Original Article

External Correlates of Adult Digital Problem-Solving Process

An Empirical Analysis of PIAAC PSTRE Action Sequences

Susu Zhang[1], Xueying Tang[2], Qiwei He[3], Jingchen Liu[4], and Zhiliang Ying[4]

[1]Department of Psychology and Statistics, University of Illinois Urbana-Champaign, IL, USA
[2]Department of Mathematics, University of Arizona, AZ, USA
[3]Interdisciplinary Program of Data Science and Analytics, Georgetown University, Washington DC, USA
[4]Department of Statistics, Columbia University, NY, USA

Abstract: Computerized assessments and interactive simulation tasks are increasingly popular and afford the collection of process data, i.e., an examinee's sequence of actions (e.g., clickstreams, keystrokes) that arises from interactions with each task. Action sequence data contain rich information on the problem-solving process but are in a nonstandard, variable-length discrete sequence format. Two methods that directly extract features from the raw action sequences, namely multidimensional scaling and sequence-to-sequence autoencoders, produce multidimensional numerical features that summarize original sequence information. This study explores the utility of action sequence features in understanding how problem-solving behavior relates to cognitive proficiencies and demographic characteristics. This is empirically illustrated with the process data from the 2012 PIAAC PSTRE digital assessment. Regularized regression results showed that action sequence features are more predictive of examinees' demographic and cognitive characteristics compared to final outcomes. Partial least squares analysis further aided the identification of behavioral patterns systematically associated with demographic/cognitive characteristics.

Keywords: process data, sequence analysis, computerized assessment, multidimensional scaling, autoencoder

Assessment of examinee proficiency using computerized simulation tasks is gaining increasing relevance in both large-scale assessments, such as the Programme for the International Assessment of Adult Competencies (PIAAC; e.g., OECD, 2012) and the Programme for International Student Assessment (PISA; e.g., OECD, 2014) surveys, and in high-stakes testing, such as the US medical licensure exam (e.g., Dillon et al., 2004). Simulation tasks are typically interactive and resemble real-life situations, requiring examinees to demonstrate the ability or skills to perform tasks that are often complex. This also introduces new measurement opportunities for the collection of process data that arise from an examinee's interaction with each task/item. Process data are commonly logged by the computer as a time-stamped sequence of actions, such as clickstreams and keystrokes, performed by an examinee in pursuit of solving an item. In carefully engineered simulation tasks, computer-logged action sequences, which explicitly document test-taking behavior, may reveal information about the examinee's response process. This affords analysis of the test-taking process at a larger scale compared to traditional think-aloud cognitive interviews, which typically involve a smaller number of examinees concurrently or retrospectively describing how they arrived at their answers (e.g., Ericsson & Simon, 1998).

Process data, on top of final scores, offer a wealth of information about individual differences, test-taking engagement, and the steps examinees take to reach their final response. Studies have demonstrated the utility of process data for a multitude of practical tasks: To start, process data can provide additional information on the measured proficiency or skills, allowing better measurement via process-incorporated scoring rules (Zhang et al., 2023) and process-based measurement models, which typically associate continuous latent proficiency (Chen, 2020; Han et al., 2022; LaMar, 2018; Liu et al., 2018; Xiao & Liu, 2024) or discrete latent skill mastery (Zhan & Qiao, 2022; Liang et al., 2022) with examinees' choices of correct/incorrect subsequent actions, observed action subsequences, or sequence length. Furthermore, analyses of behavioral characteristics associated with successful/unsuccessful final performance (e.g., Gao, Cui, et al., 2022; Gao, Zhai, et al., 2022; Greiff et al., 2015; He & von Davier, 2016; Qiao & Jiao, 2018; Qiao et al., 2023; Ulitzsch et al., 2021, 2023) can inform test validation and automated scoring. Exploratory analyses of action sequences or sequence-derived patterns, often

with cluster analysis (Eichmann et al., 2020; He et al., 2019; Gao, Cui, et al., 2022; Gao, Zhai, et al., 2022; He, Borgonovi, & Suárez-Álvarez, 2023; Hao & Mislevy, 2019; Ulitzsch et al., 2022) or with topic modeling of actions or subsequences (Fang & Ying, 2020; Xu et al., 2018), have revealed different behavioral prototypes among the examinees as they face the same task, providing insights on how individuals navigate and approach computerized tests, digital platforms encountered in daily life, collaborative problems, etc. Hidden Markov and neural language models applied to action sequences have also been shown to reveal stages or subtasks for solving a problem (Wang et al., 2023; Xiao et al., 2021; Xu et al., 2020).

The current paper aims to provide an approach to exploratory sequence analysis to understand the relationship between problem-solving behavior and the test taker's external characteristics (i.e., background variables), for example, cognitive constructs other than the measured trait, demographics, and educational or job-related outcomes. Relationships between background variables and problem-solving behavior have been documented in many prior studies; for instance, demographic variables such as gender, migration status, or socioeconomic status were found related to interaction style in the PISA 2012 complex problem-solving assessment (Eichmann et al., 2020) and navigation behavior in PISA 2018 multiple-source reading tasks (He, Borgonovi, & Suárez-Álvarez, 2023). When a pattern in the problem-solving process is found associated with a background variable, the type of insight gained differs depending on whether the specific sequential pattern is theorized to provide evidence about the measured proficiency, i.e., *construct-relevant*: When the sequential pattern is theorized to be construct-relevant, uncovering its relationship with certain external variables can be meaningful for both test validation and instruction (e.g., Abele & von Davier, 2019). Test validation often involves the formation and testing of hypotheses about how the theorized construct underlying test score should be related to external variables of interest, for instance, score differences across demographic groups, across treatments or interventions, and individuals with different outcomes of interest (AERA et al., 2014). For simulation tasks where the test-taking process is recorded and has implications for final scoring (e.g., based on an expert-defined scoring rubric), the presence of relationship between a construct-relevant sequential pattern and these external variables constitutes one source of test validity evidence (e.g., Zumbo & Hubley, 2017) that may support such hypotheses and, similarly, support the adoption of scoring rubrics that consider related behavioral evidence (Mislevy et al., 2003). From an instructional perspective, targeted treatments and training in the attempt to close performance gaps, thereby reducing educational and income disparity, often require fine-grained information on why performance gaps exist, which may be partially informed by problem-solving process data (Bergner & von Davier, 2019). Educators and policymakers benefit from understanding how different subgroups solve questions differently, how interventions and available resources might explain problem-solving differences, and how test-taking process relates to key outcomes. On the other hand, if a sequential pattern associated with a background variable is theorized to be *construct-irrelevant*, then, from a test fairness perspective (Ercikan et al., 2020), the design of the simulation task and the associated scoring rubric should be examined in terms of whether the final score is free from influence by such construct-irrelevant behavior. The analytical method introduced in the current study serves as an initial, purely exploratory approach to examining the relationship between the problem-solving process and external covariates. Hypotheses generated from these analyses can subsequently be scrutinized in more theory-oriented follow-up investigations.

Two overarching pursuits in exploratory sequence analysis for the process-background relationship are (1) to quantify the strength of association between a background variable and problem-solving process and, where there is a substantial association, (2) to extract and interpret background-relevant sequential patterns. Despite its rich information, action sequence is in a nonstandard format: On a simulation task, each examinee's observed process data come in a temporally ordered sequence of computer-logged events. This precludes the use of applicable exploratory techniques that require structured input data. Therefore, the first step of the proposed approach is to transform the process data into structured data, specifically numerical feature variables that are learned to maximally preserve original sequence information. The current study adopts two recent methods for data-driven feature extraction, namely multidimensional scaling (MDS; Tang et al., 2020) and sequence-to-sequence autoencoders (Seq2seq; Tang, Wang, et al., 2021). Both methods automatically extract numerical features from raw action sequences and do not require a priori feature engineering using domain knowledge or a term-document matrix. To quantify the strength of the association between a background variable and the problem-solving process on a simulation task, the second step builds a regression model for the background variable using the extracted sequence features, and the prediction accuracy on new samples helps quantify the amount of information on a background variable provided by the action sequences. Process features are high-dimensional and contain noise that is irrelevant to specific background variables of interest. We thus employ regularized regression to perform variable selection in the regression. Process features

extracted in a data-driven manner are high-dimensional dense vectors and lack inherent interpretations. To facilitate the identification of specific sequential patterns that explain the process-background relationship, the third step employs partial least squares analysis, which identifies a few principal variables that maximally explain the covariance between sequence features and a specified background variable. This affords inspection of how sequential pattern changes as the principal variables vary from lowest to highest. We illustrate these steps via an empirical analysis of the Problem Solving in Technology-Rich Environments (PSTRE) assessment and background questionnaire data from the 2012 PIAAC survey, but our approach can generalize to the analysis of other simulation-based assessments that collect action sequences and background information.

The rest of the paper is organized as follows. The section Motivating Example introduces the 2012 PIAAC survey and the PSTRE assessment as well as the current research questions. The next section provides a review of the literature on sequence analysis applied to PIAAC PSTRE and introduces the proposed approach to exploratory sequence analysis of the process-background relationship. The Empirical Analysis Methods and Empirical Analysis Results sections present the methods and results of the empirical study. The Discussion section provides a discussion of the empirical findings, as well as their practical implications for assessment design and interventions.

Motivating Example

The PIAAC (e.g., Schleicher, 2008) is an international large-scale assessment carried out by the Organization for Economic Co-operation and Development (OECD) to assess the cognitive and workplace skills of working-age individuals worldwide. In its first cycle in 2012, working-age individuals (16–65 years) across 25 countries and regions were measured on literacy, numeracy, and PSTRE. In addition to the three cognitive assessments, participants were also administered a background questionnaire, which collected self-reported information on their education, social background, engagement in literacy, numeracy, and use of informational and communicative technology (ICT) at home and at work, educational background, language background, employment information, and others such as health status and political efficacy (Kirsch & Thorn, 2013).

The PSTRE assessment consisted of two test blocks, with 14 items in total. For each PSTRE item, the test environment resembled commonly seen ICT platforms, such as e-mail clients, web browsers, and spreadsheets. Examinees were prompted to complete specific tasks on these interactive platforms. Under the PIAAC framework, PSTRE is defined as the use of digital technology, communication tools, and the internet to obtain and evaluate information, communicate with others, and perform practical tasks (OECD, 2012). As actual PSTRE items are unreleased, Figure 1

Figure 1. PSTRE sample item. Reprinted from Sample Questions and Questionnaire, OECD, http://www.oecd.org/skills/piaac/Problem%20Solving%20in%20TRE%20Sample%20Items.pdf.

presents an item from the OECD Education and Skills Online assessment, which illustrates the interface of the PSTRE items. An examinee works in the simulated web browser to complete the task described on the left: Five web pages (1st subfigure) are returned from a search of "Job search" and examinees are asked to bookmark all pages that do not require registration or fees. Clicking on each link will direct them to the corresponding website. For example, clicking the second link, "Work Links," directs an examinee to the second subfigure, and further clicking on "Learn More" directs the examinee to the third subfigure. To finish and exit the item, they can click on the right arrow icon ("Next") below the item instructions, and there will be a pop-out window with two options, namely confirming exit ("Next_OK") or returning to the task ("Next_Cancel"). An action sequence on the task will consist of the clicks and keystrokes by the examinee on the simulated browser, with "Start" as the initial action and "Next_OK" at the end. For example, if an examinee clicked the second link ("Work links") on the initial page, clicked "Learn More" on the next page, clicked the "Back" button on the toolbar twice to return to the home page, and clicked "Next" in the left panel and "OK" in the pop-up window, this examinee's action sequence will be recorded as "Start, Click_W2, Click_Learn_More, Toolbar_Back, Toolbar_Back, Next, Next_OK." Based on predefined scoring rubrics, the PSTRE assessment computed a final binary or polytomous score for each item, which was used to estimate individuals' PSTRE proficiency.

The PSTRE assessment measured adults' abilities to solve problems in personal, work, and civic contexts using digital environments to better understand how working-age adults utilize digital tools in practical problem-solving, thereby providing insights to policymakers and educators on fostering digital literacy and addressing skill gaps. Across several studies, it was found that PSTRE proficiency was associated with demographic characteristics as well as employment outcomes of the participants (He et al., 2019; Liao et al., 2019; Nwakasi et al., 2019). Although differing across countries that varied in policies, labor market structures, and social contexts, PSTRE proficiency was found to be positively associated with self-reported income in many countries or regions. Participation in adult education and training, which is expected to increase exposure to ICT tools, was consistently found associated with higher PSTRE proficiency. Furthermore, gender and age differences were found in PSTRE proficiency, with female and older adult participants receiving lower PSTRE scores in many countries and regions (He et al., 2021; Liao et al., 2019).

The high demand for digital literacy in most economic activities preordains the importance of PSTRE skills for the workforce. This calls for understanding of the sequential patterns in digital problem-solving that explain the differences in PSTRE scores observed for individuals with varying demographics, exposure to adult training and ICT tools, and employment outcome, as documented by the existing literature. The current study explores the relationship between how adults solve PSTRE problems, as reflected through the PSTRE action sequence data, and background variables. Six demographic variables were considered: This includes two demographic variables, namely age (in years) and gender (male or female), one employment outcome variable, namely log of country median-adjusted hourly income (log(Income)), and three variables related to education and ICT exposure, namely self-reported ICT skill use at home (ICTHome) and at work (ICTWork) and country median-adjusted total years of education (YRSEdu).[1] As some of the PSTRE tasks appeared to also involve numeracy skills (e.g., spreadsheets), performance (i.e., plausible values[2]) on numeracy was also included as an external cognitive variable. Specifically, the following research questions are addressed:

Research Question (RQ1): How related is each background variable to the 14 items' action sequences? Do action sequences provide additional information about each background variable on top of final scores?

