Exploring Response Patterns in Problem-Solving Items Using Process Data

Insights from Log Files of PSTRE in PIAAC

Qiwei (Britt) He and Matthias von Davier

Center for Global Assessment

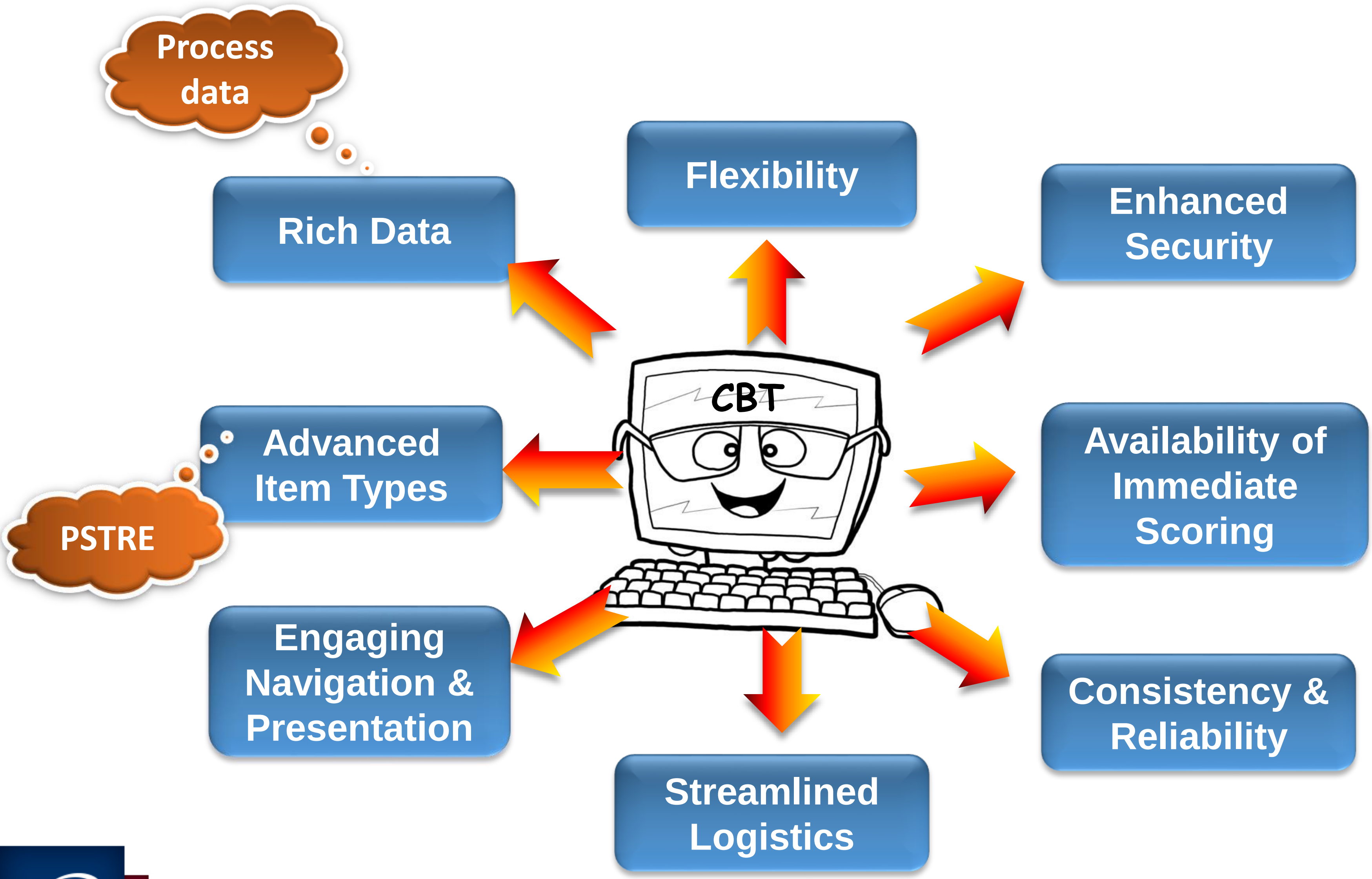
Educational Testing Service

What We Need Skills for: AIR-PIAAC Research Conference 2015, Arlington, VA, USA
December 10-11, 2015

