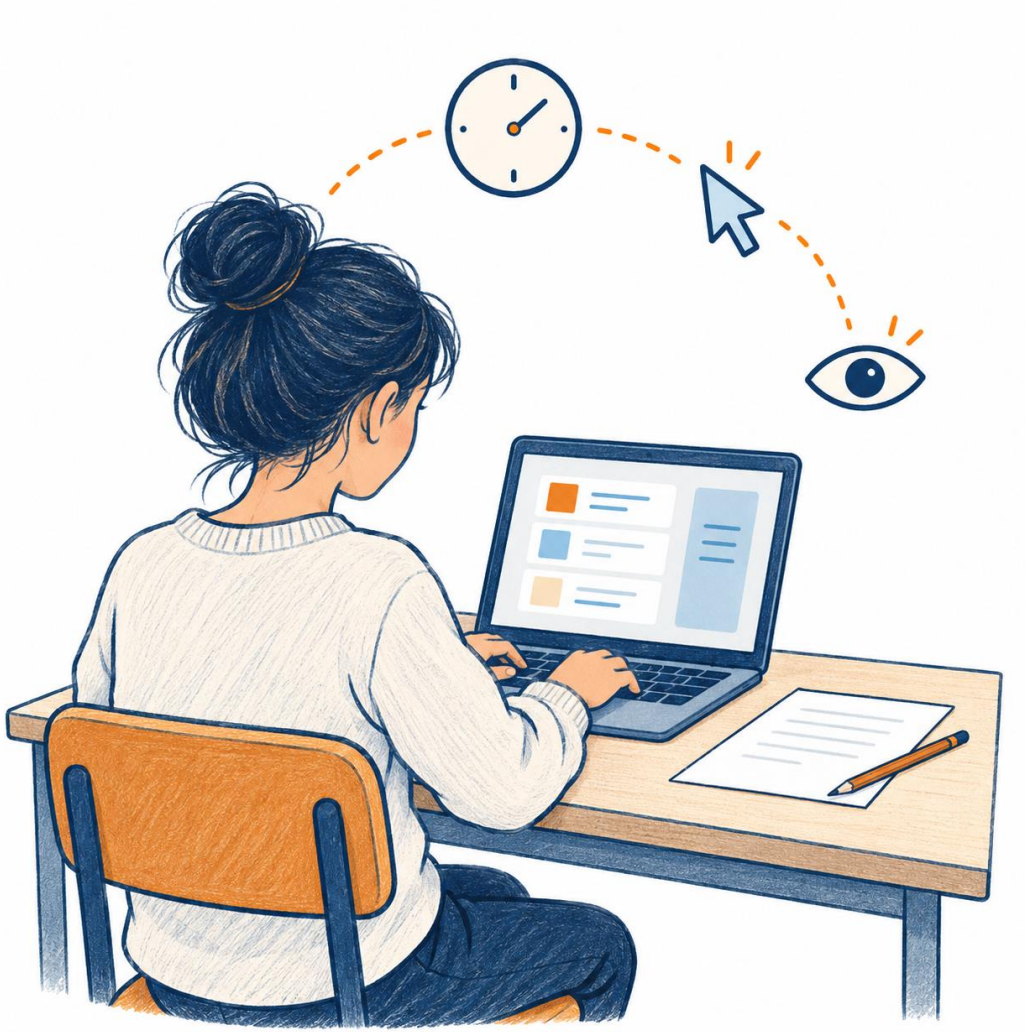
Beyond the Answer

Process Data in Educational Assessment

From response times to webcam-based attention tracking

Artur Pokropek

IFiS PAN and IBE PIB



Why do we use process data?

INSTYTUT BADAŃ EDUKACYJNYCH
Państwowy Instytut Badawczy

IFiS PAN

1 **Understand the response process**

How did the student arrive at an answer?

2 **Improve the assessment**

Which tasks produce unexpected or unhelpful behaviour?

3 **Predict information beyond the score**

Disengagement, response styles, cheating, item leakage, etc.

4 **Because the data and models are interesting**

A good reason to explore and have fun.

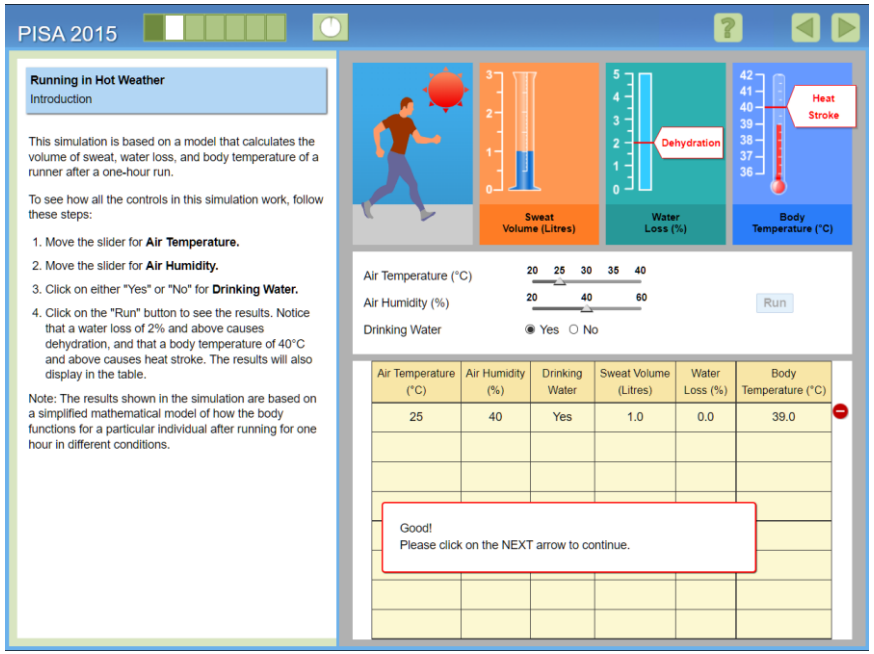
How to do it?

The interface records events. The analyst has to establish what they mean.

A raw event log

	CNTRYID	SEQID	booklet_id	item_id	event_name	event_type	timestamp	event_description
1	ZA6712_AT.data	10	L13	1	taoPIAAC	START	0	TEST_TIME=272
2	ZA6712_AT.data	10	L13	1	stimulus	initialized	317	C315B512
3	ZA6712_AT.data	10	L13	1	stimulus	cba_event	1007	<cbaloggingmodel:CBAItemLog
4	ZA6712_AT.data	10	L13	1	stimulus	cba_event	26222	<cbaloggingmodel:ImageMapLo
5	ZA6712_AT.data	10	L13	1	taoPIAAC	NEXT_INQUIRY	35797	REQUEST
6	ZA6712_AT.data	10	L13	1	taoPIAAC	NEXT_BUTTON	35797	id=nextInquiry_button
7	ZA6712_AT.data	10	L13	1	taoPIAAC	BUTTON	35801	id=nextInquiry_button
8	ZA6712_AT.data	10	L13	1	taoPIAAC	DOACTION	35802	action=tao_item.nextInquiry
9	ZA6712_AT.data	10	L13	1	stimulus	itemScoreResult	39209	<cbascoringresultmm:ItemScore
10	ZA6712_AT.data	10	L13	1	stimulus	ItemScore	39209	<cbascoringresultmm:ItemScore
11	ZA6712_AT.data	10	L13	1	stimulus	SnapshotRef	39220	http://localhost:8080/cba-runti
12	ZA6712_AT.data	10	L13	1	stimulus	END	39220	END
13	ZA6712_AT.data	10	L13	1	taoPIAAC	NEXT_ITEM	39222	REQUEST
14	ZA6712_AT.data	10	L13	1	taoPIAAC	END	39222	END
15	ZA6712_AT.data	10	L13	2	taoPIAAC	START	1	TEST_TIME=320
16	ZA6712_AT.data	10	L13	2	stimulus	initialized	212	C304B710
17	ZA6712_AT.data	10	L13	2	stimulus	cba_event	736	<cbaloggingmodel:CBAItemLog
18	ZA6712_AT.data	10	L13	2	stimulus	cba_event	133572	<cbaloggingmodel:TextBlocksLo
19	ZA6712_AT.data	10	L13	2	stimulus	cba_event	133572	<cbaloggingmodel:HighlightLog

