The 20th Annual MARC Conference
Machine Learning, Natural Language Processing, and Psychometrics

Leveraging Process Data in Large-Scale Educational Assessments with Sequence Mining

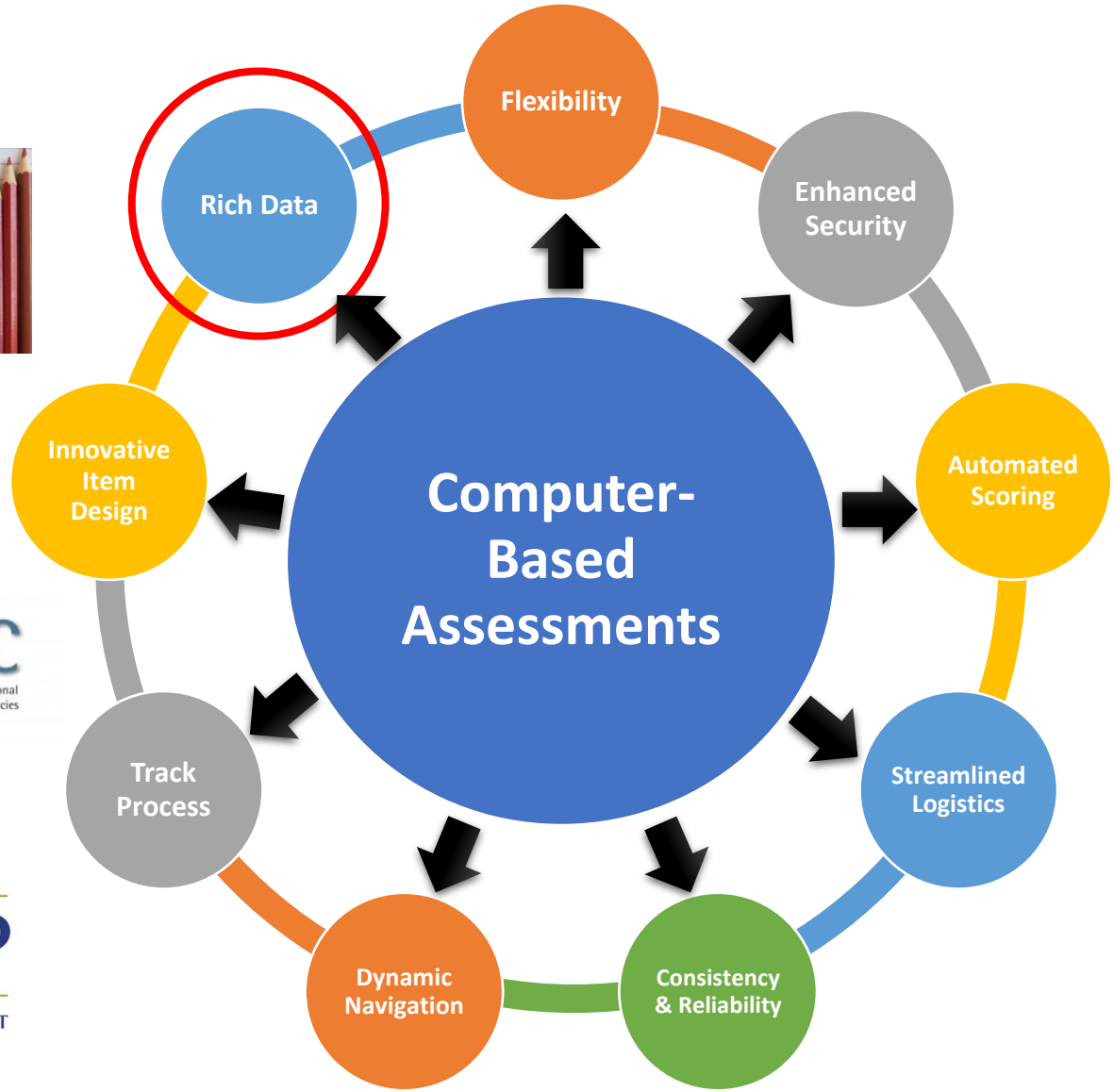
Qiwei He

Psychometric and Data Science Modeling
Educational Testing Service

November 4, 2022

Contents

Interactive Items and Process Data in Large-Scale Assessments



PIAAC
Programme for the International Assessment of Adult Competencies

NAEP
NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS

IEA ICILS

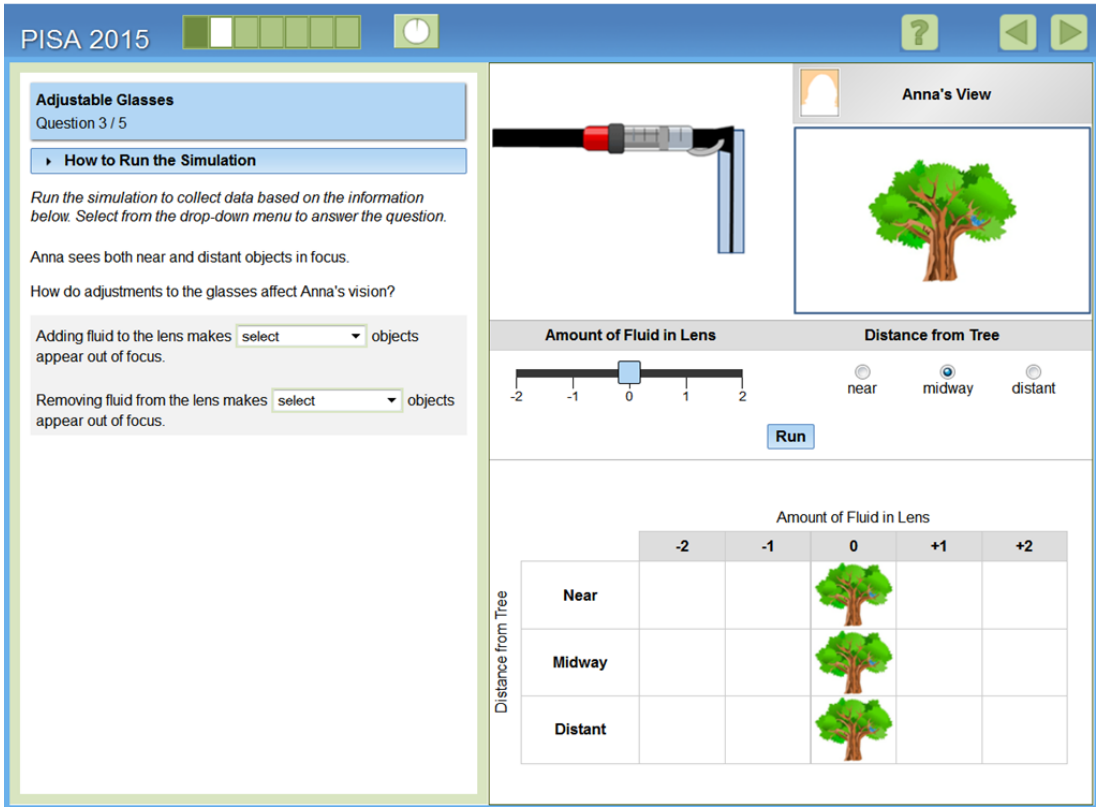
IEA TIMSS

IEA PIRLS

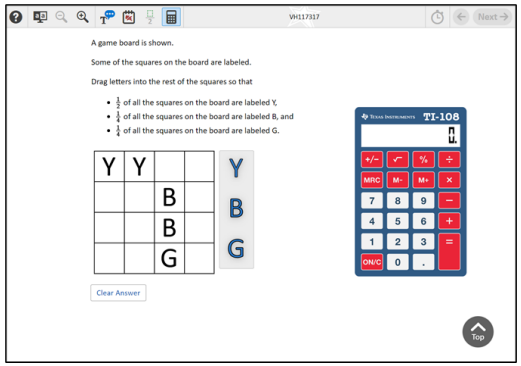
ETS

Mirroring Real-Life Behavior with Interactive Items

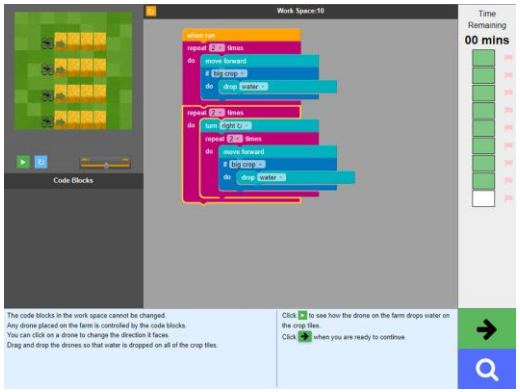
- Interactive items are becoming more widely used than in large-scale assessments.
- Recorded time-stamped action sequences contain information on type and order of performed actions as well as the time required on each action.



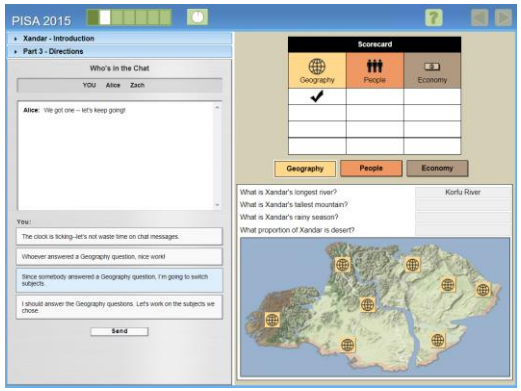
PISA 2015 Science Inquiry Interactive Item (field trial example item)



NAEP 2017 Grade 4 Interactive Math Example Item



ICILS2018 Computational Thinking Interactive Program Coding Example Item



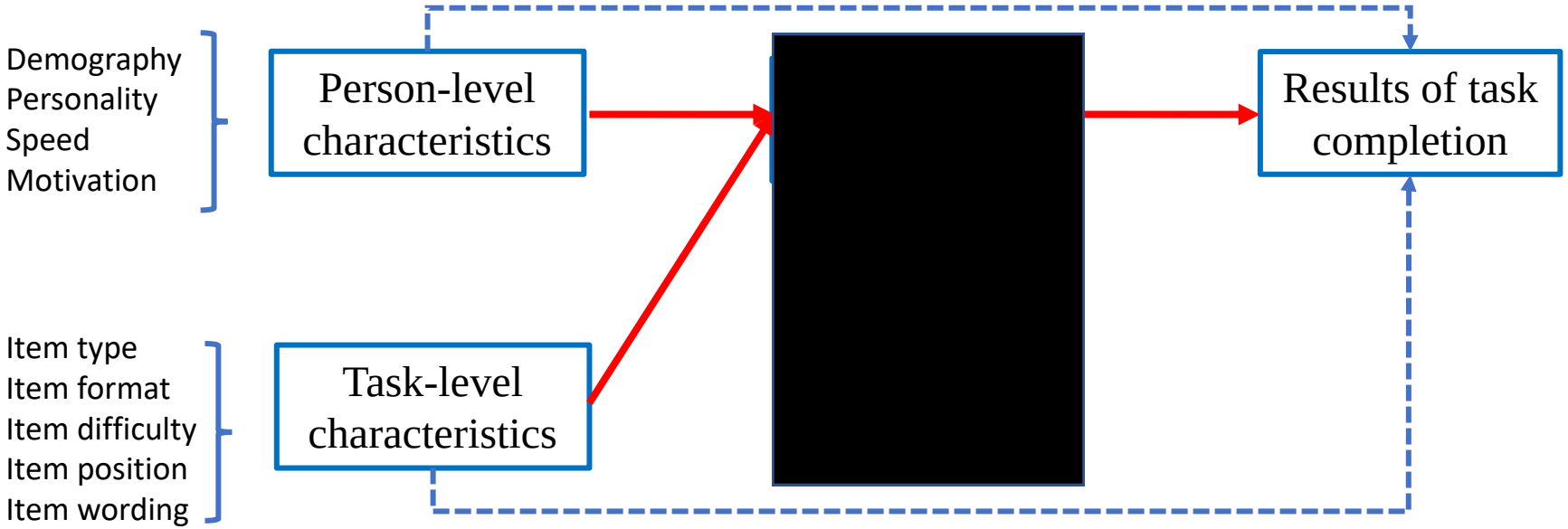
PISA 2015 Collaborative Problem-Solving Example Item

Examples from Different Subjects

Why Does Response Process Matter?

