PSYCHOMETRIKA—VOL. 89, NO. 3, 877–902
SEPTEMBER 2024
https://doi.org/10.1007/s11336-024-09963-8

POLYTOMOUS EFFECTIVENESS INDICATORS IN COMPLEX PROBLEM-SOLVING TASKS AND THEIR APPLICATIONS IN DEVELOPING MEASUREMENT MODEL

PUJUE WANG AND HONGYUN LIU

BEIJING NORMAL UNIVERSITY

BEIJING KEY LABORATORY OF APPLIED EXPERIMENTAL PSYCHOLOGY, NATIONAL DEMONSTRATION CENTER FOR EXPERIMENTAL PSYCHOLOGY EDUCATION (BEIJING NORMAL UNIVERSITY)

Recent years have witnessed the emergence of measurement models for analyzing action sequences in computer-based problem-solving interactive tasks. The cutting-edge psychometrics process models require pre-specification of the effectiveness of state transitions often simplifying them into dichotomous indicators. However, the dichotomous effectiveness becomes impractical when dealing with complex tasks that involve multiple optimal paths and numerous state transitions. Building on the concept of problem-solving, we introduce polytomous indicators to assess the effectiveness of problem states d_s and state-to-state transitions $\Delta d_{s\to s'}$. The three-step evaluation method for these two types of indicators is proposed and illustrated across two real problem-solving tasks. We further present a novel psychometrics process model, the sequential response model with polytomous effectiveness indicators (SRM-PEI), which is tailored to encompass a broader range of problem-solving tasks. Monte Carlo simulations indicated that SRM-PEI performed well in the estimation of latent ability and transition tendency parameters across different conditions. Empirical studies conducted on two real tasks supported the better fit of SRM-PEI over previous models such as SRM and SRMM, providing rational and interpretable estimates of latent abilities and transition tendencies through effectiveness indicators. The paper concludes by outlining potential avenues for the further application and enhancement of polytomous effectiveness indicators and SRM-PEI.

Key words: polytomous effectiveness indicators, SRM-PEI, process data, problem-solving.

Problem-solving ability is often considered one of the most difficult aspects of human cognition (Newell & Simon, 1972) and a crucial skill for the 21st century (Griffin & Care, 2014; OECD, 2018) . Computer-based interactive assessments are increasingly favored in large-scale international survey programs. For example, the Organization for Economic Cooperation and Development's (OECD) Programme for International Student Assessment (PISA) introduced tests for computer-based problem-solving and human-computer interactive collaborative problem-solving in 2012 and 2015, respectively (OECD, 2014, 2016) . The Assessment and Teaching of 21st Century Skills (ATC21S) initiative pioneeres the interpersonal interaction testing task, which can also assess cooperative problem-solving skills (Griffin & Care, 2014) . Computer-based interactive tests, grounded in realistic problem-solving scenarios, require respondents to engage with the scenarios and make multistep decisions towards solutions. Every action taken in addressing a problem is recorded as process data by the computer platform. These action sequences provide valuable insights into the cognitive and response mechanisms of respondents, extending beyond mere outcomes and grades (Bergner & von Davier, 2019) . They can be analyzed to extract sequence-based features for interpreting the problem-solving process (e.g., (He & von Davier,

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11336-024-09963-8.

Correspondence should be made to Hongyun Liu, Beijing Key Laboratory of Applied Experimental Psychology, National Demonstration Center for Experimental Psychology Education (Beijing Normal University), Faculty of Psychology, Beijing Normal University, No. 19, Xin Jie Kou Wai St., Hai Dian District, Beijing100875, People's Republic of China. Email: hyliu@bnu.edu.cn

2015, 2016; Tang et al., 2020)) and are key in developing measurement models for estimating latent problem-solving abilities (e.g., (Chen, 2020; Han et al., 2022; LaMar, 2018; Shu et al., 2017; Xiao & Liu, 2023)). Measurement models estimating problem-solving abilities encompass both traditional psychometric models (Liu et al., 2018; Yuan et al., 2019; Han & Wilson, 2022) and stochastic process modeling (Arieli-Attali et al., 2019; Xiao et al., 2021) . Merging the strengths of these two approaches, psychometric models incorporating stochastic process properties have also emerged (Shu et al., 2017; LaMar, 2018; Chen, 2020; Han et al., 2022; Xiao & Liu, 2023; Fu et al., 2023; Tang, 2023) . These models, considering the sequential dependency of actions, view action sequences as stochastic processes with first-order Markov properties and model the conditional probabilities of respondents' choices under each problem state (Shu et al., 2017) .

These process models are specifically designed for well-defined tasks, often utilizing the Finite State Automata (FSA) framework, a prevalent structure for interactive problem-solving tasks. In FSA tasks, the system is characterized by a finite number of states, a defined set of allowable actions, and a transition function that dictates the next state based on the action taken in the current state (Buchner & Funke, 1993) . Respondents are tasked with moving from an initial state to a target state to resolve the problem (Anderson et al., 2007) . Effective performance is achieved by determining and following the optimal path from the initial to the target state. Additionally, the occurrence of unnecessary steps is indicative of inefficiency during the knowledge application stage of the problem-solving process (Buchner & Funke, 1993; Funke, 2001) . To assess problem-solving ability, it is crucial to assess the effectiveness of each action that leads to a transition, considering the nature of FSA tasks. Then the effectiveness is integrated into the model as pre-defined parameters, subsequently facilitating the estimation of latent problem-solving abilities.

The concept and assessment of effectiveness originate in reinforcement learning, but due to the complexity of algorithmic evaluation, they have since evolved into manually assessed dichotomous indicators. LaMar (2018) used the action-value function from the reinforcement learning paradigm to evaluate the effectiveness of actions and established a measurement model using Markov Decision Processes. The action-value function calculates the expected weighted sum of future rewards for each action in a given problem state and is solved using dynamic programming algorithms. However, Lamar's approach to assessing action value is intricate, thus limiting its practicality in psychometrics. Chen (2020) determined action effectiveness based on whether an action, leading to a state-to-state transition, aligns with the optimal path. This assessment, essentially evaluating the effectiveness of transition, utilizes a dichotomous indicator with values of 0 and 1. For the PISA 2012 Ticket task (OECD, 2014) , which features a single optimal path from the initial to the target state, Chen manually evaluated the effectiveness of each transition. Transitions that are either on the optimal path or lead back to the optimal path from an incorrect one are classified as correct transitions, with a value of 1 as effectiveness. Conversely, transitions that are not on the optimal path, indicating they are on an incorrect path, are classified as incorrect transitions, marked with a value of 0 as effectiveness. Essentially, effectiveness is reduced to dichotomous correctness. Utilizing this dichotomous effectiveness, Chen applied the Nominal Response Model (Bock, 1972) in conjunction with a task difficulty parameter to fit action sequences, specifically employing the Continuous-Time Dynamic Choice Model (CTDC).

The new model modifies the values of dichotomous effectiveness and the difficulty parameter under the NRM framework. Han et al. (2022) proposed the Sequential Response Model (SRM) and employed values of 1 and -1 for correct and incorrect transitions, respectively. In SRM, effectiveness is only multiplied by the latent ability parameter, and a set of transition tendency parameters replaces the single task difficulty parameter. This expansion allows for a more nuanced estimation of transitions. Furthering the SRM, Fu et al. (2023) incorporated a log-normal action time model into the SRM to simultaneously accommodate action times. Xiao and Liu (2023) retained

the dichotomous effectiveness indicators of 0 and 1 while altering the task difficulty parameter in CTDC to transition tendency parameters in SRM, forming the State Response Measurement Model (SRMM). Comparative analysis in both simulation studies and empirical research on the Ticket task shows that SRMM outperforms CTDC.

However, the use of dichotomous effectiveness indicators often results in an oversimplification of diverse situations and limits the application of measurement models to complex tasks, primarily for three reasons: First, in scenarios with a single path to the target and multiple backward transitions, dichotomous effectiveness fails to distinguish between different types of ineffective transitions. For example, in a state transition diagram (Fig. 1a) with only one path from starting state A to target state D via intermediates B and C, the transition from C back to A (C→A) is worse than from C back to B (C→B) as it moves further from the target. However, dichotomous effectiveness does not capture the severity difference between these backward transitions. Second, when multiple paths lead to a target or there are multiple targets, calibrating dichotomous effectiveness becomes challenging since a transition can move closer to one target while simultaneously moving away from the other. For instance, in a transition diagram with two target states C and E (Fig. 1b), the optimal path is A→B→C. According to the standard of dichotomous effectiveness, the transition D→B is considered correct because it leads back to the shortest path from a non-optimal path. However, the transition D→E is more effective as it allows the task to be completed in fewer steps, illustrating a limitation in the dichotomous approach where it fails to account for the effectiveness of completing the task. Third, with a multitude of states, transitions, and optimal paths, the complexities mentioned above may coexist, complicating the effectiveness assessment of various transitions. Furthermore, the manual evaluation process becomes exceedingly labor-intensive and time-consuming, making it impractical for complex tasks. Consequently, measurement models dependent on dichotomous effectiveness indicators face significant challenges in addressing complex scenarios, highlighting the necessity for more sophisticated evaluation approaches. This situation calls for an urgent shift towards polytomous effectiveness indicators and automated evaluation methods. Such advancements are crucial for accurately differentiating and adapting to a wide range of transitions and for broadening the scope of measurement models, making them suitable for more complex FSA tasks.

In this study, we introduce a novel method for assessing the effectiveness of various state transitions, as well as a new measurement model that incorporates polytomous effectiveness indicators and is tailored for complex tasks with multiple optimal paths. Specifically, in Sect. 1, we propose a universal method for gauging the effectiveness of states and state-to-state transitions capable of complex FSA tasks. We then exemplify the derivation of polytomous effectiveness evaluation outcomes through the two real FSA tasks. In Sect. 2, we introduce a new measurement model, termed the Sequential Response Model with Polytomous Effectiveness Indicators (SRM-PEI), detailing its specification and parameter estimation methodology. In Sect. 3, we execute a simulated study to probe the accuracy of SRM-PEI estimations under various conditions within simulated tasks. In Sect. 5, we demonstrate the applicability of SRM-PEI and compare it with SRM and SRMM in two problem-solving tasks. The article concludes with a discussion in Sect. 6.

