Chapter 10
Clustering Behavioral Patterns Using Process Data in PIAAC Problem-Solving Items

Qiwei He, Dandan Liao and Hong Jiao

Abstract Technical advances provide the possibility of capturing timing and process data as test takers solve digital problems in computer-based assessments. The data collected in log files, which represent information beyond response data (i.e., correct/incorrect), are particularly valuable when examining interactive problem-solving tasks to identify the step-by-step problem-solving processes used by individual respondents. In this chapter, we present an exploratory study that used cluster analysis to investigate the relationship between behavioral patterns and proficiency estimates as well as employment-based background variables. Specifically, with a focus on the sample from the United States, we drew on a set of background variables related to employment status and process data collected from one problem-solving item in the Programme for the International Assessment of Adult Competencies (PIAAC) to address two research questions: (1) What do respondents in each cluster have in common regarding their behavioral patterns and backgrounds? (2) Is problem-solving proficiency related with respondents' behavioral patterns? Significant differences in problem-solving proficiency were found among clusters based on process data, especially when focusing on the group not solving the problem correctly. The results implied that different problem-solving strategies and behavioral patterns were related to proficiency estimates. What respondents did when not solving digital tasks correct was more influential to their problem-solving proficiency than what they did when getting them correct. These results helped us understand the relationship between sequences of actions and proficiency estimates in large-scale assessments and held the promise of further improving the accuracy of problem-solving proficiency estimates.

Q. He (✉)
Educational Testing Service, Princeton, USA
e-mail: qhe@ets.org

D. Liao
American Institutes for Research, Washington DC, USA

H. Jiao
University of Maryland, College Park, MD, USA

B. P. Veldkamp and C. Sluijter (eds.), *Theoretical and Practical Advances in Computer-based Educational Measurement*, Methodology of Educational Measurement and Assessment, https://doi.org/10.1007/978-3-030-18480-3_10

10.1 Introduction

The use of computers as an assessment 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 about the processes that test takers engage in when completing assessment tasks (He and von Davier 2016). The data collected in log files, which are unique to computer-based assessments, provide information beyond response data (i.e., correct/incorrect) that is usually referred to as *process data*. Such information is particularly valuable when examining interactive problem-solving tasks to identify the step-by-step problem-solving processes used by individual respondents.

10.1.1 Problem-Solving Items in PIAAC

As the largest and most innovative international assessment of adults, the Programme for the International Assessment of Adult Competencies (PIAAC), starting from the first cycle in 2012, has sought to assess computer, digital-learning, and problem-solving skills, which are essential in the 21st century (Organisation for Economic Co-operation and Development [OECD] 2009, 2011, 2012; Schleicher 2008). Of significance here, PIAAC is the first international household survey of skills predominantly collected using information and communication technologies (ICT) in a core assessment domain: Problem Solving in Technology-Rich Environments (PSTRE). This international survey has been conducted in over 40 countries and measures the key cognitive and workplace skills needed for individuals to participate in society and for economies to prosper (OECD 2016). Evidence has shown that process data captured in PSTRE items provide a deeper insight into the cognitive processes used by respondents when they are solving digital tasks (e.g., Goldhammer et al. 2014; Liao et al. 2019; Chen et al. 2019). This additional information helps us understand the strategies that underlie proficient performance and holds the promise of better identifying behavioral patterns by subgroups, thus helping us seek solutions for teaching essential problem-solving skills to adults with particular needs (He and von Davier 2015, 2016).

The PSTRE items that we focused on in this study were used to assess the skills required to solve problems for personal, work, and civic purposes by setting up appropriate goals and plans, as well as how individuals access and make use of information through computers and networks (OECD 2009). This new domain involved more interactive item types and was available only on computers. The construct underlying the PSTRE items describes skillful use of ICT as collecting and evaluating information for communicating and performing practical tasks such as organizing a social activity, deciding between alternative offers, or judging the risks of medical treatments (OECD 2009). To give a response in the simulated computer environments that form the PSTRE tasks, participants were required to click buttons or links, select from dropdown menus, drag and drop, copy and paste, and so on (He and von Davier 2016).

10.1.2 Employability and PSTRE Skills

Employability and the completion of computer-based testing have been shown to be positively correlated in recent research (e.g., OECD 2013b; Vanek 2017; Liao et al. 2019). Although the United States was one of the PIAAC countries with the highest accessibility to computers, internet, and advanced electronic equipment, its performance, especially in the PSTRE domain that focuses on assessing ICT skills, was far lower than expectations. According to a recent report published by the National Center for Education Statistics (Rampey et al. 2016), U.S. test takers scored lower on average than their international peers, ranking in the lowest tier in the PSTRE domain as a country, and having the largest proportion of respondents below Level 1, which is the minimum proficiency level required to complete simple problem-solving tasks in daily life (OECD 2013a). These results raise attention to adults' current PSTRE skills in the U.S. population and their employability, which is highly associated with PSTRE skills.

This chapter presents an exploratory study that used cluster analysis to investigate the relationship between behavioral patterns and proficiency estimates as well as employment-based background variables. Specifically, with a focus on the sample from the United States, we drew on a set of background variables related to employment status and process data collected from one PSTRE item in PIAAC to address two research questions: (1) What do respondents in each cluster have in common regarding their behavioral patterns and backgrounds? (2) Is problem-solving proficiency consistent across clusters, or in other words, is problem-solving proficiency related to respondents' behavioral patterns?

10.2 Method

10.2.1 Sample

The PIAAC sample was representative of the population of adults with an age range of 16–65 years old who had prior experience with computers. Those who had never used computers were excluded from the problem-solving section; the task for this scale was by default (and by definition of the construct) assessed only on a computer-based testing platform (OECD 2010). A total of 1,340 test takers in the U.S. sample who completed the PSTRE items in the second module (PS2)[1] in PIAAC were included in

[1] The PIAAC was designed in a structure of multistage adaptive testing, by routing respondents to different modules in two stages. The PSTRE domain consists of two modules (PS1 and PS2),

the present study. Of them, there were 629 female test takers (46.9%) and 711 male test takers (53.1%). The mean age was 39.2 years (*SD* = 14.0). A majority (680) of members of this sample had an educational level above high school (50.7%), whereas 534 reported completing high school (39.9%), 124 reported less than high school (9.3%), and two cases were recorded as missing (0.1%). Please note that there were 14 cases that could not be matched between the process data and the background variables[2] and thus had to be removed in further data analysis, which resulted in 1,326 test takers in the final sample.

10.2.2 Instrumentation

A total of 14 PSTRE items were administered in the PIAAC main study. We focused on the process data resulting from the task requirements of one item. The "Meeting Room Assignment" item (U02) consisted of three environments: email, web, and word processor. The task required respondents to view a number of emails, identify relevant requests, and submit three meeting room requests using a simulated online reservation site. Meanwhile, a conflict between one request and existing schedule presented an impasse that respondents had to resolve. In the interactive environment, test takers could switch among the three environments, go back and forth to understand the meeting room requests, make reservations or changes, and copy and paste the key information in the word processor environment.[3] An interim score was evaluated based only on the meeting reservation webpage. According to the scoring rule, full credit (3 points) was granted when the respondents correctly submitted all three meeting room requests, and partial credit (1 or 2 points) was given when only one or two requests were submitted successfully. No credit (0 points) was given when none of the requests was correctly fulfilled.

