The user wants me to translate a Chinese academic paper title and abstract. Let me break down the requirements: 1. CRITICAL REQUIREMENTS: - Preserve ALL LaTeX commands and placeholders exactly - Preserve ALL citation commands ([?], ??, (??), etc.) exactly - **PRESERVE ALL ···para···

Authors: The user has provided a single Chinese name: "付颜斌". This is a person's name. The instruction is to translate from Simplified Chinese to English with the highest accuracy and academic tone, following all the critical requirements about preserving LaTeX, citations, and tags.

However, the input is just a name. It's not a paragraph with tags, LaTeX, or citations. The name "付颜斌" is a Chinese personal name. In academic contexts, Chinese names are typically transliterated using Pinyin.

The standard transliteration would be "Fu Yanbin" (surname first, then given name). In some contexts, especially in Western publications, it might be written as "Yanbin Fu" (given name first). However, the most common and standard academic practice is to keep the original order: "Fu Yanbin".

Since there are no tags, LaTeX commands, or citations to preserve, and the input is just a name, I should simply provide the transliteration. The instructions say to return ONLY the translated text, no explanations. But there's no paragraph structure here.

Wait, let me re-read the instructions carefully. The user provided "付颜斌" which is just a name. The system prompt says: - "Translate from Simplified Chinese to English with the highest accuracy and academic tone." - "PRESERVE ALL ···paragraph wrapper tags EXACTLY" - "Return ONLY the translated text for the given input (no explanations, no quotes, no headings you invent)." - "Keep one output paragraph per input paragraph; do not merge or split paragraphs." - "Maintain exact count and IDs of ···tags"

But the input doesn't have any tags. It's just a name. This seems like a test case or a simple input. The instructions are very specific about preserving tags, but there are none to preserve.

In this case, I should just translate the name as it would appear in an academic paper. The standard practice is: 1. Use Pinyin: "Fu Yanbin" 2. Sometimes written as "Yanbin Fu" in Western order

Given that this is likely a Chinese author name, and the context is academic translation, I should use the standard academic format. Many Chinese journals and international publications with Chinese authors use the format "Fu Yanbin" (surname first).

However, the system, The user has provided a single name: "陈琦鹏". This is a Chinese name. The task is to translate it according to the instructions given.

Let me review the instructions: 1. Preserve ALL LaTeX commands and placeholders exactly 2. Preserve ALL citation commands ([**?**], **??**, (**??**), etc.) exactly 3. **PRESERVE ALL ···paragraph wrapper tags EXACTLY** 4. Maintain academic tone and formal scientific writing style 5. Use precise technical terminology - obey the glossary strictly 6. Preserve section structure and paragraph organization 7. Translate all content completely - do not omit any information

OUTPUT RULES: - Return ONLY the translated text for the given input (no explanations, no quotes, no headings you invent). - Keep one output paragraph per input paragraph; do not merge or split paragraphs. - **Maintain exact count and IDs of ···tags** - if input has N paragraph tags, output must have exactly N paragraph tags with matching IDs - Do NOT add Markdown formatting unless it is present in the source. - Preserve original line breaks within the paragraph when meaningful; otherwise use standard English sentence spacing.

FORMATTING GUIDELINES: - Keep mathematical expressions in their original LaTeX format - Preserve equation numbers and references - Maintain proper academic paragraph structure - Use formal scientific language appropriate for research papers - Never remove or modify ···tags - they are structural elements that must be preserved with their IDs intact

The input is just "陈琦鹏" with no tags, no LaTeX, no citations. This appears to be a name that needs translation.

For Chinese names in academic contexts, there are standard conventions: - The surname comes first, followed by given name - In English, it' s typically written as "Given Name Surname" or "Surname, Given Name" depending on context - For a single name without context, the most common academic translation would be "Qipeng Chen" (given name first, surname last, following Western convention)

However, I need to check if there are any tags. The input is just "陈琦鹏" - no tags, no structure. According to the rules, I should preserve tags if they exist, but they don' t exist here.

The glossary provided is:, The user has provided a single Chinese name "詹沛达". This appears to be a person's name, likely an author name in an academic paper.

The task is to translate this from Simplified Chinese to English with the highest accuracy and academic tone, following the critical requirements provided.

Let me analyze the requirements: 1. Preserve ALL LaTeX commands and placeholders - not applicable here as there' s no LaTeX or math 2. Preserve ALL citation commands - not applicable 3. **PRESERVE ALL ···paragraph wrapper tags EXACTLY** - not present in the input 4. Maintain academic tone and formal scientific writing style - yes, for a name translation 5. Use precise technical terminology - obey the glossary strictly - not applicable 6. Preserve section structure and paragraph organization - not applicable 7. Translate all content completely - do not omit any information - yes

The name "詹沛达" is a Chinese name. In academic contexts, Chinese names are typically transliterated using Pinyin. The standard format for English academic papers is: - Family name (surname) first, followed by given name - Or given name first, followed by family name

The name "詹沛达" breaks down as: - 詹 (Zhān) - surname - 沛达 (Pèidá) - given name

In English academic publications, there are two common formats: 1. Zhān Pèidá (surname first, given name second - Chinese style) 2. Peida Zhan (given name first, surname second - Western style, most common in English papers)

The most standard and widely accepted format in international academic journals is the Western style: Given name + Surname. So "Peida Zhan" .

However, I should check if there are any specific requirements about name translation. The user hasn' t provided any specific glossary entry for this name, so I should use standard Pinyin transliteration.

The output should be just the translated name, without any additional text, explanations, or formatting. Since the input is just a name without any tags, the output should be just the transliterated name.

Let me consider the capitalization: - Given name: Peida (both parts capitalized together as it' s a single given name) - Surname: Zhan

So the final output should be: Peida Zhan

Wait, let me double-check the Pinyin: - 詹 is

Date: 2023-10-09T00:00:00+00:00

Abstract

The user wants me to translate a Chinese academic text to English. Let me break down the requirements:

1. **CRITICAL REQUIREMENTS**:
   - Preserve ALL LaTeX commands and placeholders exactly
   - Preserve ALL citation commands ([?], **??**, (**??**), etc.) exactly

- **PRESERVE ALL ···paragraph wrapper tags EXACTLY**
- Maintain academic tone and formal scientific writing style
- Use precise technical terminology - obey the glossary strictly
- Preserve section structure and paragraph organization
- Translate all content completely - do not omit any information

2. **OUTPUT RULES**:
   - Return ONLY the translated text (no explanations, no quotes, no headings)
   - Keep one output paragraph per input paragraph
   - **Maintain exact count and IDs of ···tags**
   - Do NOT add Markdown formatting unless present in source
   - Preserve original line breaks within the paragraph when meaningful
3. **FORMATTING GUIDELINES**:
   - Keep mathematical expressions in original LaTeX format
   - Preserve equation numbers and references
   - Maintain proper academic paragraph structure
   - Use formal scientific language
   - Never remove or modify wrapper tags
4. **Glossary**:
   - 机器学习 => machine learning
   - 深度学习 => deep learning

The input text is:

行动序列作为一种典型的过程数据，可反映被试解决问题的详细步骤。鉴于行动或状态转移可区分正误，

Wait, I need to check if there are any tags in the input. Looking at the input, I don' t see any ···tags. The user said "PRESERVE ALL ···paragraph wrapper tags EXACTLY" , but there are none in the input.

Let me re-read the instructions. The user says "Remember: Mathematical content, citations, and wrapper tags must remain untouched - only translate the Chinese text inside the tags."

But the input doesn' t have any tags. So I should just translate the text as is, following the other rules.

