Feature Extraction and Ability Assessment from Process Data in Problem-Solving Tests

Authors: Han Yuting, Xiao Yue, Hongyun Liu, Liu Hongyun

Date: 2021-12-04T11:37:59+00:00

Abstract

Computer-based problem-solving tests can record in real time the detailed action traces of participants as they explore the environment and solve problems, and save them as process data. First, the analysis workflow of process data is introduced, and then, starting from problem-solving tests, research on two aspects of process data—feature extraction and ability estimation modeling—is reviewed and evaluated respectively. Future research should focus on: improving the interpretability of analysis results; incorporating more information during feature extraction; achieving ability assessment in more complex problem scenarios; emphasizing the practicality of methods; and integrating and drawing upon analytical methods from different fields.

Full Text

Feature Extraction and Ability Assessment of Process Data in Problem-Solving Tests

Han Yuting[1], Xiao Yue[2,3], Liu Hongyun[2,3]

[1]National Center for Health Professions Education Development, Peking University Health Science Center, Beijing 100191, China
[2]Beijing Key Laboratory of Applied Experimental Psychology, Beijing Normal University, Beijing 100875, China
[3]Faculty of Psychology, Beijing Normal University, Beijing 100875, China

Abstract
Computer-based problem-solving tests can record respondents' detailed behavioral traces in real time as they explore task environments and solve problems, storing these as process data. This paper first introduces the analytical workflow for process data, then reviews and evaluates research on feature extraction and ability estimation modeling for process data in the context of problem-solving

tests. Future research should focus on: improving the interpretability of analytical results; incorporating more information during feature extraction; enabling ability assessment in more complex problem scenarios; emphasizing the practical utility of methods; and integrating analytical approaches from different fields.

Keywords: computer-based problem-solving test, process data, feature extraction, ability assessment model

Problem solving refers to the cognitive processing undertaken by a problem solver to achieve a specific goal when the solution method is initially unknown (Mayer & Wittrock, 2006). Problem-solving ability is crucial in both educational and other domains. To help students adapt to a dynamically changing society, cultivating general cross-disciplinary problem-solving skills has gained widespread attention both domestically and internationally (Lu Jing, 2017). The International Society for Technology in Education (ISTE) listed "critical thinking, problem solving, and decision making" as one of six competency dimensions in its 2007 revised National Educational Technology Standards for Students (Wang Yongfeng et al., 2007). In 2014, China's Ministry of Education issued the "Opinions on Deepening Curriculum Reform in an All-Round Way and Implementing the Fundamental Task of Fostering Virtue Through Education," which for the first time proposed developing a core competency system for student development and emphasized conducting interdisciplinary thematic education activities to enhance students' problem-solving abilities.

In recent years, with increasing attention to problem-solving skill development and rapid advances in information technology, numerous international large-scale assessment programs have begun developing computer-based problem-solving assessment systems. For example, the Programme for International Student Assessment (PISA), administered by the Organisation for Economic Co-operation and Development (OECD), launched a computer-based simulation scenario problem-solving test in 2012 (OECD, 2013) and added a human-computer interactive collaborative problem-solving assessment in 2015 (OECD, 2017). In 2013, the Programme for the International Assessment of Adult Competencies (PIAAC), also under OECD, measured adults' problem-solving skills in technology-rich environments (PSTRE; Schleicher, 2008). The Assessment & Teaching of 21st Century Skills (ATC21S) project, initiated by Cisco, Intel, and Microsoft, measured students' collaborative problem-solving abilities through computer-based human-to-human interactions (Adams et al., 2015). The National Assessment of Educational Progress (NAEP) Technology and Engineering Literacy (TEL) assessments also included problem-solving ability measurement (PumpRepair; TEL, 2013).

Compared with traditional paper-and-pencil tests, computer-based problem-solving tests can use information technology to construct authentic task contexts, enable dynamic interaction between respondents and test tasks, and record respondents' real-time responses in simulated scenarios as process data (process data). Process data, elicited by specific tasks and problems, reflect the

abilities and mental processes respondents employ to solve problems and represent external manifestations of their underlying psychological activities (Yuan Jianlin, 2018). Process data record not only response outcomes but also solution steps, revealing more about respondents' thinking processes than traditional outcome data. They contain information about strategies used and errors made during problem solving, which helps distinguish low-ability respondents, identify different error types, diagnose causes of errors, and provide targeted suggestions for instructional improvement. Process data can also be used to reconstruct solution processes and identify guessing behavior. In summary, process data are valuable for understanding respondents' problem-solving behavioral patterns.

Although process data contain rich information, how to utilize and understand these data remains a pressing issue (Mislevy, 2019). Unscored process data often appear as timestamped log strings (Hao et al., 2015), recording events ranging from "clickstream" mouse events to language and text produced by respondents to complete tasks. Such log strings cannot be directly analyzed using traditional psychometric models; features reflecting latent traits must first be extracted. However, process data are voluminous and structurally complex, making it difficult to quickly and effectively filter useful information or indicators. Additionally, the temporal and multidimensional characteristics of process data pose challenges for measurement modeling. Moreover, these behavioral performances are real behavioral sequences from respondents' problem-solving processes, all bearing time stamps and exhibiting continuity and processual characteristics along the temporal dimension, potentially violating the assumption of local independence required by traditional psychometric models.

Reviewing progress in this field both domestically and internationally, researchers have recently explored how to extract more information for ability estimation from complex process data and how to establish appropriate and accurate ability assessment models to meet the needs of problem-solving ability evaluation. To help methodological researchers more conveniently understand the latest developments in process data analysis for problem-solving tests and to provide practical users with reference information on analytical workflows and method selection, this paper first briefly introduces the analytical workflow for process data. Second, it summarizes progress in feature extraction and ability assessment modeling for process data, and on this basis compares different methods' applicable scenarios, advantages, and disadvantages. Finally, it discusses future research directions based on current trends in process data analysis.

2. Analytical Workflow for Process Data

The development of information technology has made it possible to construct complex computer-interactive tests, which has stimulated demand for guiding theories on test development and performance assessment in new technological environments. Currently, large-scale computer-based problem-solving test projects including PISA and ATC21S rely on Evidence-centered Design (ECD;

Mistevy et al., 2006) theory as their overall design model. The test development and process data collection and analysis process based on ECD can be summarized into five steps shown in Figure 1 [Figure 1: see original paper], with "designing task prototypes" and "process data analysis" differing most from traditional paper-and-pencil tests. Von Davier (2017) and Mislevy (2019) have each proposed their own perspectives on the process data analysis workflow.

Computer-interactive tests developed based on ECD theory can collect rich behavioral performance data from respondents during problem solving in the form of video streams, audio streams, and simulation log files. These various forms of recorded process data can also be collectively referred to as multimodal data. Processing and analyzing multimodal data can facilitate research and understanding of individual and group-level performance (Amer et al., 2014; Morency et al., 2010; Siddiquie et al., 2013). Von Davier (2017) summarized an analytical framework for unstructured data in computer-interactive tests—computational psychometrics—based on the multimodal hierarchical approach (Khan, 2017; Khan et al., 2013). This framework combines data-driven research methods from computer science (particularly machine learning and data mining), stochastic process theory, and theory-driven psychometrics to enable real-time measurement of latent abilities. Its basic idea is illustrated in Figure 2 [Figure 2: see original paper]: First, projects are developed following ECD principles, tests are administered, and multimodal data (process data) are collected together with traditional test item data (outcome data). This test development and data collection procedure relies on theoretical input from human expert systems and is a top-down process. Then, data mining (DM) and machine learning (ML) algorithms are used for feature extraction and representation from multimodal data. If new behavioral performance features are identified, they can be considered for inclusion in subsequent psychometric model construction (Von Davier, 2017). Next, measurement models are updated, and the process is repeated with new samples. Stochastic process models can also be used if data permit, cycling through the process until measurement models are finalized.

