OECD Education Working Papers No. 205

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

**Qiwei He,
Francesca Borgonovi,
Marco Paccagnella**

<https://dx.doi.org/10.1787/650918f2-en>

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 Paper No. 205

Qiwei He, Francesca Borgonovi and Marco Paccagnella, OECD

This working paper has been authorised by Andreas Schleicher, Director of the Directorate for Education and Skills, OECD.

Qiwei He, Educational Testing Service (ETS) (qhe@ets.org);
and
Marco Paccagnella, OECD (marco.paccagnella@oecd.org)

JT03451134

OECD EDUCATION WORKING PAPERS SERIES

OECD Working Papers should not be reported as representing the official views of the OECD or of its member countries. The opinions expressed and arguments employed herein are those of the author(s).

Working Papers describe preliminary results or research in progress by the author(s) and are published to stimulate discussion on a broad range of issues on which the OECD works. Comments on Working Papers are welcome, and may be sent to the Directorate for Education and Skills, OECD, 2 rue André-Pascal, 75775 Paris Cedex 16, France.

This document, as well as any data and map included herein, are without prejudice to the status of or sovereignty over any territory, to the delimitation of international frontiers and boundaries and to the name of any territory, city or area.

You can copy, download or print OECD content for your own use, and you can include excerpts from OECD publications, databases and multimedia products in your own documents, presentations, blogs, websites and teaching materials, provided that suitable acknowledgement of OECD as source and copyright owner is given. All requests for public or commercial use and translation rights should be submitted to rights@oecd.org.

Comment on the series is welcome, and should be sent to edu.contact@oecd.org.

This working paper has been authorised by Andreas Schleicher, Director of the Directorate for Education and Skills, OECD.

www.oecd.org/edu/workingpapers

© OECD 2019

Acknowledgements

The authors would like to thank William Thorn, François Keslair, Irwin Kirsch, Kentaro Yamamoto, Matthias von Davier and Kadriye Ercikan for valuable discussions and feedback on earlier drafts of this paper. The authors would also like to thank the OECD-PIAAC (Programme for the International Assessment of Adult Competencies) team for granting access to the data source and instruments in this study and thank the Center for Global Assessments and the Center for Psychometrics, Statistics, and Data Science at the Educational Testing Service for their support in the fellowship project. Larry Hanover provided generous help by proofreading the paper. Qiwei He acknowledges financial support from the Open Society Foundation through the OECD Thomas J. Alexander Fellowship Programme. Thanks are given to Matthew Gill and Deborah Fernandez who provided editorial support.

Abstract

The Survey of Adult Skills, a product of the OECD Programme for the International Assessment of Adult Competencies (PIAAC), used computers as the main assessment deliver platform. This enabled the Programme to collect data not only on whether respondents were able to solve specific tasks, but also on how they approached the problems at hand and how much time they spent on them. This paper draws on this information to characterise individuals' problem-solving strategies using the longest common subsequence (LCS) method, a sequence-mining technique commonly used in natural language processing and biostatistics. The LCS is used to compare the action sequences followed by PIAAC respondents to a set of "optimal" predefined sequences identified by test developers and subject matter experts. This approach allows studying problem-solving behaviours across multiple assessment items.

Résumé

L'Enquête sur les compétences des adultes, un produit du Programme de l'OCDE pour l'évaluation internationale des compétences des adultes (PIAAC), a utilisé l'ordinateur comme principal outil d'évaluation. Cela a permis de recueillir des données non seulement sur la capacité des répondants à résoudre des tâches spécifiques, mais aussi sur la façon dont ils ont abordé les problèmes en question et le temps qu'ils y consacrent. Le présent document s'appuie sur cette information pour caractériser les stratégies de résolution de problèmes des individus à l'aide de la plus longue sous-séquence commune (LCS), une technique de séquençage couramment utilisée dans le traitement du langage naturel et la biostatistique. La LCS est utilisée pour comparer les séquences d'actions suivies par les répondants du PIAAC à un ensemble de séquences prédéfinies « optimales » identifiées par les développeurs de tests et les experts en la matière. Cette approche permet d'étudier les comportements de résolution de problèmes à travers de multiples items d'évaluation.

Table of Contents

Acknowledgements.....	3
Abstract.....	4
Résumé	4
1. Introduction	7
1.1. Assessing problem-solving in technology-rich environments in PIAAC	8
1.2. The value of process data in large-scale assessments	10
2. Overview of process data analysis in problem-solving tasks	12
2.1. Extracting information from log files	12
2.2. Feature selection models	14
2.3. Identifying generalised patterns across items	16
3. How to identify generalised patterns across items.....	18
3.1. Longest common subsequence (LCS)	18
3.2. Multiple predefined action sequences	21
3.3. LCS measurement indicators	23
4. From theory to action: An exploration of problem-solving strategies in the PIAAC problem-solving assessment	25
4.1. Data	25
4.2. Items in the PS2 module	26
4.3. Application of the LCS method.....	27
5. Results	28
5.1. Comparison Design	28
5.2. Problem-solving strategies across different items	29
5.3. Association between the patterns of strategies and problem-solving proficiency	34
5.4. Differences by socio-economic background in problem-solving strategies	36
6. Discussion and conclusion.....	40
6.1. Summary	40
6.2. Limitations.....	41
6.3. Future studies	42
6.4. Conclusion.....	42
References	43
Annex A. Syntax of longest common subsequence in python.....	49
Annex B. Data preparation and cleaning	50

Tables

Table 1. Sample size per country by routing path to the PS2	26
Table 2. Similarity and consistency of similarity across five sample countries	27
Table 3. Descriptive statistics of similarity and consistency of similarity	32
Table 4. Descriptive statistics of proficiency level by similarity and consistency of similarity	36
Table 5. Descriptive statistics of ICT skills.....	39

Figures

Figure 1. Example item of the PSTRE assessment (single environment)	9
Figure 2. Example items of the PSTRE assessments (multiple environments)	10
Figure 3. An example of longest common subsequence derived from two sequences	19
Figure 4. Dynamic display of length of LCS derived by two sequences	21
Figure 5. An example of calculating LCS for an item with multiple reference sequences.....	23
Figure 6. Routing path between Module 1 and Module 2 in PIAAC Main Study	25
Figure 7. Distribution of similarity and consistency of similarity across items	28
Figure 8. Mapping similarity and consistency of similarity into one matrix	29
Figure 9. Frequency of groups by similarity and consistency of similarity	30
Figure 10. A scatterplot of interaction between similarity and consistency of similarity	30
Figure 11. Similarity and consistency of similarity across five sample countries	31
Figure 12. Efficiency and consistency of efficiency across five sample countries.....	33
Figure 13. A plot matrix for interactions between efficiency and similarity	34
Figure 14. Relationship between PSTRE proficiency level and general pattern in similarity	35
Figure 15. Share of men and women in different similarity groups	37
Figure 16. Gender differences in similarity across seven items in PS2.....	37
Figure 17. Distribution of age, by similarity and consistency of similarity	38

1. Introduction

The development of computer-based assessments has facilitated the measurement of technology-related skills and knowledge (Bennett et al., 2007^[1]; Bennett et al., 2010^[2]). The Survey of Adult Skills, a product of the Programme for the International Assessment of Adult Competencies (PIAAC), has been an invaluable resource in this respect, allowing researchers and policy makers to assess the ability of individuals to solve problems in technology-rich environments (OECD, 2012^[3]; Schleicher, 2008^[4]).

PIAAC was the first international assessment of adult skills predominantly administered on a computer. The computer administration was key to the assessment framework of the Problem Solving in Technology-Rich Environments (PSTRE) domain.

The use of computers as the delivery platform enables the development of new and innovative item types, such as interactive scenario-based items, and the collection of a broader range of information, including timing data and information on how test takers engage with the computers while completing assessment tasks (He and von Davier, 2016^[5]).

The aim of this paper is to explore how the behavioural information that is stored in log files in the course of the computer-based administration can be used to characterise problem-solving strategies of respondents and, in turn, explore differences in the strategies adopted by individuals in different countries and different population subgroups.

The data contained in log files, referred to as *process data* in the present study, provide information beyond response data (i.e. whether the final response was correct or not). Such information is particularly valuable when examining interactive problem-solving tasks because the sequences of actions undertaken by test takers can be used to identify the problem-solving strategies followed by individual respondents. However, it is important to note that, as valuable as these data can be, they are currently not much more than a by-product of the features of the software used to deliver the assessment. In particular, the choice of which information was to be recorded in log files was not driven by theoretical nor analytical considerations, and similarly, items were not explicitly designed with the purpose of identifying any particular strategy with the help of process data (OECD, 2019^[6]).

In the context of large-scale assessments, items designed to test problem-solving skills generally embed the problem within a particular context or situation. Therefore, the interpretation of process data for any individual item is highly dependent on the particular task being analysed and the context in which this item is embedded. Previous research has examined the problem-solving behaviour of individuals by analysing sequence patterns of individual items (Xu et al., 2018^[7]; He and von Davier, 2016^[5]; He and von Davier, 2015^[8]; Greiff, Wüstenberg and Avvisati, 2015^[9]). However, important insights can be gained by investigating generalised patterns of respondents' behaviours across multiple tasks (i.e. multiple assessment items), embedded in different contexts and scenarios.

An important challenge for researchers willing to identify and characterise respondents' problem-solving strategies is how to define variables that have a consistent meaning across items and derive standardised measures in complex data structures across multiple items. This paper uses the longest common subsequence (LCS) method, a sequence-mining technique used in natural language processing and biostatistics (Sukkarieh, von Davier and

Yamamoto, 2012_[10]), to identify the action sequences that are most similar to predefined, optimal sequences for each item. Measurement indicators are developed in order to analyse behaviours across items and subgroups of respondents. This approach extends the research capacity from understanding adults' problem-solving behaviours in a single item to a general perspective across multiple items that form an assessment.

Based on results derived from the LCS method, the paper investigates the relationship between adults' problem-solving strategies and their proficiency level. Comparisons are made among countries, as well as among population subgroups within each country. Specifically, this study pursues three research questions:

- Do test takers adopt consistent problem-solving strategies across different items?
- What is the association between the adoption of specific patterns of problem-solving strategies and problem-solving proficiency?
- Do problem-solving strategies differ systematically by gender, age and ICT familiarity?

This paper contains six sections. The present section describes the PSTRE assessment and explains the value-added of analysing process data in international large-scale assessments. Section 2 presents a short overview of research methods that have been employed to analyse process data in the context of large-scale assessments. Section 3 illustrates the LCS method, which is then applied to the PIAAC PSTRE assessment in Section 4. Section 5 presents the results of a case study on the application of LCS application in the PIAAC assessment. Concluding remarks are presented in Section 6, as well as perspectives about future directions for research.

1.1. Assessing problem-solving in technology-rich environments in PIAAC

PIAAC surveys adults aged 16 to 65 and assesses their proficiency in literacy, numeracy and problem-solving in technology-rich environments. The study was administered in over 40 countries between 2011 and 2017.

PSTRE was an assessment domain introduced for the first time in PIAAC. It was defined as the ability to "(...) use digital technology, communication tools and networks to acquire and evaluate information, communicate with others and perform practical tasks". PSTRE items focused on the ability to solve problems for personal, work and civic purposes, by setting up appropriate goals and plans, and accessing and making use of information through computers and computer networks (OECD, 2009_[11]).

Assessing PSTRE required the use of items with peculiar characteristics. In particular, they needed to be interactive and placed in a digital framework. For these reasons, the PSTRE assessments was only administered on a computer, contrary to literacy and numeracy, which had a paper-based version for respondents unable or unwilling to take the assessment on a computer.

The interactive nature of PSTRE items makes them ideal candidates for analyses based on process data (Goldhammer, Naumann and Keßel, 2013_[12]). The assessment of PSTRE required individuals to engage with complex tasks: for instance, they were required to click buttons or links, select from dropdown menus, drag and drop, copy and paste, and so on. These multitudes of different actions (i.e. all interactions between the respondent and the testing platform) were recorded in log files.

Figure 2. Example items of the PSTRE assessments (multiple environments)

Note: Page A shows an opening screen of a job search task. Page B shows a website where relevant information regarding fees and registration can be found but is not shown without clicking further. Page C is the second page of the website shown in (B), where details on fees and registration are located in the directions for the form.

