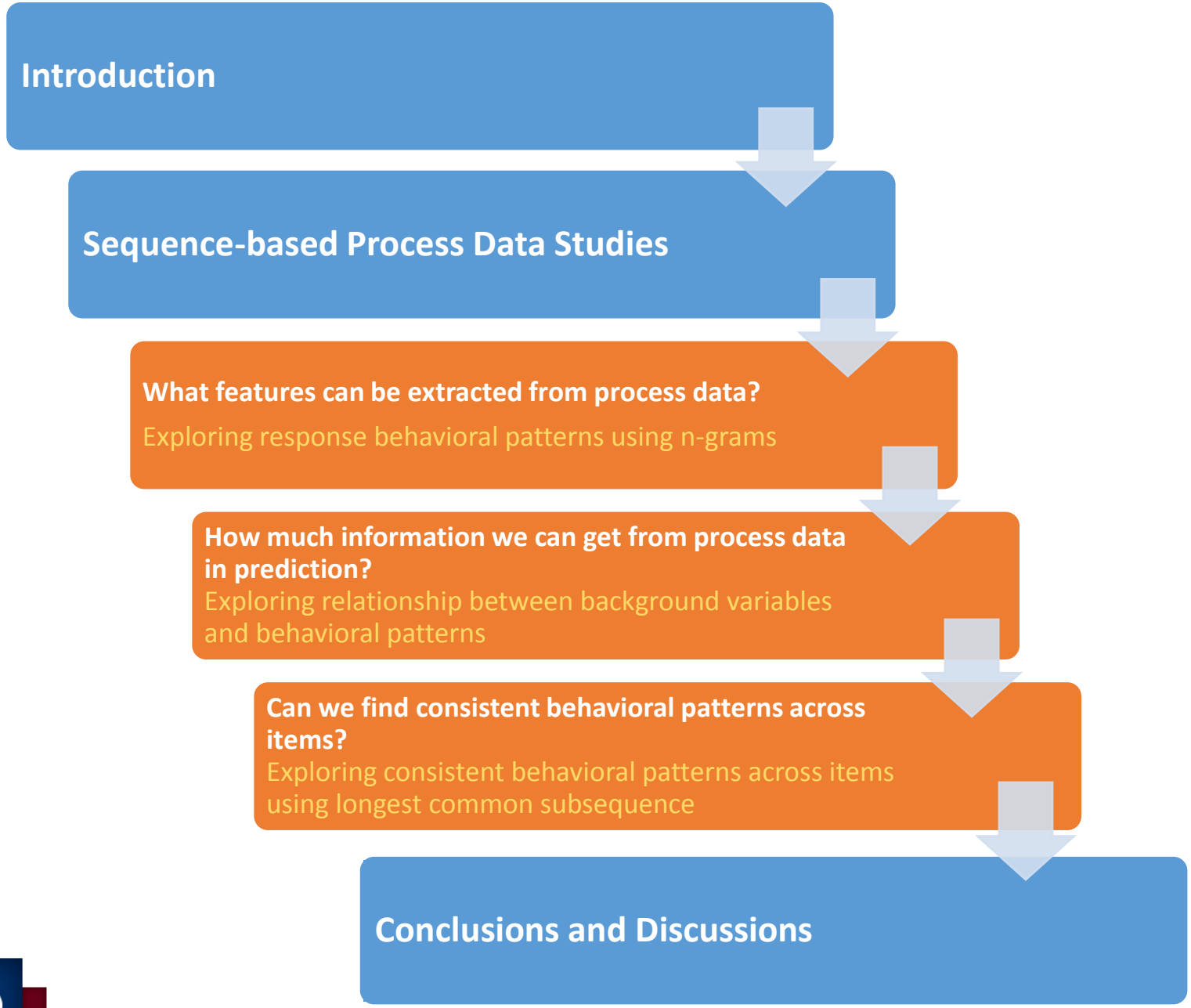
Exploring Sequence-Based Approaches Using Process Data in Large-Scale Assessments

Qiwei Britt He

Educational Testing Service

A joint conference hosted by Educational Testing Service and Educational Research Center, Ireland. @ Hotel Riu Plaza the Gresham, Dublin 1, Ireland, May 16-17 May, 2019

5/16/2019


Introduction
Sequence-based Process Data Studies
What features can be extracted from process data?
Exploring response behavioral patterns using n-grams
How much information we can get from process data in prediction?
Exploring relationship between background variables and behavioral patterns
Can we find consistent behavioral patterns across items?
Exploring consistent behavioral patterns across items using longest common subsequence
Conclusions and Discussions

Introduction

Background

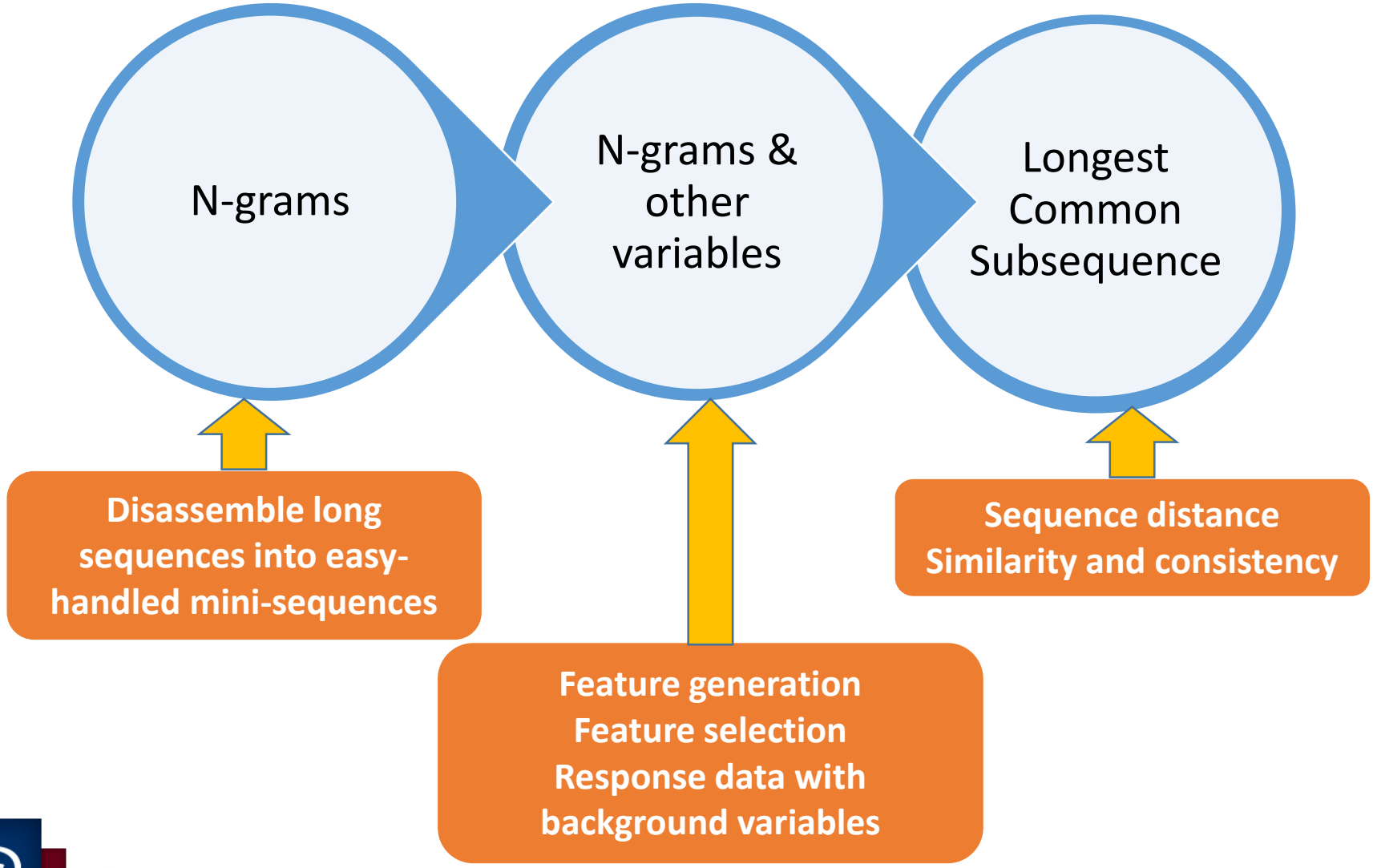
- The use of computers as the delivery platform as PISA and PIAAC enables data collection not just on whether test takers are able to solve the tasks **(response data)** but how they approach the solution and how much time their efforts take **(process data from log files)**.

- Such a new data source is especially valuable in scenario-based interactive items, which provides the possibility in **deeper understanding** about people's problem solving behaviors, **tracking the problem solving sequence**, thus, help in **detecting the reasons** of success or failure in a digital task.

Action sequences

- Similar structure between action sequences and languages.
- Motivated by the methodologies of natural language processing and text mining.
- Two approaches in sequence mining that we applied in recent studies seem promising.
  - N-grams (mini-sequences)
  - Longest common subsequence

Sequence-based process data studies



Exploring response behavioral patterns using n-grams

(He & von Davier, 2015, 2016)

N-grams Model

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"

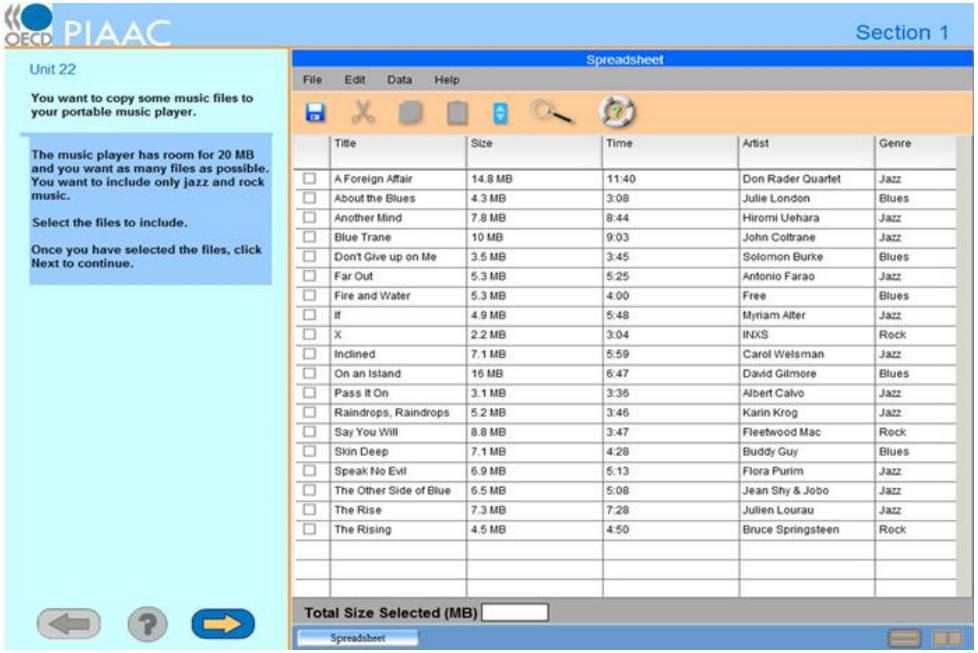
The Present Study

Characteristics	Total	US	NL	JP
N	**3926**	1340	1508	1078
Correct (%)	**2754 (70.1)**	882(65.8)	1104 (73.2)	768 (71.2)
Incorrect (%)	**1172 (29.9)**	458 (34.2)	404 (26.8)	310 (28.8)
Gender				
Female	2025	629	711	526
Male	1901	711	629	552
Age (years)				
Mean (S.D.)	39.60 (14.01)	39.21 (14.00)	40.84 (14.29)	38.35 (13.49)
Educational level				
Less than high school	615	124	401	90
High school	1493	534	590	369
Above high school	1812	680	513	619
Missing	6	2	4	0

