Commissioned Paper
January 2020

Using Log Files to Identify Sequential Patterns in PIAAC Problem Solving Environments by U.S. Adults' Employment-Related Variables

Authors:
Dandan Liao, Qiwei He, and Hong Jiao

Suggested Citation: Liao, D., He, Q., and Jiao, H. (2020). *Using Log Files to Identify Sequential Patterns in PIAAC Problem Solving Environments by U.S. Adults' Employment-Related Variables*. Retrieved [insert date] from PIAAC Gateway website: [insert link]. Washington, DC.

This project has been funded by the American Institutes for Research through a contract with the National Center for Education Statistics (NCES) of the U.S. Department of Education. This report is based on 2012 PIAAC data released in March 2016. The views expressed in this paper do not necessarily reflect the views or policies of the American Institutes for Research, National Center for Education Statistics, or the U.S. Department of Education, nor does mention of trade names, commercial products, or organizations imply endorsement of same by the U.S. Government.

AIR-PIAAC Contact:

Jaleh Soroui (AIR-PIAAC Director)
Saida Mamedova (Senior Research Analyst)
PIAACgateway.com
piaac@air.org

Author Contact:

Cambium Learning Group
Dandan Liao at dliao@air.org

Educational Testing Service
Qiwei He at qhe@ets.org

University of Maryland
Hong Jiao at hjiao@umd.edu

Using Log Files to Identify Sequential Patterns in PIAAC Problem-Solving Environments by U.S. Adults' Employment-Related Variables

Dandan Liao, Cambium Learning Group, dliao@air.org

Qiwei He, Educational Testing Service, qhe@ets.org

Hong Jiao, University of Maryland, hjiao@umd.edu

1 Introduction

Adult assessments have evolved to keep pace with the changing nature of adult literacy and learning demands. As the importance of information and communication technologies (ICT) continues to grow, measures of ICT literacy skills, digital reading, and problem solving in technology-rich environments (PSTRE) are increasingly important topics for exploration through computer-based assessment (CBA). The Program for the International Assessment of Adult Competencies (PIAAC) is the first international household survey of adult skills predominantly collected using ICT skills. This international survey is conducted in over 40 countries and measures key cognitive and workplace skills including literacy, numeracy, and problem-solving in technology-rich environments (PSTRE). These skills are not only critical to individual prosperity but are also key drivers of economic growth and societal advancement (Organization for Economic Co-operation and Development [OECD], 2013b, p. 3).

Specifically, the PSTRE assessment focuses on the ability of "using digital technology, communication tools and networks to acquire and evaluate information, communicate with others and perform practical tasks" (OECD, 2012). As digital technology has become an indispensable part of human lives, there is an increasing need for measuring the ability to solve problems in conjunction with basic computer literacy skills. The PSTRE assessment renders it

possible to measure how well adults process, analyze, and address problems for specific goals in a computer-based environment.

According to a recent report published by the National Center for Education Statistics (Rampey et al., 2016), U.S. test takers on average scored lower than test takers from other countries in the PSTRE domain (OECD, 2013b, p. 11). In particular, the U.S. sample had the largest proportion of test takers scoring below Level 1, which is the minimum proficiency level required to complete simple problem-solving tasks in daily life (OECD, 2013b, p. 21).

Some facts about specific subgroups of U.S. test takers are also concerning. Scores for millennials (adults born after 1980 and between ages 16-34) in the U.S. were among the lowest of all participating countries even though over half of them spent 35 hours per week on digital media (Goodman, Sands, & Coley, 2015; OECD, 2013b, p. 21). It was found that 41% of test takers with less than high school education chose to take the paper version of PIAAC, compared to 17% for high school graduates and 5% for those with a college degree or above (OECD, 2013b, p. 21). Further, 30% of those who reported being out of the workforce took the paper-based test as opposed to 14% for adults in the labor force (OECD, 2013b, p. 21), suggesting a correlation between skills required for completing the computerized version of the assessment and employability (Vanek, 2017).

An issue that PIAAC attempts to provide a clear picture for is the match between supply and demand for employment skills (OECD, 2016, p. 3). There has been increasing interest in exploring the relationship between proficiency levels and subgroups by employment-related variables, such as employment status and skills used at work (e.g., OECD, 2016, p. 102–103; Perry, Wiederhold, & Ackermann-Piek, 2016). However, assessment of skills is merely one step toward a more balanced labor market. Knowing which subgroups performed better is a good

starting point, but the processes that gave rise to the final proficiency levels are more informative for providing necessary education.

To bridge the gap between supply and demand and provide targeted intervention, it is important to understand which subgroups performed at a lower level and why. Specifically, how did these test takers arrive at a certain wrong answer, and how did subgroups differ in terms of problem-solving strategies? In this regard, more fine-grained investigation on multiple sources of data is needed, which cannot be easily achieved by utilizing response data alone.

Based on the results from He and von Davier's (2015, 2016) and He et al. (2018), the present study mainly focuses on employment-related variables and the U.S. sample to further identify important factors associated with problem-solving skills. Specifically, three research questions are addressed via exploring the process data from one representative PSTRE item:

(1) What features can we extract from process data discriminating best between subgroups with different employment-related variables?

(2) Clustering test takers based on features extracted from process data, what do test takers in each cluster have in common regarding employment-related variables? And what are the characteristics of the clusters with respect to employment?

(3) Are test takers' response behaviors consistent across items? What are the employment-related characteristics of the test takers that are consistent, versus those who are inconsistent in their response behaviors?

In the first research question, a "top-down" approach is employed. We start with the known groups of different employment-related variables and seek the significant differences regarding PIAAC performance among them. For those test takers with significant differences in

providing a correct answer to a representative PSTRE item, we further investigate their action sequences in this item. This offers insights about what sequential features are associated with those employment-related variables which differ significantly on their correctness to this item.

For the second research question, we utilize a "bottom-up" approach that examines the relationship between action features and employment-related variables from another perspective. We start with exploring the clustering structure of process data; in other words, based on action sequences, which test takers are more similar to each other than to those in other clusters. Then, the characteristics of the clusters are explored in terms of employment-related variables. Compared to the first research question where the sequential features of specific employment-related variables are investigated, the second question focuses on the employment profile of a group of test takers with similar sequential features.

The first two research questions both employ one representative PSTRE item. The last research question aims at the consistency of the features from process data across items, and the association between the consistency across items and employment-related variables. To get more accurate population estimates, we took survey weights into account in the calculations for all three research questions. In addition, we conducted the same analyses without survey weights as well. Only marginal differences were found between the two conditions with and without survey weights. Therefore, we reported results with survey weights only in the present study.

The following sections are organized as follows. In section 2, the datasets and the items used in the present study are introduced. In sections 3 to 5, the methods and results for each of the three research questions are demonstrated respectively. Lastly, findings and implications of the present study are discussed, shedding light on future research directions.

2 Dataset and Instruments

The current study utilized data from U.S. test takers in the main study of PIAAC 2012 (Round 1), which is a representative sample of U.S. adults ages 16 to 65. Specifically, this study focused on those test takers who took the CBA, where both responses and process data were collected for PSTRE tasks from the cognitive assessment. In addition, variables from the background questionnaires, with a focus on those related to employment and work experience[1] as well as some derived variables from the background questionnaire,[2] were used to explore the relationship between patterns extracted from process data and test takers' employment situation. Since the present study intended to explore the relationship between employment-related variables and the performance on PSTRE items, only variables related to skill use or experiences at work were included. Variables measuring skill use at home, such as ICT and numeracy skill use at home, were not considered, as work-related background variables have stronger connections to employment situation (OECD, 2016). Each of the skill use at work variables was derived from a few related original variables collected in the background questionnaire. A weighted likelihood estimate (WLE) was first obtained from these original variables, then classified into six categories. The first category included test takers who had all zero responses to the original variables, and the other five were quintiles of the WLE from low to high.

The present study used two datasets, the public-use background and cognitive response data and log file from the PIAAC 2012 study with a focus on the PSTRE items. The former dataset contains rich information about the original and derived variables from the background

[1] Relevant questions include: [D_Q13b] "How often does your job involve learning-by-doing from the tasks you perform?" and [D_Q13c] "How often does your job involve keeping up-to-date with new products or service?"
[2] Relevant derived variables include: index of use of ICT skills at home/at work, and index of learning at work.

questionnaire, cognitive response data, as well as survey weights. The original and derived background variables cover a wide spectrum of the test takers' background profile, including education, daily activities, work history, family background, income, and so on. Additionally, scored responses, total item time, timing of first action, and number of actions are available for each item in the three domains. For PSTRE items, scored responses may be polytomous or dichotomous, depending on the scoring rubrics. For each of the three domains, 10 estimated scores were provided for each test taker, which are referred to as plausible values (PVs). These PVs were estimated using item response theory (IRT) models and population models, taking the information from item responses and background variables into account simultaneously. To ensure the representativeness of the sample, survey weights were given to each test taker, i.e., a full sample weight and a number of replicate weights. For the U.S. sample, each test taker received 45 replicate weights estimated using the paired jackknife method (OECD, 2013a). These survey weights were utilized in the present study for all three research questions. The total sample size of background and cognitive response data was 5010.[3] The descriptive statistics of age, gender, and education for the target population in PIAAC were reported in Tables 1 to 3. As one can see, the age and gender of test takers are relatively evenly distributed among different levels. Based on the weighted frequency, about 42% of the test takers obtained post-secondary education, about 39% had upper secondary education, and 14% had lower secondary education or less.

[3] For the regression analyses in research question 1, all test-takers with valid PSTRE/literacy/numeracy scores and background variables were used in the analyses. There are two reasons why we do this. First is to retain the generalizability of the results to the whole U.S. population. Second, comparing the significant variables among PSTRE, literacy and numeracy enabled us to explore the uniqueness of PSTRE skills. In the later phases when we digged deeper into features from process data, we only used the sample that respond to the PSTRE assessment.

Table 1. Descriptive statistics of age for all test takers in the background and cognitive response data.

Level	Age	Frequency (%)	Weighted Frequency (%)	Standard Error of Weighted Frequency (Standard Error of Weighted Percentage)
1	24 or less	837 (16.71%)	37854352 (18.63%)	13502.76 (<0.01%)
2	25-34	1045 (20.86%)	41123724 (20.24%)	15211.24 (<0.01%)
3	35-44	978 (19.52%)	40739921 (20.05%)	12347.05 (<0.01%)
4	45-54	1084 (21.64%)	44304333 (21.81%)	14171.42 (<0.01%)
5	55 plus	1066 (21.28%)	39122045 (19.26%)	10219.47 (<0.01%)

Table 2. Descriptive statistics of gender for all test takers in the background and cognitive response data.

Level	Gender	Frequency (%)	Weighted Frequency (%)	Standard Error of Weighted Frequency (Standard Error of Weighted Percentage)
1	Male	2323 (46.37%)	99835569 (49.15%)	0.15 (<0.01%)
2	Female	2687 (53.63%)	103308805 (50.85%)	0.15 (<0.01%)

Table 3. Descriptive statistics of education level for all test takers in the background and cognitive response data.

Level	Education	Frequency (%)	Weighted Frequency (%)	Standard Error of Weighted Frequency (Standard Error of Weighted Percentage)
1	Lower secondary or less (ISCED 1,2, 3C short or less)	629 (12.55%)	28671194 (14.11%)	14661.53 (<0.01%)
2	Upper secondary (ISCED 3A-B, C long)	1977 (39.46%)	79830729 (39.30%)	23109.60 (0.01%)
3	Post-secondary, non-tertiary (ISCED 4A-B-C)	394 (7.86%)	16832923 (8.29%)	18868.59 (<0.01%)
4	Post-secondary, tertiary – professional degree (ISCED 5B)	414 (8.26%)	17347913 (8.54%)	19915.13 (<0.01%)
5	Post-secondary, tertiary – bachelor's degree (ISCED 5A)	902 (18.00%)	31595131 (15.55%)	20473.23 (0.01%)
6	Post-secondary, tertiary – master/research degree (ISCED 5A/6)	578 (11.54%)	20186736 (9.94%)	19706.35 (<0.01%)
7	Missing	116 (2.32%)	8679748 (4.27%)	25903.30 (0.01%)

The PSTRE assessment in the PIAAC 2012 study included 14 items, with 7 items in each of the two booklets. PSTRE process data from 2014 test takers were collected in the log files, including actions they took during the assessment, such as sorting, clicking menu, opening a folder, using the help function, and so on. Test takers who responded to the PSTRE items must have some prior computer experience and passed the first two stages of core computer-based assessment. Then, they were selected with 0.3333333 probability to take one PSTRE booklet as the first module. Those who got a PSTRE booklet had 0.5 probability to take another booklet as the second module; otherwise, they had 0.25 probability of taking a PSTRE booklet (OECD, 2016). With this design, 1340 out of 2014 test takers took the seven items in booklet PS1, 1342 took the seven items in booklet PS2, and 668 test takers took both booklets. The PSTRE items were generally designed in three different environments – email, web, spreadsheet, and word processor; each item involves one or two environments as summarized in Table 4. Readers are referred to Figure 2.5 in OECD (2016) regarding the design of PIAAC computer-based assessment and the percentages of test takers taking different pathways.

In the present study, item U02 was chosen as a representative item of the PSTRE domain for the first two research questions. We selected item U02 as the exemplar in this study based on criteria from both content and psychometric perspectives. From the content perspective, U02 is a PSTRE item with multiple environments (email, web, and word processor), which tends to have more diverse sequences of actions where more information can be extracted. More importantly, U02 shared multiple common environments with almost all other items in booklet PS2 (see Table 4). Sharing of the common environments by U02 provides the possibility to investigate the consistency of problem solving strategies across items by individuals, which is the focus of our third research question. From the psychometric perspective, it is a rather difficult item for the

U.S. test takers. In the U.S., 932 (70%) test takers got no credit, 294 (21%) got partial credit, and only 114 (9%) got full credit for the item U02[4]. Test takers are asked to read through a list of emails of meeting room requests in the email environment, and then try to fill out as many requests as possible in the room reservation system in a web environment. Difficult items like U02 often require more steps to complete the item successfully, thus providing more information to differentiate among test takers having different scores. Further, item U02 is located in the middle of the booklet, implying less position effect compared with items that are located at the beginning or end of the booklet.

Table 4. Summary of environments in each item.

Booklet	Order	Item	Email (MC)	Web (WB)	Spreadsheet (SS)	Word Processor (WP)
PS1	1	U01a	1			
PS1	2	U01b	1			
PS1	3	U03a		1	1	
PS1	4	U06a		1		
PS1	5	U06b		1		
PS1	6	U21		1		
PS1	7	U04a	1		1	
PS2	1	U19a	1		1	
PS2	2	U19b			1	1
PS2	3	U07		1		
PS2	4	U02	1	1		1
PS2	5	U16	1			
PS2	6	U11b	1			
PS2	7	U23	1	1		

The sample size and description of the analytical samples are summarized in Table 5. For test takers who did not complete any PSTRE booklets, PVs are estimated based on the population model, which is a combination of an IRT model and a latent regression model. The

[4] Percentages in the parentheses are weighted percentages.

latent regression model incorporated responses to the cognitive items and the background questionnaire (OECD, 2013a, Chapter 17). For research question 1, regression results from different outcome variables are compared to explore the uniqueness of PSTRE skill, robust action features for various employment-related variables are identified in section 3. Section 4 illustrates on cluster characteristics when test takers are clustered based on their action features. Based on the exploration of U02 log files in the first two research questions, the consistency of test takers' behavioral patterns is investigated in section 5, using two items from booklet PS2.

Table 5. Analytical sample size for each study.

Research Question	Analysis	Sample Size	Sample Description
1	Regression	4103	Test takers with valid PSTRE scores
1	Regression	4898	Test takers with valid literacy scores
1	Regression	4898	Test takers with valid numeracy scores
1	Regression	1340	Test takers with valid response time on U02
1	Regression	1340	Test takers with valid responses on U02
1	Feature identification	1326	Test takers with valid responses and log file entries on U02
2	Clustering	1326	Test takers with valid responses and log file entries on U02
3	Consistency	1340	Test takers with valid number of actions and response time on both U02 and U19a

3 Research Question 1: Extract Features from Process Data by Different Employment-Related Variables

The purposes of the first study were to (1) identify employment-related background variables that significantly impact the performance on PIAAC, and (2) extract action sequence features by subgroups categorized by the significant variables for the correctness of U02. We conducted this study in two phases: (1) regression analyses to select background variables that make significant impacts on general PSTRE, literacy, and numeracy skills, as well as the

response time and correctness in the example item; and (2) identify typical action sequences by different subgroups using chi-square feature selection model to explore the action sequences to differentiate the subgroups.

3.1 Select significant employment-related background variables by regression analyses

Regression analyses were conducted to examine which employment-related variables have significant impact on both person-related and item-related outcome variables. The following variables were selected as the independent variables in regression analyses from the public-use background questionnaire data file, including 20 employment-related variables and 6 demographic variables (see Table 6). The employment-related variables reflect different perspectives of the test taker's employment situation, such as employed or not, whether the test taker has a supervisor role, related work experience, computer use at work, and so on. These variables were chosen based on existing research indicating that they might potentially impact the outcome variables (e.g., Bainbridge & Lasley, 2002; Croizet & Dutrévis, 2004; Currie & Thomas, 1999; Dreher & Ryan, 2000; OECD, 2016). The demographic variables include age, gender, the test taker's education level, the test taker's parents' education level, whether the assessment was given in the test taker's native language, and the number of books at home. Table 6 summarizes the description, number of non-missing categories, and the reference category for each variable. To avoid dramatic decrease in sample size, missing responses in the selected variables were coded as another distinctive category and retained in the regression analyses. This method of coding missing responses as an extra category was popularized by Cohen and Cohen (1985) as a method for dealing with missing responses in categorical variables. Cohen and Cohen (1985) advocated that this method incorporated all the available

information into the regression analyses. All variables were recoded into dummy-coded variables with the reference category specified in Table 6.

Table 6. Summary of background questionnaire variables used in the present study.