1. Effectiveness Indicators of the States and Transitions in FSA Tasks

1.1. New Definitions of Effectiveness of States and State-to-State Transitions

Theoretically, problem-solving is a process of navigating towards the target through a series of state-to-state transitions (Newell & Simon, 1972; Mayer & Wittrock, 2006) . The criterion for dichotomous effectiveness is determined by whether the state after the transition is closer to the goal than the state before the transition. When there is a single target and an optimal

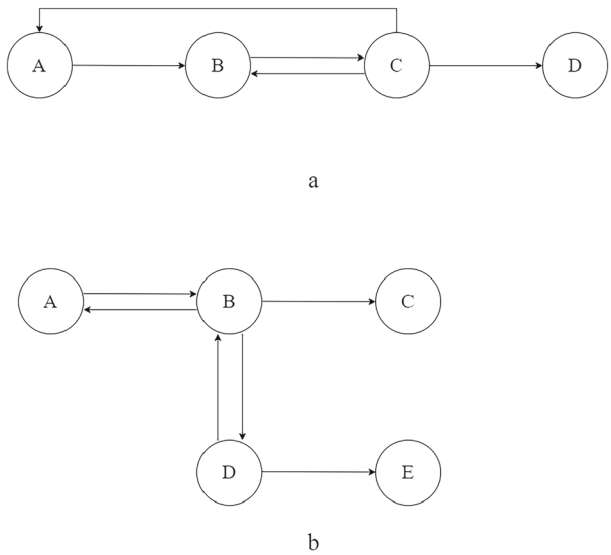

FIGURE 1.
Diagram of two scenarios with multiple backward transitions and target states.

path, the criterion reduces to whether it is consistent with the optimal path. This criterion has two limitations: first, it does not account for scenarios with multiple targets and multiple shortest paths; second, it merely assesses whether the distance to the target state is reduced without considering the extent of the change in distance. Our proposed concept of transition effectiveness quantifies the change in distance to the target before and after each transition in situations with multiple targets and paths. A transition is considered effective if it reduces the distance and inefficient if it increases the distance, aligning with the principles of evaluating problem-solving abilities in FSA tasks (Buchner & Funke, 1993; Funke, 2001) . To address these complexities, we propose two types of effectiveness indicators suitable for complex FSA tasks: (1) the effectiveness indicators d_s of the state s, (2) the effectiveness indicators $\Delta d_{s\to s'}$ of the transition $s \to s'$.

First, we define the distance between any problem state s and the target state s_{target}, denoted as d_s. The d_s is the minimal steps of transitions needed to reach the target state s_{target} from the state s. A smaller d_s means that state s is closer to target state, so it is more effective to solve the problem if the state with smaller d_s is reached In complex FSA tasks, there can be k target states that can be reached from state s, denoted as $s^{(1)}_{target}, s^{(2)}_{target}$ to $s^{(k)}_{target}$. The distances to these target states are correspondingly $d_s^{(1)}, d_s^{(2)}$ to $d_s^{(k)}$. Given the understanding of state effectiveness as the theoretical minimum distance from a state to any target state, the effectiveness of the state s is calculated as $d_s = \min (d_s^{(1)} d_s^{(2)}, \ldots, d_s^{(k)})$

In Fig. 1a, the effectiveness of states A, B, C, and D is determined by their respective shortest distances to the target state D, which are 3, 2, 1, and 0, respectively. In Fig. 1b, where the target states are C and E, both have an effectiveness of 0. The shortest distances from B to targets C and E are 1 and 2, respectively. Therefore, the effectiveness of state B is $d_B = \min \left(d_B^C, d_B^E\right) = \min (1, 2) = 1$. Similarly, the effectiveness of state D is $d_D = \min \left(d_D^C, d_D^E\right) = \min (2, 1) = 1$, and the effectiveness of state A is $d_A = \min \left(d_A^C, d_A^E\right) = \min (2, 3) = 2$

Second, to get the effectiveness of a transition $s \to s'$, we calculate two distances from state s to s_{target} (i.e., d_s) and s' to s_{target} (i.e., $d_{s'}$), respectively Then, we compute the difference between d_s and $d_{s'}$ denoted as $\Delta d_{s\to s'} = d_s - d_{s'}$ The difference, $\Delta d_{s\to s'}$ is defined as the effectiveness

indicator of the transition $s \rightarrow s'$ whose value is equal to the change in the shortest distance from the target after the transition $s \rightarrow s'$. A value greater than zero for $\Delta d_{s \rightarrow s'}$ indicates that it is closer to the target after the transition $s \rightarrow s'$, while a value less than zero implies that it is further from the target. The absolute value of $\Delta d_{s \rightarrow s'}$ indicates the number of steps closer or farther away from the target It is noteworthy that two types of effectiveness indicators are located on different ends of the indicator scales. A state with a higher value of d_s is less effective, while a transition with a higher value of $\Delta d_{s \rightarrow s'}$ is more effective.

In Fig. 1a, there are three transitions with an effectiveness of 1 that move closer to the target ($\Delta d_{A \rightarrow B} = \Delta d_{B \rightarrow C} = \Delta d_{C \rightarrow D} = 1$). Additionally, there are two transitions that move away from the target with unequal effectiveness ($\Delta d_{C \rightarrow B} = d_C - d_B = -1$, $\Delta d_{C \rightarrow A} = d_C - d_A = -2$). The transition C→A, which moves further back, has lower effectiveness. Polytomous effectiveness differentiates between various types of backward movement. In Fig. 1b, the effectiveness of three transitions close to the target is also equals 1 ($\Delta d_{A \rightarrow B} = \Delta d_{B \rightarrow C} = \Delta d_{D \rightarrow E} = 1$). Notably, $\Delta d_{B \rightarrow D} = \Delta d_{D \rightarrow B} = 0$, indicating that the distance to the target remains unchanged after these transitions. This suggests that D→B, compared to the more effective D→E, is a less efficient transition and not always the optimal choice for returning to the shortest path. Since $\Delta d_{B \rightarrow A} = -1$, the transition B→A is a worse option than B→D. The polytomous effectiveness provides a more nuanced evaluation than the dichotomous effectiveness based solely on the shortest path, better aligning with the task design.

This section presents a general framework that can automatically evaluate the polytomous effectiveness indicators of states and transitions in FSA tasks. The process of evaluating the effectiveness can be summarized in three steps: (1) Define the state space. This involves finding all the target states. During this step, states may be categorized and simplified. Simultaneously, all state transitions are defined. (2) Calculate the effectiveness d_s for all states. The shortest path can be identified using transition diagrams or search algorithms. (3) Calculate the effectiveness $\Delta d_{s \rightarrow s'}$ for all transitions. This process requires that each state within the task can reach a target state through a series of transitions. If this requirement is not met for some states or transitions, additional values may need to be assigned to their effectiveness. In the two subsequent sections, we will demonstrate the process of evaluating effectiveness for two real tasks, one with single optimal path and the other with multiple optimal paths.

1.2. Example of Evaluating the Effectiveness Indicators for a FSA Task with a Single Optimal Path

In this section, we demonstrate the calculation of effectiveness metrics using the Ticket task from PISA 2012, which is the most commonly used problem-solving task in existing models (Chen, 2020; Han et al., 2022; Xiao & Liu, 2023; Fu et al., 2023) . Taking sub-task CP038Q02 as an example, this task requires students to purchase a full-fare ticket for country trains, valid for two trips. Students have to sequentially select the correct option on the simulated ticketing interface ("COUNTRY TRAINS"→"FULL FARE"→"INDIVIDUAL"→"2 Trips"→"BUY") Before selecting "BUY", the student has the option to hit "CANCEL" to restart the task from the beginning. This task was scored in a binary fashion, depending on whether the student successfully purchased the correct ticket.

When evaluating dichotomous effectiveness in the Ticket task, previous studies have already completed much of the fundamental work for evaluating the polytomous effectiveness we propose. For the first step, Chen (2020) defined all states and transitions. Building on this, Han et al. (2022) merged states with similar error types, effectively reducing the number and complexity of states and transitions. More importantly, they illustrated all paths leading to the target state through a transition diagram, which is crucial for identifying the shortest path from each state to the target in the second step. However, some adjustments to the state categorization and the transition diagram

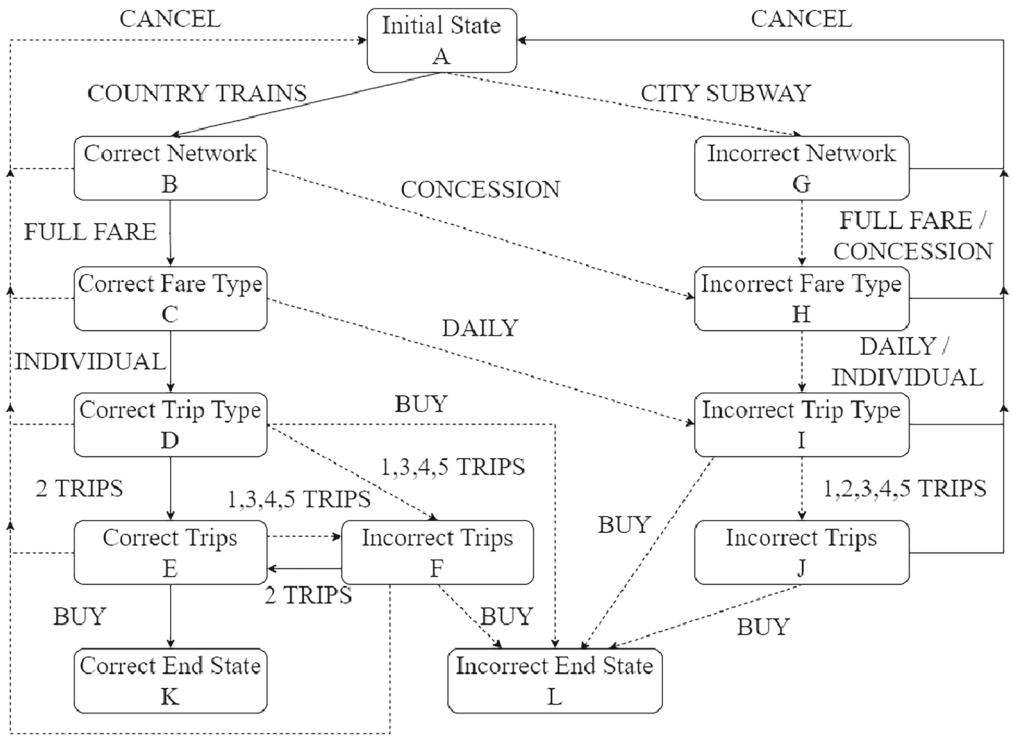

FIGURE 2.
A new transition diagram for the CP038Q02 subtask of the Ticket task in the PISA 2012. *Note*: The solid arrows represent transitions that move closer to the target state, while the dotted arrows represent transitions that do not move closer to the target state.

are still required for assessing the polytomous effectiveness. It is important to note that the Ticket task has only one target state, but also includes a non-target end state. Once the incorrect end state is reached, the problem-solving process terminates prematurely, making it impossible to reach the target state. In this case, we cannot get the shortest distance between the incorrect end state and the target state. From the task design, the incorrect end state is further from the target state than all other states, so we can set its effectiveness to be lower than any other state. Eventually, we distinguish between the correct target state and the incorrect end state to depict a new transition diagram (see Fig. 2) and then calculate the effectiveness of all transitions (see Table 1).