Research Question (RQ2): What are the specific behavioral patterns in the action sequences that explain the sequence-background association?

Essentially, RQ1 seeks to quantify the strength of association between the participants' background and the test-taking process on a task, and RQ2 seeks to interpret sequential patterns associated with the background variables, which aids in generating initial hypotheses as to whether the association of problem-solving patterns with the participants' background is construct-relevant or construct-irrelevant.

[1] In the PIAAC background survey, scores reported on ICTHome and ICTWork were derived based on examinees' responses to a series of corresponding survey items. Specifically, eight items in ICTHome (i.e., H_Q03a, H_Q03b, H_Q03c, H_Q03d, H_Q03e, H_Q03f, H_Q03g, and H_Q03h) and eight items in ICTWork (i.e., G_Q05a, G_Q05b, G_Q05c, G_Q05d, G_Q05e, G_Q05f, G_Q05g, and G_Q05h) were included. See details in Chapter 3 in PIAAC Tech Report (OECD, 2016).

[2] The PIAAC survey derived plausible values of individuals' cognitive performance (e.g., numeracy performance) based on both the responses to cognitive assessments and background variables. The numeracy performance variable used in the current study is based on the mean of the numeracy plausible values recorded in the official PIAAC data.

Analysis of PSTRE Action Sequence and Background-Process Relationship

The PIAAC PSTRE process data have been shown in multiple previous studies to provide valuable additional information on individuals' problem-solving processes beyond the final scores. For example, on tasks involving spreadsheets, He and von Davier (2016) used *n*-gram language modeling and identified subsequences related to final performance, such as the use of searching and sorting tools, and country differences in problem-solving, e.g., individuals from the Netherlands were more likely to perform double-checking. With clickstream analysis using graph-modeled clustering (Ulitzsch et al., 2021, 2023) and multidimensional scaling (Tang et al., 2020), studies consistently found that, even within the group of examinees who received the same final score, there was remarkable heterogeneity in the problem-solving process, including how the problem was approached, performing of nonessential actions for the task, and how one arrived at an incorrect response. This provided empirical support for the development of scoring rules that incorporated process information, which achieved remarkably higher measurement reliability than the final score-based proficiency estimator on the PSTRE assessment (Zhang et al., 2023). This finding was further strengthened in a recent study of He, Shi, and Tighe (2023), where the sequential problem-solving process patterns were found robust to significantly enhance the prediction power of low-skilled adults' PSTRE proficiency level via a hierarchical machine learning approach. Cluster analyses on the planning duration and interaction frequency on the 14 PIAAC PSTRE items and on the navigation trajectory in the browser-based tasks also revealed general problem-solving styles on simulation tasks (Gao, Zhai, et al., 2022) and different ways adults navigate multilayered hypertext environments to obtain information (Gao, Cui, et al., 2022).

The relationship between the PSTRE problem-solving process and background variables has also been explored in several previous studies. Some examined demographic differences in globally defined problem-solving characteristics. For example, He et al. (2021) employed the longest common subsequence approach to examine how problem-solving efficiency and sequence similarity to a reference sequence (reflecting optimal problem-solving) on the seven items in the 2nd PSTRE block relate to demographic characteristics such as gender, age, and familiarity with digital platforms. A few studies also explored the demographic differences in task-specific behavior, for instance, key actions or short subsequences of actions (i.e., *n*-grams; He & von Davier, 2016) on PSTRE tasks that were associated with participants' income, education level, and other characteristics, such as age (Liao et al., 2019). The current study investigates this process-background relationship on the task level for all 14 PSTRE items. We adopt a different approach based on extracting and examining data-driven features extracted to preserve raw action sequence information, which we describe below.

Feature Extraction From Action Sequences

Action sequences come in a nonstandard format. As a toy example, below are three arbitrarily selected participants' observed action sequences on item U06b, omitting "Start" and "Next, Next_OK" at the beginning and the end:

- Examinee 1: "Click_W4, Toolbar_Web_Back, Response_Open, Response_4, Response_Close"
- Examinee 2: "Click_W2"
- Examinee 3: "Click_W1, Toolbar_Web_Back, Click_W2"

On the same item, the number of actions performed by each individual differed. An examinee's observed sequence contains a list of temporally ordered categorical actions (e.g., "Click_W4"), with the number of possible actions ranging between 26 and 636 on the 14 items. A common approach in sequence analysis is to first transform the original variable-length, ordered, categorical sequences, which preclude most statistical methods, into rectangular data: In doing so, a set of numeric features are extracted to preserve original sequence information. There are many methods to fill this task. One approach is via a *bag-of-words* (or bag-or-phrases) model, which assumes that a sequence can be represented by the actions or subsequences (i.e., *n*-grams) that occur in them. In this case, the observed sequence data are summarized into a term-document matrix where each row is an observed sequence, and each column contains the (weighted) frequency of a particular action or length-*n* subsequence (i.e., *n*-grams) within that observed sequence. With the inclusion of *n*-grams in the term-document matrix, this approach can preserve short-term information of up to *n* consecutive steps in an examinee's action sequence, where *n* is typically small (less than 4) to keep the computations manageable. As a result, long-term dependencies in the action sequences, such as two actions that are related but more than *n* steps apart in the sequence or a long keystroke pattern that spans more than *n* words or characters, are not preserved. To reveal demographic differences in long-term or overall test-taking behavior on a task, the current paper approaches the action sequence feature extraction task using two sequence-based feature extraction methods, MDS (Tang

et al., 2020) and Seq2seq (Tang, Wang, et al., 2021). We briefly introduce the rationale behind the two methods below, and a technical description is provided in the Electronic Supplementary Material, ESM 1, Appendix I.

The goal is to extract K-dimensional numerical features that preserve original sequence information from the raw action sequences. For each observed sequence $i = 1, \ldots, N$, MDS learns a K-dimensional feature $\boldsymbol{x}_i$ by performing multidimensional scaling, an unsupervised dimension reduction technique, on a sequence pairwise dissimilarity matrix. The dissimilarity matrix is an $N \times N$ symmetric matrix, with the $[i, j]$th entry (d_{ij}) describing how *dissimilar* the action sequence of examinee i is from that of examinee j. In quantifying sequence dissimilarity, we adopt the order-based sequence dissimilarity measure (Gómez-Alonso & Valls, 2008; Tang et al., 2020), which takes into consideration the dissimilarity in both the choice of actions and their temporal ordering. The MDS features ($\boldsymbol{x}$ s) are learned so that the pairwise Euclidean distances (between every pair $\boldsymbol{x}_i, \boldsymbol{x}_j$) is as close as possible to the sequence dissimilarity (d_{ij}), so intuitively, MDS features preserve individual difference information: Features are optimized to best recover the pairwise dissimilarity of examinees' action sequences in terms of the types of actions taken and their ordering.

Seq2seq feature extraction (Tang, Wang, et al., 2021), on the other hand, aims to preserve information that can be used for sequence reconstruction: The task is achieved by training on the N observed action sequences a deep learning model that consists of two recurrent neural networks, namely an encoder that first compresses an original sequence into features ($\boldsymbol{x}$ s), followed by a decoder that subsequently predicts the original sequence based on the compression ($\boldsymbol{x}$). The Seq2seq model is learned so that, across the observed sequences, the distribution of the reconstruction (i.e., output from the decoder) is as close as possible to the original action sequence.

MDS and Seq2seq both perform sequence-based feature extraction in that the feature extraction is conducted to reconstruct original sequences or individual pairwise differences on the original sequences. Compared to term-document-matrix-based methods for sequence feature extraction, which often adopts matrix factorization (e.g., latent semantic analysis, non-negative matrix factorization; Deerwester et al., 1990; Lee & Seung, 1999) to find lower dimensional features that can reconstruct the term-document matrix, MDS and Seq2seq are expected to capture additional information beyond standalone frequencies of single actions or short subsequences, including ordering of actions and long-term effects. On the PSTRE data, feature extraction based on MDS and Seq2seq has documented performance in preserving original sequence information (e.g., Tang et al., 2020; Tang, Wang et al., 2021; Zhang et al., 2023), and software for both is available in the ProcData (Tang, Zhang, et al., 2021) R package. We chose MDS and Seq2Seq for feature extraction since the objective of both methods is to maximally preserve the original sequence information. This preservation allows for the relationships between background variables and sequential patterns to be retained as much as possible in the transformed feature space.

Quantifying Sequence-Background Association

For each examinee and item, the extracted MDS or Seq2seq features are dense K-dimensional numerical features that preserve individual differences in the action sequences. To illustrate, Table 1 presents 5-dimensional MDS and Seq2seq features extracted from U06b for the three examinees' action sequences in the toy example. In practice, the choice of dimension K should be based on cross-validation (see Tang et al., 2020) and often needs to be larger than five to adequately preserve the rich information in the action sequences. To quantify the strength of association between the test-taking process on a task and a background variable, one way is to evaluate how well the background variable can be predicted with the task's MDS or Seq2seq sequence features, a surrogate to the original sequences that are expected to well preserve their information. Because the process features are high-dimensional and contain information irrelevant to the

Table 1. Action sequence features on U06b for the three examinees, extracted using MDS (left) and with Seq2seq (right), both with $K = 5$ dimensions

	MDS			Seq2seq		
K	Examinee 1	Examinee 2	Examinee 3	Examinee 1	Examinee 2	Examinee 3
1	0.04	0.16	−0.04	−0.56	0.09	−0.62
2	0.15	−0.27	−0.15	0.90	0.08	0.89
3	0.20	−0.07	−0.04	−1.00	−0.90	−1.00
4	−0.07	−0.09	0.13	0.94	0.62	0.94
5	−0.01	−0.17	−0.21	−0.72	−0.36	−0.73

Note. k = dimension (1–5) of the MDS/Seq2seq features. The ranges of the features differed across dimensions and for MDS and Seq2Seq. For instance, across all 3,645 examinees, dimension 1 of the Seq2seq features ranged between −.92 and .27, and dimension 2 ranged between −.18 and .97.

background variable, we recommend performing cross-validation combined with variable selection to prevent overfitting. One option is fitting on the training data a regularized generalized linear model (GLM), e.g., via the R glmnet package (Friedman et al., 2009), with the background variable as the outcome, combined with a logit link function for predicting binary variables and an identity link for continuous variables. The MDS or Seq2seq process features are treated as predictors, with a weight penalty, such as the L_2 norm penalty for ridge regression, to shrink the parameter estimates toward 0. Prediction accuracy can then be evaluated on test data unseen during model training. The area under the curve (AUC) of the receiver operating characteristic curve may be used as an evaluation metric for binary outcome variables, and out-of-sample Pearson correlation (O.S.R), the correlation on test data, can be used for continuous variables. Higher prediction accuracy indicates a stronger association between the sequence patterns on the task and the background variable.

Interpretation of Sequence-Background Association

When an outcome variable is well-predicted with the sequence features of an item, interpretations can be sought to uncover the specific sequential patterns associated with the variable. Although the sequence features are high-dimensional, one may conjecture that patterns relevant to the background may be represented with just a few principal dimensions. Partial least squares (PLS; Wold et al., 2002) decomposition can perform dimension reduction on the K-dimensional sequence features to extract principal features that maximally account for the covariance between the features ($\boldsymbol{x}$) and the dependent variable (y). PLS differs from principal component analysis (PCA) in that the top M components capture the most covariance between $\boldsymbol{x}$ and y instead of the variance of $\boldsymbol{x}$ (as in PCA). The output contains M orthogonal components, the first component being a projection of the action features that maximally explain the covariance between sequence features and background, and the second orthogonal to the first one, maximally explaining the remaining unexplained sequence-background covariance, etc. For each of the M PLS components, observed action sequences can then be sorted from lowest to highest on the component score, and inspecting how the observed action sequence patterns change as the PLS component score increases can help pinpoint specific pattern(s) that explain the covariance between process features and the predicted variable.

The dimension for the PLS approximation, M, can be chosen based on the root-mean-squared error of prediction (RMSEP; e.g., Wehrens & Mevik, 2007) on cross-validation data: This starts with running PLS with large $M_{\max}$. For each $M' \in \{1, \ldots, M_{\max}\}$, the method evaluates uses the first M' PLS components to predict y in a linear regression model and computes the root-mean-squared error for the prediction (RMSEP) on validation samples, i.e.,

$$\mathrm{RMSEP}_{M'} = \sqrt{\frac{\sum_i \left(y_i - \hat{y}_i\right)^2}{N}}, \quad (1)$$

as well as the standard error of RMSEP. The number of PLS components to retain, M, was then chosen to be the smallest number of dimensions that achieves an RMSEP within 1 standard error from the minimum: Specifically, suppose that across all possible M's between 1 and 100, $M_{\min}$ achieves the lowest RMSEP. Then, this approach chooses the smallest M whose *RMSEP* is within 1 standard error above that of $M_{\min}$'s. In this case, M is often smaller than $M_{\min}$, which trades small additional y variance explained for parsimony. In the subsequent section, we illustrate this on the PIAAC data.

Empirical Analysis Methods

Data and Instruments

The current study used the PSTRE item-level sequence data and background questionnaire data from 3,645 examinees in five countries or regions, including the United Kingdom (England and Northern Ireland), Ireland, Japan, the Netherlands, and the United States. These examinees were administered all 14 PSTRE items, presented as two blocks of seven items each.[3] Action sequence data were available for each participant at the item level. Table 2 presents a brief description of the 14 PSTRE items, including task names, the types of environments involved, percent of individuals who received full credit, and descriptive statistics of the action sequences. The background variables were available in the PIAAC background questionnaire

[3] In the PIAAC computer-based assessment, examinees were routed to modules of PSTRE, literacy, or numeracy items. See details in Chapter 1 in PIAAC Tech Report (OECD, 2016). Each respondent was assigned two modules. The current study used participants who received two PSTRE modules.