Let me analyze the content: - It' s about action sequences as process data - Proposes two models: 1P-ASM and 2P-ASM based on binary logistic modeling - Compares with SRM (Sequence Response Model) based on multiclass logistic modeling - Findings: similar ability estimates, lower computational cost, 2P-ASM better than 1P-ASM

Key terms to translate: - 行动序列 => action sequences - 过程数据 => process data - 被试 => test takers / examinees / subjects

Full Text

Binary Modeling of Action Sequences in Problem-Solving Tasks: One- and Two-Parameter Action Sequence Models

FU Yanbin, CHEN Qipeng, ZHAN Peida

(School of Psychology, Zhejiang Normal University; Intelligent Laboratory of Child and Adolescent Mental Health and Crisis Intervention of Zhejiang Province; Key Laboratory of Intelligent Education Technology and Application of Zhejiang Province, Jinhua 321004, China)

Abstract

Action sequences, as a typical form of process data, reflect the detailed steps respondents take to solve problems. Given that actions or state transitions can be distinguished as correct or incorrect, this paper proposes two action sequence models with relatively low complexity based on binary logistic modeling—the one- and two-parameter action sequence models (1P-/2P-ASM). The key difference between them lies in whether the discrimination of problem states is freely estimated. Through empirical and simulation studies, we compare the performance of these two new models with the sequential response model (SRM), which is based on multinomial logistic modeling. The main findings are: (1) the two ASMs yield problem-solving ability estimates almost identical to those from the SRM; (2) the computational time for the two ASMs is significantly lower than that for the SRM; and (3) the 2P-ASM demonstrates better overall performance than the 1P-ASM. In summary, both ASMs with relatively low model complexity can effectively analyze action sequences, facilitating the practical application of action sequence data analysis.

Keywords: process data, action sequence, problem state transition, action sequence model, item response theory

---

Problem-solving refers to the process in which individuals, when faced with tasks lacking clear solutions, apply cognitive skills and activities to explore the problem space and transform the problem from an initial state to a goal state through a series of cognitive processing steps (Newell & Simon, 1972). During problem-solving, respondents must construct plans, select strategies, and evaluate whether the execution of these plans can achieve the desired state. Simultaneously, they need to monitor action outcomes against the problem goal, identify errors, and take remedial measures to adjust previous strategies promptly. Therefore, measuring problem-solving competence requires attention not only to final outcomes but also to the sequence of behaviors during the process (Liu et al., 2022). For instance, the Programme for International Student Assessment (PISA) (OECD, 2013) introduced problem-solving tests that simulate real-life scenarios. Through authentic and interactive tasks, these tests

record the dynamic changes in respondents' behaviors throughout the problem-solving process, offering a novel approach to measuring problem-solving competence. These assessments not only record students' final outcomes but also log their step-by-step operations in real-time within log files—what is known as process data. Compared to traditional outcome data, mining and analyzing process data can provide richer information for inferring students' underlying problem-solving abilities.

Current research on analytical methods for process data generated by computerized problem-solving tasks can be categorized into two main types based on research objectives: feature extraction and ability estimation modeling (Han et al., 2022; Xiao & Liu, 2023; Han et al., 2022). Feature extraction can be further divided into theory-driven and data-driven approaches. Theory-driven methods typically employ behavioral indicators defined by experts to score students' problem-solving processes (Harding et al., 2017; Rosen, 2017; Yuan et al., 2019). This approach relies on expert knowledge and represents a top-down feature extraction method. The behavioral indicators identified through theory-driven methods can serve not only as scoring criteria but also as the basis for further modeling analysis using certain measurement models (Liu et al., 2018; Zhan & Qiao, 2022; Zhang et al., 2022). However, this method often requires setting different feature extraction rules for different task contexts, resulting in high application costs. Data-driven approaches involve applying data mining, machine learning, and other algorithms to extract key information from process data, with commonly used methods including natural language processing (Hao et al., 2015; He & von Davier, 2016; He et al., 2021; Zhan et al., 2015), dimensionality reduction algorithms (Tang et al., 2020, 2021), and network analysis methods (Vista et al., 2017; Zhu et al., 2016).

Furthermore, based on whether the models utilize the sequential relationships within action sequences and whether they can obtain continuous and stable ability estimates, ability estimation modeling can be subdivided into three categories: the migration application of traditional psychometric models, stochastic process modeling, and combinations of these two approaches (Han et al., 2022). The migration application of traditional psychometric models primarily involves first extracting key indicators for task completion using feature extraction methods, then encoding respondents' specific operations or action sequences based on these indicators (e.g., coding as 1 if the operation includes a key indicator, otherwise 0), and finally analyzing the encoded data using item response theory (IRT) models or cognitive diagnosis models to estimate respondents' problem-solving abilities (Han & Wilson, 2022; Liu et al., 2018; Wilson et al., 2017; Yuan et al., 2019; Zhan & Qiao, 2022; Zhang et al., 2022; Li et al., 2020). However, this approach partially or completely neglects the sequential information in specific operations. In contrast, existing research has directly applied stochastic process modeling to action sequences, such as dynamic Bayesian networks (Levy, 2019) and hidden Markov models (Arieli-Attali et al., 2019; Bergner et al., 2017; Xiao et al., 2021). Although this method accounts for sequential information in action sequences, the estimated latent variables are typically discrete attributes

or knowledge mastery states, making it impossible to understand respondents' stable and continuous problem-solving abilities (Han et al., 2022). Additionally, some studies have proposed psychometric modeling methods that integrate stochastic process ideas (Chen, 2020; Han et al., 2022; Lamar, 2018; Shu et al., 2017; Xiao & Liu, 2023). Typically, these methods assume that different state transitions or operation transitions satisfy conditional independence given the latent problem-solving ability; for example, treating problem state transition sequences as a discrete stochastic process with first-order Markov properties (Han et al., 2022; Xiao & Liu, 2023), thereby retaining the sequential information itself while inferring continuous latent ability estimates.

To address the limitations of existing methods, Han et al. (2022) combined dynamic Bayesian networks with the nominal response model (NRM) (Bock, 1972) to propose the sequential response model (SRM). The SRM assumes that respondents' problem-solving abilities and the characteristics of state transitions jointly determine the probability of presenting a particular state transition. Compared to existing methods, the SRM not only considers the sequential information in action sequences and the uniqueness of different state transitions within a task but also provides continuous estimates of problem-solving ability, enabling a nuanced understanding of individual differences in problem-solving competence among different respondents. Similar to the NRM, the SRM treats the "multiple-choice questions" at each stage as "nominal response items," assuming that every transition option at each problem state provides measurement information, thereby assigning different parameters to each possible state transition in the task (e.g., transition propensity parameters and transition discrimination parameters). Essentially, the SRM employs multinomial (or polytomous unordered) modeling for state transitions, assuming no quantitative order among all transition options at the next stage. However, in actual problem-solving tasks, actions or state transitions can be classified as correct or incorrect: state transitions that contribute to successful task completion can be defined as correct state transitions, while those that may ultimately lead to task failure can be defined as incorrect state transitions. Therefore, all transition options at each problem state for respondents are dichotomous rather than equivalent relations without quantitative order.