According to the PIAAC technical report (OECD 2016), item U02 was one of the most difficult in the PSTRE domain, with difficulty and discrimination parameter estimates[4] of 0.78 and 1.18, respectively, ranking at difficulty level 3. In the U.S. sample, 932 (70%) test takers received no credit, 294 (22%) received partial credit,

positioned in stage 1 and stage 2, respectively. Each of the modules contains seven items without overlap to each other. The seven items within one module has a fixed position. More details about PIAAC test design refer to PIAAC technical report (OECD 2016).

[2] Process data extracted from the log file and response data from the background questionnaire could be linked with the unique respondent IDs. However, given possible technical issues in data collection, there might exist cases with only process data or only background variables. These cases had to be discarded during analysis as data could not be matched.

[3] Word processor was an optional environment instead of a compulsory one, designed to help the respondents summarize information extracted from the email requests.

[4] Two-parameter-logistic item response modeling was applied in the PIAAC data analysis to estimate the latent trait of test takers' problem-solving skills. The parameter estimates presented here are the common international parameters generally used across countries. For details on data analysis modeling in PIAAC, refer to the PIAAC technical report (OECD 2016).

Table 10.1 Descriptive statistics of number of actions and response time (in minutes) in U02

Features	Mean	SD	Min	Max
Number of actions on U02	34.06	33.89	0.00	194.00
Response time on U02	3.60	3.47	0.09	45.07

and only 114 (9%) received full credit. To explore the difference between test takers who got at least part of the item correct and those who received no credit, the polytomous scores were dichotomized by collapsing partial and full credit in the present study.

Further, the item U02 required a relatively long sequence to solve the task. On average, respondents took 34 actions over 3.6 minutes to complete the task.[5] It is also noted that the distributions of the number of actions and response time were widely spread out. As presented in Table 10.1, the longest sequence in this item used 194 actions over 45 minutes, while the shortest sequence was zero actions and 0.09 minutes: obviously a quick skipping behavior resulting in a missing response. These statistics implied that the behavioral patterns and strategies differed considerably by test takers. This sparse distribution would also impact the feature extraction, which is discussed in detail in the next section.

There are four reasons we selected item U02 as an example. First, as mentioned above, this rather difficult item could potentially provide more information to identify reasons for failure when tracking respondents' process data. Researchers have found that for an item that is difficult but not extremely so, test takers tend to demonstrate more heterogeneous use of strategies, aberrant response behavior, and variant use of response time (e.g., de Klerk et al. 2015; Goldhammer et al. 2014; Vendlinski and Stevens 2002). Second, this item consisted of multiple environments (email, web, and word processor). The action sequences are expected to be more diverse in a problem-solving item with multiple environments than an item with a single environment. Hence, it is possible to extract more information from this item. Third, item U02 had a fixed position in the middle of the PS2. Compared to items at the beginning or the end, items in the middle of a booklet are less likely to demonstrate position effect (e.g., Wollack et al. 2003). Lastly, item U02 shared environments with most items in PS2. This provided the possibility to further investigate the consistency of problem-solving strategies across items for each individual.

10.2.3 Features Extracted from Process Data

10.2.3.1 N-Gram Representation of Sequence Data

The strategy for analyzing item U02 was motivated by the methodologies and applications in natural language processing (NLP) and text mining (e.g., He et al. 2012;

[5] There is no time limitation in the PIAAC cognitive test.

Sukkarieh et al. 2012; He et al. 2017). We chose the n-grams model to disassemble the sequence of data while retaining the sequential order. As He and von Davier (2015, 2016) introduced, unigrams—analogous to the language sequences in NLP—are defined as "bags of actions," where each single action in a sequence collection represents a distinct feature. An n-gram is defined as a contiguous sequence of *n* words in text mining; similarly, when analyzing action sequences from process data, an n-gram can be defined as a sequence of *n* adjacent actions (Manning and Schütze 1999). Bigrams and trigrams are defined as action vectors that contain either two or three ordered adjacent actions, respectively. For instance, here is a typical sequence for email review actions: "MAIL_VIEWED_4, MAIL_VIEWED_2, MAIL_VIEWED_1". The unigram is each of the three separate actions (e.g., "MAIL_VIEWED_4"), a bigram is two adjacent actions as one unit, (e.g., "MAIL_VIEWED_2, MAIL_VIEWED_1"), and the trigram is three adjacent actions as one unit (e.g., "MAIL_VIEWED_4, MAIL_VIEWED_2, MAIL_VIEWED_1"). Of note is that the n-gram method was productive in creating features from sequence data without losing too much information in terms of the order in the sequence (He et al. 2018). This approach is a widely accepted tool for feature engineering in fields such as NLP and genomic sequence analysis.

A total of 34 actions (i.e., unigrams) were defined for this item and are listed in Table 10.2. The interpretation describing each action is presented as well. The frequency of sequences that contain the action by each row is shown in the right-hand column.

Besides the unigram features, we also included the total response time and the number of actions as features in the cluster analysis. These two features also showed up in a preliminary principal component analysis as the most influential features with the highest loadings. This resulted in 36 features altogether. Given concerns about the low frequency of bigrams and trigrams, the features from mini sequences were not used in the cluster analysis in this study.

10.2.3.2 Term Weights

Three types of term weights were used in the current study: sampling weights as well as between- and within-individual weights. Between-individual weights highlight how different the frequency of a certain action is among individuals, whereas within-individual weights capture how some actions are used more often than others by an individual. Regarding between-individual differences, a popular weighting method in text mining, inverse document frequency (IDF; Spärck Jones 1972), was renamed as inverse sequence frequency (ISF) and adapted for estimating the weight of each *n-gram*. ISF is defined as $ISF_i = \log(N/sf_i) \geq 0$, where N denotes the total number of sequences in the sample, which is the same as the total number of test takers, and sf_i represents the number of sequences containing action i. A large ISF reflects a rare action in the sample, whereas a small ISF represents a frequent one.

Within-individual differences had to be considered when an individual took some actions more often than others. Although more frequent sequences are more impor-

Table 10.2 Description and frequency of unigrams

No.	Features	Description	Frequency
1	FOLDER_VIEWED	View a folder	5,762
2	ENVIRONMENT_WB	Go to web environment	4,715
3	ENVIRONMENT_MC	Go to email environment	4,317
4	MAIL_VIEWED_1	View 1st email	2,725
5	HISTORY_VIEWCALENDAR	Go to calendar tab in web environment	2,190
6	MAIL_VIEWED_3	View 3rd email	1,968
7	HISTORY_RESERVATION	Go to reservation tab in web environment	1,935
8	COMBOBOX_ROOM	Choose a room when filling out a room request	1,891
9	MAIL_VIEWED_4	View 4th email	1,698
10	MAIL_VIEWED_2	View 2nd email	1,544
11	MAIL_MOVE	Move an email	1,499
12	NEXT_INQUIRY	Go to next item	1,371
13	START	Start item U02	1,326
14	COMBOBOX_START_TIME	Choose start time when filling out a room request	1,312
15	COMBOBOX_END_TIME	Choose end time when filling out a room request	1,304
16	COMBOBOX_DEPT	Choose department when filling out a room request	1,296
17	HISTORY_MEETINGROOMS	Go to meeting room details tab in web environment	1,058
18	ENVIRONMENT_WP	Go to word processor environment	987
19	SUBMIT_RESERVATION_FAILURE	Submit a reservation request unsuccessfully	987
20	SUBMIT_RESERVATION_SUCCESS	Submit a reservation request successfully	971
21	HISTORY_UNFILLED	Go to unfilled tab in the web environment	551
22	SUBMIT_UNFILLED	Submit an unfilled request	414
23	FOLDER	Do folder-related actions (i.e., create/delete a folder)	332
24	HISTORY_HOME	Click on the home button in the web environment	244
25	CHANGE_RESERVATION	Change an existing reservation	227
26	KEYPRESS	Type in word processor environment	152