An interactive science item



A sequence

Student A · 12.9 s on the item · below the science cut

start_CS608Q01	drop_destination1	drop_destination2	drop_destination3	move	end_CS608Q01
+0.5 s	**+8.2 s**	+1.4 s	+1.4 s	+2.0 s	+0.0 s

Student B · 58.1 s on the item · at or above the science cut

start_CS608Q01	drop_destination1	drop_destination2	drop_destination3	move	end_CS608Q01
+0.1 s	**+53.2 s**	+1.4 s	+1.8 s	+1.7 s	+0.0 s

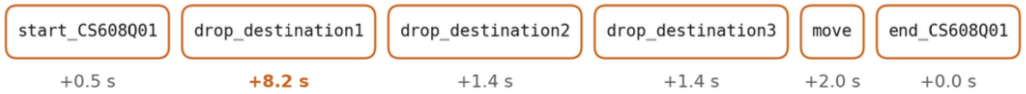
Three ways to represent behaviour

INSTYTUT BADAŃ EDUKACYJNYCH
Państwowy Instytut Badawczy

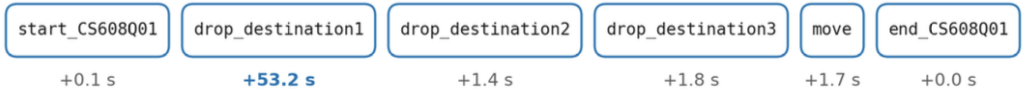
IFiS PAN

The same item, two students: identical actions, different pacing

Student A · 12.9 s on the item · below the science cut



Student B · 58.1 s on the item · at or above the science cut



1. Task-state indicators

The analyst defines the states of CS608Q01:
item opened → placing objects → submitted
and derives indicators from them. The meaning is proposed in advance and then validated. Another item may need its own states.

	A	B
time before first placement	8.2 s	53.2 s
objects placed	3	3
placements revised	0	0

2. General process indicators (classical models)

The same recipe is applied to every item: time on the item, time to the first action, number of actions, a rapid-response flag. Because each indicator is defined in the same way for every item, it is easy to average over the session. On this item the time to the first action shows the pause; the session averages no longer show on which item, or at which step, time was spent.

	A	B
on CS608Q01:		
time on the item	12.9 s	58.1 s
time to first action	8.2 s	53.2 s
actions on the item	4	4
rapid response (< 7.0 s)	no	no
averaged over the session:		
median time on an item	11.6 s	57.9 s
rapid responses	29%	0%

3. Learned representation (this paper)

Each event token is paired with the gap before it, and a sequence model reads both channels. The same code runs on every item, and the gap keeps its position in the record. The log shows where the pause occurred and how long it lasted, not what the student did during it.

gap before event (s)	start	drop1	drop2	drop3	move
Student A	0.5	8.2	1.4	1.4	2.0
Student B	0.1	53.2	1.4	1.8	1.7

One representation across many tasks

I am lazy

I would rather specify one general representation than write a new set of rules for every item.

A useful lesson from language modelling

Representations can be learned

Embeddings learn which words occur in similar contexts, instead of requiring a separate rule for every word.

Sequences preserve context

For assessment logs, we can learn from events and their timing across many tasks. Coding and validation still matter.

One representation across many tasks

I'm lazy, which makes more work for me

I want to build a single, generalized mathematical representation that can autonomously learn the underlying structure of test-taking behavior across any interface, without requiring me to explicitly define what a click or a drag means.

A useful lesson from language modelling

Representations can be learned

Embeddings learn which words occur in similar contexts, instead of requiring a separate rule for every word.

Sequences preserve context

For assessment logs, we can learn from events and their timing across many tasks. Coding and validation still matter.

Four studies, three kinds of process data

Study	Additional signal	Sample
1 PISA 2018 science	Actions and event-level timing	203,402 students 68 education systems
2a Web survey	Cursor trajectories	4,241 adults
2b Mathematics test	Cursor trajectories	1,147 main-study pupils Grades 7–8
3 Classroom mathematics	Webcam gaze + cursor + time	134 pupils in 12 classes Grade 5

Funding: NCN grant 2019/33/B/HS6/00937. Study 2b: IEA Research and Development Call 2023.

Study 1: PISA 2018 science

203,402
students

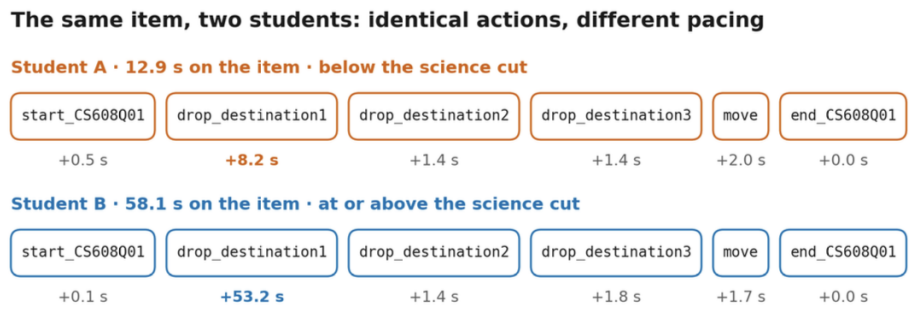
68
education systems

127
science items

Science classification

PV1SCIE ≥ 410.5
The cut for Level 2

Gender



Low self-reported effort

Post-test thermometer, EFFORT1 ≤ 5; 10.6% of respondents to that item

Parental education

Treating an event log as a sequence

The useful analogy with language is learning from context.