Outline

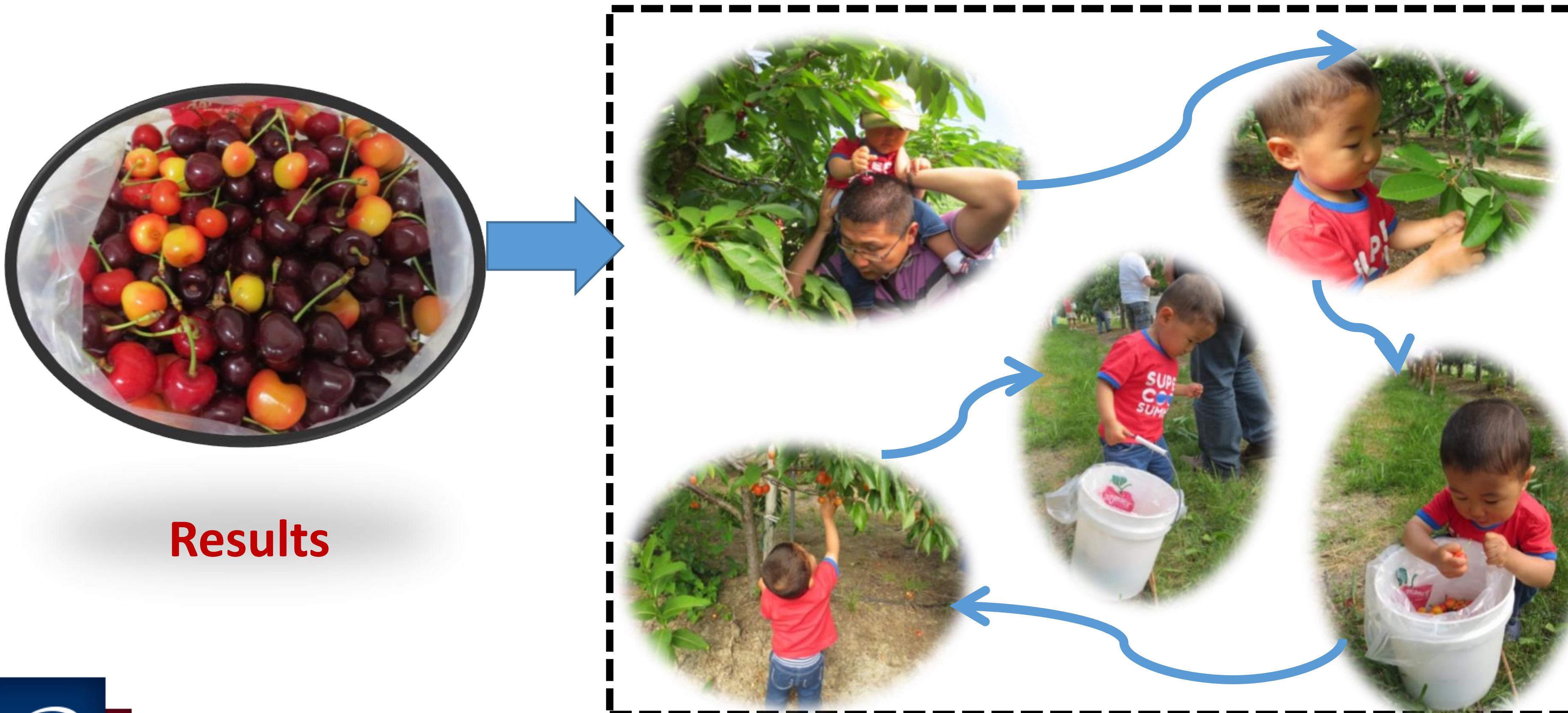
- Background
  - Computer-based testing and PSTRE items in PIAAC
  - Process data in log files
  - Motivation and research questions
- Exploring response patterns in PSTRE using process data
  - Sample and instrument
  - Methods: n-grams and sequential feature extraction model
  - Results and Discussions
- Conclusions and future work

Benefits of Computer-Based Testing



Process Data from Log Files

- In CBTs, a variety of timing and process data accompanies test performance data. This means that much more than data is available besides correctness or incorrectness.



