Exploring Process Data in Computer-Based International Large-Scale Assessments

探索过程数据在全球大规模机器考试中的应用

Qiwei He & Matthias von Davier
Educational Testing Service

Data Exploration

Contents

1. Introduction
   - Computer-based testing in large-scale assessments (LSA)
   - Problem solving items and process data
   - Research questions/hypotheses
2. Feature Extraction from Action Sequences
   - A case study in PIAAC
   - Robust features by performance groups across countries
3. Feature Generation and Selection
   - A case study in PISA
   - What features can we generate from process data?
   - What are “good” features?
4. Conclusions and Future Studies
5. Public (Released) PISA Process Data Online

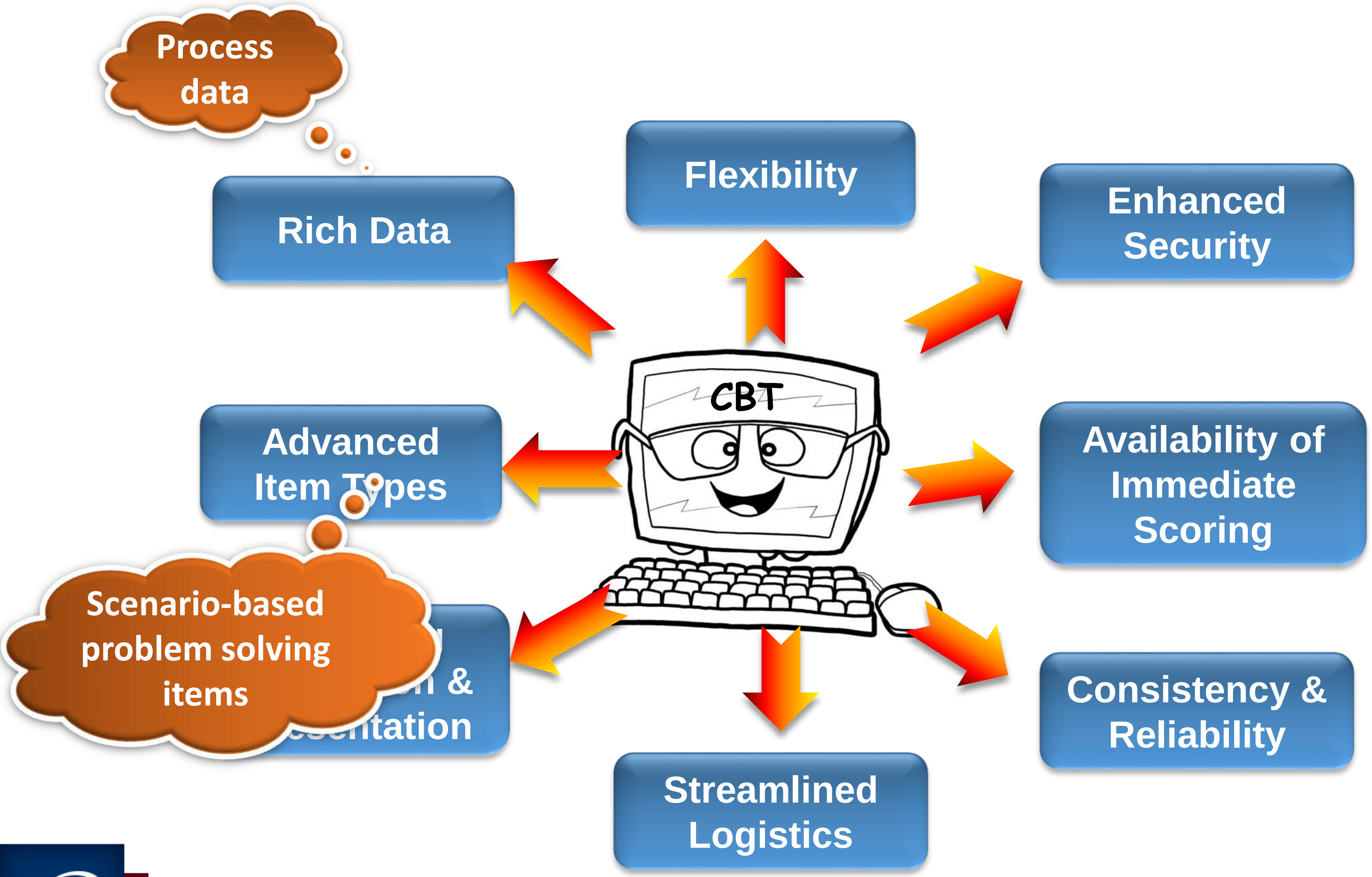
Introduction

- Computer-based testing in LSA
- Problem solving items and process data
- Research questions/hypotheses

Computer-Based Testing in LSA

- The growing interest in assessing technology-related skills and knowledge has been facilitated by the move towards delivering assessments via computer and the web, thus, making it more feasible to also assess information and communication technologies (ICT) skills.
- Several recent large-scale assessments, such as the PIAAC, PISA, and NAEP, have sought to assess computer, digital learning, and problem-solving skills, which are essential in the 21st century.
    - PIAAC adopts computer-based assessment (as one path) from the very beginning when this LSA debuts in 2012.
    - PISA 2015, for the first time, delivers the assessments of all subjects (math, reading, science, FL, CPS) via computer.

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.

Results

Problem Solving Items

- PISA 2012 defines problem-solving competence as:

> ... an individual's capacity to engage in cognitive processing to understand and resolve problem situations where a method of solution is not immediately obvious. It includes the willingness to engage with such situations in order to achieve one's potential as a constructive and reflective citizen.
>
> (OECD, 2013)

OECD (2013), PISA 2012 Assessment and Analytical Framework: Mathematics, Reading, Science, Problem Solving and Financial Literacy, OECD Publishing. http://dx.doi.org/10.1787/9789264190511-en

The Programme for International Student Assessment (PISA)

- The Programme for International Student Assessment (PISA) is a triennial international survey developed and organized by the Organization for Economic Cooperation and Development (OECD) with three core domains: math, reading and science.

- The PISA aims to evaluate education systems worldwide by testing the skills and knowledge of 15-year-old students. A total of 72 economies, over half a million students, participated the assessment in PISA 2015 cycle.

The Programme for International Student Assessment (PISA)

Assessment year	2000	2003	2006	2009	2012	2015
Subjects assessed	**Reading** Mathematics Science	Reading **Mathematics** Science Problem Solving	Reading Mathematics **Science**	**Reading** Mathematics Science	Reading **Mathematics** Science Problem Solving Financial Literacy	Reading Mathematics **Science** CPS Financial Literacy

- 2015 PISA Tests
  - Science: 184 items
  - Reading: 103 items
  - Math: 81 items
  - CPS: 117 items
  - FL: 43 items
- Each student was given a two-hour combination of these tasks.

10 items are scenario-based items with simulation environments

PISA 2015 Scientific Literacy Interactive Sample Item (1)