To proceed, we complete the final two steps by sequentially calculating the two types of effectiveness indicators. Step (2): The state effectiveness d_s distinctly and meaningfully differentiates between states A,B,C,D,E, and K on the optimal path. States G, H, I, and J on an incorrect path indicate a shortest distance of 6 transitions from the target. The other incorrect state F, as a branch on the optimal path, is only 2 steps away from the target. Since the maximum effectiveness value among states A to J is 6, we set the effectiveness of incorrect end state L as 7. Step (3): Different from the dichotomous effectiveness which can only indicate correct or incorrect, our effectiveness $\Delta d_{s\to s'}$ can be polytomously scored in the task with a single optimal path, and clearly show the change in the shortest distance to the target state after each transition. Comparing with the evaluation of Han et al. (2022), the effectiveness of correct transitions remains at a value of 1, while incorrect transitions have effectiveness values ranging from 0 to -5, with lower values indicating that the target is further away after the transitions These two types of effectiveness can be aggregated into descriptive indicators, not only providing a deeper description and evaluation

TABLE 1.
The effectiveness of all states and transitions in the CP038Q02 task.

S		S'											
Labels	States	A (5) Initial state	B (4) Correct network	C (3) Correct fare type	D (2) Correct trip type	E (1) Correct trips	F (2) Incorrect trips 1	G (6) Incorrect network	H (6) Incorrect fare type	I (6) Incorrect trip type	J (6) Incorrect trips 2	K (0) Correct end	L(7) Incorrect end
A (5)	Initial State		**COUNTRY TRAINS (1)**					CITY SUB-WAY (−1)					
B (4)	Correct Net-work	CANCEL (−1)		**FULL FARE (1)**					CONCESSION (−2)				
C (3)	Correct Fare Type	CANCEL (−2)			**INDIVIDUAL (1)**					DAILY (−3)			
D (2)	Correct Trip Type	CANCEL (−3)				**2 TRIPS (1)**	1, 3, 4, 5 TRIPS (0)						BUY (−5)
E (1)	Correct Trips	CANCEL (−4)					1, 3, 4, 5 TRIPS (−1)					**BUY (1)**	
F (2)	Incorrect Trips 1	CANCEL (−3)				**2 TRIPS (1)**							BUY (−5)

TABLE 1.
continued

S		S'											
Labels	States	A (5) Initial state	B (4) Correct network	C (3) Correct fare type	D (2) Correct trip type	E (1) Correct trips	F (2) Incorrect trips 1	G (6) Incorrect network	H (6) Incorrect fare type	I (6) Incorrect trip type	J (6) Incorrect trips 2	K (0) Correct end	L(7) Incorrect end
G (6)	Incorrect Net-work	**CANCEL (1)**							FULL FARE / CON-CES-SION (0)				
H (6)	Incorrect Fare Type	**CANCEL (1)**								INDIVIDUAL / DAILY (0)			
I (6)	Incorrect Trip Type	**CANCEL (1)**									1, 2, 3, 4, 5 TRIPS (0)		BUY (−1)
J (6)	Incorrect Trips 2	**CANCEL (1)**											BUY (−1)

The effectiveness of the states and transitions is in parentheses. The bold transitions are considered correct, while the non-bolded transitions are considered incorrect, according to Han et al. (2022).

of the problem-solving process but also serving as validation for estimating latent abilities in measurement models. The application of aggregated indicators will be discussed in detail in Sect. 5.

1.3. Example of Evaluating the Effectiveness Indicators for a FSA Task with Multiple Optimal Paths

In this section, we demonstrate the step-by-step evaluation of effectiveness indicators using a complex FSA task—the Balance Beam task—a collaborative problem-solving task from the Assessment and Teaching of 21st Century Skills (ATC21S) project (Griffin & Care, 2014) . In the Chinese version of the Balance Beam task developed by Yuan et al. (2019), two students are required to balance a beam that has four notches on each side for placing four weights (50 g, 100 g, 300 g, and 500 g). Only student A possesses all four weights at the beginning, while student B has none (see Fig. 5 in the Appendix). The testing system permits weight transfers between the students. The Balance Beam task exemplifies a complex FSA task with a multitude of intermediate states, intricate transition connections, and notably, multiple target states. There are multiple optimal paths to a target, because the order of hanging the same set of weights does not affect the final balance. The existence of multiple targets further expands the number of optimal paths. These paths are interconnected, meaning that a change in the target state during problem-solving can render a previously optimal transition suboptimal from a broader perspective. This complexity in the task structure necessitates a nuanced approach to evaluating the effectiveness of transitions and states within the problem-solving process. The procedure for automatically evaluating the effectiveness of all states and transitions in the Balance Beam task is as follows:

Step (1): Define all problem states and state transitions. Given the collaborative nature of the task, we view both sides of the balance beam as a whole, defining the state space based on the positions of the four weights. Each weight can occupy one of ten possible positions: eight on the beam and two off-beam, i.e., held by a student (see Appendix A for details). Given four weights, there are potentially 10^4 =10,000 distinct states. Consistent with the principle when defining the state space, the target state is also defined at the group level, which means that the states in which the beam is balanced are the target states shared by both individuals. Whether utilizing two, three, or four weights, there exists a wide array of combinations for hanging weights to attain balance. The Depth First Search (Cormen et al., 2022) is adept at swiftly locating all the target states in the Balance Beam task within one second. When using two, three, and four weights, there are 24, 68, and 40 target states, respectively.

A transition between states occurs when the position of any of the four weights changes. The testing platform accommodates four kinds of actions capable of inducing a position change and a state-to-state transition, which include hanging weights, removing weights, transferring weights, and shifting notches on the same side Fig. 6 illustrates four types of transitions that can occur when a single weight is moved among ten possible positions. Note that Fig. 6 is not the state transition diagram showing all possible paths in this task. Under these conditions, the task permits 168,000 possible state-to-state transitions. For any intermediate state, all target states are accessible Two students have the flexibility to change targets at any time. Thus, it is impractical to represent all states and transitions through a diagram or a table, let alone find all the optimal paths from a given state to target states.

Step (2): Evaluate the effectiveness of all states Within the context of the Balance Beam task, the effectiveness indicators d_s of the state s can be interpreted as the theoretical shortest distance between the current state and any group-level target state The state effectiveness defined under the grouplevel target does not distinguish to which student the remaining transitions belong. Obviously, the effectiveness of all target states is set as $d_{target} = 0$. Subsequently, the effectiveness of all non-target states is evaluated. Since Any intermediate state can reach all target states

before exiting the mission, we developed a rule-based algorithm which incorporates the positional encoding and edit distance. This approach can be programmed to compute the shortest distance between any intermediate and target states swiftly and precisely to avoid the labor-intensive and potentially error-prone process of manual computations. Further details of the algorithm can be found in Appendix A.

Step (3): Calculate the effectiveness of all transitions. With the effectiveness of all states determined, the effectiveness of all the transition can be obtained simply by $\Delta d_{s\to s'} = d_s - d_{s'}$. This transition effectiveness defined at the group level represents the impact of an individual's action on accomplishing the common goal shared by the two people.

In the Balance Beam task, the number of original states and transitions is too large, and the role of the original states (such as all four weights being with student A) and actions (for instance, student A passing the 50 g weight to student B) is ambiguous in terms of problem-solving. The two types of effectiveness d_s and $\Delta d_{s\to s'}$ facilitate a rapid and meaningful categorization of these numerous states and transitions. The polytomous effectiveness indicators not only reduce the number of categories for states and transitions but also enhance the interpretability of further analysis for this task. In the balance beam task, all states can be divided into 6 types based on the value of d_s from 0 to 5. Each type of state can transition from itself (e.g., 5→5), or from states that are one transition away (e.g., 5→4), resulting in 16 types of transitions according to the effectiveness of the states before and after the transition $s \to s'$. Furthermore, based on the values of transition effectiveness $\Delta d_{s\to s'}$, all transitions can be classified into 3 categories: advancing towards the target ($\Delta d_{s\to s'} = 1$), staying in the same place ($\Delta d_{s\to s'} = 0$), and moving away from the target ($\Delta d_{s\to s'} = -1$).

2. Sequential Response Model with Polytomous Effectiveness Indicators (SRM-PEI)

The new proposed indicators in Section 1 not only allow the description of response characteristics, but also enable the development of measurement models for complex problem-solving tasks. Specifically, we maintain the framework of the measurement model combined with the random process, substituting the dichotomous effectiveness indicator for the transition with the polytomous effectiveness indicator. As an example, we take the Sequential Response Model (SRM; Han et al. (2022)), a model for state transitions with a dichotomous effectiveness indicator, to illustrate how to extend a model designed for a single optimal path task to one applicable for a complex task using the new effectiveness indicator. We call it the Sequential Response Model with Polytomous Effectiveness Indicators (SRM-PEI).

2.1. Model Specification

Drawing inspiration from SRM (Han et al., 2022) , we focus on the state transitions prompted by a respondent's actions, viewing these as external manifestations of latent ability. Each state is treated as an item, while each transition originating from this state is considered a choice pertaining to that item. This structure helps us conceptualize and analyze the transitions in process data within an IRT model. The effectiveness indicators of transitions ($\Delta d_{s\to s'}$) in SRM-PEI provide a more nuanced assessment of how good or bad each transition is.

Assuming that the next state s' only depends on the current state s and the respondent's latent ability during the problem-solving process We can treat the response sequence as a discrete-time stochastic process with a conditional Markov property. Given that a state can have multiple transitions in a complex problem-solving task, we employ the effectiveness indicators to ascertain the relative superiority among all transitions. The SRM-PEI can thus be built within the framework of the NRM The SRM-PEI specifies the conditional probability of respondent i choosing to reach

state $s\prime$ when in problem state s as follows:

$$P(S_{i,l+1}{=}s' \mid S_{i,l} = s, \theta_i, \boldsymbol{\lambda}, \mathbf{D}) = \frac{\exp(\Delta d_{s\to s'} \cdot \theta_i + \lambda_{s\to s'})}{\sum_{x\in M_s} \exp(\Delta d_{s\to x} \cdot \theta_i + \lambda_{s\to x})} \tag{1}$$

where θ_i represents the latent ability of respondent i, while $\lambda_{s\to s'}$ is the tendency parameter for the transition from state s to s' and reflects the easiness of the transition. A larger value of $\lambda_{s\to s'}$ indicates a higher likelihood of making that transition. λ is a vector of tendency parameters for all transitions within the task. M_s represents the set of reachable states in the next step from the current state s; and $\Delta d_{s\to s'}$ is the effectiveness indicator calculated in the previous section, forming a vector $\mathbf{D}$ of effectiveness indicators for all transitions.