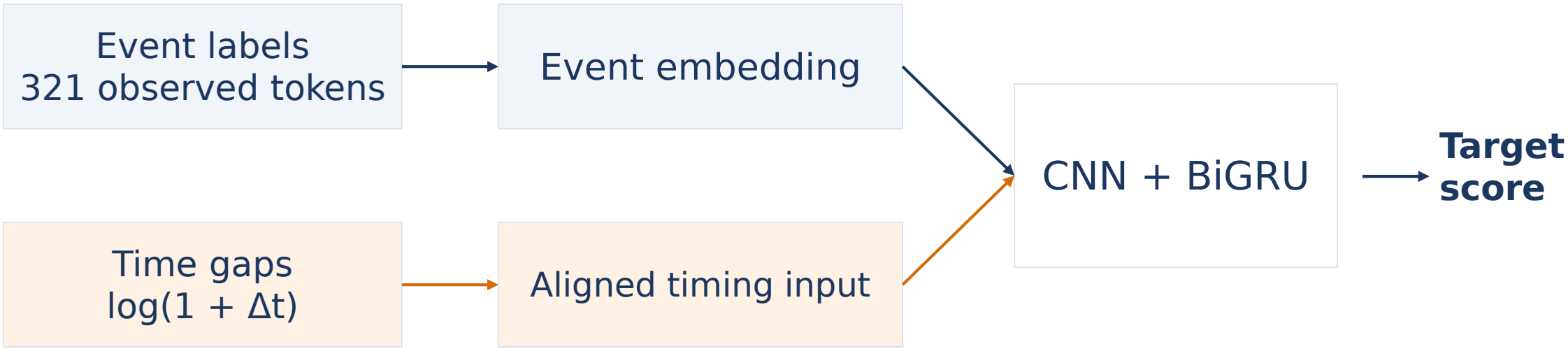
LANGUAGE	the	student	reads	carefully
ACTION LOG	ENTER	CLICK	RUN	NEXT
TIME GAPS	Δt_1	Δt_2	Δt_3	Δt_4

Event vectors and sequence models preserve context and timing.

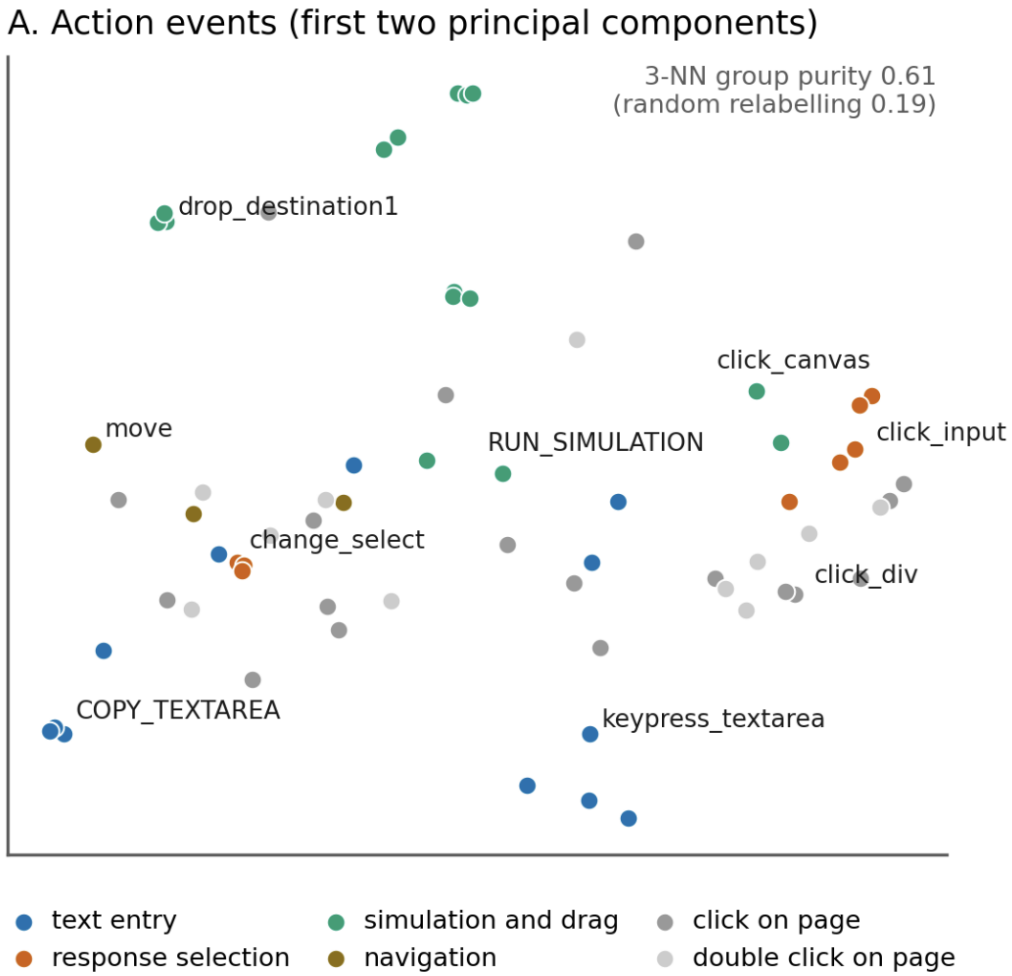
These studies use GRU / LSTM networks and embeddings, not large language models.

Events and gaps enter the model together

A common event vocabulary, with each time gap attached to its event



Learned event embeddings



Related events tend to be neighbours

Text entry, response selection and simulation actions show structure in the learned embedding space.

0.61

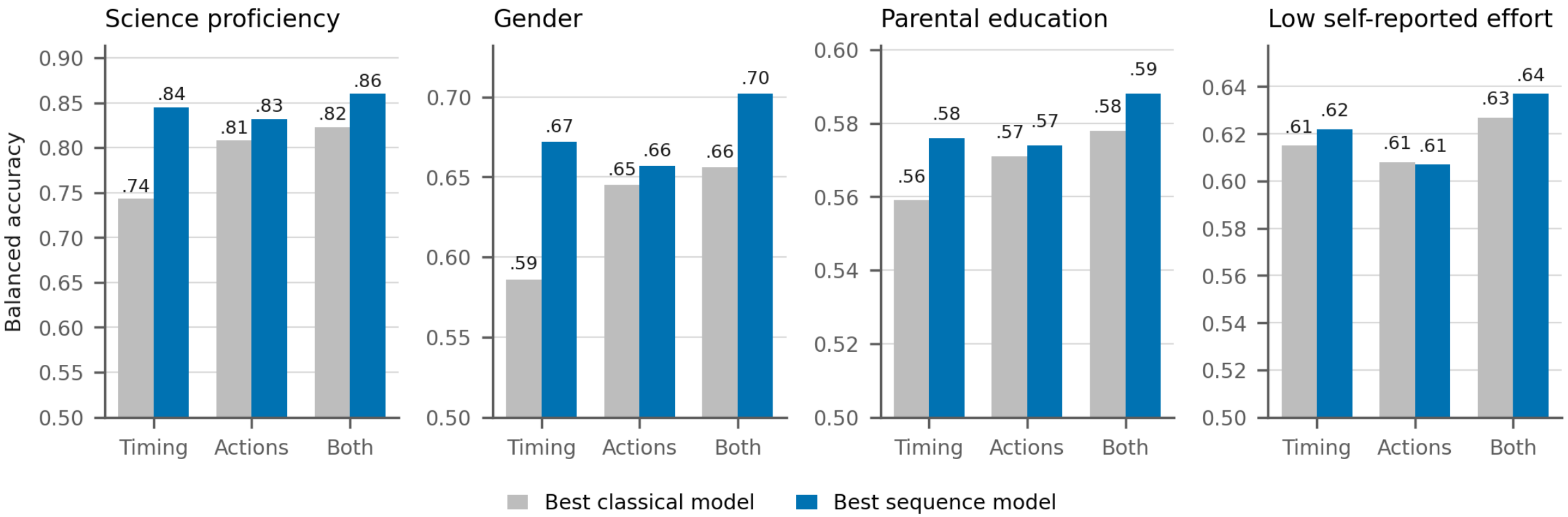
3-neighbour group purity

0.19

after random relabelling

The plot is a 2D projection of higher-dimensional vectors.