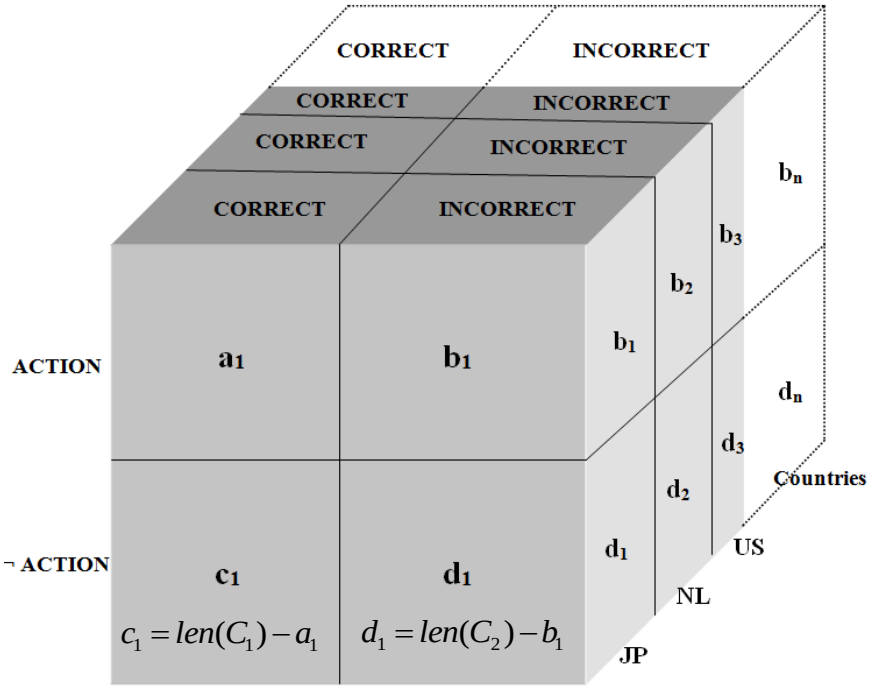
Note. US, NL and JP represent the sample from the United States, the Netherlands and Japan.

Instrument: A PSTRE Item

- The task is to identify the ID number of a specified person and send this number to a correspondent by email.
- Two environments are involved:
  - A spreadsheet environment that contains a database as the stimulus material that displays the information required to solve task.
  - An email environment to provide the response.
- The interim score is evaluated based only on the email responses.



Chi-square Feature Selection Model



$$\chi^2 = \frac{M(ad - bc)^2}{(a+b)(a+c)(b+d)(c+d)}$$

$$c = len(C_1) - a$$

$$d = len(C_2) - b$$

$$M = a + b + c + d$$

The actions with **higher chi-square scores** are **more discriminative** in classification. Therefore, we ranked the chi-square score of each action in a **descending order**. The actions ranked to the top were defined as the robust classifiers.

Feature Selection Models (2) Weighted Log Likelihood Ratio (WLLR)

- The product of probability of each action sequence and the logarithm of the ratio between conditional probability of the sequence in different performance groups.

$$WLLR(t, C_i) = P(t \mid C_i) \log \frac{P(t \mid C_i)}{P(t \mid \neg C_i)}$$

$$= P(t \mid C_i) \log \frac{P(t \mid C_i)}{Q(t \mid C_i)}$$

$P(t \mid C_i)$ the conditional probability of action t in the class C_i

$Q(t \mid C_i)$ the conditional probability of action t not in the class C_i

The higher the WLLR, the more likely the action belongs to class C_i

Conversely, the lower the WLLR, the more likely the action belongs to class $\neg C_i$

Results (1)
Features of Actions by Performance Groups

Robust Features of Actions and Action Sequences Distinguishing Correct and

	Unigrams		Bigrams			
	Actions	χ^2	Actions	χ^2		
Correct	SS	70.72	E, SS	229.99	E, S	272.49
	SS_Type_SN	68.04	SS, E	191.18	START, E, SS	226.42
	SS_So_OK	64.58	SS_So_OK, E	153.90	SS, E, E_S	211.37
	SS_So_1B	59.66	SS_So_1B, SS_So_OK	122.49	SS_So_OK, E, SS	150.25
			Type_SN, E	120.56	SS_So_1B, SS_So_OK, E	137.53
			Se, SS_Type_SN	98.21	SS, E, SS	133.85
			So, SS_So_1B	84.43	SS_Se, SS_Type_SN, E	108.55
	SS_So_2A		START, SS_Se	70.03	SS_Type_SN, E, SS	108.20
Incorrect	Next_C	892.80	TART, Next	2416.20	START, Next, FINALENDING	2420.26
	SS_Save	98.90	Next, Next_C	521.74	Next, Next_C ext	478.16
	SS_Type_PGN	33.19	Next_C, Next	504.22	START, E, Ne	399.02
	SS_H	15.75	E_S, E_S	492.26	Next_	
	SS_So_3D	14.56	E_S, E	364.66	E_S,	
	SS_So_C		S_SS	299.74	E, E_	
	E_S					
	SS_Type_PS				S, E	338.26

Correct group: using tools such as searching engine and sorting with a clear sub-goal

Incorrect group: hesitative behaviors using “cancel” a lot

Nonresponse pattern: START, Next, FINALENDING (NONRESPONSE)

Incorrect group: using “Help” function a lot and aimless save the results in the server

Results (2)
Country Level vs. Aggregate Level

Consistency Rate of Extracted Classifiers by Performance Groups Compared Between Country Level and Aggregate Level

	US	Netherlands	Japan
Correct			
Unigrams	0.88	0.88	0.63
Bigrams	0.75	0.88	0.75
Trigrams	0.75	0.88	0.75
Incorrect			
Unigrams	0.63	0.63	0.63
Bigrams	0.63	0.88	0.88
Trigrams	0.75	0.63	0.75

Mean=0.79

Mean=0.71

Results (3)
Features of Actions by Countries

Robust Features of Actions and Action Sequences Across Countries

	Unigrams		Bigrams			
	Actions	χ^2	Actions	χ^2		
US	Next_C	20.40	E, E	261.08	E, E, E	309.01
	SS_Type_FN	15.64	START, Next	39.82	E, E, Next	278.87
	E	13.25	Next, E	39.28	SS, E, E	132.21
	SS_Type_PGN	10.14	START, E	38.97	START, E, E	85.14
	SS_Save	6.22	SS_So_C, SS_Type_FN	37.63	SS_Type_FN, E, E	54.23
NL	SS_Type_FN	315.30	SS_Se, SS_Type_FN	252.93	START, SS_Se, SS_Type_GN	226.67
	SS_Type_GN	232.93	SS_Type_FN, SS_Type_FN	249.97	START, SS_Se, SS_Type_FN	226.67
	SS_Se	60.88	SS_Type_FN, E	203.30	SS_Type_FN, SS_Type_FN, E	161.66
	SS_So_3B	31.59	SS_Se, SS_Type_GN	202.10	SS_Type_FN, SS_Type_FN, SS_Type_GN	161.66
	SS_So_2A	16.15	START, SS_Se	117.42	SS_Se, SS_Type_FN, SS_Type_FN	161.66
JP	SS_Type_SM	383.58	SS_Type_SM, SS_Type_SM	308.58	SS_Type_SM, SS_Type_SM, SS_Type_SM	248.84
	SS_Type_null	123.49	SS_Type_SM, SS_So	166.12	E_S, Next, Next_C	149.25
	SS_Type_UM	70.75	E_S, Next	137.22	SS_Type_SM, SS_So, SS_So_1B	149.21
			SS_Type_SM, SS_Type_SM	116.73	SS_Type_SM, SS_Type_SM, SS_So	140.96
			SS_Type_SM, E	115.33	SS_Type_SM, SS_Type_SM, E	116.15

US: Double clicks on E-mail page

NL: More likely use full name and given names when doing searching

JP: Spelling mistakes (optimal space between first name and last name)

JP: strategy changed

Exploring relationship between background variables and behavioral patterns

He, Ling, Liu, & Ying (2019)

Research Questions

1. Study whether information from the process data could help improve the assessment of problem solving proficiency; if it can, then what is the information that can help?

2. Explore the relationship between background variables and the action sequences. How powerful is the process data to make prediction on background variables?

The Present Study

- Six countries that participated in PIAAC Round 1, including Finland, the Netherlands, Austria, Ireland, the United States and Poland.
- A total of 8,663 test takers who completed 7 PSTRE items in PIAAC PS2.
- The background variables include
  - Country
  - Age
  - Gender
  - Education level
  - Working status
  - Whether the test taker use computer at home/at work
  - Whether the test taker is an employer
  - Income level
  - Derived scores in ICT at home, ICT at work, numeracy at home, numeracy at work, reading at home, reading at work, writing at home and writing at work.

RQ1 Can information from the process data help improve the assessment of problem solving proficiency?

- Predictors include the numbers of different unigrams, bigrams, trigrams, the total number of actions, response time and the responses for each item.
- Since the total number of such predictors could be large (a few thousands), to improve prediction and interpretability of the variables, least absolute shrinkage and selection operator (LASSO) is performed.
- We carry out the estimation using training data (70% of the data) and compute the out-of-sample correlation of the PSTRE score and the predicted value as well as the mean squared error of the prediction in the testing data (the remaining 30% of the data).

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta^T x_i)^2 + \lambda||\beta||_1.$$

RQ1

Can information from the process data help improve the assessment of problem solving proficiency

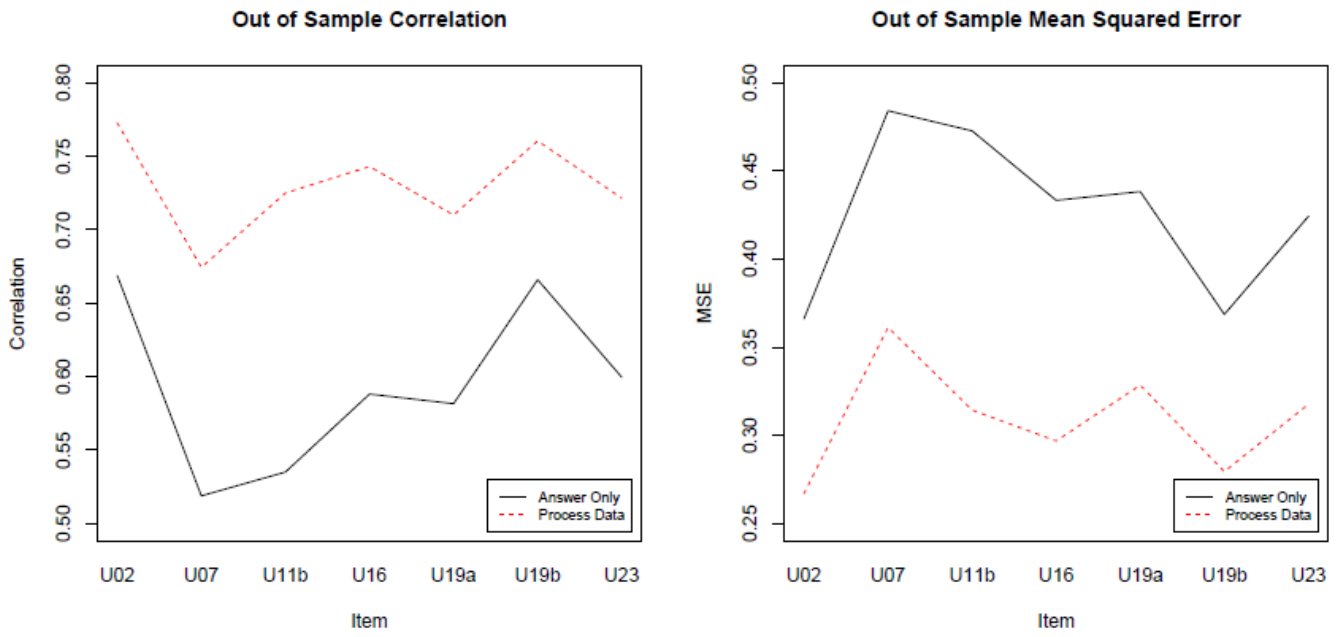


Figure 2. The average out of sample correlation (left) and mean squared error (right) using predictors extracted from only one particular item. The red dotted line shows the performance of using the process data and the black solid line shows the benchmark performance using only the responses as the predictors.