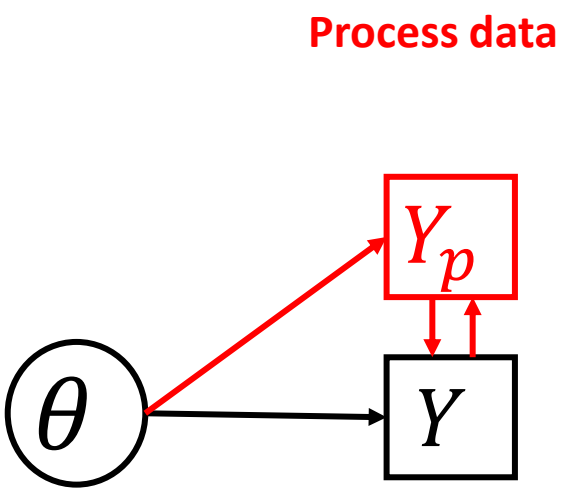
- Response process brings additional information and be helpful for aspects below:
  - **Task construction**
    - Investigating whether examinees interact with interactive tasks as intended
  - **Invariance explorations**
    - Defining invariance between group not only in terms of item difficulty but also in terms of the processes applied for solving a given task
  - **Richer description of performance**
    - Investigating not only whether examinees could solve a given task but also whether they did so efficiently and systematically
  - **Refining theories on response processes to interactive tasks**
  - **Designing tailored interventions**
    - Identifying subskills or meta competencies that examinees are lacking

A Typical Assessing Process



Heldt et al. (2020)

What is in the Black Box?



- An enormous amount of unstructured data is collected (even worse considering both performed actions and the associated time stamps).
- Potentially rich source of information on examinee behavior, leveraging its potential by meaningful aggregation is not trivial
- Calls for new methods with incorporation of data science and analytics techniques, to re-shape the future of measurement.

Sequence-Based Methods in Process Data Analysis

Sequence Mining

- Sequential pattern mining is a topic of data mining concerned with finding statistically relevant patterns between data examples where the values are delivered in a sequence.

- Sequence mining techniques include building efficient databases and indexes for sequence information, extracting the frequently occurring patterns, comparing sequences for similarity, and recovering missing sequence members.

Dong & Pei, 2007

Sequence-Based Methods in Process Data Analysis

Text and "real" clickstreams as commonly encountered types of unstructured sequence data. Borrow from sequence mining, natural language processing (NLP), and machine learning

Mini-Sequences (n-grams)

- Disassemble long sequence into manageable short pieces extracted as features in prediction and clustering

Pairwise Sequence Similarity Measures

- Take the sequence as a whole and compute the distance between each pair to create new variables for prediction and clustering

Latent Sequence Modeling

- From observed sequences to derive the latent sequence transition stage and probability

Mini Sequences (n-grams)

I am happy to give a talk today.

unigrams | bigrams | trigrams

Action sequence: STRT, SS, SS_Type_FN, E, E_S, Next, Next_OK, END	
Unigrams (8)	"START", "SS", "SS_Type_FN", "E", "E_S", "Next", "Next_OK", "END"
Bigrams (7)	"START, SS", "SS, SS_Type_FN", "SS_Type_FN, E", "E, E_S", "E_S, Next", "Next, Next_OK", "Next_OK, END"
Trigram (6)	"START, SS, SS_Type_FN", "SS, SS_Type_FN, E", "SS_Type_FN, E, E_S", "E, E_S, Next", "E_S, Next, Next_OK", "Next, Next_OK, END"

Mini Sequences (n-grams)

- Using n-grams and robust feature selection method (e.g., χ^2 selection) to identify behavioral patterns that distinguish between groups (He & von Davier, 2015, 2016)
- Using machine learning techniques for selecting key action features for predicting group membership (e.g., correctness group) (Han et al., 2019; Salles et al., 2020; Stadler et al., 2019; Ulitzsch et al., 2022)
- Using event history models for identifying key features for predicting the timing and correctness of task completion (e.g., Chen et al., 2019; Han et al., 2019)
- Using dynamic action sequence to predict whether the problem-solving process is on track (Ulitzsch et al., 2022)

Cons: Breaking up action sequences into short n-gram sequences could exploit meaningful information such as dependencies between actions

Pairwise Sequence Similarity

- **Sequence similarity to be compared with pre-defined sequences or compute pairwise distance matrix**
  - Generating efficiency and similarity indicators with longest common subsequence (LCS) (He et al., 2019, 2021)
  - Determining the similarity between students' strategies with correct ones with edit distance (Hao et al., 2015)

- **Sequence similarity with incorporation of timing information**
  - Determining common behavioral patterns in terms of the performed actions and the time elapsed in between with LCS (Ulitzsch et al, 2021; Ulitzsch, He & Pohl, 2022)
  - Clustering navigation patterns with dynamic time warping method (He et al., 2022)

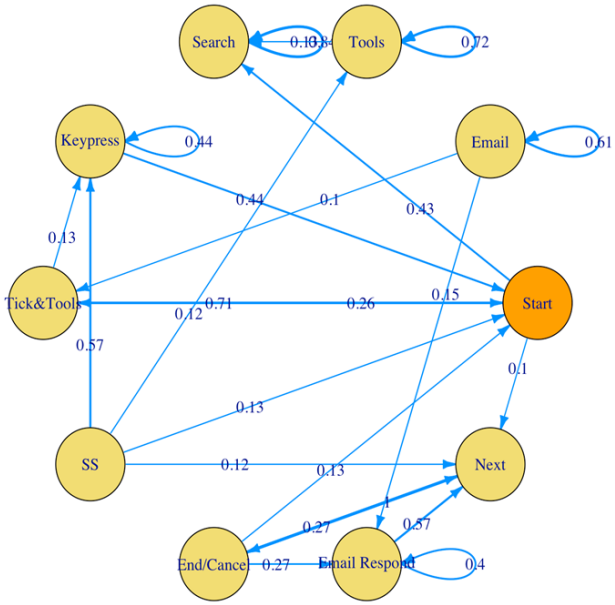
- **Dimensionality reduction and latent feature extraction**
  - Using distance measures and multidimensional scaling for parsimonious descriptions of behavioral patterns (Tang et al., 2019)
  - Deep learning from observed sequence to latent sequence features (Tang et al., 2020)

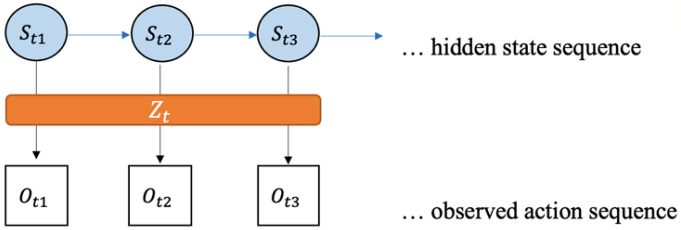
Latent Sequence Modeling

- The hidden Markov model (HMM) that has been widely used in NLP and speech recognition are also applicable in process data to extract latent states of the problem-solving process and investigate how these latent states are transitioned (Xiao et al., 2021; He et al., in preparation)

- Multigroup HMM provides the possibility to compare the transition probability between groups under the common latent sequence model (He et al., in preparation)

- Supports parsimonious modeling and description of behavioral pathways





(He, Bei, & Jiang, in preparation)

Case studies: Using Sequence Mining Techniques in Large-Scale Assessments

Clustering Reading Navigation Sequences with Dynamic Time Warping Method

In collaboration with

Dr. Francesca Borgonovi, OECD, University College London

Dr. Javier Suarez-Alvarez, University of Massachusetts

He, Q., Borgonovi, F., Suárez-Álvarez, J. (2022). Clustering Sequential Navigation Patterns in Multiple-Source Reading Tasks with Dynamic Time Warping Method. Journal of Computer-Assisted Learning. DOI: 10.1111/jcal.12748

Navigation is recognized as a key component of reading in the digital environment as readers "construct" their text through navigation and spend time retrieving information from eventually targeted texts.

Good readers tend to

- minimize their visits to irrelevant pages;
- locate necessary pages efficiently;
- choose strategies that are suited to the demands of individual tasks.

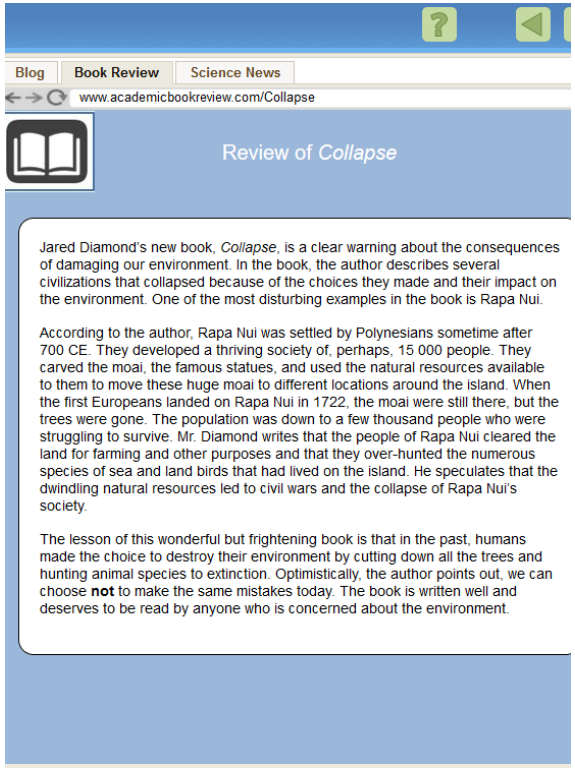
Dynamic Navigation

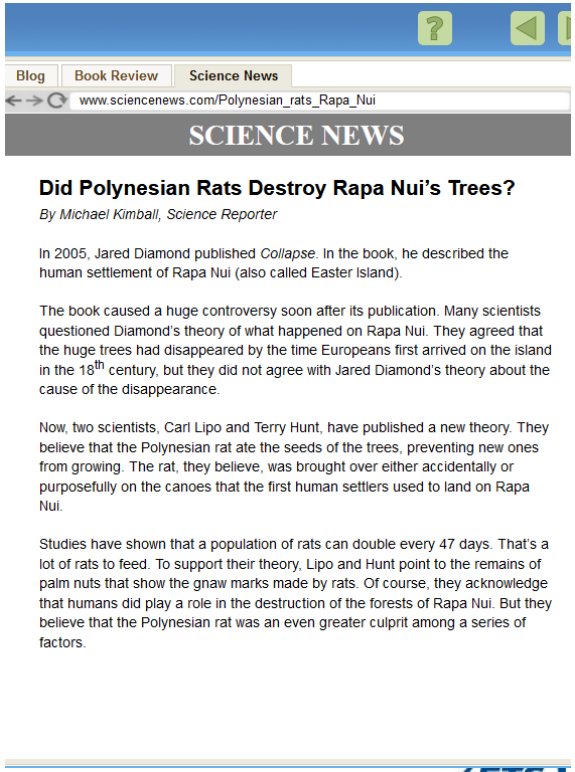
- The description of readers' navigation process demands tremendous support from log files.

- Data-driven investigations of how students transit pages in digital reading tasks and how much time they spend on each transition allow mapping sequences of navigation behaviors into students' navigation reading strategies.

Multiple Source and Dynamic Texts in PISA Reading

- Searching across multiple documents
- Integrating across texts for inferences
- Assessing the credibility of sources
- Handling conflicting information





Aims

- To identify students' navigation patterns in multiple-source reading tasks using a sequence clustering approach
- To examine how students' navigation patterns are associated with their reading performance and socio-demographic characteristics
- To showcase how the navigation sequences could be clustered on the similarity measure by dynamic time warping (DTW) methods

Dataset

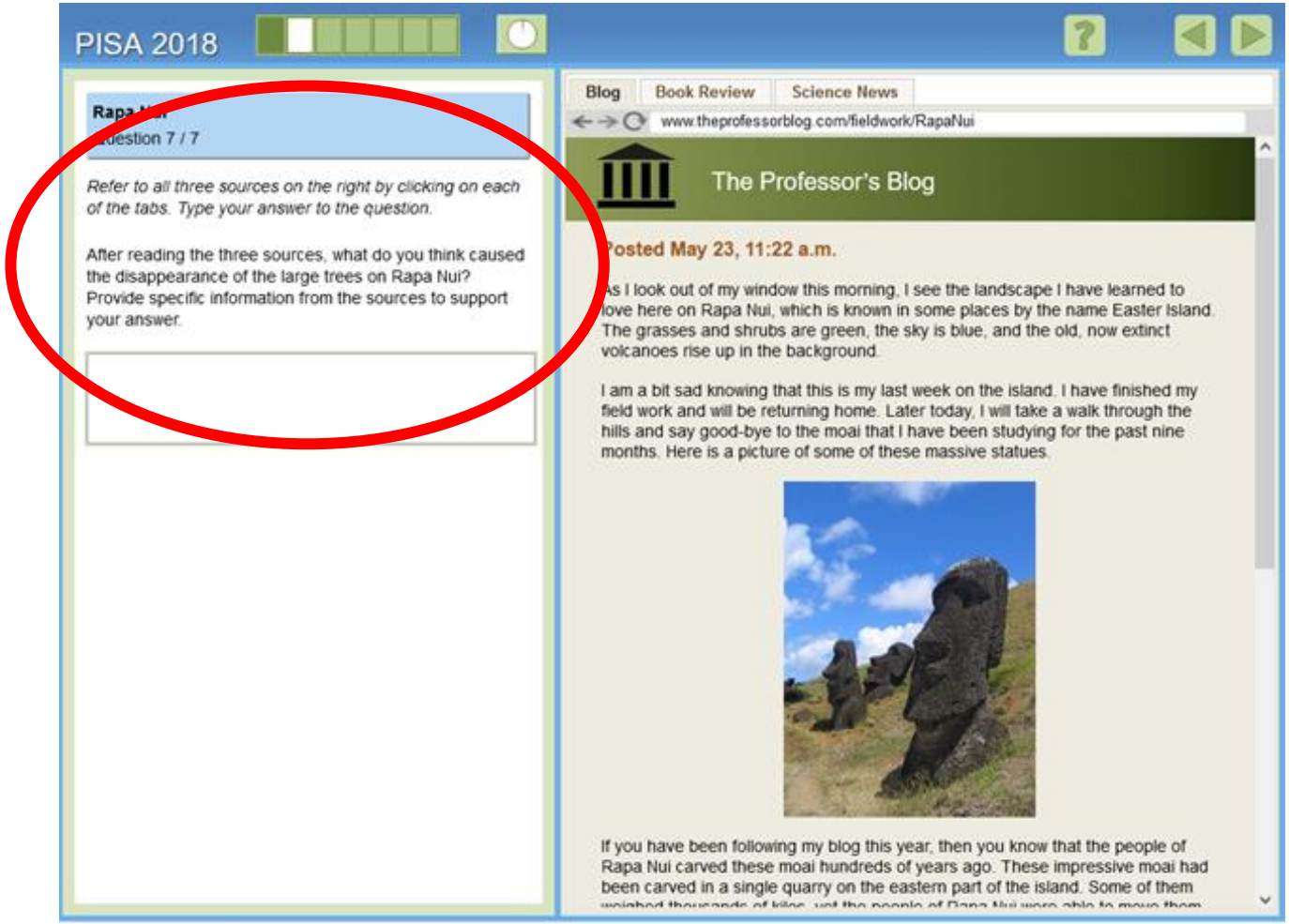
- A sample of 16,957 students from 69 countries participating in the PISA 2018