Timing drives the larger model difference



Why do we use process data?

INSTYTUT BADAŃ EDUKACYJNYCH
Państwowy Instytut Badawczy

IFiS PAN

1 **Understand the response process**

Local pauses carry information that event counts and total time can miss.

2 **Improve the assessment**

Keep event timestamps. Shorten it.

3 **Predict information beyond the score**

Identify low self-reported effort.

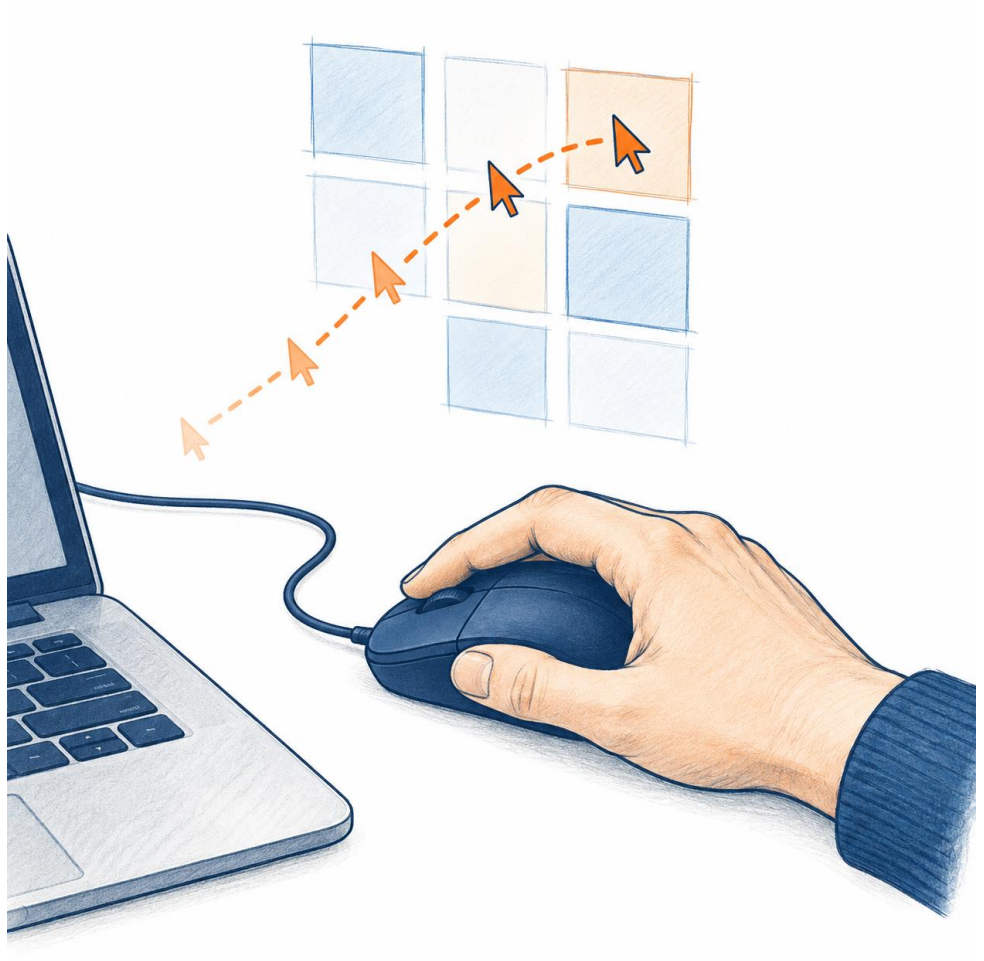
4 **Because the data and models are interesting**

A good reason to explore and have fun.

02 CURSOR TRAJECTORIES

What happens between clicks?

Two experimental studies:
a web survey



Study 2a: induced careless responding

4,241 adults from Polish opt-in panels. Desktop and laptop respondents.

Instruction	N
Regular	1,075
Motivating	2,031
Careless	**1,135**

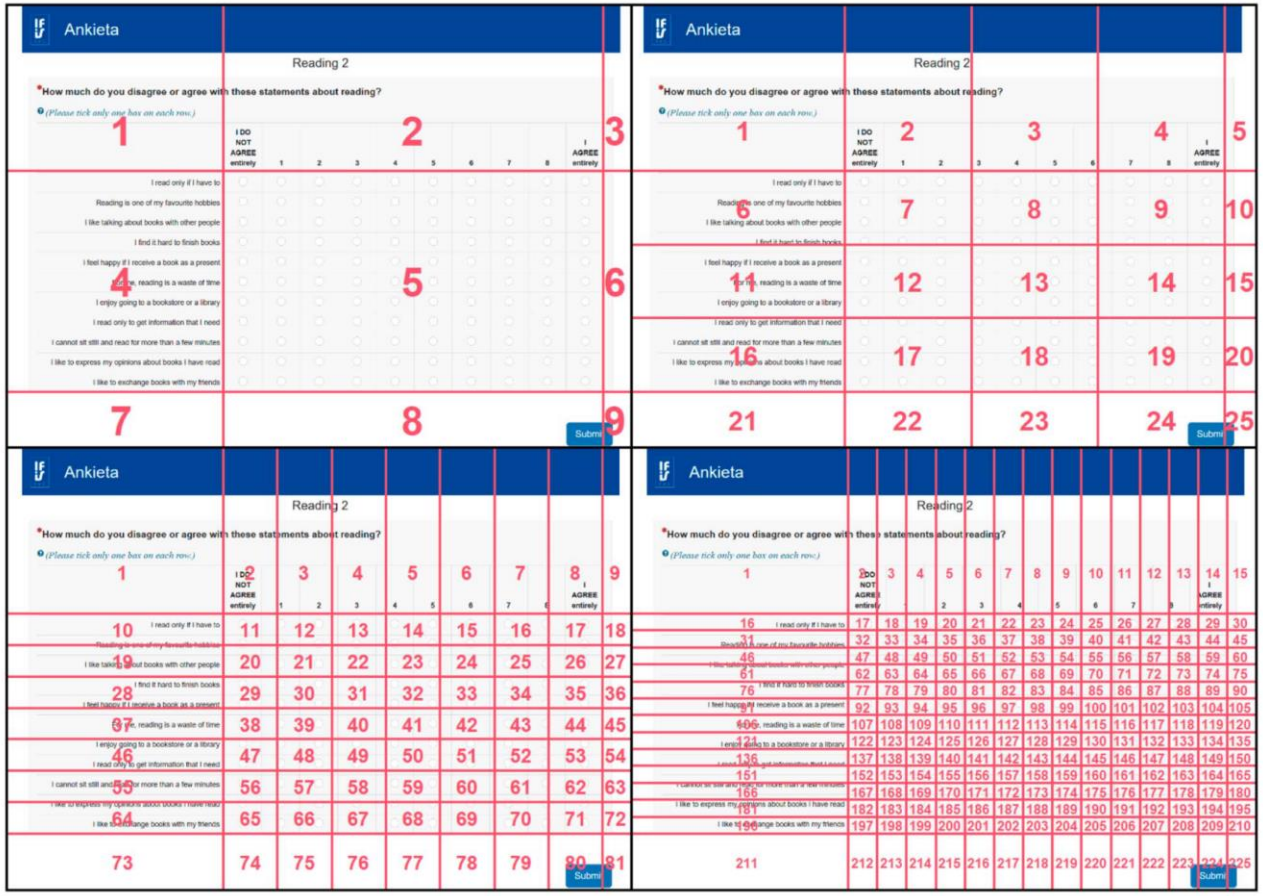
The careless instruction

“Please fill out the survey as if you didn’t feel like doing it and only participated to receive compensation. Please fill out the survey in an inattentive manner. Regardless of your answers, you will receive full compensation.”