(continued)

Table 10.2 (continued)

No.	Features	Description	Frequency
27	REPLY	Reply an email	118
28	CANCEL	Click on cancel button	111
29	HELP	Use help function	87
30	COPY	Use copy function	42
31	SEARCH	Use search function	38
32	SORT	Use sort function	21
33	PASTE	Use paste function	15
34	BOOKMARK	Do bookmark-related actions (i.e., add/delete a bookmark)	13

tant than less frequent ones for each individual, the raw frequencies of these action sequences often overestimate their importance (He and von Davier 2015). To account for within-individual differences in the importance of action sequences, a weighting function was employed $f\left(tf_{ij}\right) = 1 + \log\left(tf_{ij}\right)$, where $tf_{ij} > 0$ represents the frequency of action i in each individual sequence j (Manning and Schütze 1999). Combining the between- and within-individual weights, the final action weight can be defined as $weight(i, j) = \left[1 + \log\left(tf_{ij}\right)\right] \log(N/sf_i)$ for $tf_{ij} \geq 1$ (He and von Davier 2015, 2016). Compared to raw frequency, this weighting mechanism was applied for attenuating the effect of actions or action vectors that occurred too often to be meaningful.

The sampling weights were also taken into consideration in this study. In fact, we conducted the cluster analyses both with and without sampling weights, and the differences were marginal. Therefore, we report results only with sampling weights in the next section.

10.2.4 Clustering Sequence Data

Clustering has been widely recognized as a powerful unsupervised data mining approach for grouping similar data points. Unlike supervised learning approaches that typically train a model on known input (data and labels) to predict future outputs, unsupervised learning approaches focus on finding hidden patterns or intrinsic structures in input data (Manning and Schütze 1999). Sequence clustering aims at partitioning sequences into meaningful clusters consisting of similar sequences (Ferreira et al. 2007). It has been applied in various fields, such as gene structure exploration in biology, students' learning progression in education, and pattern recognition in industrial engineering.

To cluster sequence data, it is important to choose a clustering algorithm that is appropriate for the characteristics of the data and sequence features (Dong and Pei

2007). Some popular clustering methods include hierarchical clustering (e.g., Huang et al. 2010; Johnson 1967; Navarro et al. 1997), graph-based clustering (e.g., Kawaji et al. 2001; Felzenszwalb and Huttenlocher 2004), K-means (e.g., Bustamam et al. 2017; Gasch and Eisen 2002; Park et al. 2008), and others. Hierarchical clustering is a method of cluster analysis that, as the name indicates, seeks to build a hierarchy of clusters. Strategies for hierarchical clustering generally fall into two types. They are: (1) agglomerative (Johnson 1967)—a "bottom up" approach in which each observation starts in its own cluster and pairs of clusters are merged as one moves up the hierarchy, and (2) divisive (MacNaughton-Smith et al. 1964)—a "top down" approach in which all observations start in one cluster and splits are performed recursively as one moves down the hierarchy. Graph-based clustering algorithms generally involve two major steps. In the first step, a weighted graph is constructed from the sequences. In the second, the graph is segmented into subgraphs that correspond to the clusters (Dong and Pei 2007). K-means is one of the simplest learning algorithms to solve clustering problems. The procedure follows a straightforward way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. The main idea of the K-means algorithm is to discover K (nonoverlapping) clusters by finding K centroids ("central" points) and then assigning each point to the cluster associated with its nearest centroid (Jyoti and Singh 2011).

Our current study adopted the K-means algorithm to cluster the behavioral patterns from one PSTRE item U02 based on features extracted from process data. The reasons for choosing this algorithm can be explained from three aspects: First, K-means is efficient in terms of computational cost even with a large number of variables, which renders wider applications possible in large-scale assessments, especially for complex multidimensional data structures in process data. Second, observations can switch from one cluster to another when the centroids of the clusters are recomputed. This shows that K-means is able to recover from potential mistakes in clustering. However, it also indicates that results from K-means could be strongly influenced by the selection of initial seeds (e.g., Arthur and Vassilvitskii 2007). Therefore, the impact of selecting initial seeds should be carefully examined before interpreting the results, as we did in this study. Third, results of K-means are easily interpretable. Each observation belongs to only one cluster, and the centroids of the clusters are expressed on the scales of the variables. More details about the analytic strategy and algorithms are introduced in the next section.

10.2.5 K-Means Clustering

The K-means algorithm (Lloyd 1982) was adopted for the cluster analysis in the current study. This method starts with k arbitrary centroids and seeks to minimize the squared difference between observations in the same cluster. A cluster centroid is typically the mean or median of the points in its cluster and "nearness" is defined by a distance or similarity function. Ideally the centroids are chosen to minimize the total "error," where the error for each point is given by a function that measures the discrepancy between a point and its cluster centroid, for example, the squared distance. Note that a measure of cluster "goodness" is the error contributed by that cluster (Alphaydin 2009).

The basic K-means algorithm for finding K clusters is as follows:

1. Select K points as the initial centroids.
2. Assign all points to the closest centroid.
3. Recompute the centroid of each cluster.
4. Repeat steps 2 and 3 until the centroids do not change (or change minimally).

The first step is to define k centroids, one for each cluster. These centroids should be placed with careful consideration because different locations cause different results. The best choice is to place them as far away from each other as possible. The next step is to take each point belonging to a given data set and assign it to the nearest centroid. When no point is pending, the first step is completed and an early group membership is done. At this point we need to recalculate k new centroids. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid. This generates a loop. As a result of this loop, we could notice that the k centroids may change their location step by step until no more changes occur. In other words, the centroids do not move anymore. Finally, this algorithm aims at minimizing a function of a matrix, for instance, a squared error function (Steinbach et al. 2004).

Unlike the hierarchical algorithms that produce a nested sequence of partitions, K-means is one of the nonhierarchical algorithms that often start out with a partition based on randomly selected seeds, and then refine this initial partition (Manning and Schütze 1999). The initial cluster centers for K-means are usually picked at random. Whether the choice of initial centers is important or not depends on the structure of the set of objects to be clustered (Jyoti and Singh 2011).

In this study, we used 36 features—34 unigrams plus response time and total number of action sequences—extracted from the process data of item U02 to partition test takers into clusters using the K-means clustering method. An appropriate number of clusters, k, was selected based on the change in the total within-cluster sum of squares. As noted previously, one potential uncertainty of K-means is that the clustering results could be strongly influenced by the selection of initial seeds (e.g., Arthur and Vassilvitskii 2007). Therefore, the stability of the cluster membership was examined to maximize the generalizability of the results. Further, clusters were interpreted based on the centroids of the 36 features. We explored the homogeneous characteristics of the clusters, as well as the relationship between cluster membership and proficiency level and/or correctness of U02.