Table 2. Descriptive information of the 14 PSTRE items

Item ID	Item name	Environment	*p*	min(*L*)	max(*L*)	mean(*L*)	*N*
U01a	Party invitations	Email	.56	3	114	18.21	40
U01b	Party invitations	Email	.50	3	132	25.65	47
U02	Meeting room	Email/Web	.14	3	153	25.81	96
U03a	CD Tally	Web/SS	.39	3	51	8.96	67
U04a	Class attendance	Email/SS	.12	3	304	37.38	636
U06a	Sprained ankle	Web	.27	3	57	10.04	30
U06b	Sprained ankle	Web	.53	3	68	15.39	26
U07	Book order	Web	.49	3	79	19.05	40
U11b	Locate e-mail	Email	.23	3	256	25.17	124
U16	Reply all	Email	.61	3	267	32.98	362
U19a	Club membership	Email/SS	.76	3	357	17.64	85
U19b	Club membership	Web/SS	.48	3	396	19.38	244
U21	Tickets	Web	.38	3	77	19.74	138
U23	Lamp return	Web/Email	.37	3	138	21.62	136

Note. *p* refers to the proportion of subjects who answered the item correctly; min(*L*), max(*L*), and mean(*L*) refer to the minimum, maximum, and average sequence length across all subjects, respectively; *N* refers to the number of types of actions on the item; SS = spreadsheet.

data. Here, to reduce the nuisance introduced by country effects, we converted hourly income to US dollars based on 2012 exchange rates and adjusted income and YRSEdu by subtracting the country's median on that variable. There was missingness for a few background variables, including log(Income), ICTHome, ICTWork, and YRSEdu. For predicting a particular variable, individuals with missing values were case-wise deleted. In particular, two outcome variables, ICTWork and log(Income), contained a substantial amount of missingness across all countries or regions (31%–45%). ICTHome and YRSEdu contained a small to moderate amount of missingness across countries or regions (0%–13%). The missingness proportions on these variables for each of the five countries and regions are reported in Table E1 in ESM 1, Appendix III. *M*, *SD*, and the number of nonmissing observations of each continuous variable, as well as their pairwise Pearson correlations, are reported in Table 3. Compared to male participants, female participants on average showed lower log hourly income, ICTWork, and numeracy performance, with small Cohen's *d* effect sizes. Additionally, the correlations between each variable and the overall PSTRE performance (derived from PSTRE final scores) are also presented.

Feature Extraction From Action Sequences

In the current study, $K = 100$ features were extracted from each item and for both MDS and Seq2seq based on the action sequences of the 3,645 participants. Following feature extraction, principal component analysis (PCA; Hotelling, 1933) was applied to the 100 MDS (and similarly Seq2seq) features of each item. The purpose of performing the PCA is to decorrelate the extracted features by rotating them into directions that incrementally explain the original feature (MDS/Seq2seq) variance. This facilitates subsequent variable selection in regularized regression.

Prediction and Evaluation

To draw predictions about each background variable, for each item, three different types of predictors were considered, namely (1) the polytomous final score on the item, (2) the 100-dimensional principal sequence features extracted using the Seq2seq, and (3) the 100-dimensional principal features from MDS. For each item and background variable, 10 replications were carried out for each prediction model. In each replication, the data were randomly partitioned into three subsets, a training set (70%), a validation set (10%), and a test set (20%). The parameters of the GLM were estimated based on the training data, the optimal weight penalty for L_2 regularization was chosen to minimize the loss function on the validation data, and the prediction accuracy of the background variable was evaluated on the test data. The average prediction accuracy on the test data (i.e., O.S.R or AUC) across the 10 replications is reported.

Feature Interpretations

For select items whose process features demonstrated especially high associations with a specific background

Table 3. Descriptive statistics of the background variables and cognitive scores

	Age	log(Income)	ICTHome	ICTWork	YRSEdu	Numeracy	PSTRE
N	3,645	2,214	3,375	2,289	3,479	3,645	3,645
M	38.930	−3.581	2.076	2.081	−0.135	0.443	−0.173
SD	13.546	0.404	0.937	1.012	2.618	0.750	0.748
Correlation							
Age	—	0.269	−0.111	0.065	0.018	−0.050	−0.307
log(Income)	—	—	−0.047	0.310	0.318	0.276	0.149
ICTHome	—	—	—	0.335	0.181	0.174	0.297
ICTWork	—	—	—	—	0.22	0.176	0.245
YRSEdu	—	—	—	—	—	0.35	0.315
Numeracy	—	—	—	—	—	—	0.796

Note. N stands for the number of nonmissing observations for each variable.

variable, interpretations for the process-variable associations were further sought. PLS decomposition was applied to the 100-dimensional sequence features to find the top *M* components that maximally explain the background-feature covariance. For each of the *M* PLS components, we searched for patterns in the action sequences associated with the component. To do this, we first ranked all 3,645 examinees based on the PLS component, and then, from lowest to highest, at an interval of 50, we inspected their observed sequences. By looking at how the sequences changed as the component score increased, patterns in the action sequences associated with the PLS component were identified. We further verify the relationship between the visually spotted patterns and the PLS component by plotting how the pattern changes with the PLS component score with locally weighted scatterplot smoothing (LOWESS; Cleveland, 1981), as well as computing the Pearson correlation between the PLS score and the sequential pattern.

Empirical Analysis Results

Quantifying Background-Sequence Relationship (RQ1)

For each item, the prediction accuracy of the continuous variables and gender are reported in Figures 2 and 3. In each subplot, the *x*-axis represents the item used for prediction, the *y*-axis represents the evaluation metric, i.e., averaged test sample O.S.R or the AUC across 10 replications, and the three bars with different shades represent the prediction results based on polytomous score, 100 Seq2seq features, and 100 MDS features, respectively. As was mentioned in the item descriptions, most of the PSTRE items involved one or two of the following environments: spreadsheet (SS), web browser (Web), or e-mail box (Email), and items involving similar environments could share some common actions. Items are grouped by ICT environment(s) involved. Comparing the prediction results across items, MDS sequence features on a task tended to achieve the highest prediction accuracy on each background variable, followed by Seq2seq features (slightly lower) and polytomous final scores. It is also worth comparing the prediction accuracy of each outcome variable based on an item's polytomous scores to that based on process features. All variables tended to be predicted with consistently higher accuracy using process features than based on the final score, suggesting process features contain additional information on these external variables that were not considered in scoring. At this stage, the results on prediction accuracy can only aid with quantifying the amount of information related to each external variable. Generating initial hypotheses as to whether this information is construct-relevant, on the other hand, requires interpretation of the sequential patterns associated with the sequence features (RQ2).

Although all 14 items were designed to measure the same trait (PSTRE), there was noticeable heterogeneity across items in their strengths of associations with various background variables. Despite that features extracted from an item's action sequences frequently showed a remarkable amount of prediction power on the background variables, the magnitudes of the prediction power differed vastly, both across background variables and across items. The sequence features in general showed higher associations with participants' age and numeracy than with some of the employment outcome-related (i.e., income) and education/ICT exposure-related variables. For the same background variable, the strength of sequence-background association also showed great variability across items: Taking

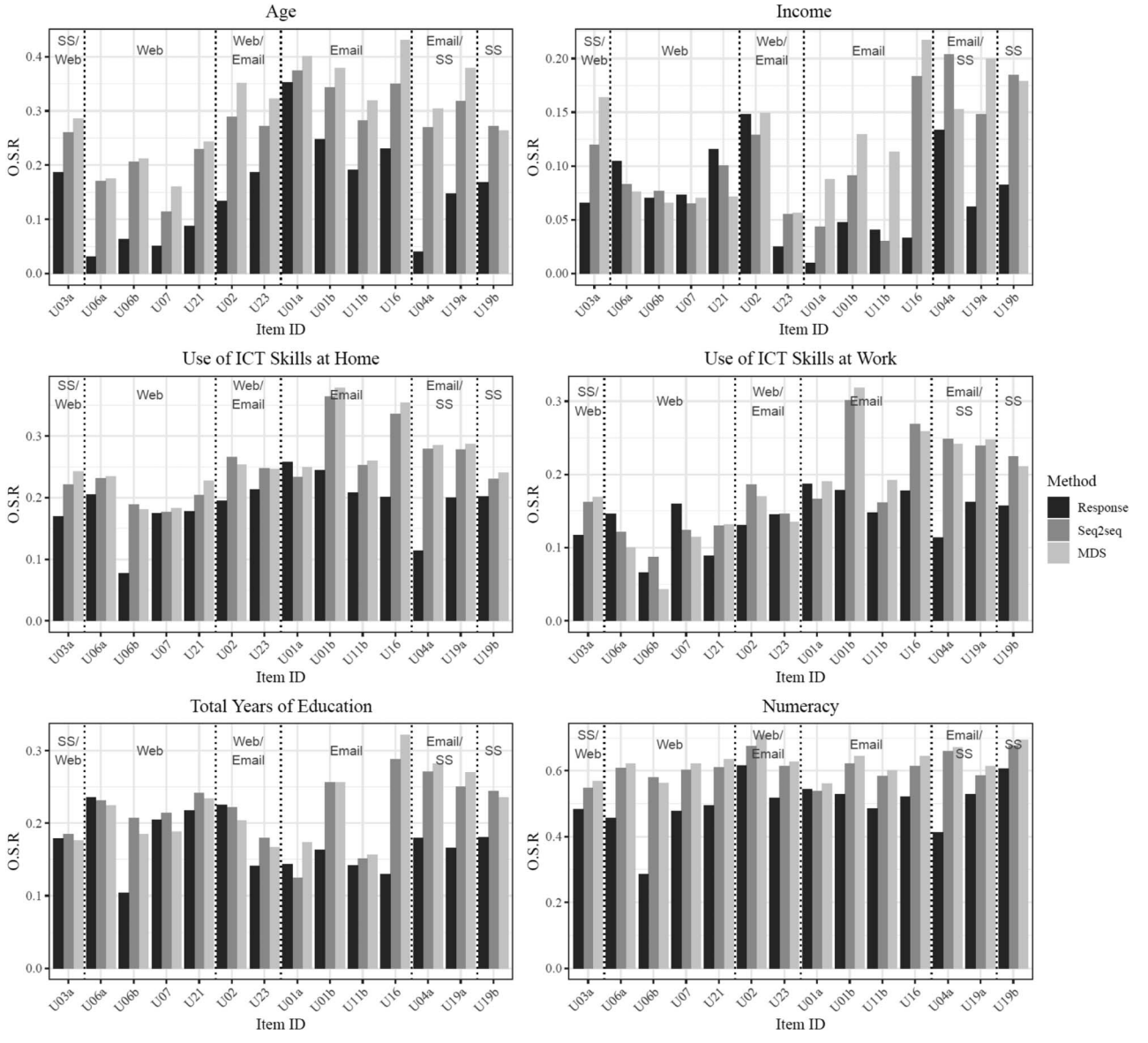

Figure 2. Prediction accuracy of continuous variables from polytomous scores, Seq2seq features, and MDS features of each PSTRE item. O.S.R stands for test-sample (out-of-sample) correlation between observed and predicted values.

age as an example, the *O.S.R* ranged from .16 (U07) to .43 (U16). This variability holds even for items sharing the same environment. An example is the prediction of ICTHome using two e-mail items, U01a and U01b, where U01b showed much higher prediction power (.37) than U01a (.24) despite that their final scores were similarly predictive of ICTHome. In subsequent interpretations of sequential patterns associated with a background variable, one may prioritize items that show a particularly strong association, for instance, U01b which showed the highest *O.S.R.* in predicting ICT use at home. Such interpretations help understand how ICT exposure relates to test-taking behavioral patterns on the item.

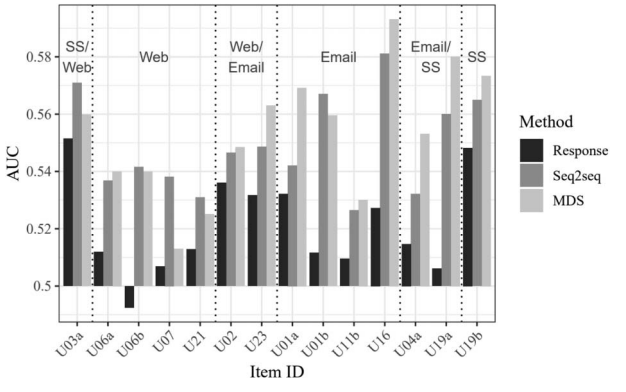

Figure 3. Prediction accuracy of gender from polytomous scores, Seq2seq features, and MDS features.