The target is the assigned instruction, rather than a diagnosis of natural carelessness.

Approximate Areas of Interest

A screen grid replaces manually labelled semantic regions.



9, 25, 81 or 225 cells

The grid resolution is a model parameter.

Crossing a cell boundary creates a new token.

Pokropek, Żółtak & Muszyński (2024), Zeitschrift für Psychologie, Figure 1

A cursor trajectory becomes a token sequence

Location changes, clicks and timestamps enter the sequence.

1-2-7-c-c-8-c-13-c-c-12-c-17-...

t_1-t_2-t_3-t_4-t_5-t_6-t_7-t_8-t_9-t_{10}-t_{11}-t_{12}-t_{13}-$t_{...}$

1 **Assign each position to a cell**

2 **Keep crossings and clicks**

3 **Attach the timestamps**

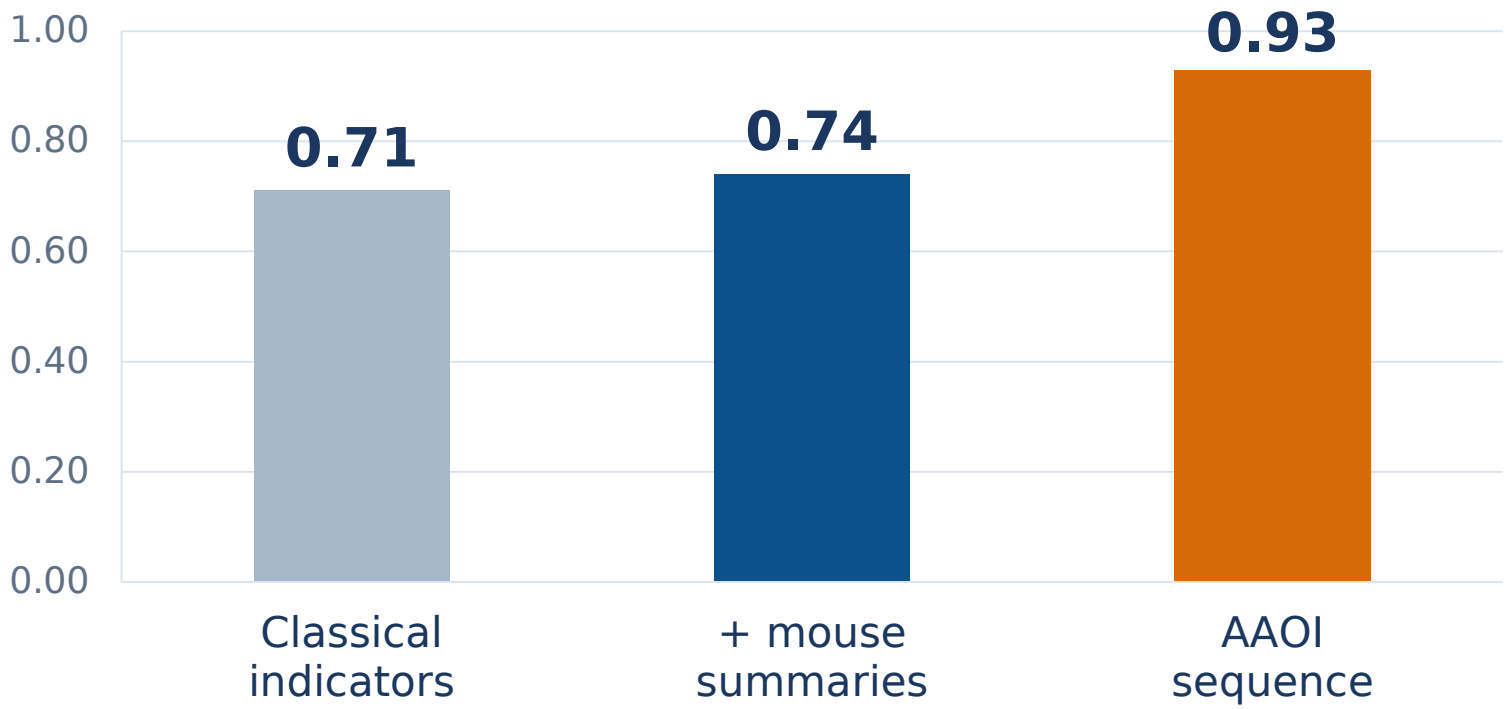
4 **Fit a sequence classifier**

The model uses both the order of movements and when they occur.

Pokropek, Żółtak & Muszyński (2024), Zeitschrift für Psychologie, Method

AAOI sequences classify the survey instruction much better

Balanced accuracy on respondent test data, mean across 10 folds



BiLSTM, 5 × 5 grid

Accuracy 0.94
F1 score 0.90
Specificity 0.96

9–225 cells give similar performance.

Pokropek, Żółtak & Muszyński (2024), Zeitschrift für Psychologie, Table 4

02 CURSOR TRAJECTORIES

What happens between clicks?

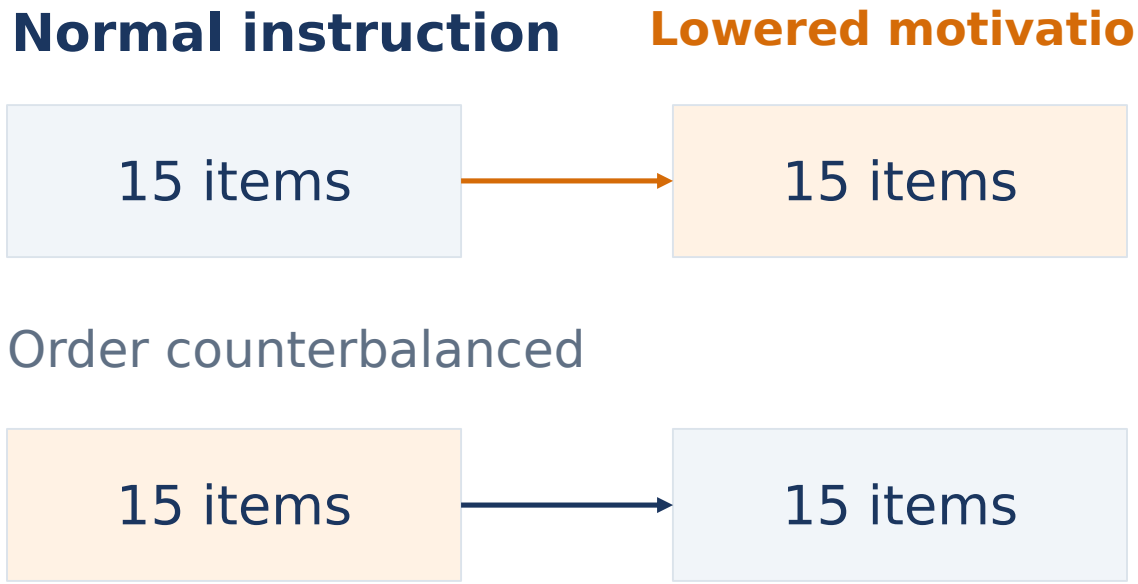
Two experimental studies:
a mathematics test

Study 2b: the same pupil in two conditions

Grades 7–8, Polish schools, April–May 2024. 30 mathematics items in two blocks.