2.2. *Model Estimation*

For flexibility and convenience in implementation, Bayesian estimation is adopted to estimate the parameters of latent ability and transition tendency. Let $\boldsymbol{\theta}$ and $\mathbf{S}$ denote the collection of latent variables for n respondents and their response sequences. The posterior probability of $\boldsymbol{\theta}$ and $\mathbf{S}$ can be expressed as follows:

$$p(\boldsymbol{\theta}, \boldsymbol{\lambda}|\mathbf{S}, \mathbf{D}) \propto p(\mathbf{S}|\boldsymbol{\theta}, \boldsymbol{\lambda}, \mathbf{D}) p(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \prod_{i=1}^{n} \prod_{l=1}^{L_i-1} p(S_{i,l+1}{=}s' \mid S_{i,l}{=}s, \theta_i, \boldsymbol{\lambda}, \mathbf{D}) p(\theta_i) p(\boldsymbol{\lambda}) \tag{2}$$

where $p(\theta_i)$ and $p(\boldsymbol{\lambda})$ are the prior distributions of the latent ability and vector of the transition tendency parameter, respectively, and are assumed to be independent of each other. To ensure model identification, the sum of all tendency parameters of transitions from the same state is constrained to be zero. To simplify Bayesian estimation, we assume that the prior distributions for the latent abilities are assumed to be standard normal distributions. The Markov chain Monte Carlo (MCMC) estimation is implemented using the Metropolis-Hastings-within-Gibbs sampling approach to empirically approximate the joint posterior distribution (Patz & Junker, 1999a,b) . The detailed sampling procedures can be found in the Appendix of Han et al. (2022).

It is crucial to assess the convergence and model fit in Bayesian estimation. We used two methods to monitor MCMC convergence: (1) the potential scale reduction factor (PSRF; Gelman & Rubin, 1992), where PSRF values approximating 1 suggest convergence; (2) Monte Carlo error (MCE; Koehler et al. (2009)) which measures the standard deviation of the sample means across chains. A smaller MCE indicates less variability between different chains, hence a higher likelihood of convergence. Following the assurance of convergence, we employ Posterior predictive checking (PPC) using the test statistics approach to evaluate the model-data fit (Gelman et al., 1996, 2014; Guttman, 1967; Rubin, 1984) . Specifically, we visually compare the observed frequencies of transitions to those obtained from the posterior predictive data. Additionally, we compute the posterior predictive p-value (*ppp*) based on the chi-square test of the two distributions. A *ppp* value close to 0.5 signifies a good model fit (Gelman et al., 2014) .

3. Simulation Study

In this section, we have performed a Monte Carlo simulation study to assess the precision of the SRM-PEI in estimating latent abilities and transition tendencies within the context of a simulated problem-solving task encompassing multiple optimal paths. This study is designed to explore the impacts of differing prior distributions, sample sizes, and lengths of response sequences on the parameter recovery performance of the SRM-PEI.

4. Simulation Design

In this simulation study, three factors were examined for their potential impact on the performance of SRM-PEI: sample sizes, sequence lengths, and prior distributions for transition tendency parameters. This resulted in a total of 16 condition: 4 (sample sizes: 200, 500, 1000, and 2000) $\times$ 2 (sequence lengths: short, long) $\times$ 2 (prior distributions: informative, non-informative). Each condition was replicated 100 times. The descriptive of the simulated problem-solving task and the effectiveness of the states and transitions are given in Supplementary Material S1

The true ability parameters were generated from a standard normal distribution for each replication. The transition tendency parameters can influence the lengths of the observed sequences (Han et al., 2022) . Based on the values of polytomous transition effectiveness and the magnitude of transition tendency parameters in the original SRM (Han et al., 2022) , we designed two sets of tendency parameters for SRM-PEI that could consistently generate differences in sequence length (see Table S2 in the Supplementary Material). To ensure model identification in SRM-PEI, a constraint was placed on $\sum_{x \in M_s} \lambda_{s \to x} = 0$. The priors for transition tendency parameters were only set for transitions with effectiveness $\Delta d_{s \to s'} < 1$, i.e., transitions not lying on the optimal paths. Tendency parameters of transitions with $\Delta d_{s \to s'} = 1$ were set to equal the opposite of the sum of the tendency parameters for transitions with $\Delta d_{s \to s'} < 1$ starting from the same state s The informative prior was a standard multivariate normal distribution $\lambda_{\Delta d_{s \to s'} < 1} \sim \text{MVN}(0, \mathbf{I})$, where $\mathbf{I}$ refers to the identity matrix. The non-informative prior only changed the standard deviation to 10, so $\lambda_{\Delta d_{s \to s'} < 1} \sim \text{MVN}(0, \mathbf{I} \cdot 100)$. The prior for the ability parameters was set as a standard normal distribution.

In MCMC, three chains were implemented with different initial values. The 10,000 samples from each chain were obtained, and the first 5,000 samples were discarded as burn-in. The remaining 15,000 samples in total were taken as the posterior distribution for each parameter. The last 500 samples from each of the three chains were used to conduct posterior predictive checks (PPC)

4.1. Results of the Simulation Study

Under all conditions, the MCMC estimates of SRM-PEI converged normally. The PSRF values for all parameters in both task settings were between 1 and 1.1, providing evidence of convergence (Brooks & Gelman, 1998; Gelman & Rubin, 1992) . Furthermore, the MCE was 0.024 for ability parameters and 0.004 for tendency parameters, which suggests negligible differences between the means of the sampling chains for each parameter and supports the assertion of convergence. In terms of model-data fit, the *ppp* value of 0.598 was close to 0.5, indicating a good fit between the model and the data. The empirical values of each transition in the observed data were consistent with the median of the posterior predictive distributions (see Figure S1 in the Supplementary Material).

The average sequence lengths of the short and long sequence conditions in this task were 17 and 45, respectively. To evaluate the accuracy of parameter estimation, four metrics were calculated under each condition: BIAS, MAE (mean absolute error), RMSE (root mean squared error) and correlation between the estimated and true values of ability and tendency parameters. The estimation accuracy of ability and tendency parameters is shown in Table 2.

In most conditions, the BIAS of latent ability parameters remains relatively small and decreases with larger sample size and longer sequence length, particularly under informative priors (standard normal distribution), where the absolute value of BIAS does not exceed 0.005. Moreover, the correlations between the estimated and true values were relatively high and seldom influenced by the prior, sequence length, and sample size. MAE and RMSE displayed more obvious differences across various factors. Generally speaking, the errors of tendency parameters

TABLE 2.
The estimation accuracy of ability and tendency parameters in SRM-PEI in the simulated problem-solving tasks with multiple optimal paths.

Sample size	Sequence length	Prior	BIAS[a]	MAE		RMSE		Correlation	
			θ	θ	λ	θ	λ	θ	λ
200	Short	Non-inf.	0.051	0.257	0.139	0.321	0.174	0.922	0.988
		Inf.	−0.003	0.256	0.127	0.321	0.161	0.920	0.990
	Long	Non-inf.	0.011	0.262	0.126	0.333	0.162	0.932	0.992
		Inf.	0.005	0.258	0.112	0.329	0.144	0.932	0.995
500	Short	Non-inf.	0.031	0.230	0.086	0.301	0.111	0.900	0.995
		Inf.	−0.001	0.229	0.074	0.301	0.096	0.900	0.997
	Long	Non-inf.	0.020	0.212	0.071	0.283	0.092	0.928	0.998
		Inf.	0.003	0.211	0.071	0.282	0.091	0.928	0.998
1000	Short	Non-inf.	0.011	0.221	0.056	0.289	0.072	0.896	0.998
		Inf.	0.001	0.220	0.055	0.288	0.070	0.897	0.998
	Long	Non-inf.	0.003	0.206	0.051	0.276	0.066	0.925	0.999
		Inf.	0.000	0.205	0.049	0.276	0.065	0.925	0.999
2000	Short	Non-inf.	0.007	0.214	0.040	0.280	0.050	0.893	0.999
		Inf.	0.001	0.213	0.040	0.279	0.051	0.894	0.999
	Long	Non-inf.	0.004	0.199	0.036	0.266	0.046	0.925	0.999
		Inf.	0.004	0.200	0.037	0.267	0.048	0.924	0.999

[a] The constraint leads to zero average bias over all tendency parameters.

were lower than those of latent abilities. The precision of the estimation was found to increase with sample size and sequence length, both of which emerged as key factors influencing MAE and RMSE Specifically, a larger sample size was associated with a more accurate estimation of tendency parameters, as reflected by the reduced MAE and RMSE. Likewise, longer sequences yielded lower MAE and RMSE for both latent abilities and tendency parameters. For complex multi-optimal-path tasks, informative priors were found to reduce estimation errors for both types of parameters, especially under conditions of a small sample size and a short sequence.

The estimation accuracy of each transition parameter was examined (see Table S3 in the Supplementary Material) Most of the tendency parameters exhibited an RMSE lower than average, and the tendency parameters with high accuracy were the important transitions in the problem-solving process. Setting informative priors can reduce the estimation errors of transition tendency parameters with low actual occurrence frequencies, especially in cases with small samples and short sequences. Some transition parameters with low occurrence frequencies might still face estimation errors due to a mismatch between the prior and the data, resulting in parameter estimates pulled towards the overall mean.