Theoretically, binary modeling is more appropriate for data with correct/incorrect distinctions. Compared to binary modeling, the relative advantage of multinomial modeling (Han et al., 2022; Xiao & Liu, 2023) is that it can incorporate richer measurement information into data analysis, but this inevitably leads to relatively higher model complexity. Higher model complexity typically means more types and quantities of parameters to estimate, greater computational burden for parameter estimation, and lower interpretability of parameter estimation results (Ma et al., 2016). Based on the principle of parsimony in model comparison and selection (Beck, 1943), this study proposes binary modeling for action sequences containing correct/incorrect information, introducing one- and two-parameter action sequence models (1P- and 2P-ASM) to reduce the complexity of action sequence analysis models and increase

computational efficiency. Meanwhile, relatively parsimonious models also help enhance the interpretability of parameter estimation results, thereby increasing the practical usability of action sequence models.

First, we elaborate on the foundation of action sequence modeling. Second, we introduce our two new models: 1P-ASM and 2P-ASM. Third, based on empirical study data, we compare the parameter estimation results of the two new models with those of the SRM to demonstrate the practical applicability of the new models and the degree of consistency with the SRM' s parameter estimation results. Fourth, through simulation studies, we explore the psychometric properties of the two new models under different simulated test conditions. Finally, we summarize the research findings and discuss study limitations and future research directions.

2. Action Sequence Modeling Foundation and the SRM

2.1 Foundation of Action Sequence Modeling

This study focuses on well-defined tasks with clear objectives and complete information, which are typically constructed using finite state automata (FSA) as prototypes. Such tasks have finite problem states, finite user input signals (i.e., actions or operations), and produce corresponding output signals through user operations, meaning they have explicit state transition rules (Buchner & Funke, 1993). Figure 1: see original paper presents an example of an FSA problem-solving task, which includes six problem states: S, A, B, C, D, and E. Here, S represents the initial problem state, E represents the goal state, and the others are intermediate states. Since this task allows respondents to return to the initial state from any intermediate state, theoretically, multiple action sequences can emerge, such as S→A→C→E, S→B→S→A→C→E, and S→B→D→E. Among these action sequences, the shortest sequence that achieves the task goal is defined as the optimal state transition sequence or optimal action sequence. For instance, the optimal state transition sequence S→A→C→E contains three state transitions: S→A, A→C, and C→E. In the figure, red solid arrows indicate correct state transitions (i.e., those that facilitate correct problem-solving), while black dashed arrows indicate incorrect state transitions (i.e., those that may ultimately lead away from the task goal).

In practice, we can treat respondents' action transitions at each problem state as answering a "multiple-choice question." Figure 1: see original paper shows the problem-solving flowchart corresponding to Figure 1: see original paper. When a respondent is at problem state S in stage 1, they must choose between two problem states, A and B, in stage 2. Similarly, when at problem state A in stage 2, they must choose among three problem states, C, D, and S, in stage 3 (where S represents returning to the initial state). At this point, we can migrate traditional IRT models, originally designed for analyzing response accuracy at the item level, to this context. For example, Han et al. (2022) applied the NRM in this way, proposing the SRM based on multinomial modeling.

2.2 Introduction to the SRM

Assume a problem-solving task contains R discrete problem states, with the set of problem states denoted as $\mathcal{S}$. Let $x_{np} \in \mathcal{S}$ represent the problem state that student n ($n = 1, \ldots, N$) is in at stage p ($p = 1, \ldots, P_n$), where N is the number of respondents and P_n is the length of the final action sequence presented by student n (which may vary across respondents). Figure 2: see original paper presents the logical diagram of the SRM, which assumes that respondents' problem-solving abilities influence their state transitions between adjacent stages. Figure 2: see original paper shows the modeling diagram of the SRM, which essentially models state transitions by assuming that respondents' problem-solving abilities affect the probability of presenting a specific state transition.

The SRM can be expressed as:

$$P(x_{n,p+1} = k \mid x_{np} = j, \theta_n) = \frac{\exp(\lambda_{jk} + I_{jk}\theta_n)}{\sum_{k' \in \mathcal{S}_j} \exp(\lambda_{jk'} + I_{jk'}\theta_n)}, \quad (1)$$

where $x_{n,p+1} \leftarrow x_{np}$ represents the observed state transition (i.e., the state transition presented by respondent n between adjacent stages p and $p+1$), j and k denote the current problem state and the selectable problem state at the next stage, respectively, $\mathcal{S}_j$ represents the set of all possible problem states that can appear at the next stage when currently at state j, θ_n is the problem-solving ability of respondent n, λ_{jk} is the state transition propensity parameter indicating the tendency to transition from state j to state k (higher values indicate that transition $j \to k$ is more likely to be presented), and I_{jk} is the state transition discrimination parameter (higher values indicate that transition $j \to k$ better discriminates problem-solving ability). The SRM assumes that, given a respondent' s latent ability, the state transitions presented at adjacent stages satisfy conditional independence. Consequently, the joint probability of the final state transition vector $\mathbf{x}_n$ presented by respondent n is:

$$P(\mathbf{x}_n \mid \theta_n) = \prod_{p=1}^{P_n - 1} P(x_{n,p+1} \mid x_{np}, \theta_n). \quad (2)$$

As a multinomial model, each state transition in the SRM contains two parameters, λ_{jk} and I_{jk}. Using Figure 1: see original paper as an example, the SRM treats the "multiple-choice question" at each stage as a "nominal response item," assuming that every option provides measurement information, resulting in a total of 22 parameters: 11 transition propensity parameters (e.g., λ_{SA}, λ_{SB}, λ_{AS}, λ_{AC}, λ_{BD}, and λ_{DE}) and 11 corresponding transition discrimination parameters. To ensure model identifiability and reduce the number of parameters to estimate, Han et al. (2022) imposed certain constraints on the SRM: (1) constraining the sum of transition propensity parameters from the current problem state j to all

optional problem states at the next stage to zero, i.e., $\sum_{k\in\mathcal{S}_j}\lambda_{jk}=0$; and (2) pre-fixing the transition discrimination parameters: $I_{jk}=1$ if $j\to k$ is a correct state transition, and $I_{jk}=-1$ if $j\to k$ is an incorrect state transition.

3. Binary Modeling of Action Sequences: 1P-ASM and 2P-ASM

3.1 Model Construction

Although the SRM employs multinomial modeling to incorporate all measurement information provided by action sequences, it still differentiates between correct and incorrect state transitions in action sequences through a pre-specified state transition discrimination parameter. For state transitions with correct/incorrect distinctions, this study adopts a binary modeling approach, using IRT models for dichotomously scored data to model action sequences, such as the one-parameter IRT model/Rasch model (Rasch, 1960) and the two-parameter IRT model (Birnbaum, 1968). Accordingly, [Figure 3: see original paper] presents the binary coding diagram for the problem-solving task corresponding to [Figure 1: see original paper], where correct state transitions are coded as 1 and incorrect state transitions as 0. In Figure 3: see original paper, we can treat the "multiple-choice question" at each stage as a "multiple-choice question with a correct answer," allowing us to draw on traditional dichotomously scored IRT models to construct action sequence models.

[Figure 4: see original paper] presents the modeling diagrams for the two ASMs. First, all state transitions in the task are binary-coded: correct state transitions are coded as 1, and incorrect state transitions as 0. The state transition vector presented by a respondent during problem-solving is thus encoded as a binary vector containing only 0 or 1 elements. For example, the optimal action sequence S→A→C→E in [Figure 1: see original paper] corresponds to the state transition vector $(SA, AC, CE)'$, which can be converted to $(1,1,1)'$. Then, based on dichotomously scored IRT models, we assume that respondents' problem-solving abilities influence the probability of presenting correct state transitions.