- Executed **at least one navigation activity** (i.e., visited at least one page beyond the default homepage) in one example task (CR551Q11) in the reading Rapa Nui unit

- Demographics:
  - 56.1% of the students were girls, which was a bit higher than the 52% of girls in the full sample in this reading unit.

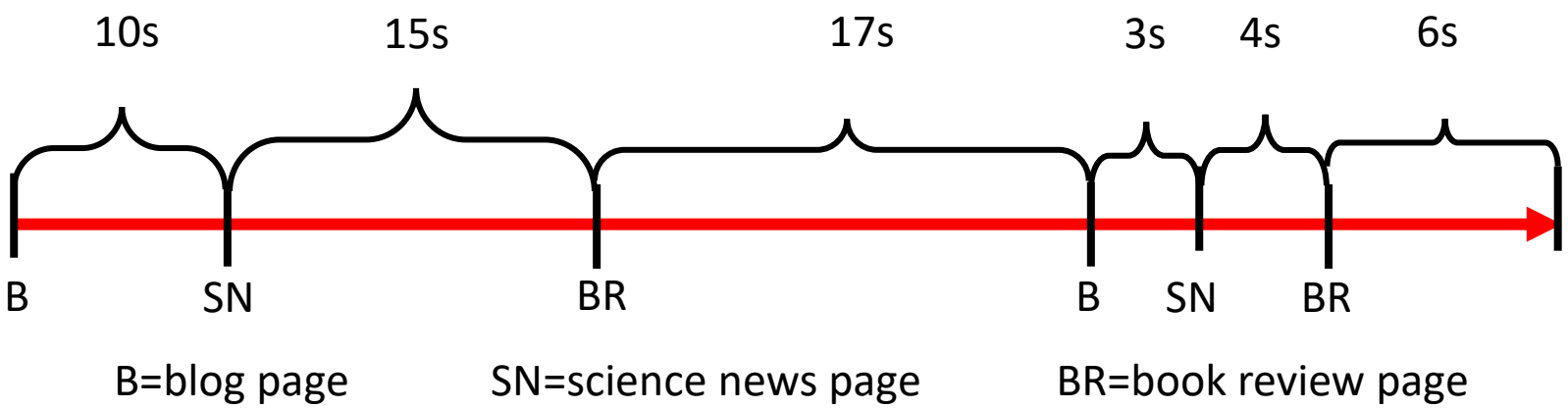
  - The index of economic, social, and cultural status (ESCS) was 0.245 in this subsample, which was also marginally higher than that of the full sample (-0.05) in this reading unit.

Instrument


PISA 2018
Question 7 / 7
Refer to all three sources on the right by clicking on each of the tabs. Type your answer to the question.
After reading the three sources, what do you think caused the disappearance of the large trees on Rapa Nui? Provide specific information from the sources to support your answer.
Blog
Book Review
Science News
www.theprofessorblog.com/fieldwork/RapaNui
The Professor's Blog
Posted May 23, 11:22 a.m.
As I look out of my window this morning, I see the landscape I have learned to love here on Rapa Nui, which is known in some places by the name Easter Island. The grasses and shrubs are green, the sky is blue, and the old, now extinct volcanoes rise up in the background.
I am a bit sad knowing that this is my last week on the island. I have finished my field work and will be returning home. Later today, I will take a walk through the hills and say good-bye to the moai that I have been studying for the past nine months. Here is a picture of some of these massive statues.
If you have been following my blog this year, then you know that the people of Rapa Nui carved these moai hundreds of years ago. These impressive moai had been carved in a single quarry on the eastern part of the island. Some of them

Clickstream Data

- Students' navigation sequences were characterized by two indicators:
  - Page sequence that tracks the page transition path
  - Time sequence that records the time duration on each visited page.



Page sequence: {B, SN, BR, B, SN, BR, END}

Time sequence: {10, 15, 17, 3, 4, 6}

Research Design

- **Step 1: Similarity measure**
  - Pairwise sequence distance computation with dynamic time warping (DTW) method
  - Execute by page sequence and time sequence respectively

- **Step 2: Sequence clustering (unsupervised machine learning)**
  - K-medoid partitioning clustering analyses on sequence distance matrix

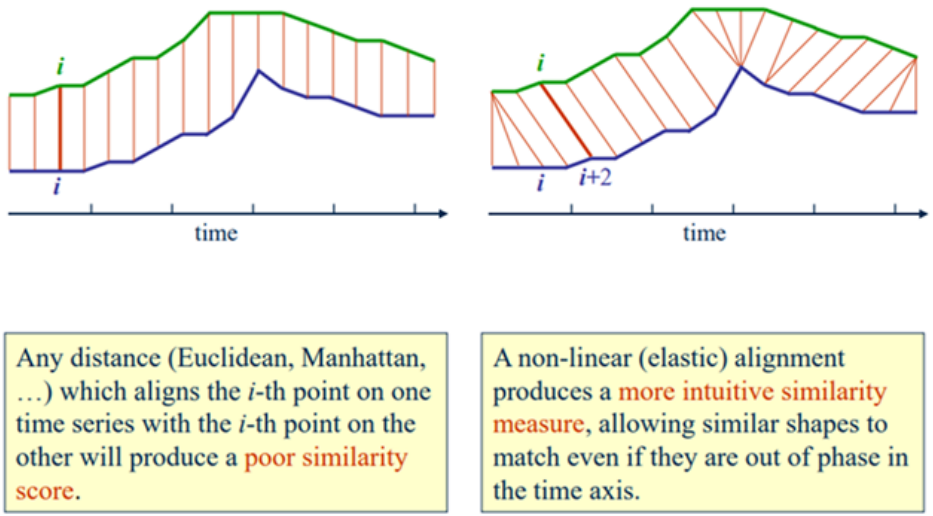
Step 1: Similarity Measure by DTW

- Dynamic time warping (Sakoe & Chiba, 1978) is a distance measure that searches the optimal warping path between two series.

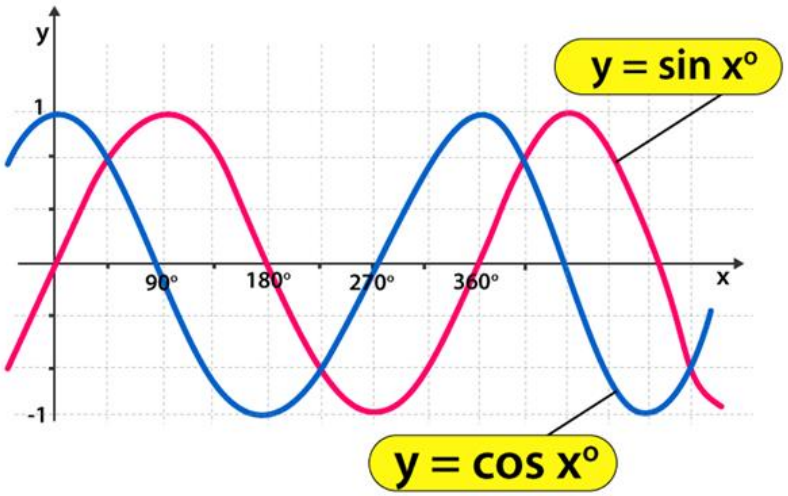
- Given sequences

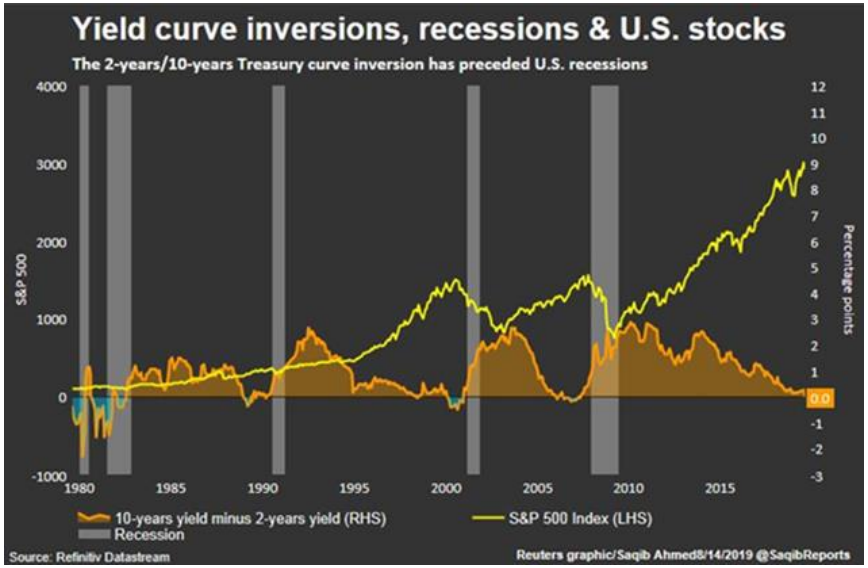
$X = \{x_1, x_2, \dots, x_n\}$ and

$Y = \{y_1, y_2, \dots, y_m\}$ with the same or different lengths, a warping path W is an alignment between X and Y, involving one-to-many mappings for each pair of elements.



Reasons to Use DTW





Dynamic Time Warping Algorithm

- The initial step of DTW algorithm is defined as:

$$DTW(i,j) = \begin{cases} \infty & if\ (i = 0\ or\ j = 0)\ and\ i \neq j \\ 0 & if\ i = j = 0 \end{cases}$$

- The recursive function of DTW is defined as

$$DTW(i,j) = min \begin{cases} DTW(i-1,j) + w_h C(i,j) \\ DTW(i,j-1) + w_v C(i,j) \\ DTW(i-1,j-1) + w_d C(i,j) \end{cases}$$

where (w_h, w_v, w_d) are weights for the horizontal, vertical and diagonal directions, respectively. $DTW(i,j)$ denotes the distance or cost between two sub-sequences $\{x_1, x_2, \dots, x_i\}$ and $\{y_1, y_2, \dots, j\}$, and $DTW(N,M)$ indicates the total cost of the optimal warping path.