RQ1 Can information from the process data help improve the assessment of problem solving proficiency

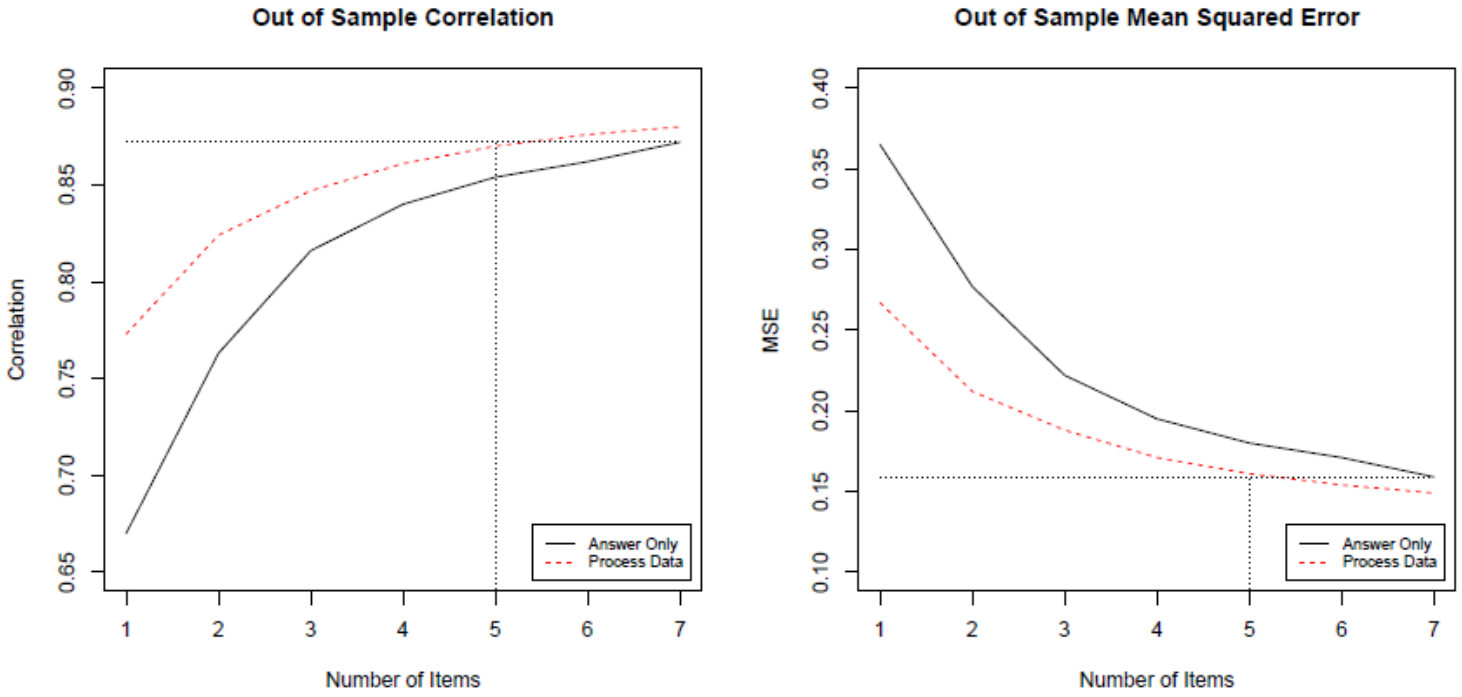


Figure 3. The average out of sample correlation (left) and mean squared error (right) using predictors extracted from different number of items. The red dotted line shows the performance of using the process data and the black solid line shows the benchmark performance using only the responses as the predictors.

RQ2 How powerful is the process data to make prediction on background variables?

- To explore the relationship between background variables and action sequences, we regress background variables on action sequences.

- This is because most of the action sequences alone contain relatively few information about a person. On the other hand, aggregating weak information from each of the action sequences may tell us more about a person.

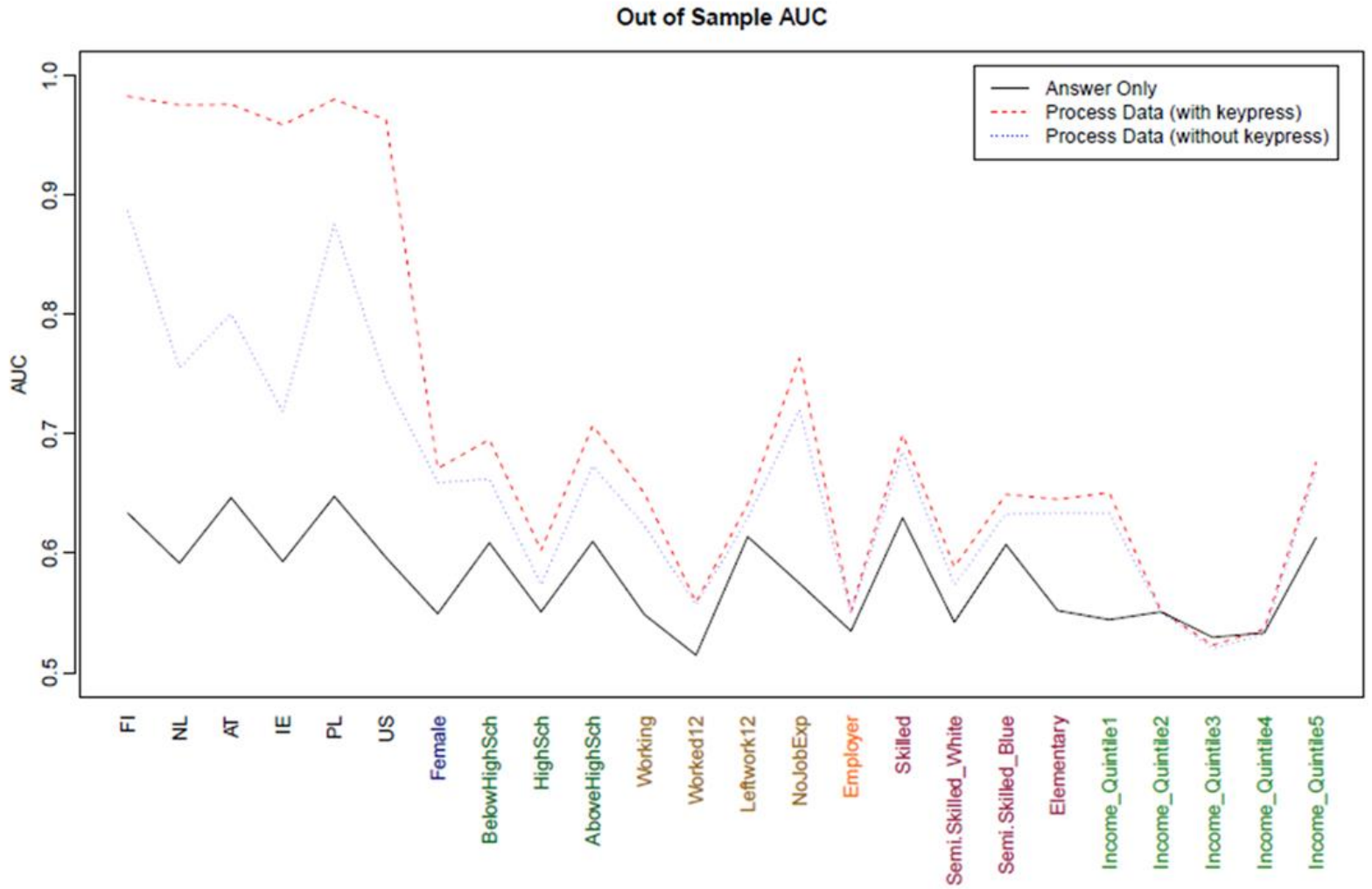
RQ2 How powerful is the process data to make prediction on background variables?

- Out-of-sample area under the receiver operating characteristics curve (AUC) was used as a measure of information.

- If there is an improvement in the AUC compared with the one using only the responses as the predictors, then the action sequences contain additional information about the background of a person. This also means there are differences in the action sequences for people with different background.

RQ2

How powerful is the process data to make prediction on background variables?

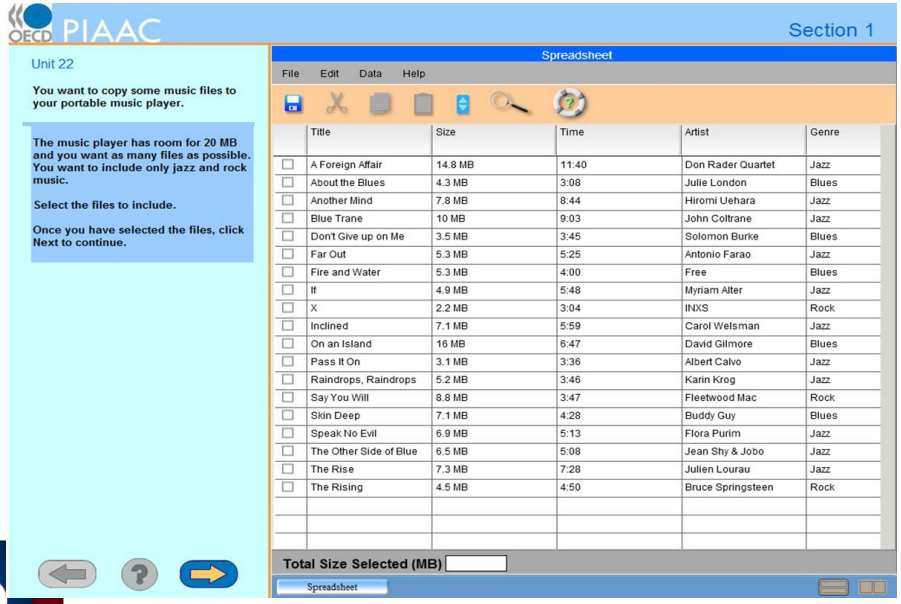


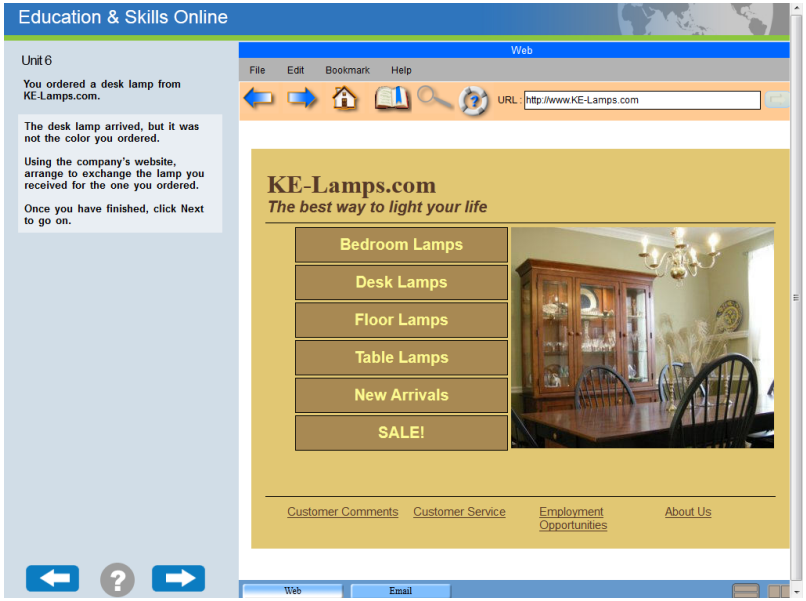
Identifying generalized patterns across multiple tasks with sequence mining

He, Borgonovi, & Paccagnella (2019)

Challenges

- With the rapid growth of advanced techniques and computer-based testing, more and more scenario-based interactive items have been used in international large-scale assessments, such as PISA, PIAAC and NAEP.
- In the context of large-scale assessments, items designed to test problem solving skills generally embed the problem within a particular context or situation.