Interpretations for Associations (RQ2)

From the results on the general prediction accuracy, it was observed that not only do sequence features provide additional information about the examinees, but the prediction powers of different items also differed widely. This calls for a closer look at action sequences on tasks with strong background associations. As the MDS features often demonstrated the highest prediction power, for each continuous variable, we performed PLS decomposition on a selection of items where the MDS features had the highest predictive power of the external variable. Depending on the item and the external variable, the number of PLS components extracted based on the RMSEP criterion differed. Table 4 presents a summary of the PLS components identified for each background variable and item, the Pearson correlation between the PLS component score with the external variable (ρ_Y), interpretations on the sequence pattern associated with the PLS component score, and the Pearson correlation between the PLS component score and the identified pattern ($\rho_{pattern}$). Relationships between PLS component scores and many variables were found to be nonlinear, in which case, the Pearson correlation will not adequately capture the relationships. For ease of presentation, Pearson correlations are reported here, but we recommend inspecting the LOWESS curves for potential nonlinear relationships. LOWESS plots for the relationships between each PLS score, background variables, and identified sequential patterns are included

Table 4. Interpretation of MDS PLS features for each external variable and correlations between PLS component score and external variable/sequence pattern

Variable (Y)	Item	PLS	ρ_Y	Sequence pattern	$\rho_{pattern}$
Age	U01a	1	.39	*Moving e-mails to correct folders*	−.75
		2	.14	*Usage of toolbar icons*	−.31
	U01b	1	.34	*Creation of new folder for e-mails*	−.84
		2	.20	Max number of consecutive clicks of e-mail folders	.40
	U16	1	.26	*Dichotomized score on U16*	−.65
		2	.26	*Usage of keyboard shortcut for copy/paste*	−.44[a]
		3	.20	Frequency of alternation between typing/not typing	.29
		4	.17	Logarithm of the number of manual keystrokes	.20
	U19a	1	.27	*Dichotomized score on U19a*	−.68
		2	.25	*Usage of dropdown menu to send e-mail*	.44
		3	.15	*Correct response to U19a without "search"*	−.74
Income	U16	1	.18	*Usage of copy/paste*	.35
	U19a	1	.17	*Usage of "search" in spreadsheet*	.63
	U19b	1	.19	*Usage of "sort" in spreadsheet*	.88
ICTHome	U01b	1	.35	*Creation of new folder for e-mails*	.84
		2	.21	*Japanese keyboard entry*	−.75
	U16	1	.31	Logarithm of keystroke count (excl. backspace)	.93
		2	.26	*Usage of keyboard shortcut for copy/paste*	.39
ICTWork	U01b	1	.29	*Creation of new folder for e-mails*	.84
YRSEdu	U04a	1	.27	Number of correct numerical entries in spreadsheet	.79
	U16	1	.18	*Usage of cc or "reply to all" to cc e-mail recipients*	.58
		2	.25	*Usage of keyboard shortcut for copy/paste*	.50
Numeracy	U04a	1	.64	Number of correct row/column titles in spreadsheet	.91
		2	.17	Logarithm of the ratio between number of correct numerical entries and number of correct row/column Title entries in the spreadsheet.	.71
		3	.16	Logarithm of the ratio between number of switches between spreadsheet and e-mail environments and The number of spreadsheet entries.	.50

Note. [a]Japanese participants were excluded due to keystroke coding differences in the log file. Binary sequential pattern variables are italicized.

in ESM 1, Appendix II. In what follows, we summarize the findings from the PLS decomposition and pattern interpretations.

Age

We performed PLS decomposition on four items whose MDS sequence features showed high associations with age, namely U01a, U01b, U16, and U19a. By inspecting the PLS components for age across items, we found the following trends:

- Senior examinees were less likely to complete the steps necessary for receiving full scores (PLS 1 of all four items). These sequential patterns are construct-relevant and are already considered in final scoring. For instance, the sequential pattern corresponding to PLS 1 on item U01a, interpreted as moving e-mails to the correct folders, was correlated with score on U01a at $\rho = .88$ and with overall PSTRE proficiency at $\rho = .58$.
- Senior examinees were more likely to perform a task, e.g., sort a spreadsheet, using text-based drop-down menus than graphical icons (e.g., an arrow icon for "sort") in the toolbar (U01a and U19a, PLS 2). These sequential patterns are clearly construct-irrelevant and do not influence final scores. For instance, the sequential pattern corresponding to PLS 2 on item U01a, interpreted as the usage of toolbar icons, was correlated with score on U01a at $\rho = -.01$ and with overall PSTRE proficiency at $\rho = -.01$.
- Senior examinees were less likely to use shortcuts when sending e-mails, e.g., using "reply all" to respond to multiple recipients and copy/pasting, and were more likely to type text contents manually (U16, PLS 2–4).
- One particular age PLS component to note was PLS 3 on item U19a, where participants were prompted to identify a person's information from a long spreadsheet. There were three ways to identify the requested row, namely to eyeball all rows, to search for the person's name directly, or (more rarely) to sort the spreadsheet alphabetically by names. While inspecting the sequences ranked on this component, it was found that examinees on the lower end of this component managed to submit the correct answer, but without using search. Examinees on the higher end, however, rarely answered correctly without using "search." This observation was confirmed as the component was found negatively related to the conditional probability of responding correctly, given that the subject did not use "search." This component increased as age increased: Without using "search" or "sort," one had to eyeball the long spreadsheet to find the row, which can be visually taxing for older adults. This pattern is speculated to reflect some construct-irrelevant variance that influenced scores, namely the ability to work with visually demanding interfaces. This speculation was supported with the additional observation that, although the 0/1 indicator for whether the examinee answered the question correctly without search/sort was correlated with the current item's score at .44, its correlation with scores on the remaining 13 items ranged between −.09 and .03, suggesting that these examinees did not perform better on other PSTRE tasks.

Income

When it comes to the participants' hourly income, the MDS features from U16, U19a, and U19b demonstrated higher prediction accuracy compared to the others. PLS decomposition was hence applied to these three items. The identified PLS components were uniformly found to reflect efficient strategies for problem-solving, specifically the use of keyboard shortcuts for copy/pasting (i.e., Ctrl + C, Ctrl + V, item U16) and the tendency to use "search" or "sort" in spreadsheets (items U19a, U19b). These patterns related to efficient problem-solving tended to contain information that is relevant to the assessed PSTRE proficiency but was not directly considered in final scoring. For instance, on item U16, regardless of whether the examinee used keyboard shortcuts for copy/pasting, full score was given if the examinee completed the task (sending the requested information via e-mail). However, among participants who received full score on the item, those who used copy/pasting had higher overall PSTRE proficiency (M = .47) versus those who did not (M = .19, two-sample t-test $t = 8.94, p < .001$, 95% confidence for PSTRE proficiency difference: $(.22, .34)$).

ICT Use at Home and at Work

PLS decomposition was performed on two items showing larger MDS features-ICTHome association, namely U01b and U16. On item U01b, it was observed that individuals with more self-reported use of ICT tools at home were more likely to create e-mail folders, which was a key step to correctly solving U01b. The second PLS component on U01b was a country artifact: Participants from Japan generally had lower usage of ICT at home, and the PLS score was associated with Japanese keystrokes. On U16, higher ICTHome individuals were generally higher on PLS components for (1) more keyboard entry (PLS 1) and (2) more usage of keyboard shortcuts for copy/pasting (PLS 2). For self-reported use of ICTWork, we performed PLS analysis on one item, U01b, and one PLS component was found, which was related to the creation of new folders for e-mails.

Total Years of Formal Education

We report the PLS analysis results and interpretations of two items with relatively high prediction power on the examinees' total years of education, namely U04a and U16. On U04a, which involved creating a spreadsheet based on the information described in an e-mail, a PLS component positively related to years of education was found. This component was associated with a higher number of correct numerical entries in the spreadsheet. On U16, two PLS components positively associated with years of education were found, the first related to cc'ing (either typing in the "cc" field or using "reply all") other recipients when drafting the e-mail, and the second related to using keyboard shortcuts for copy/paste.

Numeracy

For numeracy, we focused on the interpretation of item U04a, which required examinees to synthesize the information from the e-mail into a spreadsheet table. Participants needed to enter both the column/row titles and the numerical entries. Three PLS components positively correlated with numeracy proficiency were found. The first PLS component was associated with the correct choice of spreadsheet row and column titles. The second PLS component was associated with the ratio between the number of correct numerical entries in the spreadsheet cells and the number of correct entries in the spreadsheet's row/column titles. The third component was related to the ratio between the number of times the examinee alternated between the spreadsheet and e-mail environments and the number of times the examinee worked on the spreadsheet entries/titles. Note that the relationship between the third PLS component and numeracy proficiency was clearly nonmonotonic (see ESM 1, Appendix II Figure E10 for LOWESS plot). The score on PLS 3 was highest for individuals with a moderate level of numeracy proficiency. Those with higher numeracy proficiency might have higher working memory capacity, allowing them to fill in more spreadsheet entries before referring back to the e-mail for the information. Those with low numeracy proficiency could not synthesize the relevant information in the e-mail into a spreadsheet table, thus referring back to the e-mail less often.

Discussion

This paper introduces a sequence feature-based approach to evaluating and interpreting the relationship between action sequences on computerized simulation tasks and participants' backgrounds. MDS and Seq2seq were adopted for extracting features from raw action sequences to preserve as much information as possible. Sequence feature-based regularized regression further quantifies the strength of association between a background variable and test-taking process on a task. The results on the prediction of different background variables showed that action sequence-derived features, especially those extracted from MDS, consistently showed a higher association with background variables compared to polytomous final scores on the PSTRE items. This suggests that the sequences of actions an individual performs on a simulation task contained unique information about a variety of background variables: In these cases, individuals with different backgrounds (e.g., age, income, ICT skills, year of education, numeracy basic skills) tended to demonstrate different problem-solving patterns on a task, but the associated behavior might not have been used as part of proficiency scoring. PLS analysis on the specific items' MDS features further unveiled sequential patterns that differentiate participants on specific background variables. Many of the identified patterns, such as the tendency to arrive at a correct response without using "search" on a spreadsheet (associated with age) and the number of correct numerical entries on the table construction problem (associated with years of education and numeracy) required a holistic inspection of the full action sequence. This showcases the utility of sequence-based feature extraction methods in identifying background-related long-term and overall sequential patterns, which action- or short-subsequence-based methods may not give rise to.

Implications

Quantifying the association between action sequence-derived features and external variables offers a data-driven perspective on the evaluation, design, and scoring of simulation-based assessments, which has implications for both measurement and career counseling. To start, evaluating the prediction accuracy of a particular background variable based on extracted features presents a generic way to quantify the informativeness of the action sequence on a task for that variable: If a subset of items is to be selected for constructing a short test that can best differentiate examinees on the variable, priority may be given to items whose sequence data show a higher association. This is most relevant when the variable of interest is an external criterion variable that the test intends to predict, e.g., job performance, rather than demographic background variables such as age or gender. Potential applications to the workforce include the construction of simulation-based assessments for personnel selection (Tippins, 2015) optimizing predictive validity of job

performance, as well as the prediction of readiness (in knowledge, skills, or abilities) for a job that fits an individual's vocational interest, a predictor of job performance, satisfaction, and commitment (Nye et al., 2012). In a separate analysis, PSTRE action sequence features were found predictive of several knowledge, skills, and abilities required for the participants' self-reported occupation (e.g., computers and electronics, reading comprehension, and judgment/decision making), when we linked the occupation category to O*NET expert ratings on knowledge, skill, and abilities required on that job (Fleisher & Tsacoumis, 2012). A brief graphical summary of the prediction results is provided in Figure E11 in ESM 1, Appendix III. This may see applications in predicting the gap between job-seekers' current capabilities and their occupational interests, which can inform the choice of targeted interventions for skill development and career readiness interventions.

However, test development goes beyond maximizing predictive validity. In particular, if differences in sequential patterns are found associated with a background variable, a test developer may inspect the specific sequential patterns that contribute to the association to understand its measurement implications. This evaluation depends on whether the sequential pattern is construct-relevant, examining which requires both interpretations of the sequential pattern and gathering of additional evidence. In the current study, we attempted to identify these patterns by interpreting the PLS components, and we subsequently looked at the relationship between the identified sequential pattern and either final response, response on other PSTRE tasks, or overall PSTRE proficiency to examine (1) whether the pattern affected scoring and (2) whether it appeared construct-relevant.

For example, age was found in the current study to be related to examinee's choice to either use a text-based drop-down menu or click a toolbar graphical icon for performing a step, e.g., creating a folder or sorting. This was found uncorrelated either with final item score or with overall PSTRE proficiency. Clicking an icon or navigating through the drop-down menu achieves the same goal, but the former was found to be less common among senior participants, which may indicate lower familiarity with toolbar icons for them. When such construct-irrelevant behaviors are correlated with external background variables, test developers need to minimize their impact on the examinees' chance of successfully solving the question, as such nuisance can compromise test fairness by producing differential item functioning. Demographic differences in how a key step is approached signify the importance of universal design principles for assessment interface design (see Steinfeld & Maisel, 2012): Whereas the availability of both drop-down menu and toolbar icons allowed participants from different age groups to perform a key step despite potential differences in familiarity with one option, the observations from item U19a suggested the potential need to allow zooming or font size adjustments. When this option is unavailable, for two equally capable participants, both not knowing how to identify information from a spreadsheet using search or sort, one who is younger might be less visually burdened by scanning through an entire spreadsheet with small font sizes, thus having a higher chance of correct response.

On the other hand, the exploratory findings suggest that some action sequence patterns associated with income, education, and ICT exposure, such as the use of tools for efficient problem-solving, were associated with individual differences on PSTRE proficiency, i.e., were construct-relevant, despite that efficiency was not considered in applicable items' final scoring. This finding concurred with those found from select PSTRE assessment items in prior studies (e.g., item U02 in Liao et al., 2019), where across tasks, individuals with higher income and years of education showed a tendency to utilize appropriate tools (e.g., keyboard shortcuts, spreadsheet searching/sorting) to facilitate efficient problem-solving, and when process data were used to improve PSTRE scoring precision (Zhang et al., 2023), the process-incorporated scoring algorithm picked up on such patterns. Scoring of open-ended questions is a nontrivial task and often requires specification of behavioral evidence that is indicative of the measured proficiency (Mislevy et al., 2003). Construct-relevant behavioral patterns can often contribute to more reliable scoring, as they provide additional proficiency information, but whether they should be incorporated into the scoring rule requires broader considerations, e.g., potential scoring consequences for different subgroups, which is partially addressed by examining how a pattern relates to test takers' external characteristics.