Drawing on the one-parameter IRT model, the 1P-ASM can be expressed as:

$$P(Y_{n,j\to k}=1\mid\theta_n)=\frac{\exp(\beta_j+\theta_n)}{1+\exp(\beta_j+\theta_n)},\quad(3)$$

where $Y_{n,j\to k}=1$ indicates that respondent n presented a correct state transition between adjacent stages, β_j is the action easiness parameter, representing the ease of presenting a correct state transition when at state j, and other parameters have the same meanings as above.

Drawing on the two-parameter IRT model, the 2P-ASM can be expressed as:

$$P(Y_{n,j\to k}=1 \mid \theta_n)=\frac{\exp[\gamma_j(\beta_j+\theta_n)]}{1+\exp[\gamma_j(\beta_j+\theta_n)]}, \quad (4)$$

where γ_j is the action discrimination parameter, representing the degree to which presenting a correct state transition at state j discriminates problem-solving ability, and other parameters have the same meanings as above.

Following the local independence assumption of the SRM, the ASMs also assume that state transitions presented at adjacent stages satisfy conditional independence given respondents' latent abilities. Consequently, the joint probability of the final binary state transition vector $\mathbf{Y}_n$ presented by respondent n is:

$$P(\mathbf{Y}_n \mid \theta_n)=\prod_{p=1}^{P_n-1} P(Y_{n,x_{np}\to x_{n,p+1}} \mid \theta_n). \quad (5)$$

3.2 Comparison with Related Models

First, consistent with the SRM, the ASMs also represent psychometric modeling methods that integrate stochastic process ideas. The primary difference between them lies in their modeling logic: the former is a binary logistic model, while the latter adopts a multinomial logistic model. This difference in modeling logic leads to differences not only in model complexity but also in parameter interpretation. For instance, if we treat the transition propensity parameters in the SRM as "option-level" parameters, then the action easiness parameters in the two ASMs are "item-level" parameters. The former 刻画 the propensity to select a particular option (i.e., the propensity to present a particular state transition), while the latter 刻画 the easiness of correctly answering the item (i.e., the easiness of presenting a correct state transition). Additionally, to reduce the number of parameters to estimate, the SRM pre-fixes the state transition discrimination parameters, whereas the action discrimination parameters in the 2P-ASM are freely estimated, reflecting the degree to which different problem states (or "items") discriminate respondents' problem-solving abilities. Notably, due to the additional parameter constraints in the SRM, the number of parameters to estimate is not always greater than that of the 1P-ASM and 2P-ASM. Using problem state C in stage 3 of Figure 1: see original paper as an example, when there are only two transition options at the next stage (E and S), the SRM also requires estimating only one transition propensity parameter due to the constraint $\lambda_{CE}=-\lambda_{CS}$. At this point, the 1P-ASM also requires estimating only one action easiness parameter, while the 2P-ASM additionally requires estimating one action discrimination parameter. Of course, the number of parameters to estimate in the SRM increases with the number of transition options at the next stage, whereas the ASMs do not. Due to space limitations, the comparison between the ASMs and other models is provided in Online Appendix 1.

3.3 Bayesian Parameter Estimation

Like the SRM, the two ASMs can also be estimated using full Bayesian Markov chain Monte Carlo (MCMC) algorithms. Details are provided in Online Appendix 7.

4. Empirical Data Analysis

4.1 Task Description

Consistent with Han et al. (2022), this study also selected action sequence data from the PISA 2012 computer-based problem-solving "Tickets"task (CP038Q02) for analysis. This task requires respondents to operate a virtual ticket machine to purchase a full-fare suburban train ticket that can be used for two trips. [Figure 5: see original paper] presents the initial interface of the task, and screenshots of various stages during problem-solving are provided in Online Appendix 2. To solve the problem, respondents must first select between "city metro" or "suburban train" as the transportation mode. Second, based on the selected transportation mode, they must choose between "full-fare ticket" and "discounted ticket." Then, depending on the selected fare type, they must choose between purchasing a "day pass" or "trip ticket" ; if "trip ticket" is selected, they must also choose the number of trips (from 1 to 5). Finally, making the "purchase" decision completes the task. Respondents can return to the initial interface at any point by clicking "cancel." Consequently, the lengths of action sequences presented by different respondents to solve this task vary.

[Figure 6: see original paper] shows the problem structure of this task after decomposition, which includes 11 problem states: $\{S, A, B, C, D, E, F, G, H, I, J\}$, where S is the starting state, J is the terminal state, and the rest are intermediate states. Between two adjacent problem states, solid lines indicate correct state transitions (e.g., SA), while dashed lines indicate incorrect state transitions (e.g., SF). The optimal action sequence for this task is "Start (S) $\rightarrow$ Correct transportation type (A) $\rightarrow$ Correct discount type (B) $\rightarrow$ Correct ticket type (C) $\rightarrow$ Correct number of trips (D) $\rightarrow$ Purchase (J)," with the corresponding clicks being "Country trains" $\rightarrow$ "Full fare" $\rightarrow$ "Trip ticket" $\rightarrow$ "2 trips" $\rightarrow$ "Purchase."

further organizes the operation process from [Figure 6: see original paper] from the "multiple-choice question" perspective. The problem state at the current stage can be viewed as a "multiple-choice question" that the respondent needs to answer, with the optional problem states at the next stage treated as "options." For example, at the initial stage, respondents must choose between two "options," A and F, for "question S," where A is the correct "option" and F is the incorrect "option." For these "multiple-choice questions," the SRM treats them as nominal response items, while the ASMs treat them as dichotomously scored multiple-choice questions. For instance, if a student' s action sequence is $SABCDEDJ$, the SRM analyzes the state transition vector $(SA, AB, BC, CD, DE, ED, DJ)'$, whereas the ASM analyzes the dichotomous

state transition vector $(1, 1, 1, 1, 0, 1, 1)'$.

4.2 Data Preparation and Analysis

The raw data were downloaded from the PISA official website. Before conducting specific data analyses, the raw data were recoded according to the task structure defined in [Figure 6: see original paper], and the data were cleaned by: (1) removing action sequences that terminated early (i.e., sequences without a "purchase" click), and (2) deleting action sequences containing impossible state transitions (see Online Appendix 3, Table A2). Ultimately, action sequences were extracted from log files for 28,851 respondents, with sequence lengths ranging from 5 to 110 (mean = 6.992). The raw data contained 1,395 distinct action sequences, of which 569 sequences achieved the task goal (involving 15,408 respondents: 10,610 respondents completed the task using the optimal action sequence, while 4,798 respondents had error-correction processes during correct problem-solving). Finally, due to computational constraints and to enhance research efficiency, we used simple random sampling to select action sequences from 2,000 students for empirical analysis in this study (sequence length range: 5–46, mean = 7.03; containing 1,395 distinct action sequences, of which 569 achieved the task goal, involving 1,068 students, with 737 completing the task using the optimal action sequence).

We analyzed the data using the 1P-ASM, 2P-ASM, and SRM. For parameter estimation, we used 2 Markov chains, each with 5,000 iterations and a burn-in period of 3,000 iterations. The Potential Scale Reduction Factor (PSRF; Gelman & Rubin, 1992) was used to determine whether the MCMC algorithm had converged; PSRF values less than 1.1 indicate convergence. Additionally, we employed two fully Bayesian relative fit indices—the Watanabe-Akaike Information Criterion (WAIC; Watanabe, 2010) and leave-one-out cross-validation (LOO; Vehtari et al., 2017)—to evaluate model fit and provide evidence for model selection; smaller values indicate better model fit. Notably, because the SRM and ASMs analyze different data (the former analyzes each student' s state transition vector, while the latter analyzes the dichotomized version of this vector), their relative fit values cannot be directly compared. Therefore, we can only use relative fit indices to judge the relative fit between the two ASMs, not to compare the fit between ASMs and the SRM. To demonstrate that binary modeling achieves comparable performance to multinomial modeling, we calculated the consistency between parameter estimation results from the ASMs and the SRM. Furthermore, we used posterior predictive checking (PPC; Gelman et al., 1996) to assess absolute model fit; if a model fits the data, its posterior predictive probability (ppp) will be close to 0.5, whereas if the model does not fit, the ppp value will be < 0.025 or > 0.975. The statistics used for PPC in this study are provided in Online Appendix 4, Table A3.