PISA 2015

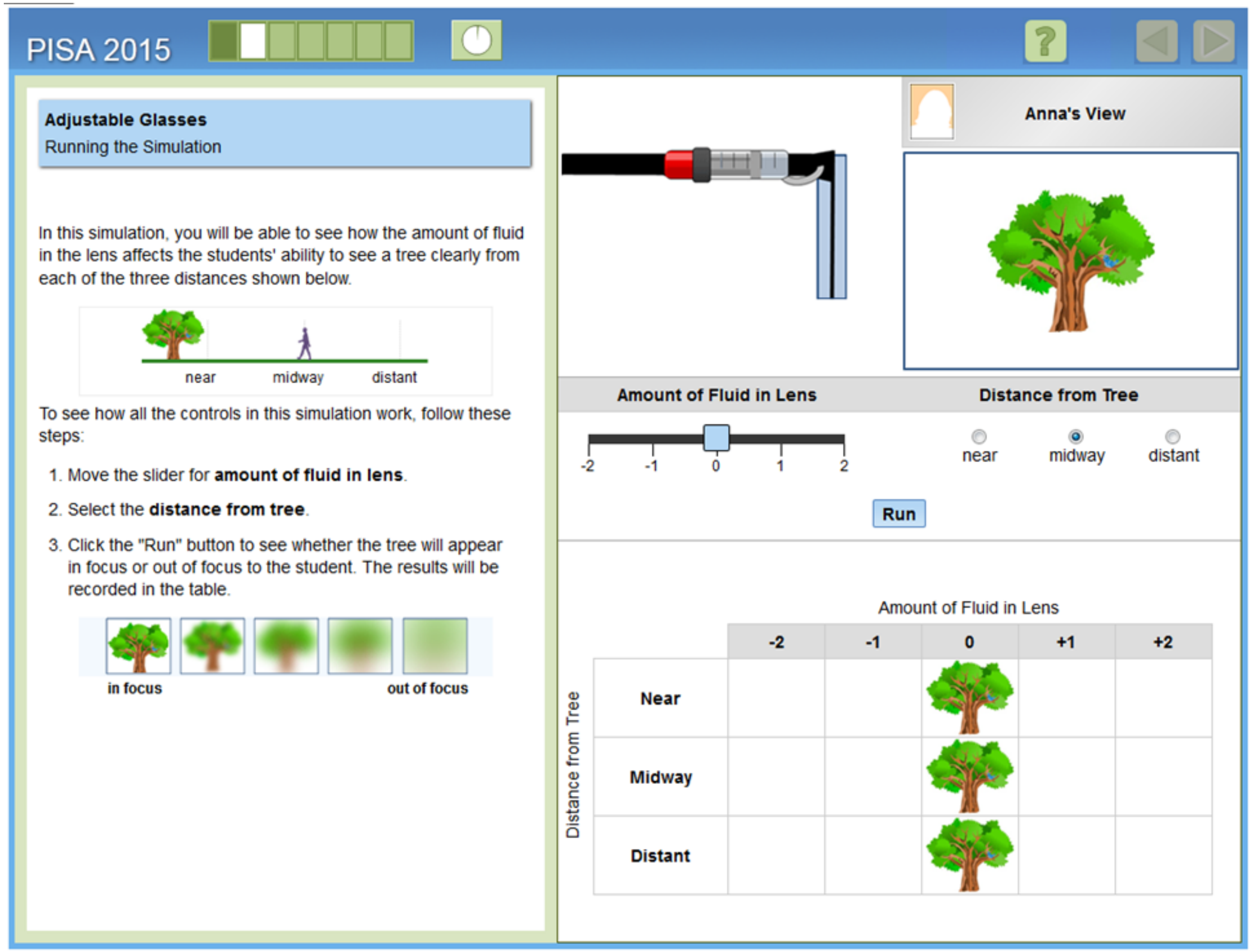
Adjustable Glasses
Introduction

Read the introduction. Then click on the NEXT arrow.

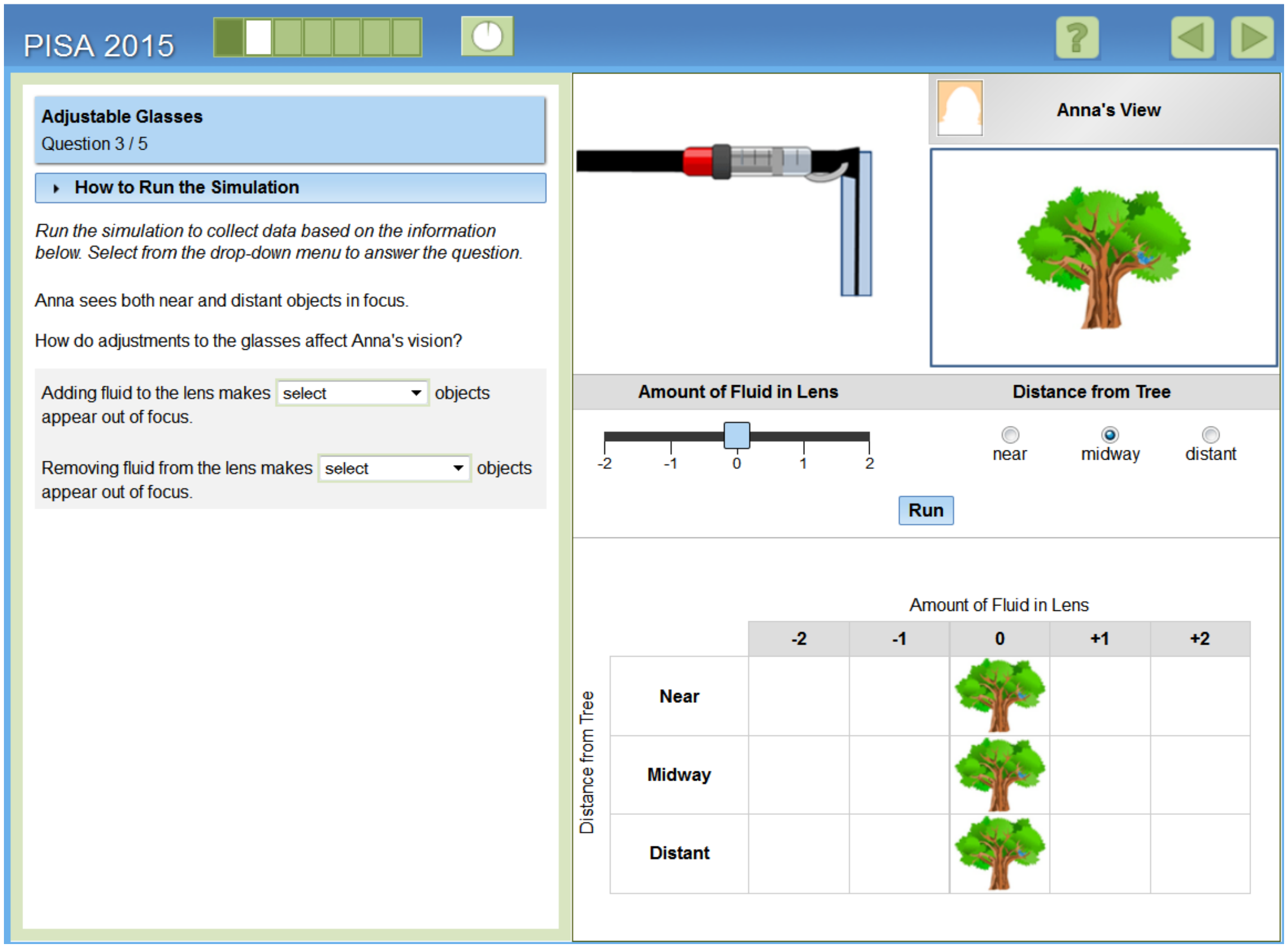
ADJUSTABLE GLASSES

A new technology, called **adjustable glasses**, has been developed to help people without access to eye doctors to correct their vision. The lenses of these glasses contain a fluid. The shape of the lens changes as the amount of fluid in the lens is adjusted.

PISA 2015 Scientific Literacy Interactive Sample Item (2)



PISA 2015 Scientific Literacy Interactive Sample Item (3)



The Programme for International Assessment of Adult Competencies (PIAAC)

- The PIAAC is a large-scale study of adult skills and life experience focusing on education and employment that was developed and organized by the OECD. This survey has been conducted in over 40 countries so far.

- The survey is implemented by
  - interviewing adults aged 16 to 65 in their homes – 5,000 individuals in each participating country.
  - answering questions via computer, although the survey can also be implemented via pencil-and-paper.
  - assessing literacy and numeracy skills and the ability to solve problems in technology-rich environments.
  - collecting a broad range of information, including how skills are used at work and in other contexts, such as the home and the community.

