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

Sequence-Based Methods in Process Data Analysis

Case Studies: Using Sequence Mining Techniques in LSA

Discussions and Outlook
Interactive Items and Process Data in Large-Scale Assessments
Computer-Based Assessments

- Rich Data
- Enhanced Security
- Automated Scoring
- Streamlined Logistics
- Dynamic Navigation
- Consistency & Reliability
- Track Process
- Innovative Item Design
- Flexibility

Innovative Item Design

Consistency & Reliability

Dynamic Navigation

Track Process

Streamlined Logistics

Automated Scoring

Enhanced Security

Rich Data

Flexibility

Computer-Based Assessments

Programs:
- PISA
- TIMSS
- PIAAC
- NAEP
- PIRLS
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)
Examples from Different Subjects

NAEP 2017 Grade 4 Interactive Math Example Item

ICILS2018 Computational Thinking Interactive Program Coding Example Item

PISA 2015 Collaborative Problem-Solving Example Item
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

Demography
- Personality
- Speed
- Motivation

Person-level characteristics

Task-level characteristics

Results of task completion

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
• 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 (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”

- **Trigrams (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., $\chi^2$ 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

**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

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.

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 weighed thousands of kilos, and the people of Rapa Nui were able to move 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 
  \( X = \{x_1, x_2, ..., x_n\} \) and 
  \( Y = \{y_1, y_2, ..., y_m\} \) 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.

Any distance (Euclidean, Manhattan, ...) which aligns the \( i \)-th point on one time series with the \( i \)-th point on the other will produce a poor similarity score.

A non-linear (elastic) alignment produces a more intuitive similarity measure, allowing similar shapes to match even if they are out of phase in the time axis.
Reasons to Use DTW
Dynamic Time Warping Algorithm

• The initial step of DTW algorithm is defined as:

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

• The recursive function of DTW is defined as

\[
DTW(i, j) = \min \left\{ \begin{array}{c}
DTW(i-1, j) + w_h C(i, j) \\
DTW(i, j-1) + w_v C(i, j) \\
DTW(i-1, j-1) + w_d C(i, j)
\end{array} \right\}
\]

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 \(\{x_1, x_2, ..., x_i\}\) and \(\{y_1, y_2, ..., y_j\}\), and \(DTW(N, M)\) indicates the total cost of the optimal warping path.

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

Silhouette Index in DTW Sequence Clustering

Number of Clusters

Silhouette Index

Page Transition Sequence

Time on Page Sequence
### Result 1: Page Sequence Clustering

<table>
<thead>
<tr>
<th>Clusters</th>
<th>Sample Size</th>
<th>Percentage</th>
<th>Reading Proficiency Score</th>
<th>Gender</th>
<th>Socio-Economic Index</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Mean</td>
<td>S.D.</td>
<td>S.E.</td>
</tr>
<tr>
<td>P1</td>
<td>11091</td>
<td>65.4%</td>
<td>574.13</td>
<td>82.05</td>
<td>0.78</td>
</tr>
<tr>
<td>P2</td>
<td>1952</td>
<td>11.5%</td>
<td>599.06</td>
<td>76.97</td>
<td>1.74</td>
</tr>
<tr>
<td>P3</td>
<td>3438</td>
<td>20.8%</td>
<td>585.07</td>
<td>72.38</td>
<td>1.23</td>
</tr>
<tr>
<td>P4</td>
<td>476</td>
<td>2.6%</td>
<td>587.75</td>
<td>85.03</td>
<td>3.90</td>
</tr>
</tbody>
</table>
## Result 2: Time Sequence Clustering