Challenges

- Insights are to be gained by investigating generalized patterns of respondents' behaviors across multiple tasks, in different context and scenarios.

- The most challenging aspect is how to define aggregate-level variables across items and derive standardized measures in complex data structures across multiple items.

Longest Common Subsequence

- This study explores the use of the Longest Common Subsequence (LCS) method (Maier, 1978; Hirschberg,1975; Chvatal & Sankoff, 1975), a sequence-mining technique used in natural language processing and biostatistics.
- The longest common subsequence was first introduced into educational assessment by Sukkarieh, Yamamoto, & von Daiver (2012) as a tool for automated scoring in multiple linguistic environment.
- The main idea of this method is simple: to identify the action sequences that are most similar to predefined, "optimal" sequences for each item.
- Measurement indicators are developed in order to analyze behaviors across items and subgroups of respondents.
- This approach extends the research capacity from understanding individuals' problem-solving behaviors in a single item to a general perspective across multiple items that form an assessment.

Research Questions

1. Do people adopt consistent problem solving strategies across different items?

2. What is the association between the adoption of specific patterns of problem-solving strategies and problem-solving proficiency?

3. Do patterns of problem-solving processes differ systematically by background variables, e.g., gender, age, and ICT familiarity?

4. How LCS methods can be used to improve the quality of items?

The Present Study

- The Programme for International Assessment of Adult Competencies (PIAAC) Round 1, problem solving in technology-rich environment (PSTRE) domain.
- Second module (fixed 7-item booklet) PSTRE, meaning each respondent has 7 PSTRE items in a row. The item is in fixed position.
- 5 countries: GBR, IRL, JPN, NLD, USA
- 8988 respondents

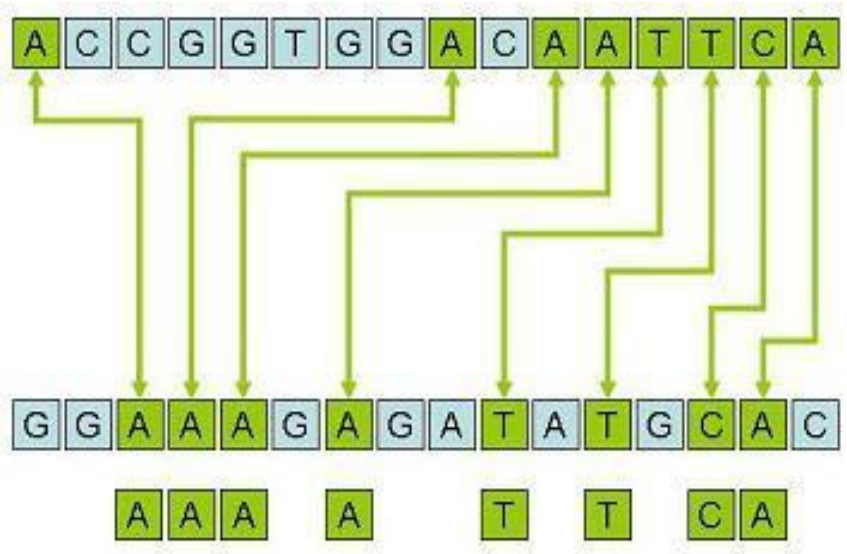
Methods – Longest Common Subsequence

S1 (observation)

Len(S1)=16
Len(S2)=16
Len(LCS)=8

S2 (reference)

LCS



- The pre-defined action sequences were built on the optimal paths designed from item developers and content experts.
- Multiple optimal paths may be designed in one item in order to solve the task.

Longest Common Subsequence (1)

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

Longest Common Subsequence (2)

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:

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

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

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

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

PDAS_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

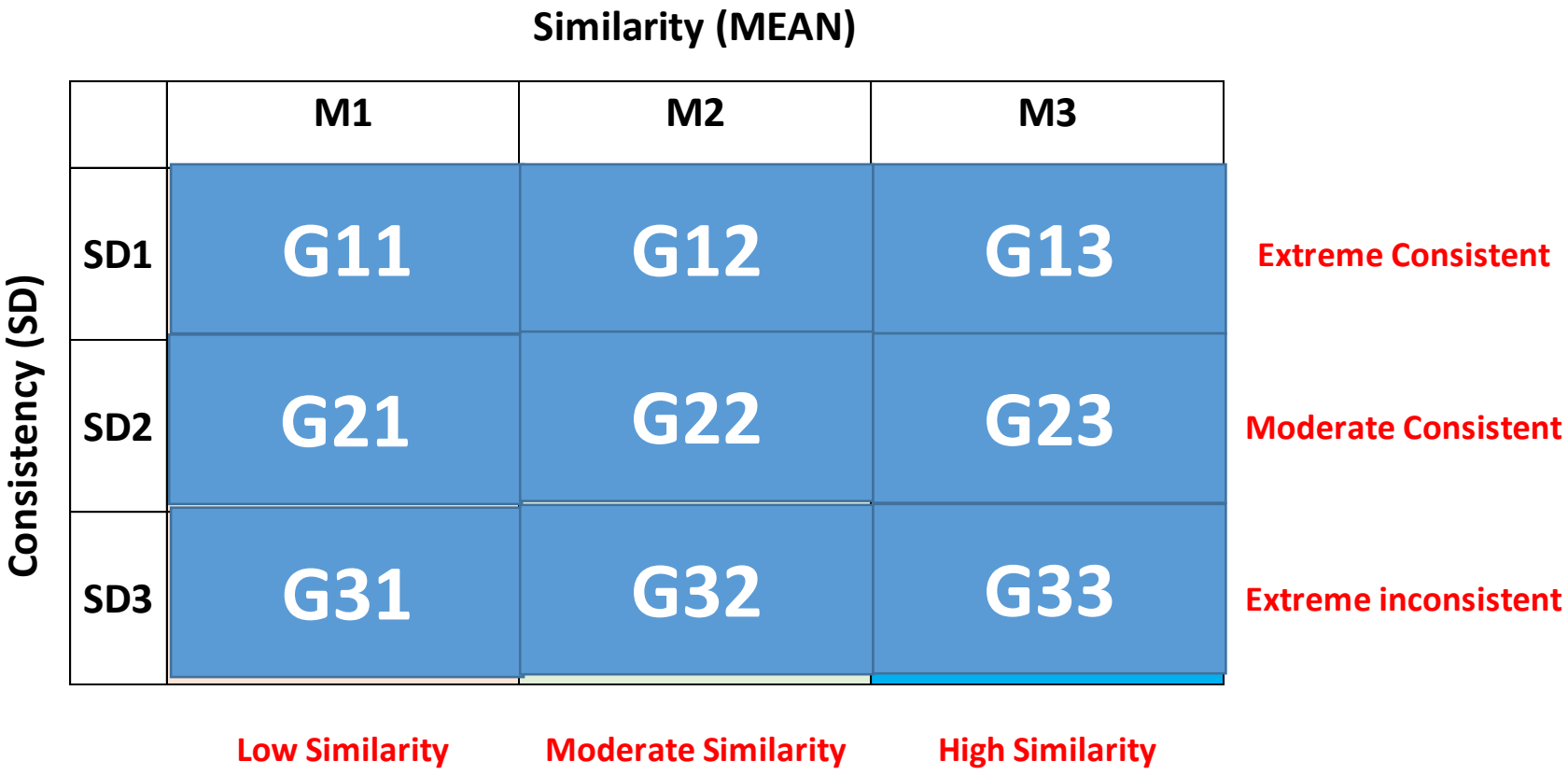
LCS1 (length=6): Start, Email, On_Email_Message, Off_Email_Message, Next, Next_OK
LCS2 (length=6): Start, Email, On_Email_Message, Off_Email_Message, Next, Next_OK
LCS3 (length=8): Start, Sort_1_B, Sort_OK, Email, On_Email_Message, Off_Email_Message, Next, Next_OK
LCS4 (length=9): Start, Menuitem_Sort,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

LCS Indicators Across Items

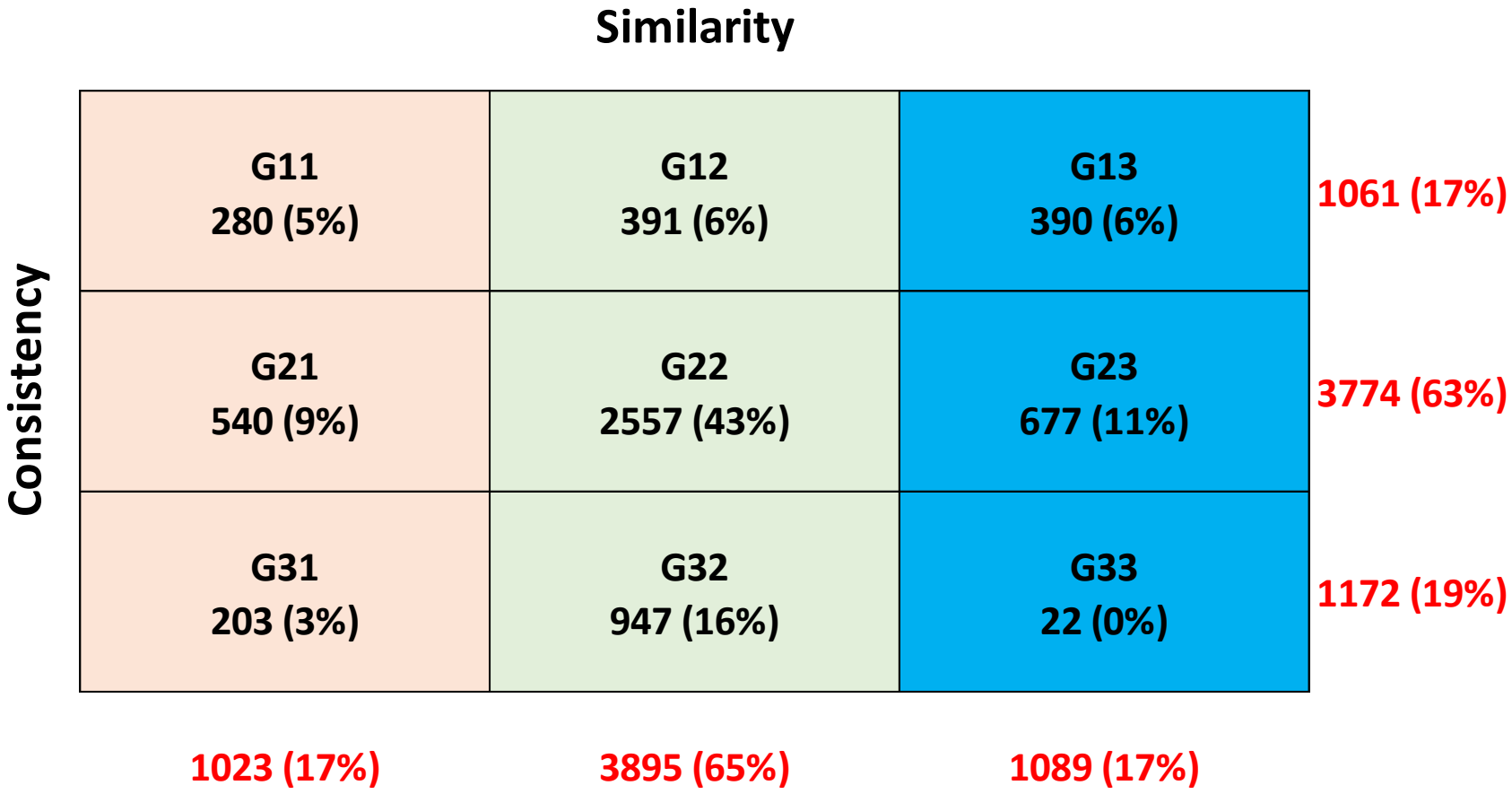
- Similarity
  - $Similarity = \text{len}(LCS)/\text{len}(PDAS)$
  - $SM = \text{Mean}(Sim_1, Sim_2, \dots, Sim_n)$
  - $SSD = \text{SD}(Sim_1, Sim_2, \dots, Sim_n)$