Log Files


```
<events>
  <event>
    <eventCounter><![CDATA[0]]></eventCounter>
    <unitId><![CDATA[XYZ/Orientation-General]]></unitId>
    <itemName><![CDATA[item1]]></itemName>
    <itemId><![CDATA[GeneralOrientationQ01]]></itemId>
    <target><![CDATA[MODULE]]></target>
    <event_name><![CDATA[onItemBegin]]></event_name>
    <time><![CDATA[1395912781102]]></time>
    <lang><![CDATA[eng-IRL]]></lang>
  </event>
  <event>
    <moduleId><![CDATA[platform]]></moduleId>
    <eventCounter><![CDATA[1]]></eventCounter>
    <unitId><![CDATA[XYZ/Orientation-General]]></unitId>
    <itemName><![CDATA[item1]]></itemName>
    <itemId><![CDATA[GeneralOrientationQ01]]></itemId>
    <target><![CDATA[MODULE]]></target>
    <event_name><![CDATA[stimulusLoaded]]></event_name>
    <time><![CDATA[1395912783070]]></time>
    <lang><![CDATA[eng-IRL]]></lang>
  </event>
```


Raw Log Files

Our research draws on process data recorded in log files in the computer-based large-scale programs to address **how sequences of actions recorded in problem-solving tasks are related to task performance**.

studentId	eventCou	itemId	event_na	target	id	lang	time
840-51-01-003-00025	0	CS633Q00	onItemBeg	MODULE		eng-USA	1397193846084
840-51-01-003-00025	1	CS633Q00	stimulusLoaded	MODULE		eng-USA	1397193846545
840-51-01-003-00025	2	CS633Q00	QuestionLoaded	MODULE		eng-USA	1397193846570
840-51-01-003-00025	3	CS633Q00	StimulusAndQuestionLoa	MODULE		eng-USA	1397193846570
840-51-01-003-00025	4	CS633Q00	onItemEnd	MODULE		eng-USA	1397193850305
840-51-01-003-00025	5	CS633Q000	click	li	next	eng-USA	1397193851030
840-51-01-003-00025	6	[illegible]	[illegible]	[illegible]		[illegible]	[illegible]
840-51-01-003-00025	7	CS633Q000	QuestionLoaded	MODULE		eng-USA	1397193851283
840-51-01-003-00025	8	CS633Q000	stimulusLoaded	MODULE		eng-USA	1397193851427
840-51-01-003-00025	9	CS633Q000	StimulusAndQuestionLoa	MODULE		eng-USA	1397193851427
840-51-01-003-00025	10	CS633Q000	click	div	roof-color	eng-USA	1397193854737
840-51-01-003-00025	11	CS633Q000	click	span	stimulus_13	eng-USA	1397193855055
840-51-01-003-00025	12	CS633Q000	click	input	roofColorRadioRed	eng-USA	1397193855061

Structured Process Dataset

Motivations

- **Obtain insights:** how these action sequences are associated with different ways of cognitive processing and to identify key actions that lead to success or failure.

- **Ongoing improvement:** The results can be useful for test developers, psychometricians, and instructors to help them better understand what distinguishes successful from unsuccessful test takers and may eventually contribute to improved task and assessment design.

The Main Elements of PIAAC

Direct Assessment

- Numeracy
- Literacy
- Reading components
- **Problem solving in technology-rich environments (PSTRE)**

Module on Skills Use

- **Cognitive skills**: re[...] writing, math and use of ICTs
- **Interaction a[...]kills**: collaboration, planning,

Background Questionnaire

- Social and linguistic background

- PSTRE items are used to assess the skills required to solve problems for personal, work, and civic purposes.
- Test takers have to set up appropriate goals and plans, and access and make use of information through computers and networks.
- More interactions are involved in the items.
- Available only in the computer-based path.

PIAAC PSTRE Sample Item (1)

OECD PIAAC

Section 1

Unit 22

You want to copy some music files to your portable music player.

The music player has room for 20 MB and you want as many files as possible. You want to include only jazz and rock music.

Select the files to include.

Once you have selected the files, click Next to continue.

