Leveraging Process Data to Assess Adults' Problem-Solving Skills: Using Sequence Mining to Identify Behavioral Patterns across Digital Tasks

Qiwei He

Educational Testing Service, 660 Rosedale Road, Princeton, NJ08541, USA

qhe@ets.org

Francesca Borgonovi

Social Research Institute, University College London, 55-59 Gordon Square, London, WC1H 0NU, United Kingdom; f.borgonovi@ucl.ac.uk

Marco Paccagnella

Organisation for Economic Co-operation and Development, 2 rue André Pascal, 75775 Paris Cedex 16, France; marco.paccagnella@ucl.ac.uk

Abstract

This paper illustrates how process data can be used to identify behavioral patterns in a computer-based problem-solving assessment. Using sequence-mining techniques, we identify patterns of behavior across multiple digital tasks from the sequences of actions undertaken by respondents. We then examine how respondents' action sequences (which we label "strategies") differ from optimal strategies. In our application, optimality is defined ex-ante as the sequence of actions that content experts involved in the development of the assessment tasks identified as most efficient to solve the task given the range of possible actions available to test-takers. Data on 7,462 respondents from five countries (the United Kingdom, Ireland, Japan, the Netherlands, and the United States) participating in the Problem Solving in Technology-Rich Environment (PSTRE) assessment, administered as part of the OECD Programme for the International Assessment of Adult Competencies (PIAAC), indicate that valuable insights can be derived from the analysis of process data. Adults who follow optimal strategies are more likely to obtain high scores in the PSTRE assessment, while low performers consistently adopt strategies that are very distant from optimal ones. Very few high performers are able to solve the items in an efficient way, i.e. by minimizing the number of actions and by avoiding undertaking unnecessary or redundant actions. Women and adults above the age of 40 are more likely to adopt sub-optimal problem-solving strategies.

Keywords:

problem-solving skills, process data, longest common subsequence, PIAAC, sequence mining.

Leveraging Process Data to Assess Adults' Problem-Solving Skills: Using Sequence Mining to Identify Behavioral Patterns across Digital Tasks

1. Introduction

The ability to solve complex problems has been identified as a critical skill for success in 21st century societies and labor markets (Organisation for Economic Co-operation and Development [OECD], 2009, 2011, 2012; Schleicher, 2008).The Survey of Adult Skills, a product of the Programme for the International Assessment of Adult Competencies (PIAAC), was the first international large-scale survey to be predominantly administered on a computer and to assess adults' problem solving in technology-rich environments (PSTRE). PIAAC surveyed representative samples of adults in more than 30 countries, assessing their information-processing skills (in literacy, numeracy and PSTRE) and collecting a wealth of background information on their education, labor market experience, and engagement with tasks requiring information-processing skills.

PSTRE is defined as the ability to "use digital technology, communication tools and networks to acquire and evaluate information, communicate with others and perform practical tasks", and "to solve problems for personal, work and civic purposes by setting up appropriate goals and plans, accessing and making use of information through computers and computer networks" (OECD, 2009).

Computer delivery of the PSTRE assessment allowed the collection of information on how respondents interacted with the testing interface, including how much time they spent on each item and which actions they took to solve them. This information is recorded in log files and is commonly referred to as "process data" (He & von Davier, 2015, 2016).

In this paper we make use of process data from the PSTRE assessment to characterize problem-solving strategies in ways that generalize across tasks building. For this exploratory study, we use PSTRE data from five high-income countries. Cross-cultural differences in the use of different cognitive processes during reasoning and problem solving have been observed, leading researchers to hypothesize that socio-historical traditions have the potential to lead to differences in cognition. For example, Nisbett (2003) illustrated how individuals in East Asia tend to think holistically, dialectically, and on the basis of their experience, whereas Westerners tend to think analytically, logically, and abstractly. Because of brain plasticity, cultural differences can lead to differences in brain activity (Han et al., 2013; Han & Ma, 2014; Huttenlocher, 2002) and behavioral responses to certain stimuli across countries (Choi et al., 1999; Han et al., 2013; Markus & Kitayama, 2010; Masuda, & Nisbett, 2001). Similarly, differences in the experiences that individuals have because of age, gender and the use of digital technologies could lead to differences in how an individual approach and consequently solves problems. While experimental results suggest that cross-cultural differences in reasoning exist, it remains to be established if differences in reasoning reflect generalized differences in cognitive processes or, rather, the ease with which individuals adopt particular strategies but remain able to mobilize the use of other strategies when in need (Nisbett, 2003). The use of process data from an interactive problem-solving assessment can become a powerful tool to better identify and characterize problem-solving processes. For example, it can reveal if individuals engage in under- or over-exploration of the problem space before providing an answer and how consistent they are in these behaviors across problem tasks that require the mobilization of different skills.

1.1. The Value of Process Data in Large-Scale Assessments

Data stored in log files, referred to as process data in the present study, contain information on the actions undertaken by test takers via the computer interface and when each action occurred [1]. Process data were originally collected to cross-validate response data but are being increasingly used to explore respondents' test-taking behaviors. For example, they have been used to infer students' reading strategies (Lee & Jia, 2014; Goldhammer et al., 2013); to traack students' inquiry-based learning processes (Lämsä et al., 2020); to identify problem solving strategies (Goldhammer et al., 2014; He et al., 2019); and to analyze the relationship between problem-solving strategies and background characteristics (Liao et al., 2019; He et al., 2018, 2019). A review of recent studies based on log file data from PIAAC can be found in Goldhammer et al. (2020).

[1] In principle, all interactions between the participant and the computer delivery platform are actions. For example, any time a mouse is clicked and any time the participant types something the participant engages is an action. However, not all actions were recorded and stored in PIAAC log files. Only actions that were meaningful for scoring an item were recorded.

Process data provide information that cannot be easily observed from response data (Guo et al., 2018; Zhu & Feng, 2015; Liu et al., 2018; Chen et al., 2019), and that can be used to characterize the approaches respondents engage in when responding to questions on a test (Goldhammer et al., 2013; Hahnel et al., 2014). Once such approaches are identified, researchers can see if they differ across different population subgroups (defined in terms of age, gender, country, socioeconomic background, migration background, or educational attainment; see for instance Eichmann et al., 2020; Liao et al., 2019; He & von Davier, 2016; Ercikan et al., 2020).

Process data have been used extensively in the educational literature in the context of educational data mining, learning analytics and artificial intelligence. For example, process data on patterns of keystrokes has been used to examine writing behavior (e.g., Deane et al., 2020; Guo et al., 2020). Process data has also been used to examine if learning can be fostered through the use of AI-operated instructional environments (e.g., Biswas et al, 2005; 2010); to monitor the development of problem-solving skills in game-based assessments (e.g., Shute et al., 2016), and how AI systems can be used to measure human intelligence (e.g., Embreston, 2014). Furthermore, the educational literature has explored how process data can be used to develop tools and methods that enable efficient processing of educational data (e.g., Paquette et al., 2020; Baker & Inventado, 2014).

1.2. Extracting Information from Process Data

Although there is broad agreement on the potential of process data, no consensus has yet been reached on how best to analyze them. Various disciplines can contribute appropriate methods, such as data mining, machine learning, natural language processing, social network analysis. These methods are typically applied when handling “big data” with complex structures.

Recent studies of process data from large-scale assessments rely on n-gram analysis to extract key actions (or subsequences of actions) associated with final item responses (He & von Davier, 2016; Stadler et al., 2019; He et al., 2018). Other studies have developed new measurement models for occurrences of actions to assess the latent traits of interest (LaMar, 2018; Liu et al., 2018), or to examine actions as recurring events (Xu et al., 2018). Finally, dissimilarity measurement has been used to extract latent variables from sequences of actions (Tang et al., 2020). These sequence-derived latent features can accurately predict the final responses of test-takers, as well as performance on other items and various cognitive traits.