A	**5**	10	6	3	1	2	4	3
	4	6	3	1	0	1	3	3
	3	3	1	0	1	1	2	4
	2	1	0	1	3	4	4	7
	1	0	1	3	6	8	9	13
		1	**2**	**3**	**4**	**3**	**2**	**5**
		B						

$A_i = \{1, 2, 3, 4, 5\}$

$B_j = \{1, 2, 3, 4, 3, 2, 5\}$

$dtw(i,j) = |A_i - B_j| + \min(D[i-1, j-1], D[i-1, j], D[i, j-1])$

Recode categorical variables to numeric variables (define B=1, BR=2, SN=3)

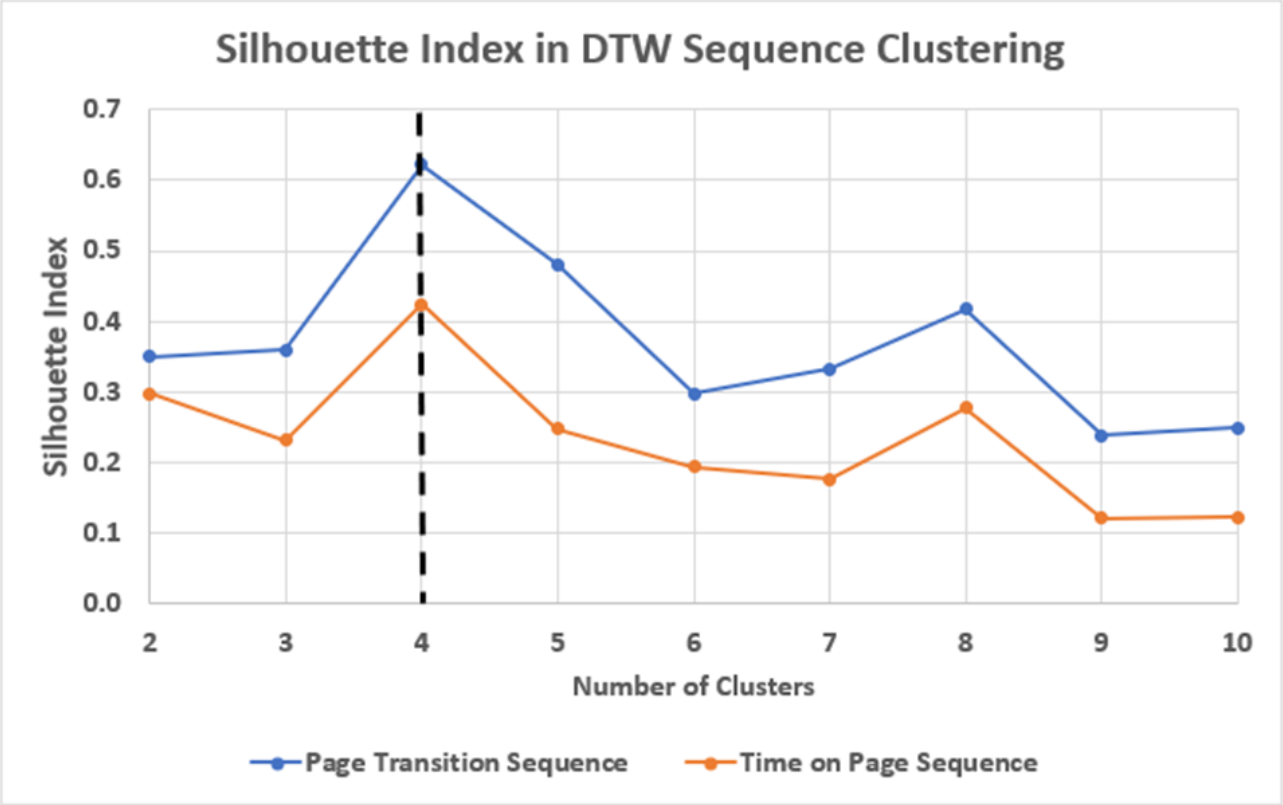
Step 2: Sequence Clustering

- K-medoids is a classical partitioning technique of clustering that splits the data set of n objects into k clusters, where the number k of clusters assumed known a priori.

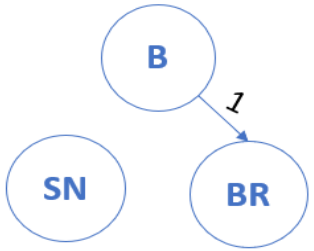
- In this study, we set k as 2 to 10 to as a priori and set the optimal number of clusters k with the silhouette index (Rousseeuw, 1987).

- The silhouette index ranges from −1 to +1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters.

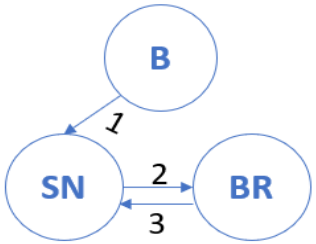
Silhouette Index for Optimal Number of Clusters



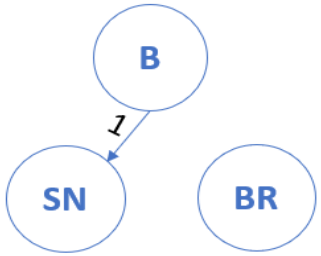
Result 1: Page Sequence Clustering



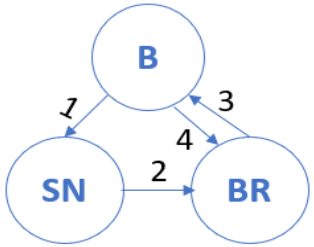
P1



P2



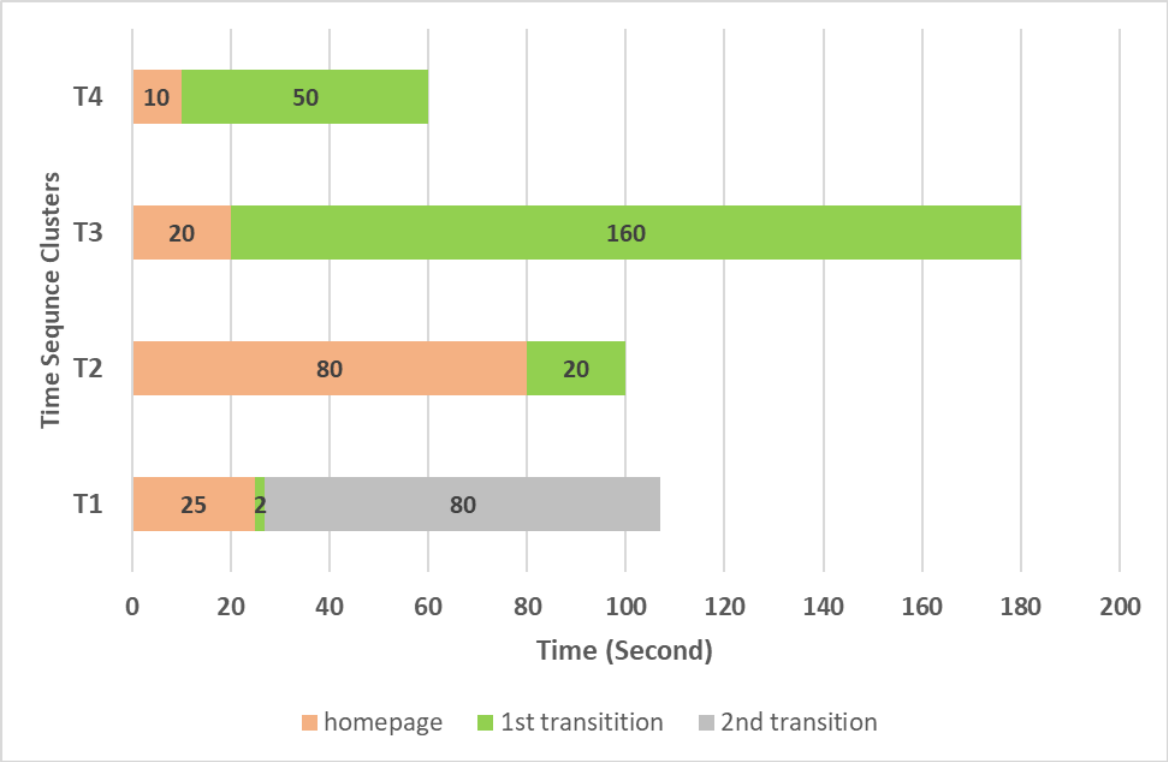
P3



P4

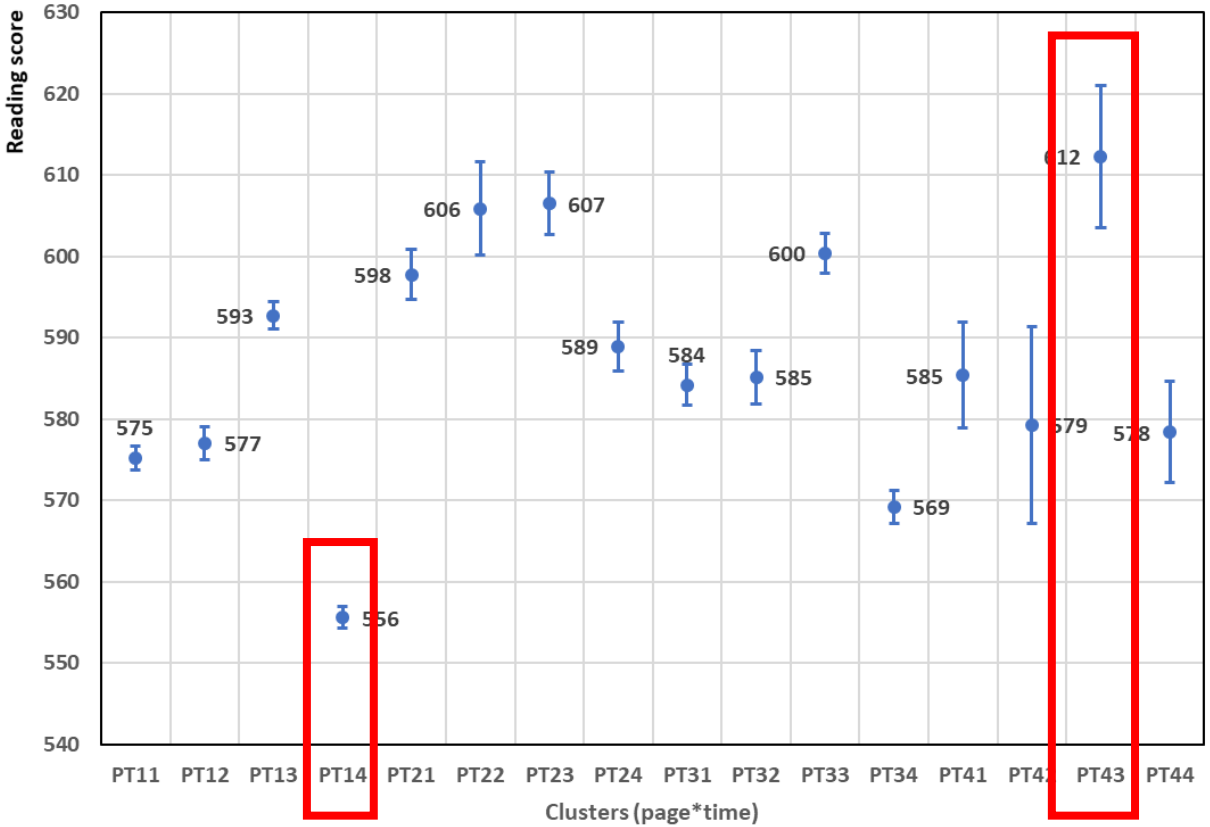
Clusters	Sample Size	Percentage	Reading Proficiency Score			Gender		Socio-Economic Index		
			Mean	S.D.	S.E.	Girl (%)	Boy (%)	Mean	S.D.	S.E.
P1	11091	65.4%	574.13	82.05	0.78	65.2%	66%	0.22	0.94	0.01
P2	1952	11.5%	599.06	76.97	1.74	11.4%	12%	0.29	0.90	0.02
P3	3438	20.8%	585.07	72.38	1.23	20.8%	20%	0.28	0.92	0.02
P4	476	2.6%	587.75	85.03	3.90	2.6%	3%	0.28	0.92	0.04

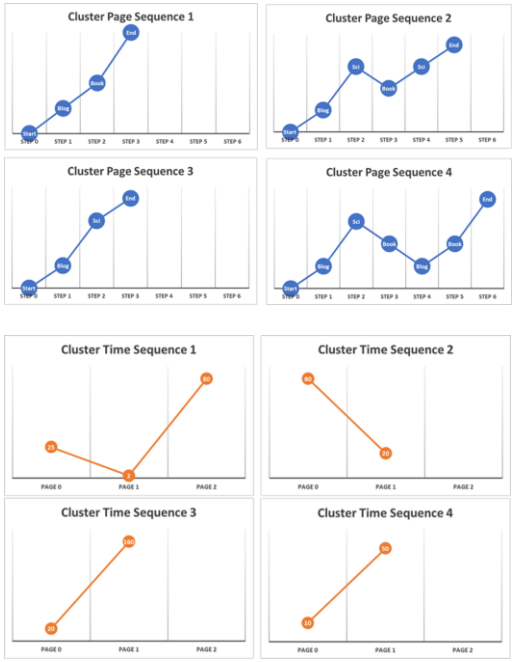
Result 2: Time Sequence Clustering



Clusters	Sample Size	Percentage	Reading Proficiency Score			Gender		Socio-Economic Index		
			Mean	S.D.	S.E.	Girl (%)	Boy (%)	Mean	S.D.	S.E.
T1	4651	27.4%	582.72	78.12	1.15	28.1%	27%	0.27	0.93	0.01
T2	2385	13.6%	583.75	79.43	1.63	13.6%	15%	0.33	0.92	0.02
T3	3675	23.5%	598.36	77.17	1.27	23.5%	19%	0.31	0.91	0.02
T4	6246	34.8%	564.66	80.83	1.02	34.8%	39%	0.16	0.94	0.01