Spreadsheet

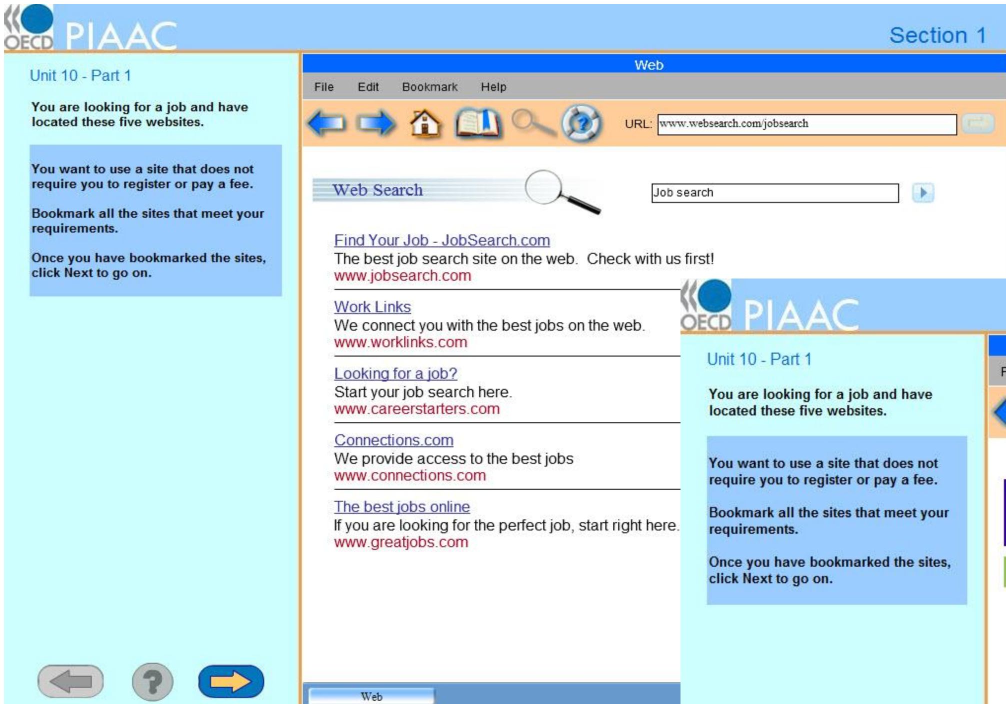
File Edit Data Help

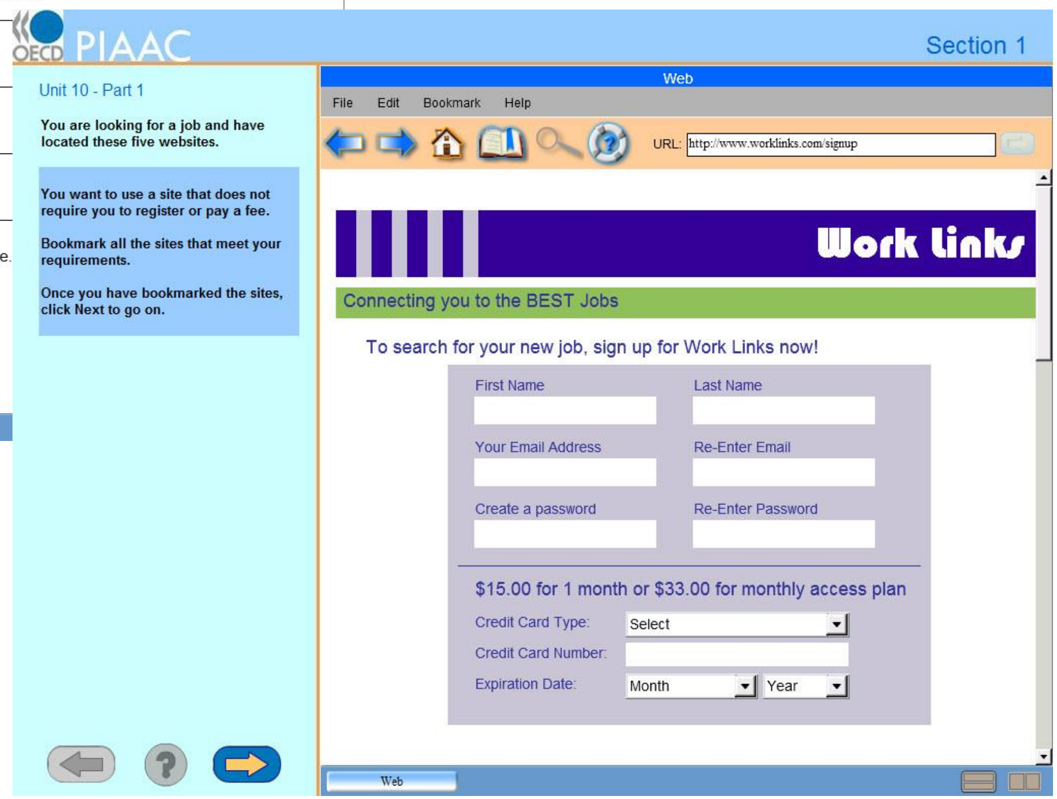
	Title	Size	Time	Artist	Genre
☐	A Foreign Affair	14.8 MB	11:40	Don Rader Quartet	Jazz
☐	About the Blues	4.3 MB	3:08	Julie London	Blues
☐	Another Mind	7.8 MB	8:44	Hiromi Uehara	Jazz
☐	Blue Trane	10 MB	9:03	John Coltrane	Jazz
☐	Don't Give up on Me	3.5 MB	3:45	Solomon Burke	Blues
☐	Far Out	5.3 MB	5:25	Antonio Farao	Jazz
☐	Fire and Water	5.3 MB	4:00	Free	Blues
☐	If	4.9 MB	5:48	Myriam Alter	Jazz
☐	X	2.2 MB	3:04	INXS	Rock
☐	Inclined	7.1 MB	5:59	Carol Welsman	Jazz
☐	On an Island	16 MB	6:47	David Gilmore	Blues
☐	Pass It On	3.1 MB	3:36	Albert Calvo	Jazz
☐	Raindrops, Raindrops	5.2 MB	3:46	Karin Krog	Jazz
☐	Say You Will	8.8 MB	3:47	Fleetwood Mac	Rock
☐	Skin Deep	7.1 MB	4:28	Buddy Guy	Blues
☐	Speak No Evil	6.9 MB	5:13	Flora Purim	Jazz
☐	The Other Side of Blue	6.5 MB	5:08	Jean Shy & Jobo	Jazz
☐	The Rise	7.3 MB	7:28	Julien Lourau	Jazz
☐	The Rising	4.5 MB	4:50	Bruce Springsteen	Rock