No.	Variable	Description	No. of categories	Reference category
1	PAIDWORK*	incidence of paid work experience	4	1. no work ever
2	C_D05	employment status	3	1. employed
3	D_Q04_T1[a]	status at this job or business	4	1. employee, not supervisor
4	D_Q08b	managing other employees - count	5	1. 1-5 people
5	D_Q12b	education to do the job satisfactorily	3	1. necessary
6	D_Q12c_RC[b]	related work experience in years	4	1. no experience
7	F_Q05a	incidence of solving simple problems	5	1. never
8	F_Q05b	incidence of solving complex problems	5	1. never
9	F_Q07b	need more training for skill use at work	2	1. yes
10	G_Q06	level of computer use	3	1. straightforward
11	ISCOSKIL4	occupational classification	4	1. skilled occupations
12	EARNMTHALLDCL	monthly earning decile including all incomes	10	1. lowest decile
13	LEARNATWORK_WLE_CA	index of learning at work	6	0. all zero response
14	ICTWORK_WLE_CA	index of use of ICT skills at work	6	0. all zero response
15	INFLUENCE_WLE_CA	index of use of influencing skills at work	6	0. all zero response
16	NUMWORK_WLE_CA	index of use of numeracy skills at work	6	0. all zero response
17	PLANNING_WLE_CA	index of use of planning skills at work	6	0. all zero response
18	READWORK_WLE_CA	index of use of reading skills at work	6	0. all zero response
19	TASKDISC_WLE_CA	index of use of task discretion at work	6	0. all zero response
20	WRITWORK_WLE_CA	index of use of writing skills at work	6	0. all zero response
21	AGEG10LFS	age in 10-year bands	5	1. 24 or less

22	PARED	highest of parents' level of education	3	1. neither parent have attained upper secondary
23	NATIVELANG	test language same as native language	2	0. test language is not native language
24	EDCAT6	highest level of formal education obtained	6	1. lower secondary or less
25	GENDER_R	gender	2	1. male
26	J_Q08	number of books	6	1. 10 books or less

[a] D_Q04_T1 provides information about whether the test taker is an employee, supervisor, self-employed, or unpaid. This variable was chosen since research has shown that the relationship between some variables are different for wage earners and self-employed workers, such as earnings and conscientiousness (OECD, 2016).
[b] PAIDWORK is recoded from NOPAIDWORKEVER (never had paid work), PAIDWORK5 (have had paid work in the last 5 years) and PAIDWORK12 (have had paid work in the last 12 months). D_Q12c_RC is recoded from D_Q12c (related work experience in years) by collapsing category 2-4.

With respect to the dependent variables, three, person-related and two, item-related variables were chosen in the present study. The three person-related dependent variables were test takers' scores on PSTRE, literacy, and numeracy, reflecting test takers' overall performance level in different domains. For item-related outcomes, we used total item response time and dichotomized scores on item U02. Sample size and descriptive statistics of the five outcome variables are reported in Table 7.

Table 7. Sample size and descriptive statistics of the outcome variables.

Variable Name	Sample Size	Minimum	Maximum	Mean	Standard Deviation
PSTRE	4103	113.56	425.01	277.98	43.11
Literacy	4898	100.94	424.33	271.84	48.28
Numeracy	4898	40.33	426.90	254.68	55.80
U02RT*	1340	0.09	45.07	3.59	3.48
U02score	1340	0	1	0.30	0.21

*U02RT is reported in minutes.

The distributions of the background variables used in the present study were checked to ensure the representativeness of this sample. The difference between the percentages of each category of the background variables from the sample with valid U02 scores and the total sample

was usually around 1-2% (see Table A1 in the Appendix for details). As such, we deemed that the differences were not substantially different. In the PIAAC 2012, item U02 was polytomously scored with 4 categories, where 0 indicates an incorrect response, 1 and 2 represent partially correct responses, and 3 is given to a fully correct response. A total of 1340 test takers had responses on this item. However, 932 out of 1340 test takers (70%) got no credit (score of 0) for this item, meaning that this is a very difficult item. Since the majority of test takers got no credit for the item, we chose to dichotomize the observed scores by collapsing partial and full credits (score categories 1 to 3) into one. In the following analyses, scores 1-3 were referred to as correct, while a score of 0 was referred to as incorrect.

Of the five outcome variables, linear regression was conducted for the four continuous outcome variables – PSTRE, literacy, and numeracy scores, as well as the item response time on U02. For the dichotomized scores on U02, a logistic regression was carried out. The regression analyses were conducted using International Association for the Evaluation of Educational Achievement (IEA)'s International Database (IDB) Analyzer (IEA, 2013) to interface with SAS 9.4 (SAS Institute, 2015). In the present study, each regression analysis was carried out using a full sample weight and 45 replicate weights, as well as 10 PVs if the outcome variable is the score from one of the three domains. The significance of the predictors was checked based on the final regression results.

3.2 Identify typical action sequences by subgroups with significant predictors

In phase 2, process data were used to understand the inherent differences among the groups of test takers obtained in phase 1. First, background and cognitive assessment data was merged with process data, and subgroups were formed based on the significant employment-

related variables. The significant level of the predictor was compared with the reference level of the predictor. For instance, two subgroups were formed based on the significant variable EARNMTHALLDCL_RC_D4: test takers in the lowest decile and test takers in the 4th decile. Then, we recoded adjacent action sequences into n-grams and employed chi-square selection model to identify robust action sequences from process data for different subgroups. The package "tm" (Feinerer, 2017) in R version 3.3.3 (R Core Team, 2017) was utilized for applying chi-square selection model to identify robust features.

An n-gram is defined as a contiguous sequence of n words in text mining; similarly, when analyzing action sequences from process data, a n-gram can be defined as a sequence of n adjacent actions. In the present study, we focused on unigrams, bigrams, and trigrams, which are adjacent action sequences of length 1, 2, and 3, respectively. Chi-square feature selection model is recommended for use in textual analysis due to its high effectiveness in finding robust keywords and for testing the similarity between different text corpora (e.g., He, Glas, Kosinski, Stillwell & Veldkamp, 2012, 2014; He, Veldkamp, & de Vries, 2012; Manning & Schütze, 1999). The definition of "robust" is different from what is defined in statistics; here robust features are generally defined as the "best" features with high information gain in NLP (Joachims, 1998).

Chi-square scores assigned to the features were ranked in descending order and those with the highest scores are defined as robust features. Specifically, frequencies and weights of certain actions[5] for different employment-related variables were used as input for the chi-square selection model. Weights were applied to frequencies of action sequences to adjust for between-

[5] Weights of certain actions indicates the term weights, instead of sample weights for the population. The terms weights depend on the rareness of certain actions or action combinations in the action sequence data pool.

individual and within-individual differences in action frequencies. Between-individual weights highlight how different the frequency of a certain action is among individuals, whereas within-individual weights capture the fact that some actions are used more often than others within an individual.

In terms of 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.

On the other hand, within-individual differences occur when an individual takes some actions more often than others. Although more frequent sequences are more important than less frequent sequences, the raw frequencies of these action sequences often overestimate their importance (He & von Davier, 2015). To account for within-individual differences in the importance of action sequences, a weighting function was employed $f(tf_{ij}) = 1 + \log(tf_{ij})$, where $tf_{ij} > 0$ represents the frequency of action i in sequence j. Combining the between- and within-individual weights, the final action weight can be defined as $\text{weight}(i, j) = [1 + \log(tf_{ij})] \log(N/sf_i)$ for $tf_{ij} \geq 1$. 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.

3.3 Results

3.3.1 Phase 1: regression analyses

The significant predictors produced by the regression analyses in phase 1 are summarized in Table 8, with respect to each of the five outcome variables. A cell of 1 indicates at least one of the dummy variables are significantly different from 0 at the significance level of 0.05, whereas a blank cell means none of the dummy variables are significant. The coefficient estimates for the significant variables are presented in Table 9. Note that the estimates for U02 score are the exponentials of the original estimates from the logistic regression.

In general, all 5 outcome variables had one significant variable in common, EDCAT6, which means that educational level is important for obtaining high scores in all three domains and on individual item responses, and it also contributes to longer item response time[6] on U02. Among the three, person-related dependent variables, more predictors were significant in predicting literacy and numeracy scores when compared with PSTRE scores. The significant variables for literacy and numeracy scores were more similar, though the three domains had 13 significant variables in common. D_Q12c_RC, the related work experience in years, and GENDER_R were significant in predicting PSTRE and numeracy but not literacy, whereas WRITWORK_WLE_CA was only important for literacy scores. As the focus of this study, PSTRE scores had one unique significant variable – READWORK_WLE_CA (more than 80% reading skill use at work). The intriguing finding was that this variable was negatively related to PSTRE scores. It is possible that

[6] It is often found that item location could affect response time distribution (e.g., Wollack, Cohen, & Wells, 2003). Items located at the beginning of the assessment tend to require more time for test takers to get familiar with the assessment structure, whereas response time for items locate at the end of the assessment may be deflated because some test takers may not have reached those items. Since U02 is located in the middle of booklet PS2, which is administered as a fixed form, item location should not be a concern for U02.

when controlling for all other variables, test takers who used a lot of reading skills at work might not need much PSTRE skills at work due to the nature of the job.

Only five variables were significant in predicting the response time on U02. The regression coefficient estimates show that test takers who are well educated, have higher levels of computer use, use more numeracy and planning skill use at work, and whose parents also obtained higher education degrees tend to spend more time on item U02. Similarly, U02 score did not have as many significant variables as the person-related outcomes either (i.e., PSTRE, literacy, numeracy scores), where only 8 variables were significant. Interestingly, not all of these variables were significant in predicting PSTRE scores. This might be because PSTRE scores are holistic measures of the PSTRE skills, which represent the common construct assessed by the 14 PSTRE items. While U02 contributes to the PSTRE scores as part of the assessment, it does not necessarily contain all aspects of the construct.

In terms of the coefficient estimates, most of them were consistent with our expectations. With respect to employment-related variables, test takers who have had paid work, have more related work experience, solve simple or complex problems more frequently, have higher level of computer use, have skilled occupations and higher monthly income, and have higher index variables tend to have higher scores in the three domains and higher odds of success in U02. For the demographic variables, we would expect younger male test takers who are well educated and have a lot of books at home to get higher scores when English was their native language (same language as the test).

However, some coefficient estimates are inconsistent with our expectations (highlighted in grey in Table 9). For example, we would expect test takers with more related work experience to perform better in general, but the estimates for experience less than a year are negative for

PSTRE and numeracy scores, which indicates that controlling for all other variables, having a short work experience is not better than having no experience for these two outcomes. For F_Q05b, coefficients for literacy and numeracy scores are also negative when comparing the highest category to the lowest, reflecting that a test taker who solves complex problems everyday might get a score lower than a test taker who never solves complex problems at work. These contradictory results may indicate some interactions among the predictors.

It is also possible that there might be some confounding variables. For instance, test takers with less than 1 year of relevant work experience had lower PSTRE scores than those without relevant experience. While in general we expect PSTRE scores to increase with more relevant experience, test takers with no experience might be more motivated to learn and apply PSTRE skills at work. However, such variables were not available from the PIAAC background questionnaire.

In the next phase where how the action sequences of the two groups differed from each other, only the significant predictors for U02score were used. For instance, the regression coefficient for the 4th decile of EARNMTHALLDCL was significant and positive, indicating that the test takers with monthly income in the 4th decile are significantly more likely to get a score of 1 than those in the 1st (lowest) decile. As such, it is of interest to investigate how the test takers with monthly income in the 4th and 1st deciles differ regarding their action sequences. In other words, what features do the two groups of test takers have in their test-taking behaviors that were associated with higher or lower chances of answering correctly to item U02.

As demonstrated in section 3.2, we conducted chi-square selection to identify the most distinguishable n-grams between the two groups. Specifically, the top five unigrams, bigrams, and trigrams with the highest chi-square scores were obtained for the focal group and the reference

group, respectively. Table 10 presents the description and frequency of the 34 unigrams used in the present study; bigrams and trigrams are all possible combinations of unigrams. These robust features were used to understand the most distinctive action sequences between the two groups of test takers. The same procedure was carried out for all significant predictors for U02score. We only demonstrated two such predictors as examples (see Tables 11 and 12); the robust features for all other predictors are reported in Tables A2 to A12 in the Appendix.

Table 8. Summary of significant predictors.

No.	Variable	Description	PSTRE	Literacy	Numeracy	U02RT	U02score
1	PAIDWORK	incidence of paid work experience		1	1		
2	C_D05	employed/unemployed/out of labor force					
3	D_Q04_T1	is an employee/supervisor/ self-employed/unpaid					
4	D_Q08b	managing how many employees					
5	D_Q12b	education level sufficient/too high/too low to do job satisfactorily	1	1	1		
6	D_Q12c_RC	related work experience in years	1		1		
7	F_Q05a	incidence of solving simple problems	1	1	1		
8	F_Q05b	incidence of solving complex problems		1	1		
9	F_Q07b	need more training for skill use at work or not		1	1		
10	G_Q06	level of computer use	1	1	1	1	
11	ISCOSKIL4	skilled/semi-skilled/elementary occupations	1	1	1		
12	EARNMTHALLDCL	monthly earning decile including all incomes	1	1	1		1
13	LEARNATWORK_WLE_CA	index of learning at work	1	1	1		
14	ICTWORK_WLE_CA	index of use of ICT skills at work	1	1	1		
15	INFLUENCE_WLE_CA	index of use of influencing skills at work		1	1		1
16	NUMWORK_WLE_CA	index of use of numeracy skills at work	1	1	1	1	
17	PLANNING_WLE_CA	index of use of planning skills at work		1	1	1	
18	READWORK_WLE_CA	index of use of reading skills at work	1				1
19	TASKDISC_WLE_CA	index of use of task discretion at work					1
20	WRITWORK_WLE_CA	index of use of writing skills at work		1			1
21	AGEG10LFS	age in 10-year bands	1	1	1		1
22	PARED	highest of parents' level of education	1	1	1	1	
23	NATIVELANG	test language same as native language or not	1	1	1		1
24	EDCAT6	highest level of formal education obtained	1	1	1	1	1
25	GENDER_R	gender	1		1		
26	J_Q08	number of books	1	1	1		

Note. A cell of one indicated that at least one level of this variable was significantly different from zero.

Table 9. Summary of unstandardized regression coefficients of significant variables.

Variable	Description	PSTRE	Literacy	Numeracy	U02RT	U02score*
Intercept	intercept	214.93	162.42	126.44	6.34	
PAIDWORK_D3	have had paid work in 5 years but not 12 months		22.37	30.59		
PAIDWORK_D4	have had paid work in 12 months			37.51		
D_Q12b_D2	a lower education level would be sufficient	5.22	3.76			
D_Q12b_D3	a higher education level would be needed	-8.17	-8.38	-9.35		
D_Q12c_RC_D2	less than 1 year of relevant work experience	-5.07		-5.49		
D_Q12c_RC_D4	more than 3 years of relevant work experience			6.76		
F_Q05a_D2	solve simple problems less than once a month		9.13			
F_Q05a_D3	solve simple problems less than once a week but at least once a month	13.53	10.91	17.92		
F_Q05a_D4	solve simple problems at least once a week but not everyday	12.58	13.80	17.11		
F_Q05a_D5	solve simple problems everyday	16.47	17.71	19.07		
F_Q05b_D5	solve complex problems everyday		-11.01	-9.14		
F_Q07b_D2	do not need more training for skill use at work		6.20	7.28		
G_Q06_D2	moderate level of computer use	9.97	7.23	8.15	0.65	
G_Q06_D3	complex level of computer use	15.44	10.35	15.96	1.25	
ISCOSKIL4_D2	semi-skilled white-collar occupations	-3.83	-4.47	-4.80		
ISCOSKIL4_D3	semi-skilled blue-collar occupations	-7.29	-7.10	-6.24		
ISCOSKIL4_D4	elementary occupations		-13.49	-14.11		
EARNMTHALLDCL_D4	4th decile of monthly earning					2.00
EARNMTHALLDCL_D9	9th decile of monthly earning		7.88			
EARNMTHALLDCL_D10	10th decile of monthly earning	10.15	12.38	11.55		

LEARNATWORK_WLE_CA_D5	more than 60% to 80% on index of learning at work	-7.60	-10.42	-9.74		
ICTWORK_WLE_CA_D2	lowest 20% on index of use of ICT skills at work		10.98	10.05		
ICTWORK_WLE_CA_D3	more than 20% to 40% on index of use of ICT skills at work	12.28	15.23	11.28		
ICTWORK_WLE_CA_D4	more than 40% to 60% on index of use of ICT skills at work	16.34	13.63	12.09		
ICTWORK_WLE_CA_D5	more than 60% to 80% on index of use of ICT skills at work	17.78	11.52	10.29		
ICTWORK_WLE_CA_D6	more than 80% on index of use of ICT skills at work	20.52	15.46	16.00		
INFLUENCE_WLE_CA_D2	lowest 20% on index of use of influencing skills at work					1.63
INFLUENCE_WLE_CA_D6	more than 80% on index of use of influencing skills at work		-10.84	-10.24		
NUMWORK_WLE_CA_D3	more than 20% to 40% on index of use of numeracy skills at work	9.12	5.70	8.76		
NUMWORK_WLE_CA_D4	more than 40% to 60% on index of use of numeracy skills at work			7.10		
NUMWORK_WLE_CA_D5	more than 60% to 80% on index of use of numeracy skills at work	8.51	8.33	10.47	0.84	
NUMWORK_WLE_CA_D6	more than 80% on index of use of numeracy skills at work	6.89		10.26		
PLANNING_WLE_CA_D5	more than 60% to 80% on index of use of planning skills at work		9.33	12.72	0.92	
READWORK_WLE_CA_D2	lowest 20% on index of use of reading skills at work					2.23
READWORK_WLE_CA_D6	more than 80% on index of use of reading skills at work	-14.44				
TASKDISC_WLE_CA_D5	more than 60% to 80% on index of use of task discretion at work					0.43

WRITWORK_WLE_CA_D4	more than 40% to 60% on index of use of writing skills at work		6.85			
WRITWORK_WLE_CA_D5	more than 60% to 80% on index of use of writing skills at work		7.93			1.68
AGEG10LFS_D2	25-34	-17.20	-12.91	-12.10		
AGEG10LFS_D3	35-44	-24.57	-17.21	-16.84		
AGEG10LFS_D4	45-54	-31.89	-22.09	-20.77		0.63
AGEG10LFS_D5	55 plus	-35.85	-23.64	-20.65		0.53
PARED_D2	at least one parent has attained secondary and post-secondary, non-tertiary	10.31	10.15	13.06	0.77	
PARED_D3	at least one parent has attained tertiary	12.33	15.38	15.93	1.28	
NATIVELANG_D2	test language same as native language	13.92	17.41	9.99		1.29
EDCAT6_D2	upper secondary (ISCED 3A-B, C long)	10.09	16.72	21.31	0.87	7.70
EDCAT6_D3	post-secondary, non-tertiary (ISCED 4A-B-C)	14.18	19.91	27.92		7.70
EDCAT6_D4	tertiary – professional degree (ISCED 5B)	17.59	27.84	35.31		5.47
EDCAT6_D5	tertiary – bachelor's degree (ISCED 5A)	24.00	35.74	45.59	1.11	10.16
EDCAT6_D6	tertiary – master/research degree (ISCED 5A/6)	28.60	44.53	55.83	1.33	15.18
GENDER_R_D2	female	-4.19		-12.89		
J_Q08_D2	11 to 25 books		6.54			
J_Q08_D3	26 to 100 books	10.45	9.84	14.52		
J_Q08_D4	101 to 200 books	13.97	13.54	21.49		
J_Q08_D5	201 to 500 books	22.49	20.20	26.62		
J_Q08_D6	More than 500 books	14.13	19.74	23.33		

Note. Coefficients reported in this table for U02score are odds ratios.

Table 10. Description and frequency of unigrams.