Mislevy (2019) argued that two basic analytical processes help explain and model process data. The first is describing evidence in given behavioral performances—that is, extracting useful information (evidence) from complex and diverse process data. This resembles the hidden process in human raters' minds when assessing respondents' complex performances. In addition to expert-specified extraction rules, this analytical procedure can be accomplished with techniques such as data mining, knowledge engineering, and computational linguistics (Bejar et al., 2016). The second is measurement modeling. In computer-based tests, we can track, accumulate, and synthesize evidence from behavioral performance processes and construct operationalized variables for target constructs. These behavioral performance features depend on respondents' latent characteristics, and their probabilistic relationships can be modeled by measurement models.

Integrating these perspectives, analyzing process data from computer-based

problem-solving tests involves two main steps: extracting interpretable information about respondents' latent abilities from process data, and using the extracted information to estimate respondents' abilities. In the information extraction stage, there are top-down approaches that rely on experts and bottom-up approaches that are data-driven. In the ability estimation stage, traditional psychometric models can be employed, or stochastic process models can be selected if data permit. The following sections review and summarize the latest research progress on these two core steps of process data analysis—feature extraction and ability assessment.

3. Feature Extraction Methods for Process Data

Currently, methods for extracting key features or meaningful behavioral indicators from problem-solving test process data mainly include theory-driven (top-down) and data-driven (bottom-up) approaches.

3.1 Top-Down Feature Extraction Methods

Top-down feature extraction methods refer to the process where, based on a problem-solving conceptual framework and combined with specific tasks, experts formulate rules to identify meaningful behavioral patterns in process data that are associated with elements of the problem-solving construct. The specific process is shown in Figure 3 [Figure 3: see original paper]: Based on the test conceptual framework, the expert panel must operationalize the construct connotation for each specific task scenario, specifying possible performance levels in the task, and then develop detailed process indicator extraction and scoring rules. Generally, multiple experts are organized to conduct iterative work on behavioral indicator design, review, and revision. After determining indicator extraction rules, they must be converted into program algorithms to enable automated extraction of process data. To ensure the validity of behavioral indicators and their scoring rules, expert panels need to clearly understand respondents' cognitive processes during task completion during the rule-writing stage. After obtaining indicator scores for respondents' process data using automated programs, domain experts should also be organized to score the extracted indicators, and inter-rater reliability between experts and between automated scoring results should be examined, with consistency measured using Kappa coefficients.

This approach is currently the mainstream scoring method for large-scale international problem-solving test systems. The PISA 2012 problem-solving test, the ATC21S collaborative problem-solving test (Adams et al., 2015), and the NAEP-TEL test (Shu et al., 2017) all employed expert-defined process data indicator extraction and scoring methods. In other studies involving process data analysis, researchers have also developed corresponding process data coding and scoring rules for different tasks (e.g., Harding et al., 2017; Rosen, 2017; Yuan et al., 2019; Zoanetti, 2010; Yuan Jianlin, 2018). However, top-down approaches

require expert panels to develop specific scoring rules for each task—that is, they suffer from task specificity and high costs.

3.2 Bottom-Up Feature Extraction Methods

Figure 3 [Figure 3: see original paper] Top-Down Feature Extraction Workflow

To address the task specificity issue of theory-driven methods, some researchers have attempted to use data-driven approaches to directly extract information from response sequences recorded in process data. Such methods are still in the preliminary exploration stage and have not formed a unified analytical paradigm; most methods borrow existing algorithms from other fields. Based on their processing philosophy and source domains, bottom-up process data feature extraction methods can be divided into three categories: methods that treat response sequences as analogous to character strings and borrow Natural Language Processing (NLP) techniques to construct indicators from response sequences (He et al., 2021; He & Von Davier, 2016); methods that use dimensionality reduction algorithms to construct low-dimensional numerical feature vectors for response sequences (Tang, Wang, et al., 2021; Tang et al., 2020); and methods that use directed graphs to represent response sequences and network indicators to characterize response features (Vista et al., 2017; Zhu et al., 2016).

3.2.1 NLP-Based Feature Extraction Methods Behavioral operation sequences recorded in process data can be encoded as timestamped string sequences (Hao et al., 2015), such as "Start, Operation 1, Operation 2, Operation 3, End." Therefore, some researchers have proposed treating operation sequences as analogous to words in natural language and using analytical methods from the NLP domain to extract information. Currently employed techniques mainly include N-Gram, edit distance, and indicators based on Longest Common Subsequence (LCS).

N-Gram is a statistical language model-based algorithm that extracts character sequences of length N from text, counts each short sequence, filters out low-frequency sequences, and forms a vector feature space for the text, where each short sequence represents one feature vector dimension. Applying N-Gram to process data involves extracting operation sequences of length N from response sequences and counting them. Some researchers have used this to identify key operation sequences. For example, He and Von Davier (2016) used N-Gram to characterize response sequences from PIAAC problem-solving items, weighted them using term frequency and inverse sequence frequency (TF-ISF) to obtain feature vectors for each operation sequence, then grouped respondents by final outcomes and used chi-square tests to identify key operation sequences associated with successful problem solving. Other researchers have assigned cognitive meaning to extracted N-Grams for further measurement modeling. For instance, Li Meijuan (2020) organized experts to assign cognitive meanings to key short operation sequences identified by N-Gram to define behavioral indicators in collaborative problem-solving tasks. Zhan and Qiao (2020) directly

assigned cognitive meanings to short operation sequences (N-Grams) in process data for diagnostic classification analysis. The method of using N-Gram to extract short operation sequences is computationally simple and easy to implement, and can also construct behavioral indicators through expert definition. However, N-Gram assumes that the occurrence of the Nth operation is only related to the preceding N-1 operations and unrelated to any other operations. Therefore, although this method considers adjacent operations, it loses most sequential information in operation sequences. Moreover, the dimensionality of feature vectors obtained this way equals the total number of all N-Grams, which becomes extremely large when many actions are possible. Additionally, N-Gram depends on how response sequences are recorded; once the encoding method changes, the form and quantity of N-Grams are affected.

For test tasks where the optimal operation sequence is known, it is natural to evaluate respondents' performance based on the similarity between their response sequences and the optimal sequence. Some researchers have borrowed edit distance and Longest Common Subsequence (LCS) from NLP to measure this similarity/difference. Edit distance, also known as Levenshtein distance, refers to the minimum number of editing operations (substitutions, insertions, or deletions) required to transform one string into another (Levenshtein, 1966). The greater the distance between two strings, the more different they are. Zhan et al. (2015) measured respondents' performance in the NAEP-TEL PumpRepair task (TEL, 2013) by comparing the Levenshtein distance between their operation sequences and the optimal sequence. Longest Common Subsequence refers to the longest common part between two given strings. He et al. (2021) constructed indicators assessing response sequence similarity and efficiency based on LCS between respondents' response sequences and optimal sequences. Methods using the distance/similarity between respondents' response sequences and optimal sequences to construct behavioral indicators are also computationally simple and easy to implement, with clear and understandable indicator meanings. However, these indicators also depend on encoding forms, and their high level of aggregation leads to loss of much useful information in process data, making it difficult to distinguish different behavioral patterns.

3.2.2 Low-Dimensional Representation of Operation Sequences Using Dimensionality Reduction Algorithms To extract all process information from response sequences, some researchers have proposed using dimensionality reduction algorithms such as autoencoders and multidimensional scaling (MDS) to obtain numerical feature vectors for response sequences. The extracted numerical vectors can be used to predict respondents' performance or improve ability estimation accuracy. Autoencoders are a class of classic artificial neural networks commonly used for dimensionality reduction, data denoising, and computer visualization (Goodfellow et al., 2016). Tang and Wang et al. (2021) used sequence-to-sequence autoencoder methods to compress respondents' response sequences into standard numerical vectors containing complex information about the original data, which can be analogized to latent abilities in Item

Response Theory (IRT) models, while decoders can be analogized to item response functions. Multidimensional scaling is another data analysis method that projects research objects from multidimensional space into an intuitive low-dimensional vector space based on pairwise distances between objects, simplifying samples or variables for positioning, analysis, and classification while preserving original relationships between objects (Luo Wenshu & Zhao Shouying, 2005). Tang et al. (2020) constructed a function to compute dissimilarity between two operation sequences, then used MDS to analyze pairwise distances between operation sequences, obtaining low-dimensional vectors for each operation sequence. Both Tang and Wang et al. (2021) and Tang et al. (2020) found that low-dimensional vectors obtained through dimensionality reduction algorithms predicted respondents' performance on other items and cognitive tests more accurately than using outcome variables alone.

