter variant by allowing discrimination parameters (directional indicators $I^{+}_{j,k}$) to be estimated rather than fixed at ± 1, and incorporating covariate-by-ability interaction terms to enable detection of non-uniform DIF. This extension leverages the model's foundational structure while expanding its analytical capabilities to address more complex DIF patterns in process data fairness evaluation.

The SRM-C is primarily designed for large-scale assessment contexts where adequate sample sizes and process data richness can support robust parameter estimation. Our simulation results indicate that the method performs optimally with sample sizes ex-

ceeding 1000 per group, particularly when combined with sequences longer than 10 transitions. While the method showed acceptable performance with 500 participants per group under favorable conditions (longer sequences, simpler DIF patterns), we recommend caution when applying the SRM-C to small-scale studies. When large samples are not feasible, researchers should ensure longer response sequences (≥ 20 transitions) to compensate for reduced sample size, as sequence length and sample size can partially offset each other in supporting parameter identification and estimation precision.

The PISA empirical analysis primarily demonstrated the method's behavior under ideal conditions — clear group differences with minimal DIF. Real-world applications may encounter more complex DIF patterns, smaller effect sizes, or violations of distributional assumptions that require additional consideration. For researchers considering the SRM-C, we recommend: (1) conducting preliminary analyses to assess the plausibility of the sparsity assumption within their specific context; (2) ensuring adequate sample sizes (≥ 1000 per group for optimal performance under all conditions, or ≥ 500 per group when sequences exceed 20 transitions, as longer sequences can partially compensate for smaller sample sizes); (3) performing comprehensive model diagnostics, including parameter convergence assessment, posterior predictive checks, and model comparison with the simplified SRM; (4) conducting sensitivity analyses with different τ values to optimize performance within the sparsity framework, particularly when facing convergence issues or estimation instability; and (5) comparing results with traditional anchor-based methods when reliable anchor items can be reasonably specified.

Funding: This research project is supported by the Fundamental Research Funds for the Central Universities of Beijing Language and Culture University (25ZX01).

Acknowledgments: The authors have no additional (i.e., non-financial) support to report.

Competing Interests: The authors have no relevant financial or non-financial interests to disclose.

Data Availability: The data and code that support the findings of this study are available in the OSF repository at Han and Ji (2025). Supplemental procedures and analyses for this study are available at Han et al. (2025).

Supplementary Materials

Type of supplementary materials	Availability/Access
Data	
Empirical study - data.	Han and Ji (2025)
Simulation study - data.	Han and Ji (2025)
Code	
Empirical study - R code.	Han and Ji (2025)

Type of supplementary materials	Availability/Access
Simulation study - R code.	Han and Ji (2025)
Material	
Supplemental procedures and analyses.	Han et al. (2025)
Study/Analysis preregistration	
The study was not preregistered.	—
Other	
No other materials available.	—

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *The standards for educational and psychological testing*. American Educational Research Association.

Banfield, J., & Wilkerson, B. (2014). Increasing student intrinsic motivation and self-efficacy through gamification pedagogy. *Contemporary Issues in Education Research, 7*(4), 291–298. https://doi.org/10.19030/cier.v7i4.8843

Bauer, D. J., Belzak, W. C. M., & Cole, V. T. (2020). Simplifying the assessment of measurement invariance over multiple background variables: Using regularized moderated nonlinear factor analysis to detect differential item functioning. *Structural Equation Modeling, 27*(1), 43–55. https://doi.org/10.1080/10705511.2019.1642754

Belzak, W. C. M., & Bauer, D. J. (2020). Improving the assessment of measurement invariance: Using regularization to select anchor items and identify differential item functioning. *Psychological Methods, 25*(6), 673–690. https://doi.org/10.1037/met0000253

Betancourt, M. (2021, May). *Sparsity blues*. GitHub. https://betanalpha.github.io/assets/case_studies/modeling_sparsity.html#2223_The_horseshoe_Population_Model

Carvalho, C. M., Polson, N. G., & Scott, J. G. (2009). Handling sparsity via the horseshoe. In D. van Dyk & M. Welling (Eds.), *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics* (pp. 73–80). PMLR.

Carvalho, C. M., Polson, N. G., & Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika, 97*(2), 465–480. https://doi.org/10.1093/biomet/asq017

Casella, G., Ghosh, M., Gill, J., & Kyung, M. (2010). Penalized regression, standard errors, and Bayesian Lassos. *Bayesian Analysis, 5*(2), 369–412. https://doi.org/10.1214/10-BA607

Chen, Y. (2020). A continuous-time dynamic choice measurement model for problem-solving process data. *Psychometrika, 85*(4), 1052–1075. https://doi.org/10.1007/s11336-020-09734-1

Chen, Y., Li, C., Ouyang, J., & Xu, G. (2023). DIF statistical inference without knowing anchoring items. *Psychometrika, 88*(4), 1097–1122. https://doi.org/10.1007/s11336-023-09930-9

Ercikan, K., & Pellegrino, J. W. (Eds.). (2017). *Validation of score meaning for the next generation of assessments: The use of response processes*. Routledge.