No.	Features	Description	Frequency
1	FOLDER_VIEWED	view a folder	5762
2	ENVIRONMENT_WB	go to web environment	4715
3	ENVIRONMENT_MC	go to email environment	4317
4	MAIL_VIEWED_1	view 1st email	2725
5	HISTORY_VIEWCALENDAR	go to calendar tab in web environment	2190
6	MAIL_VIEWED_3	view 3rd email	1968
7	HISTORY_RESERVATION	go to reservation tab in web environment	1935
8	COMBOBOX_ROOM	choose a room when filling out a room request	1891
9	MAIL_VIEWED_4	view 4th email	1698
10	MAIL_VIEWED_2	view 2nd email	1544
11	MAIL_MOVE	move an email	1499
12	NEXT_INQUIRY	go to next item	1371
13	START	start item U02	1326
14	COMBOBOX_START_TIME	choose start time when filling out a room request	1312
15	COMBOBOX_END_TIME	choose end time when filling out a room request	1304
16	COMBOBOX_DEPT	choose department when filling out a room request	1296
17	HISTORY_MEETINGROOMS	go to meeting room details tab in web environment	1058
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
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

3.3.2 Phase 2: feature identification

Table 11 presents the top five unigrams, bigrams, and trigrams for the test takers falling within the fourth and first (lowest) deciles of monthly earning. Among the unigrams, folder-related actions were found more often in the fourth decile group, such as fold, add, or delete a folder. There were a few folders in the email environment, though test takers were not required to perform any actions on them. The fourth decile group also applied more cancel-related actions, such as cancel sorting, cancel changing reservation, cancel switching to the next item, and so on. Though cancel actions are sometimes considered hesitative behaviors (He & von Davier, 2015), they could also indicate that the fourth decile group tried different options in the menu to figure out what could be done in the environment.

Other actions that the fourth decile group frequently used were actions associated with bookmarks, clicking the home button in the web environment, and help functions. The bookmarks were accessible via the dropdown menu or a button on the menu bar. Using the bookmark actions, test takers could easily access the pages that they considered important or useful. The home button was right next to the bookmark button on the menu bar, which is a convenient way to return to the main page of the web environment. The help functions were designed in both email and web environments. In the email environment, the help function provided information regarding actions taken for an email, for instance, write, reply, forward, or delete an email. In the web environment, the help function offered instructions on the menu bar items, such as home and bookmark. These robust unigrams seemed to suggest that the fourth decile group tried different options to figure out what could be done in the environment. As expected, the fourth decile group appeared to take more exploratory actions to facilitate their problem-solving process compared to the first-decile group.

Table 11. Top five features of action sequences selected for the 4th and the 1st deciles of monthly earning groups.

Group	N-gram	Action sequences	Chi-square
4th decile of monthly earning	Unigram	FOLDER	39.08
		CANCEL	16.54
		BOOKMARK	7.44
		HISTORY_HOME	4.02
		HELP	1.84
	Bigram	FOLDER_VIEWED FOLDER	24.06
		FOLDER FOLDER_VIEWED	22.93
		FOLDER FOLDER	22.67
		MAIL_VIEWED_1 MAIL_VIEWED_3	18.51
		NEXT_INQUIRY CANCEL	17.64
	Trigram	FOLDER FOLDER_VIEWED FOLDER_VIEWED	21.87
		MAIL_VIEWED_1 MAIL_VIEWED_3 MAIL_VIEWED_4	19.70
		FOLDER_VIEWED FOLDER FOLDER	18.67
		ENVIRONMENT_MC MAIL_VIEWED_1 MAIL_VIEWED_3	17.25
		MAIL_VIEWED_1 MAIL_VIEWED_3 MAIL_VIEWED_3	16.68
1st decile of monthly earning	Unigram	SEARCH	11.72
		COPY	10.90
		KEYPRESS	10.51
		PASTE	5.89
		HISTORY_VIEWCALENDAR	2.81
	Bigram	ENVIRONMENT_WB COMBOBOX_START_TIME	15.78
		COMBOBOX_END_TIME HISTORY_VIEWCALENDAR	11.20
		HISTORY_RESERVATION HISTORY_VIEWCALENDAR	10.50
		COPY KEYPRESS	10.16
		HISTORY_UNFILLED HISTORY_RESERVATION	9.79
	Trigram	HISTORY_VIEWCALENDAR HISTORY_RESERVATION HISTORY_VIEWCALENDAR	20.16
		MAIL_VIEWED_3 ENVIRONMENT_WB HISTORY_UNFILLED	16.25
		ENVIRONMENT_MC ENVIRONMENT_WB ENVIRONMENT_MC	16.13
		ENVIRONMENT_MC ENVIRONMENT_WB COMBOBOX_START_TIME	15.87
		MAIL_VIEWED_4 ENVIRONMENT_WB ENVIRONMENT_MC	14.88

The unigrams commonly adopted by the first-decile group were entirely different. The most discriminating features included search, copy, keypress (pressing a key on the keyboard),

paste, and click on the view calendar button. The search function was available in both email and web environments. However, the search function was not required to obtain a correct answer to U02, as the information in the two environments was displayed in short text or tables. The copy, keypress, and paste unigrams were used in the word processor environment solely, where test takers could take notes for the time and location of the meeting room requests and compare to the existing schedules. Similar to search, the three functions only existed to aid the synthesis of information and conflict schedules. The word processor environment was designed to assist with collecting and synthesizing conflicting information. For the view calendar button, test takers used it to retrieve the schedules for each meeting room in a certain time period. Test takers were able to see not only the existing reservations, but also the reservations they made for the meeting room requests.

The lower odds of a correct answer to U02 for the first-decile group indicated an association between these functions and lower performance in this group. One explanation for this phenomenon could be that the search function and word processor environment were rather redundant for high-performing test takers since they could collect and synthesize information more efficiently. Applying such functions might be a sign that test takers were having difficulty in comprehending or solving U02. Additionally, the view calendar button seemed to suggest that test takers in this group were still in the process of figuring out the purpose of the item instead of working on solving the problem.

Compared to the unigrams, the robust bigrams and trigrams were often more closely related for a certain group. The bigrams for the fourth decile mainly involved folder-related actions, email-viewing actions, and cancel actions; the trigrams also contained similar information. The bigram “FOLDER, FOLDER_VIEWED” was found in the trigram “FOLDER,

FOLDER_VIEWED, FOLDER_VIEWED"; the bigram "MAIL_VIEWED_1, MAIL_VIEWED_3" were also included in the robust trigram "MAIL_VIEWED_1, MAIL_VIEWED_3, MAIL_VIEWED_4". This is because bigrams with high frequencies were also likely to appear more commonly when started with or followed by another action. Further, while the five robust unigrams tended to provide unique pieces of information, the five bigrams tended to have overlap due to the increase in sequence length, as did the trigrams. For instance, the top three robust bigrams for the fourth decile group were all folder-related actions, whereas three of the top five trigrams were email-viewing actions.

These mini-sequences of the fourth decile group, along with the unigrams, demonstrated evidence that test takers in this group were working on the item and trying to understand the meeting room requests. It is worth noticing that the emails viewed by the fourth decile group were the first, third, and fourth emails (i.e., MAIL_VIEWED_1, MAIL_VIEWED_3, MAIL_VIEWED_4); the second email did not show up in any robust features. In fact, the second email was the only one irrelevant to meeting room requests among the four. Therefore, viewing only the three relevant emails was a strong indication that the test takers at least understood the goal of this item, and were able to filter out emails irrelevant to the goal.

For the first decile group, the test takers did a lot of switching among tabs in the web environment (e.g., HISTORY_VIEWCALENDAR, HISTORY_RESERVATION, HISTORY_UNFILLED), or switching among environments (e.g., ENVIRONMENT_MC, ENVIRONMENT_WB). Such switching actions indicated that the first decile did not devote much to solving the item. Instead, they seemed to be lost in the item or not interested in exploring more. Results based on unigrams, bigrams, trigrams all suggested that compared to the first decile, test takers in the fourth decile group were more engaged in solving the item.

The fourth decile group also adopted more efficient problem-solving strategies, such as bookmark and help. This is consistent with the results from regression analysis that the fourth decile group was more likely to obtain a correct answer to U02 (see Table 9).

Another example is the comparison between the robust features from the highest and lowest education groups, as presented in Table 12. Test takers in the highest education group obtained tertiary-master/research degrees, whereas the lowest education group obtained lower secondary education or less. The chi-square selection method also identified highly distinctive features for the two groups. The most discriminating unigrams for the highest education group were sorting, submitting filled reservation or unfilled request (i.e., SUBMIT_RESERVATION _SUCCESS, UNFILLED_SUBMIT), and filling out the room and the start time for the request (i.e., COMBOBOX_ROOM, COMBOBOX_START_TIME).

The sorting function was available in the email environment. Test takers could choose to sort by sender, subject, or receiver of the email. Although sorting was not a necessary step to the success of U02, well-educated test takers might consider sorting by subject as a more efficient approach to identifying the emails related to meeting room requests. The COMBOBOX-related actions showed evidence of filling out the details of meeting room requests using the dropdown menus. Then, if the requested room and time had no conflict with the existing schedule, one would receive a notice of submitting the reservation successfully. There was also one meeting room request that could not be filled given the current schedule, which needed to be recorded as well. UNFILLED_SUBMIT indicated that the test taker also submitted the unfilled request. Such actions were key to the correctness of U02, because one had to fill out the details of each room request and submit at least one reservation or unfilled request successfully to answer it correctly.

Table 12. Top five features of action sequences selected for the highest and lowest education groups.

Group	N-gram	Action sequences	Chi-square
tertiary – master/research degree	Unigram	SORT	14.04
		SUBMIT_RESERVATION_SUCCESS	7.24
		COMBOBOX_ROOM	6.88
		SUBMIT_UNFILLED	6.82
		COMBOBOX_START_TIME	6.22
	Bigram	COMBOBOX_END_TIME COMBOBOX_DEPT	20.36
		ENVIRONMENT_WB ENVIRONMENT_MC	17.06
		ENVIRONMENT_MC MAIL_VIEWED_1	16.84
		SUBMIT_RESERVATION_SUCCESS HISTORY_UNFILLED	16.73
		HISTORY_MEETINGROOMS ENVIRONMENT_MC	16.57
	Trigram	ENVIRONMENT_MC MAIL_VIEWED_1 MAIL_VIEWED_2	26.36
		HISTORY_RESERVATION HISTORY_MEETINGROOMS HISTORY_VIEWCALENDAR	23.26
		MAIL_VIEWED_1 MAIL_VIEWED_3 MAIL_VIEWED_4	20.73
		COMBOBOX_END_TIME COMBOBOX_DEPT SUBMIT_RESERVATION_SUCCESS	19.96
		COMBOBOX_START_TIME COMBOBOX_END_TIME COMBOBOX_DEPT	19.35
lower secondary or less	Unigram	MAIL_MOVE	197.12
		FOLDER_VIEWED	24.15
		PASTE	9.77
		COPY	7.73
		SEARCH	7.25
	Bigram	FOLDER_VIEWED MAIL_MOVE	159.17
		MAIL_MOVE FOLDER_VIEWED	156.81
		MAIL_VIEWED_3 MAIL_MOVE	104.55
		MAIL_VIEWED_4 MAIL_MOVE	90.67
		MAIL_MOVE MAIL_VIEWED_4	90.10
	Trigram	MAIL_MOVE FOLDER_VIEWED MAIL_MOVE	148.51
		FOLDER_VIEWED MAIL_MOVE FOLDER_VIEWED	95.11
		MAIL_VIEWED_3 MAIL_MOVE FOLDER_VIEWED	92.88
		FOLDER_VIEWED MAIL_MOVE MAIL_VIEWED_3	92.17
		MAIL_VIEWED_4 MAIL_MOVE FOLDER_VIEWED	86.50

The lowest education group, however, mainly used redundant functions. Moving emails, viewing folder, pasting, copying, and searching were the most important unigrams, which coincidently were found as robust unigrams in the first decile monthly earning group as well. Both the lowest education group and first-decile monthly earning group had lower performance on U02 compared with their peers. This finding suggested that lower-performing test takers might be prone to using these unnecessary functions (as defined by content experts), indicating they were unable to figure out a solution.

The robust bigrams and trigrams for the highest education group encompassed some action sequences that also related to filling and submitting the requests, as well as viewing emails, which were required procedures to obtain a correct answer. Some features indicating switching among tabs or environments also appeared. Though we interpreted similar actions for the first-decile group as signs of low motivation, these actions could have different meanings for another group. When combined with other robust features for the highest education group, these actions served as connections among necessary steps to finish the item, such as filling in comboboxes and submitting requests. Therefore, the highest education group did not wander around aimlessly, but in fact attempted to synthesize information from multiple environments and make a successful reservation.

Email-moving and folder-viewing actions manifested themselves again in the robust bigrams and trigrams for the lowest education group. These action sequences identified by chi-square selection method demonstrated a clear distinction between the problem-solving processes of the two groups with different education levels. While the highest education group was completing the item with clear subgoals, the lowest education group spent much time and effort moving the

emails around and viewing the folder. As a result, these discriminating features identified from the action sequences were in fact strongly associated with the performance on the item.

Some general findings from other significant variables resembled the results from the two discussed examples. As presented above, higher income, higher level on the index variables (except for TASKDISC_WLE_CA, index of use of task discretion at work), and higher educational level were associated with higher probability of answering U02 correctly. A younger test taker who took the test in the same language as his or her native language was also more likely to obtain a correct answer. Some background variables have more than one significant dummy variables, such as age and education. It is worth noticing that the features selected for the reference group did not need to be the same when the focal group changed since chi-square chose features that can best distinguish the reference and the focal groups.

Overall, groups with higher odds of a correct answer were likely to adopt the actions related to SUBMIT (submitting filled reservation or unfilled request), COMBOBOX (filling out the room and the start time for the request), help, and sort. Help and sort are two actions that might be indicative of more efficient problem-solving strategies. To complete the room requests in this item, test takers had to fill time slots for a specific room in the COMBOBOX and use one of the two submit buttons. These test takers demonstrated evidence that they went through the necessary steps to obtain correct answers to U02.

Groups with lower odds of a correct answer, however, used more actions such as MAIL_MOVE (moving email) and SUBMIT_RESERVATION_FAILURE (failure to submit a room request). The occurrence of MAIL_MOVE and SUBMIT_RESERVATION_FAILURE did not always mean that a test taker had trouble finishing an item. A test taker could have been categorizing emails, so he or she could discard those emails that were irrelevant to room

requests. If SUBMIT_RESERVATION_FAILURE was followed by some adjustments in COMBOBOX and SUBMIT_RESERVATION_SUCCESS, then the test taker made two attempts to submit a reservation and did self-correction. It is when the two actions appeared in the selected features predominantly, and not accompanied by other useful actions, that they might not be able to solve the item.

For some significant dummy variables, COMBOBOX-related actions were in fact identified as robust features for the group with lower odds of a correct answer (e.g., all zero response on INFLUENCE_WLE_CA and READWORK_WLE_CA), while for other significant variables, the selected features were mainly associated with MAIL_MOVE for the less successful group. Adopting COMBOBOX-related actions could be a sign of understanding the purpose of the item and being able to figure out how to fill out the room requests. These test takers were considered closer to the borderline of a correct answer than the groups with mostly MAIL_MOVE actions; they might have had greater potential to get a score of 1 if proper interventions were given. On the contrary, if the majority of a test taker's actions were MAIL_MOVE, he or she might have needed more detailed guidance from the initial steps to submitting the requests.

An intriguing finding is that for the lowest age group (24 or less), the MAIL_MOVE action showed up in the top five robust features quite frequently, even though this group was more likely to answer U02 correctly compared to elder age groups. That is to say, given two test takers with the same occupation, work experience, work-related skills, and so on, the one who was 24 years old or younger would have had a higher probability of a correct response than the one who was 45 to 54, or 55 or older. However, the lowest age group often had different occupations and much less work experience than test takers who were 45 and above. The skills

and experiences that the older age groups had accumulated might have enabled them to apply more efficient problem-solving strategies despite younger test takers having more advantage on information technologies. Another possible explanation is that using MAIL_MOVE was characteristic of the youngest age group as an action taken without realizing it. They could simply have been moving emails around as they went through the thinking process.

4 Research Question 2: Construct Clusters of Test Takers Based on Process data and Explore Employment-Related Characteristics of the Clusters

Compared to the first research question, the second research question focuses on investigating the latent clustering structure of process data via a bottom-up approach. The characteristics of the clusters are explored regarding employment-related variables shared among test takers in the same cluster. The sample used in this section include test takers who had entries in background questionnaire data, as well as process data and responses for U02. The sample size of the analytical sample used in this section is 1326.

4.1 Research question 2 methods

36 features were used for cluster analysis, including 34 unigrams from process data (see Table 10), total number of actions on U02, and total response time on U02 (see Table 13). Bigrams and trigrams were not considered in the cluster analysis for two reasons. One is that given the 34 unigrams, there could be more than 40000 bigrams and trigrams in the process data, which is not feasible for interpreting the clusters. The other reason is that some bigrams and trigrams would have very low frequency, as the unigram with the lowest frequency only occurred 13 times. Therefore, even if those bigrams and trigrams were used in the cluster

analysis, they were not expected to distinguish well among clusters. Similar as before, the term weights (between- and within-individual weights) and survey weights were applied to the unigrams, and the weighted frequencies and variables were used for cluster analysis.

First, we extracted features from process data to partition test takers into clusters, using k-means clustering (Lloyd, 1982). A survey of data mining techniques states that k-means has been "the most popular clustering algorithm used in scientific and industrial applications" (Berkhin, 2006). This method starts with k arbitrary centers and seeks to minimize the squared difference between observations in the same cluster. k-means clustering method has several advantages over other clustering methods, such as hierarchical clustering and expectation–maximization clustering using Gaussian mixture models. First, k-means is efficient in terms of computational cost even with a large number of variables, which renders wider applications in large-scale industries possible. Second, observations can be 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. Finally, results of k-means are easily interpretable. Each observation only belongs to one cluster, and the centroids of the clusters are expressed on the original scale of the variables. Based on these advantages, k-means was chosen to perform cluster analysis in the present study. The "kmeans" function in the "stats" package in R version 3.3.3 (R Core Team, 2017) was used to carry out the proposed cluster analyses.

An appropriate number of clusters k was selected based on the change in the total within-cluster sum of squares. One caveat of k-means is that the results would be strongly influenced by the selection of initial seeds (e.g., Arthur & Vassilvitskii, 2007). Therefore, the stability of the cluster membership was examined to ensure the generalizability of the results. Then, clusters were interpreted based on the centroids of the 36 features. The homogeneous characteristics of

the clusters were explored, as well as the relationship between cluster membership and proficiency level and/or correctness of U02. We used R for conducting cluster analyses and further investigations.

Table 13. Descriptive statistics of number of actions and response time (in minutes) on 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

4.2 Research question 2 results

First, number of clusters was chosen based on the total within-cluster sum of squares, a measure of how close the observations within a cluster were to the centroid. Figure 1 demonstrates the changes in total within-cluster sum of squares at different number of clusters. Since the changes seemed to be marginal when the algorithm formed more than three clusters, further investigation was based on the three-cluster solution.