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

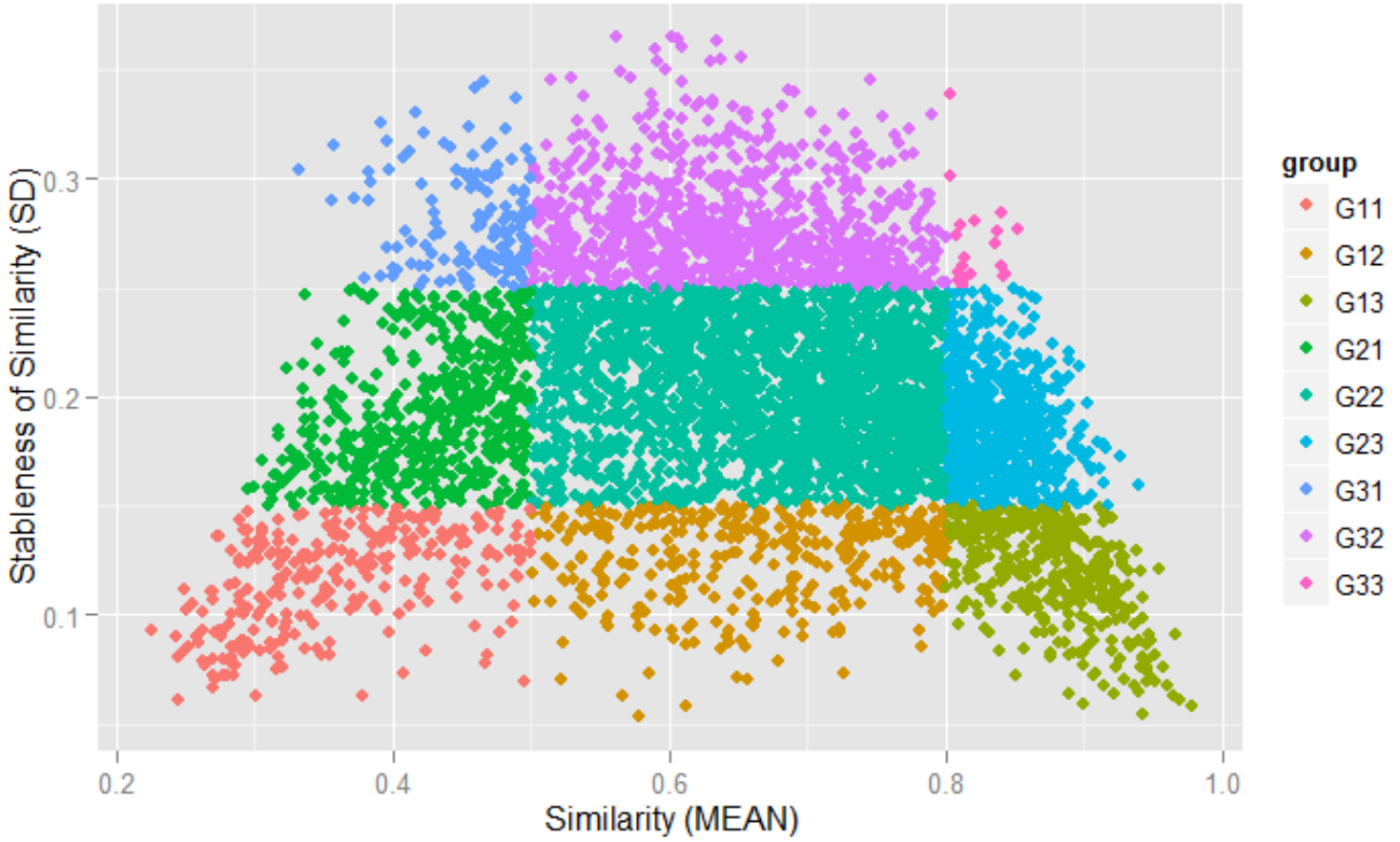
Mapping Similarity and Consistency of Similarity



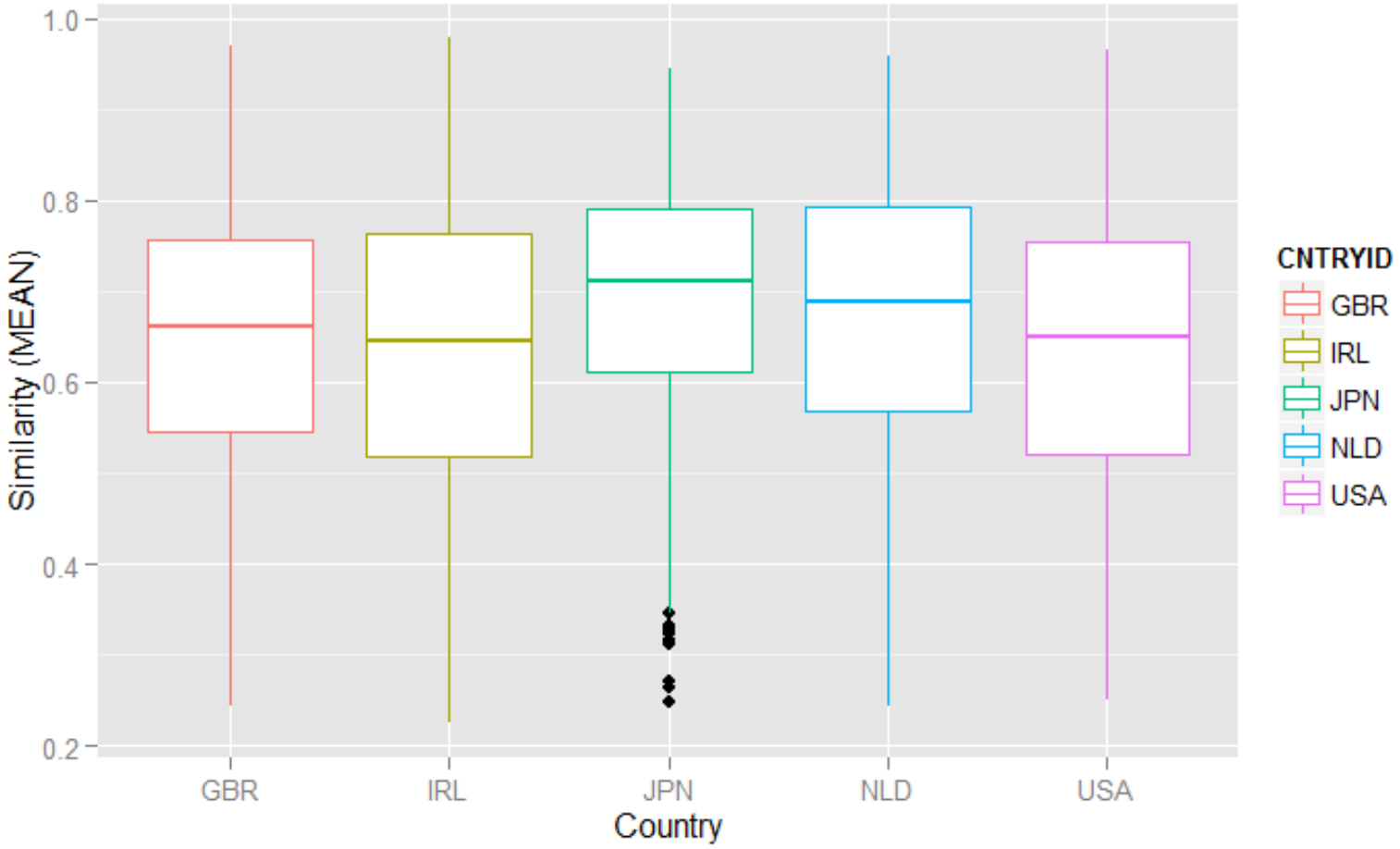
RQ1 Do people adopt consistent problem solving strategies across different items?



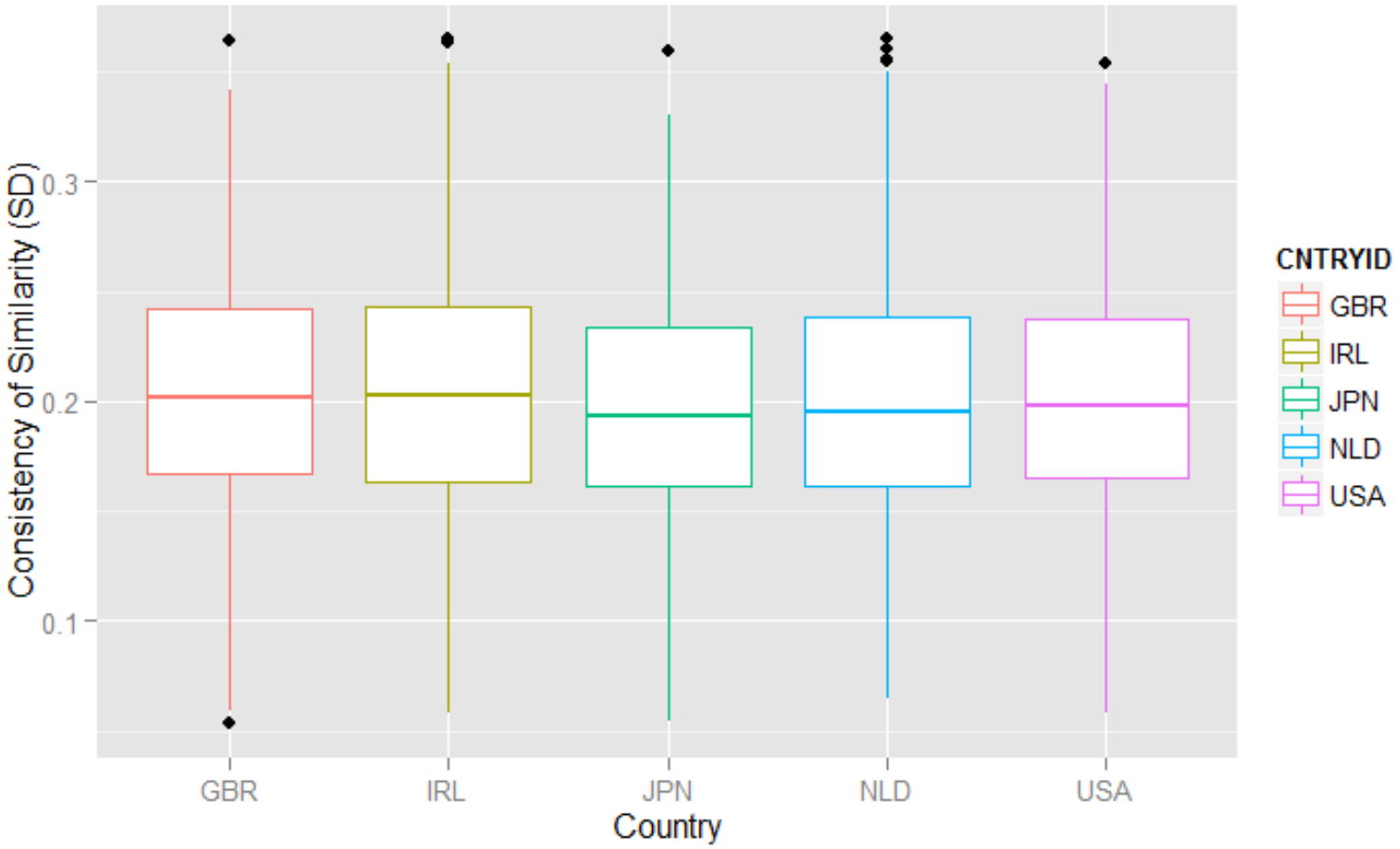
Similarity and Consistency of Similarity