Limitations and Future Directions

Some limitations are worthwhile to merit discussions. First, all analyses were conducted on merely five out of 28 countries who participated in the PIAAC 2012 cycle, although they ranked at significantly different positions in PSTRE proficiency (OECD, 2016). These five countries are relatively high-income developed countries and areas, with small differences in sample proportion by PSTRE proficiency levels, which is identical as reported in previous studies by He et al. (2021). We thus caution against the generalizations of the empirical findings, both in terms of the sequence features' prediction of various external variables and in terms of the interpretations found on sequential

patterns to the general population of working-age adults worldwide. Second, the current analysis of adult digital problem-solving was only conducted on the PIAAC PSTRE assessment, which was specifically designed to assess problem-solving in personal, work, and civic contexts in three common digital platforms (web browser, spreadsheets, e-mail client) and was a low-stakes assessment. This could limit the generalizability of the findings to adult digital problem-solving behavior in general and to high-stakes situations. Third, while this study has shed light on various aspects of the relationship between background variables and sequential patterns, it is important to acknowledge its limitations with the PLS. The identification of sequential patterns through examination of original sequences ranked by PLS remains a speculative process; the variability in the PLS components could not always be fully accounted for by the identified patterns, suggesting that there may be other sequence characteristics that were not captured. In addition, the exploratory nature of our study, based on predictions and correlations, should be considered when interpreting the results – these findings should not be construed as establishing causal relationships or conclusions about interrelationships among the variables under investigation. Moreover, it is crucial to note that the relationships observed between scores or extracted features and demographic variables do not inherently validate the score or the process data. These associations are speculative in nature and should not be interpreted as confirmatory evidence for the validity of the measures used. Future research is needed to rigorously corroborate these preliminary interpretations.

Methodologically, there is also room for future development. To start, to afford a fair comparison across items, the dimension K of the MDS and Seq2seq features was set uniformly to a large number, 100, although the selection of optimal dimension K via cross-validation would be more plausible in practice. Second, only linear models were considered in the prediction of different background and cognitive variables despite that the relationship between certain features and the dependent variables could be nonlinear. While the current empirical results suggested that MDS tended to outperform Seq2seq in the action sequence features' prediction power of external traits, one cannot conclude that MDS in general outperforms Seq2seq in preserving sequence information, due to an assortment of potential methodological confounds, including the adoption of a linear prediction model, the researcher's degrees of freedom in choosing tuning parameters during feature extraction, and sequence characteristics (e.g., action variability, length) of the PSTRE data. Finally, behavioral patterns corresponding to the PLS components were identified by visually inspecting the ranked sequences, which was labor-intensive and speculative. As a direction for future research, visualization and explainable AI methods that aid with the interpretation of action sequence features may be developed.

Electronic Supplementary Materials

The electronic supplementary material is available with the online version of the article at https://doi.org/10.1027/2151-2604/a000554.

ESM 1. Appendix I: Feature Extraction with MDS and Seq2seq. Appendix II: MDS PLS component LOWESS plots. Appendix III: Supplementary tables and figures.

References

Abele, S., & von Davier, M. (2019). CDMs in vocational education: Assessment and usage of diagnostic problem-solving strategies in car mechatronics. In M. von Davier & Y. Lee (Eds.) *Handbook of diagnostic classification models: Models and model extensions, applications, software packages* (pp. 461–488). Springer. https://doi.org/10.1007/978-3-030-05584-4_22

AERA, APA, & NCME. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

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. https://doi.org/10.3102/1076998618784700

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

Cleveland, W. S. (1981). LOWESS: A program for smoothing scatterplots by robust locally weighted regression. *The American Statistician*, *35*(1), Article 54. https://doi.org/10.2307/2683591

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the Association for Information Science and Technology*, *41*(6), 391–407. https://doi.org/10.1002/(sici)1097-4571(199009)41:6<391::aid-asi1>3.0.co;2-9

Dillon, G., Boulet, J., Hawkins, R., & Swanson, D. (2004). Simulations in the United States Medical Licensing Examination™ (USMLE™). *BMJ Quality & Safety*, *13*(Suppl 1), i41–i45. https://doi.org/10.1136/qshc.2004.010025

Eichmann, B., Greiff, S., Naumann, J., Brandhuber, L., & Goldhammer, F. (2020). Exploring behavioural patterns during complex problem-solving. *Computer Assisted Learning*, *36*(6), 933–956. https://doi.org/10.1111/jcal.12451

Ercikan, K., Guo, H., & He, Q. (2020). Use of response process data to inform group comparisons and fairness research. *Educational Assessment*, *25*(3), 179–197. https://doi.org/10.1080/10627197.2020.1804353

Ericsson, K. A., & Simon, H. A. (1998). How to study thinking in everyday life: Contrasting think-aloud protocols with descriptions and explanations of thinking. *Mind, Culture, and Activity*, *5*(3), 178–186. https://doi.org/10.1207/s15327884mca0503_3

Fang, G., & Ying, Z. (2020). Latent theme dictionary model for finding co-occurrent patterns in process data. *Psychometrika*, *85*(3), 775–811. https://doi.org/10.1007/s11336-020-09725-2

Fleisher, M. S., & Tsacoumis, S. (2012). *O* NET analyst occupational skills ratings: Procedures update*. National Center for O* NET Development.

Friedman, J., Hastie, T., & Tibshirani, R. (2009). *glmnet: Lasso and elastic-net regularized generalized linear models. R package version, 1.4*.

Gómez-Alonso, C., & Valls, A. (2008, October). A similarity measure for sequences of categorical data based on the ordering of common elements. In *Proceedings of the 5th International Conference Modeling Decisions for Artificial Intelligence, MDAI 2008, Sabadell, Spain, October 30-31, 2008* (pp. 134–145). Springer.

Gao, Y., Cui, Y., Bulut, O., Zhai, X., & Chen, F. (2022). Examining adults' web navigation patterns in multi-layered hypertext environments. *Computers in Human Behavior*, *129*, Article 107142. https://doi.org/10.1016/j.chb.2021.107142

Gao, Y., Zhai, X., Bulut, O., Cui, Y., & Sun, X. (2022). Examining humans' problem-solving styles in technology-rich environments using log file data. *Journal of Intelligence*, *10*(3), Article8. https://doi.org/10.3390/jintelligence10030038

Greiff, S., Wüstenberg, S., & Avvisati, F. (2015). Computer-generated log-file analyses as a window into students' minds? A showcase study based on the PISA 2012 assessment of problem solving. *Computers & Education*, *91*, 92–105. https://doi.org/10.1016/j.compedu.2015.10.018

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

Hao, J., & Mislevy, R. J. (2019). Characterizing interactive communications in computer-supported collaborative problem-solving tasks: A conditional transition profile approach. *Frontiers in Psychology*, *10*, Article 1011. https://doi.org/10.3389/fpsyg.2019.01011

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*, Article 104170. https://doi.org/10.1016/j.compedu.2021.104170

He, Q., Borgonovi, F., & Suárez-Álvarez, J. (2023). Clustering sequential navigation patterns inmultiple-sourcereading tasks with dynamic time warping method. *Computer Assisted Learning*, *39*(3), 719–736. https://doi.org/10.1111/jcal.12748

He, Q., Shi, Q., Tighe, E. (2023). Predicting problem-solving proficiency with hierarchical supervised models on response process. *Psychological Test and Assessment Modeling*, *65*(1), 145–178. https://www.proquest.com/scholarly-journals/predicting-problem-solving-proficiency-with/docview/2799814448/se-2

He, Q., & von Davier, M. (2016). Analyzing process data from problem-solving items with N-grams. In Y. Rosen, S. Ferrara, & M. Mosharraf (Eds.), *Handbook of research on technology tools for real-world skill development* (pp. 750–777). IGI Global. https://doi.org/10.4018/978-1-4666-9441-5.ch029

He, Q., von Davier, M., & Han, Z. (2019). Exploring process data in problem-solving items in computer-based large-scale assessments: Case studies in PISA and PIAAC. In H. Jiao, R. W. Lissitz, & A. Van Wie (Eds.), *Data analytics and psychometrics: Informing assessment practices* (pp. 53–76). Information Age Publishing.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, *24*(6), 417, 441. https://doi.org/10.1037/h0071325

Kirsch, I., & Thorn, W. (2013). *Foreword: The programme for international assessment of adult competencies: An overview* [PIAAC Technical Report] (pp. 5–24). OECD.

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

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, *401*(6755), 788–791. https://doi.org/10.1038/44565

Liang, K., Tu, D., & Cai, Y. (2023). Using process data to improve classification accuracy of cognitive diagnosis model. *Multivariate Behavioral Research*, *58*(5):969–987. https://doi.org/10.1080/00273171.2022.2157788

Liao, D., He, Q., & Jiao, H. (2019). Mapping Background variables with sequential patterns in problem-solving environments: An investigation of United States adults' employment status in PIAAC. *Frontiers in Psychology*, *10*, Article 646. https://doi.org/10.3389/fpsyg.2019.00646

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

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). A brief introduction to evidence-centered design. *ETS Research Report Series*, *2003*(1), i–29. https://doi.org/10.1002/j.2333-8504.2003.tb01908.x

Nwakasi, C. C., Cummins, P. A., Mehri, N., Zhang, J., & Yamashita, T. (2019). *Problem solving in technology-rich environments, adult education and training, and income: An International Comparison Using PIAAC Data*. Grantee Submission.

Nye, C. D., Su, R., Rounds, J., & Drasgow, F. (2012). Vocational interests and performance. *Perspectives on Psychological Science*, *7*(4), 384–403. https://doi.org/10.1177/1745691612449021

OECD (2012). *Literacy, numeracy and problem solving in technology-rich environments: Framework for the OECD survey of adult skills*. https://www.oecd.org/skills/piaac/PIAAC%20Framework%202012–%20Revised%2028oct2013_ebook.pdf

OECD (2014). *PISA 2012 results: Creative problem solving: Students' skills in tackling real-life problems* (Vol. *V*). https://www.oecd.org/pisa/keyfindings/pisa-2012-results-volume-V.pdf

OECD (2016). *Technical report of the survey of adult skills (PIAAC)* (2nd ed.). https://www.oecd.org/skills/piaac/PIAAC_Technical_Report_2nd_Edition_Full_Report.pdf

Qiao, X., & Jiao, H. (2018). Data mining techniques in analyzing process data: A didactic. *Frontiers in Psychology*, *9*, Article 2231. https://doi.org/10.3389/fpsyg.2018.02231

Qiao, X., Jiao, H., & He, Q. (2023). Multiple-group joint modeling of item responses, response times, and action counts with the Conway-Maxwell-Poisson distribution. *Journal of Educational Measurement*, *60*(2), 255–281. https://doi.org/10.1111/jedm.12349

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

Steinfeld, E., & Maisel, J. (2012). *Universal design: Creating inclusive environments*. Wiley & Sons.

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. https://doi.org/10.1007/s11336-020-09708-3

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. https://doi.org/10.1111/bmsp.12203

Tang, X., Zhang, S., Wang, Z., Liu, J., & Ying, Z. (2021). Procdata: An r package for process data analysis. *Psychometrika*, *86*(4), 1058–1083. https://doi.org/10.1007/s11336-021-09798-7

Tippins, N. T. (2015). Technology and assessment in selection. *Annual Review of Organizational Psychology and Organizational Behavior*, *2*(1), 551–582. https://doi.org/10.1146/annurev-orgpsych-031413-091317

Ulitzsch, E., He, Q., & Pohl, S. (2022). Using sequence mining techniques for understanding incorrect behavioral patterns on interactive tasks. *Journal of Educational and Behavioral Statistics*, *47*(1), 3–35. https://doi.org/10.3102/10769986211010467

Ulitzsch, E., He, Q., Ulitzsch, V., Molter, H., Nichterlein, A., Niedermeier, R., & Pohl, S. (2021). Combining clickstream analyses and graph-modeled data clustering for identifying common response processes. *Psychometrika*, *86*(1), 190–214. https://doi.org/10.1007/s11336-020-09743-0

Ulitzsch, E., Ulitzsch, V., He, Q., & Lüdtke, O. (2023). A machine learning-based procedure for leveraging clickstream data to investigate early predictability of failure on interactive tasks. *Behavior Research Methods*, *55*(3), 1392–1412.

Wang, Z., Tang, X., Liu, J., & Ying, Z. (2023). Subtask analysis of process data through a predictive model. *British Journal of Mathematical and Statistical*, *76*(1), 211–235. https://doi.org/10.1111/bmsp.12290

Wehrens, R., & Mevik, B. H. (2007). The PLS package: principal component and partial least squares regression in R. *Journal of Statistical Software*, *18*(2), 1–23. https://doi.org/10.18637/jss.v018.i02

Wold, S., Sjöström, M., & Eriksson, L. (2002). Partial least squares projections to latent structures (PLS) in chemistry. In *Encyclopedia of computational chemistry* (Vol. 3). John Wiley and Sons. https://doi.org/10/02/0470845015.cpa012

Xiao, Y., He, Q., Veldkamp, B., & Liu, H. (2021). Exploring latent states of problem-solving competence using hidden Markov model on process data. *Computer Assisted Learning 37*(5), 1232–1247. https://doi.org/10.1111/jcal.12559

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

Xu, H., Fang, G., Chen, Y., Liu, J., & Ying, Z. (2018). Latent class analysis of recurrent events in problem-solving items. *Applied Psychological Measurement*, *42*(6), 478–498. https://doi.org/10.1177/0146621617748325

Xu, H., Fang, G., & Ying, Z. (2020). A latent topic model with Markov transition for process data. *British Journal of Mathematical and Statistical*, *73*(3), 474–505. https://doi.org/10.1111/bmsp.12197

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

Zhang, S., Wang, Z., Qi, J., Liu, J., & Ying, Z. (2023). Accurate assessment via process data. *Psychometrika*, *88*(1), 76–97. https://doi.org/10.07/s11336-022-09880-8

Zumbo, B. D., & Hubley, A. M. (2017). *Understanding and investigating response processes in validation research* (Vol. 26). Springer.