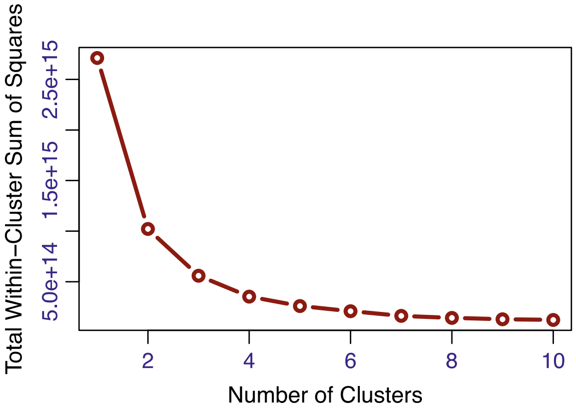


Fig. 10.1 Total within-cluster sum of squares at different number of clusters for the item U02

10.3 Results

10.3.1 Cluster Determination

The basic idea behind cluster partitioning methods, such as K-means clustering, is to define clusters such that the total intra-cluster variation (known as total within-cluster variation or total within-cluster sum of squares) is minimized. We used three methods to determine the optimal number of clusters for the data set, the elbow method (e.g., Ketchen and Shook 1996; Thorndike 1953), the average silhouette method (e.g., Kaufman and Rousseeuw 1990), and the hierarchical clustering method (e.g., Ward 1963). These three methods were chosen to provide insights about the structure of the data through visualization and statistical measures and to mutually validate the results from each.

For the elbow method, substantial drops in total within-cluster sum of squares were present when the number of clusters was set from one to three. After the "elbow point" of three, the changes became marginal despite an increasing number of clusters (see Fig. 10.1).

The average silhouette approach measures the quality of a clustering. That is, it determines how well each object lies within its cluster. A high average silhouette width indicates a good clustering. The average silhouette method computes the average silhouette of observations for different values of k. The optimal number of clusters k is the one that maximizes the average silhouette over a range of possible values for k. Given the sample in hand, the highest average silhouette width was shown when two clusters were chosen.

The hierarchical clustering method seeks to form a hierarchy of clusters either through bottom-up or top-down approach. Such an algorithm either starts with all observations in different clusters and merges them into clusters, or starts with all observations in one cluster and gradually partitions into clusters. In the current study, a bottom-up hierarchical clustering method was employed. Given the sample in hand, the results from this method were that the optimal cluster number could be between two and four. Based on mutual validation by these three methods as well as considerations on cluster sample size and interpretability of the results, we chose the three-cluster solution for further investigations. After determining the number of clusters, the clustering analysis was rerun and results were reported based on this.

As mentioned previously, K-means clustering usually starts from a random selection of initial seeds even though using more discriminative seeds that are able to separate the clusters is highly recommended as the initial partition (Arthur and Vassilvitskii 2007). Although many sets are well behaved and most initializations will result in clustering of about the same quality (Manning and Schütze 1999), it would be wise to examine the stability of cluster membership to maximize the generalizability before interpreting the clustering results.

We checked the stability of cluster membership with 100 different initial seeds in the item U02. Among the 1,326 test takers, 1,262 (95%) had no changes in cluster membership in the 100 replications. Only 64 (5%) were assigned to a different cluster in at most 10% of the replications. Overall, only 0.3% of the test-taker-replication combinations demonstrated uncertainty in the cluster membership. This suggested that the clustering results had very little dependence on initial seeds and thus the seeds could be ignored in this study.

We list the centroids of the three-cluster solution in Table 10.3. Note that the term weights and sampling weights were taken into account when the values of centroids were computed. For the 34 unigrams, values presented in Table 10.3 were based on action frequencies weighted by term weights and sampling weights; for the number of actions and response time on U02, the two features were weighted by sampling weights before computing the centroids. In general, Cluster 1 had the lowest weighted frequencies and means in almost all features and Cluster 3 had the highest ones, while Cluster 2 placed between Cluster 1 and Cluster 3. The action unigrams “NEXT_INQUIRY” and “START” had centroids at zero across all three clusters, suggesting that all test takers had taken these two actions, which led to them providing little information in the analysis. When all test takers perform the same action, the ISF of an action would be zero by definition. Thus, these two unigram features did not actually contribute in the clustering because of the zero information. As expected, the number of actions and response time appeared to be the most dominant features in clustering. The reason is probably that these two variables are of a different granularity than the others, as they summarize information for the entire sequence, rather than a partial contribution made by a single action. These two features also showed up in a preliminary principal component analysis as the most influential features with the highest loadings.

The three clusters could be interpreted as test takers with the least, medium, and most effort. The least-action cluster had the largest cluster size with 853 (64%) of the test takers in the analytical sample, the median action cluster had 398 (30%) test takers, and only 75 (6%) were in the most-action cluster (see Table 10.4). This indicated that only a small group of test takers had a great number of actions and spent a long time exploring U02, whereas the majority clustered around fewer actions and a much shorter time.

Table 10.3 Cluster centroids for a three-cluster solution

No.	Features	Clusters		
		1	2	3
1	FOLDER_VIEWED	20,365.6	40,224.5	73,644.8
2	ENVIRONMENT_WB	11,062.0	65,848.4	134,117.1
3	ENVIRONMENT_MC	11,375.0	59,744.3	124,133.6
4	MAIL_VIEWED_1	8,625.7	23,554.9	42,659.6
5	HISTORY_VIEWCALENDAR	8,296.1	61,208.4	130,657.0
6	MAIL_VIEWED_3	7,983.3	41,671.1	84,185.9
7	HISTORY_RESERVATION	6,972.4	53,034.1	117,019.4
8	COMBOBOX_ROOM	6,020.8	54,476.1	110,608.7
9	MAIL_VIEWED_4	8,606.6	35,180.8	67,087.3
10	MAIL_VIEWED_2	7,891.8	33,636.2	65,864.9
11	MAIL_MOVE	18,947.5	42,469.4	87,984.2
12	NEXT_INQUIRY	0.0	0.0	0.0
13	START	0.0	0.0	0.0
14	COMBOBOX_START_TIME	5,498.0	47,928.2	101,684.2
15	COMBOBOX_END_TIME	5,581.8	47,942.1	103,098.3
16	COMBOBOX_DEPT	5,556.0	48,052.1	101,711.1
17	HISTORY_MEETINGROOMS	5,848.3	43,725.6	108,077.0
18	ENVIRONMENT_WP	7,738.8	33,937.1	79,654.0
19	SUBMIT_RESERVATION_FAILURE	4,048.2	46,768.2	109,482.7
20	SUBMIT_RESERVATION_SUCCESS	4,797.0	42,081.0	85,547.9
21	HISTORY_UNFILLED	4,213.2	36,222.2	91,450.9
22	SUBMIT_UNFILLED	3,589.7	34,291.9	69,265.5
23	FOLDER	6,750.6	25,942.1	62,512.5
24	HISTORY_HOME	3,808.0	18,614.7	50,805.3
25	CHANGE_RESERVATION	1,522.0	23,168.2	73,968.0
26	KEYPRESS	2,880.5	12,713.7	65,743.1
27	REPLY	2,936.5	12,319.8	30,153.8
28	CANCEL	3,250.7	13,530.1	37,320.8
29	HELP	3,477.5	10,343.4	17,039.6
30	COPY	897.1	7,628.4	38,517.4
31	SEARCH	2,278.3	3,529.5	18,895.4