Despite rapid methodological developments, the study of how best behavioral patterns can be compared across multiple test items and how such patterns can be meaningfully interpreted is still at a preliminary stage. A major challenge is that features derived from process data are usually item dependent and is therefore difficult to construct variables that summarize behavior across items. This is especially the case for problem-solving items since they make extensive use of scenario-based designs. Methods that rely on disassembled variables such as n-grams (He & von Davier, 2016) or sequential factors in Hidden Markov Models (LaMar, 2018) are unlikely to be appropriate for the study of behaviors across items because such methods are generally highly specific to individual test items.

By investigating patterns of behaviors across multiple tasks that are embedded in various contexts and scenarios, we innovate on previous research that has examined respondents' problem-solving behaviors focusing only on single items or tasks (Stadler et al., 2019; Liao et al., 2019; Xu et al., 2018; Greiff et al., 2015).

Our approach for identifying behavioral patterns that are comparable across items relies on indicators based on the distance between the observed sequence of actions performed by respondents and “optimal” sequences, i.e. strategies that item developers and subject domain experts defined ex-ante as being optimal ways to solve the task.

1.3. Distance Function over Sequences

Sequential pattern mining is a data mining method designed to deal with sequential data. It extracts useful sequential patterns from a large set of sequences (Kour, 2017; Sharda et.al, 2018). A sequence database consists of sequences of ordered elements or events recorded with or without time-stamped information (Febrer-Hernandez & Hernardze-Palancar, 2012). Sequence distance functions are designed to measure sequence (dis)similarities in sequence mining. A common approach for detecting patterns in action sequences consists of converting the information contained in action sequences into distance measures (Dong & Pei, 2007). In the context of problem-solving processes, sequence distance measures can be defined to describe how action sequences differ either from each other (Tang et al., 2020) or with respect to pre-defined sequences (Hao et al., 2015; He et al., 2019). In the application of this paper, a set of pre-defined sequences were defined ex-ante by content experts involved in the design of the assessment tasks. These sequences identify optimal path(s) that test takers could follow to correctly solve a problem-solving task, i.e. they represent sequences consisting of the minimum number of actions a test taker could take given no prior knowledge of the solution space.

Character alignment-based distance functions are broadly used in sequence (dis)similarity metrics. These algorithms can be local window based or whole sequence based; they can also be edit distances or more general pairwise similarity score-based distance. For instance, the edit distance function, also called the Levenshtein distance (Levenshtein, 1965, 1966), between two sequences S_1 and S_2 defines the minimum number of edit operations (i.e., deletion, insertion and substitution) that are needed to transform S_1 into S_2 (Jurafsky & Martin, 2000). The Hamming distance between two sequences is limited to cases where the two sequences have identical lengths and is defined as the number of positions where the two sequences are different (Hamming, 1950). The Longest Common Subsequence algorithm (LCS; e.g., Hirschberg, 1975, 1977) that will be used in the current study identifies the longest subsequence that is common to two given strings. The length of the LCS is defined as the degree of closeness between the two strings. This metric was first introduced in the educational assessment literature by Sukkarieh et al. (2012), who used the LCS to detect careless errors in respondent-provided sequences to analyze the scoring of technology-based tasks in PIAAC. Tang et al. (2020) employed a dissimilarity measure first developed by Gomez-Alonso and Valls (2008) to quantify both the dissimilarity among the actions from two sequences and the count of actions that occurred uniquely in one sequence only.

BLAST and PSI-BLAST (Altschul et al., 1997), FASTA (Pearson & Lipman, 1988; Pearson, 1990), SSEARCH (Smith & Waterman, 1981) and HMM-based methods (Karplus, Barrett, & Hughey, 1998; Johnson et al, 2010) are widely used similarity searching programs that produce accurate statistical estimates to identify sequences that share significant similarity and have similar underlying structures. The information contained in distance measures developed with these methods and applications can be further aggregated by employing exploratory dimensionality reduction techniques such as principal component analysis and hierarchical clustering (Hao et al., 2015), or multidimensional scaling (Tang et al., 2020).

2. Aim and Research Questions

The main objective of this paper is to show how the behavioral information that is stored in log files can be used to characterize problem-solving strategies in ways that generalize across tasks. We establish the extent to which individuals are able to adopt optimal problem-solving strategies, whether they under- or over-explore the problem space, how consistently they adopt the same behavior when solving tasks designed to engage different skills, and if these behavioral patterns are associated with the likelihood that individuals correctly solve the task at hand.

To do so, we first measure the distance between action sequences adopted by respondents and reference action sequences (which are defined ex-ante as the optimal method or methods to solve an item) using the longest common subsequence method (LCS) (Cormen et al., 2001; Sukkarieh et al., 2012). We construct indicators to characterize the behavior of respondents across multiple PSTRE tasks, and analyze differences across countries, as well as across population subgroups within each country. We compute four indicators that summarize the response strategies, also referred to as behavioral patterns in this work and that are comparable across individuals: similarity (how close on average the observed sequence of actions was to a reference optimal sequence); consistency of similarity (how the distance between the observed and the reference sequence varied across items); efficiency (if respondents solved items using the minimum possible number of actions) and consistency of efficiency (if the ability of respondents to use the smallest number of actions possible varied across items). This study pursues three research questions:

1. To what extent adults consistently adopt strategies that are close to the optimal strategies? Do respondents in different countries tend to display different behavioral patterns, i.e. do they differ on indicators of similarity, consistency of similarity, efficiency and consistency of efficiency?

2. What is the association between the adoption of different behavioral patterns and problem-solving proficiency?
3. Do behavioral patterns adopted by respondents while solving the PSTRE assessment differ systematically by gender, age, and familiarity with ICT skills at home and work?

3. Material and Methods

3.1. PSTRE in PIAAC

The PIAAC study was administered in over 40 countries between 2012 and 2017. It consisted of a background survey designed to collect a wide range of socio-demographic information from respondents (such as their educational attainment, their labor market status, and information on familiarity with ICT and the use of digital technologies at work and in everyday life) and of an assessment of literacy, numeracy and problem solving in technology-rich environments (PSTRE). While the literacy and numeracy assessments have been administered either on paper or on a computer, by its nature the PSTRE assessment was only accessible on a computer, and adults with insufficient familiarity with digital devices were excluded from the assessment and were only tested in literacy and numeracy. The PSTRE assessment required individuals to engage with complex, interactive tasks, and the computer-based delivery allowed some of these interactions to be recorded and stored in log files (Goldhammer et al., 2013).

For each PSTRE item, the test interface tried to mimic commonly used digital platforms and environments, such as email clients, web browsers, or spreadsheets. Although most test items used in the PSTRE assessment are confidential and cannot be disclosed, a number of sample items were released (OECD, 2011) and are reported in Appendix A. The item reported in

Figure A1, for instance, required test takers to access and evaluate information in the context of a simulated job search involving navigation across three different environments. Test takers had to find and bookmark one or more websites that did not require users to register or pay a fee. In this item, process data and path tracking could be used to better assess the respondent's level of understanding of the item, beyond the mere response data (merely whether the respondent solved the item correctly or not). For example, one of the websites shown in Figure A2 meets the specified criteria, but the relevant information about fees and registration is not on the opening page. If a respondent bookmarks this site as a correct answer without clicking on the "Learn More" link to view the relevant information, shown in Figure A3, one might interpret that response differently than if the third page had been viewed. This breadth of information, combined with frameworks that specify behaviors of interest, allows us to learn more about what adults know and can do relative to the problem-solving construct as conceptualized in PIAAC.

PSTRE was conceived along three dimensions: a cognitive dimension, a technology dimension, and a task dimension (OECD, 2012). The "cognitive dimension" refers to the mental structures and processes that are activated when a person engages with a problem with the aim of solving it. Four sub-dimensions were identified: setting goals and monitoring progress, planning and self-organizing, accessing and evaluation information, and making use of information by selecting, organizing and transforming information. The "technology dimension" refers to the devices, applications and functionalities through which adults are expected to solve problems (e.g., email, web browser, or spreadsheets). The "task dimension" refers to the circumstances that trigger a person's awareness and understanding of the problem and determine the actions needed to be taken to solve the problem. For instance, a test taker may be faced with a complex issue to find out more about a medical treatment and decide to look for relevant information on the web. Single or multiple steps and/or constrains (e.g., time conflict in booking a meeting room) may be involved in the task, with implicit (e.g., move emails to different folders) and/or explicit instructions (e.g., identifying specified information in a spreadsheet). More information on the development and implementation of the PSTRE in PIAAC study refer to the PIAAC Reader's Companion (OECD, 2019a).