This method of using dimensionality reduction algorithms to obtain low-dimensional numerical vectors for operation sequences does not depend on prior knowledge or process data encoding. The obtained low-dimensional vectors contain process information and can be further used for clustering response patterns, visualization, and predicting respondents' future performance, thus serving as a general feature extraction method. However, this approach's main problem is interpretability—the low-dimensional representation vectors lack clear psychological meaning.

3.2.3 Methods Using Network Indicators to Describe Response Process Characteristics Social Network Analysis (SNA) can examine relationship structures and network characteristics through systematic analysis of relational data (Xu Wei et al., 2011). Response sequences recorded in process data are not independent activity sets; they contain information about the order of activities respondents use when solving problems. Directed graphs can visually display changes in responses, and SNA indicators can then describe characteristics of response processes. Directed graphs can represent both individual operation sequences and group response processes. For example, Zhu et al. (2016) constructed weighted directed graphs representing interdependence between operations based on each respondent's response sequences in the NAEP-TEL PumpRepair task (Wasserman & Faust, 1994). Vista et al. (2017) used task states and respondent dialogue events as network nodes and chronological order between events as connections to construct group-level network graphs for high- and low-ability groups in the ATC21S Olive Oil task. Network characteristic indicators that can describe response processes include density, centralization, local pattern features such as reciprocity and triad census (Davis & Leinhardt, 1972; Wasserman & Faust, 1994), prominence, branches, clusters, and shortest paths (Vista et al., 2016). Response process network indicators differ across respondents/respondent groups with different performance/ability levels (Zhu et al., 2016; Vista et al., 2017) and have some predictive power for respondent performance.

This approach treats response sequences as holistic processes rather than focusing on individual events. Using network graphs to represent response sequences can visually present response patterns, and SNA indicators can then describe response process characteristics. The main challenges facing this method include data complexity, requiring extensive data cleaning and preprocessing. Additionally, when using SNA indicators to describe characteristics of directed graphs of response processes, only structural network features can be obtained, losing response sequence information and being unable to capture node content information or specific response type information, making it difficult to further infer respondents' performance levels.

3.3 Summary of Feature Extraction Methods

In summary, behavioral indicators defined using top-down approaches have close correspondence with conceptual frameworks, possess interpretability and clear scores, and can be directly analyzed using psychometric models like traditional test items to obtain latent ability estimates. However, such indicator construction methods require enormous workload. Particularly, in complex tasks, experts may overlook or ignore unknown student thinking processes that have not been previously noticed, resulting in information omission and loss.

Data-driven bottom-up feature extraction methods partially solve the task specificity problem of expert-developed scoring rules. Extracted features can be used to explore behavioral pattern characteristics of different respondent groups, predict respondents' future performance, and can be used for ability estimation after expert definition, offering value for test and task development and scoring rule improvement. However, such methods may not retain all information in process data, and the relationship between obtained indicators and measured psychological traits is not clear. This paper categorizes bottom-up process data feature extraction methods for problem solving into three major classes based on source domain and processing philosophy. The above introduction reveals that these three classes each have limitations in information utilization. For example, NLP-based indicator construction methods depend on original encoding and are mostly too general with large information loss. Edit distance and LCS-based methods only apply to tasks with optimal solutions, and N-Gram methods only apply to tasks with few executable operations. Dimensionality reduction algorithms preserve entire response sequence information and can be used for predictive analysis, with some studies proposing ability estimation models using such process information (Zhang et al., 2020), but extracted features lack interpretability. Finally, network indicator methods can visualize response processes and explore response pattern characteristics across groups but struggle to capture specific operation information, and extracted features cannot be directly used for respondent ability estimation. Therefore, data-driven feature extraction methods may also face information omission problems and have interpretability issues. Very few studies use such features for ability estimation, so purely data-driven feature extraction methods have not been directly applied

to large-scale standardized test ability assessment. Characteristics of various feature extraction methods are summarized in Table 1 .

Table 1 Summary of Feature Extraction Methods for Process Data in Computer-Based Problem-Solving Tests

Method	Applicable Task Types	Information Utilization	Advantages	Disadvantages
Top-Down				
Expert-defined scoring/indicator construction rules	All task types	Construct indicator extraction and scoring rules for ability estimation	Strong theoretical basis, high interpretability, suitable for traditional measurement model analysis	High workload, task specificity, may omit information
Bottom-Up				
N-Gram	Tasks with few executable operations	Identify key operation sequences; construct behavioral indicators; obtain response sequence feature vectors	Simple computation, can construct behavioral indicators through expert definition	Loses sequential information; high dimensionality; depends on encoding
Edit distance	Tasks with optimal solution paths	Construct an indicator reflecting performance level	Simple computation, clear indicator meaning	Depends on encoding; high information loss; only for tasks with optimal paths

Method	Applicable Task Types	Information Utilization	Advantages	Disadvantages
LCS-based indicators	Tasks with optimal solution paths	Construct an indicator reflecting performance level	Simple computation, clear indicator meaning	Depends on encoding; high information loss; only for tasks with optimal paths
Dimensionality reduction	All task types	Represent response sequences with numerical feature vectors to extract all information from response sequences	Comprehensive information extraction	Lacks interpretability
Network analysis (SNA)	All task types	Visualize response processes; extract response process network features	Intuitive process visualization; can identify key operation sequences	Complex preprocessing; loses sequence and node content information; cannot directly estimate ability

4. Ability Assessment Models for Process Data

After extracting behavioral indicators/features from process data, probabilistic relationship models between these indicators and latent abilities must be constructed to enable ability estimation. Based on whether models utilize sequential relationships between indicators and whether they can obtain continuous, interpretable latent ability estimates, current methods for estimating latent abilities using process information can be divided into three categories: traditional psychometric models and their extensions, stochastic process models, and measurement models combining stochastic process ideas.

4.1 Traditional Psychometric Models and Their Extensions

Behavioral indicators obtained through expert definition directly correspond to construct elements in the test conceptual framework and can be analogized to items in traditional tests for fitting measurement models. For multidimensional

test structures, multidimensional IRT models and diagnostic classification models can be used to simultaneously estimate abilities across multiple dimensions or diagnose mastery of multiple skills (e.g., Hesse et al., 2015; Siddiq et al., 2017; Yuan et al., 2019; Zhan & Qiao, 2020). If tests are administered in group formats, multilevel models can also be fitted (Wilson et al., 2017). In addition to directly applying existing psychometric models, some researchers have extended traditional measurement models or their estimation steps based on process data characteristics (Li Meijuan et al., 2020; Liu et al., 2018; Zhang et al., 2020).

4.1.1 Multidimensional IRT Models When behavioral indicators extracted from process data correspond to multiple elements/sub-dimensions in the problem-solving operational conceptual framework (Hesse et al., 2015; OECD, 2013; Rosen, 2017), multidimensional IRT models can be used to estimate respondents' performance levels across multiple sub-dimensions. For example, some studies used the Multidimensional Random Coefficients Multinomial Logit Model (MRCMLM; Adams et al., 1997) to analyze behavioral indicators from ATC21S collaborative problem-solving tests, obtaining ability estimates for respondent groups across multiple dimensions and finding that multidimensional IRT models fit better than unidimensional IRT models estimating dimensions separately (Hesse et al., 2015; Siddiq et al., 2017). Indicator multidimensionality can correspond not only to multiple sub-dimensions of target abilities but also to different members within collaborative problem-solving groups. Yuan et al. (2019), when analyzing a collaborative problem-solving test using a "human-to-human interaction" mode with two-person groups as test units, distinguished behavioral indicators by implementer into individual and group-shared indicators and analyzed them using within-item multidimensional MRCML models, enabling estimation of individual performance and within-group influence strength.