1,147

accepted main-study records



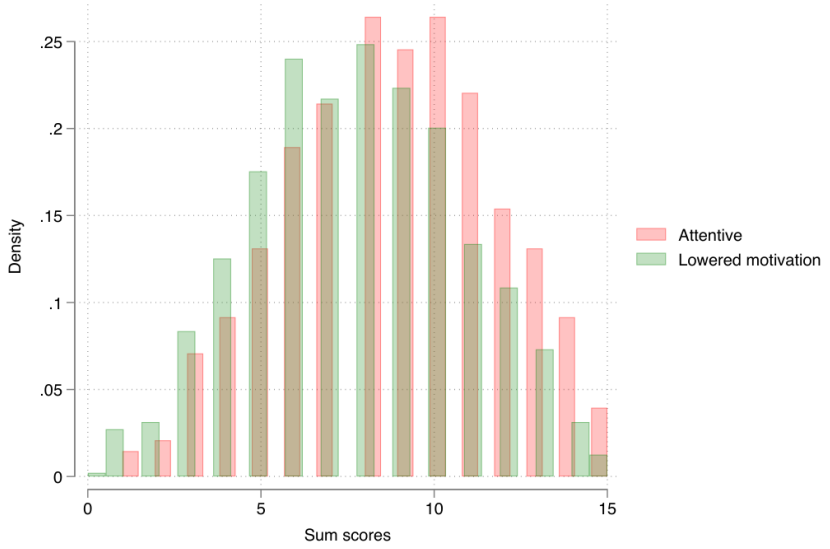
One intentionally flawed item had four correct options out of five.

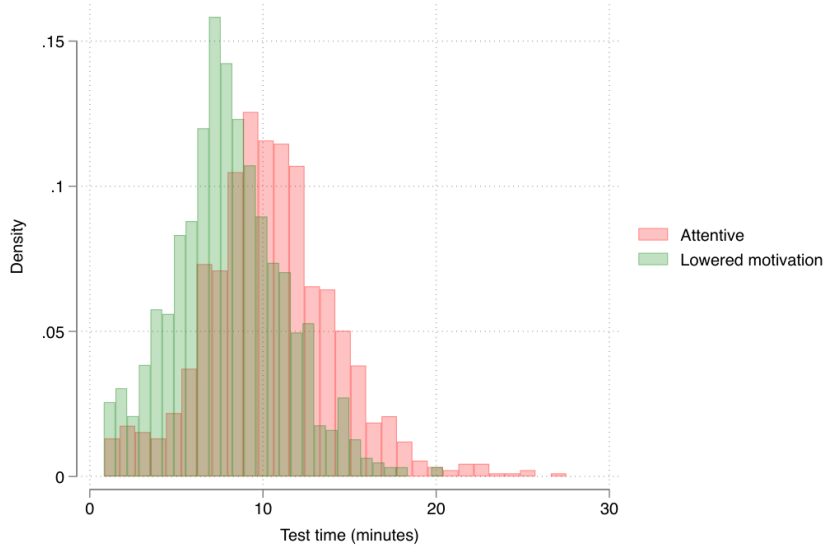
Reported models include pilot data and apply additional compliance and data-quality exclusions.

Pokropek, Żółtak & Muszyński, IEA R&D technical report, Sections 2–9

Effects of the motivation manipulation

Outcome	Attentive		Lowered motivation		Mean difference	
	Mean	SD	Mean	SD	Raw	SMD
Sum score (0–15)	8.7	3.0	7.8	3.0	1.0	0.3
Test time (minutes)	10.4	3.8	8.0	3.2	2.4	0.7

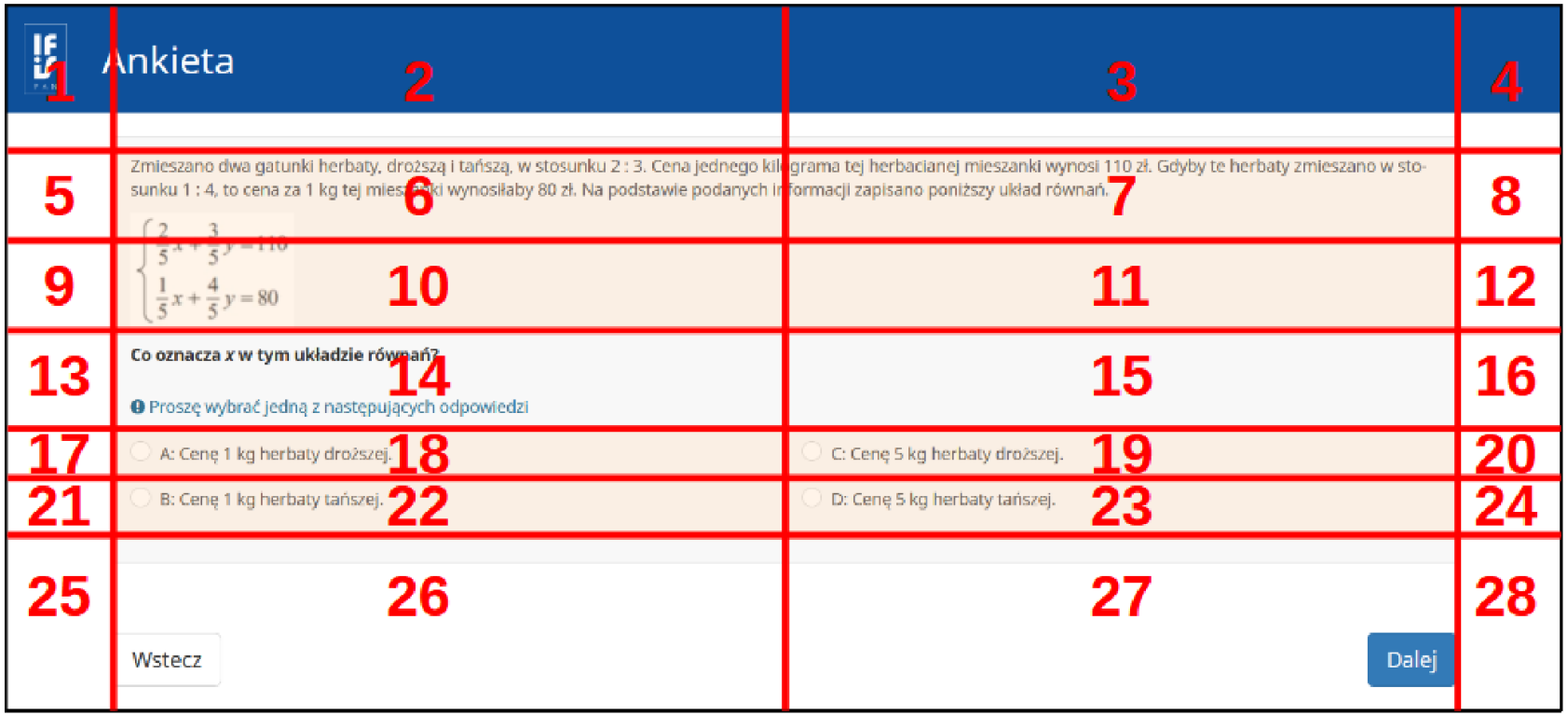




Reported differences use unrounded means. SMD = standardised mean difference.