Table 1 presents an overview of item contents, technological environments and cognitive dimensions for the seven interactive PSTRE items which we will use in the empirical analysis.

[Insert Table 1 around here.]

3.2. Data

This paper uses both response and process data on the seven items that were administered in the second module of PSTRE (PS2) in PIAAC in five countries: England/Northern Ireland (United Kingdom), Ireland, Japan, the Netherlands, and the United States. These countries were selected because they display a large range of PSTRE proficiency levels, facilitating the investigation of potential differences in problem-solving behaviors across countries with different performance levels.[2] The full distribution of PSTRE skills in these countries is reported in Appendix B.

Table 2 summarizes the characteristics of the respondents in our sample. The data refer to 7,462 respondents: 55% of them were women, and their average age was 39 years old (S.D.=13.6). Approximately half of the whole sample attained a level of education higher than a high-school diploma; this share was lowest in the Netherlands (at around 34%) and highest in

[2] Performance in PSTRE can be categorized in four levels: below level 1 (0–240), level 1 (241–290), level 2 (291–340), and level 3 (341–500). For more details, refer to OECD (2016).

Ireland (around 67%). The comparative data on the use of digital technologies at work and at home revealed a lower use in Japan than other countries, which was consistent with recent findings in the OECD Digital Economy Outlook (OECD, 2020). This report indicates that Japan is the country with the lowest average daily time spent using Internet, mobile Internet and social media in 2019. While the sample of adults that participated in PIAAC was designed to be representative of the adult population aged 16-65 living in each country, respondents could be excluded from the PSTRE assessment because they lacked sufficient familiarity with ICT devices, or because they simply decided to opt out of the computer-based assessment (CBA) and took the assessment using paper-based instruments (in which case they would only be assessed in literacy and numeracy). The percentage of exclusion from the computer-based assessment by country level, shown in the last column in Table 2, was in a range of 9.7% to 36.8%[3] among the five sample countries in the current study. On average across countries that participated in PIAAC, around 25% of the overall sample was administered the paper-based route and therefore did not participate in the PSTRE assessment (OECD, 2013).

[Insert Table 2 around here.]

Around 3% of the total PIAAC sample had only information on either process data or response data and were therefore dropped from the analysis resulting in a final sample size of 7,462 individuals in our study. Considering the complexity of multiple language input, we transferred all the actions related to keyboard input (e.g., type in “a” or use Shift & Ctrl) as a general keystroke labeled as “key”. The frequency of “key” was counted into the number of actions.

[3] In Japan, 61.8% of the respondents who completed the background questionnaire took the CBA, while 36.8% took the paper-based assessment without taking the PSTRE. This number is higher than the international average of 23.9% respondents who took the paper-based assessment (OECD, 2016). The percentage of exclusion from CBA explains the degree of potential bias of the sample in the PSTRE.

3.3. Materials

By design, all respondents were presented the seven PSTRE items in the same order. No limits in terms of time or in terms of number of actions were imposed.

Table 3 summarizes information about the characteristics of the action sequences undertaken by respondents. The first column presents the average length of action sequences by each item. Item U16 involves freestyle typing (i.e., the constructed response required to solve the item is not simply employing standardized symbols or numbers). Since U16 is the only item with this feature, it presents a relatively high peak in the length of action sequences compared to other items. The difficulty of the items can be mapped in the different levels that characterize the PSTRE scale, which is presented in the second column of Table 3.[4] The most difficult items were classified at Level 3 and the easiest were classified at Level 1. The last two columns of Table 3 report the number of reference sequences and the minimum number of actions needed to solve each item. Reference sequences are predefined optimal sequences: they were defined ex-ante by item developers and subject experts as the best ways to successfully complete the task. In our context "best" means that at any action node, reference sequences consisted of actions that were necessary to reach a successful solution to the problem presented. For any item/assessment task, it is possible to identify multiple reference sequences when alternative actions can be taken by respondents at a specific action node. Among the seven items in this study, only one item (U23) was designed to have a single optimal path of actions. Average sequence length was positively

[4] The difficulty level in PIAAC was derived by a linear transformation from the item parameters into the performance score scale. See more details in the PIAAC technical report (OECD, 2016).

correlated with the minimal number of required actions by each item. Test takers generally undertook more actions than specified in the reference sequences.

[Insert Table 3 around here.]

3.4. Methods

We analyzed test takers' actions using the LCS method (Cormen et al., 2001; Sukkarieh et al., 2012). The basic idea behind this approach is to identify differences between a reference sequence (RS) and the sequence of actions taken by a test taker to solve a particular test item (observed sequence, OS). This can be accomplished by computing for each item a measure of the distance between the observed and the reference sequence. These distances can be aggregated or compared across items and used to construct indicators that characterize the behavior of respondents during the assessment. Specifically, we propose two sets of indicators to characterize general behavioral patterns across items and subgroups of respondents: (1) similarity and consistency of similarity, and (2) efficiency and consistency of efficiency.

Similarity captures how much, on average, an individual's sequence deviates from a reference sequence (or the closest reference sequence in the case of items designed to have multiple reference sequences). It is a measure of the average degree of overlap between the sequences of actions taken by respondents across items and the corresponding reference sequences.

Consistency of similarity captures the degree to which respondents follow strategies that are consistently close or far from the relevant reference sequences across different items. It is computed as the standard deviation of the distances between observed sequences and reference sequences across items (i.e. the standard deviation of similarity). A consistent individual is someone who, over multiple items, either always follows very closely the optimal sequence, or that always deviates from it to the same extent (in other words, consistency is maximized when the distance between the observed and the reference sequence is constant across items).

Efficiency measures the ability of respondents to solve items using the minimum possible number of actions and is operationalized by the number of actions undertaken by a respondent more than the number of actions contained in the reference sequence. High efficiency indicates that no or little excess (redundant) actions occur.

Finally, we develop an indicator of consistency of efficiency. In line with consistency of similarity, consistency of efficiency measures the degree to which respondents are equally efficient (or inefficient) across different items and is computed as the standard deviation of efficiency across items.

3.4.1. Computing LCS

The LCS of a set of sequences is a subsequence whose length equals the maximum number of actions that are shared, in sequential order, with the reference sequences. For instance, suppose the observed sequence X for respondent *i* consists of the following string of actions: [A, A, B, A, D, E] (each letter represents one type of action). The reference sequence Y is [A, B, C, D, E]. The LCS is the longest string of actions that are present in sequential order in both X and Y, in this case [A, B, D, E]. The length of the LCS is then 4. The algorithm used to compute the LCS is presented in the Appendix C.

If there is no overlap between the two sequences, the length of the LCS is 0. In cases with unequal length of sequences, the maximum length of LCS equals the length of the shorter one of the two sequences.[5] As an illustration, we make the following concrete example[6] in the box below: the first line is the predefined reference sequence (RS). It is then followed by the action sequences of two different test takers' (OS1 and OS2). LCS1 and LCS2 are the two longest common subsequences derived from OS1 and OS2, respectively.

In this example, the problem is framed in a spreadsheet environment. According to the reference sequence (RS), respondents are expected to "start an item (Start), click into the spreadsheet (SS) environment, click the searching tool button (Search), type in the full name (to identify the person in request), go to the email (E) environment, type in the person ID, click on the Next button, and finally confirm to transit to the next item". The length of the RS is 8. The first test taker used 13 actions (that's the length of OS1) to solve the task, with 7 actions overlapping with the RS (and thus 7 is the length of LCS1). The second test taker only used 5 actions (the length of OS2), all of which were also contained in the RS; as a result, the length of LCS2 is 5.

[Insert Box 1 around here]

It is also possible to adapt the LCS method to fit situations in which multiple solutions for a task exist: in these contexts, the reference sequence that generates the longest LCS when paired with the observed sequence will be retained as the solution path that the respondent was most likely to follow.