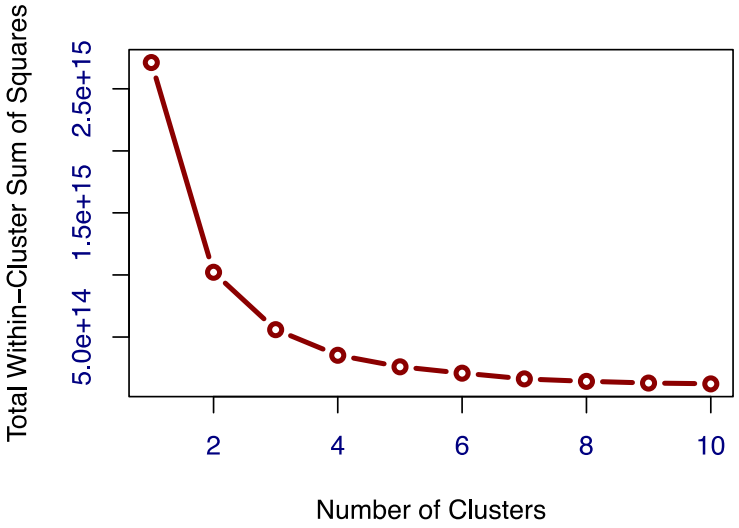

Figure 1. Total within-cluster sum of squares at different number of clusters.

We then explored the stability of cluster membership with 100 different initial seeds. Among the 1326 test takers, 1262 (95%) had exactly the same cluster membership in the 100

replications, taking into account label switching issue. Only 64 (5%) of them 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 cluster membership. This finding suggested that while initial seeds could have a large impact on the clusters, results in the present study had little dependence on initial seeds. Thus we would expect cluster membership and interpretations to be quite consistent, regardless of the seed used.

Table 14 presents the centroids of a three-cluster solution. As one can see, the centroids were not on the same scale as the raw frequency and the unweighted mean listed in Tables 10 and 13, as they were weighted by term weights and survey weights. In general, cluster 1 had the lowest weighted frequencies and means in almost all features, cluster 3 had the highest weighted frequencies and means, and cluster 2 resided in between clusters 1 and 3. The only exceptions were "NEXT_INQUIRY" and "START", where all three clusters had centroids at 0. Despite the non-zero raw frequencies and survey weights, the term weights were 0 as everyone had to take these two actions. Number of actions and response time on U02 appeared to be the most dominant features, since they were on greater scales than the other features. The three clusters could be interpreted as test takers with the least, medium, and the most actions. The least action cluster had the largest cluster size with 853 (64%) of the test takers in the analytical sample, the median action cluster included 398 (30%) test takers, and only 75 (6%) were in the most action cluster (see Table 15). This indicated that only a small group of test takers had a great number of actions and spent long time exploring U02, whereas majority of them clustered around fewer actions and much shorter time.

Table 14. Cluster centroids for a three-cluster solution.

		Clusters		
No.	Features	1	2	3
1	FOLDER_VIEWED	20365.6	40224.5	73644.8
2	ENVIRONMENT_WB	11062.0	65848.4	134117.1
3	ENVIRONMENT_MC	11375.0	59744.3	124133.6
4	MAIL_VIEWED_1	8625.7	23554.9	42659.6
5	HISTORY_VIEWCALENDAR	8296.1	61208.4	130657.0
6	MAIL_VIEWED_3	7983.3	41671.1	84185.9
7	HISTORY_RESERVATION	6972.4	53034.1	117019.4
8	COMBOBOX_ROOM	6020.8	54476.1	110608.7
9	MAIL_VIEWED_4	8606.6	35180.8	67087.3
10	MAIL_VIEWED_2	7891.8	33636.2	65864.9
11	MAIL_MOVE	18947.5	42469.4	87984.2
12	NEXT_INQUIRY	0.0	0.0	0.0
13	START	0.0	0.0	0.0
14	COMBOBOX_START_TIME	5498.0	47928.2	101684.2
15	COMBOBOX_END_TIME	5581.8	47942.1	103098.3
16	COMBOBOX_DEPT	5556.0	48052.1	101711.1
17	HISTORY_MEETINGROOMS	5848.3	43725.6	108077.0
18	ENVIRONMENT_WP	7738.8	33937.1	79654.0
19	SUBMIT_RESERVATION_FAILURE	4048.2	46768.2	109482.7
20	SUBMIT_RESERVATION_SUCCESS	4797.0	42081.0	85547.9
21	HISTORY_UNFILLED	4213.2	36222.2	91450.9
22	SUBMIT_UNFILLED	3589.7	34291.9	69265.5
23	FOLDER	6750.6	25942.1	62512.5
24	HISTORY_HOME	3808.0	18614.7	50805.3
25	CHANGE_RESERVATION	1522.0	23168.2	73968.0
26	KEYPRESS	2880.5	12713.7	65743.1
27	REPLY	2936.5	12319.8	30153.8
28	CANCEL	3250.7	13530.1	37320.8
29	HELP	3477.5	10343.4	17039.6
30	COPY	897.1	7628.4	38517.4
31	SEARCH	2278.3	3529.5	18895.4
32	SORT	949.7	4759.2	5540.5
33	PASTE	780.4	1494.6	33561.6
34	BOOKMARK	550.6	2875.4	9264.9
35	number of actions on U02	453665.4	2241595.2	5225357.2
36	response time on U02	58391.4	244418.8	475306.6

Table 15. Cluster size (percentage) of a three-cluster solution.

	Clusters		
U02score	1	2	3
0	760 (57.32%)	139 (10.48%)	23 (1.73%)
1	93 (7.01%)	259 (19.53%)	52 (3.92%)
Total	853 (64.33%)	398 (30.02%)	75 (5.66%)

Based on these clusters formed based on process data, the relationships between cluster membership and PSTRE scores and employment-related variables were investigated. Figure 2 depicted the association between clusters and PSTRE scores. As all 10 PVs showed similar pattern, the first PV (PV1) was used as an example for demonstration. Interestingly, PV1 increased from clusters 1 to 3, implying that the more action and time test takers spent on U02, the higher their PSTRE abilities were. There was a remarkable increase in PV1 from cluster 1 to cluster 2, where the 1st quartile of cluster 1 was approximately the same as the 3rd quartile of cluster 2. However, the increase in median PV1 from cluster 2 to cluster 3 was not as noticeable. Thus, there was one group with low PV1 and low amount of actions and time, one group with higher PV1 and considerable actions and time, and one group with the highest PV1 and the most effort. This pattern was consistent with our expectation that most test takers might not put much effort into solving and exploring U02, which was associated with lower performance on U02 and in the PSTRE domain overall. It is likely that the test takers disengaged intentionally after noticing that they were not able to solve the item.

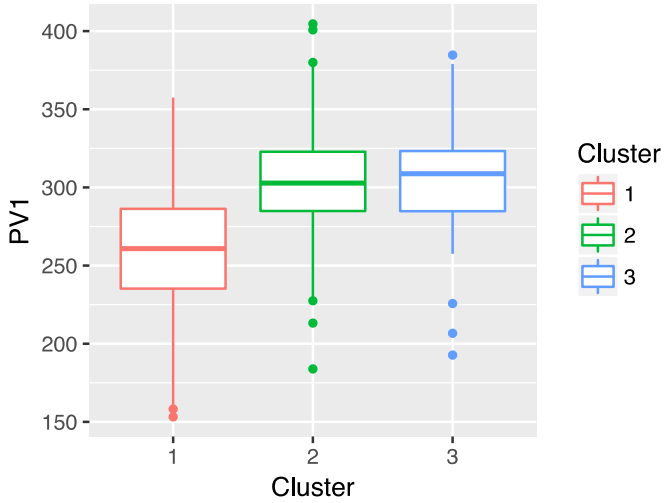

Figure 2. Boxplot of PV1 by cluster membership.

To validate whether most test takers in cluster 1 were low-performing ones who did not provide a correct answer to U02, a boxplot for PV1 distribution by cluster and by U02score was created (see Figure 3). Sample size was reported by clusters and by U02score in Table 15. Regardless of cluster membership, the distribution of PV1 for test takers with a correct answer to U02 was uniformly much higher than those with an incorrect answer. One interesting finding was that the median PV1 for test takers who answered correctly in all three clusters did not differ much, even though the median PV1 for clusters 2 and 3 were slightly higher than cluster 1 as expected. Yet, for the incorrect group, discrepancies among the median PV1 were remarkable, especially between clusters 1 and 2. Intriguingly, the incorrect group in cluster 3 had slightly lower median PV1 than cluster 2. While adopting more actions might be an indication of stronger will to explore the item more and potentially higher PSTRE skills, it could also signify that the test taker used less efficient strategy when those actions became excessive. As expected, the majority of cluster 1 did not answer correctly, since not many actions and time were taken. Clusters 2 and 3 tended to have more test takers who were successful in solving U02.

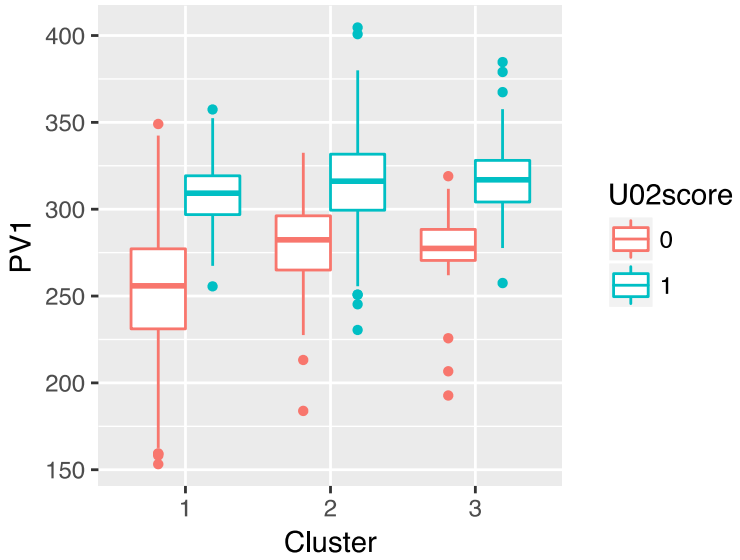

Figure 3. Boxplot of PV1 by cluster membership and U02score.

These findings suggested that the correct group can apply any number of problem-solving strategies to obtain a correct answer, and the choice of strategy did not necessarily associate with PSTRE skills. A small group of test takers in cluster 1 were able to use only a few actions to solve U02 in a short time, and their PSTRE scores were similar to those who applied much more actions. For the incorrect group, however, the number of actions and time spent on the item could be informative about a test taker's PSTRE skills. One test taker who put more effort into solving U02, even though he or she failed, was more likely to have higher PSTRE skills.

Previous explorations on the relationship between process data and PV1 and/or U02score demonstrated several patterns in process data that tended to have higher PSTRE scores or be more successful on U02. Based on the same results from cluster analysis, we further investigated the employment-related characteristics of the clusters. For instance, which earning group did test takers in each cluster fall in, and how much work-related skill use did each cluster possess, such as ICTWORK, NUMWORK, READWORK, and WRITWORK. Test takers' age and education level were also examined as background variables.

Figure 4 shows the distribution of ISCOSKIL4 in all three clusters, which is occupational classification of the test taker's last or current job. Compared to cluster 1, cluster 2 appeared to have more test takers in skilled occupations but fewer in semi-skilled blue-collar and elementary occupations. Cluster 3 included similar percentage of skilled workers as cluster 2 but more test takers semi-skilled blue-collar occupations. As such, cluster 3 in fact captured a special group of test takers who had overall the highest PSTRE scores and would take many actions and much time to solve U02. These characteristics might be more evident on test takers with semi-skilled blue-collar occupations.

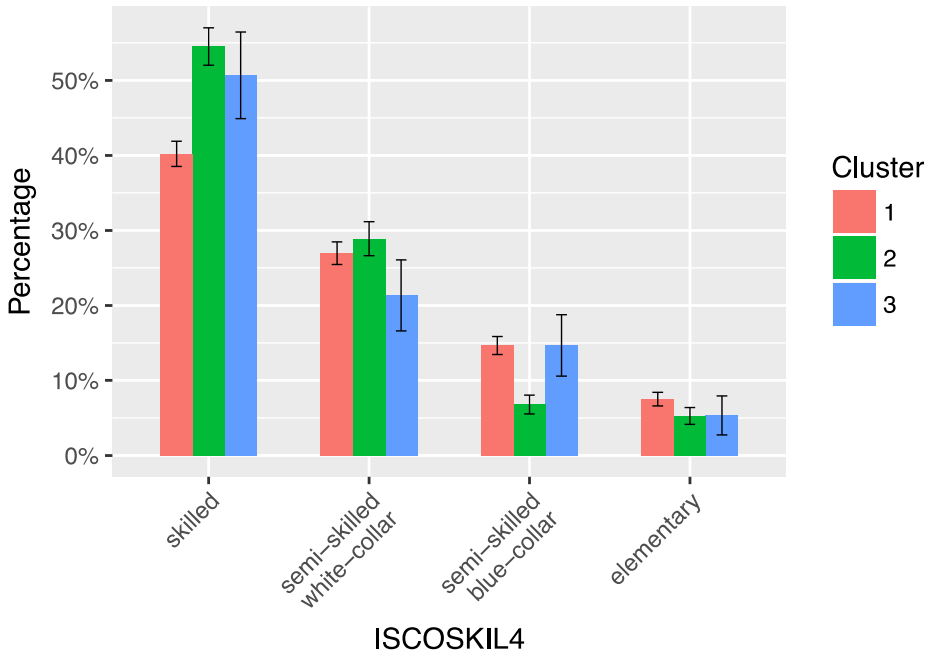

Figure 4. Distribution of ISCOSKIL4 in the three clusters[7].

Regarding monthly earning (EARNMTHALLDCL), clusters 2 and 3 were likely to have more test takers in the highest earning decile, whereas cluster 1 tended to have higher percentages in the lower earning deciles (see Figure 5). Two exceptions were 1st and 4th deciles, where most test takers were grouped in clusters 2 and 3. Despite the general pattern that earning was positively related to number of actions and response time spent on U02, some test takers who were younger or at the early stage of their careers may have lower salaries but high problem-solving capacity. This is consistent with the fact that young test takers whose salaries were in the lower deciles did have high PSTRE scores.

[7] Percentages in this plot and the following ones represent the proportion of a certain employment status/skill-use level within a cluster. Standard errors of the percentages were also included in the plots.

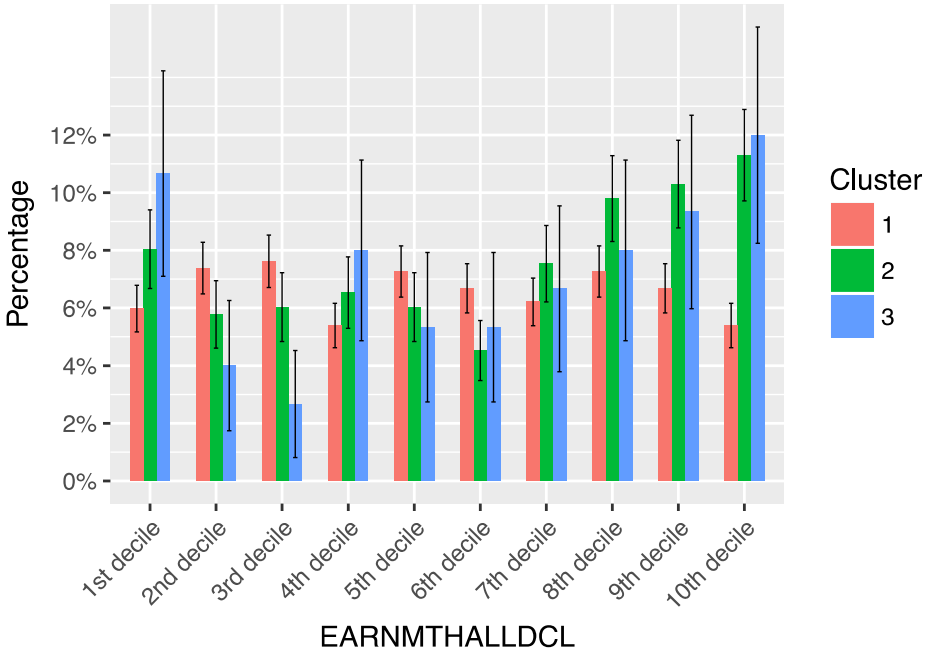

Figure 5. Distribution of EARNMTHALLDCL in the three clusters.

Work-related skill use also demonstrated a similar pattern. Cluster 1 was more likely to include test takers in the lower skill use levels (40% or lower), while more test takers with high skill use levels (40% or higher) were assigned to clusters 2 and 3. Figure 6 depicts the distribution of ICTWORK, NUMWORK, READWORK, and WRITWORK. Even though cluster 3 had the largest number of actions and the longest time, it often consisted of more test takers in skill use levels lower than cluster 2. This may be consistent with the finding that cluster 3 included a larger proportion of test takers in semi-skilled blue-collar occupations than cluster 2, which did not necessarily require higher levels of ICT, numeric, reading, or writing skill use. Compared to those in cluster 3, test takers in cluster 2 with high skill use might have learned to work efficiently, resulting in less response time and number of actions with similar PSTRE scores.