Source: Available at www.oecd.org/skills/piaac/Problem%20Solving%20in%20TRE%20Sample%20Items.pdf.

1.2. The value of process data in large-scale assessments

Process data were initially collected and used for the purpose of cross-validation of response data. For instance, in the case a large amount of missing data were observed for a particular item, timing data were used to support quality control checks, by allowing researchers to detect the possible reasons for the high incidence of missing answers and thus suggesting a more accurate approach for measurement (OECD, 2017_[13]).

More recently, the use of process data has extended beyond cross-validation, shifting from mere detection of problematic items to the exploration of respondents' test-taking behaviours (Lee and Jia, 2014_[14]; Goldhammer, Naumann and Keßel, 2013_[12]; Liao, He

and Jiao, 2019^[15]; He and von Davier, 2016^[5]; He, von Davier and Han, 2018^[16]). Indeed, process data are fuelling an entirely new strand of research. A variety of reasons makes these data particularly valuable for research purposes.

First, process data provide supplementary information that cannot be easily observed from response data. Such information has been used, for example, to identify the approaches respondents engage in when responding to questions in a test (Goldhammer, Naumann and Keßel, 2013^[12]). This information can be derived from keystrokes used to answer a question, time spent on a test question or task, and test takers' interactions with test stimulus materials (Guo et al., 2018^[17]; Zhu and Feng, 2015^[18]; Liu, Liu and Li, 2018^[19]; Chen et al., 2019^[20]).

Second, process data can yield new insights on differences in the problem-solving strategies adopted by individuals at different levels of proficiency. Moreover, it is possible to examine if there are systematic variations in the problem-solving strategies adopted by individuals with a different cultural or linguistic background (between-country variation) or across different population subgroups. Whether test takers from different cultural backgrounds adopt the same solution to a complex task, and whether the differences in strategies have an association with test takers' response style in the item-based tests, are important issues that have a potential bearing on the validity of comparisons across groups (Liao, He and Jiao, 2019^[15]; He and von Davier, 2016^[5]).

Third, process data are valuable in optimising test design and improving measurement accuracy. Leveraging information from process data has a great potential for assessment, ranging from validating or improving the validity of the test scores (Sireci and Zenisky, 2006^[21]) to furnishing evidence of new constructs that cannot be easily captured by the test scores such as metacognitive skills (van der Linden, 2005^[22]). He and von Davier (2016^[5]) demonstrate that process data can be useful in detecting nonresponse due to low engagement of test takers and can be helpful as a diagnostic method to examine items (e.g. redundant space between key information). Process data also hold the promise of developing a better understanding of individuals' thought processes and strategies during testing and, therefore, developing better tests, deeper understanding of student knowledge, skills, and competencies, and providing richer information that may guide student learning.

The motivation for pursuing innovative methods in process data analysis is promoted with the recognition of the potential value contributed by this new data source, although it is important to consider that the process data used in this paper are limited to the information collected through human-to-computer interactions. Behaviour that occurs without being recorded by a computer, for example the use of an external calculator or a note pad, remains unobservable. Section 2 will present a brief overview of the methods that have been developed during the past decade related to process data analysis in problem-solving items. It does not necessarily cover all the methods in this field, but does show tendencies that have shown up in methodology development.

2. Overview of process data analysis in problem-solving tasks

The complex structure of process data challenges traditional data analysis methods such as classic latent variable models or item response models (Rasch, 1960^[23]; Lord, 1980^[24]; Chen et al., 2019^[20]). The challenges mainly come from two aspects. First, in classic tests, the item responses are often univariate categorical variables. As the technology advances, the entire problem-solving process that occurs because of an interaction with the computer platform can be recorded. This results in a more complex format of data, containing an unordered set of responses and an ordered sequence of actions, which is hard for classic latent variable models to describe. Second, complex cognitive processes are often involved in problem-solving tasks, which are hardly completely governed by only a few latent factors, especially in the presence of high-dimensional responses. Meanwhile, including too many latent factors typically results in a model that is difficult to estimate and interpret.

Given these concerns, new methods are currently being explored and compared to extend classical latent variable models so that new data formats can be accommodated. These challenges promote interdisciplinary studies that integrate techniques from a broader range of disciplines, such as data mining, machine learning, natural language processing (NLP), social networking and sequence-mining. These methods are typically applied when handling big data with complex structures and feature in models that bridge latent relationships and observed data (Manning and Schütze, 1999^[25]; Forman, 2003^[26]), thus holding promises for allowing the analysis of the process data that share similar structures. Among them, extraction and selection of features are the two dominant directions that researchers have followed in the analysis of process data.

2.1. Extracting information from log files

Information collected in log files can be roughly categorised into three groups:

1. behavioural indicators that represent respondents' problem-solving strategies or interactions with computer
2. actions and mini action sequences that are directly extracted from test takers' process data
3. timing data, such as the duration of the test, the time spent in the simulation environment, and the time elapsed before the first interaction of the respondent with the task.

2.1.1. Behavioural indicators

Behavioural indicators are typically recorded at a higher, aggregated level; very often, user actions are considered single events, while the activities initiated in the underlying operating system are already a complex combination that involves changing the visual appearance of the user interface as well as the internal status of the software.

Although human-computer interactions are often accomplished through simple gestures or movements, in most cases these are not automated actions but involve case-based reasoning and self-regulatory processes (Azevedo, 2005^[27]; Lazonder and Rouet, 2008^[28]; Brand-Gruwel, Wopereis and Walraven, 2009^[29]; Shapiro and Niederhauser, 2004^[30]; Zimmerman, 2008^[31]; Winne and Baker, 2013^[32]; Bouchet et al., 2013^[33]). Therefore, to

perform well on computer-based problem-solving tasks present in the PIAAC PSTRE assessment – such as searching for information on the Internet, organising folders, extracting information from large datasets and shifting across software applications and windows – both basic ICT skills and higher-level problem-solving skills are required. Interacting with computers involves the process of low-level comprehension of symbolic information. Users have to decode and understand menu names or graphical icons in order to follow the appropriate chain of actions to reach a goal. In many cases, problem-solving also includes localising information, comprehending larger digital texts, or synthesising textual information from different sources or web pages.

At the same time, problem-solving tasks require higher-order thinking, such as finding new solutions, and interacting with a dynamic environment (Klieme, 2004^[34]; Mayer, 1994^[35]; Mislevy et al., 2012^[36]; Goldhammer et al., 2014^[37]). A typical example is the “vary one thing at a time” strategy (VOTAT) that was applied in Greiff, Wüstenberg and Avvisati (2015^[9]). In that study, a behavioural indicator – whether the respondent followed a VOTAT strategy or not – was found to be highly correlated with performance. Solving complex tasks requires developing a plan consisting of a set of properly arranged subgoals and performing corresponding actions to attain the final goal. This differs from solving logical or mathematical problems, where complexity is determined by reasoning requirements but not primarily by the information that needs to be accessed and used (Goldhammer, Naumann and Keßel, 2013^[12]). In this sense, one could argue that the indicators of user actions should in some systematic way map onto the subgoals a user develops and applies to successfully complete an assessment task.

2.1.2. Action sequences from process data

Identifying sequential patterns in learning activity data can be useful for discovering, understanding, and, ultimately, scaffolding human learning behaviours. The primary task, as applied in a variety of domains including education and training, is to discover patterns that are commonly shared by many sequences within a dataset (Agrawal and Srikant, 1995^[38]). Some researchers have employed sequential pattern mining to customise digitally delivered learning services (e.g. language or math courses) to individual students¹. Other researchers have employed sequential pattern mining to better understand learning behaviour in particular conditions for groups, for instance, collaborative problem-solving patterns by groups in learning environments and web usage patterns by genders (Baker and Yacef, 2009^[39]; Martinez et al., 2011^[40]; Andersen, Gulwani and Popovic, 2013^[41]; Zou et al., 2010^[42]). Ideally, these patterns provide a basis for generating models and insights about how individuals learn, solve problems, and interact with their environments. Algorithms for mining sequential patterns generally associate measures of frequency to rank identified patterns based on how often a certain pattern is repeated. Investigation of the frequency with which a pattern occurs over time can reveal additional information for interpretation. Further, these changes in occurrence patterns may help identify more important patterns that occur only at certain times or become more or less frequent than those that occur frequently but uniformly over time (Kinnebrew, Mack and Biswas, 2013^[43]).

¹ Examples include the Cognitive Tutor Authoring Tools, which learn students’ problem-solving sequences and provide adaptive guidance to students to learn math in middle and high schools (Alevan, McLaren and Sewall, 2009^[72]), or the Auto Tutor system, which stimulates a learning environment through tutorial dialogues in natural language (Graesser et al., 2004^[71]).

Various approaches can be used to conduct sequential mining. Biswas, Jeong, Kinnebrew, Sulcer, and Roscoe (2010^[44]) use hidden Markov models (Rabiner, 1989^[45]; Fink, 2008^[46]) as a direct probabilistic representation of the internal states and strategies. This methodology facilitates identification, interpretation, and comparison of student learning behaviours at an aggregate level. As with students' mental processes, the states of a hidden Markov model (HMM) are hidden, meaning that they cannot be directly observed, but produce observable output (e.g. actions in a learning environment). Poon, Kong, Wong, and Yau (2017^[47]) use sequential pattern mining (SPM) to uncover students' navigation patterns on a learning platform. This method helped to further explore the students' strategy patterns with educational implications, including quiz-performance oriented (e.g. patterns dependent on test score), resource-oriented (e.g. patterns dependent on changes before and after learning with video source) and evaluation-oriented (e.g. response patterns on surveys) sequential patterns.

As Fink (2008^[46]) pointed out, the development and spread in the use of sequential models was closely related to the statistical modelling of texts as well as the restriction of possible sequences of word hypotheses in automatic speech recognition. A number of methods from these fields can be borrowed and applied in the analysis of process. For instance, He & von Davier (2015^[8]; 2016^[5]) adopted the n-gram to disassemble the sequence of data and the chi-square feature selection model to extract action sequence patterns that facilitate differentiation between performance groups.

2.1.3. Features generated from timing data

In addition to sequential data on actions taken by respondents, computer-based assessments provide rich data on response latency or timing data. Each action log entry is associated not only with data on what a respondent did, but also on when the action took place. These timestamps can be aggregated into an overall time measure for the survey, a time measure that is specific to the task, or measures that are specific to certain types of interactions such as keystrokes, navigation behaviour, or time taken for reading instructions. Timing data at this level of resolution has led to renewed interest in how latency can be used in modelling response processes (van der Linden, Klein Entink and Fox, 2010^[48]; Demars, 2007^[49]; Weeks, von Davier and Yamamoto, 2013^[50]). The analysis of timing data from the Programme for International Student Assessment (PISA) suggests that the association between timing data and response outcomes will require item-level attributes to be included (for instance by response type such as multiple choice or open-ended questions) as well as respondent attributes to more fully utilise timing data (OECD, 2017^[13]). In addition, timing data information is expected to be valuable in conjunction with the types of actions observed in the sequence data and to help deriving features that allow predicting cognitive outcomes such as test performance (He et al., 2018^[51]; Liao, He and Jiao, 2019^[15]).

2.2. Feature selection models

In machine learning, feature selection refers to the automatic process guiding the identification of variables relevant for the problem at hand. Feature selection models can help researchers to identify robust indicators that distinguish different groups. In the context of large-scale assessments, feature selection models can help to identify if different respondents adopted the same problem-solving strategy or not. A variety of models that have been developed in "big data" fields relate to information retrieval, NLP, and data mining. In general, feature selection models are also used for mutual validation. Typically, at least two selection models are used in one study to check whether the features are

selected consistently by different methods. Cross-validation is highly recommended if only single feature selection model could be used.

As reviewed by Forman (2003_[26]) and Guyon & Elisseeff (2003_[52]), feature selection approaches can be classified in the following categories: wrappers, filters and embedded methods. Wrappers utilise the learning machine of interest as a black box to score subsets of variable according to their predictive power. Filters select subsets of variables as a pre-processing step, independently of the chosen predictor. Embedded methods perform variable selection in the process of training and are usually specific to given learning machines. The following subsections discuss these three major feature selection approaches in detail and shed light on some feature selection models that are commonly used in settings that are similar to that investigated in this paper.