Figure 1 presents an example of deriving the LCS for PSTRE item U19a. The first two RSs involve the use of the search function and both RSs have a total length of 11. The remaining

[5] See He et al. (2019) for more details on applying the LCS method to process data.

[6] Samples of actions and descriptions are available from the online PIAAC log data analyzer, which is a tool that extracts actions from the strings contained in raw log files. The log data analyzer is available at: https://dbk.gesis.org/dbksearch/sdesc2.asp?no=6712&db=e&doi=10.4232/1.12955

two RSs involve the use of the “sort” function and have a total length of 9. The action sequence of a hypothetical individual who used 25 steps to complete the task is also presented. The LCS was calculated by matching the individual action sequence with RS_1 to RS_4 separately, which resulted in LCS1 to LCS4. The LCS4 whose length equals 9 appears as the longest among the four, meaning this individual’s observed action sequence had the maximal similarity with RS_4. Therefore, it is assumed that the respondent tried to follow RS_4, and LCS4 was retained.

[Insert Figure 1 around here.]

For some problem-solving items, only one strategy can be used to reach a correct solution. In such cases, the action sequence is highly correlated with the final response. If a respondent misses a key action, his or her chances of giving the correct answer will be greatly reduced.

We finally measured the length of the observed sequence for each respondent in order to investigate whether the individual sequence was shorter or longer than necessary and construct an indicator of efficiency. All computations were performed using the Program R version 3.6.2.

3.4.2. Generating Indicators across Multiple Tasks

To aggregate the information extracted from the LCS across items, we constructed a set of indicators covering similarity, consistency of similarity, efficiency, and consistency of efficiency.

Similarity and consistency of similarity. For each item, similarity is defined as the ratio between the length of LCS (i.e., $len(LCS)$) and the length of the reference sequence (i.e.,

$len(RS)$). As it is always true that $len(LCS) \leq len(RS)$, similarity will always lie in the [0,1] interval. The higher the ratio, the more similar the observed sequence is to the reference sequence.

To capture the extent to which a respondent can consistently follow the reference sequences across different items, we looked at the distribution of similarity for each person across items. The mean of this distribution (SM) is defined as the average degree of similarity across items (and is simply labelled "similarity" in the paper). A higher value of SM indicates that, on average, a respondent solved problems by following the reference sequences closely. The standard deviation of this distribution (SSD) is used as an indicator of consistency of similarity. A low value of SSD implies that the distance between the observed and the reference sequence did not vary much for the same respondent across the seven items, while a high value suggests an inconsistent pattern, with the OS close to the RS in some items but far in others.

Efficiency and consistency of efficiency. Efficiency is defined as the ratio between the length of LCS and the length of the observed sequence (i.e., $len(OS)$) and measures to what degree the LCS and the actual observed sequence overlap. As it is always true that $len(LCS) \leq len(OS)$, the ratio between them would be a number within the [0,1] interval. A ratio close to 1 implies that a large proportion of the LCS can be matched with the OS, namely, the person solved the problem in an efficient way without performing too many actions that do not belong to the reference sequence.

The mean of this distribution (EM) is simply defined as the degree of efficiency across items. A higher value of EM indicates that a respondent on average solves problems in an efficient way (i.e., with few redundant actions). The standard deviation of this distribution (ESD) is used as an indicator for consistency of efficiency to examine whether a respondent can maintain a constant degree of efficiency across items.

3.4.3. Distribution of Measurement Indicators

The shortest possible observed action sequence (in this case the log file would record the sequence “Start, Next, Next_OK”) means the respondent skipped directly to the next item without interacting with the task at hand. We treated these cases as missing answers and code them as “Nonresponse (NR).” To avoid confusion in interpreting the results across items, all respondents with at least one occurrence of an NR sequence pattern were excluded from the LCS analysis. This resulted in the loss of 2,160 respondents, and in a final sample size of 5,302 respondents.

All four indicators (SM, SSD, EM, and ESD) were found to be approximately normally distributed (see Appendix D). To facilitate the analysis, we divided the indicators in three groups denoted by “low”, “moderate” and “high” values, where low values comprise cases more than one standard deviation below the mean, high values cases more than one standard deviation above the mean, and moderate values those between these two thresholds. The SM and SSD (i.e., average similarity and consistency of similarity) were then mapped in a matrix, classifying individuals into nine subgroups (Table 4). The first digit in the group name indicates the degree of consistency of similarity (1 is high consistency, 3 is low consistency). The second digit in the group name indicates the degree of average similarity (1 is low, 3 is high). For instance, respondents in group S13 consistently followed patterns that were very similar to the reference sequences, i.e. show high average similarity and high consistency of similarity. Analogously, efficiency and consistency of efficiency were mapped in a similar matrix, which resulted in nine groups defined by the degree of average efficiency and consistency of efficiency.

[Insert Table 4 around here.]

After classifying respondents according to the four indicators, we analyzed the relationship between behavioral patterns and performance in the PSTRE assessment, as well as the association with background variables. In order to do this, we use one-way analysis of variance (ANOVA) statistics with Bonferroni post-hoc correction.

4. Results

4.1. Behavioral patterns

To answer the first research question, we examined to what extent respondents followed patterns similar to the reference sequences and how consistently they did so across different items. Figure 2 displays the association between the similarity index (horizontal axis) and the consistency of similarity index (vertical axis). The reference lines $X = 0.51$ and $X = 0.81$ indicate thresholds for the three similarity groups (i.e., low-, moderate-, and high-similarity group), while the reference lines $Y = 0.15$ and $Y = 0.25$ indicate thresholds for the three consistency groups (i.e., low-, moderate-, and high-consistency groups). Over two-thirds of respondents belonged to the high- and moderate-consistency groups. This tendency appears more obvious in the low- and high-similarity groups. Among respondents with low consistency, the groups S31 (group with low consistency and low similarity located at the upper-left corner) and S33 (group with low consistency and high similarity at the upper-right corner) had the lowest proportion of respondents (2.6% and 0.3%, respectively). Within the high-similarity group, 98% of respondents adopted moderately or highly consistent strategies. Over 81% of respondents were in the moderate- and high-similarity groups, meaning that the large majority of respondents followed the reference sequences. Within the high-consistency group, respondents with low, moderate, and high similarity values were represented in equal proportions. However, this ratio was substantially different for the low-consistency group, where the proportion of low-, moderate-, and high-similarity groups was 14%, 84%, and 2%, respectively. Three conclusions can be drawn from results presented in Figure 2:

First, most respondents adopted strategies similar to the predefined ones. This can be seen by the small proportion of respondents in the low-similarity cells S11, S21 and S31. Second, respondents with average levels of similarity tend to display average levels of consistency (cell S22), meaning that for these respondents the distance between the observed and the reference sequences does not vary much across items. Third, respondents at the extreme of the similarity distribution, i.e. whose sequences were on average very close or very far from the reference sequence, tended to do so in a very consistent way across items (meaning that the distance between the observed and the reference sequences varied very little across items).

[Insert Figure 2 around here.]

When looking at results for individual countries, Japan and the Netherlands, the two countries with the highest average performance in the PIAAC PSTRE assessment, displayed a higher degree of similarity than the other countries. On average, respondents from these two countries followed the reference action sequence more often than respondents in countries with lower levels of performance. The one-way ANOVA on a single-factor country, reveals significant differences in similarity between the five countries, $\mathrm{F}(4,5297) = 39.52, \mathrm{P} < 0.001$; the Bonferroni post hoc correction showed that the similarity indicators on average in Japan and the Netherlands are significantly higher than in the other three countries. By contrast, no obvious country differences were observed in the indicator consistency of similarity, except for Japan, where the indicator was slightly lower than in the other countries (suggesting higher levels of consistency for Japanese respondents). The ANOVA test resulted in $\mathrm{F}(4,5297) = 9.10, \mathrm{P} < 0.001$; the Bonferroni post hoc correction showed Japan was significantly lower than the other countries.

A similar investigation was conducted on efficiency and consistency of efficiency across the five countries. Japan showed a significantly lower average efficiency and higher consistency of efficiency than the other four countries, suggesting that adults in Japan usually performed more actions than necessary. A set of boxplots of general patterns in similarity and efficiency (SM, SSD, EM, ESD) across five sample countries is provided in Appendix E.