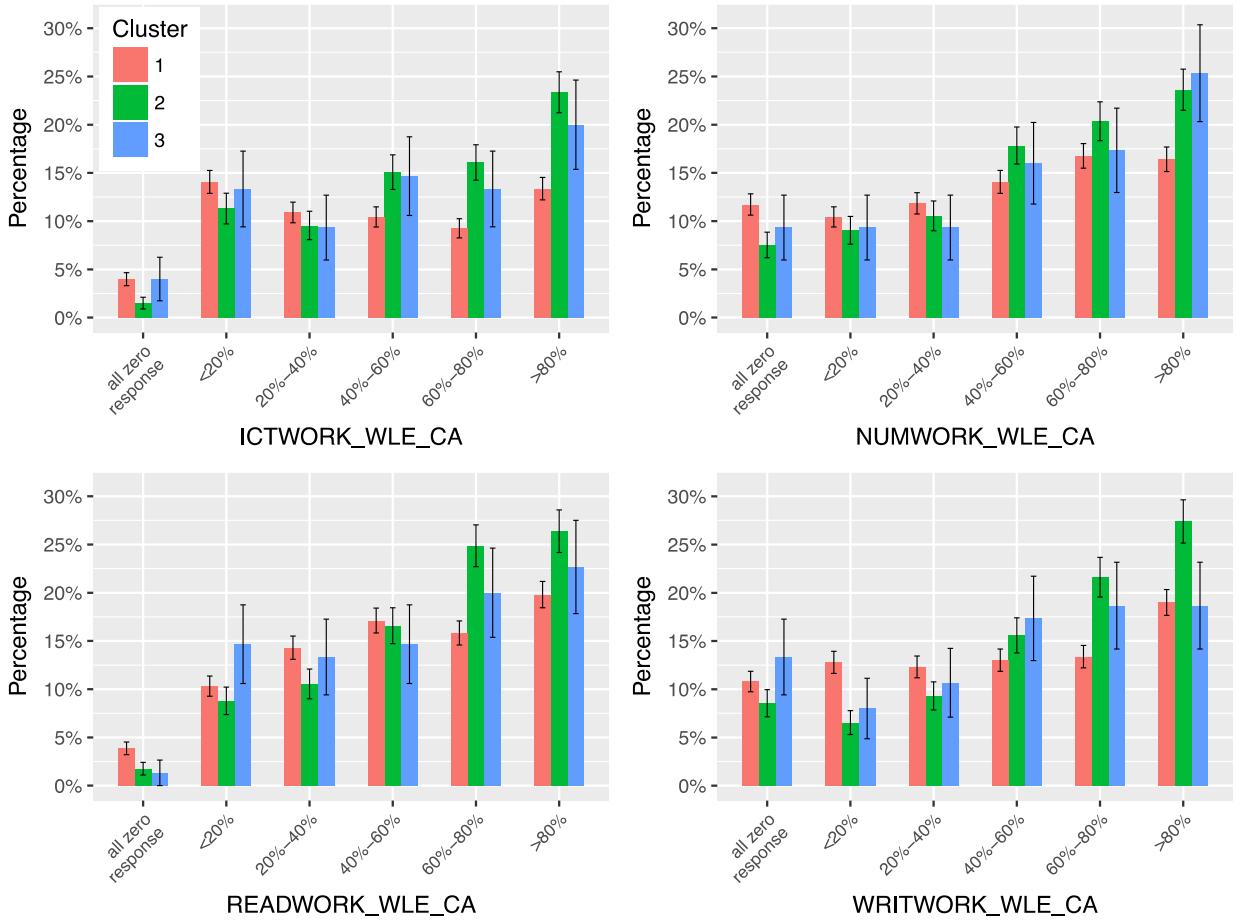

Figure 6. Distribution of ICTWORK_WLE_CA, NUMWORK_WLE_CA, READWORK_WLE_CA, and WRITWORK_WLE_CA in the three clusters.

In terms of age, over 30% test takers in cluster 3 were 24 or less, which was the highest among the three clusters (see Figure 7). Cluster 2 had the largest age group in 25-34, and most 55 or plus test takers were classified into cluster 1. This provides another explanation to the observed pattern in process data. On average, test takers in cluster 3 were younger than those in the other two clusters, and the age difference could relate to different responding behaviors in PSTRE items. Compared to the elderly, younger test takers tended to be more active in human-computer interactions, more familiar with manipulating computer systems, and learn faster when encountering new interfaces.

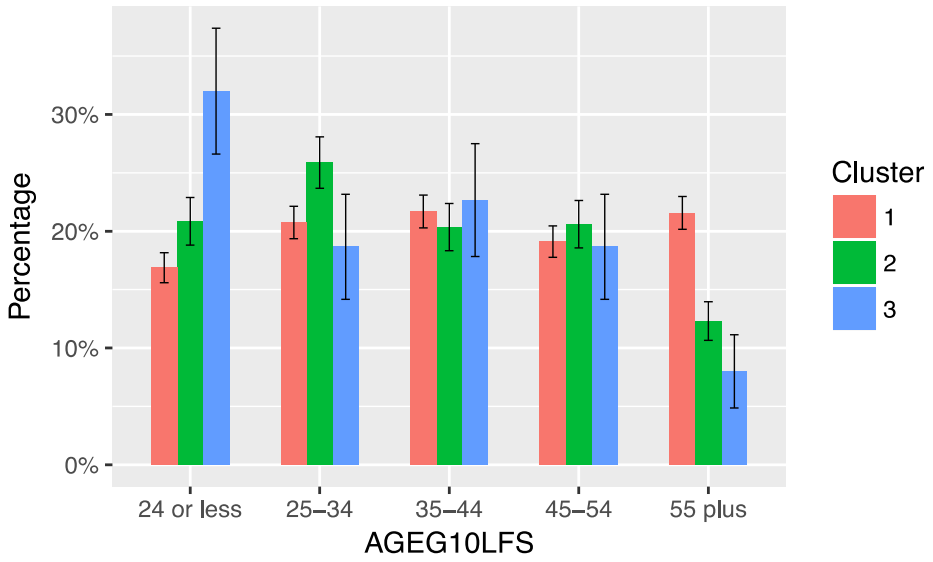

Figure 7. Distribution of AGEG10LFS in the three clusters.

Figure 8 summarizes the distribution of EDCAT6 for each cluster, which is the categorization for education with six categories. Comparing within each category, cluster 1 tended to have higher percentages in lower secondary or less and upper secondary, cluster 2 included the largest proportions of test takers with a bachelor's degree or higher, and cluster 3 had the highest percentages in tertiary professional degrees. Therefore, test takers in cluster 1 were the lowest performing group in the PSTRE domain with the lowest education level overall. Although cluster 3 demonstrated slightly higher median PV1, the percentages in the bachelor's degree or higher categories were lower than cluster 2. It turned out that test takers in cluster 3 might not possess the highest education level, but it is likely that their openness to experience enabled them to score well on PSTRE.

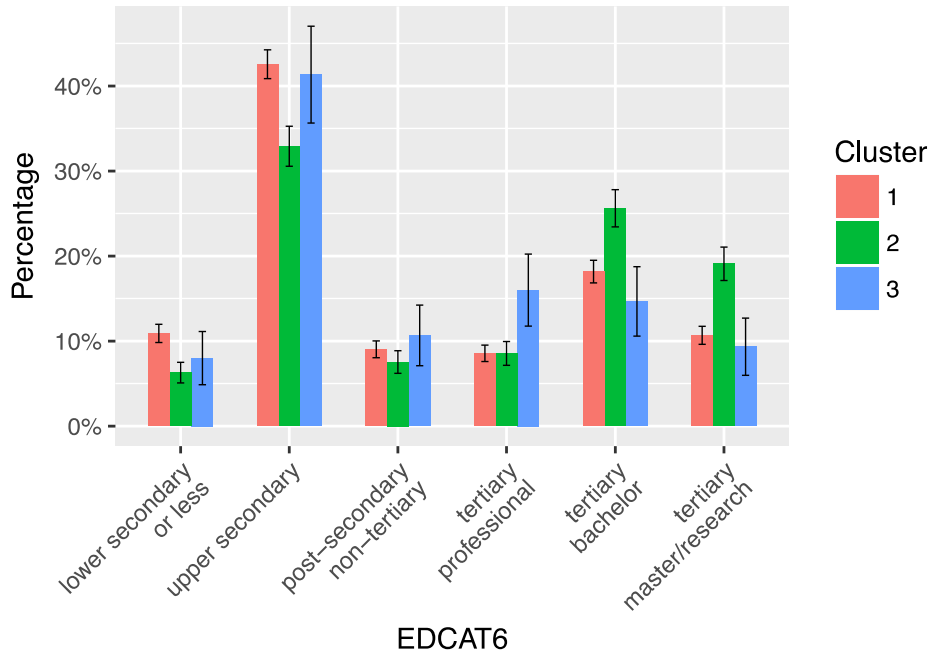

Figure 8. Distribution of EDCAT6 in the three clusters.

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

An interesting finding was that the group with the highest action frequencies, response time, and PSTRE scores often did not possess the highest income, work-related skill use, or education level. Yet, they were the youngest group with a decent proportion of skilled or semi-skilled occupations. This group was distinctive from other test takers in that they were the most

explorative or adventurous test takers who were willing to take a large number of different actions in solving a problem. This characteristic was likely to relate to higher PSTRE skills.

5 Research Question 3

The first two research questions apply top-down and bottom-up approaches respectively to examine how different employment-related variables associate with test takers' response behaviors in process data for one representative item, U02. The last research question aims at the consistency of the features from process data across items, and the association between the consistency across items and employment-related variables. In other words, do test takers tend to use similar strategies to solve different items? How can test takers be classified based on their consistency in process data across multiple items? What do test takers in such groups share in terms of their employment-related variables? These questions are discussed in this section.

5.1 Research question 3 methods

Recently, the consistency of problem-solving strategies across items has been of increasing interest to researchers. Evaluating the consistency across items renders possible capturing and modeling person-related latent characteristics, in addition to the skill level measured by the correctness of item responses. Yet difficulties arise in comparing features from process data among items. In PIAAC, items are designed in three environments – email, web, and spreadsheet, and each item has at least one environment. Items that share environments would have more actions in common than items that do not have common environments, as tasks in the same environment tend to be similar. For instance, in the spreadsheet environment, test takers are usually asked to sort or find certain information. However, the actual action sequences

are still less comparable even for items sharing environments. The fact that process data provides rich amount of information at a fine-grained level about a specific item also impedes it from being generalized across items.

One approach to solving this difficulty is to use common features across items. In this section, we chose number of actions and response time as the common features from process data. Another advantage of adopting these two variables was that in the cluster analysis conducted in section 4, number of actions and response time had the most evident impact on forming the clusters. Results from cluster analysis indicated that compared to unigrams, the latent clustering structure of process data was mainly driven by number of actions and response time of the test takers. As such, these two variables were utilized to explore the consistency of test takers' response behaviors.

The consistency of response behaviors was investigated across two items, U02 and U19a. U02 was chosen as a representative item in addressing the first two research questions. U19a was selected for two reasons. One is that U02 and U19a shared one common environment, email. Since little research was done regarding across item comparisons, we started with items that were more similar to each other. Nonetheless, future investigations can be done across more items as number of actions and response time are variables that can be generalized across all items. Another reason is that as U02 and U19a were both from booklet PS2, the analytical sample size was larger than if the two items were from different booklets. The sample size of the analytical sample in this section is 1340.

Test takers who took both items were categorized as follows. First, a test taker was classified based on whether his or her number of actions on U02 was larger than the median number of actions, denoted by (+)actions if so and (-)actions otherwise. Then, the test taker was

grouped by whether his or her response time on U02 was larger than the median response time on U02. Similarly, (+) for response time longer than median and (-) otherwise. Fully crossing the two categorizations on the two variables resulted in four groups in total, plotted in four quadrants. Then the same process was repeated for U19a. The four-quadrant categorization on the two items was detailed in Table 16.

Table 16. Four quadrants of test takers.

Quadrant	Number of Actions on U02	Response Time on U02	Number of Actions on U19a	Response Time on U19a
1	(+)actions	(+)response time	(+)actions	(+)response time
2	(-)actions	(+)response time	(-)actions	(+)response time
3	(-)actions	(-)response time	(-)actions	(-)response time
4	(+)actions	(-)response time	(+)actions	(-)response time

Based on the four quadrants of test takers on the two items, test takers were assigned into five consistency groups, including four consistent groups and one inconsistent group. For test takers who were classified into the same quadrant on both items, they were assigned to consistency groups 1 to 4 (see Table 17). For those who belonged to different quadrants on U02 and U19a, they were included in consistency group 5. The PSTRE scores and employment-related characteristics of the five groups were examined in detail.

Table 17. Consistency measure between U02 and U19a.

Consistency	Quadrant on U02	Quadrant on U19a
1	1: (+)actions (+)response time	1: (+)actions (+)response time
2	2: (-)actions (+)response time	2: (-)actions (+)response time
3	3: (-)actions (-)response time	3: (-)actions (-)response time
4	4: (+)actions (-)response time	4: (+)actions (-)response time
5	Any	Inconsistent with the quadrant on U02

5.2 Research question 3 results

Figures 9 and 10 demonstrate the scatterplots for response time vs. number of actions for U02 and U19a respectively. For U02, there were much more test takers who spent either more time and made more actions or spent less time and made less actions (quadrants 1 and 3) than the other configurations of time and actions (quadrants 2 and 4). Whereas for U19a, test takers were rather evenly distributed in the four quadrants (see Table 18). Another remarkable difference between the two items was that there appeared to be a stronger relationship between number of actions and response time for U02 than for U19a. For U02, there was a trend that more actions were associated with longer time, although the variance in response time on U02 seemed to increase when the number of actions on U02 increased. However, the association between the number of actions and response time on U19a was weaker in that more actions on U19a did not necessarily associated with longer response time.

There could be two reasons why the patterns were found to be different for U02 and U19a. First, the two items were located differently in booklet PS2. U02 was in the middle while U19a was located at the beginning of the booklet. For items like U19a, warm-up effect (e.g., van der Linden, Breithaupt, Chuah, & Zhang, 2007) could play a role in their response behaviors. That is, test takers tended to spend more time on the initial items than they actually needed. Second reason might be due to the difference in item difficulty, which was in part related to the possible approaches or paths to solving the item. There might be more approaches to a correct answer to U19a (by sorting or searching), where the relationship between number of actions and response time for different approaches could be variant. For U02, however, test takers had to follow a similar path to the correct answer, thus number of actions may be more proportional to response time.

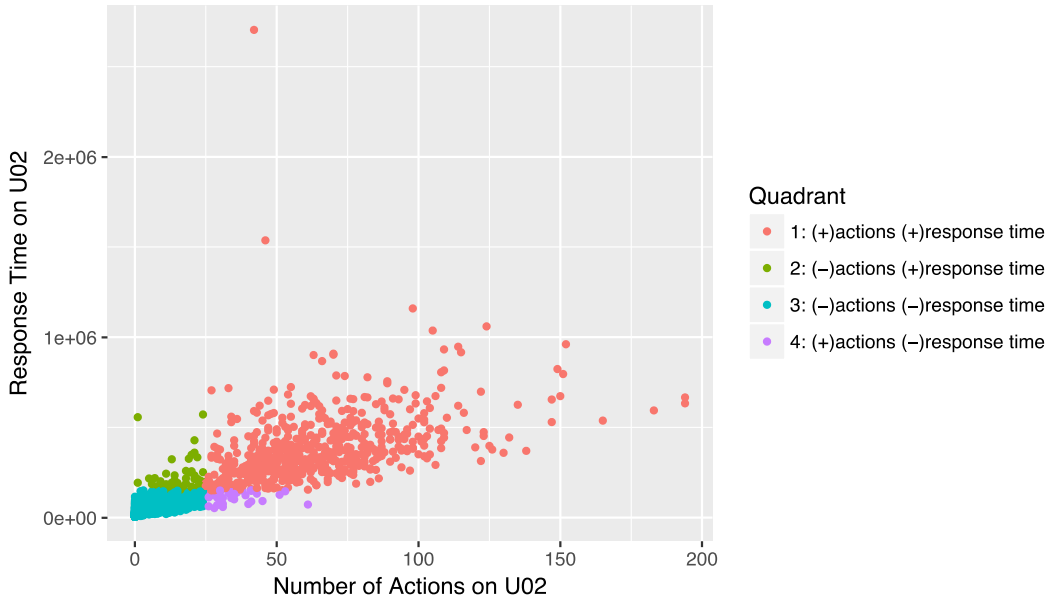

Figure 9. Response time on U02 vs. number of actions on U02.

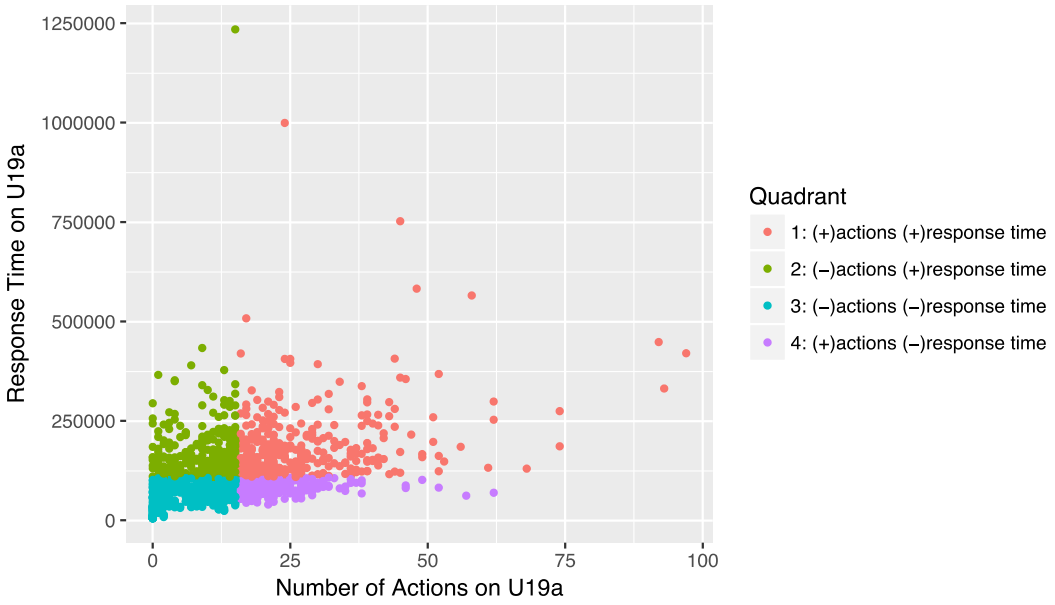

Figure 10. Response time on U19a vs. number of actions on U19a.

Table 18. Sample size of the four quadrants for U02 and U19a.

	Quadrant			
Item	1	2	3	4
U02	621	49	631	39
U19a	383	287	443	227

By comparing the four quadrants for U02 and U19a, five consistency groups were formed based on the definition in Table 17. Similar to section 4, we used PV1 to demonstrate the association between PSTRE scores and consistency groups, as the general pattern was the same across different PVs (see Figure 11). Test takers who consistently took many actions and time (consistency group 1) in general had the highest PSTRE skills, whereas those who consistently used fewer actions and shorter time (consistency group 3) had the lowest PSTRE skills overall. Test takers in consistency groups 2 and 4 tended to fall in somewhere between group 1 and group 3. The last group, where test takers were assigned to different quadrants in the two items, had a wider range of PSTRE skills, but majority of group 5 was still bounded by the medians of groups 1 and 3. Sample size of the five consistency groups was reported in Table 19. Notice that the sample size for more actions and less time or fewer actions and more time on both items (consistency groups 2 and 4) was quite small, since U02 had relatively small sample size in quadrants 2 and 4. To obtain results and interpretations with satisfactory reliability, we only included consistency groups 1, 3, and 5 for examining the relationship between these groups and employment-related variables.