The Main Elements of PIAAC

Direct Assessment

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

Module on Skills Use

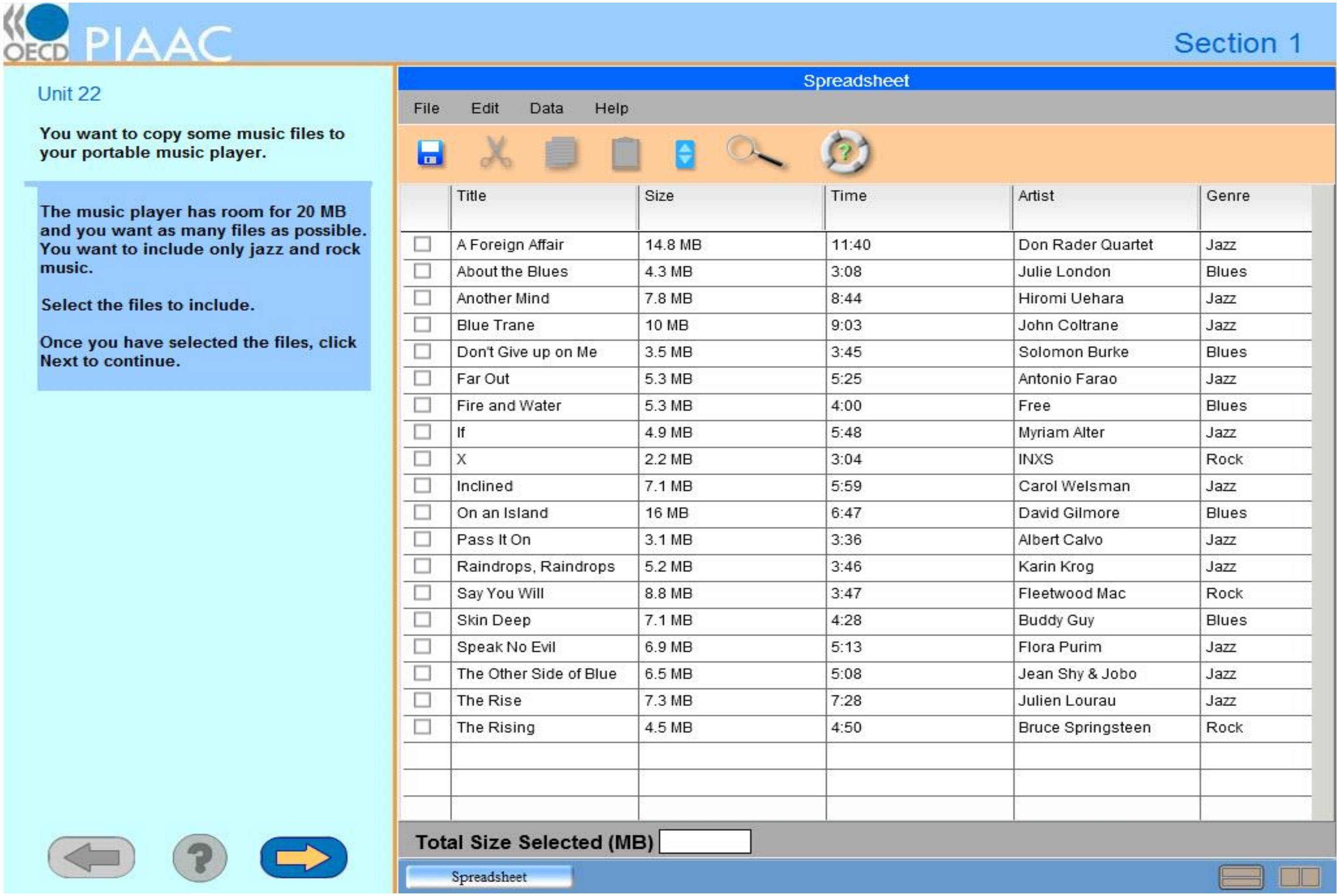
- **Cognitive skills**: re writing, math and use of ICTs
- **Interaction a kills**: collaboration, planning,

Background Questionnaire

- Social and linguistic background

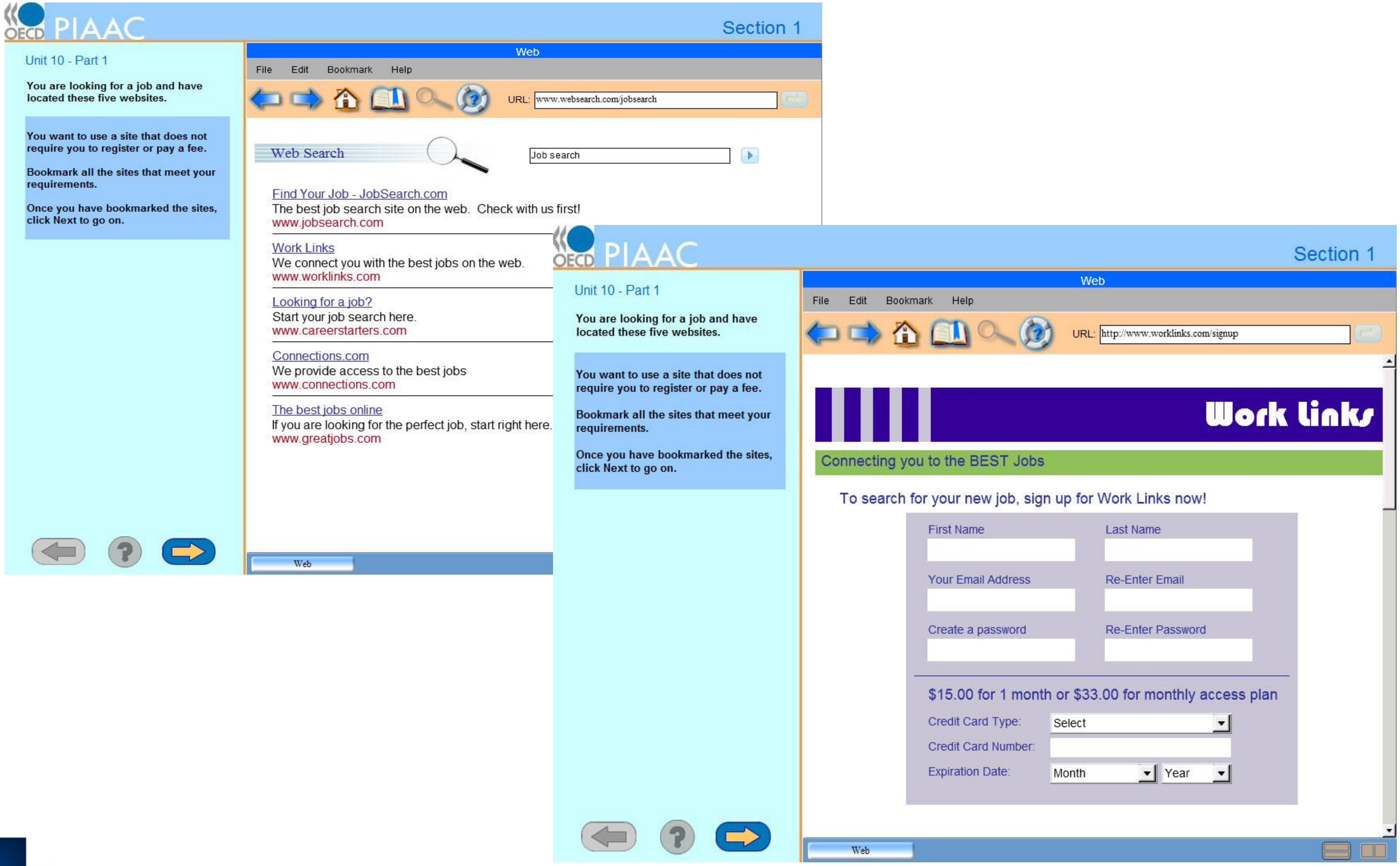
- PSTRE items are used to assess the skills required to solve problems for personal, work, and civic purposes.
- Test takers have to set up appropriate goals and plans, and access and make use of information through computers and networks.
- More interactions are involved in the items.
- Available only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

PIAAC PSTRE Sample Item (1)



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

PIAAC PSTRE Sample Item (2)



Rich Data – Raw Log File


```
      <variable name="Dehnbarkeit" userDefinedId="EndoB" value="50"/>
      <variable name="Transparenz" userDefinedId="EndoC" value="54"/>
      <variable name="Stradium" userDefinedId="ExoB" value="0"/>
      <variable name="Melium" userDefinedId="ExoC" value="0"/>
      <variable name="Ladium" userDefinedId="ExoA" value="0"/>
    </logEntry>
  </logEntry>
  <logEntry xsi:type="cbaloggingmodel:LogEntryTimeStamp" timeStamp="2011-06-23T12:29:09.859+0200">
    <logEntry xsi:type="cbaloggingmodel:SelectedPanelLogEntry" id="Scenario"/>
  </logEntry>
  <logEntry xsi:type="cbaloggingmodel:LogEntryTimeStamp" timeStamp="2011-06-23T12:29:14.843+0200">
    <logEntry xsi:type="cbaloggingmodel:MicroDynButtonPressLogEntry" button="Execute" phase="simple exploration">
      <variable name="HÃ¤rte" userDefinedId="EndoA" value="58"/>
      <variable name="Dehnbarkeit" userDefinedId="EndoB" value="54"/>
      <variable name="Transparenz" userDefinedId="EndoC" value="58"/>
      <variable name="Stradium" userDefinedId="ExoB" value="2"/>
      <variable name="Melium" userDefinedId="ExoC" value="0"/>
      <variable name="Ladium" userDefinedId="ExoA" value="0"/>
    </logEntry>
  </logEntry>
  <logEntry xsi:type="cbaloggingmodel:LogEntryTimeStamp" timeStamp="2011-06-23T12:29:19.140+0200">
    <logEntry xsi:type="cbaloggingmodel:MicroDynButtonPressLogEntry" button="Execute" phase="simple exploration">
      <variable name="HÃ¤rte" userDefinedId="EndoA" value="58"/>
      <variable name="Dehnbarkeit" userDefinedId="EndoB" value="54"/>
      <variable name="Transparenz" userDefinedId="EndoC" value="58"/>
      <variable name="Stradium" userDefinedId="ExoB" value="0"/>
      <variable name="Melium" userDefinedId="ExoC" value="0"/>
      <variable name="Ladium" userDefinedId="ExoA" value="0"/>
    </logEntry>
  </logEntry>
  <logEntry xsi:type="cbaloggingmodel:LogEntryTimeStamp" timeStamp="2011-06-23T12:29:20.796+0200">
    <logEntry xsi:type="cbaloggingmodel:MicroDynButtonPressLogEntry" button="Execute" phase="simple exploration">
      <variable name="HÃ¤rte" userDefinedId="EndoA" value="58"/>
      <variable name="Dehnbarkeit" userDefinedId="EndoB" value="58"/>
      <variable name="Transparenz" userDefinedId="EndoC" value="62"/>
      <variable name="Stradium" userDefinedId="ExoB" value="0"/>
      <variable name="Melium" userDefinedId="ExoC" value="2"/>
```