All parameters in all models achieved PSRF values less than 1.05, indicating that parameter estimation reached convergence under our settings. Additionally, Online Appendix 5 provides sampling trace plots for model parameters.

presents the model fit and computational time for the three models. First, the ppp values for all three models were close to 0.5, indicating that all models fit the data adequately. Second, the two relative fit indices showed that the 2P-ASM fit the data better than the 1P-ASM, suggesting that accounting for state transition discrimination better reflects the characteristics of this data—namely, that different state transitions have varying abilities to discriminate problem-solving competence. As mentioned earlier, the relative fit results of ASMs and SRM are not comparable. Finally, parameter estimation computational time comprehensively reflects model complexity; results showed that the SRM required the longest time, followed by the 2P-ASM, with the 1P-ASM being the shortest, confirming that binary models are indeed more parsimonious than multinomial models. The main results below focus on the two ASMs and present the consistency between the ASMs and SRM in estimating respondents' problem-solving abilities.

presents the item parameter estimation results for the two ASMs (posterior means, posterior standard deviations, and 95% highest posterior density [Bayesian credible intervals]). First, regarding the action easiness parameters, the posterior means of the easiness parameters for problem states on the correct problem-solving path (i.e., the optimal action sequence) (S, A, B, C, and D) were all greater than 0 (the posterior mean for state D in the 2P-ASM was not significantly different from zero), indicating that when respondents are on the correct path, they are more likely to continue presenting correct state transitions. Conversely, the posterior means of the easiness parameters for problem states on the incorrect path (F, G, H, and I) were all less than 0 (the posterior mean for state I in the 1P-ASM and for states H and I in the 2P-ASM were not significantly different from zero), indicating that when respondents are already on an incorrect path, they are less likely to correct their errors and turn to correct problem states (i.e., they are more likely to remain on the incorrect path). Notably, problem states E and I are error-path states both representing "selecting the wrong number of trips." Compared to other error-path states, the easiness estimates for E and I were higher, suggesting that when respondents are in these two error states, they are more likely to correct their mistakes in the next selection (i.e., choose S to return to the initial state and start over). Second, regarding the action discrimination parameters, there were some differences in discrimination across problem states. The posterior means of the discrimination parameters for states C and I were relatively high, indicating that respondents with different problem-solving abilities showed relatively large differences in the probability of presenting correct state transitions in these two states. In other words, whether students already on the correct path can select the correct number of trips, and whether students already on the incorrect path can correct their errors through "cancel," are two operations that strongly discriminate student ability. Overall, the parameter estimates reveal that when respondents are on the correct problem-solving path, they tend to remain on it; when they are on an incorrect path, they tend to continue making errors until reaching the interface for selecting the number of trips, where there is a critical

opportunity to correct mistakes.

[Figure 7: see original paper] presents scatter plots and probability density plots comparing the problem-solving ability estimates (posterior means) from the three models. First, the scatter plots show high consistency among the ability estimates from the three models (all correlation coefficients above 0.99), indicating that they measure the same latent trait and that binary modeling, like multinomial modeling, can measure respondents' problem-solving abilities through action sequence data and reflect individual differences. Second, comparing the probability density plots of the three models reveals largely consistent distributions in both high- and low-ability regions, with only slight differences in the medium-ability region (primarily for the SRM). One possible reason is that the SRM more fully utilizes the measurement information provided by different state transitions: it uses not only the information from correct state transitions but also the information from different incorrect state transitions. For example, when multiple respondents are simultaneously at problem state A, those who select the incorrect "option" S may appear to have higher problem-solving ability than those who select the incorrect "option" G. The SRM can distinguish between respondents who present AG and those who present AS, whereas the ASM treats them both as individuals who made an incorrect choice.

We selected action sequences with frequencies greater than 20 from the analysis data as typical action sequences (covering 80.1% of respondents). presents descriptive statistics for problem-solving ability estimates from the three models for these typical action sequences (sorted by the mean ability estimate from the SRM from high to low). First, the three models show consistent descriptive statistics for ability estimates across typical action sequences. For instance, respondents presenting the optimal action sequence $SABCDJ$ have the highest mean ability estimate, while those presenting the poorest action sequence $SFGHIJ$ have the lowest mean ability estimate. Second, overall, across typical action sequences, the more correct problem states appear and the fewer error problem states appear, the higher the mean ability estimate; conversely, the lower the mean ability estimate. Comparing the ASMs and SRM, we found that the ASMs produced a different ranking of mean ability estimates for two sequences compared to the SRM: the mean ability estimate for $SABCEDJ$ was slightly lower than that for $SFGHSABCDJ$. Respondents presenting $SABCEDJ$ made an error transition at state C (CE) but immediately corrected it (ED), whereas those presenting $SFGHSABCDJ$ made an error transition at the initial state and only corrected it when selecting the number of trips. The ranking difference between the ASMs and SRM for these two sequences can be interpreted from different perspectives. First, from the perspective of the number of error states or problem-solving efficiency (sequence length), respondents presenting $SABCEDJ$ should have higher mean ability estimates than those presenting $SFGHSABCDJ$; the SRM' s ranking supports this interpretation. Second, combining the action easiness parameters in , problem state C has high easiness (low difficulty), while states F, G, and H have low easiness (high difficulty). Therefore, from the perspective of the negative impact

or penalty of errors on ability estimation, the penalty for an error at state C is greater than that for errors at states F, G, and H, leading to lower mean ability estimates for respondents presenting $SABCEDJ$ compared to those presenting $SFGHSABCDJ$; the ASMs' ranking supports this interpretation.

Finally, given that relative fit indices cannot compare the fit of ASMs versus the SRM, we conducted logistic regression using the three models' problem-solving ability estimates to predict response accuracy data for the task (based on the task's scoring rule: 1 point for purchasing the correct ticket, 0 otherwise). Following logistic regression requirements, the independent variable (ability estimate) was standardized. The standardized regression coefficients for ability estimates from the SRM, 1P-ASM, and 2P-ASM were 15.285, 14.999, and 15.387, respectively (all significant at $p < 0.001$), indicating that ability estimates from all three models significantly affect task completion, with largely equivalent effect sizes (the 2P-ASM showing the largest effect, followed by the SRM, then the 1P-ASM). Additionally, the R^2 values from the regression equations for the SRM, 1P-ASM, and 2P-ASM were 0.929, 0.958, and 0.959, respectively, indicating that the models' ability estimates explain a high proportion of variance in observed data and can accurately predict students' performance on the task, with the 2P-ASM explaining the most variance, followed by the 1P-ASM, and the SRM explaining the least.