Total Size Selected (MB)

Spreadsheet

PIAAC PSTRE Sample Item (2)





Research Objectives

- **Study Purposes:**
  - To extract and detect robust sequential action patterns that are associated with success or failure on one PSTRE item.
  - To compare the extracted sequence patterns among selected countries.

- **Research Questions:**
  - How sequences of actions recorded in problem-solving tasks are related to task performance?
  - Can the key actions / action patterns that lead to success or failure be identified?

Sample

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 club member 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.

Methods

- Start, SS, SS_So, SS_So_1B, SS_So_OK, E, Next, FINALENDING
- Start, SS, E, SS, SS_Se, SS_Type_FN, E, Next, Next_C, Next, FINALENDING
- Start, Next, FINALENDING

- Similar structure between action sequences and languages.
- Motivated by the methodologies of **natural language processing and text mining.**
- Utilized feature selection models in analyzing the process data at a variety of aggregate levels.
- Evaluated the different methodologies in terms of predictive power of the evidence extracted from process data.

N-grams Model

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”
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”

Recode Next_OK, END into “FINALENDING”

Term Weights

- An **inverse sequence frequency** was applied for attenuating the effect of actions that occurred too often in the collection to be meaningful.
- A **dampened term frequency** was also used to adjust the importance of an action with multiple occurrences in a single sequence.

Dampened term frequency

Inverse sequence frequency

$$\text{weight}(i, j) = \begin{cases} [1 + \log(\text{tf}_{i,j})]\log(N / \text{sf}_i) & \\ 0 & \text{if } \text{tf}_{ij} = 0 \end{cases}$$

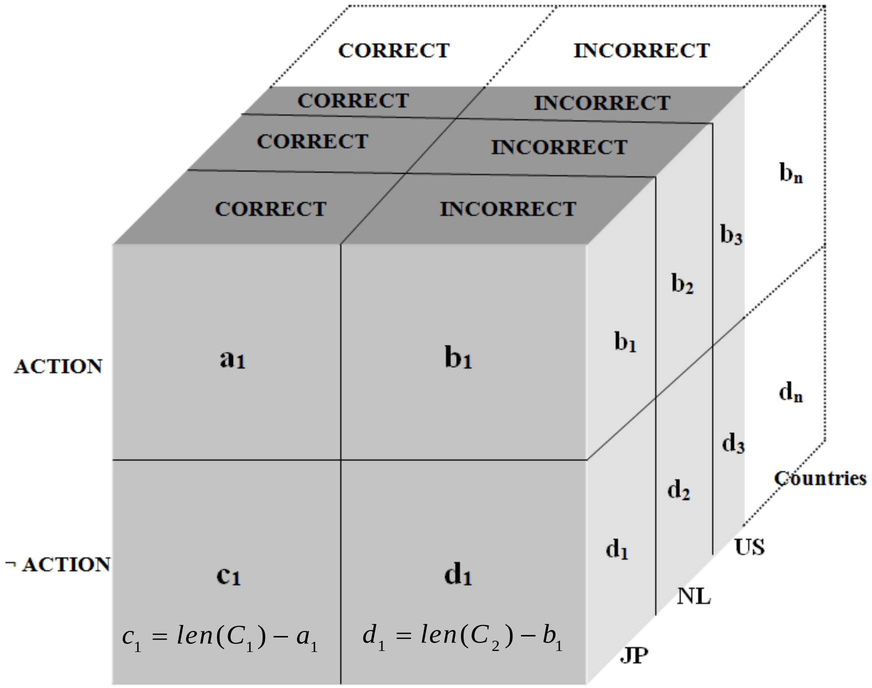
i, j action i in sequence j

$\text{tf}_{i,j}$ frequency of action i in sequence j

sf_i frequency of sequence that contains action i

N number of sequences (test takers)