One possible explanation for the lower efficiency of respondents in Japan is the fact that a typo was present in the Japanese version of items U19a and U19b: a space between two words that is optional in real life settings was required if respondents were to proceed in the context of the assessment items. This typo pushed Japanese respondents to execute redundant actions to explore the solution, for instance, to keep on making typing errors (due to the unknown space between the two words) or switch strategies from searching to sorting function. This issue might cause a lower efficiency in the Japanese group in the first two items, resulting in longer response time and more actions.

The correlation between similarity and efficiency across all the five countries showed a negative value (-0.31). These results imply that those in the high-similarity group usually undertook considerably more steps than necessary to provide a response. Few respondents were able to achieve both high efficiency and high similarity at the same time.

4.2. The Association between Behavioral Patterns and Problem-Solving Proficiency

To answer the second research question, we examined the association between response strategies, and overall proficiency in PSTRE. We found a high and positive correlation (0.79) between scores in the PSTRE assessment and the similarity indicator, implying that respondents who achieved higher scores tended to perform sequences of actions that were close to the reference sequences.

Figure 3 shows the association between similarity and PSTRE proficiency scores. The horizontal axis stands for the similarity groups nested by their consistency subgroups, and the vertical axis for respondents' PSTRE scores. A clear increasing trend is observed in the three clusters from low to high similarity with the increasing PSTRE proficiency scores.

Consistency of similarity was also correlated with proficiency, but as expected the relationship varied according to the degree of similarity: within the moderate- and high-similarity groups, adults who were more consistent (i.e., S12, S22, S13, and S23) performed slightly better than respondents in the low-consistency groups (i.e., S32 and S33). Among respondents in the low-similarity group, those who adopted more consistent sequences (group S11) showed the lowest performance score. This makes perfect sense, as these individuals consistently adopted strategies very dissimilar from the reference sequences.

[Insert Figure 3 around here.]

The relationship among proficiency in PSTRE and the four indicators derived from process data is further explored in Figure 4, which presents the associations between average similarity and consistency of similarity (plot A), as well as average efficiency and consistency of efficiency (plot B). In Figure 4 the horizontal axis stands for the degree of average similarity and efficiency and the vertical axis for consistency index, respectively. Within each set of indicators, four panels are defined by the PSTRE proficiency levels (level 3 is the highest). The mean values and standard errors by each indicator are reported in the accompanying table. The average degree of similarity significantly increased from 0.39 to 0.83 as respondents' proficiency in PSTRE increased. Adults in the two middle proficiency levels (i.e., level 1 and level 2) showed a more dispersed distribution of both similarity and consistency of similarity compared to adults in the extreme proficiency levels (e.g., below level 1 and level 3). This figure confirms that following the predefined optimal path is a good proxy of final performance in the assessment.

In contrast, we did not find a strong relationship between the degree of efficiency and proficiency in PSTRE. Adults who scored below level 1 showed a slightly more dispersed distribution of efficiency than adults in the other three levels, suggesting a mixture of patterns in efficiency in this group. Some test takers within this group frequently used very short sequences, while some performed a very large number of actions, using almost every button in the interface. The narrowest band (i.e., ESD=0.21) was noticed in the distribution of adults in proficiency level 3, who displayed the lowest degree of efficiency. This suggests that it was very difficult to achieve high scores in an efficient way, that is, minimizing the number of redundant or useless actions.

[Insert Figure 4 around here.]

4.3. Differences by Socioeconomic Background in Behavioral Patterns

To answer the third research question, we examined the relationships between the indicators of similarity and consistency of similarity and sociodemographic characteristics of respondents such as gender, age, and (self-declared) familiarity with ICT.

[Insert Table 5 around here.]

Results presented in Table 5 suggest that women were more likely than men to be in the low- and moderate-similarity groups (by approximately 4 and 3 percentage points, respectively). Men were more likely to be in the high-similarity group (by 7 percentage points). This gender gap was present in all countries, although to different degrees: the smallest difference between males and females was found in Japan and the highest in Ireland. Interestingly, women generally displayed more consistent behaviors, especially when their average similarity was low or moderate (i.e., S11, S12, S21 and S22). Men's behavior tended to be less consistent; however, men in the high similarity group tended to be more consistent than women in the same group (see e.g. cells S13 and S23).

An ANOVA analysis revealed significant age differences, $\mathrm{F}(2, 5299) = 69.53, \mathrm{P} < 0.001$. Younger respondents typically had higher levels of the similarity indicator; the average age of respondents in the high-, moderate-, and low-similarity groups was 34.7, 37.2, and 41.5 years, respectively. Respondents in group S11—characterized by high consistency and low-similarity—were the oldest (average 44 years old), and those in group S13 —characterized by high consistency and high similarity – were the youngest (average 34 years old).

Significant age differences in the degree of consistency were also found within groups defined by levels of similarity. In the low-similarity group, average age dropped from 44 to 38.5 years as the degree of consistency decreased. The opposite was found in the high-similarity group, where adults with lower levels of consistency tended to be significantly older (average 41 years old). However, because of the small sample size in S33, this result may not be sufficiently reliable to draw generalizable conclusions. The findings in socio-demographic variables (age and gender) kept in line with Hämäläinen et al. (2019) where younger and male adults were found to have a greater chance of being strong problem-solvers especially with their parents having a tertiary or upper-secondary degree.

Previous research indicates that adults who more often use ICT devices at work or at home perform better in the PSTRE assessment than those who rarely use them (Liao et al., 2019; He et al., 2018). Table 5 presents information on the extent to which engaging in tasks that require the use of ICT skills at work and at home is associated with behavioral indicators extracted from process data. Levels of ICT use at work were significantly different across adults belonging to different similarity groups in ANOVA $\mathrm{F}(2{,}3542) = 48.13, \mathrm{P} < 0.001$. Differences remained significant after correcting for the pairwise comparisons. However, no significant differences emerged by levels of consistency for the moderate- and high-similarity groups. This suggests that making more frequent use of ICT skills at work was predictive of higher similarity to the reference sequences, but not of higher consistency.

Similar results were found in the associations with the use of ICT at home. A significant difference was found across the three similarity groups in the ANOVA test, $\mathrm{F}(2{,}5012) =$

$89.31, P < 0.001$. However, differences were no longer significant in the post-hoc test that corrects for pairwise comparisons between the moderate- and high-similarity groups. This means that frequency of use of ICT at home is a less discriminating factor than use of ICT at work. Significant differences in consistency were found only for the group displaying low similarity.

5. Conclusions and Discussion

The use of digital technologies in assessment programs offers researchers and policymakers new opportunities to collect and analyze information such as process data (Mislevy et al., 2012; Kirsch et al., 2017). This promises to be beneficial for researchers, assessment specialists, and educators, providing information for interpretation and analysis to allow for better understanding of test takers' behaviors. So far, advances in understanding cognitive processes involved in problem solving have been hampered by the reliance of researchers on either measures that are easily generalizable across assessment items but lack specificity and detail (such as timing information) or measures that provide a rich characterization of the problem-solving process of test takers for a specific test item but cannot be generalized.

In this paper we exploited the LCS method to derive indicators that can capture and characterize the strategies followed by adult participating in a problem-solving assessment across a variety of items. We found consistent behavioral patterns for a large number of respondents across five countries. More consistent patterns were observed in groups performing at either high levels or low levels in the test. We also found that the degree of similarity, and the consistency of similarity across items, was a significant predictor of overall performance in the test both at the individual and country level.

We found that men were more likely than women to use strategies similar to the reference ones, while women showed more consistent patterns across items. Younger respondents were more likely than older respondents to follow reference sequences when solving digital tasks. Respondents engaging more frequently (at work or home) with tasks that require ICT skills were also more likely to follow optimal strategies. These findings are especially helpful for policymakers to better understand the general patterns by different subgroups.

Our study also has some limitations. The sample used in the current study included all respondents who were administered problem-solving items at the end of a two-part test, irrespective of whether they had received literacy items, numeracy items or other problem-solving items in the first part of the test. It is therefore possible that individuals who had been allocated problem-solving items in the first stage of the assessment might have adopted different test-taking strategies as a result of being more familiar with the assessment domain (Avvisati & Borgonovi, 2020); unfortunately, the small sample size did not allow such examination.