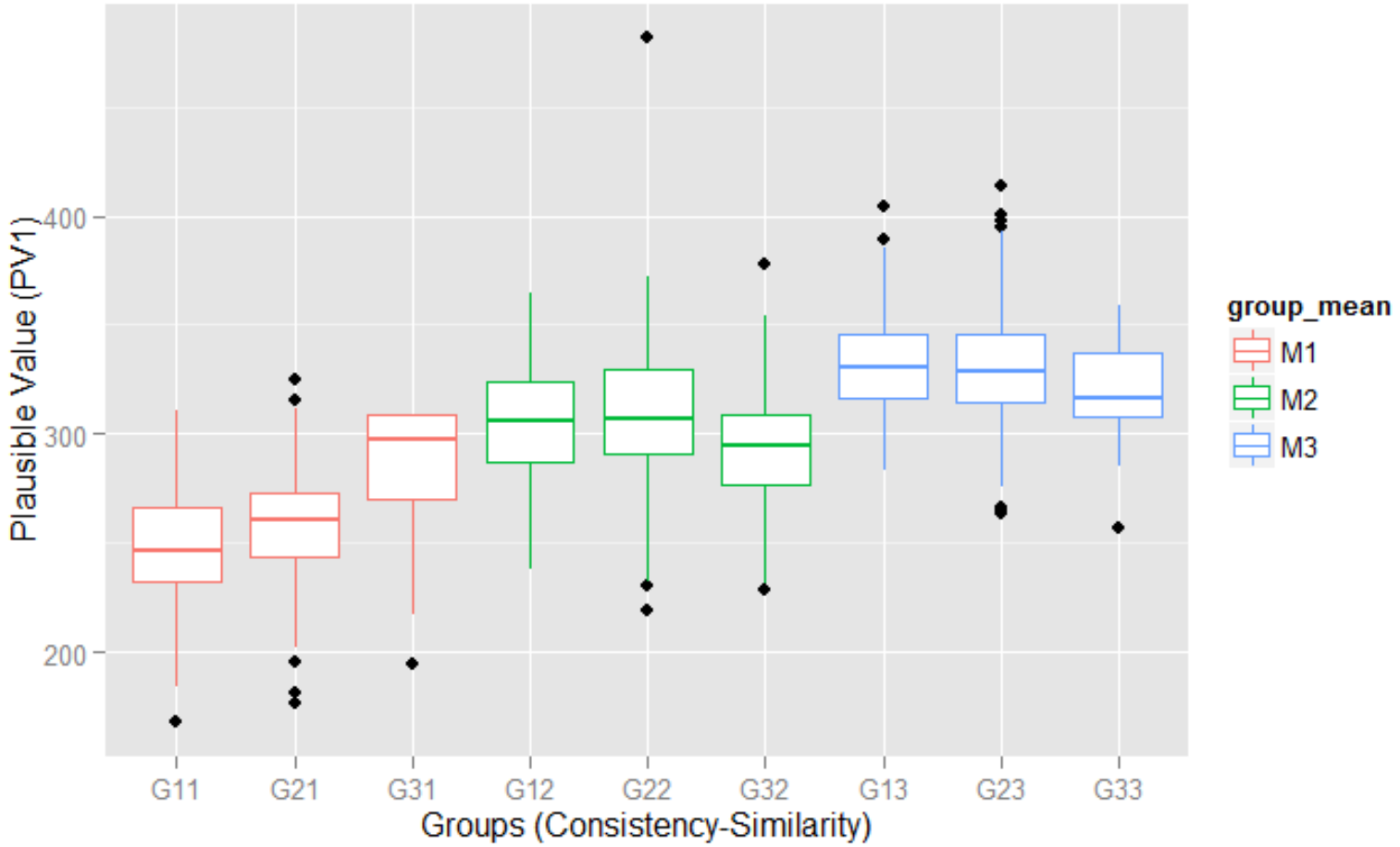
Similarity Across Countries



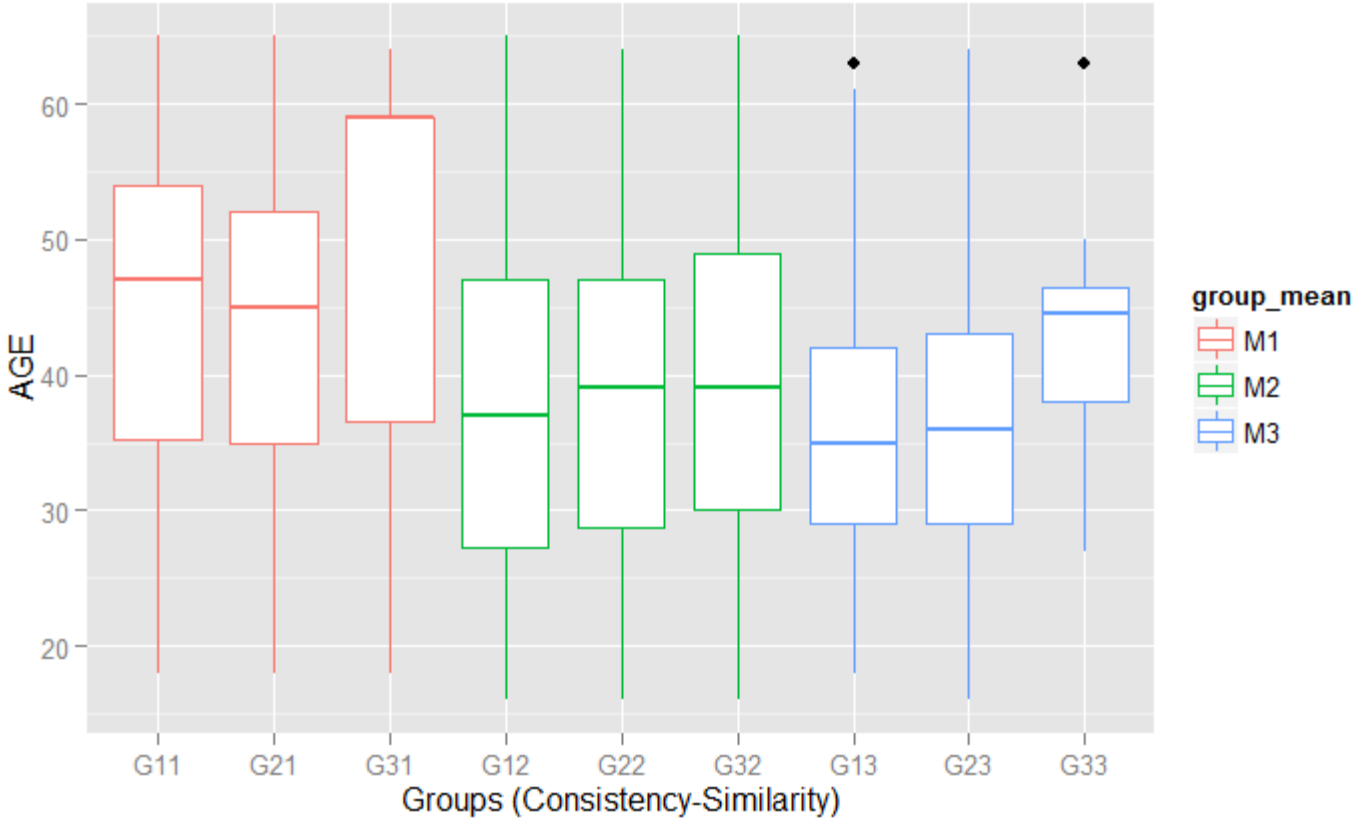
Consistency of Similarity Across Countries



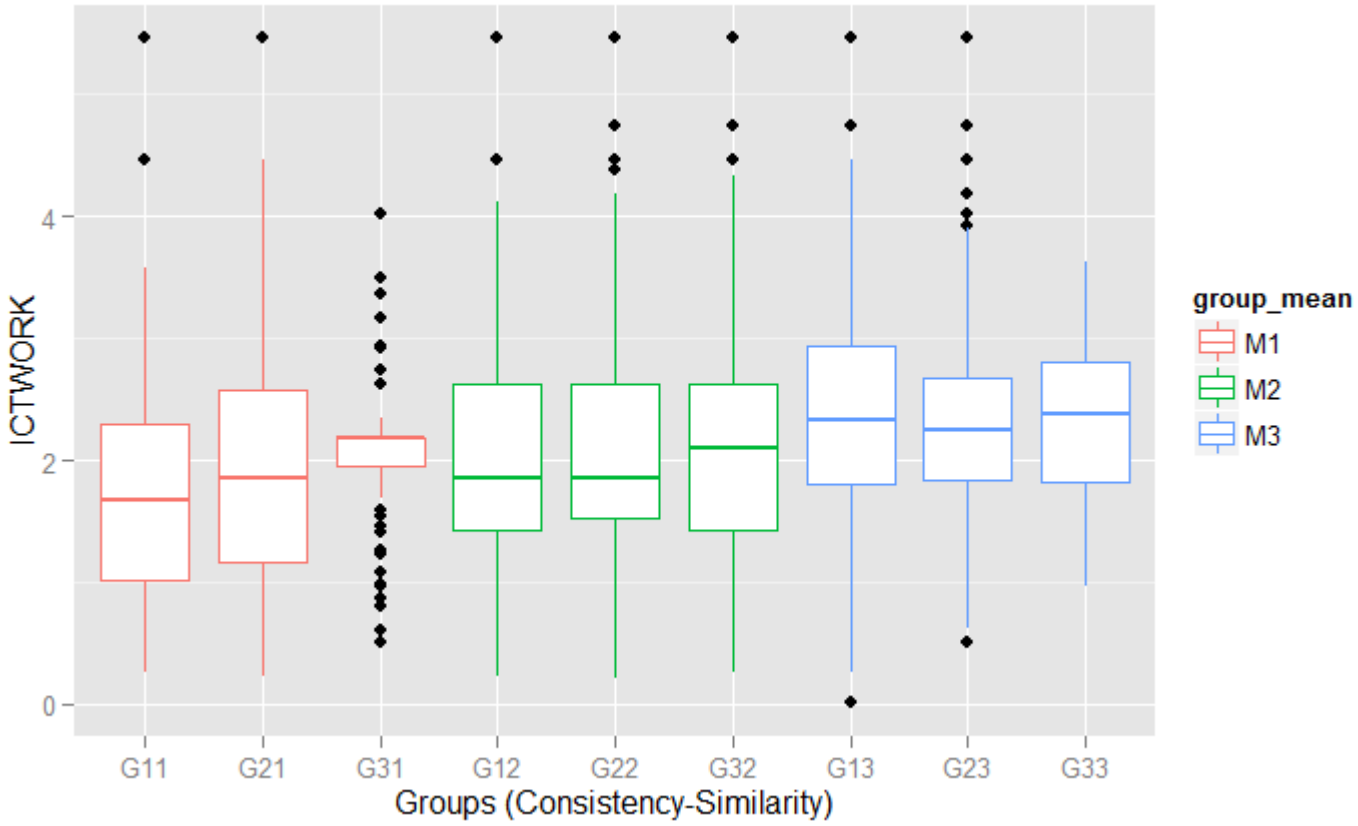
RQ2 What is the association between problem-solving strategies and proficiency?