(continued)

Table 10.3 (continued)

No.	Features	Clusters		
		1	2	3
32	SORT	949.7	4,759.2	5,540.5
33	PASTE	780.4	1,494.6	33,561.6
34	BOOKMARK	550.6	2,875.4	9,264.9
35	Number of actions on U02	453,665.4	2,241,595.2	5,225,357.2
36	Response time on U02	58,391.4	244,418.8	475,306.6
	Frequency	853	398	75

Note "NEXT_INQUIRY" and "START" show 0 cluster centroids across all three clusters. As all participants used them, they had zero term weights and did not contribute in the clustering analysis

Table 10.4 Cluster size of a three-cluster solution

U02score	Clusters		
	1	2	3
0	760	139	23
1	93	259	52
Total	853	398	75

Note The U02score has combined the partial and full credit into "1"

10.3.2 Cluster Membership and Proficiency Level

Based on the clusters derived from process data as described above, we investigated the relationships between cluster membership and PSTRE proficiency level as well as employment-related variables. To increase the accuracy of the cognitive measurement for various subpopulations and the population as a whole, PIAAC uses plausible values—which are multiple imputations—drawn from a posteriori distribution by combining the item response scaling of the cognitive items with a latent regression model using information from the background questionnaire. The "plausible value" methodology correctly accounts for error (or uncertainty) at the individual level by using multiple imputed proficiency values (plausible values) rather than assuming that this type of uncertainty is zero (for details about how the set of plausible values is generated and interpreted, refer to OECD 2016). In PIAAC, 10 plausible values (PV) are generated for each domain as the estimates of scale scores. As all 10 PVs showed similar patterns, we used only the first PV (PV1) as an example for illustration purposes. Figure 10.2 depicted the association between clusters and PSTRE proficiency level (PV1). To explore whether significant differences existed among the clusters regarding PV1, we conducted one-way analysis of variance (ANOVA). Results showed that the three clusters had significantly different proficiency levels as measured by PV1, $F(2,1323) = 254.6$, $p < 0.001$.

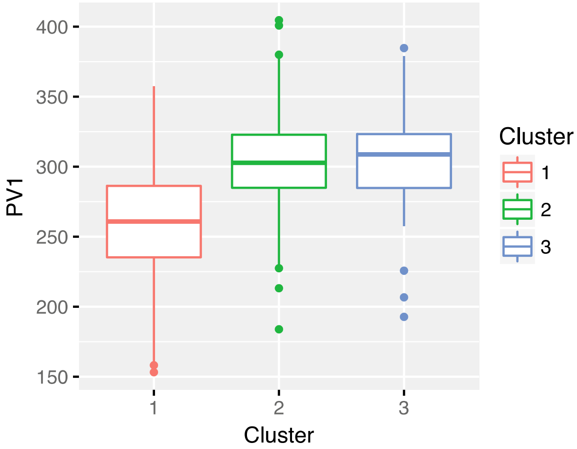
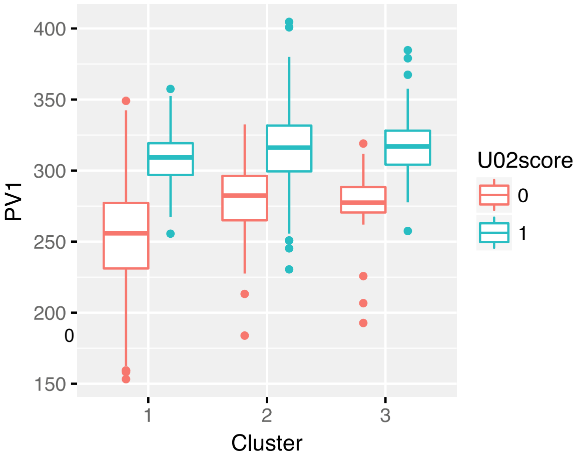


Fig. 10.2 A boxplot of PV1 by cluster membership

Given the results from ANOVA, we further conducted a post hoc pairwise comparison for the three clusters. Given concerns on the unequal sample size by clusters, the pairwise comparison *t*-test method introduced by Benjamini and Hochberg (1995) was employed. This method controls the false discovery rate (the expected proportion of false discoveries) among rejected hypotheses. The false discovery rate is a less stringent condition than the family-wise error-rate-based methods, so this method is more powerful than the others. Notably, a remarkable increase was observed in PV1 from Cluster 1 to Cluster 2, for which the first quartile of Cluster 1 was approximately at the same level as the third quartile of Cluster 2. However, the increase from Cluster 2 to Cluster 3 was marginal. Results showed no significant differences between these two groups. Given the similar proficiency level between clusters 2 and 3, but shorter action sequences and response time in Cluster 2, this might be interpreted as a higher efficiency in Cluster 2 to solve the item U02. That is, both clusters 2 and 3 were more likely to be able to solve the item, but with different strategies and paces.

We further plotted the PV1 distributions by correct and incorrect groups for item U02 for each cluster (see Fig. 10.3). The sample size by each group nested in clusters was reported in Table 10.4. As expected, the majority of those in Cluster 1 did not answer correctly, since only a few actions and a short time were taken. Clusters 2 and 3 tended to have more test takers who were successful in solving U02. In general, across the three clusters, the PV1 of test takers who responded correctly to item U02 was consistently higher than those who responded incorrectly, although the mean difference among the correct groups in pairwise comparisons was not statistically significant. This suggested that the actions or response time did not make a significant impact on how respondents correctly solved the item. Besides, the correct group in Cluster 1 actually could be interpreted as the most efficient group in finding the correct answer since they used the shortest action sequences and response times

Fig. 10.3 A boxplot of PV1 by U02score nested in clusters



across all correct groups by clusters.[6] Comparatively, a significant difference was found among the incorrect groups in the one-way ANOVA, resulting in $F(2, 919) = 55.2$, $p < 0.001$. Similar to the general pattern found in Fig. 10.3, substantial differences were found between Cluster 1 and the other two clusters, whereas little difference was found between Cluster 2 and Cluster 3.

These findings suggested that the correct group applied various problem-solving strategies to obtain a correct answer, and the choice of strategy was not necessarily associated with PSTRE skills in the correct group. As noted above, a small group of test takers in Cluster 1 was able to use only a few actions to solve U02 in a short time, and that group's PSTRE scores were similar to those who applied many more actions. While adopting more actions might be an indication of high motivation to extensively explore the item, it could also signify that the test taker used less efficient strategies when those actions became excessive. For the incorrect group, however, the number of actions and time spent on the item could be informative regarding a test taker's PSTRE skills. A test taker who put more effort into solving U02, even though he or she failed, was more likely to have higher PSTRE skills.

10.3.3 Cluster Membership and Employment-Based Background Variables

To understand the profiles for each cluster and the common characteristics that might be shared within the cluster, we further explored the relationship between problem-solving strategies and background variables. In particular, we focused

[6] In the PIAAC complex problem-solving items, multiple choice items were seldom employed. Item types such as short responses, drag-and-place, and web navigations were used to keep the guessing effect as low as possible.