4.1.2 Multilevel (Multidimensional) IRT Models Problem-solving test process data have nested structures, with process indicators nested within respondents and, in some collaborative problem-solving tests, respondents nested within groups, making them suitable for multilevel analysis. Wilson et al. (2017) added a group level to the two-level Rasch model (Kamata & Cheong, 2007; Raudenbush et al., 2003), constructing a three-level Rasch model with process indicators at level 1, respondents at level 2, and collaborative groups at level 3. They analyzed ATC21S collaborative problem-solving test data under the "Learning in Digital Networks-ICT" theme using unidimensional and multidimensional Rasch models and multilevel unidimensional and multidimensional Rasch models.

4.1.3 Diagnostic Classification Models Diagnostic Classification Models (DCMs) are a class of restricted or confirmatory latent class psychometric models that model relationships between several fine-grained discrete latent attributes and observed item responses (von Davier & Lee, 2019). Zhan and Qiao (2020)

proposed a method integrating diagnostic classification into process data analysis: treating adjacent short operation sequences (N-Grams) in response sequences as process items and converting them to 0-1 coding based on occurrence; then using problem-solving skills required to generate these operation sequences as latent attributes to construct Q-matrices for process items; and finally analyzing them using higher-order DCMs. Using higher-order DCMs to analyze process data can assess respondents' continuous latent problem-solving abilities while classifying them based on problem-solving strategies. However, using N-Grams to construct dichotomously coded process indicators loses overall sequential order information and N-Gram frequency information from response sequences. Moreover, in more complex tasks, the number of process items constructed from N-Grams is enormous, making Q-matrix specification very costly.

These studies represent new applications of existing psychometric models to process indicator analysis without proposing model improvements themselves, and all require experts to clearly define relationships between behavioral indicators and measurement constructs.

4.1.4 Modified Multilevel Mixture IRT Model To explore different strategies respondents adopt during response processes while accounting for the nested nature of process data, Liu et al. (2018) extended the Multilevel Mixture Item Response Theory (MMixIRT; Cho & Cohen, 2010) model and proposed a modified MMixIRT (mMMixIRT) model suitable for processing process data. This method first enumerates all operations in a task and pre-determines the correctness of each operation. At the process level, cumulative information (scores) of all operations are treated as process data for specific steps. At the individual level, mMMixIRT can customize design matrix A to determine what information is used for individual-level ability estimation, making it more flexible than the MMixIRT model. mMMixIRT can not only analyze response strategy category characteristics at the process level but also simultaneously estimate ability values at both process and individual levels. To avoid the problem that the assumption of normally distributed abilities within each latent class in mMMixIRT may be difficult to satisfy, Li Meijuan et al. (2020) further modified the mMMixIRT model by only distinguishing strategy categories at the process level without estimating process abilities. This exhaustive scoring approach enables mMMixIRT to utilize respondents' response data at each solution step, but this special coding method also has task specificity issues. Moreover, mMMixIRT's estimation of respondent-level ability is based only on responses at the final step, thus not incorporating sequential information from the process.

4.1.5 Two-Step Conditional Expectation Method To incorporate process information when estimating latent traits to improve estimation accuracy, Zhang et al. (2020) proposed a two-step conditional expectation method. The implementation steps are shown in Figure 4 [Figure 4: see original paper]. First, the item set is split into subsets. Let represent respondent i's response process

vector, outcome vector, and latent ability estimate (based on IRT model) from outcome vector for item subset , respectively. Process vectors can be extracted using methods such as autoencoders and MDS described previously. The construction process for a new ability estimate that integrates both outcome responses and response processes on item subset is as follows:

Step 1: Regress on to obtain .

Step 2: Regress on to obtain .

Similarly, can be obtained. Zhang et al. (2020) used MDS (Tang, Wang, et al., 2021) as the process feature extraction method and applied the two-step conditional expectation method to data from 14 PSTRE items in PIAAC 2012. Results showed that compared to estimates based solely on outcome responses, process-based latent trait estimates had higher consistency with performance on similar tasks, and fewer items were needed to achieve the same reliability level. However, this method directly uses dimensionality reduction algorithms to extract process vectors, thus sharing the interpretability issues of information utilization.

4.2 Stochastic Process Models

In problem-solving tests or similar platforms, respondents' task-solving steps can be viewed as continuous response processes along discrete time points, where response sequences are interdependent (Bellman, 1957; Puterman, 1994). Therefore, probabilistic models describing stochastic processes can be used to fit sequentially dependent process indicators and obtain latent state levels at each time point—possibly corresponding to respondents' time-varying knowledge mastery states or ability performance levels. Commonly used stochastic process analysis methods include Hidden Markov Models (HMM) and Dynamic Bayesian Networks (DBN).

4.2.1 Hidden Markov Model HMM is a probabilistic model for sequential data that describes a process where a hidden Markov chain randomly generates an unobservable state sequence, and each state then generates an observation to produce an observable sequence (Li Hang, 2011). HMM has been used to analyze process data from adaptive peer tutoring systems and adaptive tests (Arieli-Attali et al., 2019; Bergner et al., 2017). HMM can also fit observation sequences from respondents in problem-solving tests or similar systems and obtain latent state levels at various time points. Xiao et al. (2021) used HMM to analyze action sequences from two problem-solving items in PIAAC 2012, identifying latent states and state transitions. Results showed that across both items, respondents who answered correctly were more task-focused and more frequently used effective tools to solve problems, while those who answered incorrectly were more likely to use shorter action sequences and exhibit hesitant behaviors. This demonstrates that data-driven HMM methods can help researchers better understand behavioral patterns and cognitive transitions underlying respondents'

action sequences in complex problem-solving tasks.

4.2.2 Dynamic Bayesian Network Dynamic Bayesian Network (DBN) is an extension of original Bayesian networks for modeling state transitions containing temporal information and can be used to model respondents' stochastic response processes (Käser et al., 2017; Reichenberg, 2018; Reye, 2004; Rowe & Lester, 2010). Figure 5 [Figure 5: see original paper] shows a simple DBN path diagram with three time points. DBN has two basic components: one part consists of circles and rectangles corresponding to latent abilities and observed variables; the other part consists of paths (arrows) representing variable dependencies that change over time (Levy & Mislevy, 2016). As can be seen, HMM is a special case of DBN, with DBN adding paths from to latent ability at time . DBN has been applied to assessment and learning analytics: Reye (2004) demonstrated how to use the DBN framework to analyze longitudinal data, paving the way for its application in analyzing respondent learning or ability changes in intelligent tutoring systems (Reye, 2004; VanLehn, 2008) and game-based assessments (Iseli et al., 2019).

Levy (2019) combined DBN, cognitive diagnostic modeling, and process data analysis methods to analyze data from Save Patch, an educational game for rational number addition (Chung et al., 2010). The Save Patch game includes 23 progressively difficult levels, each with several types of observable responses, with each response type specified to correspond to several latent skills. Levy (2019) used DBN to analyze observation sequences and obtained results on each respondent's mastery level of each latent skill or degree of misconception holding throughout the entire game process for each attempt.

DBN can utilize information from different response sequence patterns and maintain the sequential structure of response sequences. It uses latent states to model different latent traits and skills, thereby enabling cognitive diagnosis. Both HMM and DBN yield discrete knowledge mastery states or ability states that change over the process. However, unlike intelligent tutoring tests, in psychological testing researchers generally want to obtain stable, continuous ability estimates. These conditions limit the application of DBN for assessing respondents' latent abilities using response process data in modern assessment contexts.

4.3 Measurement Models Combining Stochastic Process Ideas

Respondents' response processes in problem-solving tests are partially under their control—that is, respondents decide what steps to take in particular states. Therefore, conditional on latent ability, each respondent's response process can be viewed as a discrete-time stochastic process with conditional first-order Markov properties (Shu et al., 2017). To retain sequential relationships between process indicators in modeling while obtaining continuous latent ability estimates, researchers have proposed measurement models combining stochastic process ideas.