2.2.1. Wrapper methods

Wrapper methods offer a simple and powerful way to address the problem of variable selection, regardless of the chosen learning machine (Kohavi and John, 1997_[53]). These methods are mostly used in sequential forward selection or genetic search. They perform an exhaustive search over the space of all possible subsets of features (Guyon and Elisseeff, 2003_[52]). In their most general formulation, wrapper methods consist in using the prediction performance of a given learning machine to assess the relative usefulness of subsets of variables. Wrapper methods are especially capable when dealing with low-dimensional data, but often turn out not to be suitable for more complex, large-scale problems, as in those cases computations can quickly become intractable (Forman, 2003_[26]). As process data are typically distributed in high-dimensional structures, wrapper methods are typically not recommended for their analysis. Given the focus of this paper, an extensive description of wrapper methods is not presented here. Interested readers are directed to (Guyon and Elisseeff, 2003_[52]) for more details.

2.2.2. Filter methods

Filter methods score each potential feature according to a particular feature selection metric, and then extract the most discriminating features to distinguish target classes. Scoring involves counting the occurrences of a feature in training positive- and negative-class training examples separately, and then computing a function of these (Forman, 2003_[26]). Commonly known metrics (e.g. Chi-square statistics, accuracy, F-measure, information gain) could be used as criteria to select important features. Compared with the heavy computation involved in wrapper methods, filter methods are usually more tractable and faster. Two feature selection methods that have been successfully implemented in the analysis of process data are the chi-square selection model (CHI, e.g. (He and von Davier, 2016_[5])) and the weighted log likelihood ratio test (WLLR, e.g. (He and von Davier, 2015_[54])). A summary of feature selection metrics is contained in Forman (2003_[26]).

The use of CHI (Oakes et al., 2001_[55]) is recommended in textual analysis, due to its high effectiveness in finding robust keywords and for testing similarities between different text corpora (Manning and Schütze, 1999_[25]; He, Veldkamp and de Vries, 2012_[56]; He et al., 2014_[57]). The structural similarity between textual data and process data makes this method particularly appropriate to detect those actions or action vectors that are highly informative for distinguishing different performance groups (He and von Davier, 2016_[5]). The best classifiers are the features with the highest chi-square values.

The WLLR is defined as the product of the probability of each action sequence and the logarithm of the ratio between the conditional probabilities of the sequence in different performance groups. The WLLR of each action sequence can be used for the purpose of binary classifications, in analogy with its use in feature selection in text categorisation (He and von Davier, 2015^[8]; Nigam et al., 2000^[58]). The WLLR is usually used as an alternative model for CHI, in order to make mutual validation in feature selection.

2.2.3. Embedded methods

Embedded methods incorporate variable selection as part of the training process. Compared with wrapper methods, they are more efficient in several respects: they make better use of the available data by not needing to split the training data into a training and validation set, and they reach a solution faster by avoiding retraining a predictor for every variable subset investigated (Guyon and Elisseeff, 2003^[52]). Decision trees such as the classification and regression tree algorithm (CART), for instance, have a built-in mechanism to perform variable selection (Breiman et al., 1984^[59]).

The random forest algorithm (Breiman, 2001^[60]) is an extension of decision trees. It is an ensemble learning method that increasingly adjusts itself by randomly combining a number of single decision tree algorithms. Random forests improve overall prediction accuracy by aggregating the prediction results obtained from simple trees, thus reducing prediction variance (Dietterich, 2000^[61]). The construction of random forests (RF) starts from bootstrapping a number of samples based on the entire sample dataset. The data entries that are not chosen in each draw are called the out-of-bag data and are kept for validation purposes. For each draw, a binary decision tree is fitted on the bootstrapped sample. Growing a decision tree requires finding the best way to split the parent node on the basis of a chosen predictor variable. The “best split” is measured by the purity of the two child nodes. Classification accuracy and sum of squared residuals are typical options for measuring purity in classification and regression problems.

In a recent study by Han, He and von Davier (2019^[62]), the random forest method was employed to explore the association between response data and predictive features that were derived from process data in one problem-solving item of PISA 2012. The selected sample was initially employed to fit a random forest using 77 features. By using a recursive approach (i.e. excluding one feature at each time from the end of the variable importance rank and refitting random forest on the basis of the remaining features), the backwards elimination resulted in a reduced model with 16 features. The averaged accuracy of the reduced model was 0.842, against 0.883 for the full model.

2.3. Identifying generalised patterns across items

The short overview above provides evidence of the rapid methodological developments in the analysis of process data that occurred in the past few years. Methods that identify generalised patterns across items are still at a preliminary stage, however. In particular, there is much room for improvement in the development and application of methods that allow to easily generalise the analysis of process data across items.

Two major challenges need to be confronted in identifying generalised patterns across items. First, the features derived from process data are usually item dependent and their specificity is what makes process data valuable. However, in order to generalise across items, such specificity is typically an obstacle in the construction of aggregate-level variables. Aggregate-level variables need to reflect the characteristics of behaviour across

items while representing how individuals engaged with a specific item (otherwise dichotomous correct-incorrect aggregate-level variables would suffice). Second, it is preferable to treat action sequences in each item as a whole, rather than disassembled variables such as n-grams (He and von Davier, 2016^[5]) or sequential factors in HMMs (LaMar, 2018^[63]), because disassembled variables are mainly specific to one item.

Identifying generalised patterns across a large number of items is challenging but not impossible. One possibility is to derive indicators that are based on the distance, for each item, between the observed and a predefined strategy. This paper represents a first attempt to stimulate research in this direction. After using the longest common subsequence (LCS) method to identify the strategies most likely followed by respondent in each item, it is possible to compute the distance between such individual sequences and a reference sequence for each item. This distance is more likely to have the same meaning across different items.

3. How to identify generalised patterns across items

This section introduces a sequence-mining approach – the longest common subsequence (LCS) method – that can be used to generalise action patterns across multiple tasks. The basic idea behind the LCS approach as used in this paper is to identify how different the sequence of actions taken by a test taker to solve a particular test item is from the action sequence that subject matter experts consider optimal to solve the problem at hand. This can be done by computing the distance between individual action sequences, namely, observed sequences (OS), and expert-predefined action sequences or “reference sequences” (RS). Reference sequences are identified by expert judgements on the basis of the relative efficiency of different solution paths, and how time-consuming these are expected to be. Reference sequences are “optimal” in the sense that they represent the theoretical range of sequences that are expected to be the most efficient, or the less time-consuming, way to arrive at the solution. For any one item/assessment task, it is in principle possible to identify multiple optimal action sequences.

Once optimal sequences are defined, it is possible to compute the distance between each respondent’s sequence and the optimal one/range of optimal ones for all items in a test. The mean value of each person’s distance between the observed sequences and the reference sequences² across all test items is defined in the context of this paper as the *similarity* between individual strategies and the predefined ones. The standard deviation of the distribution of individual distances is defined as the *consistency* of problem-solving strategies. The similarity indicator in the paper illustrates how close, on average, an individual is to adopting optimal action sequences. The consistency indicator illustrates the degree to which respondents can be reliably expected to follow optimal action sequences.

3.1. Longest common subsequence (LCS)

The LCS of a set of sequences (often just two sequences) is a subsequence, whose length equals the maximum number of actions that are shared, in sequential order, with the original sequences (Cormen et al., 2001_[64]; Sukkarieh, von Davier and Yamamoto, 2012_[10]). It differs from the longest common substring approach, where the subsequence is not required to occupy sequential positions within the original sequences.

The way the method works is best illustrated with the example reported in Figure 3. The figure reports an observed sequence from one respondent *X* (ACCGGTGGACAATTCA) and a reference sequence *Y*, (GGAAGAGATATGCAC), predesigned as optimal by subject matter experts and item developers. The LCS is “AAAATTCA”, i.e. the longest string of actions that are present in a sequential order in both *X* and *Y*.

² In the case of items for which multiple optimal action sequences can be identified, in the paper we draw comparisons between the manifest actions of a respondent in that item and the optimal action sequence that most closely resemble the manifest behaviour, therefore selecting the sequence that minimises the difference between the manifest and the optimal action sequence.

Figure 3. An example of longest common subsequence derived from two sequences

Note: X is defined as an observed sequence from one respondent, while Y is defined as a reference sequence that has been predefined by item developers. The length of X and Y is equal at 16. The last row indicates the LCS derived from X and Y , which is the maximum sequence that both of X and Y possess following a sequential order. The length of the LCS is 8.

Before heading to the definition of LCS function, one important term, *prefix*, needs to be introduced. The prefix of a sequence is the first portion of the sequence. Let S be the sequence (AGCA). To record the information properly, prefixes can be denoted with the name of the sequence, followed by a subscript to indicate how many characters the prefix contains (Xia, 2007_[65]). In this paper, the prefix function is used to denote a subsequence, up to the specified maximum length, containing the initial elements of the sequence. Therefore, the prefix (A) is denoted as S_1 , the prefix (AG) is denoted as S_2 because it contains the first two elements of sequence S , and so on. The possible prefixes for sequence S are listed below:

$$S_1 = (A)$$

$$S_2 = (AG)$$

$$S_3 = (AGC)$$

$$S_4 = (AGCA)$$

The algorithm for identifying the LCS (Cormen et al., 2001_[64]) is defined as follows.

Let $X = (x_1, x_2, \dots, x_i)$ and $Y = (y_1, y_2, \dots, y_j)$ be two sequences. x_i and y_j are actions within the sequence X and Y , respectively. The prefixes of X and Y are $X_1, X_2, \dots, X_i$ and $Y_1, Y_2, \dots, Y_j$, respectively. Let $LCS(X_i, Y_j)$ represent the set of longest common subsequence of prefixes X_i and Y_j . The set of sequences is given as:

$$LCS(X_i, Y_j) = \begin{cases} \emptyset & \text{if } i = 0 \text{ or } j = 0 \\ LCS(X_{i-1}, Y_{j-1}), x_i & \text{if } x_i = y_i \\ \text{longest } (LCS(X_i, Y_{j-1}), LCS(X_{i-1}, Y_j)) & \text{if } x_i \neq y_i \end{cases} \quad (1)$$

A dynamic programming algorithm is used to find the longest subsequence common to X and Y . The algorithm works in two phases. In the first phase, the algorithm constructs an i

by j table T (see Figure 4) containing, for each table cell $t_{i,j}$, the length of the longest subsequence shared by X_i and Y_j , so that $t_{i,j} = \text{length}(LCS(X_i, Y_j))$. In the second phase, the longest subsequence is identified by stepping backwards through this table and seeing which elements increased the sequence length. The length of LCS is defined as:

$$\text{length}(LCS(X_i, Y_j)) = \begin{cases} 0 & \text{if } i = 0 \text{ or } j = 0 \\ \text{length}(i-1, j-1) + 1 & \text{if } x_i = y_i \\ \max(\text{length}(i, j-1), \text{length}(i-1, j)) & \text{if } x_i \neq y_i \end{cases} \quad (2)$$

As illustrated in Figure 4, the top-most row and left-most column are first zeroed. Trivially, if one of the sequences has no elements, the longest common subsequence is of length zero. Then, one goes through the rest of the table sequentially, filling out the table as per Formula (2). If $x_i = y_i$, the longest subsequence leading up to i, j can be extended by one element, the element is set to $t_{i,j} = \text{length}(i-1, j-1) + 1$. Otherwise, it is the same as the longest subsequence already up to the element as $t_{i,j} = \max(\text{length}(i, j-1), \text{length}(i-1, j))$. Once the table is constructed, the final element contains the length of the longest subsequence.

The highlighted numbers in Figure 4 show the path followed to derive the maximum length of the LCS, which can be followed backward from the bottom right to the top left corner when reading out an LCS. If the current symbols in X and Y are equal, they are part of the LCS, and therefore one goes both up and left (shown in bold). If not, one goes either up or left, depending on which cell has a higher number. A Python implementation of this algorithm is shown in Annex A.

Note that in the examples shown in Figure 3 and Figure 4, the length of two action sequences are the same. The maximum length of the LCS equals the length of each action sequence. In case of unequal length of sequences, the maximum length of LCS equals the length of the shorter one between the two sequences.