History
Received November 1, 2022
Revision received June 23, 2023
Accepted October 5, 2023
Published online April 24, 2024

Funding
This work has been supported by the National Science Foundation (#SES-1826540, #SES-2119938 and DMS-2310664) and Institute of Education Sciences, U.S. Department of Education, through Grant IES R305A210344 to Georgetown University.

ORCID
Susu Zhang
https://orcid.org/0000-0003-0751-6467
Xueying Tang
https://orcid.org/0000-0002-9774-2523

Jingchen Liu
Department of Statistics
Columbia University
New York, NY 10027
USA
jcliu@stat.columbia.edu

Appendix I: Feature Extraction with MDS and Seq2seq

Multidimensional Scaling

The purpose of MDS is to find a multidimensional numeric representation of the observed action sequence that can best preserve the pairwise dissimilarities between individuals (Tang et al., 2020). Denote the action sequence of individual $i \in \{1, \ldots, N\}$ on an item by $\mathbf{s}_i$, MDS finds a mapping of each $\mathbf{s}_i$ to a K-dimensional numerical representation $\boldsymbol{\theta}_i \in \mathbb{R}^K$ which minimizes

$$\sum_{i=1}^{N} \sum_{j=i+1}^{N} (d_{ij} - \|\boldsymbol{\theta}_i - \boldsymbol{\theta}_j\|)^2, \tag{1}$$

where $d_{ij} = d(s_i, s_j)$ is the dissimilarity between $\mathbf{s}_i$ and $\mathbf{s}_j$, the action sequences by examinees i and j, according to some predefined distance measure $d(\cdot)$. $\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_j\| = \sqrt{(\boldsymbol{\theta}_i - \boldsymbol{\theta}_j)'(\boldsymbol{\theta}_i - \boldsymbol{\theta}_j)}$ is the Euclidean distance between the MDS representations of examinees i and j, $\boldsymbol{\theta}_i$ and $\boldsymbol{\theta}_j$. Intuitively, MDS transforms each observation into a point in the K-dimensional Euclidean space, so that observations that are similar to each other (i.e., low on d_{ij}) remain similar in the Euclidean space (i.e., low on $\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_j\|$), and observations that are dissimilar are farther apart in the Euclidean space after the transformation.

Crucial to MDS is the choice of an appropriate distance measure, d, to capture the dissimilarity between individual observations. When observations are action sequences on interactive items, Tang et al. (2020) proposed to use an order-based sequence similarity measure (OSS; Gómez-Alonso & Valls, 2008), which allows for the quantification of the dissimilarity between ordered, categorical, variable-length event sequences. For examinee i with action sequence $\mathbf{s}_i = (s_{i1}, \ldots, s_{iL_i})$, let s_{it} be the tth action performed by the subject, L_i be the total number of actions, and L_i^a be the number of occurrences of a particular action a. The dissimilarity between the action sequences of any two examinees, d_{ij}, is given

https://doi.org/10.1027/2151-2604/a000554

by

$$\begin{aligned} d_{ij} &= \frac{f(\mathbf{s}_i, \mathbf{s}_j) + g(\mathbf{s}_i, \mathbf{s}_j)}{L_i + L_j}, \text{ where} \\ f(\mathbf{s}_i, \mathbf{s}_j) &= \frac{\sum_{a \in C_{ij}} \sum_{m=1}^{K_{ij}^a} |s_i^a(m) - s_j^a(m)|}{\max\{L_i, L_j\}}, \text{ and} \\ g(\mathbf{s}_i, \mathbf{s}_j) &= \sum_{a \in U_{ij}} L_i^a + \sum_{a \in U_{ji}} L_j^a. \end{aligned}$$

Here, C_{ij} is the set of common actions occurred in both sequences, $K_{ij}^a = \min\{L_i^a, L_j^a\}$, $s_i^a(m)$ is the serial position of the mth occurrence of event a, and U_{ij} is the set of unique actions taken by examinee i but not by j. Intuitively, $f(\mathbf{s}_i, \mathbf{s}_j)$ quantifies how the serial positions of common actions in the two examinees' action sequences differ, and $g(\mathbf{s}_i, \mathbf{s}_j)$ counts the number of actions unique to each examinee. In turn, the dissimilarity between i and j, d_{ij}, takes into account the differences both in the ordering of the same actions and the types of actions taken.

After calculating the dissimilarities (d_{ij}) between each pair of individuals' action sequences, the optimization problem that minimizes Equation (1) is solved to find $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N$. As the transformed $\boldsymbol{\theta}$'s preserve as much as possible the pairwise dissimilarities in the observed action sequences, $\boldsymbol{\theta}_i$ could be seen as a $K-$dimensional latent feature vector which contains information of the original action sequence $\mathbf{s}_i$.

Sequence-to-sequence autoencoder

Seq2seq is another method that extracts K-dimensional numerical features from the action sequence $\mathbf{s}_i$ (Tang et al., 2021). Commonly used for information compression from phrases or sentences in natural languages, an autoencoder seeks to encode categorical event sequences into lower-dimensional latent vectors, which can then be used to restore the original sequences. A Seq2seq autoencoder is an artificial neural network with two main components, an encoder function $\phi(\cdot)$ that transforms the original input sequence $\mathbf{s}_i$ into a fixed-dimensional latent vector $\boldsymbol{\theta}_i$, and a decoder function $\psi(\cdot)$ that maps the latent vector $\boldsymbol{\theta}_i$ to a reconstructed version of the original sequence, $\hat{\mathbf{s}}_i$. The action sequence feature extraction procedures proposed in Tang et al. (2021) employed a recurrent neural

network-based autoencoder with multiple hidden layers. The model is structured as follows.

In the encoding stage, each action in the sequence is mapped to a $K-$dimensional continuous representation (i.e., an embedding, $\mathbf{e}_{it}$) based on the surrounding context using an embedding layer (Mikolov et al., 2013). Following the action embedding, a recurrent layer is applied to the embedded sequences. The recurrent layer creates a $K-$dimensional hidden state ($\boldsymbol{\theta}_{it}$) for each time step $t = 1, \ldots, L_i$. The hidden state at time t, $\boldsymbol{\theta}_{it}$ is a function of the hidden state at the previous time step, $\boldsymbol{\theta}_{it-1}$, and the embedded action performed at time t, $\mathbf{e}_{it}$, that is, $\boldsymbol{\theta}_{it} = f(\boldsymbol{\theta}_{it-1}, \mathbf{e}_{it})$. Different choices for the recurrent function, $f(\cdot)$, have been proposed, including the long short-term memory (LSTM; Hochreiter & Schmidhuber, 1997) architecture and the gated recurrent unit (GRU; Cho et al., 2014) architecture. For feature extraction from action sequences, Tang et al. (2021) adopted the GRU for the recurrent function, which learns from the observed data to update or reset the hidden states depending on the context and the action type. In this way, the recurrent states ($\boldsymbol{\theta}_{i1}, \ldots, \boldsymbol{\theta}_{iL_i}$) accumulate information in the action sequence over time, and the hidden state at the final time step, $\boldsymbol{\theta}_{iL_i}$, summarizes the information contained in the entire action sequence. We simplify the notation for the last recurrent state (i.e., the encoder output, $\boldsymbol{\theta}_{iL_i}$) to $\boldsymbol{\theta}_i$, which will be used to reconstruct the observed sequence in the decoding stage.

The first layer of the decoding stage is also a recurrent layer with L_i time steps, with initial hidden state $\mathbf{y}_{i1} = \mathbf{0}$ and the tth hidden state obtained by $\mathbf{y}_{it} = f'(\mathbf{y}_{it-1}, \boldsymbol{\theta}_i)$. Here, f' is another recurrent function with a different set of parameters from f. Note that the decoder recurrent states are obtained by feeding in the same encoder output, $\boldsymbol{\theta}_i$, for L_i times and accumulating the information over time through $f'(\cdot)$. The last layer of the decoder is a softmax layer, where, for each time point $1 \leq t \leq L_i$, the decoder hidden state at time t ($\mathbf{y}_{it}$) is used to predict the probability that individual i takes each possible action. Intuitively, the decoder aims at reconstructing the distribution of the original action sequence $\mathbf{s}_i$ using the encoder output $\boldsymbol{\theta}_i$.

A graphical illustration of the Seq2seq is presented in Figure 1. Action sequence features extracted with Seq2seq are $\boldsymbol{\theta}_i$, the last state of the encoder which summarizes the information throughout the entire sequence and is used in the decoder to reconstruct the original sequence.

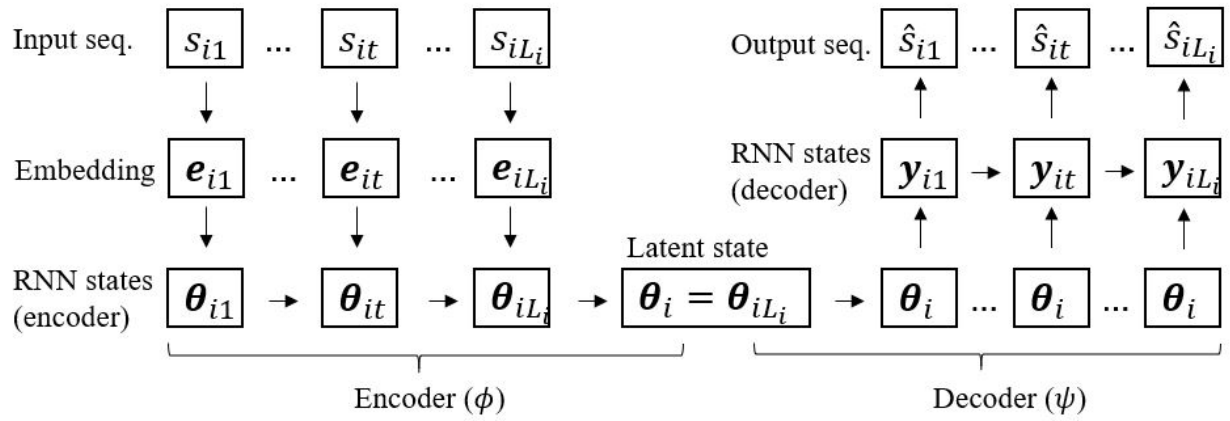

Figure E1

An illustration of the structure of the sequence-to-sequence autoencoder.

References

Cho, K., Van Merriënboer, B., Bahdanau, D., & Bengio, Y. (2014). On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.

Gómez-Alonso, C., & Valls, A. (2008). A similarity measure for sequences of categorical data based on the ordering of common elements. *International Conference on Modeling Decisions for Artificial Intelligence*, 134–145.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

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.

Appendix II: MDS PLS component LOWESS plots

Age

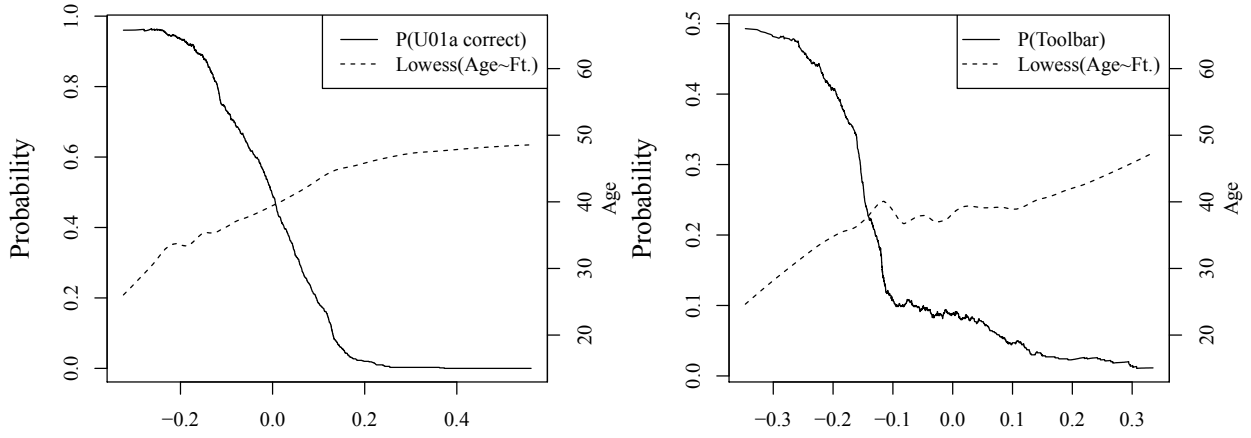

Figure E2

Distribution of U01a PLS component scores with respect to examinees' age. Dashed lines: LOWESS plot of age against the PLS component score (y-axis ticks on the right); solid lines: LOWESS plot for sequence pattern against PLS score (y-axis ticks on the left).