4.3.1 Markov IRT Model Shu et al. (2017) proposed a Markov-IRT model that conditions on latent ability and uses operation transitions (i.e., two adjacent operations in response sequences) as observed variables. To retain frequency information about operation transitions, they proposed both polytomously scored and dichotomously scored operation transitions. Under the dichotomous scoring framework, letting represent operation transition, scored 1 when correct and 0 when incorrect, the probability that respondent selects operation transition can be expressed as:

$$P(a_{ijj} = 1|\theta_i) = \frac{\exp(\beta_{jj} + \alpha_{jj}\theta_i)}{1 + \exp(\beta_{jj} + \alpha_{jj}\theta_i)}$$

where represents the tendency to select operation transition, and links the transition to latent trait . As can be seen, this formula has the form of a two-parameter IRT (2PL-IRT) model. To include low-frequency operation transitions while ensuring estimation accuracy, Shu et al. (2017) also proposed a higher-order Markov-IRT model that groups operation transitions to reduce the impact of data sparsity caused by extremely low frequencies of some transitions. When using the Markov-IRT model for analysis, all possible operation transitions are used to construct indicators, and the repetition count of each operation transition is considered in scoring, fully utilizing information carried by the operation and transition space. However, this method's analysis object is the scored operation transition frequency matrix, which does not retain the sequential order of operation transitions. Moreover, directly using operation transitions to represent response processes has limitations—for example, in some tasks, the same operation transition in different problem states may lead to completely opposite results.

4.3.2 Continuous-Time Dynamic Choice Model To simultaneously consider event history and occurrence time in analysis, Chen (2020) treated respondents' response processes as marked point processes and proposed a parametric method for marked point processes—the continuous-time dynamic choice (CTDC) model. In the CTDC model, the choice of event type at the next moment is modeled using conditional density functions that depend on respondents' latent problem-solving ability and task difficulty , with a multinomial logit form:

$$f_j(j|t, \mathcal{F}_{jt}, \theta, \beta_j) = \frac{\exp(\beta_j + \theta V_{jj}(\mathcal{F}_{jk}))}{\sum_{i \in S_j(\mathcal{F}_{jk})} \exp(\beta_j + \theta V_{ji}(\mathcal{F}_{jk}))}$$

where represents the set of immediately possible event types at time , and is an effectiveness measure for event type and task characteristics , with 1 for effective and 0 for ineffective. The timestamp of the next operation is modeled using a ground intensity function that depends on respondents' behavioral speed trait and has an exponential form:

$$\lambda_j(t|\mathcal{F}_{jt}, t, \gamma_j) = \exp(\gamma_j + \tau)$$

The two latent traits—problem-solving ability and behavioral speed —follow a bivariate normal distribution. The CTDC model can estimate each respondent's problem-solving ability and operation speed based on process data from one or multiple tasks through specification of event history information. However, although this model incorporates time information, it actually models latent ability and response speed separately, only assuming they follow a multivariate normal distribution. Moreover, this method's analysis of task characteristics and response processes is not deep enough—each task has only one difficulty parameter and cannot distinguish unique attributes of each event type in the response process.

4.3.3 Markov Decision Process Measurement Model Markov Decision Process (MDP) is an uncertainty decision model based on longitudinal cost-benefit analysis (Puterman, 1994) that includes four elements: goals, motivation, task understanding (beliefs), and problem-solving ability. LaMar (2018) explored methods for using MDP as a measurement model to infer individual characteristics from actions and problem states recorded in process data in complex decision-making problem tasks, proposing the Markov Decision Process Measurement Model (MDP-MM). For a task with state set , MDP-MM describes the conditional probability of taking action in state (LaMar, 2018):

$$p(a|s, \beta_j) = \frac{\exp(\beta_j Q(s, a|\beta_j))}{\sum_{a' \in A} \exp(\beta_j Q(s, a'|\beta_j))}$$

where is analogous to latent ability in IRT and follows a log-normal distribution. is a recursive function representing the value of an action, containing both immediate rewards (scores) for the current action and expected scores for subsequent steps. Simulation studies show that MDP-MM can clearly separate datasets generated under "high ability-low motivation" conditions from those generated under "low ability-high motivation" conditions. LaMar (2018) also used MDP-MM to analyze actual data from a microbial game, finding that ability estimates had significant positive correlations with post-test scores. However, MDP-MM has many restrictions. When used, reasonable reward parameterizations must be defined for various operations and/or outcomes based on specific tasks. If reward parameters are freely estimated, reward values may have opposite directions from the construct, making unable to represent respondent ability.

4.3.4 Sequential Response Model To fully utilize problem-solving test process data for estimating respondents' latent ability levels, Han et al. (2021) proposed using problem state sequences to represent complete response process information for well-structured problem scenarios and developed a Sequential Response Model (SRM) that can analyze entire problem state sequences. SRM

assumes that respondents' choice of next state is related to their latent ability and current state . The model has a multinomial logit form:

$$P(S_{i,t+1} = x_j | S_{i,t} = x_j, \theta_i, \lambda, \mathcal{R}) = \frac{\exp(\lambda_{x_j,x_j} + I_{x_j,x_j} \cdot \theta_i)}{\sum_{x_h \in M_{x_j}} \exp(\lambda_{x_j,x_h} + I_{x_j,x_h} \cdot \theta_i)}$$

where is the state transition parameter representing the tendency to transition from state to state ; is an indicator function that equals 1 when the state transition is correct and -1 otherwise; represents the set of all possible states at the next moment when the current state is ; and are task-specific preset rules. Han et al. (2021) validated the feasibility and reasonableness of SRM for estimating respondent latent abilities and item state transition parameters using process data from the PISA 2012 problem-solving test "Ticket" task. SRM can effectively analyze complete response sequences, yielding item characteristic parameters (state transition parameters) that provide useful information for understanding task features and respondent ability estimates with interpretability that helps understand ability levels of different response patterns. However, reasonable application of SRM requires well-defined state sequences; definition methods for problem states and state transitions in ill-structured problems still need further exploration.

4.3.5 Summary of Measurement Models Combining Stochastic Process Ideas Except for MDP-MM, such models mainly apply to simple test scenarios with limited operation sets, requiring exhaustive enumeration of all behaviors in tasks and prior judgment of each behavior's correctness (or effectiveness) by experts, while MDP-MM requires prior definition of reward parameters before recursively calculating action values. Operation transitions in Markov-IRT, event types in CTDC, actions in MDP-MM, and state transitions in SRM are all different representations of behaviors. Judgments of behavioral correctness (or effectiveness) are reflected in scoring for Markov-IRT and represented by coefficients in multinomial logit models for the other three models: in CTDC and in SRM. Their differences lie in: Markov-IRT only preserves order between adjacent operations, while the other three models use state representation that contains (partial) historical behavior information; only CTDC utilizes response time among these models, but CTDC can only obtain overall task difficulty parameters, while Markov-IRT and SRM can obtain tendency parameters for each behavior.

4.4 Overall Evaluation of Current Process Data Ability Assessment Models

In summary, to estimate latent ability levels from observed indicators using ability assessment models, reasonable construction of the correspondence between indicators and latent abilities is essential. As described in Section 3, this process currently still requires expert experience (either before or after analysis).

The interpretability of different assessment models depends on the strength of assumptions between their utilized indicators and latent structures. Psychometric models focus primarily on latent ability estimation. In addition to direct applications of traditional measurement models, some researchers have proposed improvements to existing models or estimation steps. Process indicators used in such models generally have relatively strong correspondence with latent abilities, yielding highly interpretable results (except for the two-step conditional expectation method). However, they are limited by local independence assumptions and do not contain sequential information between indicators in analysis. Stochastic process models focus on modeling response processes, preserving response path information, but have weaker assumptions between indicators and latent structures. Sometimes they first use data-driven models to obtain latent state levels before theoretical interpretation, and they do not focus on stable, continuous latent ability estimates. In educational and psychological testing, the primary purpose is to obtain valid estimates of respondents' latent traits. From this perspective, stochastic process models struggle to meet the need for valid estimation of stable, continuous ability traits in educational and psychological testing. Finally, measurement models combining stochastic process ideas have advantages of both: their analysis objects are action sequences in tasks, they can preserve action sequential order, and they have experts specify indicator coefficients or scoring methods aligned with ability direction, giving them some interpretability. Thus they can utilize relatively complete response process information to obtain continuous latent ability estimates. However, such models require exhaustive enumeration of all actions in tasks and are mostly suitable for simple tasks with limited operation sets. Therefore, there remains room for further research on how to fully utilize response process information to more accurately assess respondents' latent abilities while maintaining scientific rigor and interpretability of results. Applicable scenarios, advantages, disadvantages, and actual datasets and analytical software tools used in research for each model are summarized in Table 2 .