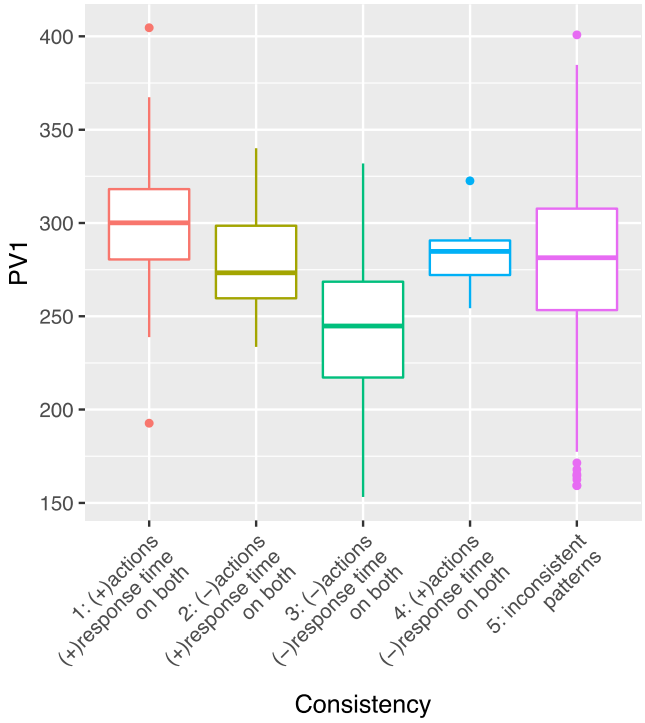

Figure 11. Boxplot of PV1 by consistency groups.

Table 19. Sample size of the five consistency groups.

	Consistency				
	1	2	3	4	5
Sample size	198	11	265	6	860

Figure 12 depicts the distribution of ISCOSKIL4 by consistency groups 1, 3, and 5. Test takers who applied many actions and spent much time on both items (consistency group 1) had the largest proportion of skilled occupations, while those who spent fewer actions and less time on both (consistency group 3) tended to have semi-skilled or elementary occupations. Group 5, the inconsistent group, had a similar distribution as group 3 regarding the semi-skilled occupations, but higher proportions of test takers with skilled occupations and lower proportions of test takers with elementary occupations. One interpretation of this finding was that group 1 contained more test takers with typical response behaviors from skilled occupations, group 3 represented the common behaviors from semi-skilled or elementary occupations, and those who

had less typical response behaviors were classified into group 5. Following this interpretation, test takers with skilled occupations consistently applied more actions and used more time in solving the items, whereas semi-skilled or elementary workers were inclined to spend fewer actions and shorter time.

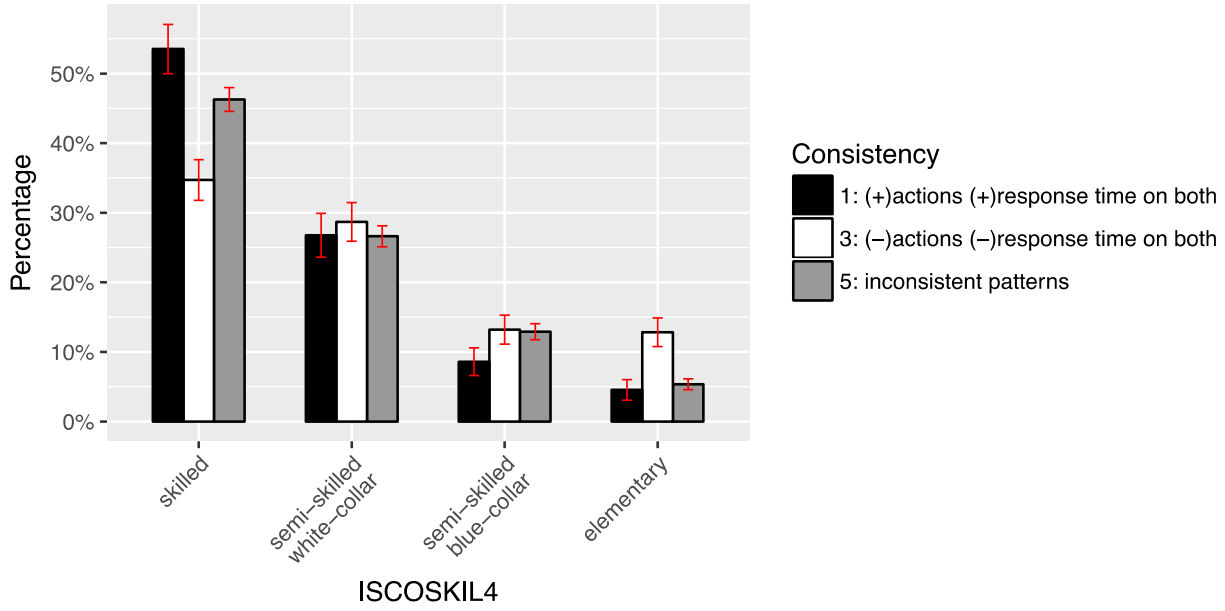

Figure 12. Distribution of ISCOSKIL4 in the three consistency groups.

Regarding monthly earning, the distribution in three consistency groups was displayed in Figure 13. Comparing the two groups with consistent patterns on both items (groups 1 and 3), it was found that test takers who employed a large amount of actions and time on both items (group 1) had higher percentages in the 4^{th}, 7^{th} to 10^{th} deciles, especially the 8^{th} and 10^{th} deciles if the uncertainty was taken into account. In other words, test takers who consistently put more effort into the items had overall higher monthly income than those who did not spend much time and actions. Interestingly, the inconsistent group (group 5) also had high percentages in 7^{th} to 9^{th} deciles. This might be an indication that group 5 encompassed more diverse background profiles, rather than a mix of profiles of test takers from the other two consistency groups.

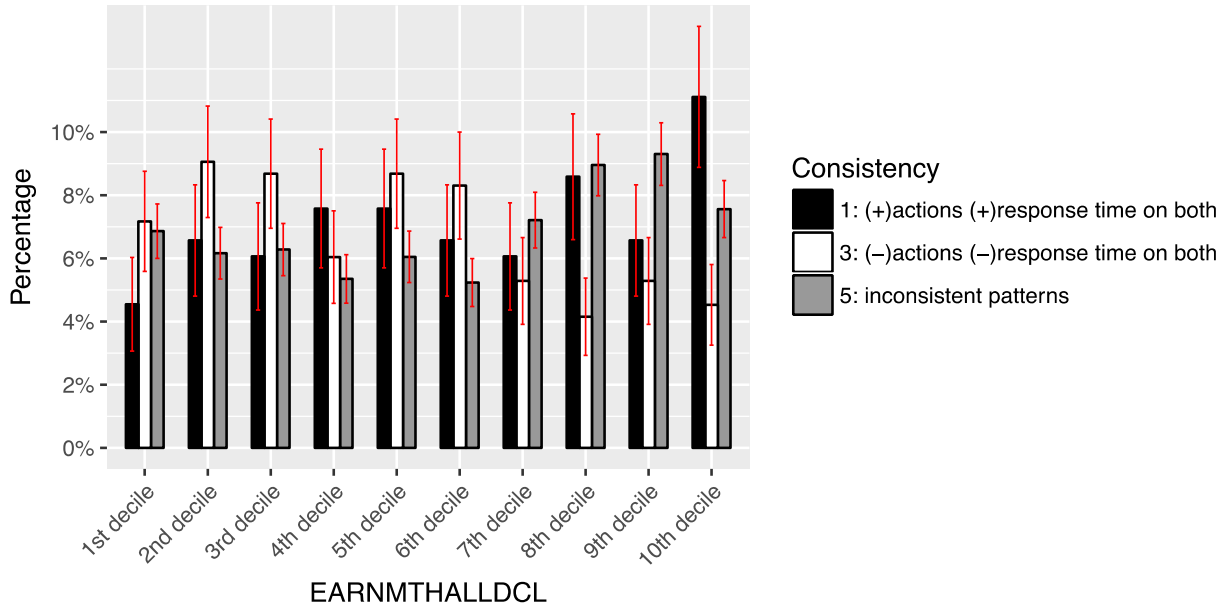

Figure 13. Distribution of EARNMTHALLDCL in the three consistency groups.

The association between consistency and work-related skill use was also of interest to us. Figure 14 shows four panels for the distribution of ICTWORK_WLE_CA, NUMWORK_WLE_CA, READWORK_WLE_CA, and WRITWORK_WLE_CA in the three consistency groups. One general observation was that test takers that did hard work on both items (group 1) tended to have higher work-related skill use (40% and above), and those who did not make much effort (group 3) consisted of more test takers with lower skill use (below 40%). The percentages of group 5 often ranged somewhere between the percentages of groups 1 and 3. This might suggest that there were multiple subgroups in consistency group 5 with diverse profiles.

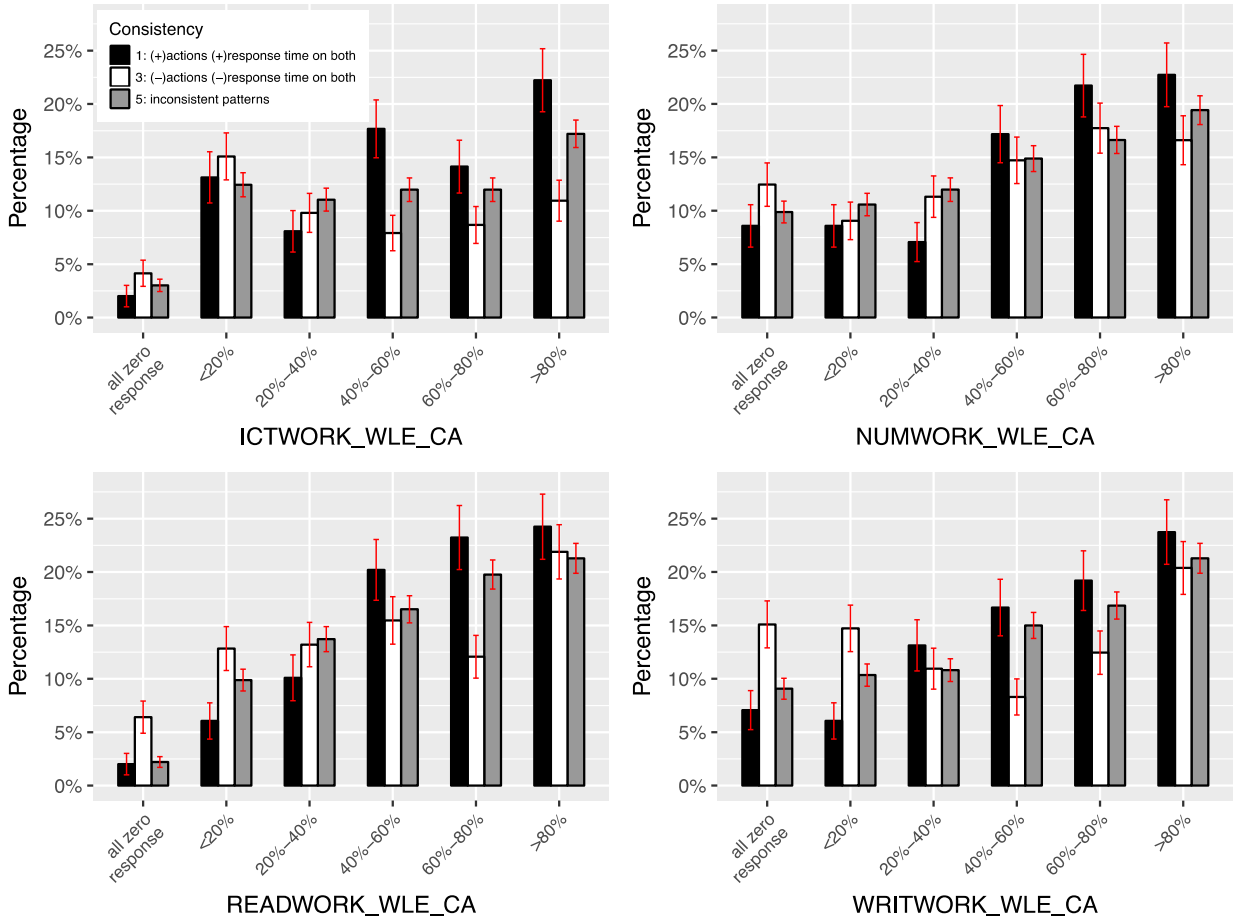

Figure 14. Distribution of ICTWORK_WLE_CA, NUMWORK_WLE_CA, READWORK_WLE_CA, and WRITWORK_WLE_CA in the three consistency groups.

Lastly, we examined the distribution of age and education. Overall, the distribution of age appeared to be similar across the three consistency groups, except that consistency group 3 had more test takers within 35-44 age band compared to the other groups. Test takers who constantly spent a lot of effort in solving the two items tended to be better-educated than those who chose otherwise in both items. Compared to those put in less effort, more test takers who took many actions and long time on both items had bachelor, master or research degrees. The finding regarding the inconsistent group was similar in that it might contain more than one distinguishable subgroup.

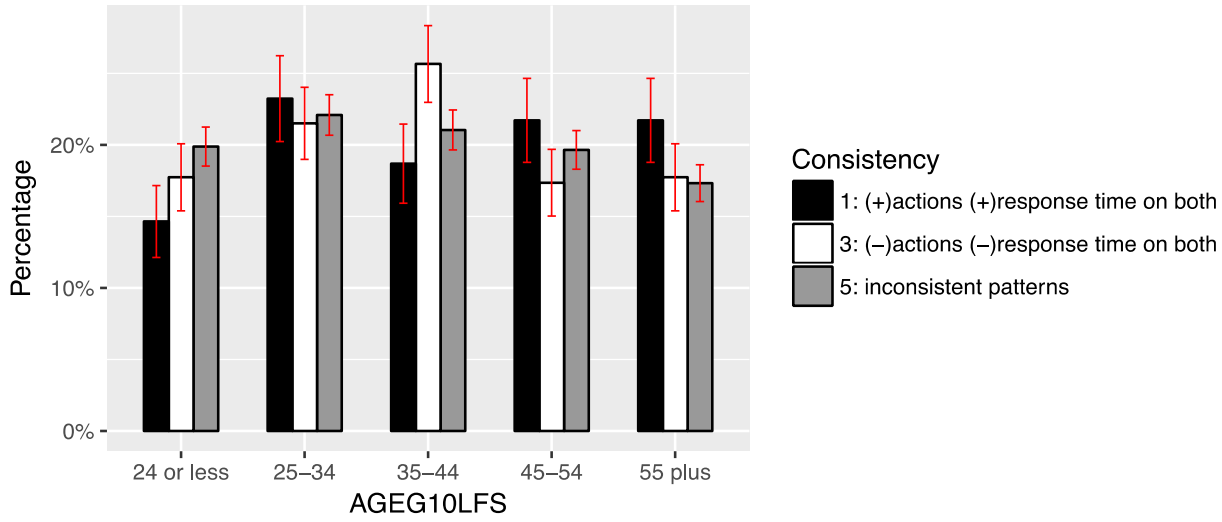

Figure 15. Distribution of AGEG10LFS in the three consistency groups.

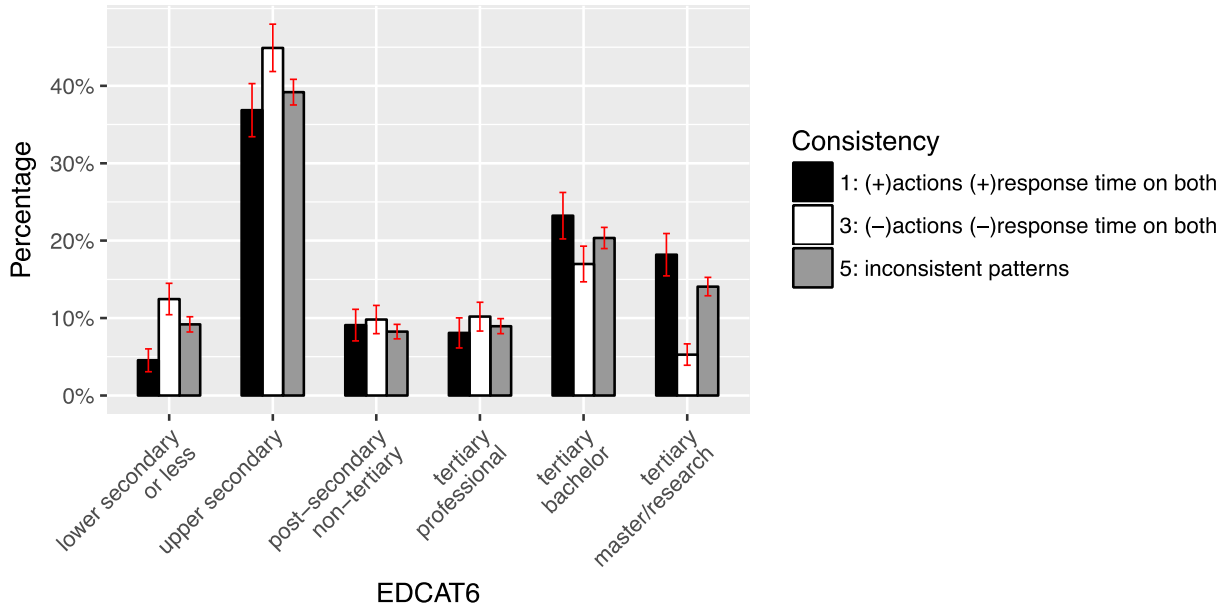

Figure 16. Distribution of EDCAT6 in the three consistency groups.

In summary, test takers were classified into five consistency groups based on their process data from U02 and U19a. We mainly focused on exploring the characteristics of three groups due to sample size constraint. Test takers who consistently put considerable effort in solving the two items were more likely to have skilled occupations, higher income, higher work-related skill use, higher age, and better education. Those who did not spend much time and

actions on either item demonstrated the opposite characteristics. The last group we evaluated involved test takers who had inconsistent patterns on the two items. This group tended to show a blend of characteristics of the two most extreme consistency groups, though in some cases it was found to encompass a more diverse variety of profiles due to potential distinctive subgroups.

6 Discussion

6.1 Summary and implications

The current study presented a comprehensive exploration of the relationship between process data and employment-related variables by conducting three sub-studies. These results raise questions about adults' PSTRE skills in the U.S. population, and whether their employability is associated with PSTRE skills. With the coming era of high technology, it is of importance to ensure people get readiness for work in the digital environment. The use of process data not only provides more detailed information about how people solve the problems and the reasons of their success and failure, but also helps identify the groups that are in shortage of skills in employment. These findings tremendously support the policy makers to make better decisions on allocating educational resources the groups who need to master new skills for work readiness and also outline the training direction for people who are in urgent need to improve competences in the adult education and foster the life-long learning.

This study tackled with the association between process data and employment-related variables from three aspects. First, what employment-related variables significantly affected the success on PSTRE items, what features can be extracted from the groups formed based on significant variables, and how were the features different between groups. Second, what was the latent clustering structure of process data, and how were the clusters related to employment-

related variables. The first aspect was a top-down approach, whereas the second aspect was a bottom-up approach. Lastly, whether test takers' choice of response behaviors or problem-solving strategies were consistent across multiple items, and what employment-related characteristics were shared among different consistency groups.

In general, most significant variables and their regression coefficients were consistent with our expectations. Test takers who were well-educated and young, and had more work experience and higher work-related skill use, tended to have higher scores in the three domains and higher odds of success in the example item. Similar to findings in He and von Davier (2015, 2016), test takers who had higher income, work-related skill use, and education level demonstrated clear subgoals in solving the item.

The most important implication of the present study was that features identified from process data shed light on how much intervention a certain group of test takers might need. There was clear evidence from process data for the steps to read emails, filter the irrelevant email, and submit requests. Given sufficient evidence for each required step, further analyses could potentially determine at which specific step an intervention was needed. It also provides the possibility of scoring complex items like PSTRE items base on process data in the future.