Log-Data Restructuring

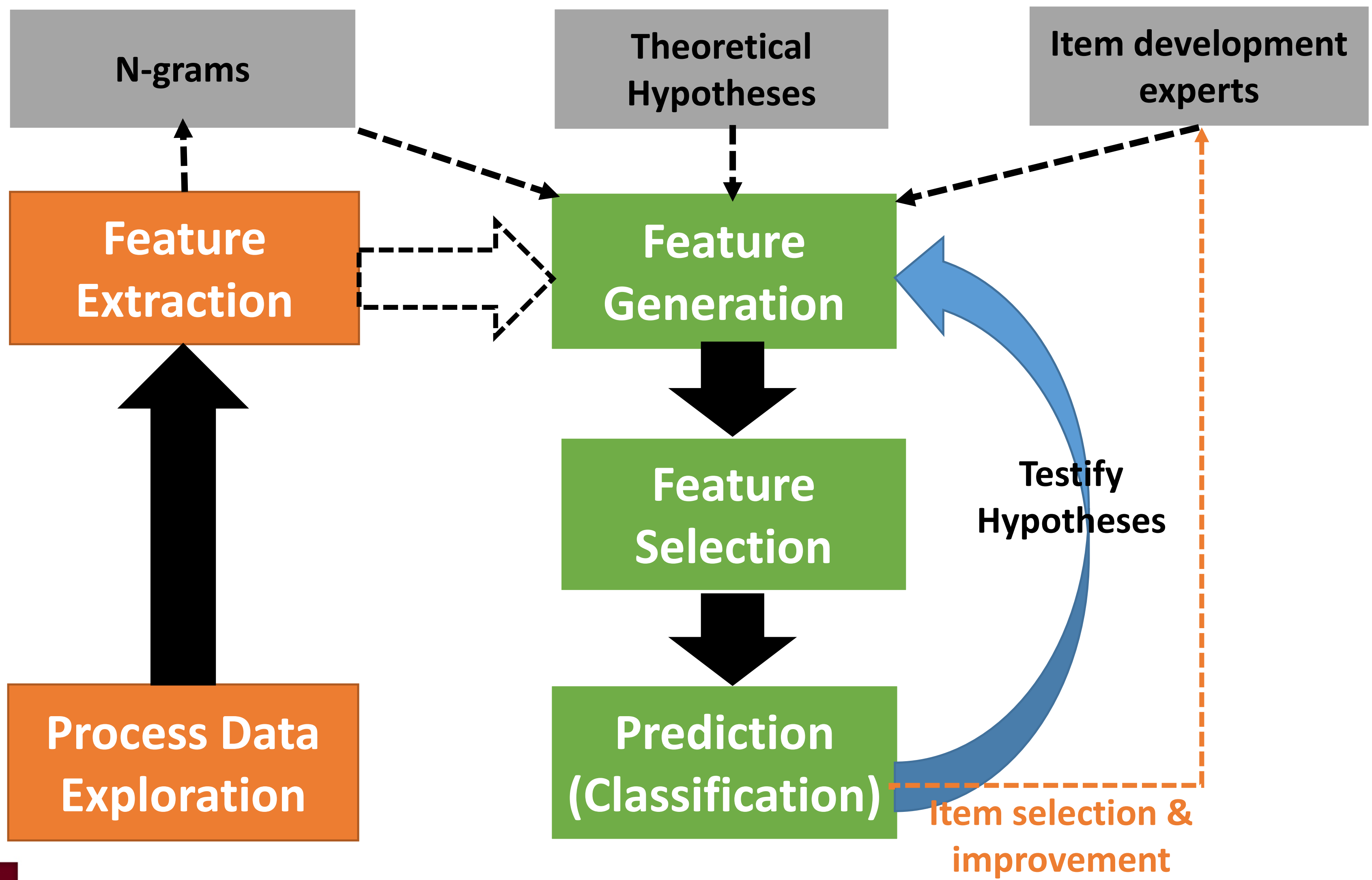
studentId	eventCou	itemId	event_name	target	id	lang	time
840-51-01-003-00025	0	CS633Q00	onItemBegin	MODULE		eng-USA	1397193846084
840-51-01-003-00025	1	CS633Q00	stimulusLoaded	MODULE		eng-USA	1397193846545
840-51-01-003-00025	2	CS633Q00	QuestionLoaded	MODULE		eng-USA	1397193846570
840-51-01-003-00025	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	CS633Q000	onItemBegin	MODULE		eng-USA	1397193851104
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

Research Questions

- **WHAT CAN WE LEARN FROM PROCESS DATA?**
  - What features can we generate from the process data?
  - What are the good features?
  - How can we select/group them?

- **WHAT CAN PROCESS DATA DO FOR US?**
  - Can process data make effect in assessment?
  - How can process data testify hypotheses from theory (e.g., theory in cognitive process)?
  - Can process data improve the measurement?

Structure of a Process Data Study



Finding the Treasure in the Data Lake

Feature Extraction from Action Sequences

- A case study in PIAAC
- Robust features by performance groups across countries

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 are sequences of actions recorded in problem-solving tasks 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 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.

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.

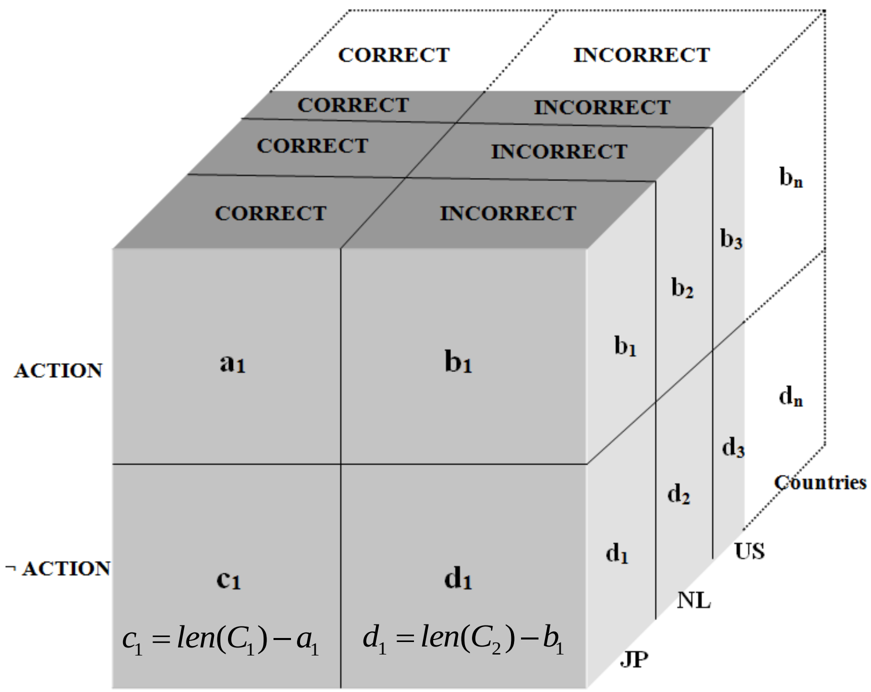
Dampened term frequency

Inverse sequence frequency

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

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.

$$
\begin{aligned}
WLLR(t, C_i) &= P(t \mid C_i) \log \frac{P(t \mid C_i)}{P(t \mid \neg C_i)} \\
&= P(t \mid C_i) \log \frac{P(t \mid C_i)}{Q(t \mid C_i)}
\end{aligned}
$$

$P(t \mid C_i)$ 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 is typical of class C_i

Conversely, the lower the WLLR, the more likely the action belongs to class Ø C_i

Measuring the Power of Learning.™

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
			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	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		[illegible]	
	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	[illegible]	[illegible]
	SS_Se	60.88	SS_Type_FN, E	203.30	[illegible]	[illegible]
	SS_So_3B	31.59	SS_Se, SS_Type_GN	202.10	[illegible]	[illegible]
	SS_So_2A	16.15	START, SS_Se	117.42	[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	[illegible]	166.12	E_S, Next, Next_C	149.25
	SS_Type_UM	70.75	[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

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

Feature Generation and Selection

- A case study in PISA
- What features can we generate from process data?
- What are “good” features?