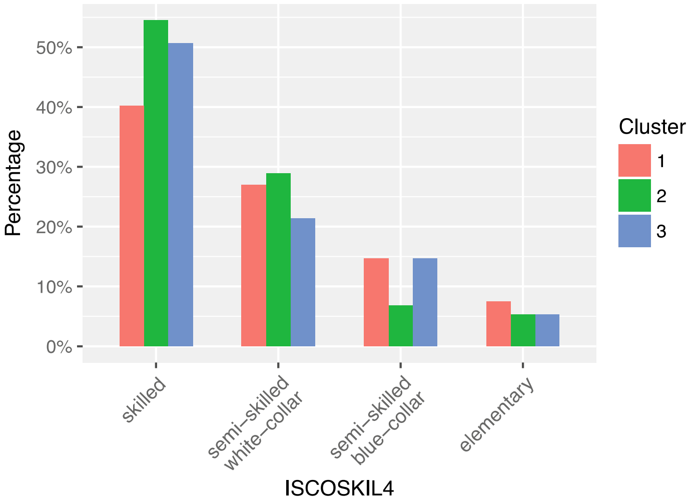


Fig. 10.4 Distribution of ISCOSKIL4 in the three clusters (Percentages in plots represent the percentages of a certain employment status/skill-use level in a cluster.)

on employment-related characteristics: variables related to occupational category, monthly income, work-related skills, age, and education.

Figure 10.4 shows the distribution of the degree of skills (ISCOSKIL4) in all three clusters, which indicates the occupational classification of the test taker's last or current job. Cluster 2 appeared to have the largest proportion of those in skilled and semi-skilled white-collar occupations, while Cluster 1 had the smallest proportion. In contrast, clusters 1 and 3 both had the largest proportion in semi-skilled blue-collar occupations, while Cluster 2 had the smallest. As expected, Cluster 1 had the largest proportion in elementary-level occupations, while the other two clusters shared equally low proportions.

As for the monthly earnings variable (EARNMTHALLDCL) in Fig. 10.5, Cluster 2 and Cluster 3 showed a substantial proportion in the highest earning deciles, from 7th to 10th, whereas Cluster 1 tended to have higher percentages in the lower earning deciles. Two exceptions were found in the first and fourth deciles, in which most test takers were grouped in Cluster 2 and Cluster 3. Despite the general pattern that earnings were mainly positively related to the number of actions and response time spent on the item, some test takers who were younger or at the early stage of their careers may have had lower salaries but higher problem-solving capacity.

Variables regarding work-related skill use also demonstrated similar patterns. Figure 10.6 depicts the distribution of variables for skills related to work: ICT (ICTWORK), numeracy (NUMWORK), reading (READWORK), and writing (WRITWORK). These skills were each divided into five levels in intervals of 20 percentage points, plus a separate category for nonresponses. Cluster 1 was more likely to include test takers in the lower skill-use levels (<40%), while more test takers with high skill use levels (>40%) were in Cluster 2 and Cluster 3. Notably, even though those in

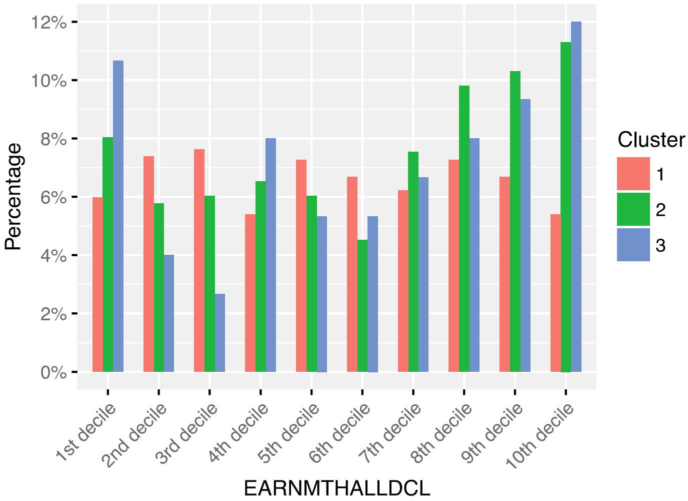


Fig. 10.5 Distribution of EARNMTHALLDCL in the three clusters

Cluster 3 had the largest number of actions and spent the longest time on the item, this group consisted of more test takers in the lower skill levels than Cluster 2. This is also consistent with the finding in the occupational classification that Cluster 3 included a large proportion of test takers in semi-skilled blue-collar occupations, which did not necessarily require higher levels of ICT, numeric, reading, or writing skill use. In addition, considering the item context related to a practical working environment—a meeting room reservation, which is more or less an assistant-style task—this item might not have required as high a level of skills as did other more complex items.

Figure 10.7 exhibits the distribution of the three clusters by five age groups: 24 or less, 25–34, 35–44, 45–54, and 55 plus. Over 30% of test takers in Cluster 3 were younger than 24 years old, representing the highest proportion for this age group. Cluster 2 had the highest proportion in the 25–34 age group, while Cluster 1 had the largest proportion in the oldest group (over 55). This finding provided another perspective for interpreting the pattern observed from process data. Since a large proportion of test takers in Cluster 3 were younger than test takers in the other two clusters, different behaviors could be expected. Compared to the older test takers, younger test takers tended to be more active in human-computer interactions, more familiar with manipulating computer systems, and learned faster when encountering new interfaces. Furthermore, they were expected to exhibit more curiosity about exploring the item, which could increase the number of actions and response time.

Lastly, we took educational backgrounds into consideration. Figure 10.8 presents the distribution of six education levels (EDCAT6) for each cluster. Cluster 1 had the highest percentage for those with a lower/upper secondary level or less of education, whereas Cluster 3 had the highest percentages in the postsecondary and tertiary

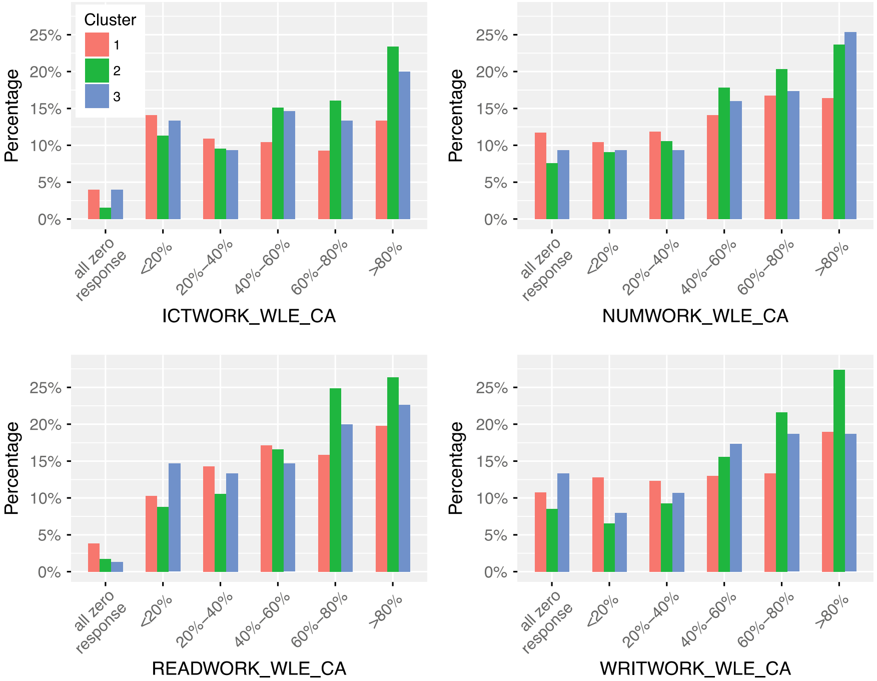


Fig. 10.6 Distribution of ICTWORK_WLE_CA, NUMWORK_WLE_CA, READWORK_WLE_CA, and WRITWORK_WLE_CA in the three clusters