Figure 4. Dynamic display of length of LCS derived by two sequences

		0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
		∅	A	C	C	G	G	T	G	G	A	C	A	A	T	T	C	A
0	∅	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	G	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	G	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	A	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
4	A	0	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2
5	A	0	2	2	2	2	2	2	2	2	2	2	3	3	3	3	3	3
6	G	0	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
7	A	0	3	3	3	3	3	3	3	3	3	3	3	4	4	4	4	4
8	G	0	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
9	A	0	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
10	T	0	4	4	4	4	4	4	4	4	4	4	4	4	5	5	5	5
11	A	0	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
12	T	0	5	5	5	5	5	5	5	5	5	5	5	5	6	6	6	6
13	G	0	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6
14	C	0	6	6	6	6	6	6	6	6	6	6	6	6	6	6	7	7
15	A	0	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	8
16	C	0	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8

Note: Figure 4 adopts the same example shown in Figure 3. The observed sequence X and the reference sequence Y follows the horizontal and vertical direction, respectively. The current length of LCS by the two sequences is shown in each cell. The LCS is bolded and labelled in blue. The highlighted grey cells present the trace of deriving LCS from the two sequences.

3.2. Multiple predefined action sequences

As shown in Figure 3, to derive the LCS for each respondent per item, it is important to identify:

- (1) the observed sequence, that is, the sequence of action that a given individual has followed to solve a specific item
- (2) the reference sequence, that is, the expert-predefined action sequences for a particular item (which are not restricted to be unique)

In the case of large-scale assessments, action sequences denote the patterns of actions that individuals undergo while solving an item. As most problem-solving items allow respondents to take multiple paths to arrive at the same response, it is possible to have multiple “optimal” sequences. Different paths can be, of course, more or less efficient (e.g. they may require a different number of actions or take a different amount of time). For instance, to identify a specific entry in a spreadsheet, respondents can use the search or sort function, or even scroll up and down to spot the entry merely by eyes without using any tool. To adapt the LCS method to such a more complex situation with multiple solutions for one task, Formula (1) should be extended as follows.

Let $X = (x_1, x_2, \dots, x_i)$ be the observed individual action sequence and let $\mathbf{Y} = \{Y_k\}$ be a set of reference sequences. Let $Y_k = (y_{k1}, y_{k2}, \dots, y_{kj})$ be a certain reference sequence within $\mathbf{Y}$. The prefixes of X and Y_k are $X_1, X_2, \dots, X_i$ and $Y_{k1}, Y_{k2}, \dots, Y_{kj}$, respectively. Following Formula (1), $LCS(X_i, Y_{kj})$ can be derived by an individual action sequence matched with each predefined sequence by k times. The longest $LCS(X_i, Y_{kj})$ is defined as

the individual's eventual LCS. The Y_{kj} that creates the longest LCS is the approach that the individual followed with the maximal similarity.

$$LCS(X, Y) = \text{longest} \left(LCS(X_i, Y_{kj}) \right) \quad (3)$$

If the LCSs have maximal equal length, they are labelled $LCS(X, Y) = 0$, because it is not possible to distinguish which potential reference sequence the respondent followed. To a large extent, it is possible that the respondent did not follow any reference sequence, but rather created a new possible path to solve the problem. This often occurred in a short-observed action sequence, when no key actions could be matched with the reference sequence.

Figure 5 presents an example of deriving the LCS from multiple solutions in one PSTRE item (U19a) in PIAAC. Two environments are involved in this item: a spreadsheet and an email. Test takers must find the correct information in the spreadsheet page, input the information in an email page, and then send such information to a specified agent. Four sequences (i.e. RS_1 to RS_4) were defined by experts as proper ways to solve this task, all equally optimal³. The first two RSs show the strategy of using the search function. Regardless of whether the search function is accessed from the toolbar or menu item, the length of the RS is the same at 11, meaning that to solve this item by using the search tool, one has to take at least 11 steps. Analogously, the third and fourth RS indicate the solving strategy via the sort function. The length of RS_3 and RS_4 both equal 9, meaning that to use the sort function, one has to follow at least nine steps. A typical individual action sequence to solve this problem is presented in the observation box. This specific person used 25 steps to complete the task. 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 has the maximal similarity with RS_4. Therefore, LCS4 is defined as this individual's problem-solving solution in the item U19a. The actions highlighted in bold font in the observed sequence match the actions in the LCS4, the eventual LCS in this item.

³ Two sequences are equally optimal if at each decision node in which respondents need to take an action, two actions are equally optimal. In other words, the respondent cannot expect one or the other to lead to fewer completion steps or to a faster solution path.

Figure 5. An example of calculating LCS for an item with multiple reference sequences

RS_1: searching from toolbar (length=11) Start, Toolbar_SS_Find, On_SearchBox, Off_SearchBox, Search_OK, SS_SEARCH, Email, On_Email_Message, Off_Email_Message, Next, Next_OK
RS_2: searching from menu item (length=11) Start, MenuItem_Find, On_SearchBox, Off_SearchBox, Search_OK, SS_SEARCH, Email, On_Email_Message, Off_Email_Message, Next, Next_OK
RS_3: sorting from toolbar (length=9) Start, Toolbar_SS_Sort, Sort_1_B, Sort_OK, Email, On_Email_Message, Off_Email_Message, Next, Next_OK
RS_4: sorting from menu item (length=9) Start, MenuItem_Sort, Sort_1_B, Sort_OK, Email, On_Email_Message, Off_Email_Message, Next, Next_OK
OBSERVATION (length=25) Start, Toolbar_SS_Help, Menu_SS_Edit, Menu_SS_Data, MenuItem_Sort, Sort_1_B, Sort_1A, Sort_OK, SS_Sort_1B , Email, On_Email_Message, Off_Email_Message, SS, On_Email_Message, Off_Email_Message, Email, On_Email_Message, Off_Email_Message, Toolbar_E_Send, On_Email_Message, Off_Email_Message, Next, On_Email_Message, Off_Email_Message, Next_OK
LCS1 (length=6): Start, Email, On_Email_Message, Off_Email_Message, Next, Next_OK LCS2 (length=6): Start, Email, On_Email_Message, Off_Email_Message, Next, Next_OK LCS3 (length=8): Start, Sort_1_B, Sort_OK, Email, On_Email_Message, Off_Email_Message, Next, Next_OK LCS4 (length=9): Start, MenuItem_Sort, Sort_1_B, Sort_OK, Email, On_Email_Message, Off_Email_Message, Next, Next_OK

Note: The actions highlighted in bold font are the eventual LCS derived for the certain respondent.

Some problem-solving items, on the other hand, require following a single unique strategy in order to arrive at the correct solution (Hao, Shu and Davier, 2015^[66]; Han, He and von Davier, 2019^[62]). In such a case, the action sequence is highly correlated with the final response. It implies that if a respondent missed a key action, their chances of giving the correct answer would be greatly reduced. For items designed in this way, it is straightforward to compute the LCS between individual's sequence and the unique reference sequence. However, the items with only one solution would produce little information on individuals' problem-solving strategies because of the strict item-design constraints.

3.3. LCS measurement indicators

In order to aggregate the information extracted from the LCS across items and groups, a set of measurement indicators are developed, covering two major dimensions.

3.3.1. Similarity and consistency of similarity

For each item, similarity of the individual action sequence to the reference sequence is defined as the ratio between the length of LCS and the length of reference sequence. In other words, similarity measures how big a proportion of the common subsequence can be found in the reference sequence. As it is always true that $len(LCS) \leq len(RS)$, similarity will always lie in $[0,1]$. The extreme value 1 indicates a perfect match between the individual sequence and the reference sequence, in which case the length of the LCS will be the same as the length of the reference sequence. Conversely, when the ratio equals 0, no overlap between the LCS and reference sequence exists, thus the individual action sequence would have no match to the reference sequence. In general, the higher the ratio is, the more similar the individual sequence to the reference sequence.

$$\text{Similarity} = \text{len}(LCS)/\text{len}(RS) \quad (4)$$

The consistency indicator aims to capture to what extent a respondent is able to follow the reference sequences across different items. In order to do so, it is first necessary to look at the distribution of similarities for each person across items. The mean of this distribution (SM) is defined as the similarity across n items. A higher value of SM indicates that a respondent on average solves problems by following the reference sequences closely

$$SM = \text{Mean}(Sim_1, Sim_2, \dots, Sim_n) \quad (5)$$

In order to see whether a respondent follows the predefined sequence in a consistent way, the standard deviation of this distribution (SSD) is used as an indicator for consistency of similarity. A lower value of consistency implies that the person follows a stable pattern across items, while a higher value suggests an unstable pattern. For instance, a person might solve one problem with a high similarity to the reference sequence but show a high dissimilarity for a second item.

$$SSD = SD(Sim_1, Sim_2, \dots, Sim_n) \quad (6)$$

3.3.2. Efficiency and consistency of efficiency

Efficiency of a solution strategy (or of a sequence of action) is also an important element. The definition of efficiency used in this paper does not make use of timing information. Rather, efficiency is defined as the ratio between the length of LCS and the length of observed sequence (OS). This indicator measures to what degree the LCS and the actual observed sequence overlap. As it is always true that $\text{len}(LCS) \leq \text{len}(OS)$, the ratio between them would be a number within the range $[0,1]$. 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 optimal sequence.

$$\text{Efficiency} = \text{len}(LCS)/\text{len}(OS) \quad (7)$$

Analogously to the definition of the similarity indicator stated in subsection 3.3.1, it is possible to analyse the distribution of efficiencies (for the various items solved by a given individual), and compute the mean (EM) and the standard deviation (ESD) of this distribution.

$$EM = \text{Mean}(Eff_1, Eff_2, \dots, Eff_n) \quad (8)$$

$$ESD = SD(Eff_1, Eff_2, \dots, Eff_n) \quad (9)$$

4. From theory to action: An exploration of problem-solving strategies in the PIAAC problem-solving assessment

This section illustrates how the LCS method can be applied to explore general patterns in multiple problem-solving items, using process data on seven PSTRE items from five countries that participated in PIAAC. The LCS method is used here to investigate the relationship between adults' problem-solving strategies and their proficiency level, and to make comparisons among countries and population subgroups within each country. Specifically, this and the following two sections will tackle three research questions:

- Do people adopt consistent problem-solving strategies across different items?
- What is the association between the adoption of specific patterns of problem-solving strategies and problem-solving proficiency?
- Do patterns of problem-solving processes differ systematically by gender, age and ICT familiarity?

4.1. Data

Respondents participating in PIAAC were randomly assigned to two of the three cognitive domains assessed in PIAAC. After being randomly assigned to a literacy, numeracy or PSTRE modules, they were further randomly assigned to a second module, with the only restriction that they could not take two literacy or numeracy modules. It was instead possible to be assigned to two (different) PSTRE modules. The routing paths from/to different domains in Module 1 and Module 2 are shown in Figure 6. More details on the PIAAC study design are available in (OECD, 2016_[67]). Each literacy and numeracy module consisted of two stages, with each stage containing a number of testlets of varying difficulty. In each stage, only one testlet was delivered to a respondent. The PSTRE modules were organised in a different way. Each module contained a fixed set of tasks: seven in Module 1 and seven in Module 2, which resulted in 14 PSTRE items in total. No item present in Module 1 was present in Module 2, and vice versa. In this paper, the seven PSTRE items belonging to Module 1 are labelled as item cluster “PS1”, and the seven PSTRE items in Module 2 as item cluster “PS2”. PS1 was always administered as the first module, and PS2 was always administered as the second module.

Figure 6. Routing path between Module 1 and Module 2 in PIAAC Main Study

Note: PS1 and PS2 represent the PSTRE item clusters in Module 1 and Module 2, respectively.

Source: Technical Report of the Survey of Adult Skills (PIAAC) (Second Edition) (OECD, 2016_[67]).

This paper focuses on respondents who took PS2. The analysis is restricted to the following five countries: Ireland, Japan, the Netherlands, England/N. Ireland (United Kingdom) and the United States. Table 1 reports information on the sample size per country by routing path to the PS2, i.e. LIT-PS2, NUM-PS2 and PS1-PS2. The sample consists of 7 749 respondents in total, among whom, 3 938 (51%) were routed from either the literacy or the numeracy domain, and 3 811 (49%) respondents took both PS1 and PS2.