Sample and Instrument

- This study focused on one representative item, **Climate Control**, which was released in PISA 2012 and intended to assess problem-solving skills.
- Process data (n = 30,224) is extracted from the item's log-file recorded in tests. Information such as students' strategies and behaviors needs to be mined out from the raw data.

1
2
3

CLIMATE CONTROL

You have no instructions for your new air conditioner. You need to work out how to use it.

You can change the top, central and bottom controls on the left by using the sliders (-▫-). The initial setting for each control is indicated by ▲.

By clicking APPLY, you will see any changes in the temperature and humidity of the room in the temperature and humidity graphs. The box to the left of each graph shows the current level of temperature or humidit

- This is a harder item – **Level 4** on the problem-solving scale.

- Students must engage with the machine, and use the feedback and information uncovered to reach a solution: it is an **interactive** problem.

APPLICATION

Top Control

-- - ▲ + ++

Central Control

-- - ▲ + ++

Bottom Control

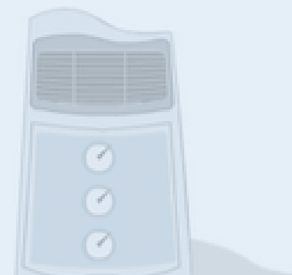
- ▲ + ++

Temperature

25

Humidity

25



APPLY

RESET

DIAGRAM

Question 1: CLIMATE CONTROL *CP025Q01*

Find whether each control influences temperature and humidity by changing the sliders. You can start again by clicking RESET.

Draw lines in the diagram on the right to show what each control influences.

To draw a line, click on a control and then click on either Temperature or Humidity. You can remove any line by clicking on it.