Table 2 Summary of Ability Assessment Models for Process Data in Computer-Based Problem-Solving Tests

Model Category	Specific Model	Applicable Scenarios	Process Indicator Requirements	Advantages	Disadvantages	Empirical Datasets	Software Tools
Psychometric Models	Multidimensional IRT (Hesse et al., 2015; Siddiq et al., 2017; Yuan et al., 2019)	Multidimensional test structure	Need to predefine relationships between indicators and dimensions	Theoretically grounded; estimated latent abilities have clear psychological meaning	Limited by indicator definition, may omit information; cannot analyze behavioral sequences	ATC21S collaborative problem-solving test	ConQuest 2.0 (Wu et al., 2007)
	Multilevel IRT (Wilson et al., 2017)	Group collaborative tests	Need to predefine relationships between indicators and constructs	Can simultaneously estimate individual and group abilities; can classify strategies at process level	Task-specific unique coding; individual ability estimates only use final step information	ATC21S-ICT test	Mplus (Muthén & Muthén, 1998-2015)

Model Category	Specific Model	Applicable Scenarios	Process Indicator Requirements	Advantages	Disadvantages	Empirical Datasets	Software Tools
	Cognitive Diagnosis Models (Zhan & Qiao, 2020)	Path-clear, enumerable tasks	Q-matrix specified indicators	Can assess continuous latent ability while diagnosing strategies	Loses overall sequence and frequency information; high Q-matrix specification cost	PISA 2012 "Ticket" item CP038Q01	GDINA R package (Ma & de la Torre, 2020)
	Modified Multilevel Mixture IRT (Liu et al., 2018; Liu et al., 2020)	Simple tasks with limited operation sets	Predetermine correctness of each optional operation; use cumulative coding	Comprehensive information utilization; can estimate abilities at both levels and classify strategies	Task-specific coding; individual ability only uses final step	PISA 2012 "Ticket" items CP038Q01, CP038Q02	TAM R package (Robitzsch et al., 2020)

Model Category	Specific Model	Applicable Scenarios	Process Indicator Requirements	Advantages	Disadvantages	Empirical Datasets	Software Tools
	Two-step Conditional Expectation (Zhang et al., 2020)	No special requirements	Process information feature vectors	Incorporates process information in latent trait estimation	Interpretability issues with utilized process information	PIAAC 2012 PSTRE test	glmnet R package (Friedman et al., 2009); Procdata R package (Tang et al., 2021)
Stochastic Process Models	Hidden Markov Model (Bergner et al., 2017; Xiao et al., 2021)	Latent states change over process	Temporally continuous indicators	Preserves sequential structure; uses latent states to model traits for cognitive diagnosis	Cannot obtain stable continuous ability estimates like psychometric models	Adaptive peer tutoring system; PIAAC 2012 PSTRE test	depmixS4 R package (Visser & Speekenbrink, 2010)

Model Category	Specific Model	Applicable Scenarios	Process Indicator Requirements	Advantages	Disadvantages	Empirical Datasets	Software Tools
	Dynamic Bayesian Network (Levy, 2019)	Temporally continuous indicators with clear indicator-trait relationships	Clear correspondence between indicators and latent traits	Preserves sequential structure; can model multiple skills for cognitive diagnosis	Cannot obtain stable continuous ability estimates	Save Patch educational game	Bayes Net Matlab toolbox (Murphy, 2001); gRain R package (Højsgaard, 2012)
Hybrid Models	Markov IRT (Shu et al., 2017)	Simple tasks with limited operation sets; operation transition correctness constant throughout	Process indicators are operation transitions; need predetermine correctness and score	Considers both correct/incorrect operations and frequencies; comprehensive information	Only preserves adjacent operation order; limited applicability	NAEP-TEL PumpRepair task	MIRT software (Haberman, 2013)

Model Category	Specific Model	Applicable Scenarios	Process Indicator Requirements	Advantages	Disadvantages	Empirical Datasets	Software Tools
	Continuous-time Dynamic Choice (Chen, 2020)	Simple tasks with limited events	Pre-determine effectiveness of each event; obtain timestamps	Can estimate problem-solving ability and operation speed from process data	Only one difficulty parameter per task; cannot distinguish unique event attributes	PISA 2012 problem-solving test	nnet R package (Venables & Ripley, 2002)
	MDP Measurement Model (LaMar, 2018)	Well-structured tasks with clear state and action sets	Pre-define reasonable reward parameters for actions/outcomes	Uses reinforcement learning to estimate ability considering multi-step information	Many parameters; free estimation may yield unreasonable values	Public educational game Microbes	C++ custom program

Model Category	Specific Model	Applicable Scenarios	Process Indicator Requirements	Advantages	Disadvantages	Empirical Datasets	Software Tools
	Sequential Response Model (Han et al., 2021)	Well-structured problems with optimal strategies	Predetermine correctness of each state transition	Can use complete sequences; obtains ability and transition parameters	Data preprocessing for ill-structured problems needs exploration	PISA 2012 "Ticket" item CP038Q02	R Bayesian estimation program

5. Issues and Prospects

To obtain valid ability estimates from computer-based problem-solving tests, scientifically and reasonably analyzing process data is essential. Process data analysis generally involves two parts: feature extraction and ability assessment model construction. This paper introduces the latest methodological research on these two aspects and summarizes the applicable scenarios, advantages, and disadvantages of each method. This can serve as a reference for methodological researchers to quickly grasp new developments in process data analysis methods for problem-solving tests, promote methodological innovation, and provide practical users with guidance on selecting appropriate methods for data analysis, offering direction for future research. Current research on how to extract process data features and use process data to assess respondents' latent abilities is still in its initial stages. Based on the above summary, several aspects can be improved.

5.1 Interpretability Issues in Process Data Analysis

Ensuring psychological-level interpretability at all stages of process data analysis is a noteworthy topic important for ensuring test result fairness, validity, and objectivity. When extracting features from process data, bottom-up approaches can directly obtain digital representations of response sequences or key features, but the association mechanisms between these indicators and target psychological variables are relatively difficult to explain and understand. When

modeling process indicators, estimated latent ability levels should match the measured latent construct levels. Researchers analyzing process data should follow ECD theory's "evidence-based reasoning" concept, combine psychological theory when extracting evidence, focus on the psychological meaning of evidence indicators, and try to use highly interpretable algorithms for modeling. Moreover, to deeply explore the cognitive processing of problem solving using process data, test developers, domain experts, and psychometric experts still need to jointly participate in decision-making. For distinguishing and interpreting error strategies, bottom-up approaches can first extract features containing error information, followed by cluster analysis. Different feature combinations may reflect different strategy types, but clustering results still require expert interpretation.

5.2 Feature Extraction Should Incorporate More Information

While ensuring interpretability of extracted features, as much valuable information as possible should be extracted from process data. Current utilization of process data is mostly based on behavioral performance information, with only a few studies using time or language information recorded in process data (Chen, 2020; Yuan Jianlin, 2019). Future research should consider how to incorporate these multimodal information sources beyond behavioral performance into measurement models for more accurate ability estimation. Additionally, for application in large-scale standardized tests, regardless of information extraction method, automatic extraction and scoring of indicators should be achievable. Automatic indicator extraction and reasonable scoring for multimodal data also present certain challenges.

5.3 Enabling Ability Assessment in More Complex Problem Scenarios

Current stochastic process models and measurement models combining stochastic process ideas all assume that, conditional on respondents' latent abilities, response processes have (conditional) first-order Markov properties. This holds in simple test scenarios but may be violated in complex, highly interactive dynamic problem scenarios. As shown in the "Empirical Datasets" column of Table 2, currently available empirical datasets for researchers are not rich, mostly concentrated in three large-scale test projects: PISA, PIAAC, and ATC21S. Notably, the PISA problem-solving test "Ticket" item is frequently used, mainly because this item type has a simple structure. This also reflects current models' limitations in analyzing complex tasks. Therefore, while proposing demands for developing more complex tests, methodological researchers should also provide corresponding data analysis and processing methods. Additionally, covariates may affect respondent performance in process-based tests. For example, research shows that factors such as problem-solving persistence and openness significantly affect students' performance on problem-solving tests in digital environments (Yuan Jianlin et al., 2016). Future research could also consider constructing assessment models for process data that include covariates to further

improve ability estimation accuracy.