Fu, Y., Zhan, P., Chen, Q., & Jiao, H. (2023). Joint modeling of action sequences and action time in computer-based interactive tasks. *Behavior Research Methods, 56*, 4293–4310. https://doi.org/10.3758/s13428-023-02178-2

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). CRC Press.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science, 7*(4), 457–472. https://doi.org/10.1214/ss/1177011136

Goodman, M., Finnegan, R., Mohadjer, L., Krenzke, T., & Hogan, J. (2013). *Literacy, numeracy, and problem solving in technology-rich environments among US adults: Results from the program for the international assessment of adult competencies 2012: First look* (No. NCES 2014-008). National Center for Education Statistics.

Griffin, P., McGaw, B., & Care, E. (Eds.). (2012). *Assessment and teaching of 21st century skills*. Springer.

Han, Y., & Ji, F. (2025). *Bayesian sampler for the Sequential Response Model with Covariates (SRM-C)* [OSF project page containing data and analysis/sampler code for study]. Open Science Framework. https://osf.io/e3pqv/overview

Han, Y., Ji, F., Chen, Y., Gan, K., & Liu, H. (2026). *Supplementary Materials to* "Analyzing group differences and measurement fairness in process data: A sequential response model with covariates" [Supplemental procedures and analyses]. PsychOpen GOLD. https://doi.org/10.23668/psycharchives.21776

Han, Y., Liu, H., & Ji, F. (2022). A sequential response model for analyzing process data on technology-based problem-solving tasks. *Multivariate Behavioral Research, 57*(6), 960–977. https://doi.org/10.1080/00273171.2021.1932403

Han, Y., & Wilson, M. (2022). Analyzing student response processes to evaluate success on a technology-based problem-solving task. *Applied Measurement in Education, 35*(1), 33–45. https://doi.org/10.1080/08957347.2022.2034821

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika, 57*(1), 97–109. https://doi.org/10.1093/biomet/57.1.97

Hesse, F., Care, E., Buder, J., Sassenberg, K., & Griffin, P. (2015). A framework for teachable collaborative problem solving skills. In P. Griffin & E. Care (Eds.), *Assessment and teaching of 21st century skills: Methods and approach* (pp. 37–56). Springer.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129–145). Lawrence Erlbaum Associates.

Huang, P. H. (2018). A penalized likelihood method for multi-group structural equation modelling. *British Journal of Mathematical & Statistical Psychology, 71*(3), 499–522. https://doi.org/10.1111/bmsp.12130

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90*(430), 773–795. https://doi.org/10.1080/01621459.1995.10476572

LaMar, M. M. (2018). Markov decision process measurement model. *Psychometrika, 83*(1), 67–88. https://doi.org/10.1007/s11336-017-9570-0

Levy, R., & Mislevy, R. J. (2016). *Bayesian psychometric modeling*. CRC Press.

Li, J., Zhang, B., Du, H., Zhu, Z., & Li, Y. M. (2015). Metacognitive planning: Development and validation of an online measure. *Psychological Assessment, 27*(1), 260–271. https://doi.org/10.1037/pas0000019

Liu, H., Liu, Y., & Li, M. (2018). Analysis of process data of PISA 2012 computer-based problem solving: Application of the modified multilevel mixture IRT model. *Frontiers in Psychology, 9*, Article 1372. https://doi.org/10.3389/fpsyg.2018.01372

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Lawrence Erlbaum Associates.

Magis, D., Béland, S., Tuerlinckx, F., & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods, 42*(3), 847–862. https://doi.org/10.3758/BRM.42.3.847

Magis, D., Tuerlinckx, F., & De Boeck, P. (2015). Detection of differential item functioning using the lasso approach. *Journal of Educational and Behavioral Statistics, 40*(2), 111–135. https://doi.org/10.3102/1076998614559747

Mayer, R. E. (1992). *Thinking, problem solving, cognition* (2nd ed). Freeman.

Mayer, R. E., & Wittrock, M. C. (2006). Problem solving. In P. A. Alexander & P. H. Winne (Eds.), *Handbook of educational psychology* (2nd ed., pp. 287–303). Lawrence Erlbaum Associates.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), *Frontiers in Econometrics* (pp. 105–142). Academic Press.

Mislevy, R. J., Oranje, A., Bauer, M. I., von Davier, A. A., Hao, J., Corrigan, S., Hoffman, E., DiCerbo, K., & John, M. (2014). *Psychometric considerations in game-based assessment*. GlassLabGames.