5. Simulation Study

5.1 Research Design, Data Generation, and Analysis

A simulation study was conducted to further explore the psychometric performance of the two ASMs under ideal testing conditions. It is important to emphasize that the ASMs themselves cannot generate the action sequences presented by respondents in solving tasks (they can only generate 0-1 vectors). Therefore, the SRM was used as the data-generating model for action sequences in the simulation study. The problem-solving task structure from the empirical study ([Figure 6: see original paper]) was used to generate action sequence data. The simulation study manipulated two variables: sample size (100, 200, and 500) and action sequence length (short and long). Following Han et al. (2022) and Fu et al. (2022), we manipulated action sequence length in the SRM by adjusting the transition propensity parameters for the "cancel" operation (e.g., $A \rightarrow S$): larger parameter values produce longer sequences. Detailed steps for generating action sequences are provided in Online Appendix 6. Ultimately, the generated short and long action sequences had average lengths of approximately 10.5 and 20.2, respectively. Additionally, to reduce the impact of random error, we repeated the data generation process 50 times for each of the six simulation conditions.

For the generated data, we estimated parameters using the SRM, 1P-ASM, and 2P-ASM, following the same procedure as in the empirical study, with PSRF used as the convergence criterion. Since the SRM and ASMs have different mod-

eling logics, only the problem-solving ability parameters have the same meaning and are comparable; other parameters have different meanings and cannot be compared. Therefore, we evaluated model performance for problem-solving ability estimates from two perspectives. First, regarding parameter estimation accuracy, we used Bias and root mean square error (RMSE) to examine the recovery of problem-solving ability parameters across the three models:

$$\text{Bias}_r = \frac{1}{N}\sum_{n=1}^{N}(\hat{\theta}_{nr} - \theta_{nr}), \quad \text{RMSE}_r = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(\hat{\theta}_{nr} - \theta_{nr})^2},$$

where θ_{nr} and $\hat{\theta}_{nr}$ represent the "true" and estimated ability parameters for respondent n in replication r, respectively. Additionally, we calculated the correlation coefficient Cor between true and estimated values.

Second, regarding parameter estimation consistency, we used consistency bias (CBias) and consistency root mean square error (CRMSE) to examine the agreement between ability estimates from the two ASMs and those from the data-generating SRM:

$$\text{CBias}_r = \frac{1}{N}\sum_{n=1}^{N}(\hat{\theta}_{nr}^{\text{ASM}} - \hat{\theta}_{nr}^{\text{SRM}}), \quad \text{CRMSE}_r = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(\hat{\theta}_{nr}^{\text{ASM}} - \hat{\theta}_{nr}^{\text{SRM}})^2},$$

where $\hat{\theta}_{nr}^{\text{SRM}}$ and $\hat{\theta}_{nr}^{\text{ASM}}$ represent ability estimates from the SRM and ASM in replication r, respectively. We also calculated the correlation coefficient $Ccor$ between the two models' estimates.

Additionally, we computed the average runtime (ART) for parameter estimation across conditions to evaluate model efficiency and reflect model complexity:

$$\text{ART} = \frac{1}{R}\sum_{r=1}^{R} T_r,$$

where T_r represents the computational time for parameter estimation in replication r. To ensure comparability of computational time results, all programs for the three models were run on the same server (Intel® Xeon® Gold 6266C CPU @ 3.00 GHz and 64 GB RAM).

First, across all conditions, all parameters in the three models achieved PSRF values less than 1.1, indicating convergence.

presents the recovery and computational time for problem-solving ability parameters across the three models under different simulation conditions. First, sample size had minimal impact on the recovery of ability parameters; longer

average sequence length led to better recovery. From another perspective, sequence length reflects the number of "item" samples—the longer the sequence, the larger the item sample size, and the more accurate the inference about respondents' ability values. Second, as the data-generating model, the SRM should theoretically show the best recovery, followed by the 2P-ASM, then the 1P-ASM. However, the overall differences were minimal (in most conditions, the RMSE for the 1P-ASM was less than 0.05 higher than that of the SRM, and the Cor was less than 0.02 lower). Finally, across all conditions, the 1P-ASM required the shortest computational time, followed by the 2P-ASM, with the SRM requiring the longest time. This result aligns with the empirical study findings, demonstrating that binary modeling significantly reduces parameter estimation time compared to multinomial modeling while only slightly decreasing estimation precision.

presents the consistency between the two ASMs and the SRM in problem-solving ability parameter estimation across different simulation conditions. Overall, both ASMs showed high consistency with the SRM, with the 2P-ASM demonstrating higher consistency than the 1P-ASM. Notably, as sequence length increased, the consistency between the 1P-ASM and SRM decreased slightly (longer sequences produce greater differences in discrimination across problem states), while the consistency between the 2P-ASM and SRM increased slightly. A possible reason is that the 1P-ASM is relatively simple, constraining all problem states to have equal discrimination. When sequences are short (fewer "items"), the negative impact of this constraint is smaller than when sequences are long. The 2P-ASM is relatively complex, freely estimating discrimination for all problem states. As sequence length increases, the differences in discrimination across problem states increase, better aligning with the 2P-ASM' s assumptions.

6. Summary and Discussion

Compared to traditional response accuracy data, process data such as action sequences can provide richer information about how respondents solve problems. However, the non-standardized format of action sequence data (i.e., varying data lengths across respondents) also poses challenges for the direct application of traditional psychometric models. To address the limitations of existing methods, Han et al. (2022) combined dynamic Bayesian networks with the NRM to propose the SRM. Similar to the NRM, the SRM uses multinomial logistic modeling, assigning different parameters to each possible state transition in the task, resulting in high model complexity. Given that state transitions in problem-solving tasks have correct/incorrect distinctions rather than equivalence relations without quantitative order, this paper proposes two action sequence models based on binary logistic modeling with relatively low model complexity—the 1P-ASM and 2P-ASM. Unlike the SRM, which migrates the NRM to action sequence analysis, the 1P-ASM and 2P-ASM migrate the simpler one- and two-parameter IRT models to action sequence analysis, respectively.

Empirical study results found that: (1) the problem-solving ability estimates

from the two ASMs and the SRM had correlations close to 1, indicating they measure the same latent trait; (2) the computational time for the two ASMs was significantly lower than that for the SRM, suggesting lower model complexity for the ASMs; (3) parameter estimates revealed task characteristics in this study: when respondents are already on the correct problem-solving path, they tend to remain on it, whereas when they are on an incorrect path, they tend to continue making errors; and (4) unlike the SRM and 1P-ASM, which fix discrimination parameters, the 2P-ASM can provide discrimination parameters for presenting correct state transitions at current problem states, helping to identify relatively important problem states (such as states C and I in the empirical study) and enabling data analysts to better understand the task itself.

Simulation study results found that: (1) even when not the data-generating model, the two ASMs could provide accurate parameter estimates; (2) the computational time for both ASMs was lower than that for the SRM, with the relative advantage being more pronounced under small sample size conditions; (3) the problem-solving ability estimates from both ASMs showed high agreement with those from the SRM, with the agreement between the 2P-ASM and SRM being relatively higher; and (4) the length of the problem state transition sequences presented by respondents was a primary factor affecting parameter estimation recovery for both ASMs and the SRM: longer sequences contain more information, leading to more accurate estimation of problem-solving abilities.

In summary, the two ASMs proposed in this paper based on binary logistic modeling can effectively analyze action sequence data, providing almost identical estimates of respondents' problem-solving abilities as the SRM while significantly reducing computational time. Combining the results of the simulation and empirical studies, we believe the 2P-ASM demonstrates better overall performance than the 1P-ASM. However, the more parsimonious 1P-ASM is recommended when sample size is small (e.g., 100 respondents) or the task is simple (requiring fewer operations to solve).