The grid follows the item structure

Example: 28 cells spanning the stem, lead-in, response options and margins



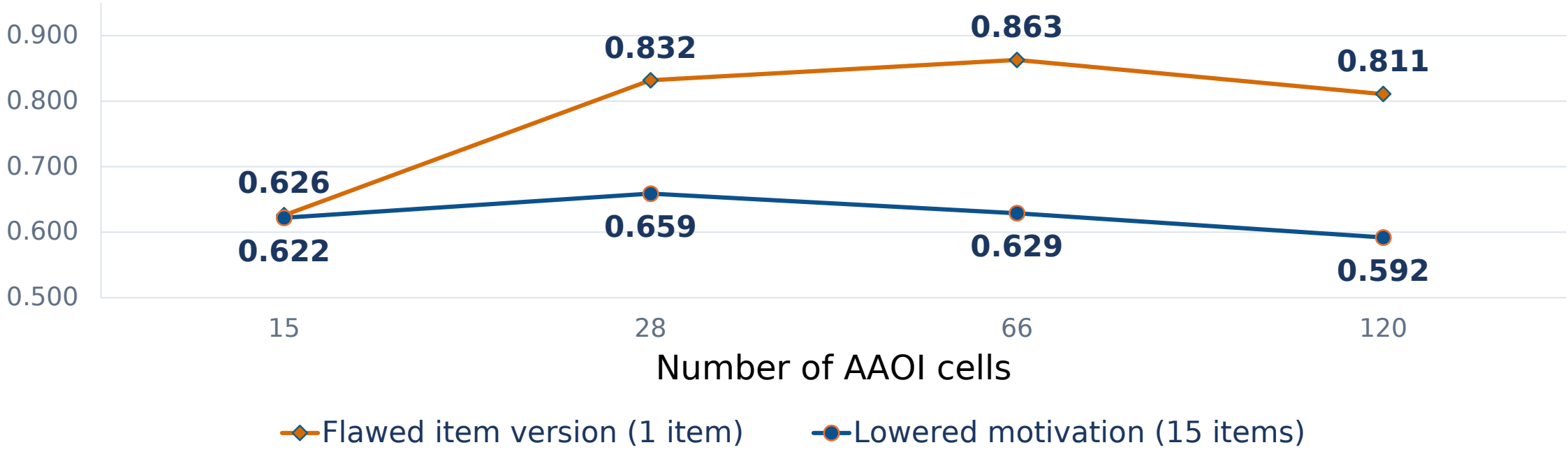
Comparable regions across item layouts

The grid adapts to the task's functional areas.

Screen layout remains an input.

A flawed item is easier to classify than the motivation condition

Balanced accuracy on the report’s held-out test data



The highest reported values are 0.863 for the item version and 0.659 for motivation.

Pokropek, Żółtak & Muszyński, IEA R&D technical report, Tables 12 and 14

Why was motivation harder to classify?

The instruction did not always change behaviour

183

16.0% of pupils reported low effort in the regular condition

287

25.0% reported careful responding when asked to guess

Condition-specific exclusions; 13 pupils met both criteria.

An average shift is easier to detect

Scores and time changed (SMD 0.3 and 0.7), but the two conditions still overlapped substantially.

IRT evidence was mixed

Difficulty shifts were present. AIC and SABIC favoured equal discriminations; BIC favoured full invariance.

Assigned instructions are an imperfect label for a pupil’s actual engagement.

Why do we use process data?

INSTYTUT BADAŃ EDUKACYJNYCH
Państwowy Instytut Badawczy

IFiS PAN

1 **Understand the response process**

Cursor paths help describe how respondents move through a task or response grid.

2 **Improve the assessment**

Screen for item problems: flawed-item classification reached balanced accuracy 0.863.

3 **Predict information beyond the score**

Survey instruction was easier to classify than reduced motivation in mathematics.

4 **Because the data and models are interesting**

A good reason to explore and have fun.

Pokropek, Żółtak & Muszyński (2024; 2025)

03 WEBCAM GAZE

What happened during the pause?

A period without clicks can include reading, thinking, writing or distraction.

Study 3: webcam gaze in the classroom

134 pupils, 12 classes

Grade 5 in Poland, December 2025
Eight mathematics tasks
School laptops, sessions up to 45 minutes

Recorded together

Response times
Cursor movements
Webcam gaze through RealEye

Effort criterion

“How much effort did you put into solving the tasks?”

Not at all
A little
Moderately
Very much

133 responses, including 36 in the two lowest categories

How the webcam estimates gaze

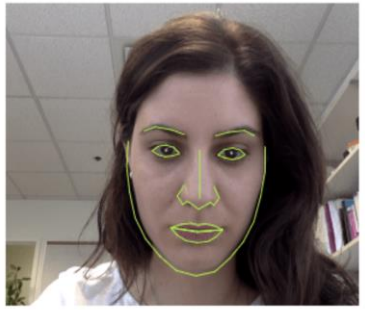
INSTYTUT BADAŃ EDUKACYJNYCH
Państwowy Instytut Badawczy