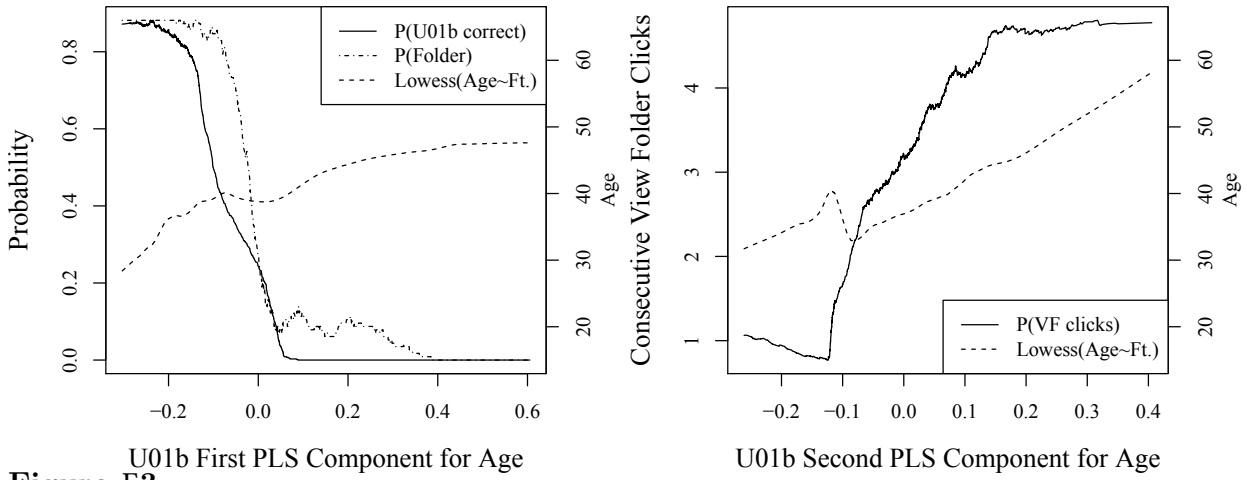

Figure E3

Distribution of U01b PLS component scores with respect to examinees' age.

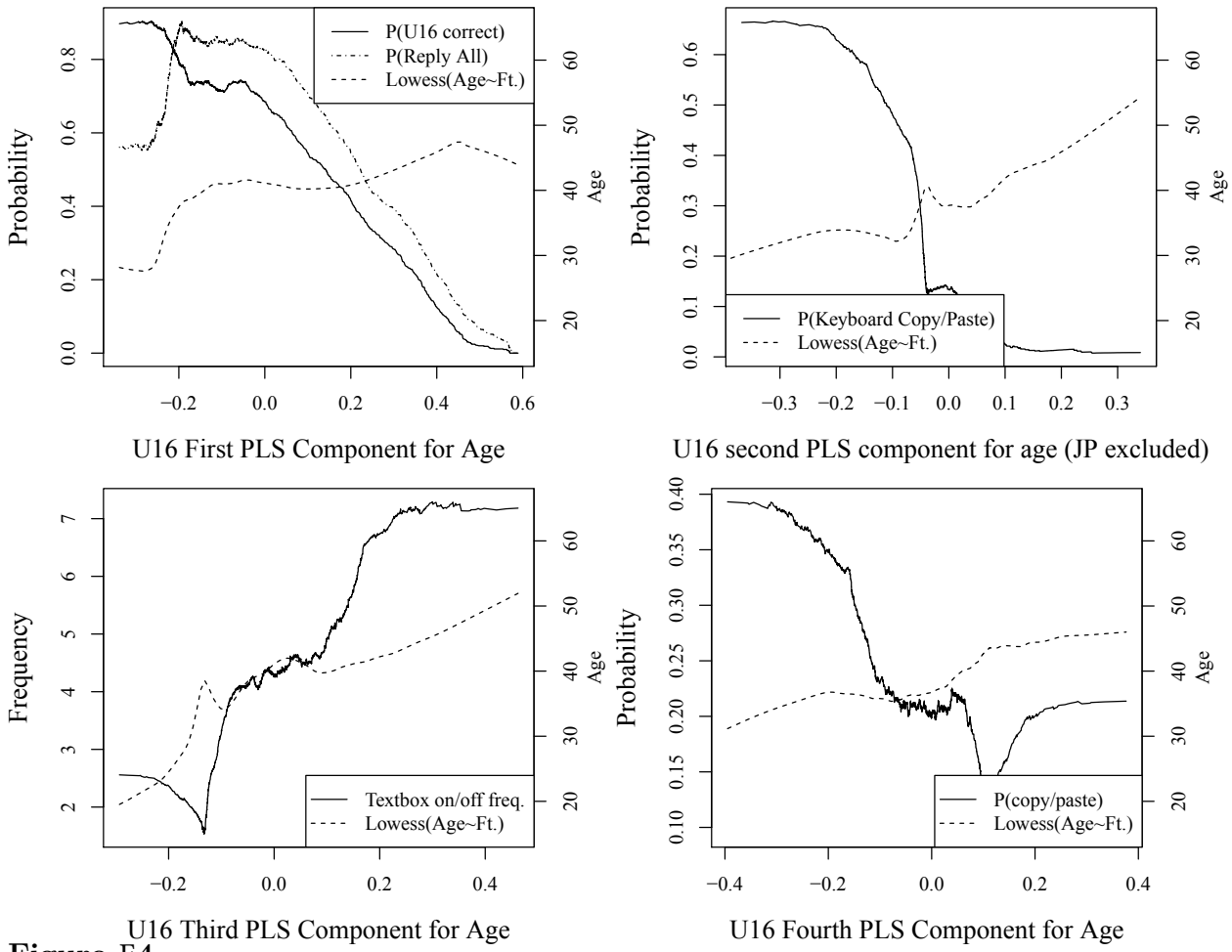

Figure E4

Distribution of U16 PLS component scores with respect to examinees' age.

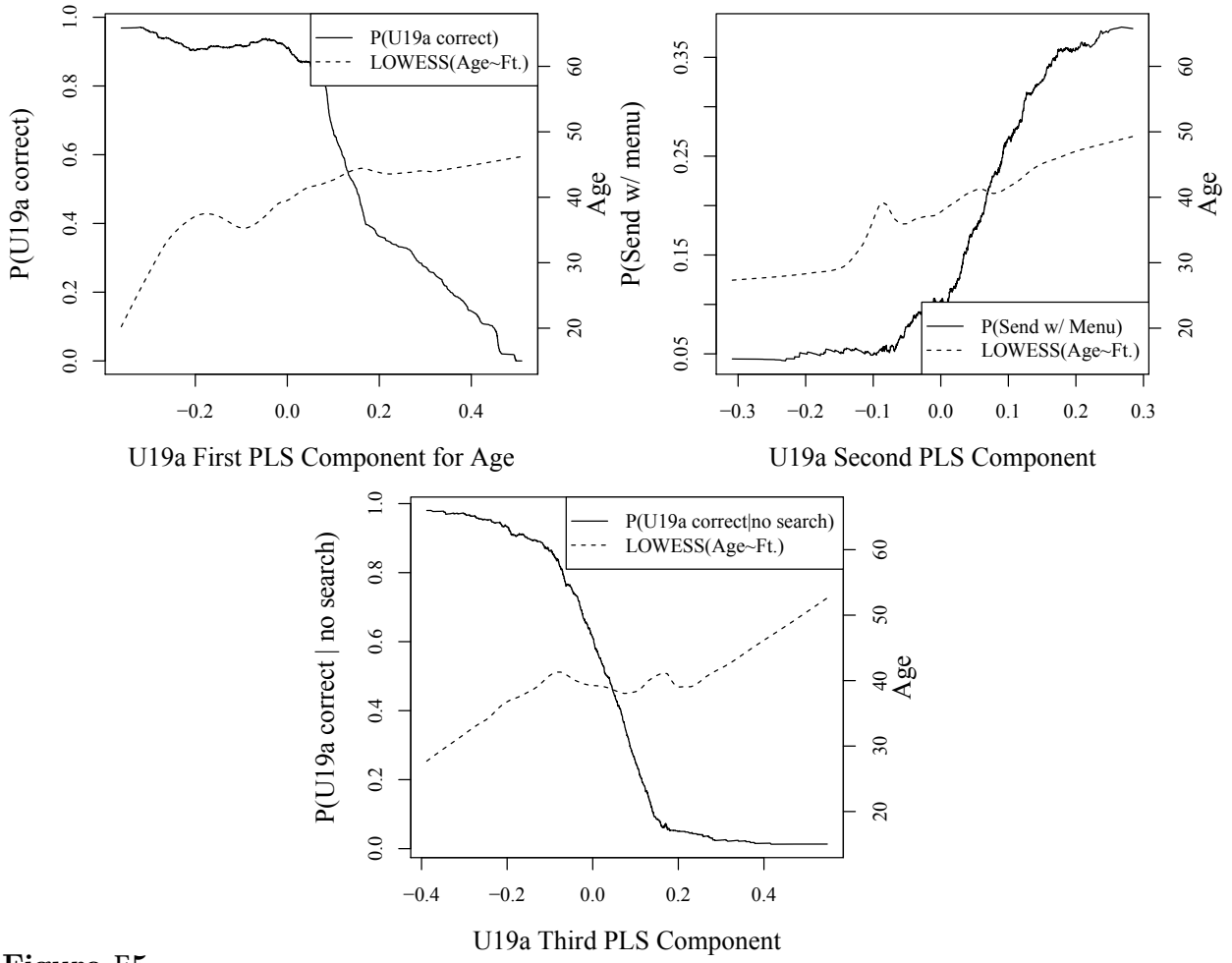

Figure E5

Distribution of U19a PLS component scores with respect to examinees' age.

Income

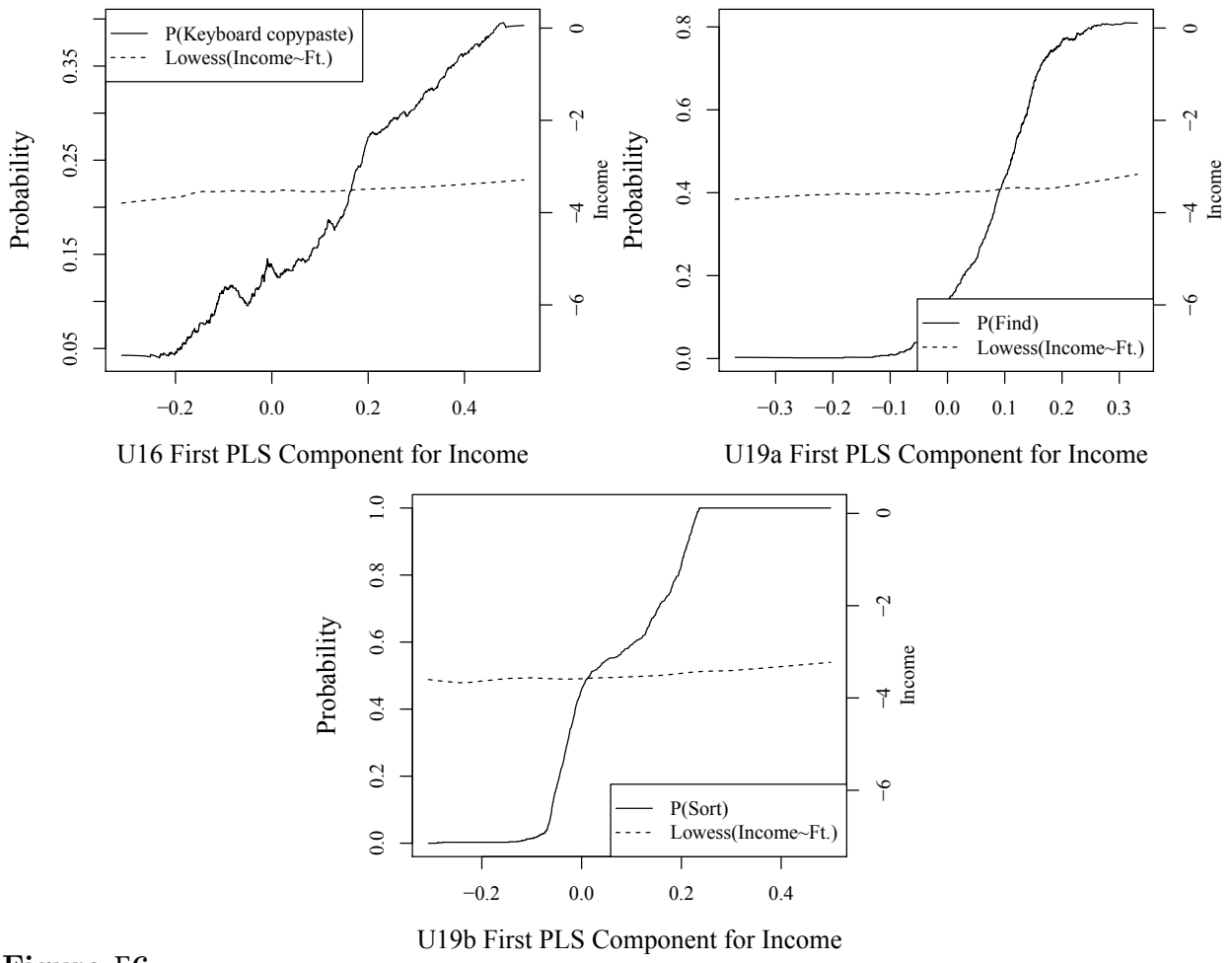

Figure E6

Distribution of PLS component scores with respect to examinees' income on items U16, U19a, and U19b.