Table 1. Sample size per country by routing path to the PS2

	LIT-PS2	NUM-PS2	PS1-PS2	Total
England/N. Ireland (UK)	602	661	1 200	2 463
Ireland	322	367	665	1 354
Japan	263	260	555	1 078
Netherlands	407	376	727	1 510
United States	344	336	664	1 344
Total	1 938	2 000	3 811	7 749

Note: LIT and NUM represent Literacy and Numeracy, respectively. PS1 and PS2 represent the PSTRE item clusters in Module 1 and Module 2, respectively.

Source: PIAAC log files (www.oecd.org/skills/piaac/log-file) (accessed 1 April 2019).

These five countries display a significant range of PSTRE proficiency levels in PIAAC, facilitating the investigation of potential differences in problem-solving behaviours by countries at different performance levels. In particular, Japan and the Netherlands were among the top performers in PSTRE, with average scores of 294 and 286, respectively (far above the OECD average of 278), while England/N.Ireland (United Kingdom), Ireland and the United States all scored close to the average. These five countries are also located in three different continents, allowing for comparison across different cultures and languages.

4.2. Items in the PS2 module

The analysis uses process data on the seven items that were part of the PS2 cluster. All respondents faced the same items in the same order, and were required to give responses to all of them. There was no time limitation, neither at the level of individual items, nor for the overall assessment. Table 2 summarises information about the nature and characteristics of the seven items examined. The items make use of a variety of digital environments (e.g. email, web and spreadsheet). For instance, U19a uses two environments, email and spreadsheet. The email environment is shared with four other items, U02, U16, U11b and U23, while the spreadsheet environment is shared with item U19b. The column RS presents the number of expert-predefined action sequences for each item. Only item U23 was designed to have a single optimal correct path. The last two columns present the international item parameters for each item. The item parameters were calibrated from the two-parameter-logistic model (Lord, 1980_[24]), in which parameter *a* is defined as a discrimination (slope) parameter to distinguish high- and low-performance groups, while parameter *b* is defined as a difficulty (location) parameter to indicate the degree of difficulty for a certain item.

Table 2. Similarity and consistency of similarity across five sample countries

	Environment				RS	Difficulty Level	Item Parameters	
	Email	Web	Word Processor	Spreadsheet			<i>a</i>	<i>b</i>
U19a	Y			Y	4	1	1.414	-1.367
U19b			Y	Y	4	2	1.072	-0.677
U07		Y			2	2	1.104	-0.237
U02	Y	Y	Y		5	3	1.184	0.784
U16	Y				16	1	1.377	-0.773
U11b	Y				18	3	0.471	0.774
U23	Y	Y			1	2	0.533	-0.052

Source: Technical Report of the Survey of Adult Skills (PIAAC) (Second Edition) (OECD, 2016_[67]).

4.3. Application of the LCS method

The LCS method presented in Section 3 was applied to the process data just described. For each item, each individual's action sequence was calculated against all the experts-predefined reference sequences to derive the longest common subsequence. The measurement indicators by two dimensions – similarity and efficiency (see details in Section 3.3) were calculated per person per item based on the determined LCS and its corresponding reference sequence. The process dataset was linked with the cognitive and background dataset through a unique respondent ID.

Notably, the shortest observable action sequence by an individual was “Start, Next, Next_OK”. In this case, the respondent skipped immediately to the next item, without interacting at all with the task at hand. These cases were treated as missing answers in the current analysis and coded as “Nonresponse (NR)” when calculating the LCS.

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 6 007 respondents in total being used for the reporting of the LCS results.

5. Results

5.1. Comparison Design

Applying the LCS method resulted in a distribution of each measurement indicator across the items. Figure 7 displays the distribution of average similarity (SM) in panel (A) and the distribution of consistency of similarity (SSD) in panel (B), using data on the 6 007 respondents who did not have any nonresponse sequence patterns in the process data. The mean and standard deviation for the similarity distribution were 0.65 and 0.15, respectively. Thresholds of one SD below and above the mean were used to divide the similarity index in three levels. Those with a similarity degree ranging from 0 to 0.5 were placed in the low similarity group. Those with a similarity degree within the range of 0.5 to 0.8 were placed in the moderate similarity group, and the remaining individuals whose similarity degree ranged from 0.8 to 1.0 were placed in the high similarity group.

A similar method was used to determine the thresholds for consistency of similarity. The mean and standard deviation for the consistency of similarity distribution was 0.20 and 0.05, respectively. Those with a consistency degree ranging from 0 to 0.15 were placed in the high-consistency group, those with a similarity degree within the range of 0.15 to 0.25 were placed in the moderate-consistency group, and the remaining individuals whose consistency ranged from 0.25 to 1.0 were placed in the low-consistency group.

Figure 7. Distribution of similarity and consistency of similarity across items

Note: The distribution of similarity and consistency of similarity were used to define thresholds for categorising people into three groups for comparison purpose. The mean and standard deviation for similarity (A) is 0.65 and 0.15, respectively. The mean and standard deviation for consistency of similarity (B) is 0.20 and 0.05, respectively. The thresholds to determine the categories were set as $\text{Mean} \pm 1\text{SD}$.

Source: PIAAC log files (www.oecd.org/skills/piaac/log-file) (accessed 1 April 2019).

Similarity and consistency of similarity were then mapped in a matrix, classifying individuals in nine subgroups (Figure 8). Note that the first digit in the group name indicates the consistency of similarity degree (1 is consistent, 3 is inconsistent). The rows (whose labels start with SD) in Figure 8 reflect this. The second digit in the group name indicates the degree of (average) similarity (1 is low, 3 is high). The columns (whose labels start with M) reflect this. For instance, G13 meant people in this group consistently followed patterns that were highly similar to the reference sequences.

Figure 8. Mapping similarity and consistency of similarity into one matrix

		Similarity (MEAN)			
		M1	M2	M3	
Consistency (SD)	SD1	G11 Extreme Consistent Low Similarity	G12 Extreme Consistent Moderate Similarity	G13 Extreme Consistent High Similarity	Extreme Consistent
	SD2	G21 Moderate Consistent Low Similarity	G22 Moderate Consistent Moderate Similarity	G23 Moderate Consistent High Similarity	Moderate Consistent
	SD3	G31 Extreme Inconsistent Low Similarity	G32 Extreme Inconsistent Moderate Similarity	G33 Extreme Inconsistent High Similarity	Extreme Inconsistent
		Low Similarity	Moderate Similarity	High Similarity	

Note: M indicates Mean. SD indicates standard deviation. Similarity degree is shown by columns, while the consistency of similarity degree is shown by rows.

5.2. Problem-solving strategies across different items

The first research question asked whether people adopt consistent problem-solving strategies across different items. In order to answer that, it is possible to explore the association between the similarity and the consistency of similarity. The frequency of each cell corresponding to the groups of similarity and consistency of similarity is reported in Figure 9. Figure 10 displays the association between the similarity index (horizontal axis) and the consistency of similarity (vertical axis) in a scatterplot. The majority of respondents were located in the high-consistency and moderate-consistency groups, while only a small proportion were located in the low-consistency group. This tendency appeared more obvious in the low and high similarity groups. In the extreme inconsistent group, G31 and G33 had the lowest proportion of respondents, consisting of only 3% and 0.03%, respectively, of the whole sample. Within the high similarity group, 98% respondents were located in the extreme or moderate consistent group, while only 2% were located in the low-consistency group (the small group in pink at the top right of the figure). Over 83% of respondents were located in the moderate and high similarity groups, meaning that the majority of respondents followed the reference sequences. Notably, nested within the high-consistency group, the low, moderate and high similarity subgroups were represented in equal proportions, close to one-third by each subgroup. However, this ratio was substantially different for the extreme inconsistent group, where the representation of low, moderate, and high similarity groups were 18%, 80% and 2%, respectively. It suggested that starting from the moderate similarity group, respondents could have unstable similarity degree depending on different items.

Figure 9. Frequency of groups by similarity and consistency of similarity

		Similarity			
Consistency	G11	G12	G13		
	280 (5%)	391 (6%)	390 (6%)		1 061 (17%)
	G21	G22	G23		
	540 (9%)	2557 (43%)	677 (11%)		3 774 (63%)
	G31	G32	G33		
	203 (3%)	947 (16%)	22 (0%)		1 172 (19%)
				1 023 (17%)	3 895 (65%)
					1 089 (17%)

Note: Group name per cell corresponds to the cells in Figure 8.

Figure 10. A scatterplot of interaction between similarity and consistency of similarity

Note: The group names correspond to the cell names in Figure 8.

Source: PIAAC log files (www.oecd.org/skills/piaac/log-file) (accessed 1 April 2019).

The analysis also investigated whether patterns for similarity and consistency of similarity were identical across countries. Figure 11 exhibits the results across five countries. The general patterns in similarity were found to vary across countries. Japan and the Netherlands, the two countries with the highest level of performance, were characterised by a higher degree of similarity than the other countries. It suggested that respondents from these two countries on average followed the predefined action sequence more often than respondents in countries with lower levels of performance. By contrast, no country differences were observed in consistency of similarity.

Figure 11. Similarity and consistency of similarity across five sample countries

Note: (A) and (B) present boxplots for similarity and consistency of similarity across five countries. GBR, IRL, JPN, NLD, USA represent the England and Northern Ireland, Ireland, Japan, the Netherlands and United States, respectively.

Source: PIAAC log files (www.oecd.org/skills/piaac/log-file) (accessed 1 April 2019).

Statistical results further supported these findings. In the one-way analysis of variance (ANOVA), significant difference in similarity was found among the five countries, $F(4, 6002) = 32.705, P < 0.001$; the Bonferroni post hoc correction showed Japan and the Netherlands were significantly higher than the other three countries. However, there was no significant difference between Japan and the Netherlands. In the ANOVA test for the consistency of similarity across countries, $F(4, 6002) = 3.48, P = 0.07$; the results showed a not significant result, meaning the five countries performed equally consistently. Detailed descriptive statistics of similarity and consistency of similarity for each of the five countries examined are reported in Table 3.

Table 3. Descriptive statistics of similarity and consistency of similarity

		Sample Size	Similarity		Consistency of Similarity	
			Mean	Variance	Mean	Variance
England/N. Ireland (UK)	Ireland	1 885	0.645	0.022	0.204	0.003
Ireland		856	0.635	0.026	0.203	0.003
Japan		1 174	0.692	0.018	0.197	0.003
Netherlands		1 157	0.672	0.022	0.200	0.003
United States		935	0.633	0.024	0.199	0.003

Source: PIAAC log files (www.oecd.org/skills/piaac/log-file) (accessed 1 April 2019).

Notably, in the high-consistency levels, all the countries showed identical patterns, with the majority of respondents located in the moderate similarity group, a smaller number of respondents in the low similarity group, and an even smaller proportion in the high similarity group. However, large variances were observed in the low-consistency level. High-performing countries, especially Japan and the Netherlands, had the majority of respondents in the high similarity group; only 2-4% respondents of the whole country sample were located in the low similarity group. On the contrary, in countries with lower average performance (Ireland and the United States), the largest proportion belonged to the low similarity group and the smallest proportion to the high similarity group.

A similar investigation was also conducted on the efficiency and consistency of efficiency across the five countries (see Figure 12). As mentioned in section 3.3.2, efficiency is an indicator that captures whether respondents solved the problems in a concise way, without performing actions that are redundant (as they do not appear in the reference sequences). On this metric, Japan showed a significantly lower efficiency, meaning that respondents usually followed longer sequences.

Figure 12. Efficiency and consistency of efficiency across five sample countries

Note: (A) and (B) present boxplots for efficiency and consistency of efficiency across five countries. GBR, IRL, JPN, NLD, USA represent the England and Northern Ireland (UK), Ireland, Japan, the Netherlands and United States, respectively.

Source: PIAAC log files (www.oecd.org/skills/piaac/log-file) (accessed 1 April 2019).

To explore the interaction between similarity and efficiency, we plotted the distributions of these two measurement scales in the nine-cell matrix corresponding to the comparison design in Figure 13. A trade-off between efficiency and similarity is visible in this table

matrix. Interestingly, the efficiency and similarity distribution had large overlaps in the low similarity groups regardless of the consistency degree. These overlaps were dramatically reduced in the moderate similarity groups. Almost no overlaps can be observed in the high similarity group. The results imply that those in the high similarity group usually made quite a few more steps than necessary. Few respondents were able to achieve high efficiency and high similarity.