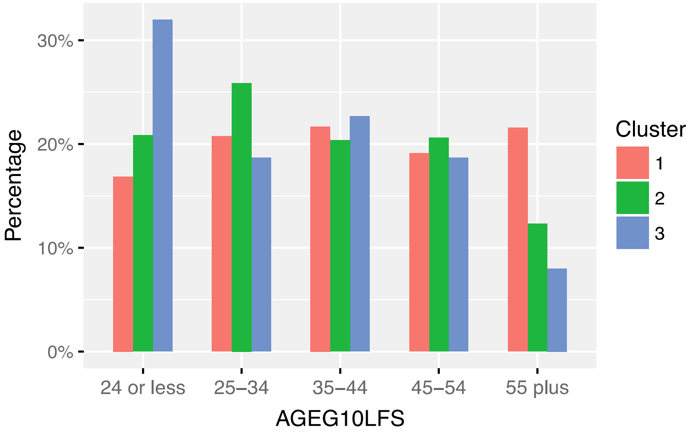


Fig. 10.7 Distribution of AGEG10LFS in the three clusters

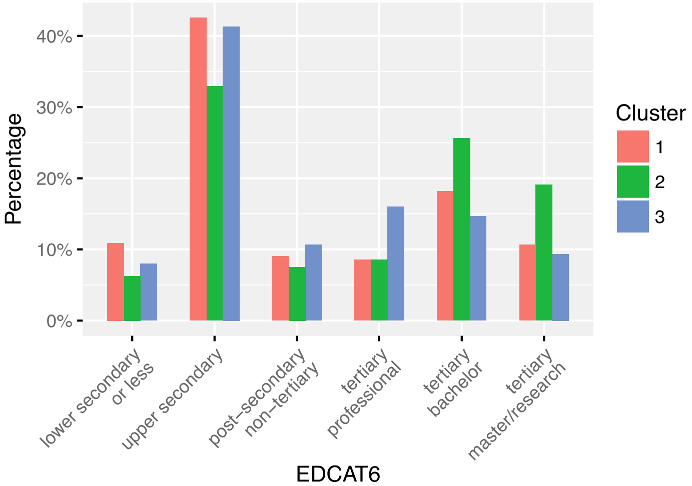


Fig. 10.8 Distribution of EDCAT6 in the three clusters

professional degree levels. Cluster 2 had the greatest proportions of test takers with a bachelor's degree or higher. Thus, test takers in Cluster 1 were the lowest performing group in PSTRE, with the lowest education level overall. Although Cluster 3 provided a slightly higher median PV1, the percentages in the bachelor's degree or higher categories were the lowest. It turned out that test takers in Cluster 3 might not possess the highest education level, but the openness to experience enabled them to score well in PSTRE.

10.4 Discussion

The current study shows an example of clustering students' action sequences and response times, which reveals how these data can provide important information beyond students' item responses. Such information has the potential to help educators understand student motivation and the specific struggles that students might have during a task, and could shed light on the effectiveness of a particular teaching practice.

To summarize, we grouped test takers into three clusters based on 36 features extracted from process data. We found that more actions and longer response time in general were associated with higher PSTRE scores, but such a pattern was more evident when the test takers did not answer U02 correctly. In other words, it was possible to obtain a correct answer with various strategies, but when test takers answered incorrectly, process data could be informative about the extent to which interventions would be needed. In fact, this finding was reiterated when we conducted the same clustering method on other problem-solving items. In the examination of process data patterns with background variables, it was found that test takers who did not put much effort into solving the item tended to work in semi-skilled or elementary occupations, have lower monthly income and lower work-related skill use, be of a higher age, and have lower education. This group of test takers might be in need of further education or intervention.

An interesting finding was that the group with the highest action frequencies, response time, and PSTRE scores did not necessarily possess the highest income, work-related skill use, or education level. The youngest group with longer response time and action sequences was distinct from other test takers in that these individuals were the most explorative or adventurous test takers and were willing to engage in a large number of different actions in solving a problem. This characteristic was likely to relate to higher PSTRE skills.

Besides the merits of this study, some limitations are also worth discussing. First, the response time and number of actions seemed to play a dominant role in the clustering in the current study. It would be worthwhile to try the standardized variables of response time and number of actions in the future study to check whether different results may occur.

Second, the information contributed from the key actions (unigrams) might be difficult to distinguish and not show up very clearly in this clustering analysis. Previous studies have shown that the mini sequences of bigrams and trigrams are more informative than unigrams and were robust classifiers in distinguishing subgroups (He and von Davier 2015). We also extracted bigrams and trigrams for item U02 in this study. However, because of the sparse distribution of action combinations, which resulted in over 40,000 n-grams altogether, it would be very challenging to use this large number of features in its entirety in the cluster analysis. Meanwhile, given the low frequency (lower than five times) of the majority of bigrams and trigrams, we had to exclude them from further cluster analysis to ensure the reliability of calculation. A substantial increase in sample size would help enhance the frequency of mini sequences in future studies in clustering.

Third, we conducted 100 replications with different initial seeds to determine the cluster membership in this study, but need to make cross-validation to examine the clustering performance in further studies. As in this study we mainly explored the relationship between behavioral patterns and proficiency level, a formal classification based on clustering results is not that essential. However, clustering is usually regarded as a “prelude” to classification (Dong and Pei 2007). It would be more appropriate to include a further classification or other means of validation to better evaluate the cluster results.

Fourth, the current study focused on only one PSTRE item. It is not clear yet whether the respondent may choose consistent strategies in solving other items with similar environments. It would be interesting to further examine the consistency of each individual across different items to better generalize the findings from the current study. Some explorations that have been done in this direction may benefit the further analysis in consistency investigation. For instance, He et al. (2019) used the longest common subsequence (LCS) method, a sequence-mining technique commonly used in natural language processing and biostatistics to compare the action sequences followed by PIAAC respondents to a set of "optimal" predefined sequences identified by test developers and subject matter experts, which allows studying problem solving behaviors across multiple assessment items.

Finally, this exploratory study was conducted only based on the U.S. sample in the PSTRE domain of PIAAC. It would be desirable to include multiple countries in a future study to examine the cross-country differences in a general way. Further, it would also be interesting to use process data to explore the relationship between problem-solving skills and numeracy and literacy to better understand the consistency of test takers' behavioral patterns in different domains.

In conclusion, from this study, we have learned that different problem-solving strategies and behavioral patterns may influence proficiency estimates, and are more impactful in the groups that fail to give correct responses than the groups that succeed in answering correctly. Additionally, groups with different backgrounds may show different problem-solving patterns. This suggested that various solutions would need to be properly adapted to different groups to improve their problem-solving skills. In future studies, we recommend researchers further explore the methods to better model the relationship between behavioral patterns and proficiency estimates in large-scale assessments, and challenge other researchers to develop models in estimating problem-solving proficiency more accurately by possibly integrating the new data source from process data.

References

Alpaydin, E. (2009). *Introduction to machine learning*. Cambridge, MA: MIT Press.

Arthur, D., & Vassilvitskii, S. (2007). K–means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on discrete algorithms* (pp. 1027–1035). Society for Industrial and Applied Mathematics.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B, 57,* 289–300.