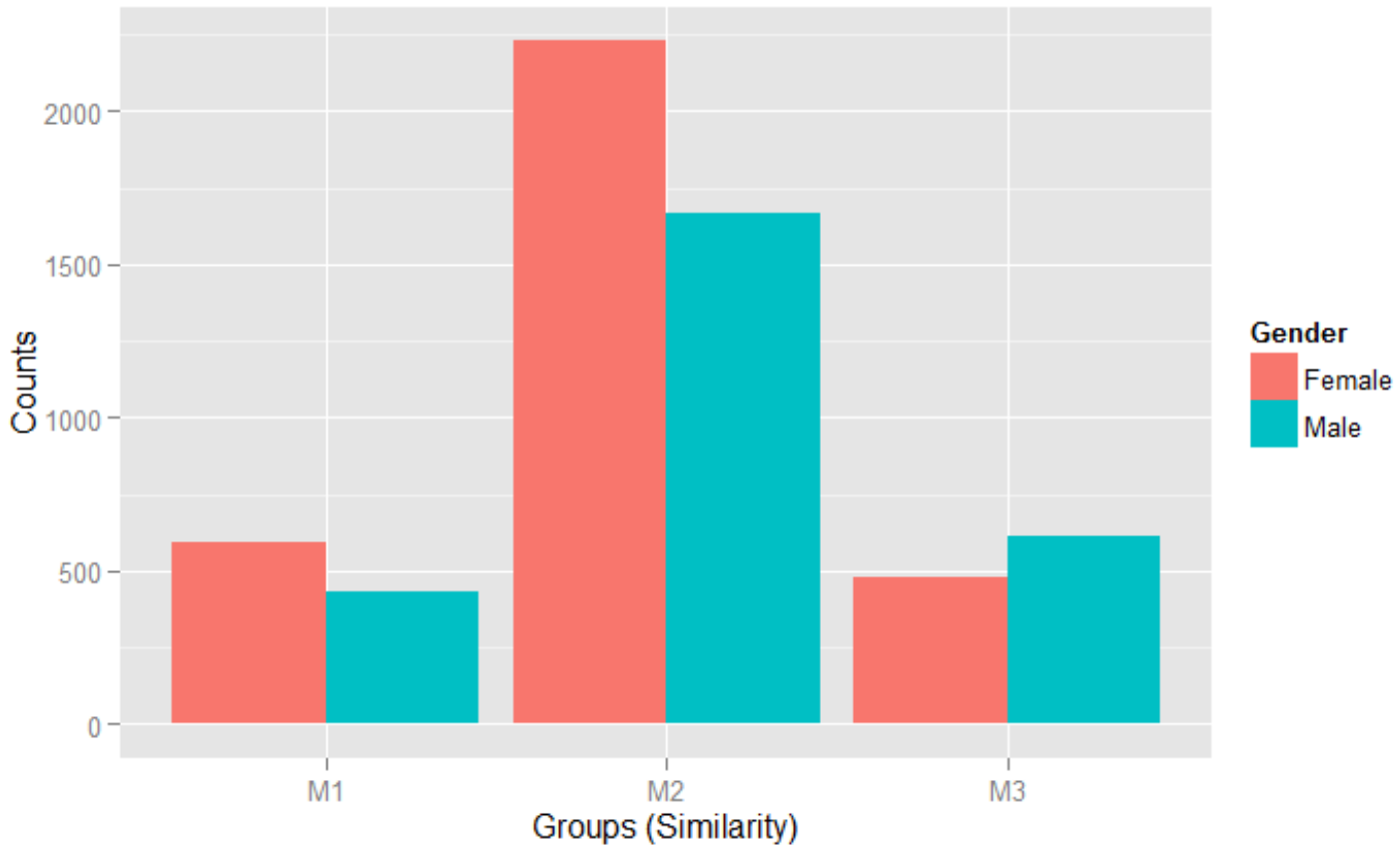
RQ3 Do patterns of problem-solving processes differ systematically by background variables?