Overall, groups with different levels of employment-related variables often demonstrated quite distinctive characteristics with respect to test-taking behaviors. Nevertheless, actions indicative of low PSTRE skill for one group does not necessarily mean the same for another group. Therefore, it is important to establish a basic understanding of the common action sequences that a group would take before further analyses, or making decisions on the necessary training and interventions.

Features extracted from process data were also closely related to their employment-related variables. In research question 2, the latent clustering structure of process data was explored, and the relationship between such clustering and employment-related variables was demonstrated. Test takers were found to be clustered in three groups, representing three levels of effort spent on U02. Those who took a large amount of different actions and response time tended to have higher PSTRE scores, especially for test takers who were not able to provide a correct answer for U02. A small group of test takers were found to have high levels of openness to experience (McCrae, 1987). They often had the highest action frequencies, response time, and PSTRE scores, but did not possess the highest income, work-related skill use, or education level. For test takers who did not put much effort in solving the items, the majority of them had semi-skilled or elementary occupations, lower monthly income, lower work-related skill use, higher age, and lower education.

Number of actions and response time were found to be the most dominant features from process data for U02, as indicators of endeavor. In research question 3, we evaluated the characteristics of test takers who demonstrated consistent or inconsistent patterns across multiple items in the same booklet, with a focus on number of actions and response time. Test takers who consistently spent fewer actions and shorter time on the items possessed similar profiles as those who did not put much effort into U02, indicating that test takers with such characteristics tended to choose similar levels of effort when solving different problems.

When other features from process data cannot be easily extracted, number of actions and response time could also be informative regarding the extent to which interventions were needed, especially when test takers did not answer the item correctly. Those who only did a few actions quickly might need more training than those who spent more effort solving the item. In

particular, test takers who did not make much effort in solving the items tended to have semi-skilled or elementary occupations, lower monthly income, lower work-related skill use, higher age, and lower education. This group of test takers was considered in need of further education or intervention.

Similar profiles were found among test takers who were consistently careless on neither item. Test takers were free to adopt any test-taking strategies throughout the test; in fact, majority of test takers demonstrated some changes in their behaviors as captured by number of actions and response time. On the other hand, the present study showed empirical evidence that test takers switched test-taking behaviors, possibly due to changes in the location of the item and nature of the item. The association between number of actions and response time was less evident for the studied item located at the beginning of the booklet. For the item located at the middle of the booklet, the relationship between the two features became clearer. Other factors that could impact the relationship between number of actions and response time were changes in item content, difficulty, or possible problem-solving strategies.

6.2 Limitations and future work

Limitations from the present study also lay the foundation for future explorations. This study presents methods and approaches to analyze process data from PIAAC technology-rich environments. Results from cluster analysis in research question 2 seem to suggest the presence of different levels of interactions in the sample. This gives rise to the question whether, compared to a categorical variable, a continuous latent variable might be more appropriate to model the patterns in process data. Both of these two approaches can be used, depending on the different purpose of research. Cluster analysis was employed in the present study because the

research purpose was to classify people with similar patterns into groups. One can also choose among models with continuous latent variables, such as latent regression analysis, to make inferences about the characteristics of the process data. Moreover, bigrams and trigrams were eliminated from the cluster analyses due to their low frequencies in the sample, thus the order of the actions was not taken into account. It would be of interest to include such information into cluster analyses and investigate the impact on the latent clustering structure.

For the last research question, we explored the consistency of problem-solving behaviors across two items where consistency was defined based on the median of two most dominant features from process data. This is because the exploration of process data is highly context dependent (Rupp, Gushta, Mislevy, & Shaffer, 2010). A slight change in the task environment could result in completely different actions. Admittedly, even if the assessment environments were the same, how the task was implemented may also have an impact on the consistency. Further, extracting higher-level problem-solving strategies from the actual actions could be subjective. Therefore, in the present study, we explored the consistency issue through features that are common across items without changing the granularity of the data. It is worth exploring whether other consistency measures could be developed to provide more sights regarding the consistency of test-taking behaviors. We focus on the test takers' general patterns, for instance, whether a person generally responds quickly, or a person is cautious and checks every step carefully, etc.

Only two representative items were used in the present study due to the complexity and the amount of process data an item typically has, especially items with the appropriate level of difficulty. It would be beneficial to explore more items to obtain more generalizable and more conclusive results. One such exploration is done by He, Borgonovi, and Paccagnella (2019),

where the authors introduced longest common subsequence approach for the distance between individual sequence and predefined action sequence to evaluate the consistency of test takers' strategies across PSTRE items. The distance between individual and predefined action sequence provides standard indicators across items for measuring whether people follow the predefined strategy in a consistent way or not. The other alternative would be to develop a model with a latent variable that describes the consistency of test takers' behaviors.

To summarize, this study provides critical evidence of relationships between employment-related background variables and sequential patterns in PSTRE based on the U.S. sample in PIAAC. It also provides information to education policy makers to explore the test-taking behaviors by different employment-related subgroups, thus helping to find an optimal solution to improve their PSTRE skills via a tailored approach. We recommend continuing to explore the generalizability of results presented in this study across PSTRE items in future studies and to make comparisons across countries and language groups.

References

Allison, P. D. (2001). *Missing data: Sage university papers series on quantitative applications in the social sciences (07–136)*. Thousand Oaks, CA.

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.

Bainbridge, W. L., & Lasley, T. J. (2002). Demographics, diversity, and K-12 accountability: The challenge of closing the achievement gap. Education and Urban Society, 34(4), 422-437.

Berkhin, P. (2006). A survey of clustering data mining techniques. In *Grouping multidimensional data* (pp. 25-71). Springer, Berlin, Heidelberg.

Cohen, J., & Cohen, P. (1985). *Applied multiple regression and correlation analysis for the behavioral sciences* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Croizet, J. C., & Dutrévis, M. (2004). Socioeconomic status and intelligence: Why test scores do not equal merit. Journal of Poverty, 8(3), 91-107.

Currie, J., & Thomas, D. (1999). Early test scores, socioeconomic status and future outcomes (No. w6943). National bureau of economic research.

Dreher, G. F., & Ryan, K. C. (2000). Prior work experience and academic achievement among first-year MBA students. Research in Higher Education, 41(4), 505-525.

Feinerer, I. (2017). Introduction to the tm Package Text Mining in R. Retrieved from https://cran.r-project.org/web/packages/tm/vignettes/tm.pdf

Goodman, M., Sands, A., & Coley, R. (2015). America's skills challenge: Millennials and the future. Princeton, NJ: Educational Testing Service. Retrieved from http://www.ets.org/s/research/30079/asc-millennials-and-the-future.pdf

He, Q., Borgonovi, F., & Paccagnella, M. (2019, under review). *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., Glas, C. A., Kosinski, M., Stillwell, D. J., & Veldkamp, B. P. (2014). Predicting self-monitoring skills using textual posts on Facebook. *Computers in Human Behavior*, *33*, 69-78.

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., & 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.

International Association for the Evaluation of Educational Achievement (IEA). (2013). *IDB Analyzer* (Computer software and manual). Hamburg, Germany: IEA.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. *Machine Learning: ECML-98 Lecture Notes in Computer Science*, *1398*, 137-142.

Jones, M. P. (1996). Indicator and stratification methods for missing explanatory variables in multiple linear regression. *Journal of the American Statistical Association*, *91*(433), 222-230.

Kingston, N. M., & Dorans, N. J. (1984). Item location effects and their implications for IRT equating and adaptive testing. *Applied Psychological Measurement*, *8*(2), 147-154.

Leary, L. F., & Dorans, N. J. (1985). Implications for altering the context in which test items appear: A historical perspective on an immediate concern. *Review of Educational research*, *55*(3), 387-413.

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* (Vol. 999). Cambridge, MA: MIT press.

Marks, G. N. (2006). Are between-and within-school differences in student performance largely due to socio-economic background? Evidence from 30 countries. *Educational Research*, *48*(1), 21-40.

McCrae, R. R. (1987). Creativity, divergent thinking, and openness to experience. *Journal of Personality and Social Psychology*, *52*(6), 1258-1265.

Organization 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

Organization for Economic Co-operation and Development. (2013b). *Time for the U.S. to reskill?* Retrieved from http://www.oecdilibrary.org/education/time-for-the-u-s-to-reskill_9789264204904-en. OECD Publishing. doi:10.1787/9789264204904-en

Organization for Economic Co-operation and Development. (2016). *Skills matter: Further results from the survey of adult skills.* Retrieved from https://www.oecd.org/skills/piaac/Skills_Matter_Further_Results_from_the_Survey_of_Adult_Skills.pdf

R Core Team (2017). *R: A language and environment for statistical computing. R Foundation for Statistical Computing*. Vienna, Austria. URL https://www.R-project.org/.

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 http://nces.ed.gov/pubsearch

SAS Institute. (2015). *Base SAS 9.4 Procedures Guide*. SAS Institute.

Schmitt, N., Keeney, J., Oswald, F. L., Pleskac, T. J., Billington, A. Q., Sinha, R., & Zorzie, M. (2009). Prediction of 4-year college student performance using cognitive and noncognitive predictors and the impact on demographic status of admitted students. *Journal of Applied Psychology*, *94*(6), 1479-1497.

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

van Der Linden, W. J., Breithaupt, K., Chuah, S. C., & Zhang, Y. (2007). Detecting differential speededness in multistage testing. *Journal of Educational Measurement*, *44*(2), 117-130.

Vanek, J. (2017). *Using the PIAAC framework for problem solving in technology-rich environments to guide instruction: An introduction for adult educators.* Retrieved from

http://45f7d2w3u2c1bb3951cm1kxu.wpengine.netdna-cdn.com/wp-content/uploads/2017/05/PSTRE_Guide_Vanek_2017.pdf

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.

Wößmann, L. (2005). Educational production in East Asia: The impact of family background and schooling policies on student performance. *German Economic Review*, *6*(3), 331-353.

Appendix

Table A1. Difference in the percentages of each category of background variables between the whole sample and the sample with U02 response.

	Levels											
Variables	0	1	2	3	4	5	6	7	8	9	10	99
PAIDWORK		1.35	1.42	-0.62	-4.40							2.25
C_D05		-4.86	-1.44	4.08	0.04							2.18
D_Q04_T1		0.13	-1.17	-1.49	-1.81							4.34
D_Q08b		-1.22	0.11	-0.86	-0.51	-0.33						2.81
D_Q12b		-3.28	-0.03	-0.15								3.47
D_Q12c		-0.76	-1.97	-0.46	0.13							
F_Q05a		1.97	0.45	0.38	-2.88	-4.83						4.92
F_Q05b		0.75	-1.02	-1.52	-1.70	-1.38						4.87
F_Q07b		-0.83	-4.05									4.88
G_Q06		-2.61	-4.92	-0.02								7.54
ISCOSKIL4		-5.49	-0.85	0.61	0.73							5.00
EARNMTHALLDCL		-0.44	-0.99	-0.13	-0.74	0.46	-0.32	0.63	-0.19	-1.32	-1.27	4.33
LEARNATWORK_WLE_CA	0.68	0.75	0.09	-2.12	-1.25	-1.54						3.39
ICTWORK_WLE_CA	0.71	-1.79	-0.70	-2.18	-0.95	-2.47						7.38
INFLUENCE_WLE_CA	0.58	1.33	0.21	-2.04	-1.86	-2.88						4.68
NUMWORK_WLE_CA	1.89	-0.91	-1.03	-1.14	-1.32	-2.17						4.68
PLANNING_WLE_CA	1.59	-0.40	-1.29	0.02	-1.40	-3.19						4.68
READWORK_WLE_CA	0.36	0.80	0.78	-1.63	-2.94	-2.05						4.68
TASKDISC_WLE_CA	0.75	0.10	-1.85	-3.27	0.03	-0.48						4.72
WRITWORK_WLE_CA	1.75	0.75	-1.38	-1.65	-1.15	-2.98						4.68
AGEG10LFS		-2.18	-1.34	-1.85	2.08	3.30						
PARED		4.18	-3.25	-4.49								3.57
NATIVELANG	3.05	-5.25										2.20
EDCAT6		3.22	-0.04	-0.79	-0.69	-2.24	-1.63					2.17
GENDER_R		-0.51	0.51									
J_Q08		3.48	-1.41	-1.00	-2.16	-0.97	-0.24					2.29

Table A2. Top five features of action sequences selected for the groups with lowest 20% or all zero response on INFLUENCE_WLE_CA.

Group	N-gram	Action sequences	Chi-square
lowest 20% on INFLUENCE_WLE_CA	Unigram	COPY	6.46
		REPLY	5.01
		SEARCH	4.56
		HELP	3.38
		PASTE	3.18
	Bigram	SUBMIT_RESERVATION_SUCCESS HISTORY_MEETINGROOMS	11.12
		SUBMIT_UNFILLED ENVIRONMENT_MC	8.10
		MAIL_VIEWED_3 FOLDER_VIEWED	7.31
		HISTORY_VIEWCALENDAR COMBOBOX_ROOM	6.77
		ENVIRONMENT_WB SUBMIT_UNFILLED	6.56
	Trigram	COMBOBOX_DEPT ENVIRONMENT_MC ENVIRONMENT_WB	12.09
		SUBMIT_RESERVATION_SUCCESS HISTORY_MEETINGROOMS	9.62
		HISTORY_VIEWCALENDAR HISTORY_UNFILLED SUBMIT_UNFILLED	9.58
		ENVIRONMENT_MC COMBOBOX_DEPT SUBMIT_RESERVATION_SUCCESS	9.42
		HISTORY_MEETINGROOMS ENVIRONMENT_MC ENVIRONMENT_WB	9.20
		ENVIRONMENT_MC	
all zero response on INFLUENCE_WLE_CA	Unigram	MAIL_MOVE	17.83
		FOLDER_VIEWED	9.48
		CANCEL	7.97
		SUBMIT_RESERVATION_FAILURE	1.58
		HISTORY_HOME	0.54
	Bigram	SUBMIT_RESERVATION_SUCCESS SUBMIT_RESERVATION_FAILURE	29.75
		MAIL_MOVE FOLDER_VIEWED	23.45
		ENVIRONMENT_WB HISTORY_HOME	21.24
		MAIL_VIEWED_3 MAIL_MOVE	20.17
		MAIL_VIEWED_2 MAIL_MOVE	18.77
	Trigram	FOLDER_VIEWED MAIL_VIEWED_1 MAIL_MOVE	30.13
		MAIL_MOVE FOLDER_VIEWED FOLDER_VIEWED	23.99
		ENVIRONMENT_WB COMBOBOX_ROOM COMBOBOX_START_TIME	23.23
		COMBOBOX_DEPT SUBMIT_RESERVATION_SUCCESS SUBMIT_RESERVATION_FAILURE	22.58
		ENVIRONMENT_MC NEXT_INQUIRY CANCEL	22.00

Table A3. Top five features of action sequences selected for the groups with lowest 20% or all zero response on READWORK_WLE_CA.

Group	N-gram	Action sequences	Chi-square
lowest 20% on READWORK_WLE_CA	Unigram	HELP	4.29
		CANCEL	3.69
		COPY	3.59
		HISTORY_VIEWCALENDAR	3.28
		SEARCH	2.00
	Bigram	ENVIRONMENT_MC MAIL_VIEWED_3	9.10
		HISTORY_VIEWCALENDAR HISTORY_UNFILLED	5.45
		HISTORY_RESERVATION COMBOBOX_ROOM	5.41
		ENVIRONMENT_WB HISTORY_VIEWCALENDAR	5.15
		CHANGE_RESERVATION CHANGE_RESERVATION	4.58
	Trigram	ENVIRONMENT_MC ENVIRONMENT_WB HISTORY_VIEWCALENDAR	6.70
		ENVIRONMENT_WB HISTORY_RESERVATION HISTORY_VIEWCALENDAR	6.59
		ENVIRONMENT_WB HISTORY_RESERVATION COMBOBOX_ROOM	6.27
		ENVIRONMENT_MC MAIL_VIEWED_3 ENVIRONMENT_WB	6.06
		START MAIL_VIEWED_1 MAIL_VIEWED_2	5.64
all zero response on READWORK_WLE_CA	Unigram	MAIL_MOVE	28.76
		FOLDER_VIEWED	4.66
		SUBMIT_RESERVATION_FAILURE	4.58
		HISTORY_HOME	4.45
		REPLY	1.07
	Bigram	HISTORY_UNFILLED HISTORY_UNFILLED	35.28
		MAIL_VIEWED_3 ENVIRONMENT_MC	26.10
		FOLDER_VIEWED MAIL_MOVE	22.78
		MAIL_MOVE FOLDER_VIEWED	22.25
		HISTORY_HOME HISTORY_MEETINGROOMS	20.55
	Trigram	ENVIRONMENT_WB COMBOBOX_END_TIME COMBOBOX_DEPT	47.51
		FOLDER_VIEWED MAIL_MOVE FOLDER_VIEWED	35.91
		MAIL_VIEWED_3 MAIL_MOVE MAIL_VIEWED_2	34.97
		SUBMIT_RESERVATION_FAILURE COMBOBOX_DEPT SUBMIT_RESERVATION_FAILURE	33.24
		COMBOBOX_ROOM SUBMIT_RESERVATION_FAILURE COMBOBOX_DEPT	33.24

Table A4. Top five features of action sequences selected for the groups with 60%-80% or all zero response on TASKDISC_WLE_CA.

Group	N-gram	Action sequences	Chi-square
60%-80% on TASKDISC_WLE_CA	Unigram	FOLDER	2.17
		HISTORY_HOME	1.87
		KEYPRESS	1.52
		HELP	0.88
		MAIL_MOVE	0.63
	Bigram	HISTORY_RESERVATION HISTORY_VIEWCALENDAR	2.58
		COMBOBOX_END_TIME ENVIRONMENT_MC	2.47
		MAIL_VIEWED_4 MAIL_VIEWED_3	2.13
		ENVIRONMENT_WB ENVIRONMENT_WP	1.94
		ENVIRONMENT_WP ENVIRONMENT_MC	1.91
	Trigram	COMBOBOX_START_TIME COMBOBOX_END_TIME ENVIRONMENT_MC	2.16
		ENVIRONMENT_MC ENVIRONMENT_WB COMBOBOX_DEPT	2.15
		COMBOBOX_END_TIME ENVIRONMENT_MC ENVIRONMENT_WB	1.74
		ENVIRONMENT_WB COMBOBOX_DEPT SUBMIT_RESERVATION_SUCCESS	1.60
		ENVIRONMENT_WB HISTORY_RESERVATION HISTORY_VIEWCALENDAR	1.44
all zero response on TASKDISC_WLE_CA	Unigram	SUBMIT_RESERVATION_FAILURE	3.94
		COMBOBOX_DEPT	0.69
		CHANGE_RESERVATION	0.41
		ENVIRONMENT_WP	0.19
		COMBOBOX_END_TIME	0.18
	Bigram	CANCEL HISTORY_MEETINGROOMS	78.31
		ENVIRONMENT_WB COMBOBOX_END_TIME	75.33
		COMBOBOX_DEPT COMBOBOX_DEPT	34.89
		COMBOBOX_ROOM ENVIRONMENT_MC	30.89
		START NEXT_INQUIRY	30.31
	Trigram	ENVIRONMENT_WB COMBOBOX_END_TIME COMBOBOX_DEPT	138.78
		MAIL_VIEWED_4 MAIL_VIEWED_4 NEXT_INQUIRY	96.79
		ENVIRONMENT_WP ENVIRONMENT_WB COMBOBOX_END_TIME	95.64
		HISTORY_RESERVATION COMBOBOX_START_TIME ENVIRONMENT_MC	95.64
		MAIL_VIEWED_1 ENVIRONMENT_WP ENVIRONMENT_WP	95.64

Table A5. Top five features of action sequences selected for the groups with 60%-80% or all zero response on WRITWORK_WLE_CA.