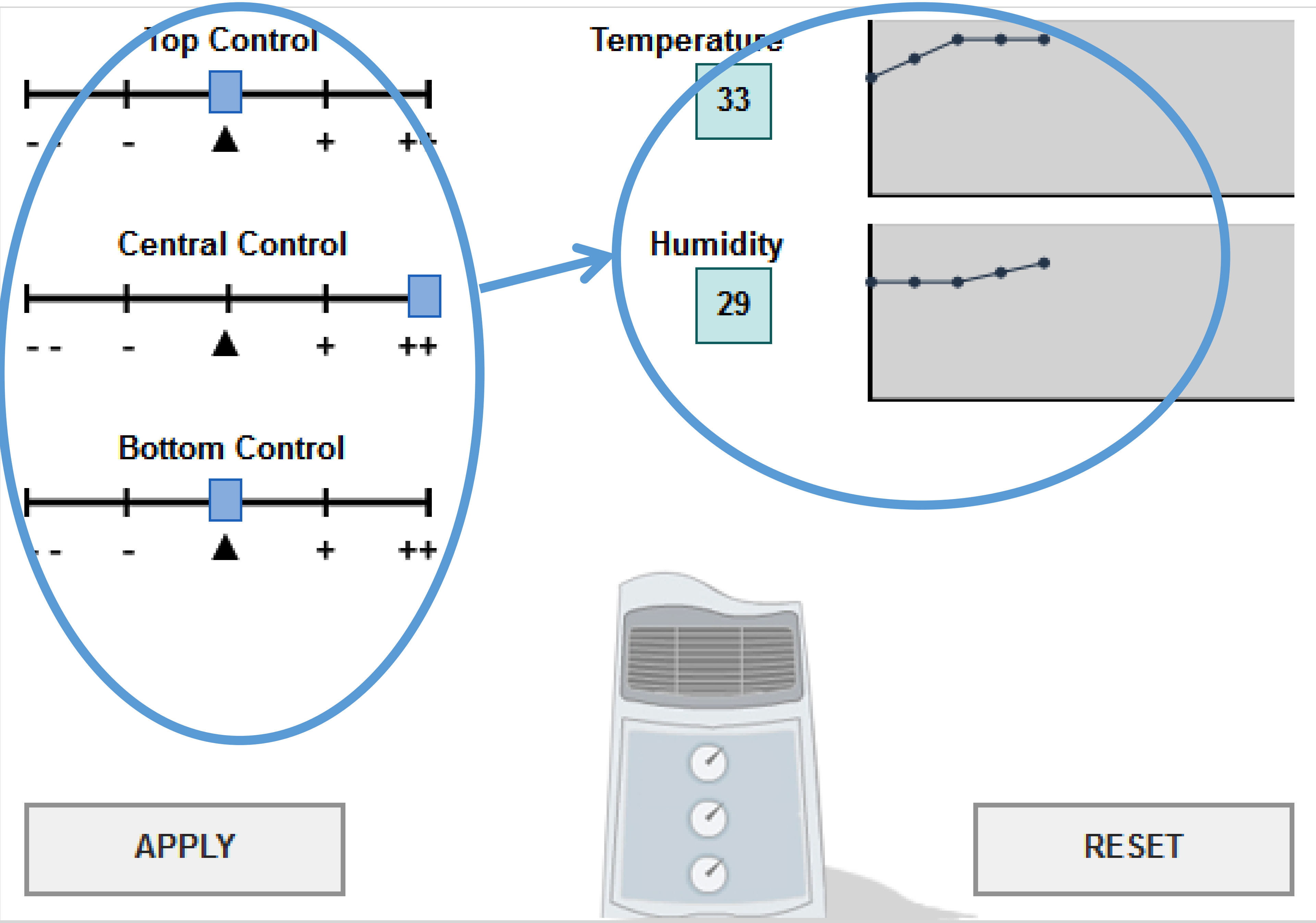
Top Control

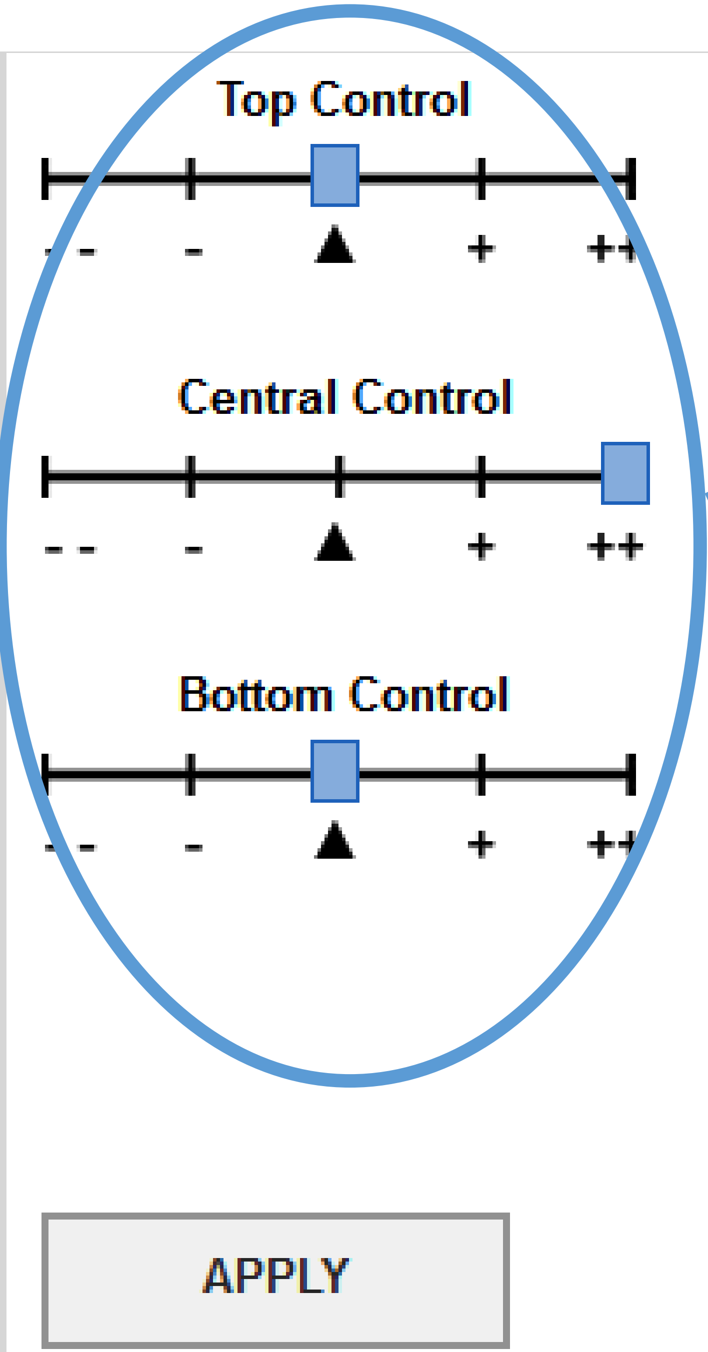
Central Control

Bottom Control

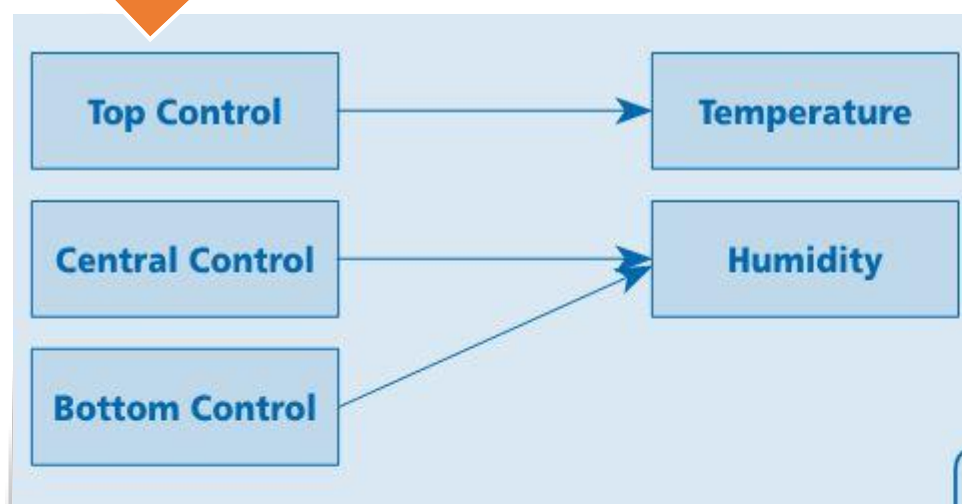
Temperature

Humidity

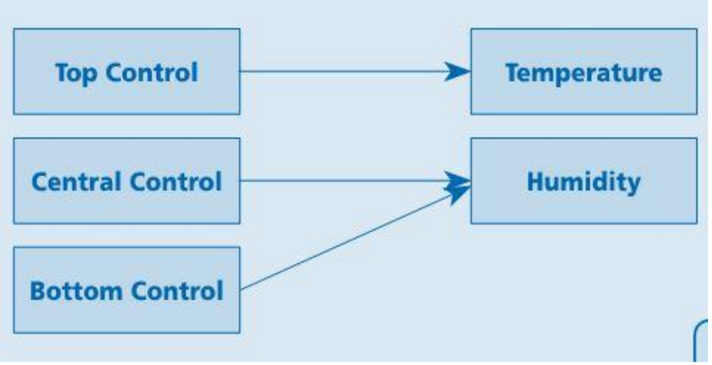

Top Control
- -
-
+
++
Central Control
- -
-
+
++
Bottom Control
- -
-
+
++
Temperature
33
Humidity
29
APPLY
RESET



- Full credit for this question requires that the causal diagram is correctly completed.
- Partial credit for this question is given if the student explores the relationships among variables efficiently, by varying only one input at a time, but fails to correctly represent them in a diagram.
- We define binary responses in analysis by setting partial credit as wrong.



Process Data



time	event_type	top_setting	central_setting	bottom_setting	temp_value	humid_value	diag_state
1288.1	start	NULL	NULL	NULL	NULL	NULL	NULL
1291.9	reset					25	NULL
1338.4	apply					28	NULL
1346.8	apply					33	NULL
1350.1	apply					36	NULL
1354.5	apply					36	NULL
1361.1	apply					36	NULL
1361.1	reset					25	NULL
1375.3	Diagra					NULL	'000000
1376.2	Diagra					NULL	'000000
1400.1	Diagra						'000000
1402.1	Diagra						'000001
1406.8	Diagra						'000001
1408.4	Diagra					NULL	'000101
1410.2	Diagram					NULL	'000101
1410.6	Diagram					NULL	'100101
1416.1	end	NULL	NULL	NULL	NULL	NULL	NULL

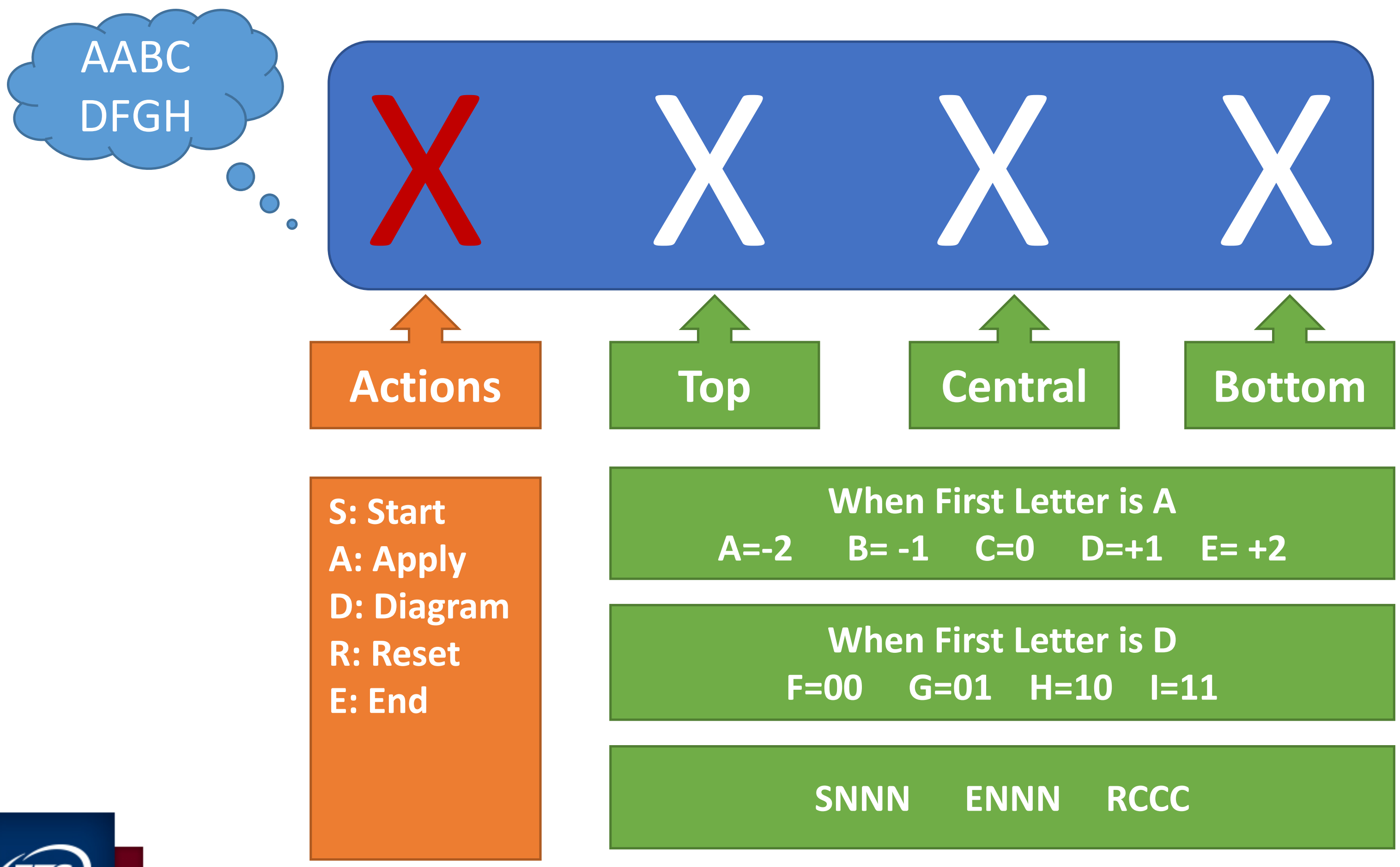
- Binary indicators at corresponding location for temperature / humidity for each control bar.

- Correct response: 10 01 01

meaning:
top-> temperature (1), humidity (0)
central-> temperature (0), humidity (1)
low-> temperature (0), humidity (1)

Recoded Sequences

- Each action is recoded into a 4-letter sequence, to show the action and current status.



Feature Generation

- We generated predictive features from three categories:
  - N-gram action sequences.
  - Solving strategies and behaviors buried under process data.
  - Time information.