5.4 Moving from Theoretical Research to Practical Application

Theoretical research on process data analysis methods needs practical testing of its actual effectiveness. On one hand, regardless of how complex the measurement models or data mining techniques used in analytical methods are, they should ultimately serve practical purposes. As shown in the last column of Table 2, most existing assessment models have corresponding parameter estimation software or packages, but for new models developed specifically for process data, custom programs may be needed for parameter estimation, creating high barriers to model application. Therefore, developers of new models should be encouraged to publicly release parameter estimation code or develop user-friendly software packages to facilitate model use and dissemination. On the other hand, to facilitate practical users, test developers could also consider developing user-friendly problem-solving test systems based on existing analytical methods that can realize automatic scoring of process data and immediate generation of ability assessment results and knowledge/skill diagnostic reports.

5.5 Integration and Cross-Fertilization of Analytical Methods Across Fields

This paper focuses on reviewing feature extraction and ability assessment research for problem-solving tests. Currently, feature extraction methods and ability assessment models are not perfectly matched. Most features extracted through data-driven approaches cannot be applied to ability assessment models because they have not established correspondence with latent abilities. However, the main purpose of psychological testing is to accurately measure respondents' latent abilities. Researchers should develop more feature extraction methods that can be applied to ability assessment models. In addition to problem-solving ability, measurement of many other higher-order abilities has also been computerized, such as critical thinking (Liu et al., 2016; Song & Sparks, 2019), creative thinking, and disciplinary literacy. Adaptive learning and tutoring systems also often include ability assessment. Process data analysis for problem-solving tests is one of the most studied test types. Because problem-solving tests focus more on ability evaluation, research on measurement model construction is also relatively rich, while other test types have more limited research on ability assessment model innovation. On one hand, the analytical 思路 for problem-solving test process data is referential for other test types—for example, the general process of defining indicators through expert systems is similar. On the other hand, each test type has its particularities. For instance, problem-solving tests or disciplinary literacy tests focus more on accurate ability estimation, while some tests focus more on response processes (e.g., critical thinking tests focus more on argumentation processes). Therefore, when borrowing analytical methods across fields, specific circumstances should be considered.

References

李航. (2012). 统计学习方法. 北京: 清华大学出版社.
Li Hang. (2012). *Statistical Learning Methods*. Beijing: Tsinghua University Press.

李美娟. (2020). 基于过程数据的合作问题解决评分和测量模型研究 (博士学位论文). 北京师范大学.
Li Meijuan. (2020). *Research on Scoring and Measurement Models for Collaborative Problem Solving Based on Process Data* (Doctoral dissertation). Beijing Normal University.

李美娟, 刘玥, 刘红云. (2020). 计算机动态测验中问题解决过程策略的分析: 多水平混合 IRT 模型的拓展与应用. 心理学报, 52(4), 528–540.
Li Meijuan, Liu Yue, & Liu Hongyun. (2020). Analysis of problem-solving process strategies in computer-based dynamic testing: Extension and application of multilevel mixture IRT models. *Acta Psychologica Sinica*, *52*(4), 528–540.

陆璟. (2017). 基于 log 数据的国际学生评估项目 (PISA) 问题解决能力研究 (博士学位论文). 华东师范大学, 上海.
Lu Jing. (2017). *Research on PISA Problem-Solving Competence Based on Log Data* (Doctoral dissertation). East China Normal University, Shanghai.

骆文淑, 赵守盈. (2005). 多维尺度法及其在心理学领域中的应用. 中国考试, (4), 27–30.
Luo Wenshu & Zhao Shouying. (2005). Multidimensional scaling and its application in psychology. *China Examinations*, *(4)*, 27–30.

王永锋, 王以宁, 何克抗. (2007). 从 "学习使用技术" 到 "使用技术学习"——解读新版美国 "国家学生教育技术标准". 电化教育研究, (12), 82–85.
Wang Yongfeng, Wang Yining, & He KeKang. (2007). From "learning to use technology" to "using technology to learn": Interpreting the new U.S. National Educational Technology Standards for Students. *e-Education Research*, *(12)*, 82–85.

徐伟, 陈光辉, 曾玉, 张文新. (2011). 关系研究的新取向: 社会网络分析. 心理科学, 34(2), 499–504.
Xu Wei, Chen Guanghui, Zeng Yu, & Zhang Wenxin. (2011). A new approach to relationship research: Social network analysis. *Journal of Psychological Science*, *34*(2), 499–504.

袁建林. (2018). 基于行为过程表现测量合作问题解决能力的研究 (博士学位论文). 北京师范大学.
Yuan Jianlin. (2018). *Measuring Collaborative Problem-Solving Competence Based on Behavioral Process Performance* (Doctoral dissertation). Beijing Normal University.

袁建林, 刘红云, 张生. (2016). 数字化测验环境中学生问题解决能力影响因素分析——以 PISA 2012 为例. 中国电化教育, (8), 74–81.
Yuan Jianlin, Liu Hongyun, & Zhang Sheng. (2016). Analysis of factors influencing students' problem-solving ability in digital testing environments: The

case of PISA 2012. *China Educational Technology*, *(8)*, 74–81.

Adams, R., Vista, A., Scoular, C., Awwal, N., Griffin, P., & Care, E. (2015). Automatic coding procedures for collaborative problem solving. In P. Griffin & E. Care (Eds.), *Assessment and teaching of 21st century skills: Methods and approach* (pp. 115–132). Dordrecht: Springer.

Adams, R. J., Wilson, M., & Wang, W. C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, *21*(1), 1–23.

Amer, M. R., Siddiquie, B., Khan, S., Divakaran, A., & Sawhney, H. (2014). Multimodal fusion using dynamic hybrid models. In *IEEE winter conference on applications of computer vision* (pp. 556–563). New York, NY: IEEE.

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.

Bejar, I. I., Mislevy, R. J., & Zhang, M. (2016). Automated scoring with validity in mind. In A. A. Rupp & J. P. Leighton (Eds.), *The Wiley handbook of cognition and assessment* (pp. 226–246). Hoboken, NJ: Wiley-Blackwell.

Bellman, R. (1957). A markovian decision process. *Journal of Mathematics and Mechanics*, *6*(5), 679–684.

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.

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

Cho, S. J., & Cohen, A. S. (2010). A multilevel mixture IRT model with an application to DIF. *Journal of Educational and Behavioral Statistics*, *35*(3), 336–370.

Chung, G. K. W. K., Baker, E. L., Vendlinski, T. P., Buschang, R., Delacruz, G. C., Michiuye, J. K., & Bittick, S. J. (2010, April). Testing instructional design variations in a prototype math game. In R. Atkinson (Chair), *Current perspectives from three national R&D centers focused on game-based learning: Issues in learning, instruction, assessment, and game design.* Poster session presented at the Annual Meeting of the American Educational Research Association, Denver, CO.

Davis, J. A., & Leinhardt, S. (1972). The structure of positive interpersonal relations in small groups. In J. Berger (Ed.), *Sociological theories in progress* (Vol. 2, pp. 218–251). Boston, MA: Houghton Mifflin.

Friedman, J., Hastie, T., & Tibshirani, R. (2009). Glmnet: Lasso and elastic–net regularized generalized linear models [R package version]. Retrieved August 4, 2021, from https://cran.r-project.org/web/packages/glmnet/

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. Cambridge, MA: MIT Press.

Haberman, S. J. (2013). A general program for item-response analysis that employs the stabilized Newton–Raphson algorithm (No. ETS RR-13-32). Princeton, NJ: Educational Testing Service.

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.