Of course, as binary models, the ASMs have certain theoretical limitations compared to the SRM. For example, before analyzing action sequence data with ASMs, the sequences must be binary-coded, treating all incorrect state transitions as "equivalent," which inevitably loses the differentiated information provided by different incorrect state transitions. Additionally, because ASMs model dichotomously coded action sequence data, we cannot generate action sequences by specifying model parameters.

Although this paper proposes two models that can effectively analyze action sequence data, several limitations warrant further investigation. First, like the SRM, the ASMs assume that respondents' problem-solving ability is unidimensional. However, some problem-solving tasks may require respondents to use multiple dimensions of problem-solving ability. Future research could attempt to propose multidimensional action sequence models (Shu et al., 2017). Second, in addition to recording the problem states respondents are in during each problem-solving stage, process data also record timestamps for each stage. Us-

ing timestamp information, we can calculate the time respondents spend on each state transition, i.e., action times (Fu et al., 2022). Currently, in item-level data analysis, numerous studies have analyzed item response times and jointly analyzed them with item response accuracy data (e.g., van der Linden, 2006, 2007; Man et al., 2022; Peng et al., 2022; Zhan et al., 2018, 2022). Future research could attempt to combine action time data with action sequence data to further mine the rich information contained in process data (Fu et al., 2022). Third, respondents must select one option from the transition alternatives at the next stage to continue the task; when unsure which option to choose, they may select randomly. Future research could attempt to migrate three-parameter IRT models that include guessing parameters to address potential guessing issues in action sequence data. Finally, due to space, time, and resource constraints, the simulation study manipulated a limited number of variables or levels, 未能充分挖掘 ASM 在不同理想测验条件下的表现. Future research could manipulate additional variables (e.g., task complexity [including more problem states]) to further explore the psychometric performance of ASMs.

References

Arieli-Attali, M., Ou, L., & Simmering, V. R. (2019). Understanding test takers' choices in a self-adapted test: A hidden Markov modeling of process data. *Frontiers in Psychology*, 10, 83.

Beck, L. W. (1943). The principle of parsimony in empirical science. *The Journal of Philosophy*, 40(23), 617−633. https://doi.org/10.2307/2019692

Bergner, Y., Walker, E., & Ogan, A. (2017). Dynamic Bayesian network models for peer tutoring interactions. In A. A. von Davier, M. Zhu, & P. C. Kyllonen (Eds), *Innovative assessment of collaboration* (pp. 249−268). Cham: Springer.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee' s ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397−124). Reading, MA: Addison-Wesley.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 37(1), 29−51. https://doi.org/10.1007/BF02291411

Buchner, A., & Funke, J. (1993). Finite-state automata: Dynamic task environments in problem-solving research. *The Quarterly Journal of Experimental Psychology*, 46(1), 83−118.

Chen, Y. (2020). A continuous-time dynamic choice measurement model for problem-solving process data. *Psychometrika*, 85(4), 1052−1075.

Fu, Y., Zhan, P., Chen, Q., & Jiao, H. (2022). Joint modeling of action sequences and action times in problem-solving tasks. *PsyArXiv*. Retrieved from psyarxiv.com/e3nbc

Gelman, A., Meng, X.-L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 6, 733−760.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7, 457−511.

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), 911−932.

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.

Han, Y., Xiao, Y., & Liu, H. (2022). Feature extraction and ability estimation of process data in the problem-solving test. *Advances in Psychological Science*, 30(6), 1393−1409.

Hao, J., Shu, Z., & von Davier, A. (2015). Analyzing process data from game/scenario-based tasks: An edit distance approach. *Journal of Educational Data Mining*, 7(1), 33−50.

Harding, S. M. E., Griffin, P. E., Awwal, N., Alom, B. M., & Scoular, C. (2017). Measuring collaborative problem solving using mathematics-based tasks. *AERA Open*, 3(3), 1−19.

He, Q., Borgonovi, F., & Paccagnella, M. (2021). Leveraging process data to assess adults' problem-solving skills: Using sequence mining to identify behavioral patterns across digital tasks. *Computers & Education*, 166, 104170.

He, Q., & von Davier, M. (2016). Analyzing process data from problem-solving items with N-grams: Insights from a computer-based large-scale assessment. In R. Yigal, F. Steve, & M. Maryam (Eds.), *Handbook of research on technology tools for real-world skill development* (pp. 749−776). Hershey, PA: Information Science Reference.

Hoffman, M. D., & Gelman, A. (2014). The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(1), 1593−1623.

LaMar, M. M. (2018). Markov decision process measurement model. *Psychometrika*, 83(1), 67−88.

Levy, R. (2019). Dynamic Bayesian network modeling of game-based diagnostic assessments. *Multivariate Behavioral Research*, 54(6), 771−794.

Li, M., Liu, Y., & Liu, H. (2020). Analysis of the problem-solving strategies in computer-based dynamic assessment: The extension and application of multilevel mixture IRT model. *Acta Psychologica Sinica*, 52(4), 528−540.

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, 1372.

Liu, Y., Xu, H., Chen, Q., & Zhan, P. (2022). The measurement of problem-solving competence using process data. *Advances in Psychological Science*, 30(3), 522−535.

Ma, W., Iaconangelo, C., & de la Torre, J. (2016). Model similarity, model selection, and attribute classification. *Applied Psychological Measurement*, 40(3), 200−217.

Man, K., Harring, J. R., & Zhan, P. (2022). Bridging models of biometric and psychometric assessment: A three-way joint modeling approach of item responses, response times and gaze fixation counts. *Applied Psychological Measurement*, 46(5), 361−381.

Newell, A., & Simon, H. A. (1972). *Human problem solving* (Vol. 104, No. 9). Englewood Cliffs, NJ: Prentice-hall.

OECD. (2013). *PISA 2012 assessment and analytical framework: Mathematics, reading, science, problem solving and financial literacy*. OECD Publishing. http://dx.doi.org/10.1787/9789264190511-en

Peng, S., Cai, Y., Wang, D., Luo, F., & Tu, D. (2022). A generalized diagnostic classification modeling framework integrating differential speediness: Advantages in psychological and educational testing. *Multivariate Behavioral Research*, 57(6), 940−959.

Rasch, G. (1960). On general laws and the meaning of measurement in psychology. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability: Held at the Statistical Laboratory, University of California, June 20-July 30, 1960* (Vol. 4, p. 321). University of California Press.

Rosen, Y. (2017). Assessing students in human-to-agent settings to inform collaborative problem-solving learning. *Journal of Educational Measurement*, 54(1), 36−53.

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.

Tang, X., Wang, Z., He, Q., Liu, J., & Ying, Z. (2020). Latent feature extraction for process data via multidimensional scaling. *Psychometrika*, 85(2), 378−397.

Tang, X., Wang, Z., Liu, J., & Ying, Z. (2021). An exploratory analysis of the latent structure of process data via action sequence autoencoders. *British Journal of Mathematical and Statistical Psychology*, 74(1), 1−33.

van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics*, 31(2), 181−204.

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, 72(3), 287−308.

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

Vista, A., Care, E., & Awwal, N. (2017). Visualising and examining sequential actions as behavioural paths that can be interpreted as markers of complex behaviours. *Computers in Human Behavior*, 76, 656−671.

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11(12), 3571−3594.

Wilson, M., Gochyyev, P., & Scalise, K. (2017). Modeling data from collaborative assessments: In digital interactive social networks. *Journal of Educational Measurement*, 54(1), 85−102.

Xiao, Y., He, Q., Veldkamp, B., & Liu, H. (2021). Exploring latent states of problem-solving competence using hidden Markov model on process data. *Journal of Computer Assisted Learning*, 37(5), 1232−1247.