Group	N-gram	Action sequences	Chi-square
60%-80% on WRITWORK_WLE_CA	Unigram	CANCEL	12.84
		SORT	12.34
		BOOKMARK	5.55
		SUBMIT_UNFILLED	5.41
		SEARCH	4.88
	Bigram	SUBMIT_RESERVATION_SUCCESS COMBOBOX_ROOM	16.28
		COMBOBOX_END_TIME COMBOBOX_DEPT	15.44
		ENVIRONMENT_WB SUBMIT_UNFILLED	14.59
		COMBOBOX_END_TIME SUBMIT_RESERVATION_SUCCESS	13.67
		CHANGE_RESERVATION ENVIRONMENT_MC	12.38
	Trigram	ENVIRONMENT_WB HISTORY_VIEWCALENDAR HISTORY_RESERVATION	17.00
		HISTORY_VIEWCALENDAR ENVIRONMENT_MC NEXT_INQUIRY	16.99
		MAIL_VIEWED_4 ENVIRONMENT_WB HISTORY_VIEWCALENDAR	16.20
		MAIL_VIEWED_1 MAIL_VIEWED_1 ENVIRONMENT_WB	15.83
		ENVIRONMENT_WB ENVIRONMENT_MC FOLDER_VIEWED	15.70
all zero response on WRITWORK_WLE_CA	Unigram	MAIL_MOVE	121.84
		FOLDER_VIEWED	12.97
		COPY	11.99
		PASTE	2.66
		MAIL_VIEWED_2	1.97
	Bigram	MAIL_MOVE FOLDER_VIEWED	91.26
		FOLDER_VIEWED MAIL_MOVE	89.43
		MAIL_MOVE MAIL_VIEWED_1	74.40
		MAIL_VIEWED_4 MAIL_MOVE	72.36
		MAIL_MOVE MAIL_VIEWED_3	69.92
	Trigram	FOLDER_VIEWED MAIL_MOVE MAIL_VIEWED_1	88.62
		MAIL_VIEWED_2 MAIL_VIEWED_2 MAIL_VIEWED_2	81.90
		MAIL_MOVE FOLDER_VIEWED MAIL_MOVE	76.13
		MAIL_MOVE MAIL_VIEWED_1 MAIL_MOVE	68.14
		MAIL_MOVE MAIL_VIEWED_4 MAIL_MOVE	59.57

Table A6. Top five features of action sequences selected for the groups with age 45-54 and age 24 or less.

Group	N-gram	Action sequences	Chi-square
age 45-54	Unigram	HISTORY_HOME	71.57
		HELP	51.53
		FOLDER	44.96
		REPLY	31.92
		COPY	11.36
	Bigram	FOLDER_VIEWED ENVIRONMENT_MC	61.98
		HISTORY_HOME HISTORY_HOME	56.88
		HELP FOLDER_VIEWED	46.19
		FOLDER FOLDER_VIEWED	42.16
		START NEXT_INQUIRY	41.40
	Trigram	FOLDER_VIEWED FOLDER FOLDER_VIEWED	49.10
		HISTORY_HOME HISTORY_HOME ENVIRONMENT_MC	46.57
		MAIL_VIEWED_3 MAIL_VIEWED_4 ENVIRONMENT_WB	42.33
		FOLDER_VIEWED FOLDER_VIEWED ENVIRONMENT_MC	41.07
		MAIL_VIEWED_1 MAIL_VIEWED_3 MAIL_VIEWED_4	40.06
age 24 or less	Unigram	MAIL_MOVE	32.37
		HISTORY_UNFILLED	11.04
		CHANGE_RESERVATION	6.10
		SUBMIT_RESERVATION_FAILURE	3.41
		SUBMIT_UNFILLED	2.96
	Bigram	MAIL_VIEWED_3 MAIL_MOVE	56.40
		MAIL_VIEWED_1 MAIL_MOVE	41.25
		MAIL_MOVE MAIL_VIEWED_2	40.61
		MAIL_MOVE MAIL_VIEWED_4	38.06
		MAIL_VIEWED_4 MAIL_MOVE	37.37
	Trigram	START MAIL_VIEWED_1 MAIL_MOVE	54.83
		MAIL_VIEWED_3 MAIL_MOVE FOLDER_VIEWED	51.72
		MAIL_MOVE MAIL_VIEWED_4 MAIL_MOVE	43.26
		MAIL_VIEWED_2 MAIL_VIEWED_2 MAIL_VIEWED_2	43.11
		MAIL_VIEWED_1 MAIL_MOVE FOLDER_VIEWED	42.37

Table A7. Top five features of action sequences selected for the groups with age 55 or more and age 24 or less.

Group	N-gram	Action sequences	Chi-square
age 55 or more	Unigram	HISTORY_HOME	95.57
		HELP	69.64
		REPLY	56.68
		FOLDER	53.56
		SORT	22.21
	Bigram	HISTORY_HOME HISTORY_HOME	103.26
		FOLDER_VIEWED ENVIRONMENT_MC	84.40
		FOLDER_VIEWED FOLDER	52.83
		HELP FOLDER_VIEWED	48.34
		FOLDER_VIEWED REPLY	39.27
	Trigram	HISTORY_HOME HISTORY_HOME HISTORY_HOME	75.62
		FOLDER_VIEWED FOLDER_VIEWED ENVIRONMENT_MC	60.11
		FOLDER_VIEWED FOLDER_VIEWED FOLDER	41.40
		FOLDER_VIEWED ENVIRONMENT_MC ENVIRONMENT_MC	39.70
		HELP FOLDER_VIEWED FOLDER_VIEWED	37.36
age 24 or less	Unigram	MAIL_MOVE	40.37
		PASTE	10.44
		COPY	6.60
		ENVIRONMENT_WB	4.88
		HISTORY_MEETINGROOMS	3.82
	Bigram	MAIL_MOVE MAIL_VIEWED_2	43.08
		FOLDER_VIEWED MAIL_MOVE	34.00
		MAIL_VIEWED_1 MAIL_MOVE	32.85
		MAIL_MOVE MAIL_VIEWED_3	31.82
		MAIL_MOVE FOLDER_VIEWED	29.20
	Trigram	MAIL_MOVE MAIL_VIEWED_2 FOLDER_VIEWED	34.83
		ENVIRONMENT_MC ENVIRONMENT_WB COMBOBOX_DEPT	33.78
		MAIL_VIEWED_2 FOLDER_VIEWED MAIL_MOVE	33.19
		MAIL_MOVE MAIL_VIEWED_1 MAIL_MOVE	32.86
		ENVIRONMENT_WB COMBOBOX_DEPT ENVIRONMENT_MC	28.72

Table A8. Top five features of action sequences selected for the groups with test language same as native language or not.

Group	N-gram	Action sequences	Chi-square
test language same as native language	Unigram	SORT	5.64
		PASTE	4.37
		HELP	3.00
		SUBMIT_UNFILLED	2.77
		FOLDER	2.29
	Bigram	HISTORY_VIEWCALENDAR ENVIRONMENT_WP	17.51
		SUBMIT_RESERVATION_FAILURE HISTORY_VIEWCALENDAR	10.59
		FOLDER FOLDER_VIEWED	8.49
		MAIL_VIEWED_3 NEXT_INQUIRY	8.43
		COMBOBOX_DEPT COMBOBOX_START_TIME	7.32
	Trigram	ENVIRONMENT_WB HISTORY_MEETINGROOMS ENVIRONMENT_MC	11.37
		FOLDER_VIEWED FOLDER_VIEWED ENVIRONMENT_WB	10.48
		ENVIRONMENT_WB HISTORY_VIEWCALENDAR HISTORY_UNFILLED	10.42
		HISTORY_UNFILLED ENVIRONMENT_MC ENVIRONMENT_WB	10.13
		HISTORY_VIEWCALENDAR ENVIRONMENT_WP ENVIRONMENT_WB	9.33
test language not the same as native language	Unigram	SEARCH	16.28
		BOOKMARK	2.29
		MAIL_MOVE	1.16
		HISTORY_MEETINGROOMS	0.67
		COMBOBOX_DEPT	0.29
	Bigram	NEXT_INQUIRY KEYPRESS	141.04
		COMBOBOX_ROOM HISTORY_HOME	75.69
		BOOKMARK HISTORY_VIEWCALENDAR	67.68
		HISTORY_RESERVATION BOOKMARK	67.68
		SEARCH KEYPRESS	49.98
	Trigram	HISTORY_HOME HISTORY_MEETINGROOMS HISTORY_VIEWCALENDAR	134.08
		SUBMIT_RESERVATION_FAILURE ENVIRONMENT_MC ENVIRONMENT_WP	132.41
		COMBOBOX_END_TIME COMBOBOX_ROOM ENVIRONMENT_MC	130.71
		FOLDER_VIEWED SEARCH SEARCH	129.22
		MAIL_VIEWED_2 ENVIRONMENT_WP ENVIRONMENT_WB	128.67

Table A9. Top five features of action sequences selected for upper secondary and the lowest education groups.

Group	N-gram	Action sequences	Chi-square
upper secondary	Unigram	SUBMIT_RESERVATION_FAILURE	6.88
		COMBOBOX_ROOM	5.08
		HISTORY_HOME	5.03
		COMBOBOX_START_TIME	3.49
		BOOKMARK	3.21
	Bigram	MAIL_VIEWED_4 MAIL_VIEWED_2	13.92
		FOLDER FOLDER_VIEWED	13.77
		HISTORY_HOME HISTORY_HOME	13.14
		FOLDER_VIEWED FOLDER	12.18
		ENVIRONMENT_WB ENVIRONMENT_MC	12.15
	Trigram	ENVIRONMENT_WB HISTORY_MEETINGROOMS ENVIRONMENT_MC	15.66
		ENVIRONMENT_MC MAIL_VIEWED_1 MAIL_VIEWED_2	15.18
		MAIL_VIEWED_1 MAIL_VIEWED_3 MAIL_VIEWED_4	14.87
		HISTORY_RESERVATION HISTORY_MEETINGROOMS HISTORY_VIEWCALENDAR	14.28
		FOLDER_VIEWED FOLDER FOLDER_VIEWED	13.07
lower secondary or less	Unigram	SEARCH	32.49
		MAIL_MOVE	30.71
		CANCEL	14.37
		COPY	12.53
		PASTE	11.72
	Bigram	SEARCH FOLDER_VIEWED	54.79
		HISTORY_UNFILLED HISTORY_UNFILLED	54.12
		CANCEL MAIL_MOVE	53.27
		CHANGE_RESERVATION ENVIRONMENT_WP	52.96
		MAIL_MOVE FOLDER	51.19
	Trigram	FOLDER_VIEWED MAIL_MOVE MAIL_VIEWED_3	74.80
		MAIL_VIEWED_4 FOLDER_VIEWED MAIL_VIEWED_4	72.11
		MAIL_MOVE MAIL_VIEWED_3 MAIL_MOVE	55.93
		HISTORY_UNFILLED HISTORY_RESERVATION HISTORY_UNFILLED	52.49
		MAIL_VIEWED_4 FOLDER_VIEWED MAIL_VIEWED_1	52.28

Table A10. Top five features of action sequences selected for post-secondary, non-tertiary and the lowest education groups.

Group	N-gram	Action sequences	Chi-square
post-secondary, non-tertiary	Unigram	HISTORY_HOME	19.52
		SORT	8.20
		SUBMIT_RESERVATION_FAILURE	3.59
		BOOKMARK	3.29
		COMBOBOX_ROOM	1.85
	Bigram	HISTORY_HOME HISTORY_HOME	26.85
		ENVIRONMENT_WB ENVIRONMENT_MC	13.68
		SUBMIT_RESERVATION_SUCCESS HISTORY_VIEWCALENDAR	11.67
		FOLDER_VIEWED REPLY	11.06
		ENVIRONMENT_MC MAIL_VIEWED_1	10.32
	Trigram	ENVIRONMENT_MC MAIL_VIEWED_1 MAIL_VIEWED_2	18.60
		ENVIRONMENT_WB ENVIRONMENT_WB ENVIRONMENT_MC	18.46
		ENVIRONMENT_WB ENVIRONMENT_MC FOLDER_VIEWED	16.61
		ENVIRONMENT_WB HISTORY_MEETINGROOMS ENVIRONMENT_MC	16.19
		MAIL_VIEWED_1 MAIL_VIEWED_3 MAIL_VIEWED_4	16.03
lower secondary or less	Unigram	SEARCH	23.04
		MAIL_MOVE	19.31
		PASTE	6.78
		HISTORY_UNFILLED	2.58
		FOLDER_VIEWED	2.25
	Bigram	MAIL_VIEWED_2 MAIL_MOVE	32.76
		MAIL_MOVE MAIL_VIEWED_1	25.84
		MAIL_VIEWED_3 MAIL_MOVE	25.74
		MAIL_MOVE MAIL_VIEWED_2	25.42
		MAIL_VIEWED_1 MAIL_MOVE	21.10
	Trigram	MAIL_MOVE MAIL_VIEWED_2 MAIL_MOVE	32.53
		FOLDER_VIEWED MAIL_MOVE MAIL_VIEWED_2	26.88
		MAIL_VIEWED_1 MAIL_MOVE MAIL_VIEWED_1	23.64
		MAIL_VIEWED_2 MAIL_MOVE FOLDER_VIEWED	23.23
		MAIL_VIEWED_3 MAIL_MOVE FOLDER_VIEWED	21.26

Table A11. Top five features of action sequences selected for tertiary – professional degree and the lowest education groups.

Group	N-gram	Action sequences	Chi-square
tertiary – professional degree	Unigram	COMBOBOX_ROOM	4.48
		SUBMIT_RESERVATION_SUCCESS	4.02
		COMBOBOX_DEPT	3.66
		COMBOBOX_START_TIME	3.65
		COMBOBOX_END_TIME	3.26
	Bigram	ENVIRONMENT_WB ENVIRONMENT_MC	13.20
		MAIL_MOVE ENVIRONMENT_WP	11.95
		ENVIRONMENT_WB ENVIRONMENT_WB	10.87
		ENVIRONMENT_MC MAIL_VIEWED_1	10.34
		SUBMIT_RESERVATION_FAILURE HISTORY_MEETINGROOMS	9.98
	Trigram	ENVIRONMENT_WB ENVIRONMENT_WB ENVIRONMENT_MC	19.62
		MAIL_VIEWED_1 MAIL_VIEWED_4 ENVIRONMENT_WB	18.65
		ENVIRONMENT_MC MAIL_VIEWED_1 MAIL_VIEWED_2	18.29
		ENVIRONMENT_WB HISTORY_MEETINGROOMS ENVIRONMENT_MC	17.63
		MAIL_VIEWED_3 MAIL_VIEWED_1 MAIL_VIEWED_4	15.67
lower secondary or less	Unigram	COPY	30.70
		MAIL_MOVE	28.30
		FOLDER	22.22
		PASTE	20.33
		FOLDER_VIEWED	12.20
	Bigram	MAIL_VIEWED_2 MAIL_MOVE	31.13
		MAIL_MOVE MAIL_VIEWED_3	28.74
		FOLDER_VIEWED MAIL_MOVE	24.85
		MAIL_MOVE MAIL_VIEWED_1	22.55
		MAIL_MOVE MAIL_VIEWED_2	22.37
	Trigram	MAIL_MOVE MAIL_VIEWED_2 MAIL_MOVE	38.57
		FOLDER_VIEWED MAIL_MOVE MAIL_VIEWED_3	38.13
		MAIL_MOVE MAIL_VIEWED_1 MAIL_MOVE	30.98
		MAIL_VIEWED_2 MAIL_MOVE FOLDER_VIEWED	29.04
		MAIL_VIEWED_1 MAIL_MOVE MAIL_VIEWED_1	28.07

Table A12. Top five features of action sequences selected for tertiary – bachelor's degree and the lowest education groups.

Group	N-gram	Action sequences	Chi-square
tertiary – bachelor's degree	Unigram	REPLY	7.10
		COMBOBOX_ROOM	6.50
		COMBOBOX_START_TIME	5.30
		SUBMIT_RESERVATION_SUCCESS	5.14
		COMBOBOX_DEPT	4.76
	Bigram	MAIL_VIEWED_4 MAIL_VIEWED_1	16.49
		MAIL_VIEWED_1 MAIL_VIEWED_3	16.24
		ENVIRONMENT_WB ENVIRONMENT_MC	16.13
		HISTORY_MEETINGROOMS ENVIRONMENT_MC	14.32
		MAIL_VIEWED_4 MAIL_VIEWED_2	13.87
	Trigram	HISTORY_RESERVATION HISTORY_MEETINGROOMS HISTORY_VIEWCALENDAR	20.75
		ENVIRONMENT_WB HISTORY_MEETINGROOMS ENVIRONMENT_MC	20.25
		ENVIRONMENT_MC MAIL_VIEWED_1 MAIL_VIEWED_2	17.86
		MAIL_VIEWED_1 MAIL_VIEWED_3 MAIL_VIEWED_4	16.25
		MAIL_VIEWED_4 MAIL_VIEWED_1 ENVIRONMENT_WB	15.25
lower secondary or less	Unigram	MAIL_MOVE	109.87
		PASTE	59.04
		FOLDER_VIEWED	14.04
		COPY	13.82
		SEARCH	8.05
	Bigram	MAIL_VIEWED_1 MAIL_MOVE	101.94
		FOLDER_VIEWED MAIL_MOVE	98.18
		MAIL_MOVE MAIL_VIEWED_2	90.67
		MAIL_MOVE MAIL_VIEWED_3	90.67
		MAIL_MOVE FOLDER_VIEWED	83.84
	Trigram	FOLDER_VIEWED MAIL_MOVE MAIL_VIEWED_2	103.42
		MAIL_MOVE FOLDER_VIEWED MAIL_MOVE	93.20
		MAIL_MOVE MAIL_VIEWED_3 FOLDER_VIEWED	88.33
		MAIL_VIEWED_3 FOLDER_VIEWED MAIL_MOVE	88.33
		MAIL_VIEWED_3 MAIL_MOVE FOLDER_VIEWED	88.13