Result 3: Association Between Navigation Patterns and Reading Skills

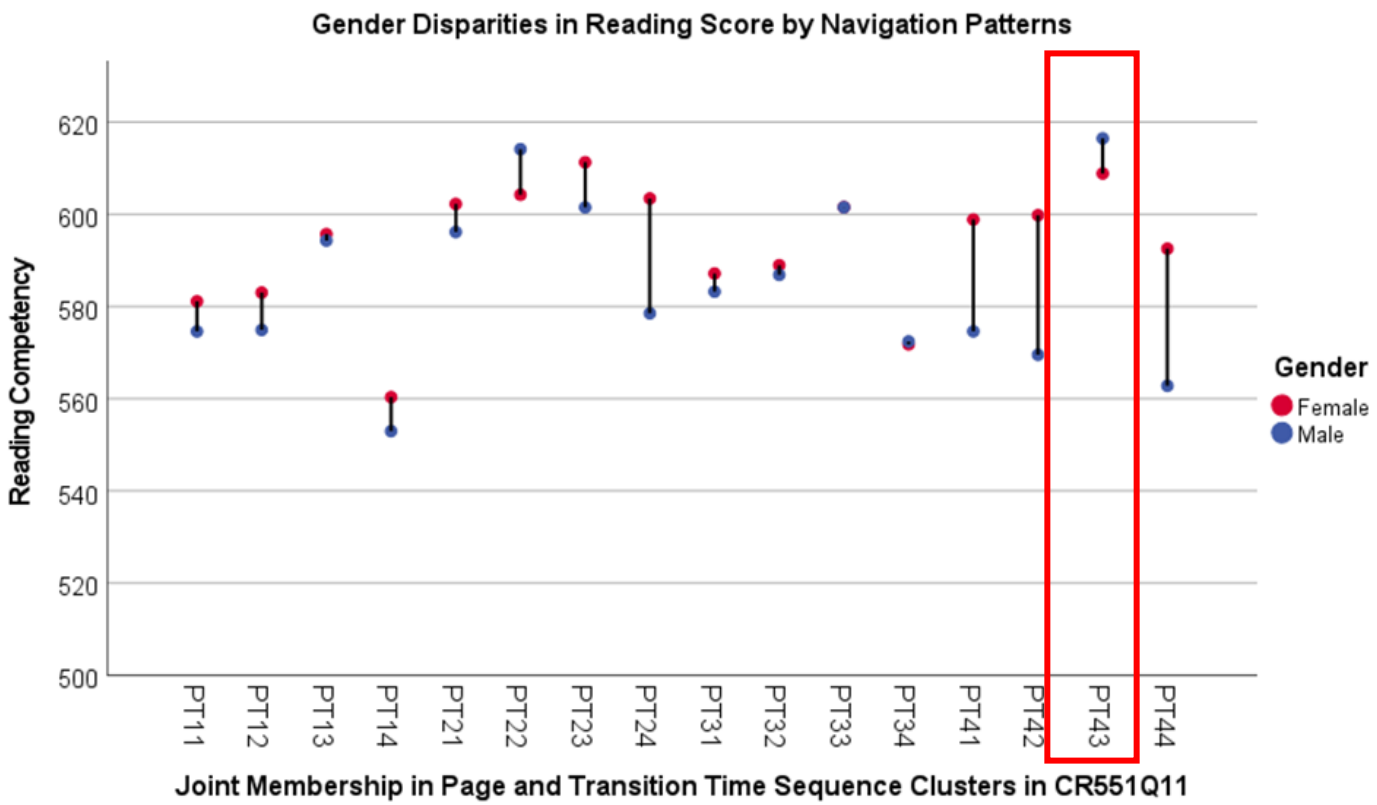




Result 4: Gender Disparities

Girls were more likely to achieve higher scores than boys when longer navigation sequences were used with shorter reading time on transited pages.

However, boys were more likely to achieve higher scores than girls when they spent longer time reading either homepage or transited page along with comprehensive navigation paths through the multiple pages.



Summary

- This study draws on sequential process data from a multiple-source reading task to showcase how students' navigation strategies could be identified via sequence clustering on the DTW similarity measure.
- The DTW method features in identifying the optimal warping path between the two sequences, which cares more about the sequence shapes across time windows.
- The navigation sequence patterns were moderately associated with students' reading proficiency levels.
- Students who visited all the pages and spent more time reading without rush transitions scored higher in reading than those students with less focused navigation.

Evaluating Consistency of Behavioral Patterns Across Multiple Tasks Using Process Data

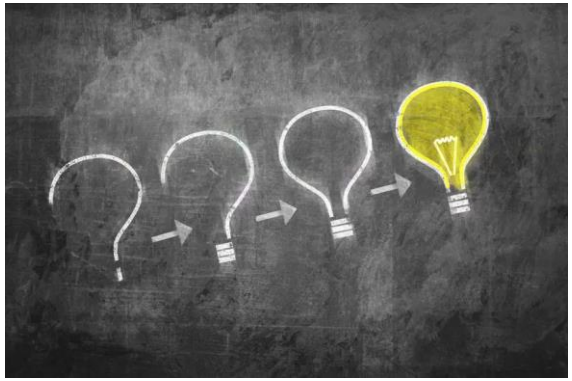
An Empirical Study in PIAAC

In collaboration with
Dr. Dandan Liao, Cambium Assessment
Dr. Hok Kan Ling, Queen's University
Dr. Hong Jiao, University of Maryland

He, Q., Liao, D., Ling, H. K., & Jiao, H. (accepted). Evaluating Consistency of Behavioral Patterns across Multiple Tasks Using Process Data: A case study in PIAAC. Invited book chapter in forthcoming book L. Khorramdel, M. von Davier, K. Yamamoto. (Eds.) *Innovative Computer-based International Large-Scale Assessments – Foundations, Methodologies and Quality Assurance Procedures*. Springer.

PIAAC

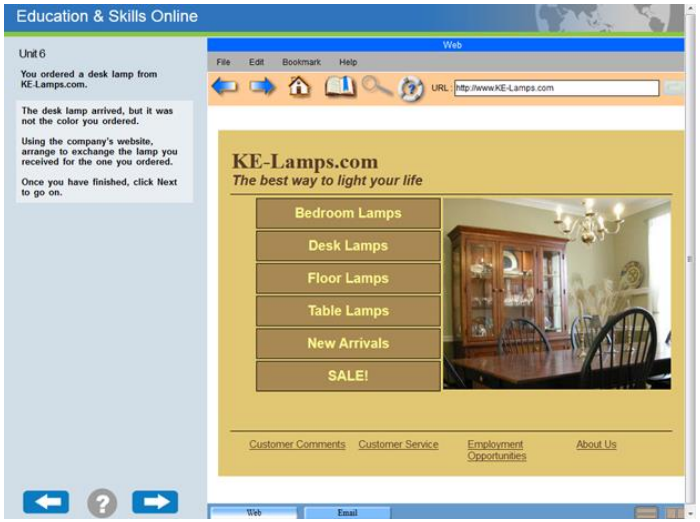
- The Programme for the International Assessment of Adult Competencies (PIAAC) is the first International study with a focus on assessing adults' skills carried out by OECD.
- Assess the cognitive and workplace skills of working-age (16 – 65 years) individuals worldwide.
- Problem Solving in Technology-Rich Environments (PSTRE)

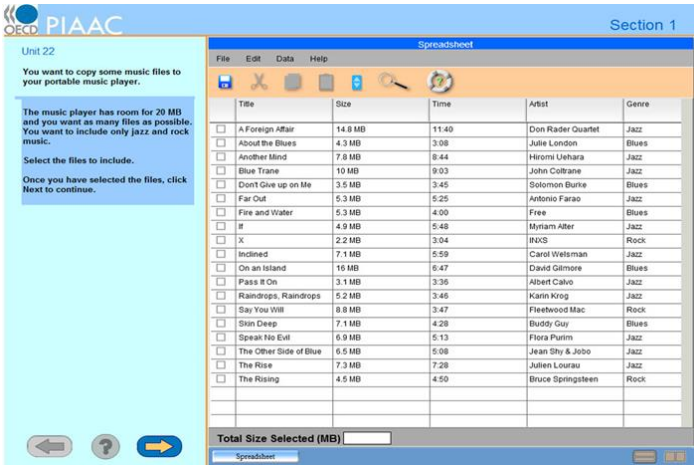


Introduction

- Because of the complexity and high dimensional structure of process data, most studies focus on analyzing **<u>one single</u>** item. (e.g., He & von Davier, 2016; Han et al., 2019; Ulitzsch et al., 2021)

- Evaluating the behavioral consistency across items renders possible capturing and modeling person-related latent characteristics. (e.g., Liao et al., 2019; He et al., 2019, 2021)





Research Objectives

- To investigate whether the consistent behavioral patterns could be identified by process data features

What are the consistent behavioral patterns?

- To examine the association among the consistency of behavioral patterns with cognitive competency and background variables.

What kind of people are in each behavioral pattern?

Three sub-studies

Study 1:

Using aggregate-level response process variables: number of actions and response time.

Study 2:

Using action sequences with similarity computation.

Study 3:

Mapping the behavioral patterns by problem-solving competency and background variables.

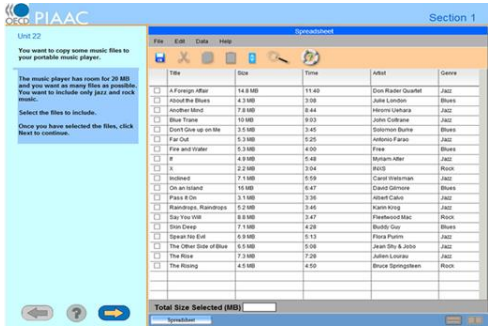
Sample

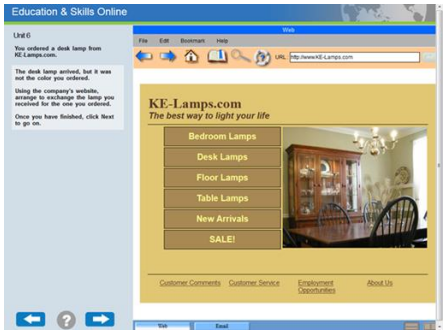
- A total of 1,340 test takers in the U.S. sample
- Routed to the second module of PSTRE items in PIAAC 2012.
- 629 female test takers (46.9%) and 711 male test takers (53.1%).
- The mean age was 39.2 years (SD = 14.0).
- 680 test takers (50.7%) had an educational level above high school.
- Three cases had technical issues, removed
- 1,337 test takers in the final sample.

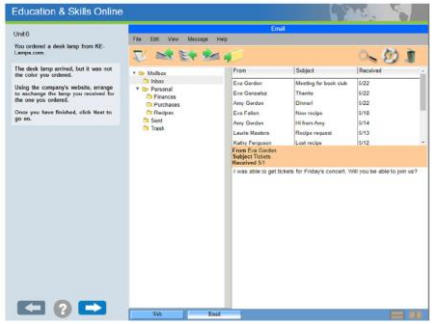
Instruments

Item ID	Item Sequence	Content	Environment		
			Spreadsheet	Web	Email
U19a	Item 1	Club Membership	*		*
U19b	Item 2	Club Membership	*		*
U07x	Item 3	Book Order		*	
U02x	Item 4	Meeting Room		*	*
U16x	Item 5	Reply All			*
U11b	Item 6	Locate Email			*
U23x	Item 7	Lamp Return		*	*

Note: The item position is fixed in PS2.







Study 1: Using Aggregate Level Variables (number of actions, total response time)

Why These Two Variables?

- The number of actions and response time bear important information on behavioral patterns.

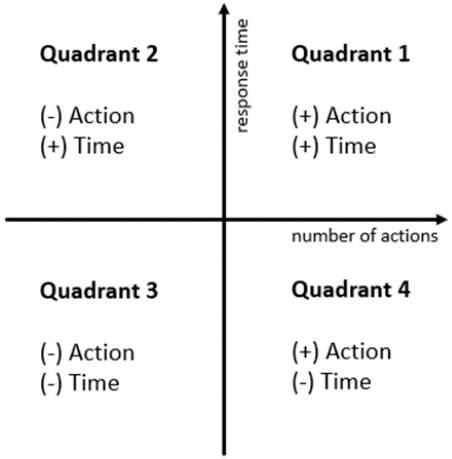
- Evidence has shown that response time is highly correlated with the problem-solving process, skipping behaviors, engagement, and performance (e.g., Ulitzsch et al., 2019; de Boeck & Jeon, 2019; Goldhammer et al., 2014; Liao et al., 2020; Naumann & Goldhammer, 2017; Sahin & Colvin, 2020).

- The variables response time and the number of actions were found as the most informative factors in a previous cluster analysis of respondents' behaviors using the same sample (He, Liao, & Jiao, 2019).

Study 1: Using Aggregate Level Variables (number of actions, total response time)

Item ID	Number of Actions					Response Time (Minutes)				
	Median	Mean	SD	Min	Max	Median	Mean	SD	Min	Max
U19a	15	15.85	11.76	0	97	1.80	2.08	1.40	0.09	20.58
U19b	11	17.80	24.28	0	260	2.83	3.24	2.49	0.08	26.37
U07x	13	14.24	12.00	0	90	1.66	1.88	1.29	0.05	9.45
U02x	24	33.93	33.82	0	194	2.54	3.59	3.48	0.09	45.07
U16x	14	20.07	21.25	0	191	1.85	2.26	1.83	0.06	15.93
U11b	15	24.15	27.54	0	271	1.21	1.59	1.49	0.06	19.43
U23x	12	15.88	16.02	0	183	1.36	1.66	1.72	0.07	42.25

Note: The items are presented in the order of their position in PS2.