Moreover, the current study does not make use of timing information. This additional information could be used in the future to better characterize respondents' "efficiency": shorter action sequences could in fact entail more time expended overall (von Davier et al., 2019). Respondents who have the same action sequences may differ in the time interval between actions (Ulitzsch et al., in press).

The LCS method identifies a single strategy that the respondent was most likely to have followed. However, in practice, respondents often use a combination of strategies: they might switch from one strategy to another, or they might use multiple strategies to validate their solutions. It would be interesting to refine the present LCS approach in future studies to accommodate mixed membership of problem-solving strategies. In addition, the LCS method used in this study is just one of several possible choices. Alternative choices such as edit distance (Hao et al., 2015) and optimal symbol alignment distance (Herranz et al., 2011) could also be used and might give similar results. However, these measures are often more computationally demanding and are less well suited for visualizing the strategy that has been chosen by respondents.

Finally, different orthographical input of alphabetic-based and character- based languages may also impact the length of sequences in problem-solving process such as typing in key information in searching function or giving a constructed response to an open-ended question. The frequency of keystrokes in character-based language such as Japanese may require more actions than the alphabetic-based languages because an extra step to select an appropriate homophone is usually involved in typing actions (Ercikan et al., 2020). The effect of language variation has been minimized in the PIAAC tasks by using standardized symbols and numbers, and only one item required freestyle typing. Overall, we did not find any significant difference in the sequence length by language input.

Future studies could further refine the LCS method developed in this paper and integrate information from process data with item response models to get a more accurate estimate of respondents' latent traits. While the analysis of process data can add valuable insights and expand the value of the information gathered in the context of large-scale assessments, it is important to be mindful of the technical and ethical implications that stem from the use of such data and give a meaningful interpretation of the results.

Future research could also try to use the cognitive diagnostic model to examine the relationship between each cognitive dimension and problem-solving skills and identify subskills that need a further improvement.

Besides, cultural and country differences might be reflected in test takers' process patterns. This also leads to new research questions. For instance, whether test takers from different cultural backgrounds or different countries adopt the same solution to a digital task, and whether the differences in strategies have an association with test takers' responses to assessment items, or to items contained in the background questionnaire that is administered alongside the assessment. We believe that process data will play an important role in providing a new angle to explore these issues of cultural diversity.

As valuable as process data can be, they are currently not much more than a byproduct of the features of the software used to deliver the assessment. The choice of which information was to be recorded in log files was not driven by theoretical or analytical considerations, and similarly, items were not explicitly designed with the purpose of identifying any strategy with the help of process data (OECD, 2019b). The research potential of process data can be maximized if their use for secondary analysis of this type were considered at the item development stage. Effort should be made to design items that allow for clear identification of different solution strategies; ideally, these different strategies would be mapped to cognitive theories that researchers might be interested to test with the aid of process data, thus better supporting adult learning and assessments (OECD, 2019b).

References

Altschul, S., Madden, T. L., Schaffer, A. A., Zhang, J., Miller, W. & Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, *25*(17), 3389–3402.

Avvisati, F., & Borgonovi, F. (2020). Learning mathematics problem solving through test practice: A randomized field experiment on a global scale. *Educational Psychology Review*. https://doi.org/10.1007/s10648-020-09520-6

Baker R.S., & Inventado P.S. (2014). Educational Data Mining and Learning Analytics. In: Larusson J., White B. (eds) *Learning Analytics*. Springer, New York, NY. https://doi.org/10.1007/978-1-4614-3305-7_4

Biswas, G., Jeong, H., Kinnebrew, J. S., Sulcer, B., & Roscoe, R. O. D. (2010). Measuring self-regulated learning skills through social interactions in a teachable agent environment. *Research and Practice in Technology Enhanced Learning*, *5*(2), 123-152. https://doi.org/10.1142/S1793206810000839

Biswas, G., Leelawong, K., Schwartz, D., Vye, N., & The Teachable Agents Group at Vanderbilt (2005). Leaning by teaching: A new agent paradigm for educational software. *Applied Artificial Intelligence, 19* (3-4), 363-392, https://doi.org/10.1080/08839510590910200

Chen, Y., Li, X., Liu, J., & Ying, Z. (2019). Statistical analysis of complex problem-solving process data: An event history analysis approach. *Frontiers in Psychology*, *10*, 486.

Choi, I., Nisbett, R. E., & Norenzayan, A. (1999). Causal attribution across cultures: Variation and universality. Psychological Bulletin, 125(1), 47–63.

Cormen, T. H., Leiserson, C. E., Rivest, R. L., & Stein, C. (2001). *Introduction to algorithms (2nd ed.).* MIT Press.

Deane, P., Wilson, J., Zhang, M., Li, C., van Rijn, P., Guo, H., Roth, A., Winchester, E., & Richter, T. (2020). The Sensitivity of a Scenario-Based Assessment of Written Argumentation to School Differences in Curriculum and Instruction. *International Journal of Artificial Intelligence in Education.* https://doi.org/10.1007/s40593-020-00227-x.

Dong, G., & Pei, J. (2007). *Sequence Data Mining*. New York, NY: Springer.

Eichmann, B., Goldhammer, F., Greiff, S., Brandhuber, L., & Naumann, J. (2020). Using process data to explain group differences in complex problem solving. *Journal of Educational Psychology*. https://doi.org/10.1037/edu0000446

Embretson, S. E. (2004). *Measuring human intelligence with artificial intelligence.* Georgia Institute of Technology.

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

Febrer-Hernandez, J. K. & Hernandez-Palancar, J. (2012). Sequential pattern mining algorithms review. *Intelligent Data Analysis, 16*, 451-466.

Goldhammer, F., Naumann, J., & Keβel, Y. (2013). Assessing individual differences in basic computer skills: Psychometric characteristics of an interactive performance measure. *European Journal of Psychological Assessment*, *29*(4), 263–275.

Goldhammer, F., Naumann, J., Selter, A., Toth, K., Rolke, H., & Klieme, E. (2014). The time on task effect in reading and problem solving is moderated by task difficulty and skill:

Insights from a computer-based large-scale assessment. *Journal of Educational Psychology*. https://doi.org/10.1037/a0034716

Goldhammer, F., Hahnel, C., & Kroehne, U. (2020). Analyzing log file data from PIAAC. In D. B. Maehler &B. Rammstedt (Eds.), *Large-scale cognitive assessment: Analyzing PIAAC data*. Cham: Springer.

Gómez-Alonso, C., & Valls, A. (2008). A similarity measure for sequences of categorical data based on the ordering of common elements. In V. Torra & Y. Narukawa (Eds.), *Modeling decisions for artificial intelligence* (pp. 134–145). Berlin: Springer.

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.

Guo, H., Deane, P. D., van Rijn, P. W., Zhang, M., & Bennett, R. E. (2018). Modeling basic writing processes from keystroke logs. *Journal of Educational Measurement*, *55*(2), 194–216.

Guo, H., Zhang, M., Deane, P., & Bennett, R. (2020). Effects of Scenario-Based Assessment on Students' Writing Processes. *Journal of Educational Data Mining*, *12*(1), 19-45. https://doi.org/10.5281/zenodo.3911797

Hahnel, C., Goldhammer, F., Naumann, J., & Kröhne, U. (2016). Effects of linear reading, basic computer skills, evaluating online information, and navigation on reading digital text. *Computers in Human Behavior*, *55*, 486-500.

Hämäläinen, R., de Wever B., Nissinen, K., Cincinnato, S. (2019) What makes the difference – PIAAC as a resource for understanding the problem-solving skills of Europe's higher-education adults, *Computers & Education, 129*, 27-36.

Hamming, R. W. (1950). Error detecting and error correcting codes. *The Bell System Technical Journal, 29* (2): 147–160.