Figure 13. A plot matrix for interactions between efficiency and similarity

Note: G11 = Extreme Consistent/Low Similarity; G12 = Extreme Consistent/Moderate Similarity; G13 = Extreme Consistent/High Similarity; G21 = Moderate Consistent/Low Similarity; G22 = Moderate Consistent/Moderate Similarity; G23 = Moderate Consistent/High Similarity; G31 = Extreme Inconsistent/Low Similarity; G32 = Extreme Inconsistent/Moderate Similarity; G33 = Extreme Inconsistent/High Similarity.

Source: PIAAC log files (www.oecd.org/skills/piaac/log-file) (accessed 1 April 2019).

5.3. Association between the patterns of strategies and problem-solving proficiency

The second research question investigated whether the general patterns and consistency of these general patterns were related to respondents' proficiency levels. The analysis focuses on the relationship between PSTRE proficiency and the similarity degree as well as the consistency of similarity. As Figure 14 shows, the PSTRE proficiency level apparently differs by the three similarity groups, which were differentiated by colour - red, green, and blue for low, moderate, and high similarity groups, respectively. A significant association was detected between proficiency and similarity.

Among respondents with moderate and high similarity level, the highly consistent (G12) and the moderately-consistent (G22) displayed similar proficiency levels; a t-test showed there was no significant differences between these two groups. However, proficiency was lower for the low-consistency group (G32) with moderate and high similarity. In the low similarity group, opposite results were observed. The high-consistency group showed the lowest proficiency level, while the low-consistency group showed a significantly higher proficiency level than the other two. This implies that the general patterns in similarity did make an impact on proficiency level. The consistency of similarity was also significantly correlated to the proficiency level, but the relationship varied by similarity degree. In the high similarity group, higher consistency was associated with higher performance, while in the lower similarity group the opposite was true.

Table 4 reports the descriptive statistics of proficiency levels by similarity and consistency of similarity.

Figure 14. Relationship between PSTRE proficiency level and general pattern in similarity

Note: Red, green, and blue clusters represent low, moderate, and high similarity groups, respectively. G11 = Extreme Consistent/Low Similarity; G12 = Extreme Consistent/Moderate Similarity; G13 = Extreme Consistent/High Similarity; G21 = Moderate Consistent/Low Similarity; G22 = Moderate Consistent/Moderate Similarity; G23 = Moderate Consistent/High Similarity; G31 = Extreme Inconsistent/Low Similarity; G32 = Extreme Inconsistent/Moderate Similarity; G33 = Extreme Inconsistent/High Similarity.

Source: PIAAC log files (www.oecd.org/skills/piaac/log-file) (accessed 1 April 2019).

Table 4. Descriptive statistics of proficiency level by similarity and consistency of similarity

	Sample Size	Mean	Variance	Similarity Group	Sample Size	Mean	Variance
G11	280	240.75	702.08	M1	1 023	252.99	737.43
G21	540	255.22	558.09				
G31	203	263.93	931.02				
G12	391	298.74	863.24	M2	3 895	298.74	739.35
G22	2 557	301.69	737.76				
G32	947	290.77	607.02				
G13	390	330.48	444.30	M3	1 089	329.16	504.97
G23	677	328.79	531.97				
G33	22	317.10	618.99				

Note: G11 = Extreme Consistent/Low Similarity; G12 = Extreme Consistent/Moderate Similarity; G13 = Extreme Consistent/High Similarity; G21 = Moderate Consistent/Low Similarity; G22 = Moderate Consistent/Moderate Similarity; G23 = Moderate Consistent/High Similarity; G31 = Extreme Inconsistent/Low Similarity; G32 = Extreme Inconsistent/Moderate Similarity; G33 = Extreme Inconsistent/High Similarity

Source: PIAAC log files (www.oecd.org/skills/piaac/log-file) (accessed 1 April 2019).

5.4. Differences by socio-economic background in problem-solving strategies

This section further investigates the relationships between the general patterns in similarity and the socio-demographic characteristics of respondents. In particular, the purpose is to identify differences in problem-solving strategies across different socio-demographic groups, defined by gender, age, income and familiarity with ICT.

5.4.1. Gender

Figure 15 illustrates the share of men and women in the different similarity groups across the seven items in PS2. Women are over-represented in the low and moderate similarity groups, while men are more likely to be present in the high similarity group. This finding is confirmed by Figure 16, where average similarity of male and female respondents is depicted for each item. On average, males showed higher similarity than females in each item.

On the other hand, women showed a significantly higher degree of consistency, $T(6005) = 1.65, P = 0.003$, implying that they generally followed a more stable pattern of problem-solving strategies than males, no matter whether their problem-solving strategies were farther or closer away from the predefined reference sequences.

Figure 15. Share of men and women in different similarity groups

Note: M1, M2, and M3 represent the low, moderate and high similarity groups, respectively.

Source: PIAAC log files (www.oecd.org/skills/piaac/log-file) (accessed 1 April 2019).

Figure 16. Gender differences in similarity across seven items in PS2

Source: PIAAC log files (www.oecd.org/skills/piaac/log-file) (accessed 1 April 2019).

5.4.2. Age

Significant age differences were found across similarity groups in an ANOVA analysis, $F(2, 6004) = 87.92, P < 0.001$. Figure 17 presents the age distribution of the nine groups identified in Figure 8, nested by degree of similarity and consistency of similarity. Respondents from the low similarity groups were the oldest on average. They were 4.6 and 7.7 years older on average than respondents in the moderate and high similarity groups, respectively, suggesting that younger individuals were more likely to choose expert-predefined strategies or strategies close to such sequences. No significant differences were found between consistent groups nested in low and moderate similarity groups. The high similarity group was an exception in this respect. The extreme inconsistent group (with an average 39 years old) was found to be significantly older than the moderate and extreme consistent groups (average 35 years old). However, because of the small sample size (22 respondents) in G33, this result may not be sufficiently reliable to make a general conclusion.

Figure 17. Distribution of age, by similarity and consistency of similarity

Note: Red, green and blue clusters represent low, moderate and high similarity groups respectively.

G11 = Extreme Consistent/Low Similarity; G12 = Extreme Consistent/Moderate Similarity; G13 = Extreme Consistent/High Similarity; G21 = Moderate Consistent/Low Similarity; G22 = Moderate Consistent/Moderate Similarity; G23 = Moderate Consistent/High Similarity; G31 = Extreme Inconsistent/Low Similarity; G32 = Extreme Inconsistent/Moderate Similarity; G33 = Extreme Inconsistent/High Similarity

Source: PIAAC log files (www.oecd.org/skills/piaac/log-file) (accessed 1 April 2019).

5.4.3. Differences related to familiarity with ICT

Previous research indicates that adults who use ICT devices at work or at home often perform better in the PSTRE assessment than those who rarely use ICT devices (Liao, He and Jiao, 2019^[15]; He et al., 2018^[51]). Table 5 presents descriptive statistics on the extent to which being involved in tasks that require the use of ICT skills at work and at home is associated with group membership. Levels of ICT use at work are significantly different across adults belonging to different similarity groups in ANOVA $F(2, 4021) = 49.32, P <$

0.001.⁴ Differences remain 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 predefined sequences, but not of higher consistency.

Table 5. Descriptive statistics of ICT skills

	ICT Skills at Work		ICT Skills at Home	
	Mean	Variance	Mean	Variance
Low Similarity	1.88	0.80	1.87	0.72
Moderate Similarity	2.09	0.86	2.21	0.70
High Similarity	2.36	0.88	2.38	0.67

Note: The derived weighted likelihood estimate variables ICT skills at work and ICT skills at home from the background questionnaire were used in this study. The missing data from the background questionnaire were not excluded from the statistical analyses.

Source: PIAAC log files (www.oecd.org/skills/piaac/log-file) (accessed 1 April 2019).

Similar results were found in the association with the use of ICT at home. A significant difference was found across the three similarity groups in the ANOVA test, $F(2,5602) = 96.80, P < 0.001$.⁵ However, differences turned out to be 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 only found for the group displaying low similarity.

⁴ The missing data in the background variables were excluded in the ANOVA analysis.

⁵ The missing data in the background variables were excluded in the ANOVA analysis.

6. Discussion and conclusion

6.1. Summary

PIAAC is considered by many as the most innovative large-scale assessment of adult skills under many respects (Kirsch and Lennon, 2017_[68]). In particular, the choice of computers as the primary delivery platform allowed for the collection of data not just on whether respondents were able to solve a specific assessment task but also on how they approached the solution and how much time they spent on it. The increasing availability of process data provides researchers with a unique opportunity to gain a deeper understanding of how students and adults engage with and respond to a variety of traditional and innovative items (He and von Davier, 2016_[5]; Kirsch et al., 2013_[69]).

This paper exploits the LCS method, commonly used in natural language processing and biostatistics, to derive indicators that can reliably capture and characterise the strategies followed by respondents across a variety of problem-solving assessment items. In the empirical application of the method, consistent behavioural patterns were found for a large number of respondents across countries. More consistent patterns were observed in groups performing either extremely well or extremely poorly on the test. It was also found that the degree of similarity, and the consistency of similarity across items, was a significant predictor of overall performance in the PSTRE assessment. Countries with high average PSTRE performance showed higher degrees of similarity, meaning that more adults in these countries chose strategies similar to the predefined reference ones.

The LCS approach, introduced in Sukkarieh et al. (2012_[10]), was employed as an automated scoring tool for multiple languages in open-ended questions in international large assessments such as PISA. This method applied in the current study can also be used as an automatic tool to extract respondents' strategy in problem-solving efforts. For instance, given the predefined reference sequences, the input of each individual action sequences could be automatically categorised and the system could output the maximum likely strategy for each respondent. It holds promise in getting a rapid strategy extraction from respondents' process data, which can greatly enhance the analytical potential of the data. In addition, the LCS method was also shown to be an efficient way to use process data to check item quality and provide valuable information in further item development. For instance, the distance between the individual action sequences and predefined sequences could be used as an additional factor in detecting items with different item functioning (DIF; (Lord, 1980_[24])) that is, examines the measurement equivalence across groups (e.g. country, gender and so on), which is essentially important in international large-scale assessments.

The sequences of actions followed to solve problem-solving items were found to be significantly different by gender, age and ICT familiarity. Males were more likely to use strategies similar to the predefined ones than females, while females showed more consistent patterns across items. Younger respondents were more likely to follow the predefined sequences when solving digital tasks than older respondents. Respondents engaging more frequently with tasks that require ICT skills were also more likely to follow optimal strategies.

These findings are especially helpful for policy makers to get a better picture about the general patterns by different subgroups. In particular, they allow for to identify subgroups

with special need for improvement and plan appropriate training plans for those subgroups. In addition, the comparable scales based on LCS results made it possible to compare sequences and problem-solving strategies across countries, thus facilitating researchers' and policy makers' efforts to understand the deep reasons behind the differences in information processing skills that normally emerge from large-scale assessments.

6.2. Limitations

Besides these positive results, some limitations also merit discussion. First, as mentioned in section 4.2, the sample used in the current study included respondents who were routed to PS2 only. To maintain the maximum sample size for each item, we did not separate the respondents into groups by the routing paths (i.e. routing from literacy or numeracy, or routing from PS1). Therefore, the potential pre-knowledge effect was not taken into consideration. The respondents who took both PS1 and PS2 could have already accumulated knowledge on item structures and instructions and become familiar with the PSTRE environments in the PS1 in the first module. One might expect this group to use shorter action sequences and have a shorter processing time to solve the items in PS2. It would be interesting to analyse this pre-knowledge effect in the further study to compare whether groups by different routing paths have the identical patterns in problem-solving and how much the pre-knowledge could make an impact on respondents' general patterns and proficiency level.

Second, time stamps of actions are also available in process data. The time elapsed between the occurrences of two actions may provide additional information about respondents and can be useful in cognitive assessments. The current study does not make use of this information. This could affect, for instance, the interpretation of the "efficiency" indicator: respondents might have taken shorter action sequences, but might at the same time have spent more time on the item. Incorporating response time information in the analysis of process data is a promising avenue for future research.