N-grams

Recoded Sequences

SNNN, RCCC, ADDD, ADDE, ADEE, AEEE, AEDD, RCCC, DFFF, DFFG, DFGG, DHGG, ENNN

Action Sequences

S, R, A, A, A, A, A, R, D, D, D, D, E

Uni - Gram

S(1), R(2), A(5), D(4), E(1)

Tri – Gram

SRA(1), RAA(1), AAA(3), ARD(1), DDD(2), DDE(1)

Bi - Gram

SR(1), RA(1), AA(4), AR(1), RD(1), DD(3), DE(1)

Strategy Indicators

Recoded Sequences
SNNN, RCCC, ADDD, ADDE, ADEE, AEEE, AEDD, RCCC, DFFF, DFFG, DFGG, DHGG, ENNN

Action Sequences
S, R, A, A, A, A, A, R, D, D, D, D, E

AD Predictor

AD Predictor indicates the behavior of applying the simulation and plotting diagrams on an aggregate level.

Strategy Indicators
RCCC, ADDD, ADDE, ADEE, AEEE, AEDD

VOTAT

Vary-one-thing-at-a-time

Timing Features

- **A time:** Total time spent on applying simulation.
- **D time:** Total time spent on linking diagrams.
- **R time:** The time between applying RESET and the last action before RESET.
- **E time:** The time between END and the last action before END.
- **Total time:** Total time spent on the whole item.
- **time_bf_action:** The time between Start and the first action, suggesting the "reading time" on the item.

Generated Features

Table 1 *A Total of 77 Features Generated from Process Data to Predict Student's Performance*

Uni-gram (3)	D, R, A
Bi-gram (16)	DD, AA, RA, AR, AD, DA, AE, SD, SA, DR, DE, RD, RE, RR, SR, SE
Tri-gram (48)	ADD, AAR, SRD, DDR, AAE, DRE, AAA, ARD, SDR, ADE, RAA, RRE, DDD, DAR, ARR, DAA, RDA, RRA, DAD, SDA, RRR, AAD, RAD, RRD, ADR, ARE, DRR, RDE, DRR, SRA, ADA, SAR, SRE, ARA, RAR, SDE, DRA, RDD, RDR, SDD, DAE, SAR, DDA, DRD, SRR, SAA, SAD, RAE
Timing Features (6)	D time, A time, R time, E time, total time, time_bf_action.
Strategy Indicators (4)	AD predictor, VOTAT group, VOTAT num, n_actions.

Feature Selection (1)

- Some features are highly correlated with or even ”nested” into others. Interactions among features are complicated.
- Many features are categorical variables.
- Random Forest (Breiman, 2001) can tackle highly correlated features and complex interaction at the same time.
- RF also provides a way (permutation variable importance) to select “important” variables in terms of prediction performance.
- RF is possible and efficient for us to do predictive modeling based on a large number of categorical variables.
- RF has a more efficient way to do cross-validation and tune the parameters.

Feature Selection (2)

Table 1 *"Backward Elimination" Algorithm for Feature Selection*

For each **run**, (the total runs are ***m***)

Randomly choose ***p percent*** of sample data.

Run the RF based on all variables and choose tuning parameters by out-of-bag (OOB) error.

For each **round** in **this run**,

- Leave the least ***k*** important features in the rank of VI obtained from previous fit out. (If it is the first run, leave features out from the model with all features.)
- Fit the remaining features, choose new tuning parameters and then measure the permutation VI.
- Record the OOB error rate of this fit.
- Go to the next round if there are more than *k* features in this model. Otherwise, get out of this run.

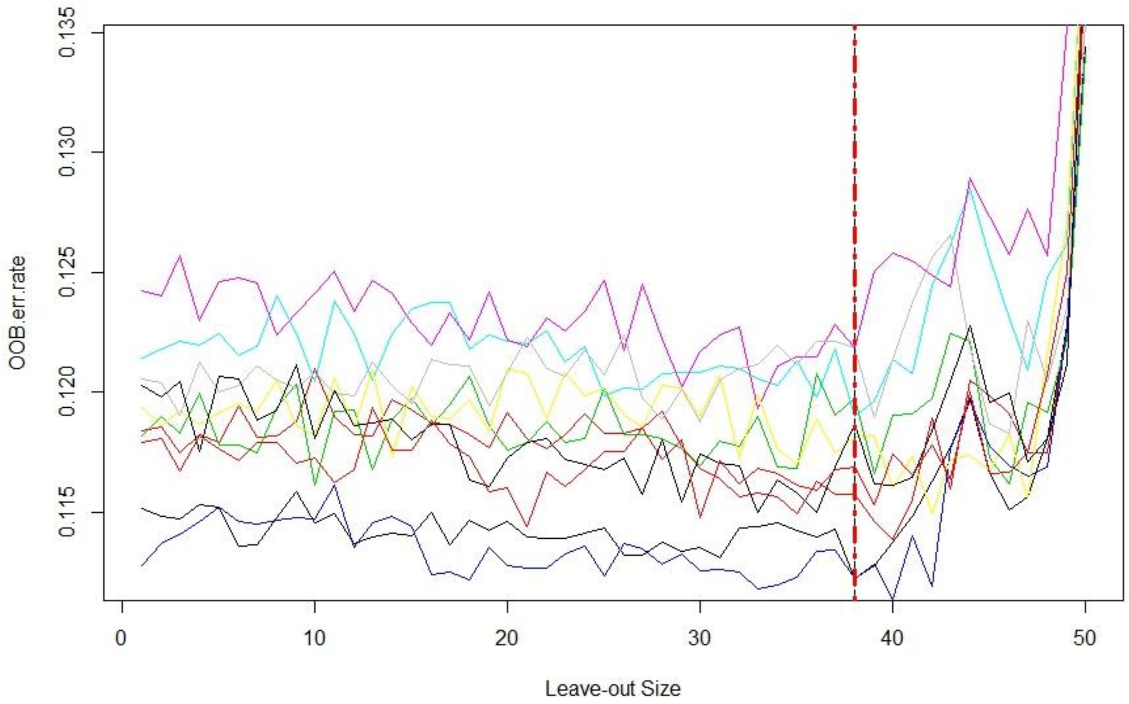
Plot the relation b/w leave-out size and corresponding OOB error rate in this run.

Go to the next run until m runs are finished.

END

Feature Selection (3)

Figure 1. *10-Out-of-30 Runs of "Backward Elimination"*

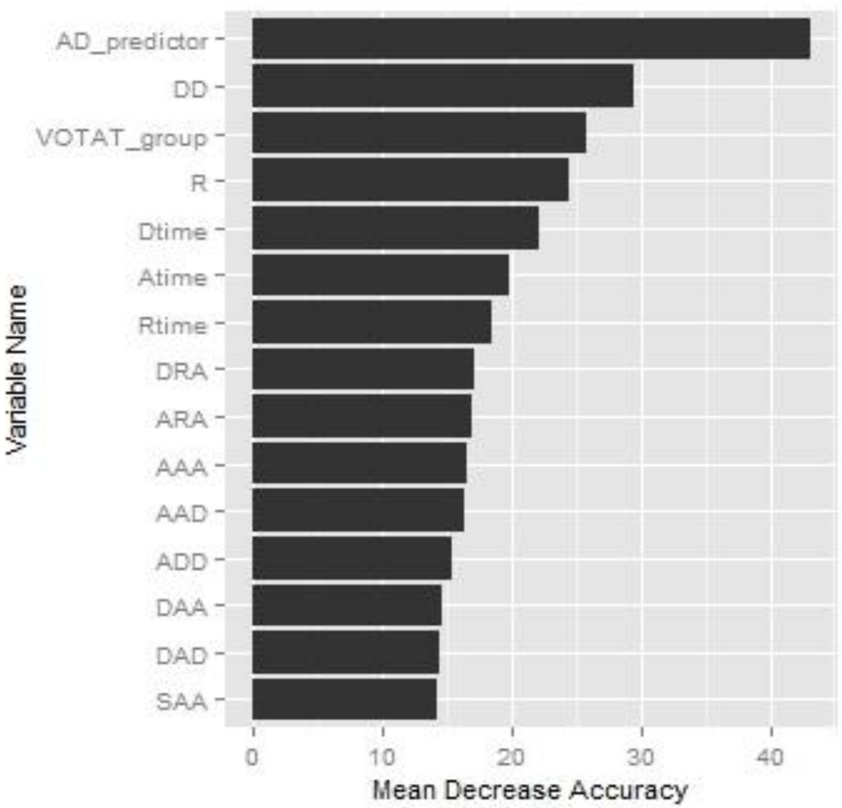


Note. The figure shows 10 out of 30 runs of Backward Elimination. The red threshold indicates how many variables should be eliminated for each run. It was obtained by "elbow" point where the OOB error rate starts increase after leaving out certain number of variables. For caution's sake, the red threshold is the "elbow" point with the smallest number of eliminated variables among 30 runs.