Han, S., Northoff, G., Vogeley, K., Wexler, B. E., Kitayama, S., & Varnum, M. E. W. (2013). A cultural neuroscience approach to the biosocial nature of the human brain. *Annual Review of Psychology, 64,* 335-359

Han, S., & Ma, Y. (2014). Cultural differences in human brain activity: A quantitative meta-analysis. *NeuroImage*, *99*(1), 293-300.

Han, Z., He, Q., & von Davier, M. (2019). Predictive feature generation and selection using process data from PISA interactive problem-solving items: An application of random forests. *Frontiers in Psychology*, *10*, 1421.

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

He, Q., Borgonovi F., & Paccagnella, M. (2019). Using process data to understand adults' problem-solving behaviour in the Programme for the International Assessment of Adult Competencies (PIAAC): Identifying generalised patterns across multiple tasks with sequence mining, OECD Education Working Papers, No. 205. OECD Publishing. https://doi.org/10.1787/650918f2-en

He, Q., & von Davier, M. (2015). Identifying feature sequences from process data in problem-solving items with n-grams. In A. van der Ark, D. Bolt, S. Chow, J. Douglas, & W. Wang

(Eds.), *Quantitative Psychology Research: Proceedings of the 79th Annual Meeting of the Psychometric Society* (pp. 173–190). Springer.

He, Q., & von Davier, M. (2016). Analyzing process data from problem-solving items with n-grams: Insights from a computer-based large-scale assessment. In Y. Rosen, S. Ferrara, & M. Mosharraf (Eds.) *Handbook of research on technology tools for real-world skill development* (pp. 749–776). Information Science Reference.

He, Q., von Davier, M., & Han, Z. (2018). Exploring process data in computer-based international large-scale assessments. In H. Jiao, R. Lissitz, & A. van Wie (Eds.), *Data analytics and psychometrics: Informing assessment practices* (pp. 53–76). Information Age Publishing.

Herranz, J., Nin, J., & Sole, M. (2011). Optimal symbol alignment distance: A new distance for sequences of symbols. *IEEE Transactions on Knowledge and Data Engineering, 23*(10), 1541–1554.

Hirschberg, D. S. (1975). A linear space algorithm for computing maximal common subsequences. *Communications of the Association for Computing Machinery, 18*, 341-343.

Hirschberg, D. S. (1977). Algorithms for the longest common subsequence problem. *Journal of the Association for Computing Machinery, 24*(4), 664-675.

Huttenlocher, P. R. (2002). *Neural plasticity: The effects of environment on the development of the cerebral cortex.* Cambridge, MA: Harvard University Press.

Johnson, L. S., Eddy, S. R., Portugaly, E. (2010). Hidden markov model speed heuristic and iterative hmm search procedure. *BMC Bioinformatics, 11*, 431.

Jurafsky, D., & Martin, J. H. (2009). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Upper Saddle River, NJ: Pearson Prentice Hall.

Karplus, K., Barrett, C., & Hughey, R. (1998). Hidden Markov models for detecting remote protein homologies. *Bioinformatics*, *14*, 846–856.

Kirsch, I., & Lennon, M. L. (2017), PIAAC: A new design for a new era. *Large-Scale Assessments in Education*, *5*(11). https://doi.org/10.1186/s40536-017-0046-6

Kour, A. (2017). Sequential rule mining, methods and techniques: A review. I*nternational Journal of Computational Intelligence Research*, *13*(7). 1709-1715.

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

Lämsä, J., Hämäläinen, R., Koskinen, P., Viiri, J., Mannonen, J. (2020). The potential of temporal analysis: Combining log data and lag sequential analysis to investigate temporal differences between scaffolded and non-scaffolded group inquiry-based learning processes, *Computers & Education*, *143*. https://doi.org/10.1016/j.compedu.2019.103674.

Lee, Y. H., & Jia, Y. (2014). Using response time to investigate students' test-taking behaviors in a NAEP computer-based study. *Large-Scale Assessments in Education*, *2*(8). https://doi.org/10.1186/s40536-014-0008-1

Levenshtein, V. I. (1965). Binary codes capable of correcting spurious insertions and deletions of ones. *Problems of Information Transmission*, *1*(1), 8–17.

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

Liao, D., He, Q., & Jiao, H. (2019). Mapping background variables with sequential patterns in problem-solving environments: An investigation of U.S. adults' employment status in PIAAC. *Frontiers in Psychology, 10*, 646.

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*.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT Press.

Markus, H. R., & Kitayama, S. (2010). Cultures and selves. A cycle of mutual constitution. *Perspectives on Psychological Science*, *5,* 420-430.

Masuda, T., & Nisbett, R. E. (2001). Attending holistically versus analytically: Comparing the context sensitivity of Japanese and Americans. *Journal of Personality and Social Psychology*, *81*(5), 922–934.

Mislevy, R. J., Behrens, J. T., Dicerbo, K. E., & Levy, R. (2012). Design and discovery in educational assessment: Evidence-centered design, psychometrics, and educational data mining. *Journal of Educational Data Mining*, *4*(1), 11–48.

Nisbett, R.E. (2003). *The geography of thought*. New York: Free Press.

OECD (2009). PIAAC problem solving in technology-rich environments: A conceptual framework, OECD Education Working Paper No. 36. OECD Publishing.

OECD (2011). Released items from the Programme for the International Assessment of Adult Competencies (PIAAC). OECD Publishing. http://www.oecd.org/skills/piaac/samplequestionsandquestionnaire.htm

OECD (2012). *Literacy, numeracy and problem solving in technology-rich environments: Framework for the OECD Survey of Adult Skills.* OECD Publishing. https://doi:10.1787/9789264128859-en

OECD (2013). *The Survey of Adult Skills: Reader's Companion,* OECD Publishing. http://dx.doi.org/10.1787/9789264204027-en

OECD (2016). *Technical report of the Survey of Adult Skills (PIAAC) (Second Edition).* OECD Publishing, Paris.

OECD (2019a). *The Survey of Adult Skills: Reader's Companion, Third Edition, OECD Skills Studies,* OECD Publishing, Paris. https://doi.org/10.1787/f70238c7-en.

OECD (2019b). *Beyond proficiency: Using log files to understand respondents' behaviour in PIAAC.* OECD Publishing, Paris.

OECD (2020). *OECD Digital Economy Outlook 2020*. OECD Publishing, Paris. https://doi.org/10.1787/bb167041-en.

Paquette, L., Ocumpaugh, J., Li, Z., Andres, A., & Baker, R. (2020). Who's Learning? Using Demographics in EDM Research. *Journal of Educational Data Mining, 12*(3), 1-30. https://doi.org/10.5281/zenodo.4143612

Pearson, W. R. (1990). Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods in Enzymology, 183*, 63–98.

Pearson, W. R., & Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences of the United States of America, 85*, 2444–2448.

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

Sharda, R., Delen, D., & Turban, E. (2018). *Business Intelligence, Analytics, and Data Science: A Managerial Perspective.* New York, Pearson Education, Inc.

Shute, V. J., Wang, L., Greiff, S., Zhao, W., & Moore, G. (2016). Measuring problem solving skills via stealth assessment in an engaging video game. *Computers in Human Behavior, 63*, 106-117. https://doi.org/10.1016/j.chb.2016.05.047.

Sireci, S. G., & Zenisky, A. L. (2006). Innovative item formats in computer-based testing: In pursuit of improved construct representation. In S. M. Downing & T. S. Haladyna (Eds.), *Handbook of test development* (pp. 329–347), Erlbaum.

Smith, T. F., & Waterman, M. S, (1981). Identification of Common Molecular Subsequences. *Journal of Molecular Biology, 147*, 195–197.

Stadler M., Fischer, F., and Greiff, S. (2019) Taking a closer look: An exploratory analysis of successful and unsuccessful strategy use in complex problems. *Frontiers in Psychology, 10*, 777.

Sukkarieh, J. Z., von Davier, M., & Yamamoto, K. (2012). *From biology to education: Scoring and clustering multilingual text sequences and other sequential tasks*. (Research Report No. RR-12-25). Educational Testing Service.

Tang, X., Wang, Z., He, Q., Liu, J., & Ying, Z. (2020). Latent feature extraction for process data via multidimensional scaling. Psychometrika. https://doi.org/10.1007/s11336-020-09708-3.