Third, we excluded the nonresponse sequence pattern "Start, Next, Next_OK" from the LCS analysis, which resulted in a loss of over 2 000 individuals from the sample. It would be valuable to explore more these data, focusing for instance on engagement and item difficulty that could be highly correlated with skipping behaviour. For the nonresponse pattern, the time between each action, especially the reaction time for the first action (i.e. time between "Start" and "Next"), was not yet taken into account in this study. That meant we could not distinguish whether the respondent quickly skipped the item because of low engagement or made an effort but eventually skipped it because it seemed too hard. We recommend including the time interval between actions in future studies to better distinguish nonresponse patterns by process data.

Fourth, the LCS method identifies a single strategy that the respondent was most likely to have followed. However, in practice, respondents are often using a combination of strategies: they might switch from one strategy to the other, or they might use multiple strategies to validate their solutions. LCS model could be refined in order to accommodate mixed membership of strategies in future studies.

In addition, the LCS method used in this study is just one of the possible choices. Other choices such as edit distance (Hao, Shu and Davier, 2015^[66]) and optimal symbol alignment distance (Herranz, Nin and Sole, 2011^[70]) can also be used and give similar results. However, these measures are often more computationally demanding, and are less well suited for visualising the strategy that has been chosen by respondents.

6.3. Future studies

Future studies will concentrate on how to integrate the information derived from process data with the response model to get a more accurate estimate on respondents' latent traits. For instance, the strategy and sequence that one individual used in a specific item could be added as auxiliary information to the item response model to better distinguish correct and incorrect groups. A parallel study based on PS1 could also be conducted to compare whether the consistency of problem-solving strategies would be maintained in another context. For instance, the task related to meeting room arrangement was parallel, designed in two clusters. It would be interesting to examine the consistency of using similar strategies for the same content by looking at respondents who took both items. In addition, it would also be interesting to identify the general patterns for each respondent, and thus cluster them by common properties. For instance, one could explore whether there were increasing or decreasing patterns in similarity degree by respondents, that is, whether the respondent started from low similarity and ended with high similarity through a continuous learning process, or the other way round, possibly because of fatigue or decreasing engagement. There is indeed ample evidence that performance on the assessment was typically lower in the second module (OECD, 2019^[6]).

Future work may also focus on the challenging task of identifying patterns that have broad theoretical and face value meaning, and that allow to test a set of hypotheses and research questions. While most research in assessment was quantitative in nature (trying to make sense of differences in achievement levels), the analysis of process data (in particular data referring to sequences and strategies) is inherently qualitative, although highly formalised.

6.4. Conclusion

The increasing use of computer-based assessments will offer more opportunities to analyse process data. This will be beneficial for researchers, assessment specialists and educators, and will allow to better understand test takers' behaviours. This paper illustrated how process data can be used to identify adults' problem-solving behaviours and introduced a sequence-mining technique that can be applied to identify general patterns of behaviour across multiple assessment tasks. 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.

The research potential of process data will be maximised if their analysis is already taken into account at the item development stage. Effort should be made to design items that allow for the clear identification of different solution strategies; ideally, these different strategies should be mapped to cognitive theories that researchers might be interested to test with the aid of process data (OECD, 2019^[6]).

References

- Agrawal, R. and R. Srikant (1995), *Mining sequential patterns*, IEEE Comput. Soc. Press, [38]
<http://dx.doi.org/10.1109/ICDE.1995.380415>.
- Aleven, V., B. McLaren and J. Sewall (2009), "Scaling Up Programming by Demonstration for Intelligent Tutoring Systems Development: An Open-Access Web Site for Middle School Mathematics Learning", *IEEE Transactions on Learning Technologies*, Vol. 2/2, pp. 64-78, [72]
<http://dx.doi.org/10.1109/TLT.2009.22>.
- Andersen, E., S. Gulwani and Z. Popovic (2013), *A trace-based framework for analyzing and synthesizing educational progressions*, ACM Press, New York, NY, USA, [41]
<http://dx.doi.org/10.1145/2470654.2470764>.
- Azevedo, R. (2005), "Using Hypermedia as a Metacognitive Tool for Enhancing Student Learning? The Role of Self-Regulated Learning", *Educational Psychologist*, Vol. 40/4, pp. 199-209, [27]
http://dx.doi.org/10.1207/s15326985ep4004_2.
- Baker, R. and K. Yacef (2009), "The State of Educational Data Mining in 2009: A Review and Future Visions.", *Journal of Educational Data Mining*, Vol. 1/1, pp. 3-16, <https://eric.ed.gov/?id=EJ1115366> [39]
 (accessed on 22 May 2019).
- Bennett, R. et al. (2010), "Measuring Problem Solving with Technology: A Demonstration Study for NAEP", *The Journal of Technology, Learning, and Assessment*, Vol. 8/8, <http://www.jtla.org> (accessed on 22 May 2019). [2]
- Bennett, R. et al. (2007), *Problem Solving in Technology-Rich Environments A Report From the NAEP Technology-Based Assessment Project (NCES 2007-466)*, US Department of Education National Center for Education Statistics, Washington, D.C., [1]
https://nces.ed.gov/nationsreportcard/pdf/studies/2007466_1.pdf (accessed on 22 May 2019).
- Biswas, G. et al. (2010), "Measuring Self-Regulated Learning Skills Through Social Interactions in a Teachable Agent Environment", *Research and Practice in Technology Enhanced Learning*, Vol. 05/02, pp. 123-152, <http://dx.doi.org/10.1142/S1793206810000839>. [44]
- Bouchet, F. et al. (2013), "Clustering and Profiling Students According to Their Interactions with an Intelligent Tutoring System Fostering Self-Regulated Learning", *Journal of Educational Data Mining*, Vol. 5/1, pp. 104-146, <https://jedm.educationaldatamining.org/index.php/JEDM/article/view/32> [33]
 (accessed on 22 May 2019).
- Brand-Gruwel, S., I. Wopereis and A. Walraven (2009), "A descriptive model of information problem solving while using internet", *Computers & Education*, Vol. 53/4, pp. 1207-1217, [29]
<http://dx.doi.org/10.1016/j.compedu.2009.06.004>.
- Breiman, L. (2001), "Random Forests", *Machine Learning*, Vol. 45/1, pp. 5-32, [60]
<http://dx.doi.org/10.1023/A:1010933404324>.
- Breiman, L. et al. (1984), *Classification and regression trees*, Chapman and Hall/CRC. [59]

- Chen, Y. et al. (2019), “Statistical Analysis of Complex Problem-Solving Process Data: An Event History Analysis Approach”, *Frontiers in Psychology*, Vol. 10, p. 486, <http://dx.doi.org/10.3389/fpsyg.2019.00486>. [20]
- Cormen, T. et al. (2001), *Introduction to Algorithms*, MIT Press, Cambridge, MA, <https://cdn.manesht.ir/19908/Introduction%20to%20Algorithms.pdf> (accessed on 22 May 2019). [64]
- Demars, C. (2007), “Changes in Rapid-Guessing Behavior Over a Series of Assessments”, *Educational Assessment*, Vol. 12/1, pp. 23-45, <http://dx.doi.org/10.1080/10627190709336946>. [49]
- Dietterich, T. (2000), *Ensemble Methods in Machine Learning*, <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.228.2236> (accessed on 22 May 2019). [61]
- Fink, G. (2008), *Markov models for pattern recognition : from theory to applications*, Springer, Berlin, Germany. [46]
- Forman, G. (2003), “An Extensive Empirical Study of Feature Selection Metrics for Text Classification”, *Journal of Machine Learning Research*, Vol. 3, pp. 1289-1305, http://www.jmlr.org/papers/volume3/forman03a/forman03a_full.pdf (accessed on 22 May 2019). [26]
- Goldhammer, F., J. Naumann and Y. Keßel (2013), “Assessing Individual differences in Basic Computer Skills: Psychometric characteristics of an interactive performance measure”, *European Journal of Psychological Assessment*, Vol. 29, pp. 263-275, https://www.uni-frankfurt.de/67186348/Goldhammer_et_al_EJPA_2013.pdf (accessed on 22 May 2019). [12]
- Goldhammer, F. et al. (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*, <http://dx.doi.org/10.1037/a0034716>. [37]
- Graesser, A. et al. (2004), “AutoTutor: A tutor with dialogue in natural language”, *Behavior Research Methods, Instruments, & Computers*, Vol. 36/2, pp. 180-192, <http://www.autotutor.org> (accessed on 22 May 2019). [71]
- Greiff, S., S. Wüstenberg and F. Avvisati (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*, Vol. 91/C, pp. 92-105, <http://dx.doi.org/10.1016/j.compedu.2015.10.018>. [9]
- Guo, H. et al. (2018), “Modeling Basic Writing Processes From Keystroke Logs”, *Journal of Educational Measurement Summer*, Vol. 55/2, pp. 194-216, [http://dx.doi.org/10.1111/\(ISSN\)1745-3984](http://dx.doi.org/10.1111/(ISSN)1745-3984). [17]
- Guyon, I. and A. Elisseeff (2003), “An Introduction to Variable and Feature Selection”, *Journal of Machine Learning Research*, Vol. 3/Mar, pp. 1157-1182, <http://www.jmlr.org/papers/v3/guyon03a.html> (accessed on 22 May 2019). [52]
- Han, Z., Q. He and M. von Davier (2019), *Predictive Feature Generation and Selection from Process Data in PISA Simulation-Based Environment: An Implementation of Tree-based Ensemble Methods*, mimeo. [62]
- Hao, J., Z. Shu and A. Davier (2015), “Analyzing Process Data from Game/Scenario-Based Tasks. An Edit Distance Approach”, *Journal of Educational Data Mining*, Vol. 7/1, pp. 33-50, <https://jedm.educationaldatamining.org/index.php/JEDM/article/view/JEDM072> (accessed on 22 May 2019). [66]

- He, Q. et al. (2014), "Predicting self-monitoring skills using textual posts on Facebook", *Computers in Human Behavior*, Vol. 33, pp. 69-78, <http://dx.doi.org/10.1016/j.chb.2013.12.026>. [57]
- He, Q. et al. (2018), *Exploring relationship between sequence patterns in solving digital tasks and background variables: an empirical study using log data in PIAAC*. [51]
- He, Q., B. Veldkamp and T. de Vries (2012), "Screening for posttraumatic stress disorder using verbal features in self narratives: A text mining approach", *Psychiatry Research*, Vol. 198/3, pp. 441-447, <http://dx.doi.org/10.1016/j.psychres.2012.01.032>. [56]
- He, Q. and M. von Davier (2016), "Analyzing Process Data from Problem-Solving Items with N-Grams. Insights from a computer-based large-scale assessment", in Rosen Y., Ferrara S., M. (ed.), *Handbook of Research on Technology Tools for Real-World Skill Development*, Information Science Reference, <http://dx.doi.org/10.4018/978-1-4666-9441-5.ch029>. [5]
- He, Q. and M. von Davier (2015), "Identifying Feature Sequences from Process Data in Problem-Solving Items with N-Grams", in van der Ark L., Bolt D., Wang WC., Douglas J., C. (ed.), *Quantitative Psychology Research. Springer Proceedings in Mathematics & Statistics*, vol. 140, Springer, Cham, http://dx.doi.org/10.1007/978-3-319-19977-1_13. [8]
- He, Q., M. von Davier and Z. Han (2018), "Exploring Process Data in Computer-based International Large-scale Assessments", in Jiao, H., R. Lissitz and A. Van Wie (eds.), *Data analytics and psychometrics : informing assessment practices*, Information Age Publishing, Charlotte, NC. [16]
- Herranz, J., J. Nin and M. Sole (2011), "Optimal Symbol Alignment Distance: A New Distance for Sequences of Symbols", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 23/10, pp. 1541-1554, <http://dx.doi.org/10.1109/TKDE.2010.190>. [70]
- Kinnebrew, J., D. Mack and G. Biswas (2013), *Mining Temporally-Interesting Learning Behavior Patterns*, <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.670.697&rep=rep1&type=pdf> (accessed on 22 May 2019). [43]
- Kirsch, I. and M. Lennon (2017), "PIAAC: a new design for a new era", *Large-scale Assessments in Education*, Vol. 5/1, p. 11, <http://dx.doi.org/10.1186/s40536-017-0046-6>. [68]
- Kirsch, I. et al. (2013), "On the Growing Importance of International Large-Scale Assessments", in *The Role of International Large-Scale Assessments: Perspectives from Technology, Economy, and Educational Research*, Springer Netherlands, Dordrecht, http://dx.doi.org/10.1007/978-94-007-4629-9_1. [69]
- Klieme, E. (2004), "Assessment of cross-curricular problem-solving competencies", in Moskowitz, J. and M. Stephens (eds.), *Comparing learning outcomes : international assessments and education policy*, Routledge, London. [34]
- Kohavi, R. and G. John (1997), "Wrappers for feature subset selection", *Artificial Intelligence*, Vol. 97/1-2, pp. 273-324, [http://dx.doi.org/10.1016/S0004-3702\(97\)00043-X](http://dx.doi.org/10.1016/S0004-3702(97)00043-X). [53]
- LaMar, M. (2018), "Markov Decision Process Measurement Model", *Psychometrika*, Vol. 83/1, pp. 67-88, <http://dx.doi.org/10.1007/s11336-017-9570-0>. [63]
- Lazonder, A. and J. Rouet (2008), "Information problem solving instruction: Some cognitive and metacognitive issues", *Computers in Human Behavior*, Vol. 24/3, pp. 753-765, <http://dx.doi.org/10.1016/j.chb.2007.01.025>. [28]