Bustamam, A., Tasman, H., Yuniarti, N., Frisca, & Mursidah, I. (2017). Application of K-means clustering algorithm in grouping the DNA sequences of hepatitis B virus (HBV). In *AIP Conference Proceedings* (Vol. 1862, No. 1, p. 030134). AIP Publishing.

Chen, Y., Li, X., Liu, J., & Ying, Z. (2019). Statistical analysis of complex problem–solving process data: An event history analysis approach. *Frontiers in Psychology, 10*. https://doi.org/10.3389/fpsyg.2019.00486.

de Klerk, S., Veldkamp, B. P., & Eggen, T. J. (2015). Psychometric analysis of the performance data of simulation-based assessment: A systematic review and a Bayesian network example. *Computers & Education, 85,* 23–34.

Dong, G., & Pei, J. (2007). *Sequence data mining* (Vol. 33). Berlin: Springer Science & Business Media.

Felzenszwalb, P. F., & Huttenlocher, D. P. (2004). Efficient graph-based image segmentation. *International Journal of Computer Vision, 59*(2), 167–181.

Ferreira, D., Zacarias, M., Malheiros, M., & Ferreira, P. (2007). Approaching process mining with sequence clustering: Experiments and findings. In *International conference on business process management* (pp. 360–374). Berlin, Germany: Springer.

Gasch, A. P., & Eisen, M. B. (2002). Exploring the conditional coregulation of yeast gene expression through fuzzy K-means clustering. *Genome Biology, 3*(11), research0059-1.

Goldhammer, F., Naumann, J., Stelter, A., Tóth, K., Rölke, H., & Klieme, E. (2014). The time on task effect in reading and problem solving is moderated by task difficulty and skill: Insights from a computer-based large-scale assessment. *Journal of Educational Psychology, 106*(3), 608.

He, Q., Borgonovi, F., & Paccagnella, M. (2019, forthcoming). *Using process data to understand adults' problem-solving behaviours in PIAAC: Identifying generalised patterns across multiple tasks with sequence mining. OECD Research Paper.*

He, Q., & von Davier, M. (2015). Identifying feature sequences from process data in problem-solving items with n-grams. In A. van der Ark, D. Bolt, S. Chow, J. Douglas, & W. Wang (Eds.), *Quantitative psychology research: Proceedings of the 79th annual meeting of the psychometric society* (pp. 173–190). New York, NY: Springer.

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

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

He, Q., Veldkamp, B. P., & de Vries, T. (2012). Screening for posttraumatic stress disorder using verbal features in self narratives: A text mining approach. *Psychiatry Research, 198*(3), 441–447.

He, Q., Veldkamp, B. P., Glas, C. A. W., & de Vries, T. (2017). Automated assessment of patients' self-narratives for posttraumatic stress disorder screening using natural language processing and text mining. *Assessment, 24*(2), 157–172.

Huang, Y., Niu, B., Gao, Y., Fu, L., & Li, W. (2010). CD-HIT Suite: A web server for clustering and comparing biological sequences. *Bioinformatics, 26*(5), 680–682.

Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika, 32*(3), 241–254.

Jyoti, K., & Singh, S. (2011). Data clustering approach to industrial process monitoring, fault detection and isolation. *International Journal of Computer Applications, 17*(2), 41–45.

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data: An introduction to cluster analysis*. Hoboken, NJ: John Wiley and Sons.

Kawaji, H., Yamaguchi, Y., Matsuda, H., & Hashimoto, A. (2001). A graph-based clustering method for a large set of sequences using a graph partitioning algorithm. *Genome Informatics, 12,* 93–102.

Ketchen, D. J., & Shook, C. L. (1996). The application of cluster analysis in strategic management research: An analysis and critique. *Strategic Management Journal, 17*(6), 441–458.

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

Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory, 28*(2), 129–137.

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

MacNaughton-Smith, P., Williams, W. T., Dale, M. B., & Mockett, L. G. (1964). Dissimilarity analysis: A new technique of hierarchical sub-division. *Nature, 202*(4936), 1034.

Navarro, J. F., Frenk, C. S., & White, S. D. (1997). A universal density profile from hierarchical clustering. *Astrophysical Journal, 490*(2), 493.

Organization for Economic Co-operation and Development. (2009). *PIAAC problem solving in technology-rich environments: A conceptual framework* (OECD Education Working Paper No. 36). Paris, France: Author.

Organisation for Economic Co-operation and Development. (2010). *New millennium learners project: Challenging our views on ICT and learning*. Paris, France: Author.

Organisation for Economic Co-operation and Development. (2011). *PISA 2009 results: Students on line: Digital technologies and performance* (Vol. VI.) http://dx.doi.org/10.1787/9789264112995-en.

Organisation for Economic Co-operation and Development. (2012). *Survey of adult skills (PIAAC)*. Available at http://www.oecd.org/skills/piaac/.

Organisation for Economic Co-operation and Development. (2013a). *Technical report of the survey of adult skills (PIAAC)*. Retrieved from http://www.oecd.org/skills/piaac/_technical%20report_17oct13.pdf.

Organisation for Economic Co-operation and Development. (2013b). *Time for the U.S. to reskill?* Paris, France: OECD Publishing. https://doi.org/10.1787/9789264204904-en.

Organisation for Economic Co-operation and Development. (2016). *Skills matter: Further results from the survey of adult skills*. http://dx.doi.org/10.1787/9789264258051-en. https://www.oecd.org/skills/piaac/Skills_Matter_Further_Results_from_the_Survey_of_Adult_Skills.pdf.

Park, S., Suresh, N. C., & Jeong, B. K. (2008). Sequence-based clustering for web usage mining: A new experimental framework and ANN-enhanced K-means algorithm. *Data & Knowledge Engineering, 65*(3), 512–543.

Rampey, B. D., Finnegan, R., Goodman, M., Mohadjer, L., Krenzke, T., Hogan, J., & Provasnik, S. (2016). *Skills of U.S. unemployed, young, and older adults in sharper focus: Results from the program for the international assessment of adult competencies (PIAAC) 2012/2014: First look* (NCES Report No. 2016–039). U.S. Department of Education. Washington, DC: National Center for Education Statistics. Retrieved from https://nces.ed.gov/pubs2016/2016039.pdf.

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

Spärck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation, 28*(1), 11–21.

Steinbach, M., Ertöz, L., & Kumar, V. (2004). The challenges of clustering high dimensional data. *New directions in statistical physics* (pp. 273–309). Berlin, Germany: Springer.

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

Thorndike, R. L. (1953). Who belongs in the family? *Psychometrika, 18*(4), 267–276.

Vanek, J. (2017). *Using the PIAAC framework for problem solving in technology-rich environments to guide instruction: An introduction for adult educators*. Retrieved from https://piaac.squarespace.com/s/PSTRE_Guide_Vanek_2017.pdf.

Vendlinski, T., & Stevens, R. (2002). Assessing student problem-solving skills with complex computer-based tasks. *Journal of Technology, Learning and Assessment, 1*(3).

Ward, J. H., Jr. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association, 58*(301), 236–244.

Wollack, J. A., Cohen, A. S., & Wells, C. S. (2003). A method for maintaining scale stability in the presence of test speededness. *Journal of Educational Measurement, 40*(4), 307–330.