5. Empirical Study

In this section, we demonstrate the applicability of SRM-PEI by analyzing empirical data from two tasks, one with a single optimal path (i.e., the Ticket task) and the other with multiple optimal paths (i.e., the Balance Beam task). Specifically, we examine whether SRM-PEI can distinguish response patterns of different abilities and provide rational estimates of transition tendencies. In addition, we conduct model comparisons between SRM-PEI and other process models on both tasks

5.1. Ticket Task

5.1.1. Data Description and Analysis Process for the Ticket Task To demonstrate the application of the SRM-PEI in conjunction with effectiveness indicators, we utilized it to analyze log file data from the sub-task CP038Q02 of the TICKET unit in PISA 2012. After excluding data that did not align with the transition diagram, we analyzed sequences from 31,906 students. The lengths of these sequences ranged from 2 to 110, with an average of 6.983 and a median of 6. For comparative purposes, we also implemented the original SRM (Han et al., 2022) and the SRMM (Xiao & Liu, 2023). We estimated transition tendency parameters and latent abilities for all three models. While SRM and SRMM relied on dichotomous effectiveness and Bayesian estimation as outlined in their respective studies, SRM-PEI used the polytomous transition effectiveness derived in Sect. 1. Given that effectiveness in SRM-PEI is akin to discrimination parameters in the NRM, extreme negative values (e.g., $\Delta d_{D\to L} = -5$) were considered impractical. The pre-experiment show that directly using the effectiveness of the minimum value of -5 harms the model-data fit of SRM-PEI and leads to unreasonable transition characteristic curves. We scaled the effectiveness indicators to a range of -1 to 1 before integrating them into SRM-PEI. For the Bayesian estimation process, we employed three chains with 10,000 sampling iterations each, discarding the initial 5,000 as burn-in. The priors for latent ability and transition tendency parameters were set to standard normal and standard multivariate normal distributions, respectively. The approach for assessing convergence and model-data fit mirrored that used in the simulation study. To compare the models, we utilized various indices such as the Deviance Information Criterion (DIC; Spiegelhalter et al. (1998)), and Pseudo-Bayes Factor (PsBF; Geisser & Eddy, 1979; Gelfand & Dey, 1994). For DIC, lower values suggesting a model that provides a better fit without unnecessary complexity. According to Levy & Mislevy (2016, p. 246), a PsBF value greater than 3 is considered to provide positive, or even stronger, evidence in favor of Model 1 over Model 2.

The latent abilities estimated by the other two models, SRM and SRMM, were utilized to validate the SRM-PEI. Another essential part of our analysis involved determining whether latent abilities from SRM-PEI could account for the overall problem-solving performance in PISA 2012. To this end, we selected ten items (CP018Q04T, CP018Q05, CP025Q01, CP025Q02, CP036Q01, CP036Q02, CP036Q03, CP038Q01, CP038Q02, and CP038Q03) and used Rasch models to estimate overall problem-solving performance. In addition, we conducted an extensive calculation of effectiveness indicators for all states and transitions present in the sequences. This analysis led to the derivation of four aggregated indicators, which were employed to validate the latent abilities estimated by SRM-PEI. Three of these indicators were based on the effectiveness of transitions: the proportion of transitions that approach the target state (i.e., $\Delta d_{s\to s'} = 1$), transitions that maintain the same distance from the target state ($\Delta d_{s\to s'} = 0$), and transitions that move away from the target state ($\Delta d_{s\to s'} < 0$). The fourth indicator was state-based, reflecting the average shortest distance to the target state across all states in a given sequence. These indicators provided a comprehensive view of the students' problem-solving processes, further substantiating the validity of the latent abilities estimated by SRM-PEI.

5.1.2. Results of the Empirical Study on the Ticket Task The Bayesian estimation metrics for the three models, as shown in Table 3, proved the robustness of the MCMC estimates for the three models echoing the findings of previous studies (Han et al., 2022; Xiao & Liu, 2023; Fu et al., 2023) . The Potential Scale Reduction Factor (PSRF) for all parameters was below 1.1, and the Monte Carlo Error (MCE) for ability and tendency parameters was small. These indicators confirm that the MCMC for the three models has successfully converged. Furthermore, the posterior predictive p-value (ppp) for all three models was close to 0.5, suggesting an excellent fit between the models and the observed data. Most notably, two evaluation indices for model comparisons (DIC and PsBF) strongly supported the superiority of the SRM-PEI over the SRM

TABLE 3.
Model comparison of three models in the two empirical studies.

	Max PSRF	MCE		*ppp*	DIC	PsBF
		θ	λ			
Ticket task						
SRM-PEI	1.045	0.037	0.002	0.441	202188.986	
SRM	1.050	0.036	0.002	0.443	203780.763	3529.466
SRMM	1.060	0.051	0.002	0.457	230881.946	34080.304
Balance Beam task						
SRM-PEI	1.050	0.016	0.018	0.458	42791.019	
SRM-v1	1.060	0.022	0.023	0.334	42978.018	151.365
SRM-v2	1.058	0.018	0.024	0.650	43183.545	361.788
SRMM-v1	1.066	0.019	0.050	0.241	47100.057	6,273.488
SRMM-v2	1.061	0.021	0.040	0.243	48485.016	5,657.929

and SRMM in modeling the Ticket task. This indicates that process models embedded with polytomous effectiveness parameters are a better fit than those with dichotomous effectiveness parameters.

The posterior estimates for the transition tendency parameters from SRM-PEI are shown in Table 4 The transition parameters obtained for SRM and SRMM aligned closely with those reported in the original studies by Han et al. (2022) and Xiao and Liu (2023). When grouping by transitions under the same state, the ranking order of the transition tendency parameters estimated by SRM-PEI was consistent with those derived from SRM and SRMM. Figure 3 displays characteristic curves for each group of transitions fitted by SRM-PEI, revealing that the transition tendency parameters assisted by polytomous effectiveness accurately portrayed the probabilities in a manner that reflects the inherent dynamics of the Ticket task. These curves demonstrated that students with higher abilities were more likely to engage in actions that brought them closer to the target state or returned to the initial state from the incorrect path. In contrast, students with lower abilities tended to engage in actions that enter or stay on the incorrect path. This distinction is crucial for understanding the variations in problem-solving abilities among students, as reflected in their choices during the task. It underscores the effectiveness of SRM-PEI in accurately capturing these subtle differences.

Table 5 summarizes the marginal posterior distributions for latent abilities alongside corresponding response sequences, focusing on the five patterns that received the highest and lowest abilities from SRM-PEI. Compared with the states represented by letters, the implementation of state effectiveness encoding markedly enhanced the ability to discern students' proximity to the target. Notably, behavioral patterns closely aligning with the optimal path were associated with the highest problem-solving abilities. In contrast, those students who initially pursued incorrect paths and faced difficulties in redirecting towards the correct path were assigned the lowest abilities.

The latent ability estimated by SRM-PEI demonstrated a very high correlation with the ability estimates from both SRM ($r = 0.987^{***}$) and SRMM ($r = 0.975^{***}$). Furthermore, the correlation of the problem-solving abilities as assessed in PISA 2012 with latent abilities from SRM-PEI was marginally higher ($r = 0.608^{***}$) compared to those obtained from SRM ($r = 0.601^{***}$) and SRMM ($r = 0.607^{***}$). This finding suggests that SRM-PEI may provide a slightly more accurate representation of students' problem-solving abilities. In addition, the latent abilities of SRM-PEI revealed significant correlations with aggregated variables that describe the problem-solving process. These correlations are indicative of the model's nuanced understanding

TABLE 4.
Marginal posterior distributions for the transition tendency parameters of SRM-PEI for the Ticket task.

Transition	Transition effectiveness	Scaled transition effectiveness	Frequency	Mean	Median	SD	95%HPDL	95%HPDU
AB	1	1	29095	0.547	0.531	0.009	0.530	0.564
AG	−1	−1	14667	−0.547	−0.531	0.009	−0.564	−0.530
BA	−1	−1/3	938	−1.406	−1.555	0.023	−1.452	−1.362
BC	1	1	24083	1.742	1.665	0.019	1.705	1.778
BH	−2	−1	3989	−0.336	−0.109	0.019	−0.373	−0.299
CA	−2	−0.5	584	−1.564	−1.661	0.028	−1.620	−1.511
CD	1	1	20048	1.610	1.553	0.024	1.560	1.657
CI	−3	−1	3424	−0.047	0.107	0.022	−0.090	−0.005
DA	−3	−0.6	383	−1.345	−1.415	0.041	−1.423	−1.263
DE	1	1	16938	1.885	1.797	0.027	1.832	1.936
DF	0	−1	1010	0.045	−0.455	0.027	−0.010	0.098
DL	−5	0.6	1708	−0.585	0.073	0.036	−0.655	−0.515
EA	−4	−1	767	−0.826	−0.909	0.035	−0.895	−0.757
EF	−1	1	16204	−0.551	1.491	0.025	−0.600	−0.505
EK	1	0.2	1064	1.376	−0.581	0.026	1.328	1.426
FA	−3	−0.6	298	−0.698	−0.945	0.044	−0.786	−0.616
FE	1	1	1583	0.455	0.875	0.046	0.368	0.546
FL	−5	−1	825	0.243	0.070	0.038	0.171	0.322
GA	1	1	1900	−0.725	−0.629	0.014	−0.752	−0.699
GH	0	−1	12619	0.725	0.629	0.014	0.699	0.752
HA	1	1	1603	−0.876	−0.778	0.014	−0.903	−0.850
HI	0	−1	14906	0.876	0.778	0.014	0.850	0.903
IA	1	1	4295	0.085	0.303	0.013	0.058	0.110
IJ	0	−1	9170	−0.388	0.194	0.012	−0.413	−0.366
IL	−1	−1	4590	0.303	−0.497	0.010	0.283	0.323
JA	1	1	1264	−0.081	0.228	0.018	−0.116	−0.046
JL	−1	−1	2946	0.081	−0.228	0.018	0.046	0.116

95%HPDL: Lower bound of 95% highest posterior density interval. 95%HPDU: Upper bound of 95% highest posterior density interval.

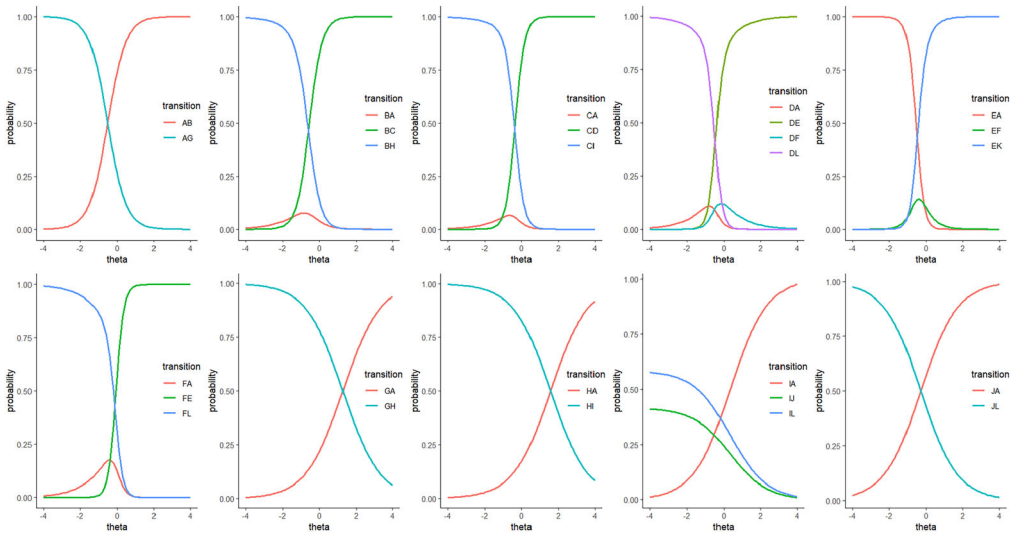

FIGURE 3.
State transition characteristic curves for all transitions under each state of the Ticket task as estimated by SRM-PEI

of students' problem-solving strategies. Specifically, students with higher abilities were more likely to make progress towards the target (indicated by $\Delta d_{s\to s'} = 1, r = 0.976^{***}$), and less inclined to maintain a constant distance ($\Delta d_{s\to s'} = 0, r = -0.916^{***}$) or move backwards ($\Delta d_{s\to s'} < 0, r = -0.843^{***}$). They generally exhibited a shorter average distance to the target throughout their sequence of actions ($r = -0.984^{***}$). These findings demonstrate at the behavioral level that students with high abilities assessed by SRM-PEI tend to engage in more actions that are conducive to problem-solving and fewer actions that are detrimental to it. Moreover, they exhibit a preference for choosing paths that are easier to approach the targets throughout the problem-solving process. This interpretation underscores the importance of considering both the directionality of actions (towards or away from the target) and the overall strategic approach in assessing problem-solving abilities. The SRM-PEI's ability to capture these aspects highlights its utility in providing a comprehensive evaluation of problem-solving skills in educational assessments.