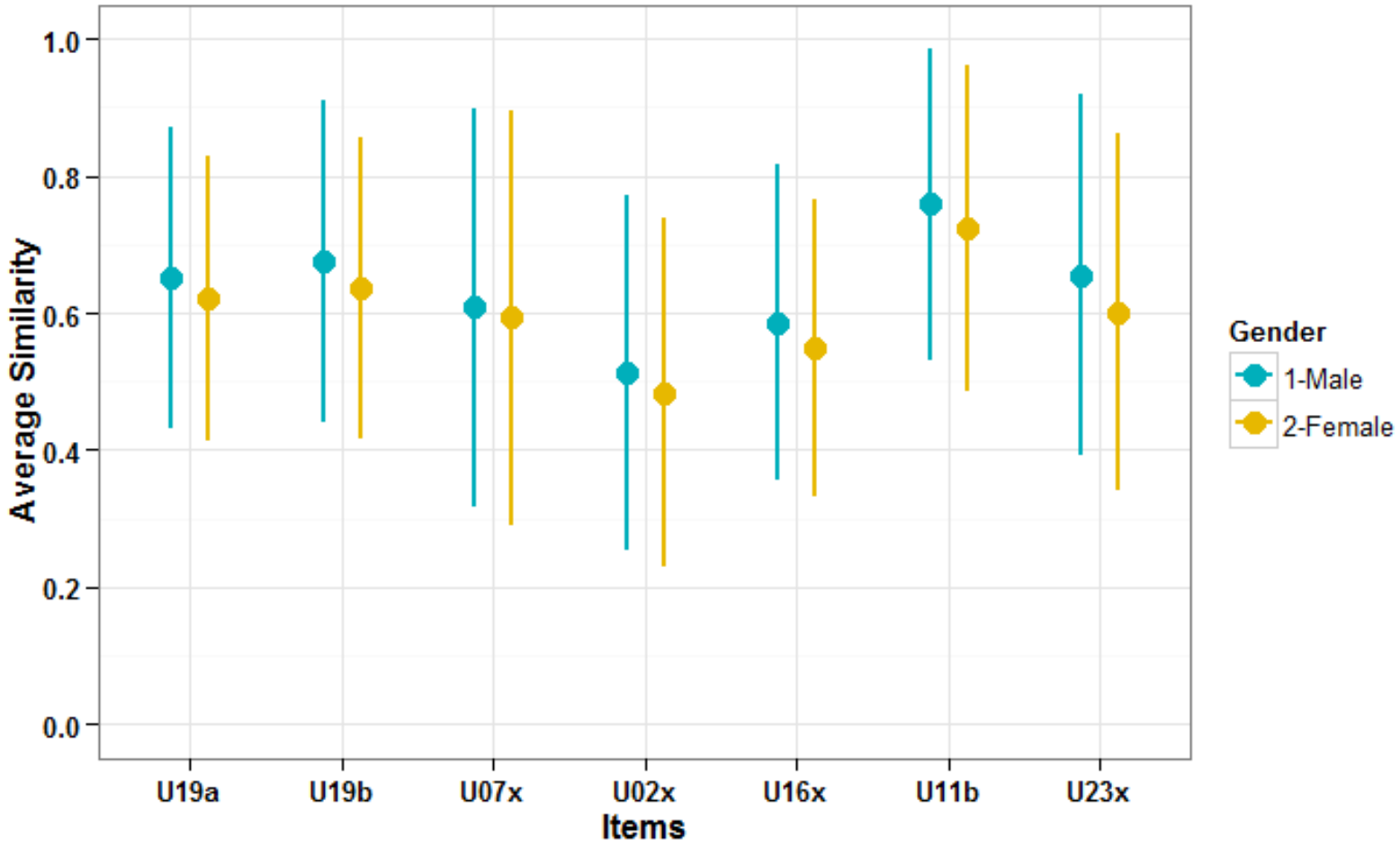
Similarity Measure with ICTWORK



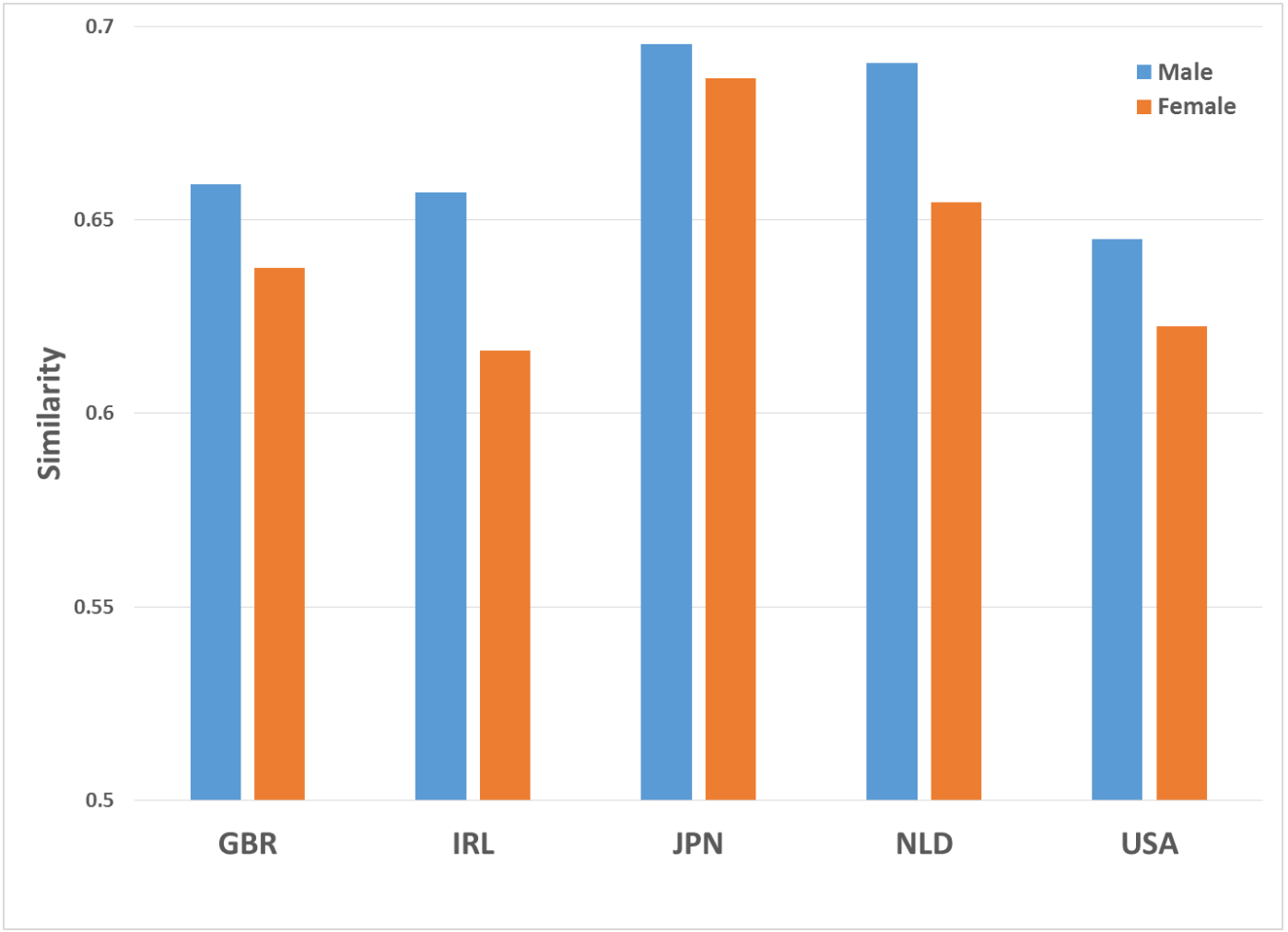
Similarity Measure with Gender



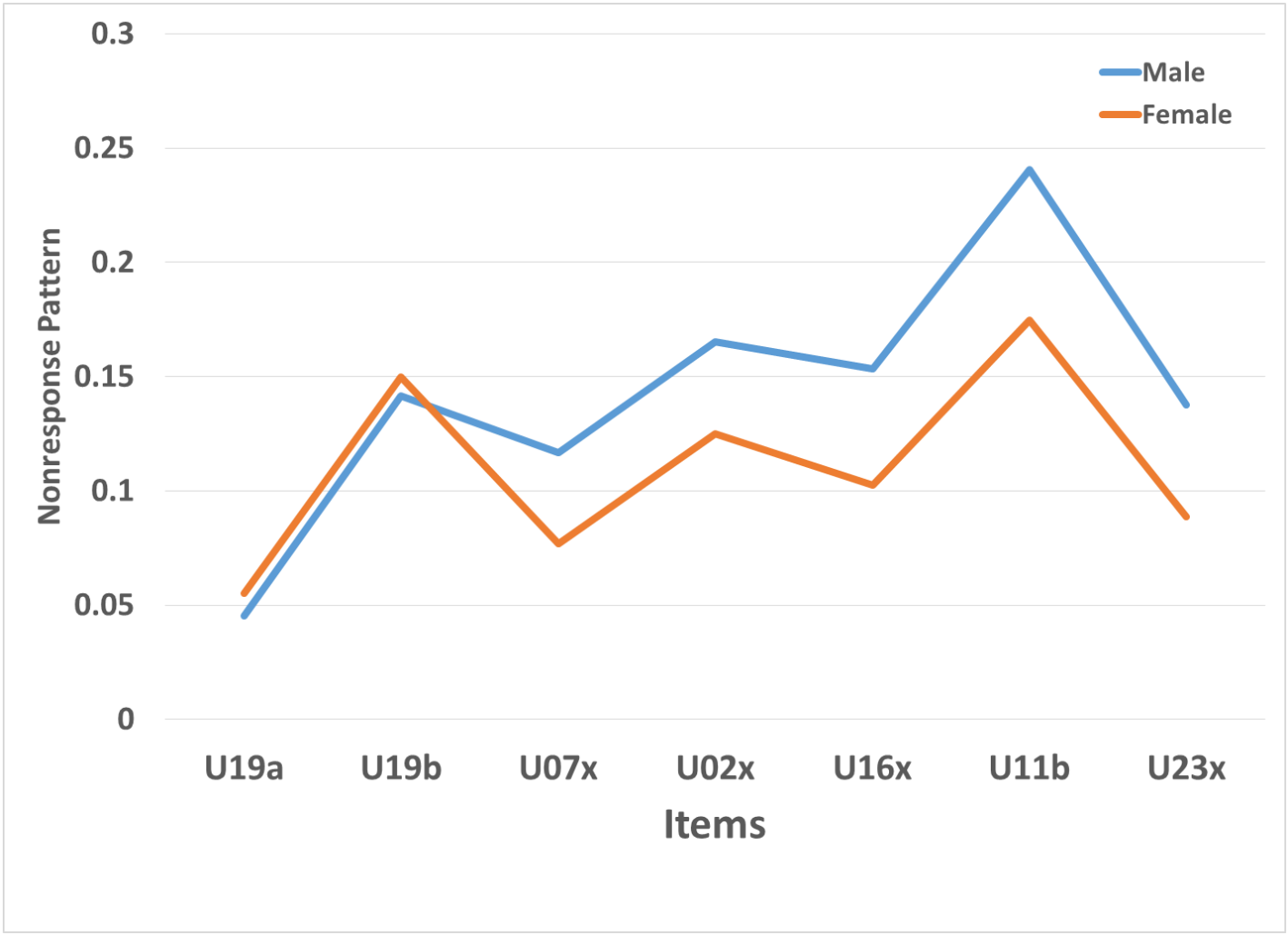
Comparisons of Similarity Between Gender by Items



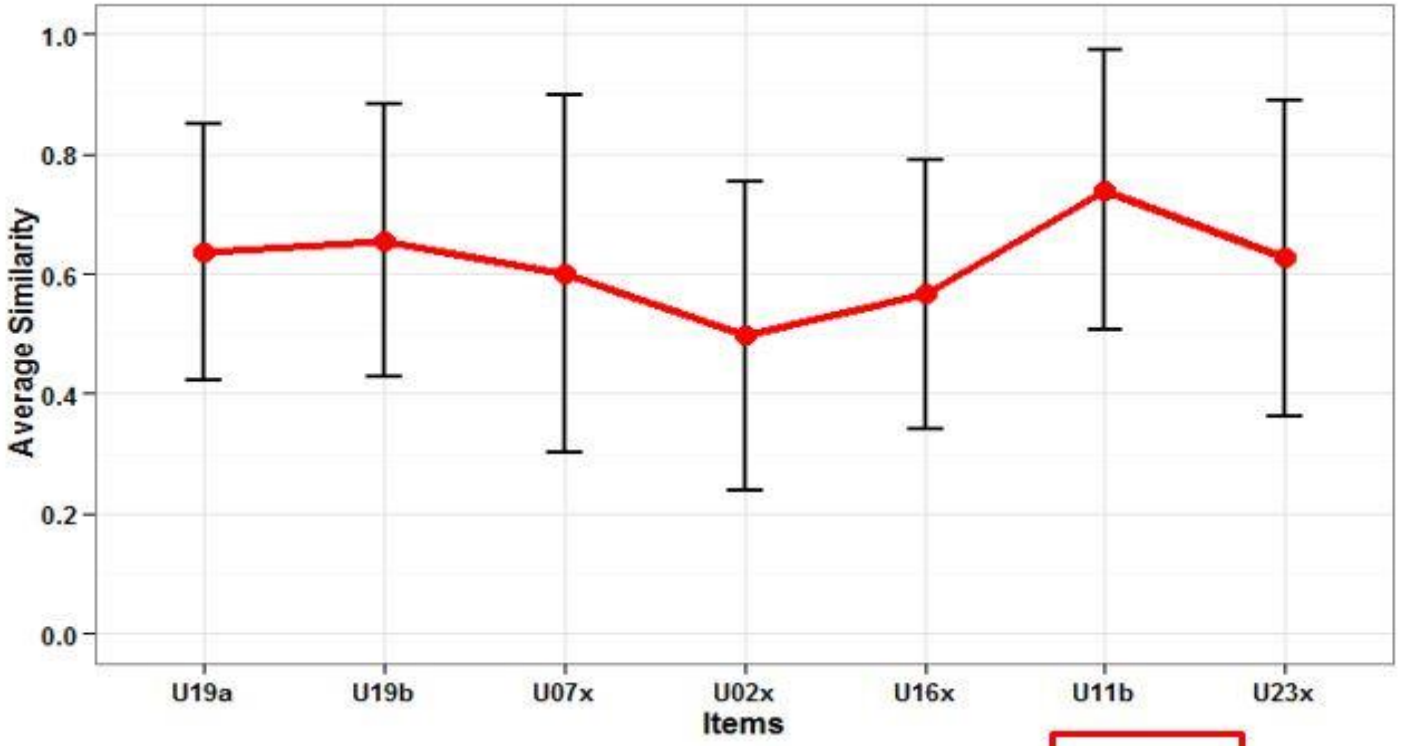
Comparisons of Similarity Between Gender by Countries



Persistence (Nonresponse Patterns) Between Gender

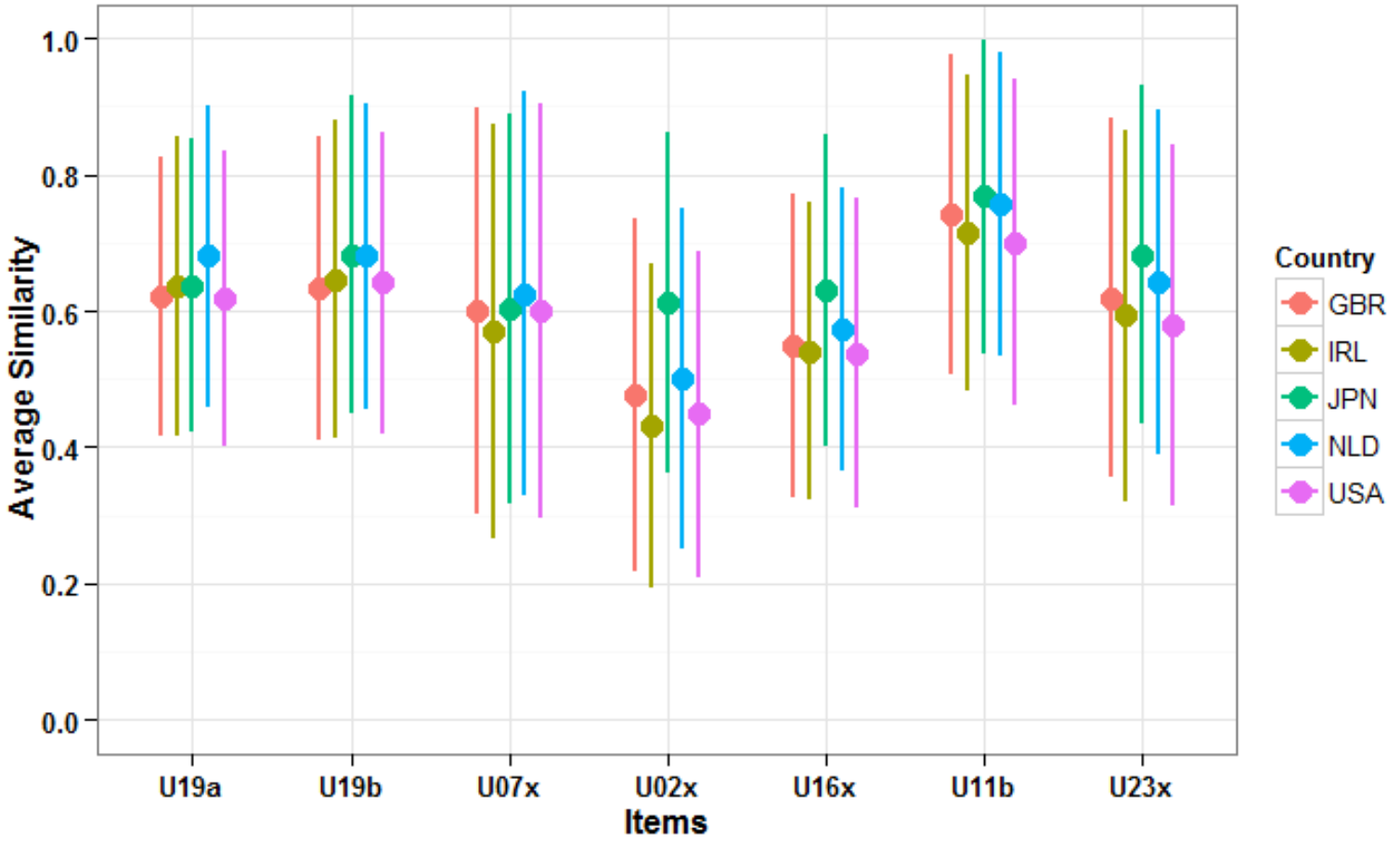


RQ4 How LCS methods can be used to improve the quality of items?



	U19a	U19b	U07x	U02x	U16x	U11b	U23x
a-slope	1.414	1.072	1.104	1.184	1.377	0.471	0.533
b-difficulty	-1.367	-0.677	-0.237	0.784	-0.773	0.774	-0.052
RP67 level	1	2	2	3	1	3	2

Comparisons of Similarity Across Countries by Items



Conclusions & Discussions

Discussion and Conclusion

- The sequence-based approaches hold a great promise in process data analysis.
- N-grams method is more helpful in checking the item quality and understanding test takers' behaviors on specific items.
- Longest common subsequences method provides the possibility to generalize factors that are associated with test takers' problem-solving behaviors across multiple items.
- The sequence-based approaches are also promising to automatically identify test takers' strategies and detect the DIF items and check differences between groups (e.g., countries, gender, background variables).
- Response time and time interval between actions would also be interesting to be added in the future study.

Selected Publications

- Han, Z., He, Q., & von Davier, M. (2019, under review). Predictive feature generation and selection using process data in PISA simulation-based environment: An application of tree-based ensemble methods. *Frontiers in Psychology*.
- He, Q., Liao, D., & Jiao, H. (2019, in press). Clustering behavioral patterns using process data in PIAAC problem-solving items. In C. Sluiter & B. Veldkamp (Eds.), Theoretical and practical advances in computer-based educational measurement. Springer.
- 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., & 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.
- He, Q., von Davier, M., & Han, Z. (2018). Exploring Process Data in Computer-based International Large-scale Assessments. In H. Jiao, R. Lissitz, & A. van Wie (Eds.), Data Analytics and psychometrics: Informing Assessment Practices (pp. 53-76). Charlotte, NC: Information Age Publishing.
- 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. doi:10.3389/fpsyg.2019.00646
- Liao, D., He, Q., & Jiao, H. (2019, under review). 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.
- Tang, X., Wang, Z., He, Q., Liu, J. & Ying, Z. (2019, under review). Latent feature extraction for process data via multidimensional scaling. Applied Psychological Measurement.

Thank you very much!

For further information and suggestions, please contact

Dr. Qiwei (Britt) He
qhe@ets.org