ICTHome

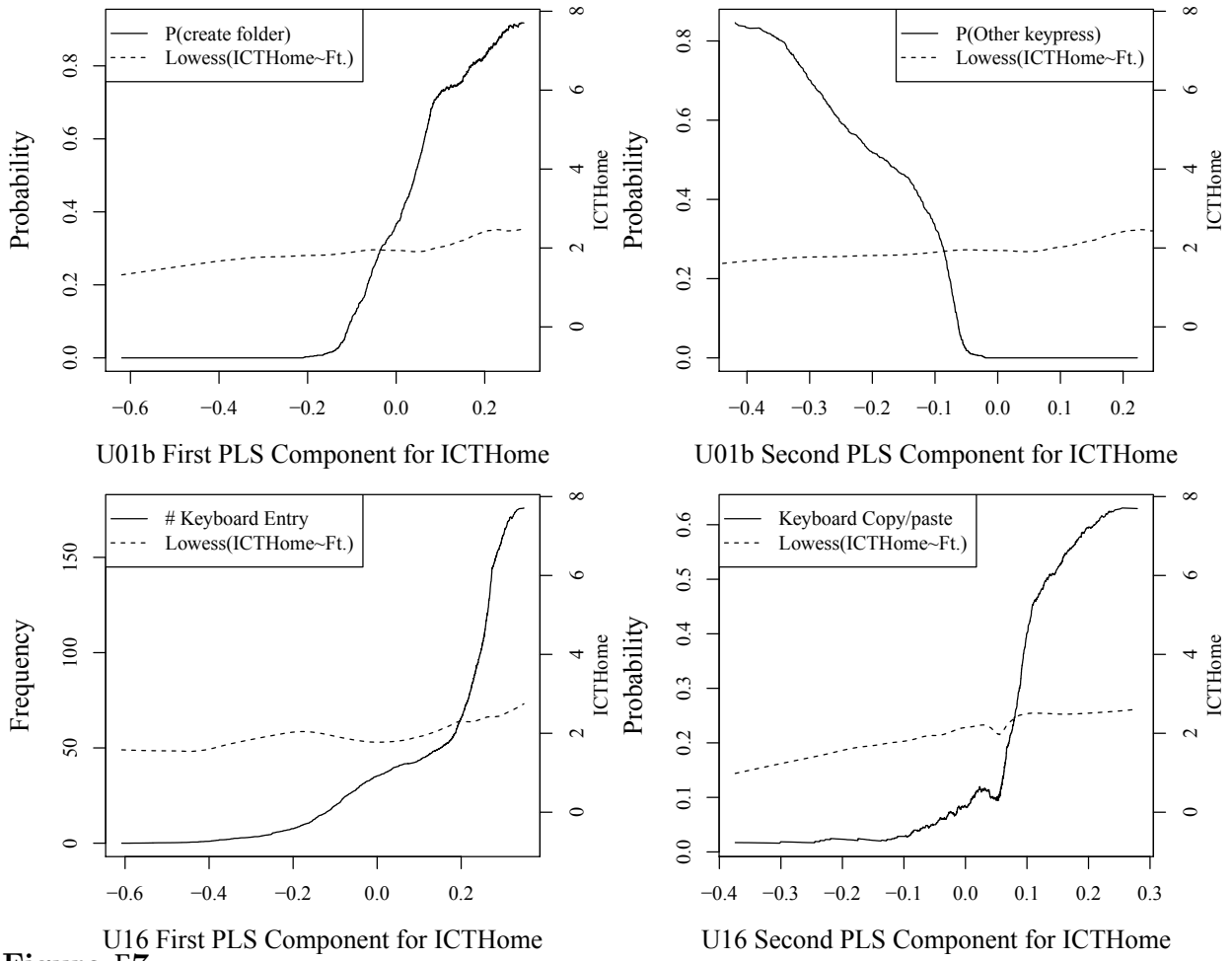

Figure E7

Distribution of PLS component scores with respect to examinees' ICT usage at home on items U01b and U16.

ICTWork

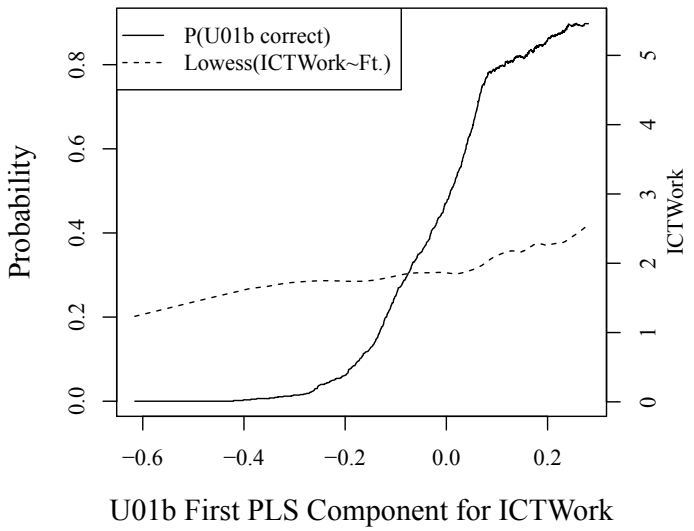

Figure E8

Distribution of PLS component scores with respect to examinees' ICT usage at work on item U01b.

YRSEdu

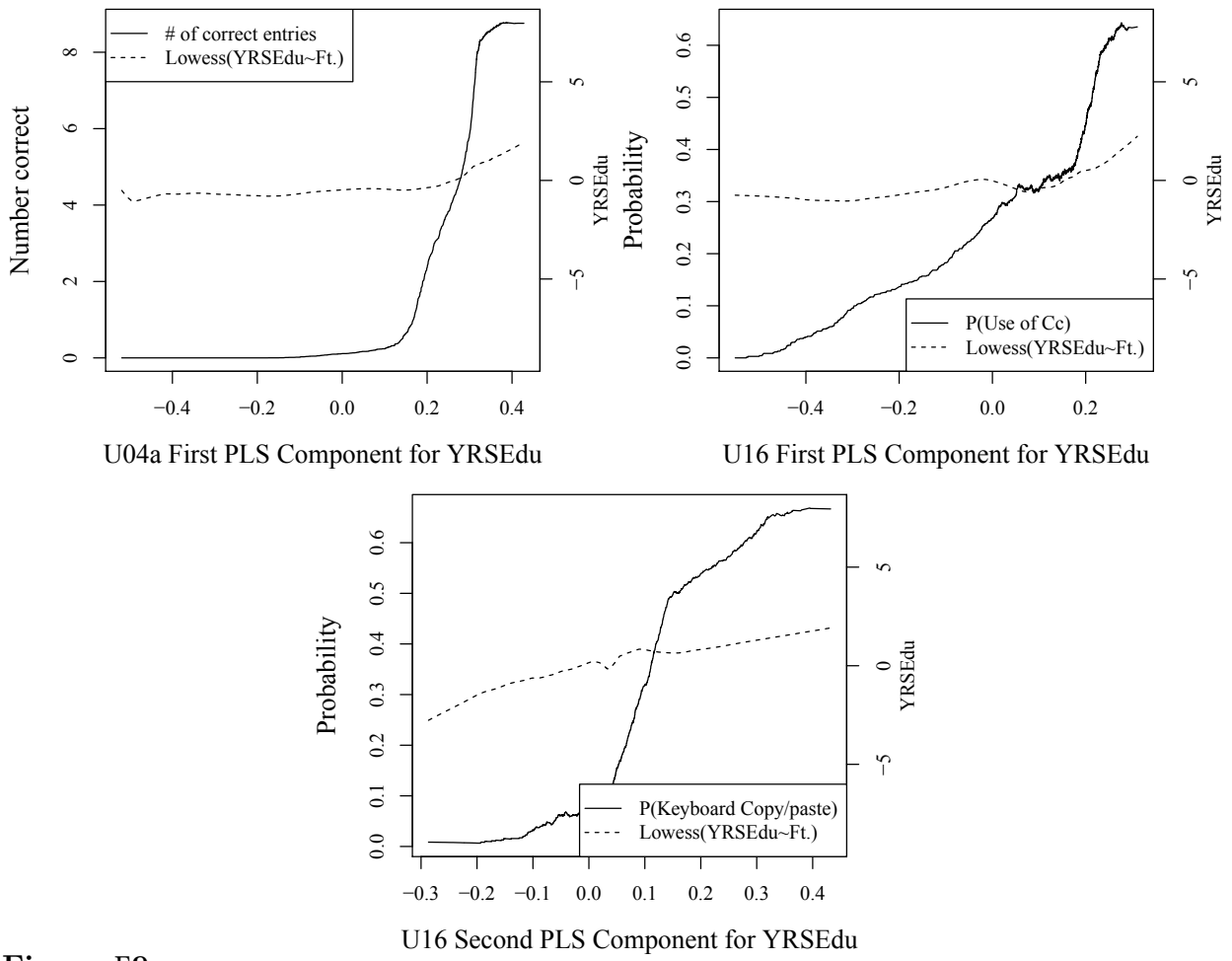

Figure E9

Distribution of PLS component scores with respect to examinees' education level on items U04a and U16.

Numeracy

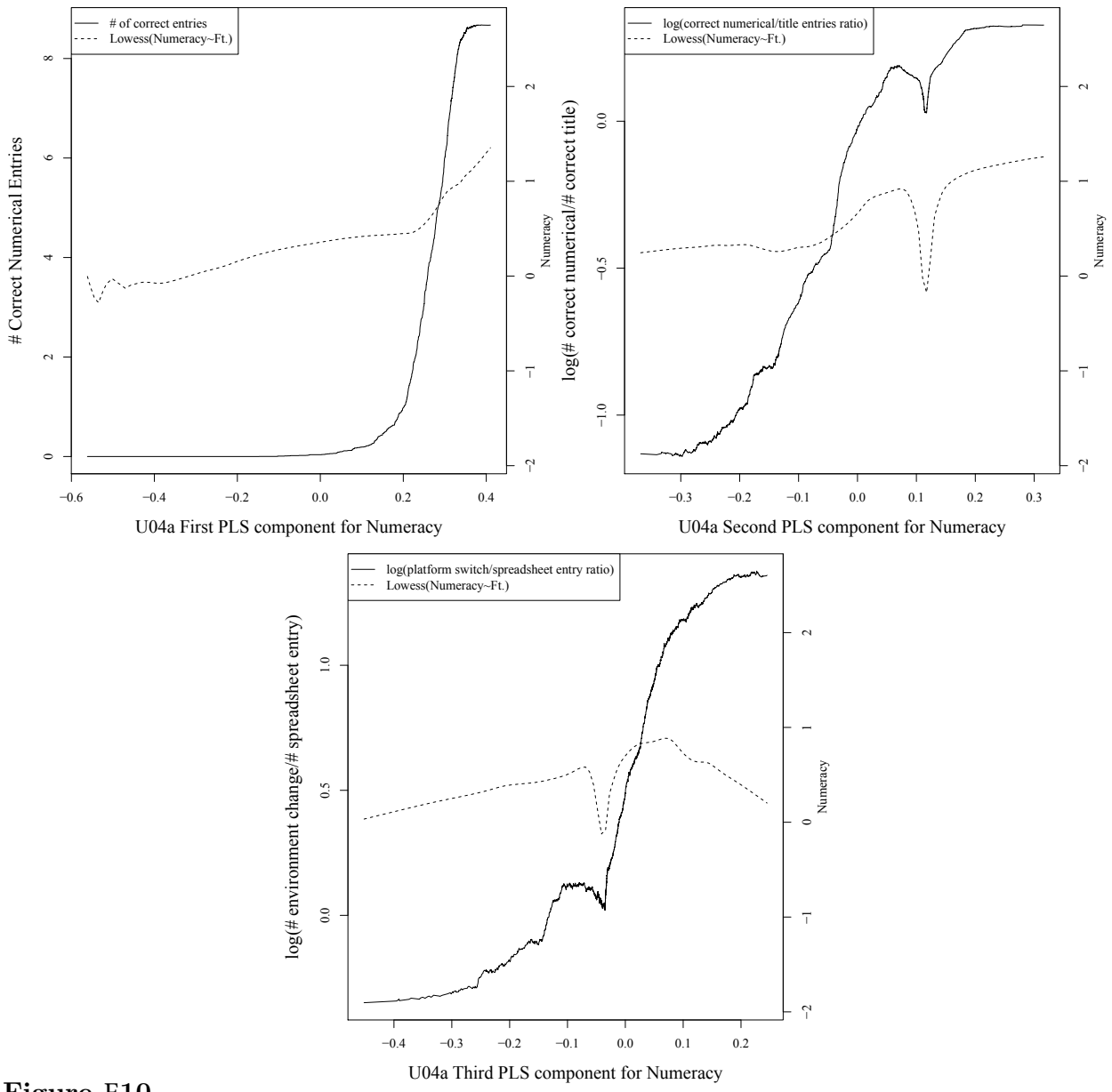

Figure E10

Distribution of PLS component scores with respect to examinees' numeracy proficiency on item U04a.

Appendix III: Supplementary tables and figures

Country or Region	ICTHome	ICTWork	log(income)	YRSEdu
GB	0.08	0.39	0.39	0.07
IE	0.09	0.42	0.45	0.00
JP	0.09	0.36	0.34	0.00
NL	0.01	0.31	0.37	0.00
US	0.10	0.37	0.41	0.13

Table E1

Missingness proportion on background variables by country or region.

Figure E11

*Out-of-sample correlation between observed and predicted values of the participant's occupation's knowledge (K), skill (S), and ability (A) requirements, based on O*NET expert ratings. Pink bars are the out-of-sample correlations based on 14 PSTRE items' final responses. Pink and orange bar combined are the out-of-sample correlations based on 14 items' MDS features (first 20 principal components) combined.*