TABLE 5.
Marginal posterior distributions for the top five and bottom five abilities estimated by SRM-PEI and corresponding response patterns for the Ticket task.

Sequence of states	Sequence coded by state effectiveness [a]	Frequency	Mean	Median	SD	95% HPDI
ABCDEK	543210	11111	0.827	0.773	0.703	(−0.467, 2.229)
ABCDE	54321	248	0.760	0.709	0.722	(−0.579, 2.197)
ABCD	5432	5	0.692	0.633	0.745	(−0.666, 2.185)
ABCDABCDEK	5432543210	92	0.684	0.635	0.572	(−0.366, 1.827)
ABCDABCDE	543254321	1	0.626	0.578	0.568	(−0.386, 1.783)
AGH	566	66	−0.910	−0.880	0.726	(−2.364, 0.462)
AGHIJ	56666	186	−0.934	−0.898	0.639	(−2.216, 0.269)
AGHI	5666	137	−0.964	−0.926	0.700	(−2.365, 0.352)
AGHIJL	566667	1665	−1.073	−1.033	0.606	(−2.286, 0.061)
AGHIL	56667	4964	−1.134	−1.093	0.673	(−2.479, 0.126)

95%HPDI: 95% highest posterior density interval.
[a] The end of the number 0 means that the correct ticket was bought, the end of the number 7 means that the wrong ticket was bought, and the end of the other numbers means that the task was quit midway.

5.2. *Balance Beam Task*

5.2.1. Data Description and Analysis Process for the Balance Beam Task Students from eighth and ninth grades across six schools in three regions of China participated in the Chinese version of the Balance Beam task developed by Yuan et al. (2019). In this study, we only used records from the sub-task requiring two weights to balance the beam. The states and transitions in the records inconsistent with the system settings (as detailed in Sect. 1.3) were excluded from the analysis. After data cleaning, there were a total of 422 groups with 167 successfully completing the task. On average, each group executed 33 transitions.

Based on the classification of effectiveness values in Sect. 1.3, there are 6 types of states and 16 types of state transitions. Additionally, Yuan's testing system allows students to exit the test either midway through or after task completion. As a result, the extra termination states (marked as #) were added. Unlike the incorrect end state in the Ticket task, we defined the effectiveness of the transitions leading to the termination states in two cases: Exiting the system after reaching any target state (the transition is denoted as 0→#) was considered correct with an effectiveness of 1 (i.e., $\Delta d_{0\to\#} = 1$). Since the task could not be continued after the termination, the transitions from the 5 types of non-target states ($d_s = 5, 4, \ldots, 1$) to the termination state were the incorrect early termination. A lower value of effectiveness than any other transition was assigned to these 5 transitions of early termination ($\Delta d_{s\to\#} = -2$ for any state s if $d_s > 0$), indicating them as the least preferable among the 22 (i.e., 16+1+5) transition types. In this study, we classified transitions based on the values of state effectiveness before and after the transition (22 categories), rather than using the original state representations from the task interface (168,000 categories) or the values of transition effectiveness (4 categories, i.e., $\mathbf{\Delta} d_{s\to s'} = -2, -1, 0, 1$). By doing so, we ensured a manageable number of transition tendency parameters, with grouped transitions more likely to conform to the premise of equal difficulty, thereby sharing the same tendency parameters in the SRM-PEI.

Since Sect. 1.3 has defined the target states and the two types of effectiveness at the group level for the Balance Beam task, the measurement model combined with these effectiveness was designed to estimate the collective problem-solving ability of groups. To facilitate a comparison between SEM-PEI and the models utilizing dichotomous effectiveness, we adapted the polyto-

mous effectiveness into dichotomous form. Specifically, transitions that progress closer to the targets (i.e., $\Delta d_{s\to s'} = 1$) retained their effectiveness value of 1, denoting correct transitions. In contrast, transitions that lead away from the targets (i.e., $\Delta d_{s\to s'} < 0$) were considered incorrect. For SRM and SRMM, these transitions were reassigned values of -1 or 0 as dichotomous effectiveness, respectively. Transitions that keep a constant distance from the target (i.e., $\Delta d_{s\to s'} = 0$) were bifurcated as either correct or incorrect, leading to two distinct versions of both SRM (termed SRM-v1 and SRM-v2) and SRMM (termed SRMM-v1 and SRMM-v2). The latent abilities estimated by these four versions offered a basis for validating the SRM-PEI. Mirroring the approach used for the Ticket task, we computed the proportion of each of the four types of transitions, categorized by their effectiveness values, within every sequence. This computation also included an assessment of the average distance from the target throughout the problem-solving process. For the Balance Beam task with multiple targets, we defined the nearest targets as the target states closest to the current state. Moreover, we quantified the average number of nearest targets, and the proportion of transitions that either augmented or reduced this number. To evaluate the efficacy of the SRM-PEI, we scrutinized the correlations between these seven aggregated indicators, based on both state and transition effectiveness, and the latent abilities estimated by the SRM-PEI. These correlations were integral to understanding the validity and interpretability of the SRM-PEI in measuring group problem-solving abilities.

5.2.2. Results of the Empirical Study on the Balance Beam Task All models met the criteria of convergence and good model-data fit as shown in Table 3. The trace plots for the ability and transition tendency parameters specific to SRM-PEI are displayed in Figures S2 and S3 in the Supplementary Material The model comparison metrics indicated a preference for SRM-PEI over the two versions of SRM and SRMM. These preliminary findings suggest that, considering the complexity of the model, SRM-PEI is more apt at predicting data in complex problem-solving tasks that encompass multiple optimal paths.

Upon evaluating the transition tendency parameters estimated by SRM-PEI (see Table 6) and by two versions of SRM and SRMM (see Table S4 in the Supplementary Material), the assignment schemes of the effectiveness for transitions maintaining the same position (i.e., $\Delta d_{s\to s'} = 0$) played an important role. When these transitions were considered incorrect, their tendency parameter rankings generally decreased A comparison of the rankings for the tendency parameters across each group of transitions revealed that SRM-v1 aligned most closely, albeit not identically, with the estimations of SRM-PEI. Another crucial observation was that assigning a small effectiveness value of -2 to the five types of early termination transitions did not compromise the rationality and order of the tendency parameters in SRM-PEI. These transitions uniformly received the lowest rankings in terms of transition tendency parameters, signifying a strong preference to avoid prematurely ending the task. This inference was supported by the low percentages of groups that opted for early termination (0.8%, 1%, 4%, 0.6%, and 0.5%).

Figure 4 displays the characteristic curves for each group of transitions fitted by SRM-PEI. These curves demonstrated the nuanced capabilities of SRM-PEI, especially with its incorporation of polytomous transition effectiveness, in precisely fitting probability curves aligned with the task design. For states that were not the target, groups with lower abilities tended to exhibit a higher likelihood of either terminating the task prematurely or moving away from the target. In contrast, groups with medium abilities generally engaged in actions that maintain the current distance from the target. Most notably, groups with the highest abilities demonstrated a pronounced propensity to make transitions that brought them closer to the target. Upon reaching the target state, the pattern of responses shifted. Groups with the highest abilities are most likely to correctly conclude the task and exit the system, a behavior indicative of successful task completion. However, groups with lower abilities might persist in actions like passing weights (which did not affect the balance) or hanging or removing weights (which could disrupt the balance). These observed behaviors and

TABLE 6.
Marginal posterior distributions for the transition tendency parameters of SRM-PEI for the Balance Beam task.

Transition	Transition effectiveness	Frequency	Mean	Median	SD	95%HPDL	95%HPDU
5→#	−2	1	−1.742	−1.722	0.418	−2.538	−0.911
5→5	0	44	0.124	0.131	0.306	−0.462	0.766
5→4	1	84	1.618	1.616	0.274	1.074	2.132
4→#	−2	44	−2.203	−2.182	0.187	−2.606	−1.870
4→5	−1	85	−1.549	−1.549	0.134	−1.813	−1.289
4→4	0	1911	1.981	1.970	0.079	1.849	2.152
4→3	1	1349	1.770	1.757	0.117	1.557	1.999
3→#	−2	233	−1.759	−1.753	0.120	−2.000	−1.546
3→4	−1	1394	0.304	0.303	0.051	0.208	0.400
3→3	0	2870	1.096	1.094	0.048	1.001	1.184
3→2	1	1473	0.358	0.362	0.099	0.184	0.575
2→#	−2	2	−1.918	−1.921	0.146	−2.209	−1.668
2→3	−1	1329	1.217	1.217	0.069	1.077	1.351
2→2	0	1404	1.001	1.004	0.059	0.878	1.109
2→1	1	528	−0.300	−0.295	0.109	−0.522	−0.095
1→#	−2	5	−0.900	−0.909	0.176	−1.249	−0.584
1→2	−1	404	0.945	0.943	0.089	0.788	1.126
1→1	0	338	0.528	0.529	0.087	0.355	0.702
1→0	1	219	−0.572	−0.567	0.132	−0.835	−0.322
0→ 1	−1	53	0.540	0.538	0.160	0.238	0.817
0→0	0	24	0.484	0.493	0.129	0.225	0.729
0→#	1	119	−1.024	−1.018	0.170	−1.352	−0.685

95%HPDL: Lower bound of 95% highest posterior density interval.
95%HPDU: Upper bound of 95% highest posterior density interval.

the corresponding probability curves underscore the effectiveness of SRM-PEI in capturing the likelihood of various transitions accurately.