Xiao, Y., & Liu, H. (2023). A state response measurement model for problem-solving process data. *Behavior Research Methods*, Online First.

Yuan, J., Xiao, Y., & Liu, H. (2019). Assessment of collaborative problem solving based on process stream data: A new paradigm for extracting indicators and modeling dyad data. *Frontiers in Psychology*, 10, 369.

Zhan, P., Jiao, H., & Liao, D. (2018). Cognitive diagnosis modeling incorporating response times. *British Journal of Mathematical and Statistical Psychology*, 71(2), 262−286.

Zhan, P., Man, K., Wind, S. A., & Malone, J. (2022). Cognitive diagnosis modeling incorporating response times and fixation counts: Providing comprehensive feedback and accurate diagnosis. *Journal of Educational and Behavioral Statistics*, 47(6), 736−776.

Zhan, P., & Qiao, X. (2022). Diagnostic classification analysis of problem-solving competence using process data: An item expansion method. *Psychometrika*, 87(4), 1529−1547.

Zhan, S., Hao, J., & Davier, A. V. (2015). Analyzing process data from game/scenariobased tasks: An edit distance approach. *Journal of Educational Data Mining*, 7(1), 33−50.

Zhang, S., Wang, Z., Qi, J., Liu, J., & Ying, Z. (2022). Accurate assessment via process data. *Psychometrika*, 88(1), 76−97.

Zhu, M., Shu, Z., & von Davier, A. A. (2016). Using networks to visualize and analyze process data for educational assessment. *Journal of Educational Measurement*, 53(2), 190−211.

Online Appendices

Appendix 1. Comparison of ASMs with Existing Models

In addition to the SRM, the state response model proposed by Xiao and Liu (2023) also employs multinomial modeling. Table A1 presents a comparison among the SRM, state response model, and ASM. First, since both the SRM and state response model use multinomial modeling, both involve the occurrence probabilities of each "option." The difference is that the SRM allows these probabilities to vary across options, while the state response model assumes they are equal and evenly distributed among incorrect options. Therefore, the state response model can be considered a constrained version of the SRM. Second, when all problem states in a task have exactly $K = 2$ options, the three models become completely equivalent. Additionally, it is worth noting that, similar to the SRM, the state response model also imposes constraints on some model parameters, so the number of parameters to estimate is not always greater than that of the ASM.

Furthermore, some process data analysis studies have used forms similar to the one- or two-parameter IRT models used in the 1P-/2P-ASM. For example, Han and Wilson (2022) applied mixed Rasch models or mixed partial credit models to process data analysis, enabling not only the estimation of students' latent abilities but also exploratory classification of their problem-solving processes. The Markov IRT model proposed by Shu et al. (2017) also has a form similar to the two-parameter IRT model (or partial credit model) used in the 2P-ASM. However, the main difference between these two models and the ASM (as well as the SRM and state response model) is that the former analyze data in the form of numerical matrices converted from action sequences with standardized data formats, while the ASM analyzes non-standardized format data that retains temporal information and has varying lengths across individuals. For instance, to ensure all respondents have data of the same length, the former often converts repeated but temporally ordered specific operation sequences into frequency information and uses polytomous scoring models for data analysis, but this conversion loses important temporal sequence information.

Appendix 2. Introduction to the PISA 2012 Tickets Task

Figure A1 shows screenshots of the PISA 2012 Tickets task, which includes three sub-questions. For CP038Q02: purchasing a full-fare suburban train ticket valid for two trips, respondents who meet the task requirements receive 1 point, while those who do not answer or fail to meet the requirements receive 0 points. Figure A2 shows the correct action path for completing this task.

Appendix 3. Abnormal Action Sequences in PISA 2012 Tickets Task Data

Table A2 shows an example of an abnormal action sequence deleted from the empirical study data. In the table, Cnt represents country ID, SchoolID represents

school ID, and StdID represents student ID. The abnormal action sequence is highlighted in bold. The action sequence that follows the task' s state transition rules should be: Country_{trains} → Full_{Fare} → Daily → Cancel → Country_{trains} → Full_{Fare} → Individual Trip_2 → Buy. During the recording of this respondent' s operations, the system made an error, causing the respondent' s action sequence to be recorded in reverse order and repeated once. Due to the large volume of data in the empirical study, it was impractical to correct all action sequences in the dataset individually, so all sequences that did not conform to the task' s preset rules were deleted.

Appendix 4. Posterior Predictive p-value (ppp) Calculation Logic

In this study, model fit was evaluated using posterior predictive p-values (ppp). We selected the sum of observed values ($O(\cdot)$) as the test statistic. Table A3 presents the calculation rules for the SRM, 1P-ASM, and 2P-ASM. Notably, observed state transitions are categorical data. We need to compare the true values and replicated values for each of the 4,000 MCMC samples. Therefore, the true and replicated values of state transitions are recoded as 0 or 1, where 1 represents a correct state transition and 0 represents an incorrect state transition. The ppp value is the proportion of replications where the O statistic for the replicated data exceeds that for the observed data. If the model fits the data, the ppp value will be close to 0.5.

Appendix 5. Parameter Estimation Trace Plots and Posterior Distribution Plots from Empirical Study

[Figures A3-A8 showing trace plots and posterior distributions for various parameters]

Appendix 6. Supplementary Content for Simulation Study

The specific steps for generating all respondents' action sequences are as follows: (1) Define the optimal action sequence and all correct/incorrect state transitions for the task based on Figure 6; (2) Generate all model parameters sequentially, where (a) respondents' "true" problem-solving ability parameters are randomly generated from a standard normal distribution $\theta \sim N(0,1)$, (b) discrimination parameters I for correct and incorrect state transitions are set to 1 and -1, respectively, and (c) "true" values for state transition propensity parameters λ_{jk} are set by referencing both the empirical estimates (see Online Appendix Table A8) and the simulation settings from Han et al. (2022). Table A4 presents the "true" values for all state transition propensity parameters under short and long sequence conditions; (3) Substitute all "true" parameter values into the SRM to calculate the probability matrix for all respondents presenting all state transitions, with rows representing respondents and columns representing state transitions; (4) Set all respondents to start from the initial state S. According to the task structure in Figure 6, at state S, generate the stage 1 to stage 2 state

transition (i.e., whether stage 2 selects A or F) randomly from a categorical distribution based on the respondent' s probabilities of presenting SA and SF. If A is selected, then at state A, generate the stage 2 to stage 3 state transition (i.e., whether stage 3 selects B, G, or S) randomly from a categorical distribution based on the respondent' s probabilities of presenting AB, AG, and AS. Continue this process until reaching the goal state J, completing the generation of this respondent' s action sequence. Repeat this process to generate action sequences for all respondents.

Appendix 7. Supplementary Notes on Parameter Estimation and Robustness Analysis Results

This study used the Rstan package in R software for MCMC parameter estimation. Rstan uses the No-U-Turn Sampler (NUTS) (Hoffman & Gelman, 2014) as the default sampling method. Tables A5-A7 and Figure A9 present robustness analysis results for parameter estimation under uninformative and informative prior distributions. The results show that parameter estimates for both models are highly robust regardless of the amount of information in the prior distributions. All parameter estimates in the main text used informative prior distributions. Parameter estimation code and example data are shared at https://osf.io/3y2xr/?view_{only}=7bc05393a51f472aa2462214ba588063

Appendix 8. State Transition Propensity Parameter Estimates for SRM from Empirical Study

[Table A8 showing SRM parameter estimates]

Note: Figure translations are in progress. See original paper for figures.