Ulitzsch, E., He, Q., Ulitzsch, V., Nichterlein, A., Molter, H., Niedermeier, R., Pohl, S. (in press). Combining clickstream analyses and graph-modeled data clustering for identifying common response process using time-stamped action sequence. *Psychometrika.*

van der Linden, W. J. (2005). *Linear models for optimal test design*. New York, NY: Springer.

von Davier, M., Khorramdel, L., He, Q., Shin, H., & Chen, H. (2019). Developments in psychometric population models for data from innovative items. *Journal of Educational and Behavioral Statistics, 44*(6), 671–705.

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,* 478–498.

Zhu, M., & Feng, G. (2015). An exploratory study using social network analysis to model eye movements in mathematics problem solving. In *Proceedings of the Fifth International Conference on Learning Analytics and Knowledge* (pp. 383–387). ACM.

Box 1

Example of a Reference Sequence and Longest Common Sequences

RS (length=8): Start, SS, Search, Type_FullName, E, InputID, Next, Next_OK

OS1 (length=13): Start, E, SS, E, SS, Search, Type_FullName, Sort, Next, Next_Cancel, E, Next, Next_OK

LCS1 (length=7): Start, SS, Search, Type_FullName, E, Next, Next_OK

OS2 (length=5): Start, E, InputID, Next, Next_OK

LCS2 (length=5): Start, E, InputID, Next, Next_OK

Table 1

Item Concepts and Cognitive Dimensions of the Seven PSTRE Items in PIAAC PS2

Item ID	Item Content	Environments			Cognitive Dimensions			
		Email	Web	Spreadsheet	Goal setting and monitoring progress	Planning, self-organizing	Acquiring and evaluating information	Making use of information
U19a	Club Membership	X		X		X		X
U19b	Club Membership			X	X	X	X	
U07	Book Order		X				X	
U02	Meeting Room	X	X		X		X	
U16	Relay All	X				X		
U11b	Locate Email	X					X	X
U23	Lamp return	X	X			X	X	

Note: The items are ordered according to the order or appearance in the assessment.

Table 2

Sample Descriptions by Countries, PSTRE Proficiencies and Background Variables

	Sample Size in Current Study	Sample Proportion by PSTRE Level				PSTRE Score	Gender	Age	Education	ICT at Home	ICT at Work	Percentage of Exclusion from CBA on Country Level
		Below level 1	Level 1	Level 2	Level 3	Mean (S.D.)	Female (%)	Mean (S.D.)	Above high school (%)	Mean (S.D.)	Mean (S.D.)	
GBR	2,317	0.18	0.43	0.35	0.04	276.4 (39.0)	60%	39.4 (13.5)	42%	2.0 (0.9)	2.1 (1.0)	14.1%
IRL	1,289	0.20	0.46	0.31	0.04	273.8 (39.2)	55%	37.6 (12.3)	67%	2.1 (0.9)	2.2 (1.0)	30.7%
JPN	1,041	0.12	0.28	0.48	0.12	294.4 (42.5)	49%	38.0 (13.5)	57%	1.5 (0.9)	1.7 (0.9)	36.8%
NLD	1,487	0.14	0.37	0.42	0.07	285.8 (40.1)	53%	40.8 (14.3)	34%	2.3 (0.8)	2.1 (0.9)	9.7%
USA	1,328	0.20	0.42	0.33	0.05	274.9 (41.3)	53%	39.1 (14.0)	50%	2.3 (0.9)	2.1 (1.1)	14.9%
Total	7,462	0.17	0.39	0.38	0.06	280.1 (40.8)	55%	39.0 (13.6)	50%	2.1 (0.9)	2.1 (1.0)	21.2%

Note: S.D. refers to standard deviation. GBR=England/Northern Ireland (United Kingdom), IRL=Ireland, JPN=Japan, NLD=Netherlands, USA=United States. N=7,462, corresponding to test takers in PIAAC in the five countries who were assigned to PS2. The last column, percentage of exclusion from CBA on country level indicates the percentage of respondents per country who completed the Background Questionnaire took the paper-based assessment because of being opted out (personally refused CBA) or excluded (failed in ICT Core Assessment) from the CBA, thus not took PSTRE (data source: PIAAC Tech Report, (OECD, 2016)).

Table 3

Sequence Characteristics of the Seven PSTRE Items in PIAAC PS2 (N=7,462)

Item ID	Average Sequence Length	Difficulty Level	Number of Reference Sequences	Minimal Number of Required Actions
U19a	19.63	1	4	9
U19b	21.18	2	4	12
U07	18.08	2	2	18
U02	53.02	3	5	25
U16	97.71	1	16	8
U11b	29.61	3	18	10
U23	28.51	2	1	17

Note: Reference sequences indicate the expert-predefined action sequences for each item. The minimal number of actions indicates the least number of actions to correctly solve the item. The items are ordered according to the order of appearance in the assessment. The keystrokes are counted into the number of actions. The key strokes in different languages were not counted into the minimal number of required actions in U16.

Table 4

Matrix of Similarity and Consistency of Similarity

		Average Similarity (MEAN)		
		M1 (Low Similarity)	M2 (Moderate Similarity)	M3 (High Similarity)
Consistency (SD)	SD1 (High Consistency)	S11 High Consistency Low Similarity	S12 High Consistency Moderate Similarity	S13 High Consistency High Similarity
	SD2 (Moderate Consistency)	S21 Moderate Consistency Low Similarity	S22 Moderate Consistency Moderate Similarity	S23 Moderate Consistency High Similarity
	SD3 (Low Consistency)	S31 Low Consistency Low Similarity	S32 Low Consistency Moderate Similarity	S33 Low Consistency High Similarity

Note: M indicates the mean level of similarity across items per person. SD indicates the standard deviation of similarity across items per person. The degree of average similarity is shown by columns. The degree of consistency of similarity is shown by rows.

Table 5

Descriptive Statistics of Gender, Age, and ICT Skills by Similarity Levels

	Gender		Age		ICT Skills at Work		ICT Skills at Home	
	Male (%)	Female (%)	Mean (SE)	S.D.	Mean (SE)	S.D.	Mean (SE)	S.D.
Low Similarity	**16.7%**	**20.8%**	**41.51 (0.45)**	**14.17**	**1.87 (0.04)**	**0.94**	**1.90 (0.04)**	**0.82**
S11	4.6%	6.0%	43.79 (0.82)	13.85	1.72 (0.07)	0.89	1.69 (0.06)	0.89
S21	9.6%	12.0%	41.11 (0.59)	14.14	1.86 (0.05)	0.94	1.91 (0.04)	0.83
S31	2.5%	2.8%	38.45 (1.22)	14.33	1.97 (0.11)	0.95	2.02 (0.08)	0.97
Moderate Similarity	**62.0%**	**65.2%**	**37.15 (0.23)**	**13.32**	**2.16 (0.02)**	**0.98**	**2.25 (0.02)**	**0.91**
S12	6.1%	6.9%	37.31 (0.72)	13.34	2.19 (0.06)	0.99	2.23 (0.05)	0.84
S22	40.9%	42.4%	37.30 (0.28)	13.36	2.14 (0.09)	0.99	2.26 (0.02)	0.98
S32	15.0%	15.9%	36.69 (0.46)	13.24	2.15 (0.04)	1.03	2.21 (0.03)	0.83
High Similarity	**21.3%**	**14.0%**	**34.66 (0.36)**	**10.91**	**2.40 (0.04)**	**0.98**	**2.38 (0.03)**	**0.80**
S13	9.0%	5.1%	33.85 (0.57)	11.03	2.40 (0.06)	1.08	2.46 (0.04)	0.81
S23	12.1%	8.5%	35.03 (0.46)	10.79	2.38 (0.04)	0.92	2.37 (0.03)	0.79
S33	0.2%	0.4%	41.40 (2.61)	10.13	2.47 (0.20)	0.72	2.47 (0.24)	0.95

Note: SE refers to standard error. S.D. refers to standard deviation. The consistency subgroups nested within each similarity group are ordered in a range from high consistency (e.g., S11) to low consistency (e.g., S31).