- In each item, individual was assigned to the quadrant based on the number of actions and total response time compared with the median value of the whole sample in this item.
- For those showing in the same quadrant across more than half the assigned items (i.e., 4 out of 7 items), they were identified as consistent pattern and labeled as the corresponding quadrant (G1-G4), otherwise they were assigned in the inconsistent group (G5).

$$Q_1 = \{1, 2, 1, 3, 1, 1, 1\}$$
$$G = \{5, 1, 1, 0\}$$

$$Q_2 = \{1, 2, 1, 3, 4, 4, 3\}$$
$$G = \{2, 1, 2, 2\}$$

Study 1: Using Aggregate Level Variables (number of actions, total response time)

	G1	G2	G3	G4	G5
	A+T+	A-T+	A-T-	A+T-	
Number of respondents	488	32	477	38	302
Percentage	36.5%	2.4%	35.7%	2.8%	22.6%

Note: G1 to G4 represent consistent behavioral groups, while G5 represent inconsistent behavioral group. G1 indicates the group with long action sequences and long time, G2 indicates the group with short action sequences and long time, G3 indicates the group with short action sequences and short time, and G4 indicates the group with long action sequences and short time.

Study 1: Using Aggregate Level Variables (number of actions, total response time)

Percentage agreement by five (in)consistency groups

	U19a	U19b	U07x	U02x	U16x	U11b	U23x
U23x							1.00
U11b						1.00	0.62
U16x					1.00	0.54	0.58
U02x				1.00	0.65	0.67	0.70
U07x			1.00	0.66	0.55	0.58	0.60
U19b		1.00	0.43	0.47	0.44	0.48	0.46
U19a	1.00	0.42	0.37	0.40	0.37	0.37	0.42

G1 average percentage agreement = 0.51

	U19a	U19b	U07x	U02x	U16x	U11b	U23x
U23x							1.00
U11b						1.00	0.53
U16x					1.00	0.44	0.38
U02x				1.00	0.44	0.34	0.16
U07x			1.00	0.28	0.47	0.41	0.44
U19b		1.00	0.44	0.41	0.59	0.34	0.34
U19a	1.00	0.50	0.59	0.16	0.50	0.50	0.50

G2 average percentage agreement = 0.42

	U19a	U19b	U07x	U02x	U16x	U11b	U23x
U23x							1.00
U11b						1.00	0.67
U16x					1.00	0.69	0.65
U02x				1.00	0.75	0.77	0.73
U07x			1.00	0.78	0.66	0.68	0.71
U19b		1.00	0.63	0.68	0.60	0.60	0.61
U19a	1.00	0.46	0.51	0.51	0.45	0.47	0.45

G3 average percentage agreement = 0.62

	U19a	U19b	U07x	U02x	U16x	U11b	U23x
U23x							1.00
U11b						1.00	0.50
U16x					1.00	0.37	0.37
U02x				1.00	0.34	0.16	0.39
U07x			1.00	0.37	0.32	0.37	0.37
U19b		1.00	0.37	0.16	0.39	0.58	0.63
U19a	1.00	0.71	0.47	0.05	0.47	0.66	0.58

G4 average percentage agreement = 0.41

	U19a	U19b	U07x	U02x	U16x	U11b	U23x
U23x							1.00
U11b						1.00	0.53
U16x					1.00	0.44	0.38
U02x				1.00	0.44	0.34	0.16
U07x			1.00	0.28	0.47	0.41	0.44
U19b		1.00	0.44	0.41	0.59	0.34	0.34
U19a	1.00	0.50	0.59	0.16	0.50	0.50	0.50

G5 average percentage agreement = 0.42

- Respondents using short action sequence and short response time were the most consistent across multiple items.
- Respondents (A-T+ or A+T-) were less stable to remain in the same quadrant across items.

Study 2: Using Action Sequences for Similarity Computation

Study 2: Using Action Sequences for Similarity Computation

- Generating process sequence indicators across multiple tasks

- Sequence distance between individual observed sequence (OS) and predefined reference sequences (RS) with longest common subsequences (LCS; He et al., 2021)

He, Q., Borgonovi, F., Paccagnella, M. (2021). Leveraging process data to assess adults' problem-solving skills: Identifying generalized behavioral patterns with sequence mining. Computers and Education, 166, 104170.

Longest common subsequence

i	j	0	1	2	3	4	5	6
		y_j	B	D	C	A	B	A
0	x_i	0	0	0	0	0	0	0
1	A	0	↑0	↑0	↑0	↖1	←1	↖1
2	B	0	↖1	←1	←1	↑1	↖2	←2
3	C	0	↑1	↑1	↖2	←2	↑2	↑2
4	B	0	↖1	↑1	↑2	↑2	↖3	←3
5	D	0	↑1	↖2	↑2	↑2	↑3	↑3
6	A	0	↑1	↑2	↑2	↖3	↑3	↖4
7	B	0	↖1	↑2	↑2	↑3	↖4	↑4

Let $X = (x_1, x_2, \dots, x_i)$ and $Y = (y_1, y_2, \dots, y_j)$ be two sequences. x_i and y_j are actions within the sequence X and Y, respectively. The prefixes of X and Y are $X_{1,}\ X_{2,}\ , \dots, X_i$ and $Y_{1,}\ Y_{2,}\ , \dots,\ Y_j$, respectively. Let $LCS\ (X_i, Y_j)$ represent the set of longest common subsequence of prefixes X_i and Y_j. The set of sequences is given as:

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

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

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

LCS Computation Example

RS_1: searching from toolbar (length=11)
Start, Toolbar_SS_Find, On_SearchBox, Off_SearchBox, Search_OK, SS_SEARCH, Email, On_Email_Message, Off_Email_Message, Next, Next_OK

RS_2: searching from menu item (length=11)
Start, Menuitem_Find, On_SearchBox, Off_SearchBox, Search_OK, SS_SEARCH, Email, On_Email_Message, Off_Email_Message, Next, Next_OK

RS_3: sorting from toolbar (length=9)
Start, Toolbar_SS_Sort, Sort_1_B, Sort_OK, Email, On_Email_Message, Off_Email_Message, Next, Next_OK

RS_4: sorting from menu item (length=9)
Start, Menuitem_Sort, Sort_1_B, Sort_OK, Email, On_Email_Message, Off_Email_Message, Next, Next_OK

OBSERVATION (length=25)
Start,Toolbar_SS_Help,Menu_SS_Edit,Menu_SS_Data,Menuitem_Sort,Sort_1_B,Sort_1A,Sort_OK,SS_Sort_1Ba,Email,On_Email_Message,Off_Email_Message,SS,On_Email_Message,Off_Email_Message,Email,On_Email_Message,,,,,,,Off_Email_Message,Toolbar_E_Send,On_Email_Message,Off_Email_Message,Next,On_Email_Message,Off_Email_Message,Next_OK

LCS1 (length=6): Start, Email, On_Email_Message, Off_Email_Message, Next, Next_OK
LCS2 (length=6): Start, Email, On_Email_Message, Off_Email_Message, Next, Next_OK
LCS3 (length=8): Start, Sort_1_B, Sort_OK, Email, On_Email_Message, Off_Email_Message, Next, Next_OK
LCS4 (length=9): Start, Menuitem_Sort, Sort_1_B, Sort_OK, Email, On_Email_Message, Off_Email_Message, Next, Next_OK

LCS Indicators

- Similarity
  - $Similarity = \text{len}(LCS)/\text{len}(RS)$
  - $SM = \text{Mean}(Sim_1, Sim_2, \dots, Sim_n)$
  - $SSD = \text{SD}(Sim_1, Sim_2, \dots, Sim_n)$

The mean of this distribution (SM) is defined as the average degree of similarity across items. A higher value of SM indicates that, on average, a respondent solved problems by following the reference sequences closely.

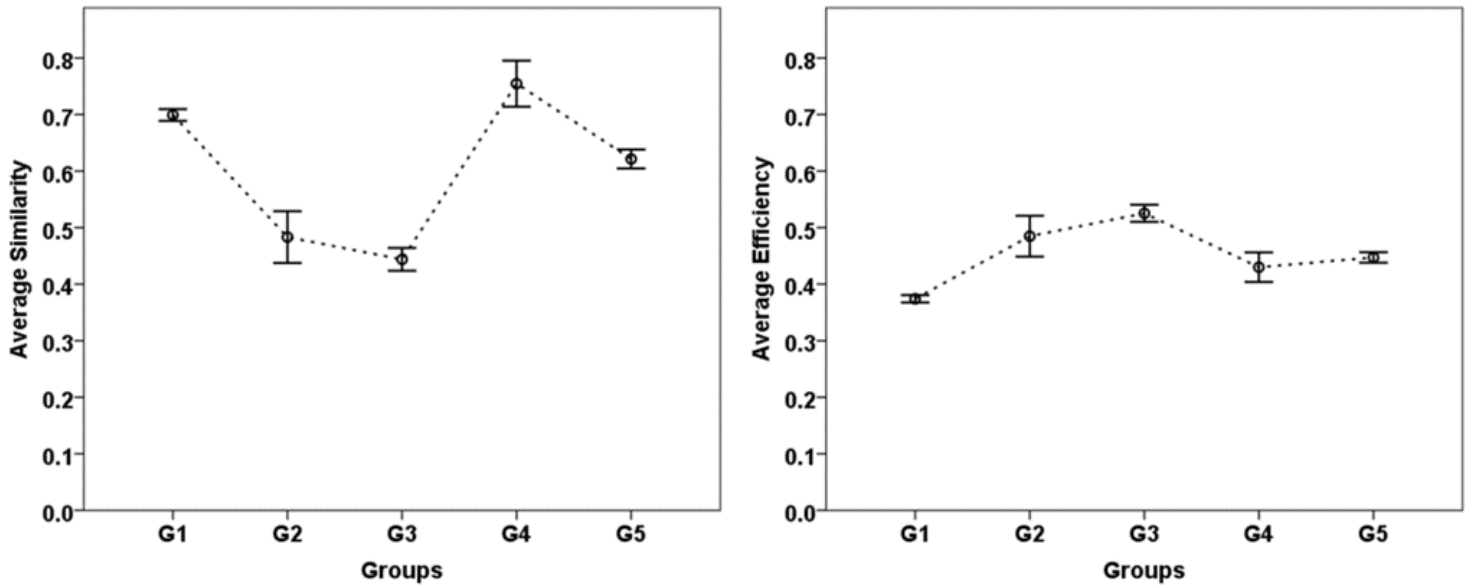
The standard deviation of this distribution (SSD) is used as an indicator of consistency of similarity to show how much the observed sequence is far away from the predefined ones.

- Efficiency
  - $Efficiency = \text{len}(LCS)/\text{len}(OS)$
  - $EM = Mean(Eff_1, Eff_2, \dots, Eff_n)$
  - $ESD = \text{SD}(Eff_1, Eff_2, \dots, Eff_n)$

The mean of this distribution (EM) is simply defined as the degree of efficiency across items. A higher value of EM indicates that a respondent on average solves problems in an efficient way (i.e., with few redundant actions).

The standard deviation of this distribution (ESD) is used as an indicator of consistency of efficiency that varies across items.