Table 7 showcases sequences associated with the highest and lowest abilities estimated by SRM-PEI. With the help of state effectiveness, we could clearly observe the problem-solving process in complex tasks in which the massive original states were challenging to represent by letters. High-ability groups efficiently located and followed optimal paths, demonstrating proficient problem-solving processes. Conversely, low-ability groups wandered in states further from the targets than the initial state and finally terminated the test process. States with an effectiveness value of 4, which were one step further from the targets than the initial state with an effectiveness value of 3, typically resulted from an incorrect action such as the improper transfer or suspension of one weight.

The latent problem-solving abilities of groups estimated by SRM-PEI exhibited a very high correlation with those derived from SRM ($r = 0.922$*** for SRM-v1 and $r = 0.931$*** for SRM-v2) and SRMM ($r = 0.888$*** for SRMM-v1 and $r = 0.924$*** for SRMM-v2), indicating strong consistency across these models. In terms of aggregated indictors, groups with higher abilities demonstrated a greater likelihood of advancing towards the target ($\Delta d_{s\to s'} = 1$, $r = 0.845$***). They were less prone to maintaining the same distance ($\Delta d_{s\to s'} = 0$, $r = -0.502$***), retreating one step ($\Delta d_{s\to s'} = -1$, $r = -0.574$***), or terminating the task prematurely ($\Delta d_{s\to s'} = -2$, $r = -0.381$***). These groups also showed a shorter average distance to targets throughout the sequence ($r = -0.928$***), suggesting efficient progression towards task completion. Furthermore, they tended to focus on fewer targets ($r = -0.422$***),

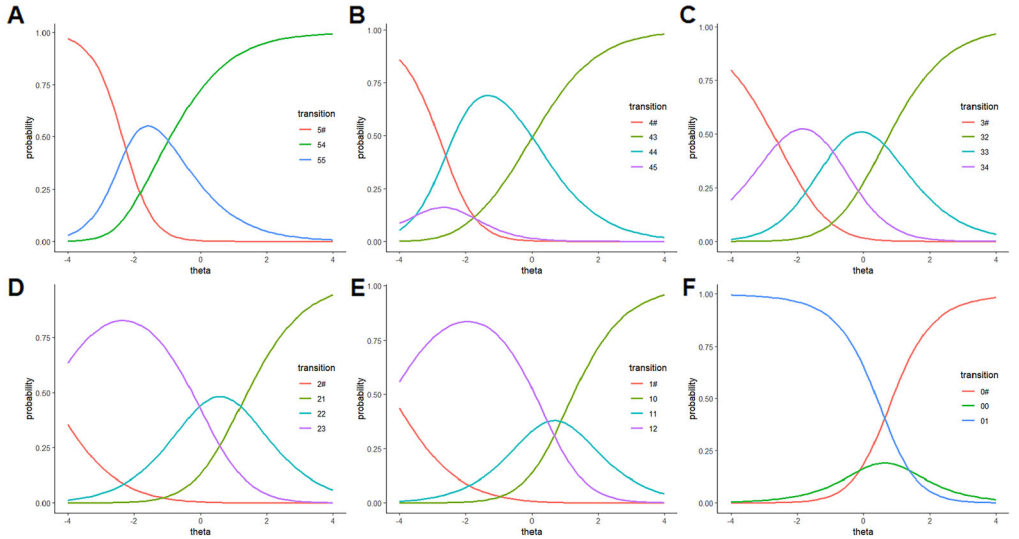

FIGURE 4.
State transition characteristic curves for all transitions in each state of the Balance Beam task as estimated by SRM-PEI.

TABLE 7.
Marginal posterior distributions for the top five and bottom five abilities estimated by SRM-PEI and corresponding response patterns for the Balance Beam task.

Sequences coded by state effectiveness [a]	Frequency	Mean	Median	SD	95% HPDI
3210#	11	2.085	2.055	0.606	(0.932, 3.288)
32100#	1	1.681	1.660	0.494	(0.754, 2.675)
33210#	2	1.672	1.653	0.503	(0.717, 2.682)
321210#	1	1.549	1.534	0.439	(0.672, 2.369)
322110#	1	1.531	1.511	0.438	(0.709, 2.430)
33334444444#	1	−1.025	−1.031	0.304	(−1.607, −0.421)
323434444333444444444#	1	−1.035	−1.036	0.331	(−1.710, −0.426)
334343443444444#	1	−1.062	−1.064	0.360	(−1.792, −0.391)
3334444#	1	−1.102	−1.108	0.366	(−1.853, −0.399)
344444#	1	−1.474	−1.490	0.386	(−2.218, −0.709)

95%HPDI: 95% highest posterior density interval.
[a] The state effectiveness for the termination state (#) is not defined.

and take more actions that reduced ($r = 0.735$***) rather than increasing the number of nearest targets ($r = -0.471$***). These findings underscore the strong alignment between the latent ability as estimated by SRM-PEI and the actual performance in the problem-solving process. The results affirm that SRM-PEI effectively characterizes the procedural aspects of evaluation, highlighting its utility in assessing complex problem-solving skills.

6. Discussion

In this study, we innovated a method that assesses the effectiveness indicators of problem states d_s and transitions $\Delta d_{s\to s'}$ in problem-solving tasks. Moreover, we proposed a measurement

model named the sequential response model with polytomous effectiveness indicators (SRM-PEI) Through simulation and empirical studies, we demonstrated that the effectiveness indicators and SRM-PEI are capable of estimating latent problem-solving ability in various types of tasks.

Following the concepts of problem-solving and the characteristics of interactive tasks, we defined the effectiveness of a state as the theoretical shortest distance from the target state and the effectiveness of a transition between two states as the change in the theoretical shortest distance to the target. To facilitate the application, we proposed a general algorithm for computing the effectiveness indicators and illustrated the calculation process and results using two real tasks. Applied to measurement models and sequence-based aggregated features, we established an accessible methodology that promotes standardization and accuracy in the interactive problem-solving tests. Our proposed effectiveness indicators demonstrate several primary benefits. First, the automated nature of the evaluation method allows for rapid computation of the effectiveness of all states and transitions through straightforward programming. Second, our evaluation approach is not reliant on response data collection and can be performed once the task simulation system is designed or the task is planned out. Third, in the context of complex tasks with multiple states and transitions, the effectiveness indicators assist in simplifying and categorizing states and transitions. The indicator values provide a clear semantic understanding of different categories, as illustrated in the Balance Beam task. Fourth, in the simple task with a single optimal path, such as Ticket, the polytomous effectiveness indicators offer a more detailed classification and richer information compared to dichotomous effectiveness. Fifth, in the reinforcement learning framework that first proposed action effectiveness (i.e., (LaMar, 2018)), there is also a state-value function, closely related to the action value function, which computes the expected weighted rewards in the future for a given state, but the calculation is complex and difficult to use. Our proposed state effectiveness has a concise meaning and low computational complexity to evaluate the value of states, which enriches the applicable scope and improves the usability of effectiveness indicators.

The SRM-PEI model leverages the full potential of the effectiveness indicator $\Delta d_{s\to s'}$ for evaluating and differentiating various types of state transitions, especially filling the gaps in the modeling and analysis of process data in complex problem-solving tasks. Furthermore, the SRM-PEI model introduces a new tool for data analysis that can facilitate the development of more intricate and realistic problem-solving interactive tests, as well as the evaluation of higher-level cognitive abilities. Both simulation and empirical studies demonstrated that the SRM-PEI model provides a comprehensive characterization of the easiness and probability of occurrence for a vast number of state transitions. The transition characteristic curves not only serve to further evaluate whether the model accurately fits the data, but also provide a detailed and intuitive description of the difficulty of each transition within the task. These curves can be utilized to study the adaptability of SRM-PEI to various tasks, inform the design of scoring based on transition tendencies, and validate the scores. In this study, in order to test the validity of latent abilities, we also innovatively created many aggregation indicators based on two effectiveness indicators, and all of them have a very high correlation with the latent ability estimated by SRM-PEI.

Unlike the values of dichotomous effectiveness indicators in CTDC, SRM, and SRMM, which merely categorize responses as correct or incorrect, the effectiveness indicators $\Delta d_{s\to s'}$ provide more practical meaning, as their value signifies the distance toward the target that a transition affords. Furthermore, the SRM-PEI broadens the applicability of these types of models to encompass more complex problem-solving tasks. The aforementioned models—CTDC, SRM, and SRMM—while insightful, are best suited to the problem-solving task with a single optimal path. In contrast, the introduction of effectiveness indicators allows SRM-PEI to navigate inherently complex tasks with multiple optimal paths, as demonstrated by the Balance Beam task. Therefore, the development of polytomous effectiveness and SRM-PEI signals a meaningful progression in the analysis of process data in complex problem-solving tasks. We also simplify the polytomous effectiveness of transitions to dichotomous version to makes SRM and SRMM

capable of the Balance Beam task. From a different perspective, the SRM can be considered a special case of the SRM-PEI with restricted effectiveness indicators. If the ability and tendency parameters in SRM-PEI are reparametrized, then both the CTDC and SRMM can also be viewed as special cases of the SRM-PEI with restricted effectiveness indicators. Additionally, two issues need to be considered when estimating models with the polytomous effectiveness $\Delta d_{s\to s'}$: First, this study categorizes transitions in the Balance Beam task based on the values of effectiveness, which implies the assumption that all transitions with the same value of $\Delta d_{s\to s'}$ have equal difficulty and can be estimated with the same transition tendency parameter. This assumption needs to be evaluated for its applicability to different tasks. Second, when some values of $\Delta d_{s\to s'}$ are excessively small as demonstrated in the Ticket task, it is feasible to scale the original values to a range appropriate for the NRM framework to achieve a better model-data fit of SRM-PEI.

This study serves as an initial exploration, and several areas warrant further research in the future. (1) Effectiveness indicators can be leveraged across different levels of process data analysis and integrated with a wider range of models and analytic methodologies. In the framework of computational psychometrics put forward by von Davier (2017) for unstructured data in computer-based interactive assessments, effectiveness indicators could be used not only in measurement models, but also in sequence-based analysis approaches. (2) In addition to the two approaches demonstrated on the Ticket task and the Balance Beam task, there are various ways to assign effectiveness values to incorrect end or termination states and transitions leading to these states. The impact of different effectiveness values on SRM-PEI or other methods is also worth exploring. (3) There are numerous possibilities to improve the SRM-PEI. Though two empirical studies have all been conducted on a single sub-task, SRM-PEI is capable of handling multiple tasks like other psychometrics process models. After classifying states and transitions with the same evaluation procedure, SRM can be utilized to analyze the total ability across multiple sub-tasks. From a methodology perspective SRM-PEI could also be extended to a multidimensional form to estimate the abilities of two individuals, as well as two distinct types of abilities in collaborative problem-solving tasks (Yuan et al., 2019; Li et al., 2023)

Funding No funding was received to assist with the preparation of this manuscript. The authors have no competing interests to declare that are relevant to the content of this article.