Feature Selection (4)

- Taking the intersection of selected variables across 30 runs, we obtained 22 variables. Pair-wised correlations were examined and 15 variables were finally kept.

Figure 2. *Permutation Variable Importance for Model with the 15 Variables*



Measuring the Power of Learning.™

Feature Selection (5)

Table 2. *Averaged Predictive Performance of Full Model v. Simple Model by 10-fold Cross-validation*

	77-feature Model	15-feature Model
Accuracy	.892 (.006)	.857 (.006)
Cohen's Unweighted Kappa	.784 (.012)	.714 (.012)

15 features were selected from 77 with 0.04 loss in prediction accuracy.

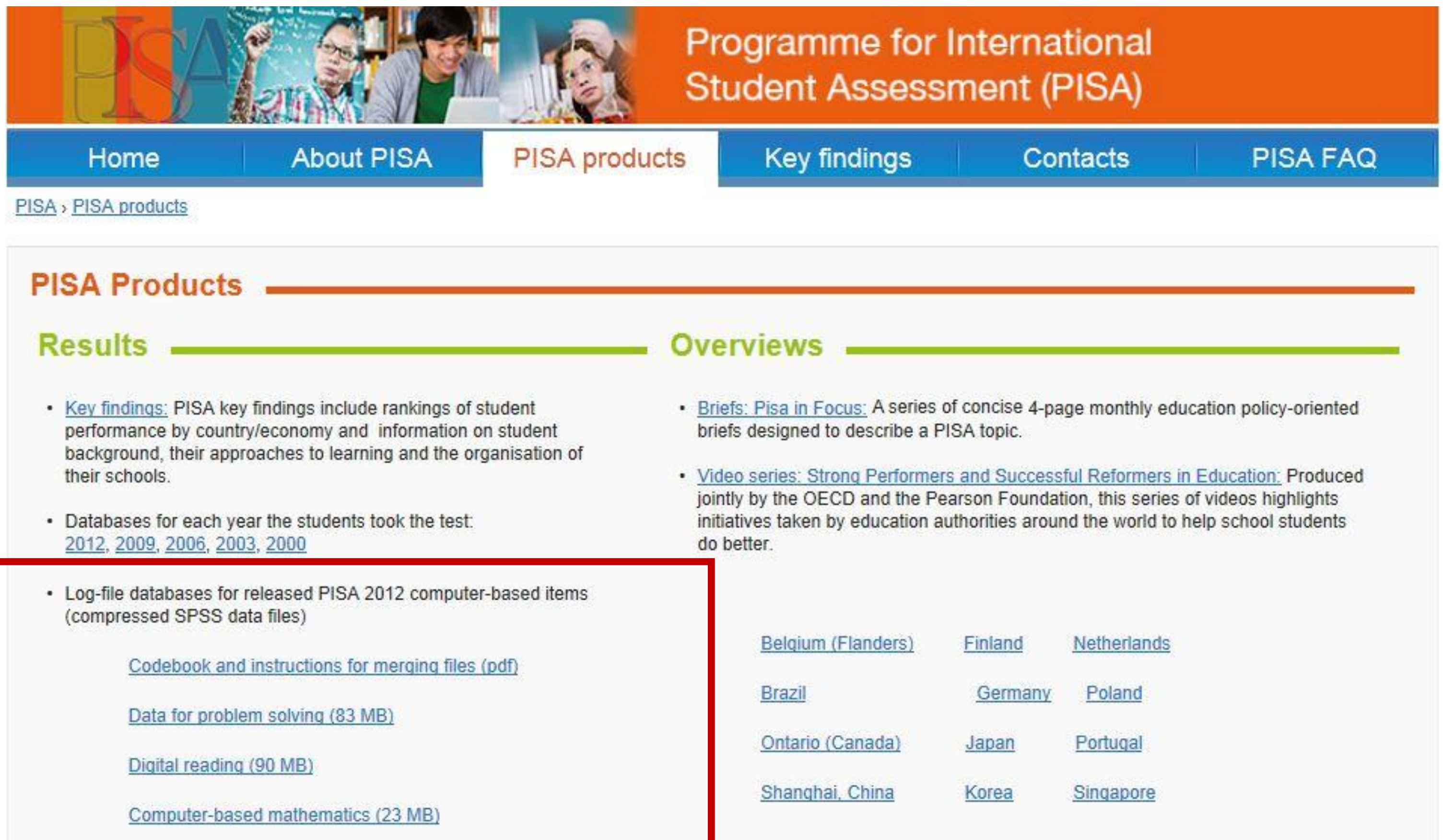
Conclusions and Future Studies

Conclusions and Future Studies

- 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 studies 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.
- Future studies may focus on adapting existing methods for sequence data mining and develop a generalized toolkit for process data analysis.
- Background variables can be taken into consideration in the future studies, to compare the process patterns between different countries and subgroups.

Public PISA Process Data (1)

http://www.oecd.org/pisa/pisaproducts



Programme for International Student Assessment (PISA)

Home | About PISA | PISA products | Key findings | Contacts | PISA FAQ

PISA › PISA products

PISA Products

Results

- Key findings: PISA key findings include rankings of student performance by country/economy and information on student background, their approaches to learning and the organisation of their schools.
- Databases for each year the students took the test: 2012, 2009, 2006, 2003, 2000
- Log-file databases for released PISA 2012 computer-based items (compressed SPSS data files)

  Codebook and instructions for merging files (pdf)

  Data for problem solving (83 MB)

  Digital reading (90 MB)

  Computer-based mathematics (23 MB)

Overviews

- Briefs: Pisa in Focus: A series of concise 4-page monthly education policy-oriented briefs designed to describe a PISA topic.
- Video series: Strong Performers and Successful Reformers in Education: Produced jointly by the OECD and the Pearson Foundation, this series of videos highlights initiatives taken by education authorities around the world to help school students do better.

Belgium (Flanders)	Finland	Netherlands
Brazil	Germany	Poland
Ontario (Canada)	Japan	Portugal
Shanghai, China	Korea	Singapore

Public PISA Process Data (2)

cnt	schoolid	StIDStd	event	time	event_num...	event_value
ARE	0000048	01205	START_ITEM	876.10	1	NULL
ARE	0000048	01205	END_ITEM	886.90	2	NULL
ARE	0000048	01217	START_ITEM	96.00	1	NULL
ARE	0000048	01217	click	128.10	2	robq3text
ARE	0000048	01217	click	131.80	3	robq3text
ARE	0000048	01217	click	142.30	4	Start-Reset
ARE	0000048	01217	click	153.70	5	robq3text
ARE	0000048	01217	click	156.50	6	robq3text
ARE	0000048	01217	END_ITEM	160.40	7	NULL
ARE	0000261	06649	START_ITEM	546.80	1	NULL
ARE	0000261	06649	click	557.30	2	robq3text
ARE	0000261	06649	END_ITEM	574.60	3	NULL
ARE	0000261	06644	START_ITEM	293.10	1	NULL
ARE	0000261	06644	click	350.00	2	Start-Reset
ARE	0000261	06644	click	352.70	3	robq3text
ARE	0000261	06644	END_ITEM	409.80	4	NULL
ARE	0000313	07960	START_ITEM	99.60	1	NULL
ARE	0000313	07960	END_ITEM	101.30	2	NULL
ARE	0000313	07943	START_ITEM	206.60	1	NULL
ARE	0000313	07943	click	221.90	2	Start-Reset
ARE	0000313	07943	click	279.70	3	robq3text
ARE	0000313	07943	END_ITEM	314.10	4	NULL
ARE	0000313	07955	START_ITEM	49.10	1	NULL
ARE	0000313	07955	click	54.30	2	Start-Reset
ARE	0000313	07955	END_ITEM	82.40	3	NULL

References

- Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009) Random forests. *The elements of statistical learning: Data mining, inference, and prediction* (pp. 587-604). New York, NY: Springer.
- He, Q., & von Davier, M. (2016). Analyzing Process Data from Problem-Solving Items with N-Grams: Insights from a Computer-Based Large-Scale Assessment. In Y. Rosen, S. Ferrara, & M. Mosharraf (Eds.) *Handbook of Research on Technology Tools for Real-World Skill Development* (pp. 750-777). 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
- Strobl, C., Boulesteix, A., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance for random forests. *BMC Bioinformatics*, 9-307.

Harvest from Data Exploration

Thank you very much!

For further information and suggestions, please contact

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