- Lee, Y. and Y. Jia (2014), "Using response time to investigate students' test-taking behaviors in a NAEP computer-based study", *Large-scale Assessments in Education*, Vol. 2/1, p. 8, <http://dx.doi.org/10.1186/s40536-014-0008-1>. [14]
- Liao, D., Q. He and H. Jiao (2019), "Mapping Background Variables With Sequential Patterns in Problem-Solving Environments: An Investigation of United States Adults' Employment Status in PIAAC", *Frontiers in Psychology*, Vol. 10/646, <http://dx.doi.org/10.3389/fpsyg.2019.00646>. [15]
- Liu, H., Y. Liu and M. Li (2018), "Analysis of Process Data of PISA 2012 Computer-Based Problem Solving: Application of the Modified Multilevel Mixture IRT Model", *Frontiers in Psychology*, Vol. 9, p. 1372, <http://dx.doi.org/10.3389/fpsyg.2018.01372>. [19]
- Lord, F. (1980), *Applications of item response theory to practical testing problems*, L. Erlbaum Associates, Broadway, Hillsdale, NJ, <https://eric.ed.gov/?id=ED312280> (accessed on 22 May 2019). [24]
- Manning, C. and H. Schütze (1999), *Foundations of statistical natural language processing*, MIT Press, Cambridge, MA, <https://mitpress.mit.edu/books/foundations-statistical-natural-language-processing> (accessed on 22 May 2019). [25]
- Martinez, R. et al. (2011), *Analysing frequent sequential patterns of collaborative learning activity around an interactive tabletop*, <http://www.prometheanworld.com/> (accessed on 22 May 2019). [40]
- Mayer, R. (1994), "Problem Solving Teaching and Testing", in Husén, T. and T. Postlethwaite (eds.), *The International encyclopedia of education*, Pergamon, Oxford, England] ;New York, <https://www.worldcat.org/title/international-encyclopedia-of-education/oclc/29797711> (accessed on 22 May 2019). [35]
- Mislevy, R. et al. (2012), "Design and discovery in educational assessment: Evidence-centered design, psychometrics, and educational data mining", *Journal of Educational Data Mining*, Vol. 4/1, pp. 11-48, <https://jedm.educationaldatamining.org/index.php/JEDM/article/view/22> (accessed on 22 May 2019). [36]
- Nigam, K. et al. (2000), "Text Classification from Labeled and Unlabeled Documents using EM", *Machine Learning*, Vol. 39/2-3, p. 1'34, <http://www.kamalnigam.com/papers/emcat-mlj99.pdf> (accessed on 22 May 2019). [58]
- Oakes, M. et al. (2001), *A Method Based on the Chi-Square Test for Document Classification.*, ACM, New York, NY, https://www.researchgate.net/publication/221300440_A_Method_Based_on_the_Chi-Square_Test_for_Document_Classification (accessed on 23 May 2019). [55]
- OECD (2019), *Beyond Proficiency: Using Log Files to Understand Respondent Behaviour in the Survey of Adult Skills*, OECD Skills Studies, OECD Publishing, Paris, <https://dx.doi.org/10.1787/0b1414ed-en>. [6]
- OECD (2017), *PISA 2015 Technical Report*, OECD Publishing, Paris, <https://www.oecd.org/pisa/sitedocument/PISA-2015-technical-report-final.pdf> (accessed on 22 May 2019). [13]
- OECD (2016), *Technical Report of the Survey of Adult Skills (PIAAC) (2nd Edition)*, OECD Publishing, Paris, https://www.oecd.org/skills/piaac/PIAAC_Technical_Report_2nd_Edition_Full_Report.pdf (accessed on 22 May 2019). [67]
- OECD (2012), *Literacy, numeracy and problem solving in technology-rich environments : framework for the OECD survey of adult skills.*, OECD Publishing, Paris. [3]

- OECD (2009), *PIAAC Problem Solving in Technology-Rich Environments: A Conceptual Framework*, OECD Publishing, <https://dx.doi.org/10.1787/220262483674>. [11]
- Poon, L. et al. (2017), *Mining Sequential Patterns of Students' Access on Learning Management System*, Springer, http://dx.doi.org/10.1007/978-3-319-61845-6_20. [47]
- Rabiner, L. (1989), "A tutorial on hidden Markov models and selected applications in speech recognition", *Proceedings of the IEEE*, Vol. 77/2, pp. 257-286, <http://dx.doi.org/10.1109/5.18626>. [45]
- Rasch, G. (1960), *Probabilistic models for some intelligence and achievement tests*, Danish Institute for Educational Research, Copenhagen. [23]
- Schleicher, A. (2008), "Piaac: A New Strategy for Assessing Adult Competencies", *International Review of Education*, Vol. 54/5-6, pp. 627-650, <http://dx.doi.org/10.1007/s11159-008-9105-0>. [4]
- Shapiro, A. and D. Niederhauser (2004), "Learning from Hypertext: Research Issues and Findings", in Jonassen, D. (ed.), *Handbook of research on educational communications and technology*, Lawrence Erlbaum, Mahwah, NJ, <http://www.compassproject.net/sadhana/teaching/readings/23.pdf> (accessed on 22 May 2019). [30]
- Sireci, S. and A. Zenisky (2006), "Innovative Item Formats in Computer-Based Testing: In Pursuit of Improved Construct Representation", in Downing, S. and T. Haladyna (eds.), *Handbook of test development*, Lawrence Erlbaum Associates, Mahwah, NJ, US, <https://psycnet.apa.org/record/2006-01815-014> (accessed on 23 May 2019). [21]
- Sukkarieh, J., M. von Davier and K. Yamamoto (2012), "FROM BIOLOGY TO EDUCATION: SCORING AND CLUSTERING MULTILINGUAL TEXT SEQUENCES AND OTHER SEQUENTIAL TASKS", *ETS Research Report Series*, Vol. 2012/2, pp. i-43, <http://dx.doi.org/10.1002/j.2333-8504.2012.tb02307.x>. [10]
- van der Ark L. et al. (eds.) (2015), *Identifying Feature Sequences from Process Data in Problem-Solving Items with N-Grams*, Springer. [54]
- van der Linden, W. (2005), *Linear Models for Optimal Test Design*, Springer New York, New York, NY, <http://dx.doi.org/10.1007/0-387-29054-0>. [22]
- van der Linden, W., R. Klein Entink and J. Fox (2010), "IRT Parameter Estimation With Response Times as Collateral Information", *Applied Psychological Measurement*, Vol. 34/5, pp. 327-347, <http://dx.doi.org/10.1177/0146621609349800>. [48]
- Weeks, J., M. von Davier and K. Yamamoto (2013), *Using response time data to inform the coding of missing responses*. [50]
- Winne, P. and R. Baker (2013), "The Potentials of Educational Data Mining for Researching Metacognition, Motivation and Self-Regulated Learning.", *Journal of Educational Data Mining*, Vol. 5/1, pp. 1-8, <https://eric.ed.gov/?id=EJ1115337> (accessed on 22 May 2019). [32]
- Xia, X. (2007), *Bioinformatics and the cell : modern computational approaches in genomics, proteomics, and transcriptomics*, Springer. [65]
- Xu, H. et al. (2018), "Latent Class Analysis of Recurrent Events in Problem-Solving Items", *Applied Psychological Measurement*, Vol. 42/6, pp. 478-498, <http://dx.doi.org/10.1177/0146621617748325>. [7]

- Zhu, M. and G. Feng (2015), *An exploratory study using social network analysis to model eye movements in mathematics problem solving*, ACM Press, New York, New York, USA, [18]
<http://dx.doi.org/10.1145/2723576.2723591>.
- Zimmerman, B. (2008), "Investigating Self-Regulation and Motivation: Historical Background, Methodological Developments, and Future Prospects", *American Educational Research Journal*, [31]
Vol. 45/1, pp. 166-183, <http://dx.doi.org/10.3102/0002831207312909>.
- Zou, M. et al. (2010), "Sequential Pattern Analysis of Learning Logs: Methodology and Applications", in [42]
Romero, C. et al. (eds.), *Handbook of Educational Data Mining*, CRC Press,
<http://dx.doi.org/10.1201/b10274>.

Annex A. Syntax of longest common subsequence in python

```
from numpy import zeros

seq_a = "ACCGGTGGACAATTCA"
seq_b = "GGAAAGAGATATGCAC"

t = zeros((len(seq_a), len(seq_b)))

# Forward pass -- populate lengths
for i in range(1, len(seq_a)):
    for j in range(1, len(seq_b)):
        if seq_a[i] == seq_b[j]:
            t[i][j] = t[i-1][j-1]+1
        else:
            t[i][j] = max(t[i-1][j], t[i][j-1])

# Backwards pass -- extract string
LCS = ""

while i > 0 and j > 0:
    if t[i][j] == t[i-1][j]:
        i = i-1
    elif t[i][j] == t[i][j-1]:
        j = j-1
    else:
        LCS = seq_a[i] + LCS
        i = i-1
        j = j-1

print(LCS)
```

Annex B. Data preparation and cleaning

The preparatory phase was the most time-consuming of the whole procedure. The phase included data cleaning and validation, restructuring process data, dataset linking, codebook development and predefining action sequences with experts.

First, the process dataset and corresponding cognitive dataset were cross validated. Process data for PS2 were extracted from the raw log files and structured into time-stamped action sequences for each individual. In general, the action sequences began with action “START” and ended with “NEXT_OK” for a given PSTRE item. However, due to unexpected technical issues, errors were possible during data collection. For instance, some action sequences might not be able to be closed as a full loop, or the process data could not be matched with the response data. For such partially missing records, we tried to recover the information first. If the recovered information was insufficient to show the full picture of the person’s action sequences and responses, the records were removed from the study. This data cleaning process provided an extra source of data validation and information recovery for missing data. Keystroke actions were ignored, because their processing in multiple languages would have been too complicated. Keystrokes are nonetheless included in the counting of the total number of actions per respondent per item.

Next, codebook development, one of the core tasks in process data analysis, was performed. The event names and action labels used in log files were originally designed for item developers to debug errors rather than for research purposes. Thus, these records were usually not research-friendly. For instance, clicking on the toolbar search function was labelled as “clickbox1” in the raw log file, which was not easily understood or used. It was then recorded in a more meaningful way as “Toolbar_search” to facilitate further study. Consequently, the codebook played a role as a “bridge” to link from the raw log records and provide meaningful understanding. The unique event names for one item from a raw log file on average ranged from 200 to 1 500, including keystroke actions, which is a very challenging total to examine, with no very efficient way to make the computation. The codebook was condensed with a small range of meaningful actions to a range of 30 to 60 events per item. To ensure the actions included in the codebook contained sufficient information and our understanding on the event names was correct, a thorough validation check was performed with the test platform developers and test developers and experts in designing problem-solving items.

The last step for preparation was to work together with experts and item developers to define the optimal action sequences for each item. Here, the reference sequences meant the sequences that the respondents were supposed to follow from the test developers’ point of view. These predefined sequences were not necessarily the unique solutions for individuals to solve the problems, but were considered the optimal way when initially designed by test developers.