Data Available The data analyzed in the empirical example of this study are available on this project's Open Science Framework (OSF) page: https://osf.io/fw82q/.

Code Availability The codes are available on this project's Open Science Framework (OSF) page: https://osf.io/fw82q/.

Appendix A. Algorithm for Automatically Calculating State Effectiveness in the Balance Beam Task

In the Balance Beam task, the ten potential positions for each weight are categorized into four groups: (1) Positions 1–4: Positioned on side A of the beam; (2) Position 5: Not suspended on side A; (3) Position 6: Not suspended on side B; (4) Positions 7–10: Positioned on side B of the

beam. Figure 5 illustrates the transition of each weight among ten positions through four types of operations: (1) removing a weight from the beam; (2) hanging an unhung weight; (3) passing a weight to the other student; and (4) shifting the position of a weight on the same side. Each arrow represents an operation that can lead to a transition. Through this figure, we can easily find the minimum number of transitions between any two positions for one weight. Since an operation can only alter the position of one weight once, the shortest distance between states s and s' equals the sum of the minimum number of operations required for each of the four weights to change its position from state s to s'. Then, we can quickly and accurately calculate the shortest distance $d_s^{(k)}$ between a state s and the target state $s_{target}^{(k)}$ using the state code and rules to change the position according to Fig. 6. Finally, we select the minimum distance $d_s = \min\left(d_s^{(1)}, d_s^{(2)}, \ldots, d_s^{(k)}\right)$ as the effectiveness indicator d_s of the state s

During the process of programming the calculations mentioned above, the position of each weight can be assigned a unique number from one to ten. Therefore, any given state in the Balance Beam task can be encoded by a sequence of four numbers, a representation we refer to as the state code. For one weight, calculating the shortest distance between any two positions can be simplified by several rules. The R code for evaluating the effectiveness of states for the Balance Beam task that requires the use of two weights to achieve balance is available at https://osf.io/fw82q/.

In the example of the code, the four positions for hanging weights on the balance beam on student A's side are coded as 1 to 4, and the four positions on student B's side are coded as -1 to -4. The unhung weights are coded as 0.5 when in student A's hand and -0.5 when in student B's hand. In the initial state, all four weights are in the hand of A, and the state code is (0.5, 0.5, 0.5, 0.5). The effectiveness of the initial state is equal to 3, which means that the balance state using two weights can be achieved after a minimum of three transitions. Another example is that Student B holds the 50 g and 100 g weights and Student A has hung the 300 g weight at position 1 and the 500 g weight at position 2. This state is at a minimum distance of 2 from the balance state.

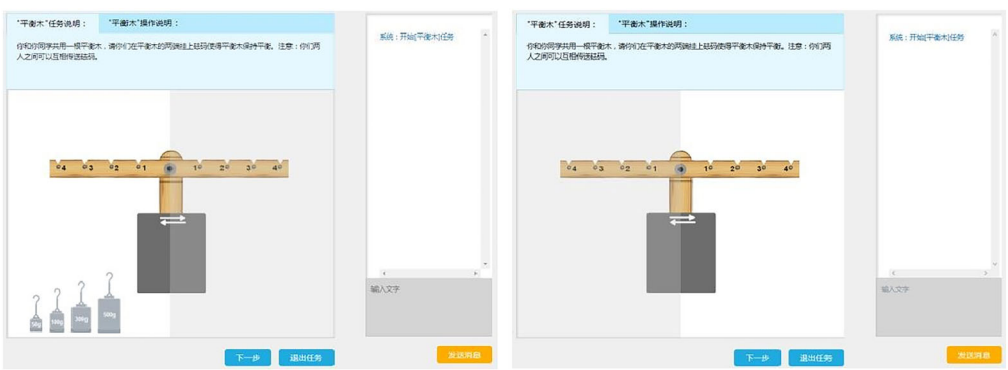

FIGURE 5.
The interface of the initial state in the Chinese version of the Balance Beam task.

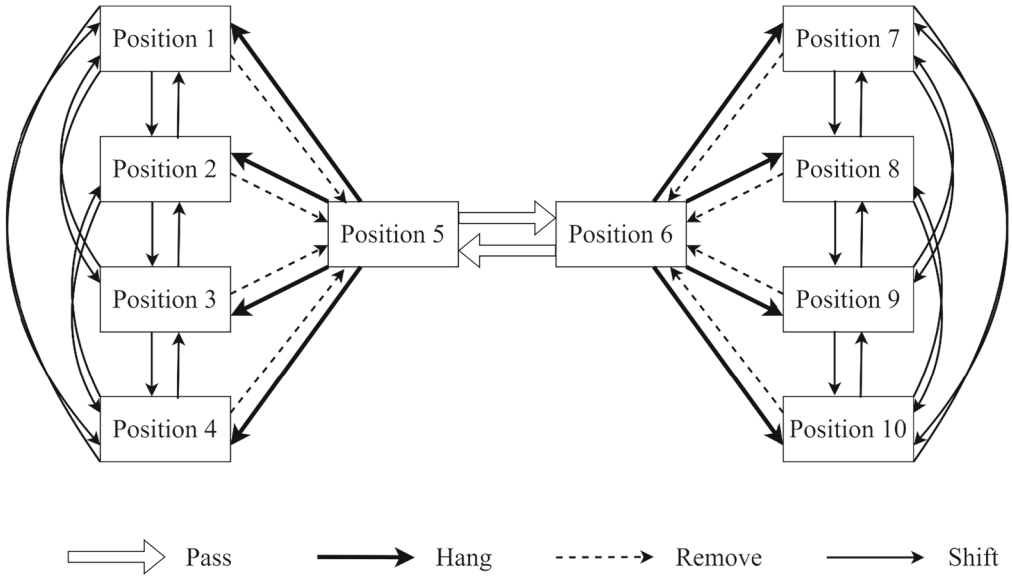

FIGURE 6.
The diagram for the four types of transitions that can occur when a weight moves among ten possible positions in the Balance Beam task.

References

Anderson, J. R., Funke, J., & Plata, G. (Eds.). (2007). *Cognitive psychologic* (6 Aufl.). Spektrum Akademischer Verlag http://www.gbv.de/dms/bs/toc/529836963.pdf

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. https://doi.org/10.3389/fpsyg.2019.00083

Bergner, Y., & von Davier, A. A. (2019). Process data in NAEP: Past, present, and future. *Journal of Educational and Behavioral Statistics, 44*(6), 706–732.

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

Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics, 7*, 434–455.

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. https://doi.org/10.1007/s11336-020-09734-1

Cormen, T. H., Leiserson, C. E., Rivest, R. L., & Stein, C. (2022). *Introduction to algorithms* (4th ed., pp. 563–572). Cambridge: MIT Press.

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

Funke, J. (2001). Dynamic systems as tools for analysing human judgement. *Think Reason, 7*, 69–89.

Geisser, S., & Eddy, W. F. (1979). A predictive approach to model selection. *J Am Stat Assoc, 74*, 153–160.

Gelfand, A. E., & Dey, D. K. (1994). Bayesian model choice: Asymptotics and exact calculations. *Journal of the Royal Statistical Society Series B, 56*, 501–514.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (3rd ed.). Boca Raton: Chapman & Hall/CRC Press.

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*(4), 457–472. https://doi.org/10.1214/ss/1177011136

Griffin, P., & Care, E. (2014). *Assessment and teaching of 21st century skills: Methods and approach*. New York, NY: Springer.

Guttman, I. (1967). The use of the concept of a future observation in goodness-of-fit problems. *Journal of the Royal Statistical Society: Series B (Methodological), 29*(1), 83–100. https://doi.org/10.1111/j.2517-6161.1967.tb00676.x

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.
He, Q., & von Davier, M. (2015). Identifying feature sequences from process data in problem-solving items with n-grams. In L. A. van der Ark, D. M. Bolt, W.-C. Wang, J. A. Douglas, & S.-M. Chow (Eds.), *Quantitative psychology research* (pp. 173–190). Berlin: Springer. https://doi.org/10.1007/978-3-319-19977-1_13
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 Y. Rosen, S. Ferrara, & M. Mosharraf (Eds.), *Handbook of research on technology tools for real-world skill development* (pp. 750–777). Pennsylvania: IGI Global.
Koehler, E., Brown, E., & Haneuse, J. P. A. (2009). On the assessment of monte carlo error in simulation-based statistical analyses. *The American Statistician, 63*(2), 155–162. https://doi.org/10.1198/tast.2009.0030
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*. Cambridge: CRC Press.
Li, M., Liu, H., Cai, M., & Yuan, J. (2023). Estimation of individuals' collaborative problem solving ability in computer-based assessment. *Education and Information Technologies*. https://doi.org/10.1007/s10639-023-12271-w
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.
Mayer, R. E., & Wittrock, M. C. (2006). Problem solving. In P. A. Alexander & P. H. Winne (Eds.), *Handbook of educational psychology* (2nd ed., pp. 287–304). Mahwah: Erlbaum.
Newell, A., & Simon, H. A. (1972). *Human problem solving* (Vol. 104). Englewood Cliffs: Prentice-Hall.
OECD (2014). *PISA 2012 results: Creative problem solving: Students' skills in tackling real-life problems* (Vol. V). OECD.
OECD. (2016). *PISA 2015 Assessment and analytical framework: Science*. Reading, mathematic and financial literacy: PISA. OECD Publishing. https://doi.org/10.1787/9789264255425-en
OECD. (2018). *The future of education and skills: Education 2030*. Paris: OECD Publishing.
Patz, R. J., & Junker, B. W. (1999). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics, 24*(4), 342–366. https://doi.org/10.3102/10769986024004342
Patz, R. J., & Junker, B. W. (1999). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics, 24*(2), 146–178. https://doi.org/10.3102/10769986024002146
Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applies statistician. *The Annals of Statistics, 12*, 1151–1172. https://doi.org/10.2307/2240995
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.
Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van der Linde, A. (1998). *Bayesian deviance, the effective number of parameters, and the comparison of arbitrarily complex models*. MRC Biostatistics Unit: Technical report.
Tang, X. (2023). A latent hidden Markov model for process data. *Psychometrika*. https://doi.org/10.1007/s11336-023-09938-1
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.
von Davier, A. A. (2017). Computational psychometrics in support of collaborative educational assessments. *Journal of Educational Measurement, 54*(1), 3–11.
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, 1–20*. https://doi.org/10.3758/s13428-022-02042-9
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. https://doi.org/10.3389/fpsyg.2019.0036

Manuscript Received: 5 AUG 2023
Published Online Date: 9 APR 2024