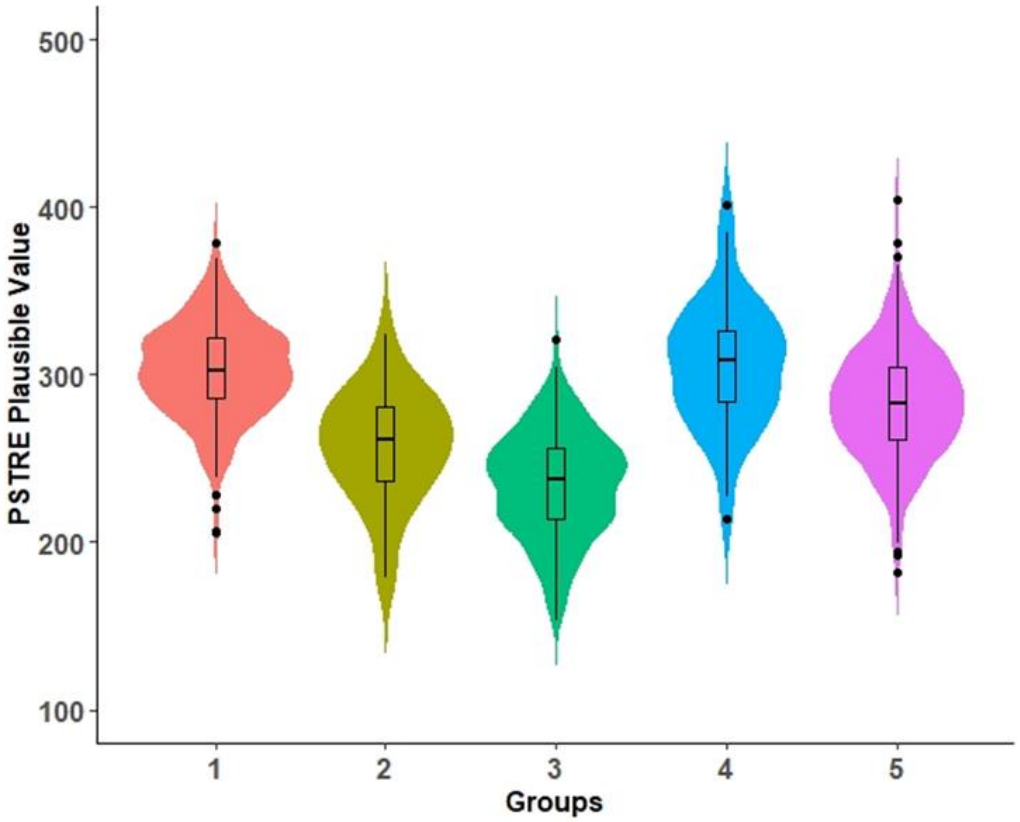
Study 2: Using Action Sequences for Similarity Computation



The association among two LCS indicators and (in)consistency groups derived from response time and number of actions

Study 3: Mapping Behavioral Patterns by Problem-Solving Competency and Background Variables

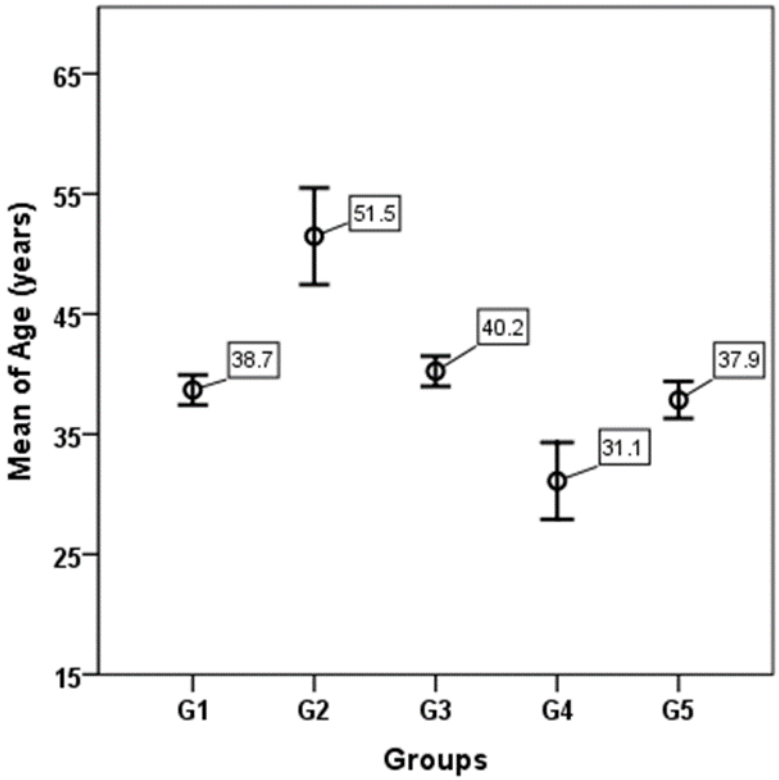
Study 3: Mapping Behavioral Patterns by Problem-Solving Competency and Background Variables



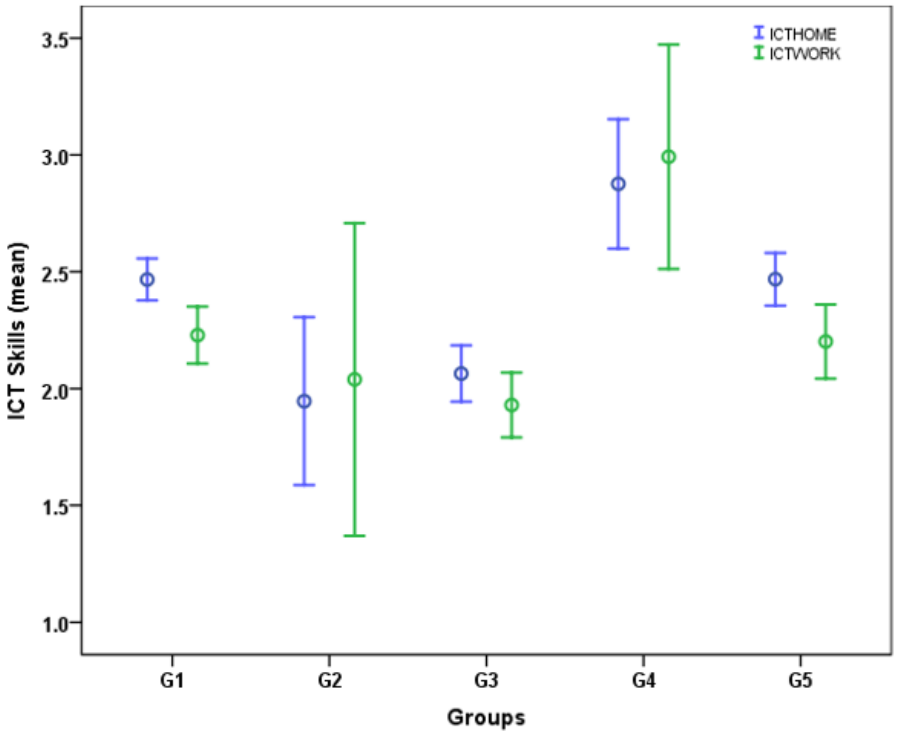
The distribution of problem-solving proficiency by five (in)consistency groups

A bimodal distribution in G1 and G3, suggesting respondents in these groups possibly had mixed proficiency levels while having similar behaviors.

Study 3: Mapping Behavioral Patterns by Problem-Solving Competency and Background Variables

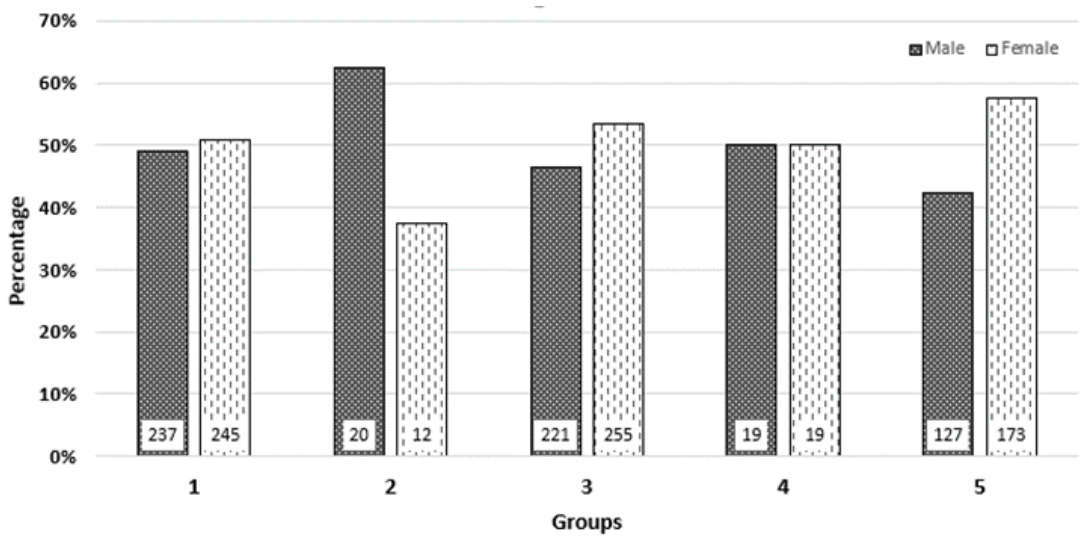


The distribution of age by behavioral patterns



The distribution of use of ICT at home and at work by behavioral patterns

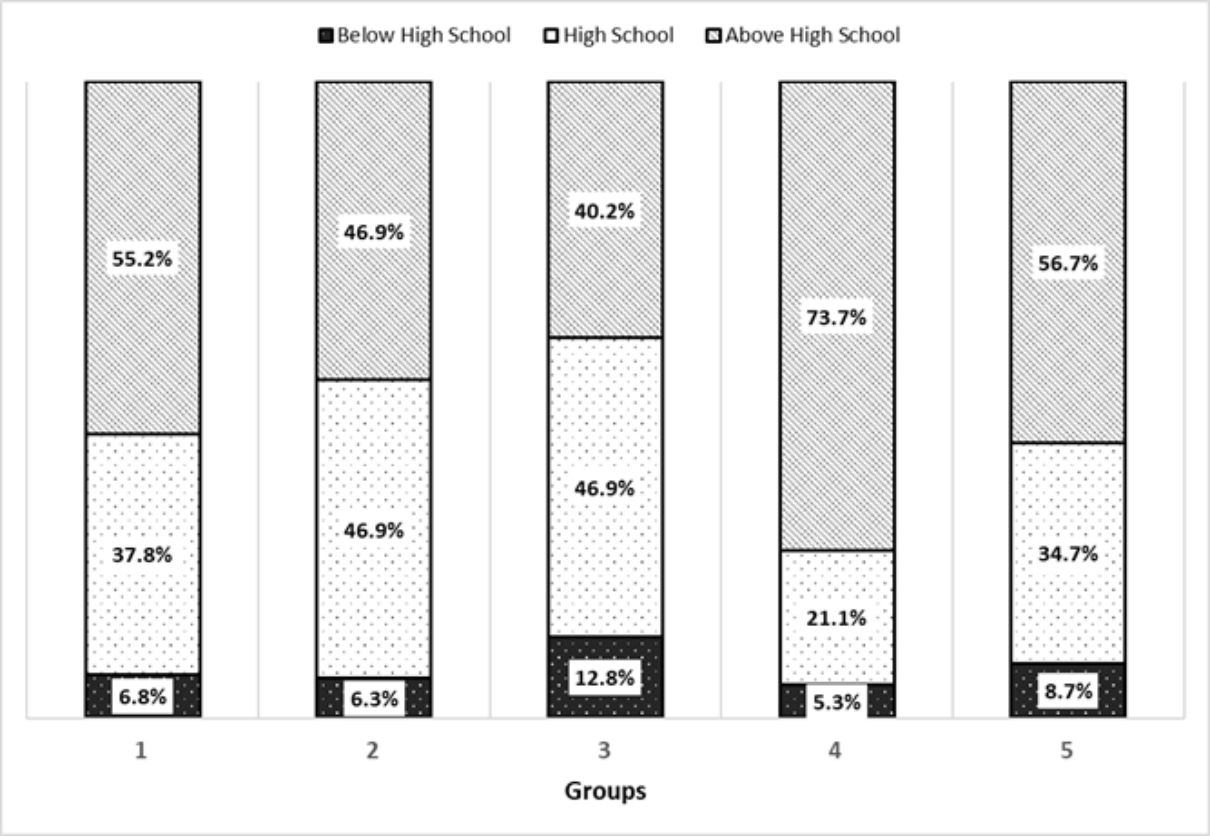
Study 3: Mapping Behavioral Patterns by Problem-Solving Competency and Background Variables



A higher proportion of males (over 60%) than females were found in the group consistently used short action sequences and long response times.

The distribution of gender by behavioral patterns

Study 3: Mapping Behavioral Patterns by Problem-Solving Competency and Background Variables



The distribution of respondents' educational level by behavioral patterns

Summary

- This study drew on the process aggregate-level variables and action sequences to assess the consistency of test-takers' behavior patterns across multiple interactive items, and also associated with competency level and background variables.

- Around 80% of respondents showed consistent patterns (in over half of the tasks) by the two dimensions, response time and the number of actions.

- Respondents who consistently followed A+T- pattern were found the highest problem-solving competency, the youngest, highest education level and adopted the most similar strategies with optimal sequences.

- From response processes, high-need groups could be identified for better attention for intervention (e.g., adults showing A-T+ may be highly motivated but need further support)

Discussion and Outlook

How Predictable the Process Variables in PUF?

- Dilemma:
  - How many variables and on which level of process data need to be released?
    - Research supports: the more the better?
    - Country: confidentiality, cost, operational work
    - Limited data could be released

- Three process variables in the public released file of PIAAC and PISA
  - Number of actions
  - Total time
  - Time for the first action

He, Q., Shi, Q., Borgonovi, F., & Paccagnella, M. (in preparation). Evaluating Prediction Power of Process Variables on Problem-Solving Proficiency Levels with Machine Learning Methods.