Han, Y., Liu, H., & Ji, F. (2021). A sequential response model for analyzing process data on technology-based problem-solving tasks. *Multivariate Behavioral Research*. Advance online publication. https://doi.org/10.1080/00273171.2021.1932403

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.

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*, 104155.

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.

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

Højsgaard, S. (2012). Graphical independence networks with the gRain package for R. *Journal of Statistical Software*, *46*(10), 1–26.

Iseli, M. R., Koenig, A. D., Lee, J. J., & Wainess, R. (2010). *Automatic assessment of complex task performance in games and simulations* (CRESST Report 775). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Kamata, A., & Cheong, Y. F. (2007). Multilevel rasch models. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution rasch models: Extensions and applications* (pp. 217–232). New York, NY: Springer.

Käser, T., Klingler, S., Schwing, A. G., & Gross, M. (2017). Dynamic Bayesian networks for student modeling. *IEEE Transactions on Learning Technologies*, *10*(4), 450–462.

Khan, S., Cheng, H., & Kumar, R. (2013). A hierarchical behavior analysis approach for automated trainee performance evaluation in training ranges. In D.

D. Schmorrow & C. M. Fidopiastis (Eds.), *Foundations of augmented cognition: Proceedings of HCI international 2013* (pp. 60–69). Berlin: Springer.

Khan, S. M. (2017). Multimodal behavioral analytics in intelligent learning and assessment systems. In A. A. von Davier, M. Zhu, & P. C. Kyllonen (Eds.), *Innovative assessment of collaboration* (pp. 173–184). Cham: Springer.

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

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, *10*(8), 707–710.

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

Levy, R., & Mislevy, R. J. (2016). *Bayesian psychometric modeling.* Boca Raton, FL: Chapman & Hall/CRC Press.

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, O. L., Mao, L., Frankel, L., & Xu, J. (2016). Assessing critical thinking in higher education: The HEIghten approach and preliminary validity evidence. *Assessment & Evaluation in Higher Education*, *41*(5), 677–694.

Lunn, D., Spiegelhalter, D., Thomas, A., & Best, N. (2009). The BUGS project: Evolution, critique and future directions. *Statistics in Medicine*, *28*(25), 3049–3067.

Ma, W., & de la Torre, J. (2020). GDINA: An R package for cognitive diagnosis modeling. *Journal of Statistical Software*, *93*(14), 1–26.

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, NJ: Erlbaum.

Mislevy, R. J. (2019). Advances in measurement and cognition. *The ANNALS of the American Academy of Political and Social Science*, *683*(1), 164–182.

Mislevy, R. J., Steinberg, L. S., Almond, R. G., & Lukas, J. F. (2006). Concepts, terminology, and basic models of evidence-centered design. In D. M. Williamson, R. J. Mislevy, & I. I. Bejar (Eds.), *Automated scoring of complex tasks in computer-based testing* (pp. 15–48). Mahwah, NJ: Lawrence Erlbaum.

Morency, L. P., de Kok, I., & Gratch, J. (2010). A probabilistic multimodal approach for predicting listener backchannels. *Autonomous Agents and Multi-Agent Systems*, *20*(1), 70–84.

Murphy, K. P. (2001). The bayes net toolbox for matlab. *Computing science and statistics*, *33*(2), 1024–1034.

Muthén, L. K., & Muthén, B. O. (2011). *Mplus user's guide* (6th ed.). Los Angeles, CA: Muthén and Muthén.

Muthén, L. K., & Muthén, B. O. (2015). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén and Muthén.

Organisation for Economic Co-operation and Development. (2013). *PISA 2012 assessment and analytical framework: Mathematics, reading, science, problem solving and financial literacy*. Paris: OECD Publishing.

Organisation for Economic Co-operation and Development. (2017). *PISA 2015 assessment and analytical framework: Science, reading, mathematic, financial literacy and collaborative problem solving* (Rev. ed.). Paris: OECD Publishing.

Puterman, M. L. (1994). *Markov decision processes*. New York, NY: Wiley.

Raudenbush, S. W., Johnson, C., & Sampson, R. J. (2003). A multivariate, multilevel Rasch model with application to self–reported criminal behavior. *Sociological Methodology*, *33*(1), 169–211.

Red Hill Studios. (n.d.). *Lifeboat to mars*. Retrieved August 4, 2021, from http://www.pbskids.org/lifeboat

Reichenberg, R. (2018). Dynamic Bayesian networks in educational measurement: Reviewing and advancing the state of the field. *Applied Measurement in Education*, *31*(4), 335–350.

Reye, J. (2004). Student modelling based on belief networks. *International Journal of Artificial Intelligence in Education*, *14*(1), 63–96.

Robitzsch, A., Kiefer, T., & Wu, M. (2020). TAM: Test analysis modules [R package version 3.5-19]. Retrieved August 4, 2021, from http://CRAN.R-project.org/package=TAM

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

Rowe, J. P., & Lester, J. C. (2010). Modeling user knowledge with dynamic Bayesian networks in interactive narrative environments. In G. M. Youngblood & V. Bulitko (Eds.), *Proceedings of the sixth AAAI conference on artificial intelligence and interactive digital entertainment* (pp. 57–62). Menlo Park, CA: AAAI Press.

Schleicher, A. (2008). Piaac: A new strategy for assessing adult competencies. *International Review of Education*, *54*(5), 627–650.

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.

Siddiq, F., Gochyyev, P., & Wilson, M. (2017). Learning in digital networks – ICT literacy: A novel assessment of students' 21st century skills. *Computers &*

Education, *109*, 11–37.

Siddiquie, B., Khan, S., Divakaran, A., & Sawhney, H. (2013). Affect analysis in natural human interaction using joint hidden conditional random fields. In *Proceedings of the 2013 IEEE international conference on multimedia and expo (ICME)* (pp. 1–6). New York, NY: IEEE.

Song, Y., & Sparks, J. R. (2019). Measuring argumentation skills through a game-enhanced scenario-based assessment. *Journal of Educational Computing Research*, *56*(8), 1324–1344.

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.

Tang, X., Zhang, S., Wang, Z., Liu, J., & Ying, Z. (2021). Procdata: An R package for process data analysis. *Psychometrika*, *86*(4), 1058–1083.

Technology and Engineering Literacy. (2013). *Technology and engineering literacy assessments*. Retrieved August 4, 2021, from https://nces.ed.gov/nationsreportcard/tel/

VanLehn, K. (2008). Intelligent tutoring systems for continuous, embedded assessment. In C. Dwyer (Ed.), *The future of assessment: Shaping teaching and learning* (pp. 113–138). Mahwah, NJ: Erlbaum.

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S-Plus* (4th ed.). New York, NY: Springer.

Visser, I., & Speekenbrink, M. (2010). depmixS4: An R package for hidden Markov models. *Journal of Statistical Software*, *36*(7), 1–21.

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.

Von Davier, A. A. (2017). Computational psychometrics in support of collaborative educational assessments. *Journal of Educational Measurement*, *54*(1), 3–11.

von Davier, M., & Lee, Y. S. (2019). *Handbook of diagnostic classification models: Models and model extensions, applications, software packages*. New York, NY: Springer.

Walker, E., Rummel, N., & Koedinger, K. R. (2009). CTRL: A research framework for providing adaptive collaborative learning support. *User Modeling and User-Adapted Interaction*, *19*(5), 387–431.

Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications*. Cambridge: Cambridge University Press.

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

Wu, M., Adams, R. J., Wilson, M., & Haldane, S. A. (2007). *ConQuest: Generalised item response modelling software* (version 2.0). Camberwell: ACER Press.

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.

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., & Qiao, X. (2020). A diagnostic classification analysis of problem-solving competence using process data. *PsyArXiv*. Retrieved August 4, 2021, from https://doi.org/10.31234/osf.io/wtyae

Zhang, S., Wang, Z., Qi, J., Liu, J., & Ying, Z. (2020). Accurate assessment via process data. Retrieved August 4, 2021, from http://www.columbia.edu/~zw2393/publication/process_{{{data

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.

Zoanetti, N. (2010). Interactive computer based assessment tasks: How problem-solving process data can inform instruction. *Australasian Journal of Educational Technology*, *26*(5), 585–606.

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