Mitchell, T. J., & Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association, 83*(404), 1023–1032. https://doi.org/10.1080/01621459.1988.10478694

Muthén, B. (1985). A method for studying the homogeneity of test items with respect to other relevant variables. *Journal of Educational Statistics, 10*(2), 121–132. https://doi.org/10.3102/10769986010002121

Muthén, B., & Lehman, J. (1985). Multiple group IRT modeling: Applications to item bias analysis. *Journal of Educational Statistics, 10*(2), 133–142. https://doi.org/10.3102/10769986010002133

Muthén, B. O., Kao, C. F., & Burstein, L. (1991). Instructionally sensitive psychometrics: Application of a new IRT-based detection technique to mathematics achievement test items. *Journal of Educational Measurement, 28*(1), 1–22. https://doi.org/10.1111/j.1745-3984.1991.tb00340.x

National Center for Education Statistics. (2014). *NAEP TEL Wells sample item*. http://nces.ed.gov/nationsreportcard/tel/wells_item.aspx

OECD. (2014). *PISA 2012 results: Creative problem solving: Students' skills in tackling real-life problems* (Vol. 5). OECD Publishing.

OECD. (2016). *Technical report of the survey of adult skills (PIAAC)* (2nd ed.). OECD Publishing.

OECD. (2017). *PISA 2015 technical report*. OECD Publishing. https://www.oecd.org/pisa/data/2015-technical-report/PISA2015_TechRep_Final.pdf

Park, T., & Casella, G. (2008). The Bayesian Lasso. *Journal of the American Statistical Association, 103*(482), 681–686. https://doi.org/10.1198/016214508000000337

Piironen, J., & Vehtari, A. (2017). Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Statistics, 11*(2), 5018–5051. https://doi.org/10.1214/17-EJS1337SI

Polson, N. G., & Scott, J. G. (2011). Shrink globally, act locally: Sparse Bayesian regularization and prediction. In J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith & M. West (Eds.), *Bayesian statistics 9* (pp. 501–538). Oxford University Press.

R Core Team. (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika, 53*(4), 495–502. https://doi.org/10.1007/BF02294403

San Martín, E. (2016). Identification of item response theory models. In W. J. van der Linden (Ed.), *Handbook of item response theory* (Vol. 2, pp. 127–150). CRC Press.

Schauberger, G., & Mair, P. (2020). A regularization approach for the detection of differential item functioning in generalized partial credit models. *Behavior Research Methods, 52*(1), 279–294. https://doi.org/10.3758/s13428-019-01224-2

Schleicher, A. (2008). PIAAC: A new strategy for assessing adult competencies. *International Review of Education, 54*(5–6), 627–650. https://doi.org/10.1007/s11159-008-9105-0

Shu, Z., Bergner, Y., Zhu, M., Hao, J., & von Davier, A. A. (2017). An item response theory analysis of problem-solving processes in scenario-based tasks. *Psychological Test and Assessment Modeling, 59*(1), 109–131.

Shute, V. J., & Moore, G. R. (2017). Consistency and validity in game-based stealth assessment. In H. Jiao & R. W. Lissitz (Eds.), *Technology enhanced innovative assessment: Development, modeling, and scoring from an interdisciplinary perspective* (pp. 31–51). Information Age Publishing.

Siddiq, F., Gochyyev, P., & Wilson, M. (2017). Learning in digital networks – ICT literacy: A novel assessment of students' 21st century skills. *Computers & Education, 109*, 11–37. https://doi.org/10.1016/j.compedu.2017.01.014

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B: Statistical Methodology, 64*(4), 583–639. https://doi.org/10.1111/1467-9868.00353

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*(4), 361–370. https://doi.org/10.1111/j.1745-3984.1990.tb00754.x

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika, 51*(4), 567–577. https://doi.org/10.1007/BF02295596

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67–113). Lawrence Erlbaum Associates.

Tutz, G., & Schauberger, G. (2015). A penalty approach to differential item functioning in Rasch models. *Psychometrika, 80*(1), 21–43. https://doi.org/10.1007/s11336-013-9377-6

Xiao, Y., & Liu, H. (2024). A state response measurement model for problem-solving process data. *Behavior Research Methods, 56*(1), 258–277. https://doi.org/10.3758/s13428-022-02042-9

Xiao, Y., Veldkamp, B., & Liu, H. (2022). Combining process information and item response modeling to estimate problem-solving ability. *Educational Measurement: Issues and Practice, 41*(2), 36–54. https://doi.org/10.1111/emip.12474

Yuan, K. H., Liu, H., & Han, Y. (2021). Differential item functioning analysis without a priori information on anchor items: QQ plots and graphical test. *Psychometrika, 86*(2), 345–377. https://doi.org/10.1007/s11336-021-09746-5

Zhan, P., & Qiao, X. (2022). Diagnostic classification analysis of problem-solving competence using process data: An item expansion method. *Psychometrika, 87*(4), 1529–1547. https://doi.org/10.1007/s11336-022-09855-9