Machine Learning Binary Prediction

- Study 1: Use three process variables (number of actions, total time, time for the first action) in the PUF (n = 5,292, from 5 countries, focus on PS2)

	LR		RF		SVM	
	Mean	*SD*	*Mean*	*SD*	*Mean*	*SD*
Accuracy	0.81	0.02	0.84	0.01	0.83	0.00
Sensitivity	0.84	0.02	0.85	0.01	0.85	0.01
Specificity	0.77	0.02	0.82	0.02	0.79	0.02
PPV	0.83	0.01	0.86	0.01	0.85	0.01
NPV	0.78	0.02	0.81	0.01	0.80	0.01
F1	0.84	0.02	0.86	0.01	0.85	0.00
AUROC	0.88	0.02	0.92	0.00	0.90	0.00

Note. LR = logistic regression; RF = random forest; SVM = support vector machine; PPV = positive predictive value; NPV = negative predictive value; AUC = area under the ROC curve; BLv1 = below level 1; Lv1 = level 1; Lv2 = Level 2; Lv3 = Level 3.

Machine Learning Binary Prediction

- Study 2: Use two generated variables similarity and efficiency (n = 5,292, from 5 countries, focus on PS2)

	LR		RF		SVM	
	Mean	*SD*	*Mean*	*SD*	*Mean*	*SD*
Accuracy	0.83	0.01	0.83	0.01	0.83	0.01
Sensitivity	0.77	0.02	0.78	0.02	0.78	0.02
Specificity	0.87	0.01	0.86	0.02	0.87	0.01
PPV	0.81	0.01	0.81	0.02	0.82	0.02
NPV	0.84	0.01	0.84	0.01	0.84	0.01
F1	0.79	0.01	0.79	0.00	0.80	0.01
AUROC	0.91	0.01	0.91	0.01	0.91	0.01

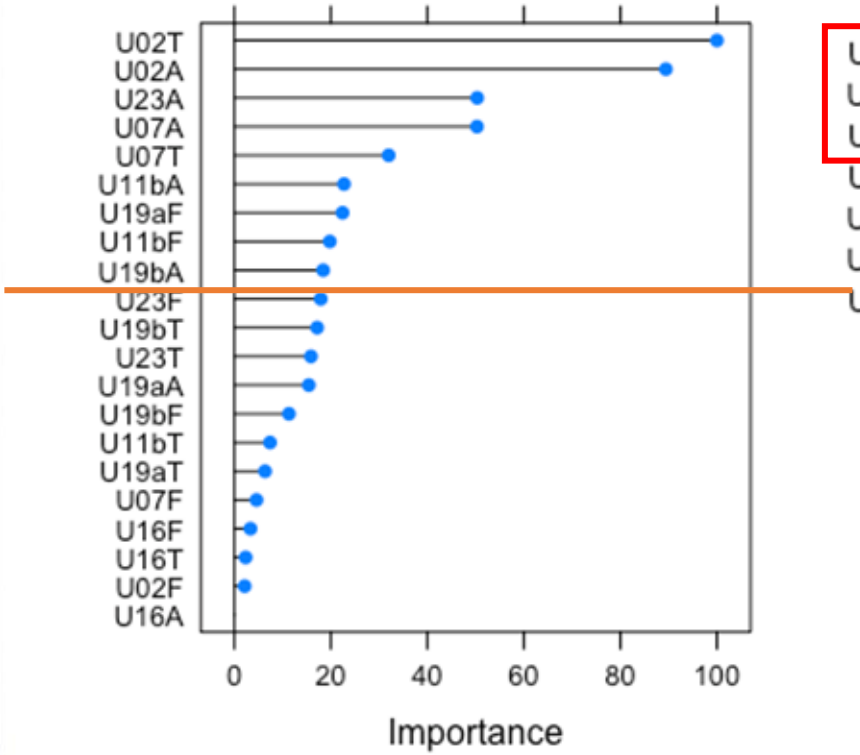
Note. LR = logistic regression; RF = random forest; SVM = support vector machine; PPV = positive predictive value; NPV = negative predictive value; AUC = area under the ROC curve; BLv1 = below level 1; Lv1 = level 1; Lv2 = Level 2; Lv3 = Level 3.

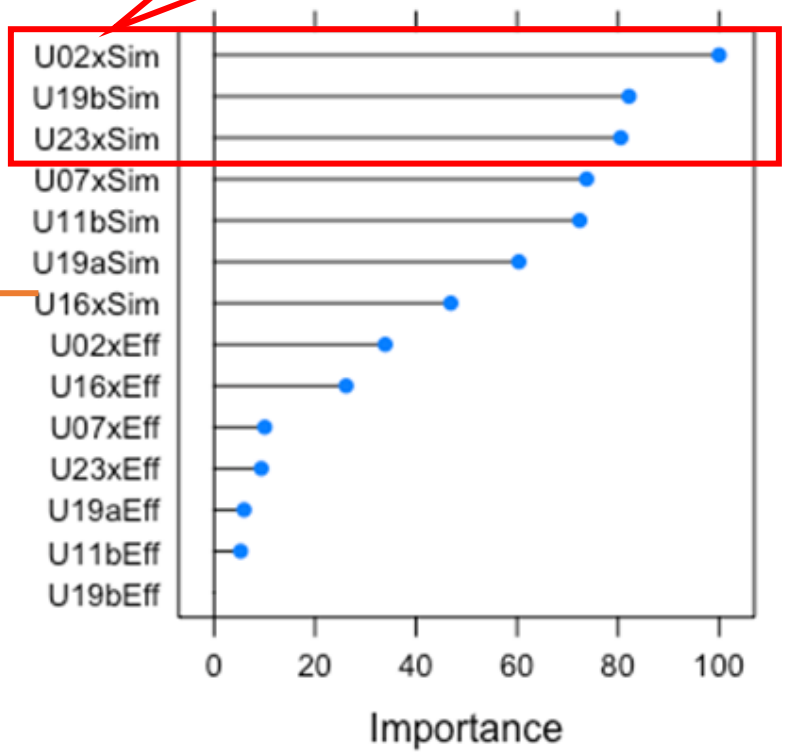
Machine Learning Binary Prediction

- Importance of variables

	a	*b*
Unit 19a	1.414	-1.367
Unit 19b	1.072	-0.677
Unit 07	1.104	-0.237
Unit 02	1.184	0.784
Unit 16	1.377	-0.773
Unit 11b	0.471	0.774
Unit 23	0.533	-0.052

The items with high discrimination in behaviors (not necessary to be highly discriminative ones in IRT model).





Implications

- The three process variables (number of actions, total time, time for the first action) are very predictive (over 80% accuracy) to distinguish low and high adults with problem-solving skill. These three variables need to continue included in the PUF.

- The generated variable similarity (distance measure between observed sequence and predefined sequence) is a very predictive variable and also could be considered to use in further operational work and recommended to be included in the PUF. (would need content experts' help).

- The importance of variables in prediction could be used as an efficient way to identify the high-discriminative items from process data perspective, (not necessary to be consistent as item discrimination parameters estimated in IRT), which could be helpful in interactive problem-solving items design.

Final Words

- Process data has been arousing increasing interests and attention in educational assessment and learning analytics.

- One-time assessment is not the target of education but tracking the students' progress and providing helpful supports. Process data could be given a broader scope for dynamic and longitudinal study.

- Post-assessment process data analysis could be progressed to dynamic process data analysis throughout the assessment. It could be a new trend in the future studies to bring new meaning for adaptive testing, not only from responses but also from interactions.

Reference

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

De Boeck, P., & Jeon, M. (2019). An overview of models for response times and processes in cognitive tests. *Frontiers in psychology, 10*, 102.

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

Goldhammer, F., Naumann, J., Stelter, A., Tóth, K., Rölke, H., & Klieme, E. (2014). The time on task effect in reading and problem solving is moderated by task difficulty and skill: Insights from a computer-based large-scale assessment. *Journal of Educational Psychology, 106*(3), 608–626.

Han, Z., He, Q., & von Davier, M. (2019). Predictive Feature Generation and Selection Using Process Data from PISA Interactive Problem-Solving Items: An Application of Random Forests. *Frontiers in Psychology, 10*: 1421.

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

He, Q., Bei, N., & Jiang, Y. (in preparation). Identifying Latent State Transitions with Multigroup Hidden Markov Model on Process Data.

He, Q., Bei, N., Jiang, Y., Kaldes, G., & Tighe, E.L. (in preparation). Understanding Adults' Latent Problem-Solving Strategies by Different Literacy Competence Levels with Hidden Markov Model

He, Q., Liao, D., Ling, H. K., & Jiao, H. (accepted). Evaluating Consistency of Behavioral Patterns across Multiple Tasks Using Process Data: A case study in PIAAC. Invited book chapter in forthcoming book L. Khorramdel, M. von Davier, K. Yamamoto. (Eds.) *Innovative Computer-based International Large-Scale Assessments – Foundations, Methodologies and Quality Assurance Procedures*. Springer.

He, Q., Borgonovi, F., & Paccagnella, M. (2021). Leveraging process data to assess adults' problem-solving skills: Identifying generalized behavioral patterns with sequence mining. *Computers and Education, 166*, 104170.

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

He, Q., Borgonovi, F., Suárez-Álvarez, J. (2022). Clustering Sequential Navigation Patterns in Multiple-Source Reading Tasks with Dynamic Time Warping Method. *Journal of Computer-Assisted Learning*. DOI: 10.1111/jcal.12748

He, Q., Shi, Q., Borgonovi, F., & Paccagnella, M. (in preparation). Evaluating Prediction Power of Process Variables on Problem-Solving Proficiency Levels with Machine Learning Methods.

He, Q., & von Davier, M. (2016). Analyzing Process Data from Problem-Solving Items with N-Grams: Insights from a Computer-Based Large-Scale Assessment. In Y. Rosen, S. Ferrara, & M. Mosharraf (Eds.) *Handbook of Research on Technology Tools for Real-World Skill Development* (pp. 749-776). Hershey, PA: Information Science Reference.

He, Q., & von Davier, M. (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: Springer.

Reference

He, Q., Liao, D., & Jiao, H. (2019). Clustering behavioral patterns using process data in PIAAC problem-solving items. In B. P. Veldkamp and C. Sluijter (eds.), *Theoretical and Practical Advances in Computer-based Educational Measurement* (pp. 189-212), Methodology of Educational Measurement and Assessment (book series), Springer.

Heldt, M., Massek, C., Drossel, K., & Eickelmann, B. (2020). The relationship between differences in students' computer and information literacy and response times: an analysis of IEA-ICILS data. *Large-scale Assessments in Education, 8*(1), 1-20.

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

Liao, D., He, Q., & Jiao, H. (2020). Using log files to identify sequential patterns in PIAAC problem solving environments by U.S. adults' employment. *National Center for Education Statistics (NCES) commissioned research report*. https://static1.squarespace.com/static/51bb74b8e4b0139570ddf020/t/5e41972c89f5bb0179fcb999/1581356847181/2020_Liao_He_Jiao_Log-Files-Sequential-Patterns.pdf

Naumann, J., & Goldhammer, F. (2017). Time-on-task effects in digital reading are non-linear and moderated by persons' skills and tasks' demands. *Learning and Individual Differences*, 53, 1-16.Rousseeuw, 1987

Salles, F., Dos Santos, R., & Keskpaik, S. (2020). When didactics meet data science: Process data analysis in large-scale mathematics assessment in France. *Large-scale Assessments in Education, 8*(1), 1-20.

Sahin, F., & Colvin, K. (2020). Enhancing response time thresholds with response behaviors for detecting disengaged examinees. *Large-scale Assessment in Education, 8*(1), 1–24.

Sakoe, H. & Chiba. (1978). Dynamic programming algorithm optimization for spokenword recognition. Acoustics, Speech and Signal Processing, *IEEE Transactions on 26*(1), 43–49.

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

Tang, X., Wang, Z., He, Q., Liu, J., & Ying, Z. (2020). Latent feature extraction for process data via multidimensional scaling. *Psychometrika, 85*(2), 378-397.

Tang, X., Wang, Z., Liu, J., & Ying, Z. (2021). An exploratory analysis of the latent structure of process data via action sequence autoencoders. *British Journal of Mathematical and Statistical Psychology, 74*(1), 1-33.

Ulitzsch, E., He, Q., Pohl, S. (2022). Using sequence mining techniques for understanding incorrect behavioral patterns on interactive tasks. *Journal of Educational and Behavioral Statistics, 47*(1), 3-35.

Ulitzsch, E., He, Q., Ulitzsch, V., Nichterlein, A., Molter, H., Niedermeier, R., Pohl, S. (2021). Combining clickstream analyses and graph-modeled data clustering for identifying common response process using time-stamped action sequence. *Psychometrika, 86*, 190-214.

Ulitzsch, E., Ulitzsch, V., He, Q., & Lüdtke, O. (2022). A machine learning-based procedure for leveraging clickstream data to investigate early predictability of failure on interactive tasks. *Behavior Research Methods*, 1-21.

Xiao, Y., He, Q., Veldkamp, B., & Liu, H. (2021). Exploring Latent States of Problem-Solving Competence Using Hidden Markov Modeling on Process Data. *Journal of Computer-Assisted Learning, 37*(5), 1232-1247.

Thank you very much!

Qiwei He

qhe@ets.org
heqw.britt0328@gmail.com