<table>
<thead>
<tr>
<th>Clusters</th>
<th>Sample Size</th>
<th>Percentage</th>
<th>Reading Proficiency Score</th>
<th>Gender</th>
<th>Socio-Economic Index</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Mean</td>
<td>S.D.</td>
<td>S.E.</td>
</tr>
<tr>
<td>T1</td>
<td>4651</td>
<td>27.4%</td>
<td>582.72</td>
<td>78.12</td>
<td>1.15</td>
</tr>
<tr>
<td>T2</td>
<td>2385</td>
<td>13.6%</td>
<td>583.75</td>
<td>79.43</td>
<td>1.63</td>
</tr>
<tr>
<td>T3</td>
<td>3675</td>
<td>23.5%</td>
<td>598.36</td>
<td>77.17</td>
<td>1.27</td>
</tr>
<tr>
<td>T4</td>
<td>6246</td>
<td>34.8%</td>
<td>564.66</td>
<td>80.83</td>
<td>1.02</td>
</tr>
</tbody>
</table>

![Chart showing time sequence clustering with sample sizes and percentage distribution across different clusters.](chart)

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

The chart illustrates the time sequence clustering with sample sizes and percentage distribution across different clusters. The reading proficiency score, gender, and socio-economic index are also provided for each cluster.
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. Dandand Liao, Cambium Assessment
Dr. Hok Kan Ling, Queen’s University
Dr. Hong Jiao, University of Maryland

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 one single 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

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

What are the consistent behavioral patterns?

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

<table>
<thead>
<tr>
<th>Item ID</th>
<th>Item Sequence</th>
<th>Content</th>
<th>Environment</th>
</tr>
</thead>
<tbody>
<tr>
<td>U19a</td>
<td>Item 1</td>
<td>Club Membership</td>
<td>*</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>*</td>
</tr>
<tr>
<td>U19b</td>
<td>Item 2</td>
<td>Club Membership</td>
<td>*</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>*</td>
</tr>
<tr>
<td>U07x</td>
<td>Item 3</td>
<td>Book Order</td>
<td></td>
</tr>
<tr>
<td>U02x</td>
<td>Item 4</td>
<td>Meeting Room</td>
<td>*</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>*</td>
</tr>
<tr>
<td>U16x</td>
<td>Item 5</td>
<td>Reply All</td>
<td></td>
</tr>
<tr>
<td>U11b</td>
<td>Item 6</td>
<td>Locate Email</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>*</td>
</tr>
<tr>
<td>U23x</td>
<td>Item 7</td>
<td>Lamp Return</td>
<td>*</td>
</tr>
</tbody>
</table>

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)

<table>
<thead>
<tr>
<th>Item ID</th>
<th>Number of Actions</th>
<th>Response Time (Minutes)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Median</td>
<td>Mean</td>
</tr>
<tr>
<td>U19a</td>
<td>15</td>
<td>15.85</td>
</tr>
<tr>
<td>U19b</td>
<td>11</td>
<td>17.80</td>
</tr>
<tr>
<td>U07x</td>
<td>13</td>
<td>14.24</td>
</tr>
<tr>
<td>U02x</td>
<td>24</td>
<td>33.93</td>
</tr>
<tr>
<td>U16x</td>
<td>14</td>
<td>20.07</td>
</tr>
<tr>
<td>U11b</td>
<td>15</td>
<td>24.15</td>
</tr>
<tr>
<td>U23x</td>
<td>12</td>
<td>15.88</td>
</tr>
</tbody>
</table>

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)

<table>
<thead>
<tr>
<th></th>
<th>G1</th>
<th>G2</th>
<th>G3</th>
<th>G4</th>
<th>G5</th>
</tr>
</thead>
<tbody>
<tr>
<td>A+T+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A-T+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A-T-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A+T-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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 represents an 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

- 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)

**Longest common subsequence**

Let \( X = (x_1, x_2, \ldots, x_i) \) and \( Y = (y_1, y_2, \ldots, y_j) \) be two sequences. \( x_i \) and \( y_j \) are actions within the sequence \( X \) and \( Y \), respectively. The prefixes of \( X \) and \( Y \) are \( X_1, X_2, \ldots, X_i \) and \( Y_1, Y_2, \ldots, Y_j \), 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:

\[
LCS(X, Y) = \text{longest } LCS(X_i, Y_{k,j})
\]