Feature Selection Models (1)
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	TART, 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
	[illegible]	[illegible]	Type_SN, E	120.56	SS_So_1B, SS_So_OK, E	137.53
	[illegible]	[illegible]	Se, SS_Type_SN	98.21	SS, E, SS	133.85
	[illegible]	[illegible]	So, SS_So_1B	84.43	SS_Se, SS_Type_SN, E	108.55
	SS_So_2A	[illegible]	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 xt	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	[illegible]
	SS_So_3D	14.56	E_S, E	364.66	E_S,	[illegible]
	SS_So_C	[illegible]	S, SS	299.74	E, E	[illegible]
	E_S	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	SS_Type_PS	[illegible]	[illegible]	[illegible]	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			
Mean=0.79	Unigrams	0.88	0.88	0.63
	Bigrams	0.75	0.88	0.75
	Trigrams	0.75	0.88	0.75
	Incorrect			
Mean=0.71	Unigrams	0.63	0.63	0.63
	Bigrams	0.63	0.88	0.88
	Trigrams	0.75	0.63	0.75

Results (3)
Features of Actions by Countries

Robust Features of Actions and Action Sequences Across Countries

	Unigrams		Bigrams			
	Actions	χ^2	Actions	χ^2	[illegible]	[illegible]
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	STA[illegible]	[illegible]
	SS_Se	60.88	SS_Type_FN, E	203.30	SS_Type_FN[illegible]	[illegible]
	SS_So_3B	31.59	SS_Se, SS_Type_GN	202.10	SS_Type_FN[illegible]	[illegible]
	SS_So_2A	16.15	START, SS_Se	117.42	SS_Se, SS_Ty[illegible]	[illegible]
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_S[illegible]S_So	166.12	E_S, Next, Next_C	149.25
	SS_Type_UM	70.75	E[illegible]	137.22	SS_Type_SM, SS_So, SS_So_1B	149.21
	[illegible]	[illegible]	[illegible]	116.73	SS_Type_SM, SS_Type_SM, SS_So	140.96
	[illegible]	[illegible]	[illegible]	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

ETS

Results (4)
Correlation between CHI and WLLR

Correlation between CHI and WLLR in Different Performance Groups by N-grams

	Correct	Incorrect
Unigrams	0.74	0.60
Bigrams	0.87	0.98
Trigrams	0.88	0.94

- The CHI and WLLR scores were **moderately correlated in the unigrams and highly correlated in the bigrams and trigrams** in both the correct and incorrect groups.

- It also proves that the mini-sequences (**bigrams and trigrams**) are more informative in process data analysis compared with single actions (unigrams).

Conclusions and Implications (1)

- With increasing use of computer-based assessments, process data play an increasingly important role in tracking test takers' thinking and action sequences, which is specially helpful in analyzing problem-solving items.

- The pilot study presented what we think is a promising method to analyze process data and extract robust sequence features that are informative for differentiating between performance groups.

Conclusions and Implications (2)

- The studies also demonstrate that process data can be useful in detecting nonresponse due to low engagement of test takers and checking item designs, especially in the field test.

- We explored these topics as part of the ongoing improvement of the software platform used in the PIAAC assessment for test delivery.

Future Studies

- The future study will focus on adapting existing methods for sequence data mining and develop a generalized toolkit for process data analysis.
- We recommend including background characteristics and timing data in the analysis of process data to further explore their interaction effects on performance.
- Explorations of how process data may inform adaptive testing appear to be a potential valuable research direction.

Acknowledgements

- We are thankful for all the supports from the Center for Global Assessment and Data Analysis Group in ETS for the helpful information and data preparations.

Related Publications

- 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. Doi:10.4018/978-1-4666-9441-5.ch029.

- 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. Doi: 10.1007/978-3-319-19977-1_13

Thank you very much!

For further information and suggestions, please contact

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

Dr. Matthias von Davier
mvondavier@ets.org