![LCS Matrix]

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

\[
\text{length}(LCS(X_i, Y_j)) = \begin{cases} 
0 & \text{if } i = 0 \text{ or } j = 0 \\
\text{length}(i - 1, j - 1) + 1 & \text{if } x_i = y_i \\
\max(\text{length}(i, j - 1), \text{length}(i - 1, j)) & \text{if } x_i \neq y_i 
\end{cases}
\]
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, 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**
  - \( \text{Similarity} = \frac{\text{len}(LCS)}{\text{len}(RS)} \)
  - \( SM = \text{Mean}(Sim_1, Sim_2, \ldots, Sim_n) \)
  - \( SSD = \text{SD}(Sim_1, Sim_2, \ldots, Sim_n) \)

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**
  - \( \text{Efficiency} = \frac{\text{len}(LCS)}{\text{len}(OS)} \)
  - \( EM = \text{Mean}(Eff_1, Eff_2, \ldots, Eff_n) \)
  - \( ESD = \text{SD}(Eff_1, Eff_2, \ldots, Eff_n) \)

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

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

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

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)

<table>
<thead>
<tr>
<th></th>
<th>LR</th>
<th>RF</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.81</td>
<td>0.02</td>
<td>0.84</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.84</td>
<td>0.02</td>
<td>0.85</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.77</td>
<td>0.02</td>
<td>0.82</td>
</tr>
<tr>
<td>PPV</td>
<td>0.83</td>
<td>0.01</td>
<td>0.86</td>
</tr>
<tr>
<td>NPV</td>
<td>0.78</td>
<td>0.02</td>
<td>0.81</td>
</tr>
<tr>
<td>F1</td>
<td>0.84</td>
<td>0.02</td>
<td>0.86</td>
</tr>
<tr>
<td>AUROC</td>
<td>0.88</td>
<td>0.02</td>
<td>0.92</td>
</tr>
</tbody>
</table>

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)

<table>
<thead>
<tr>
<th></th>
<th>LR</th>
<th>RF</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Mean</td>
<td>Mean</td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.83</td>
<td>0.83</td>
<td>0.83</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.77</td>
<td>0.78</td>
<td>0.78</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.87</td>
<td>0.86</td>
<td>0.87</td>
</tr>
<tr>
<td>PPV</td>
<td>0.81</td>
<td>0.81</td>
<td>0.82</td>
</tr>
<tr>
<td>NPV</td>
<td>0.84</td>
<td>0.84</td>
<td>0.84</td>
</tr>
<tr>
<td>F1</td>
<td>0.79</td>
<td>0.79</td>
<td>0.80</td>
</tr>
<tr>
<td>AUROC</td>
<td>0.91</td>
<td>0.91</td>
<td>0.91</td>
</tr>
</tbody>
</table>

Note. LR = logistic regression; RF = random forest; SVM = support vector machine; PPV = positive predictive value; NPV = negative predictive value; AUROC = 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

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

<table>
<thead>
<tr>
<th>Unit</th>
<th>a</th>
<th>b</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unit 19a</td>
<td>1.414</td>
<td>-1.367</td>
</tr>
<tr>
<td>Unit 19b</td>
<td>1.072</td>
<td>-0.677</td>
</tr>
<tr>
<td>Unit 07</td>
<td>1.104</td>
<td>-0.237</td>
</tr>
<tr>
<td>Unit 02</td>
<td>1.184</td>
<td>0.784</td>
</tr>
<tr>
<td>Unit 16</td>
<td>1.377</td>
<td>-0.773</td>
</tr>
<tr>
<td>Unit 11b</td>
<td>0.471</td>
<td>0.774</td>
</tr>
<tr>
<td>Unit 23</td>
<td>0.533</td>
<td>-0.052</td>
</tr>
</tbody>
</table>
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


Reference


Thank you very much!

Qiwei He

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