IFiS PAN

1 Camera frames → **2** Face and eye features → **3** 9-point calibration → **4** Estimated gaze

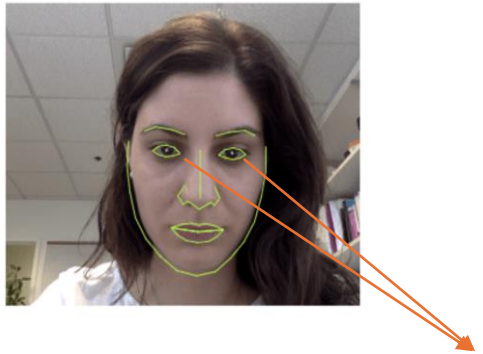
Ordinary laptop webcam

Model estimates eye position



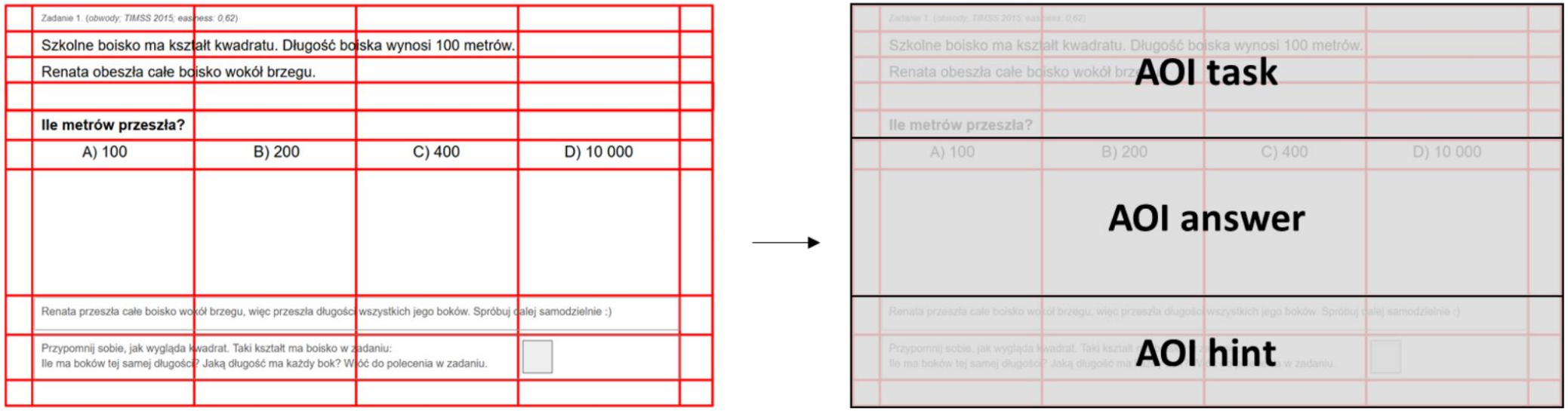
Maps eye features to the screen

Coordinates (x, y, t) and quality



AOIs

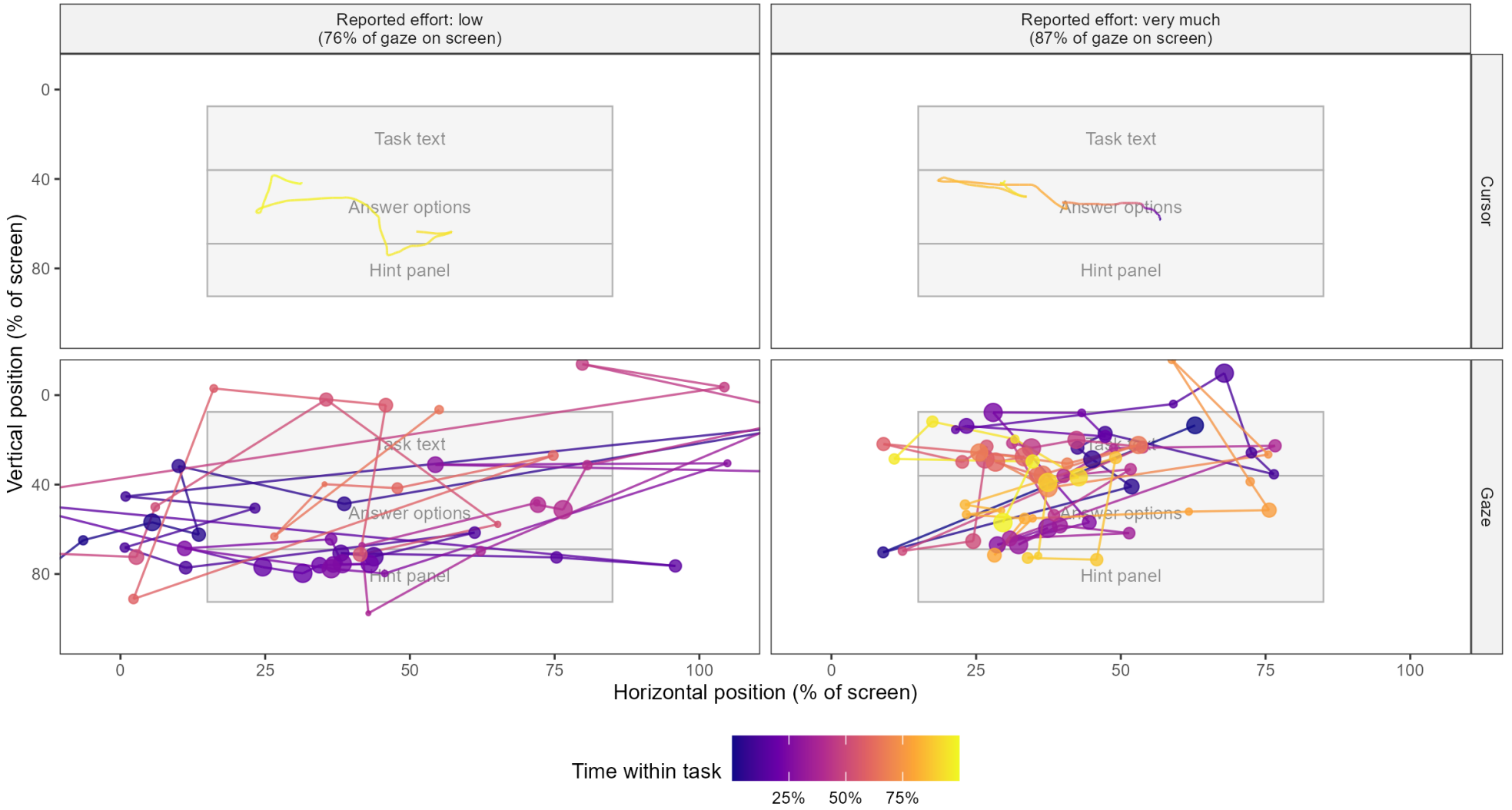
Areas of interest on the task screen



Webcams do not allow for high precision

Two students, two different recordings

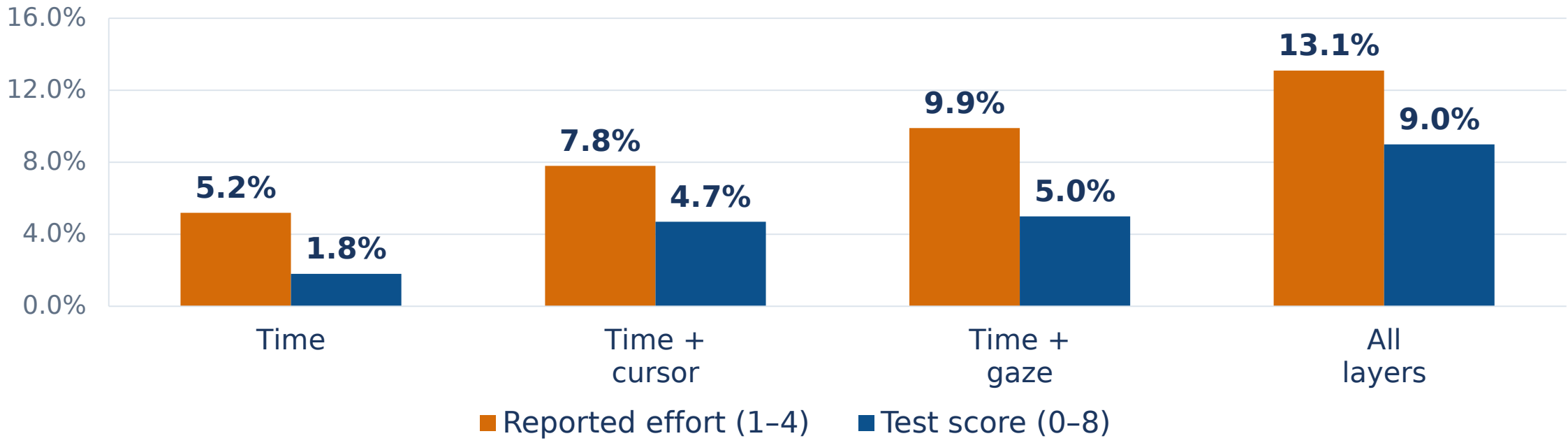
Same task. Cursor at the top, gaze at the bottom.



This pair illustrates what the channels record. It does not establish a group difference.

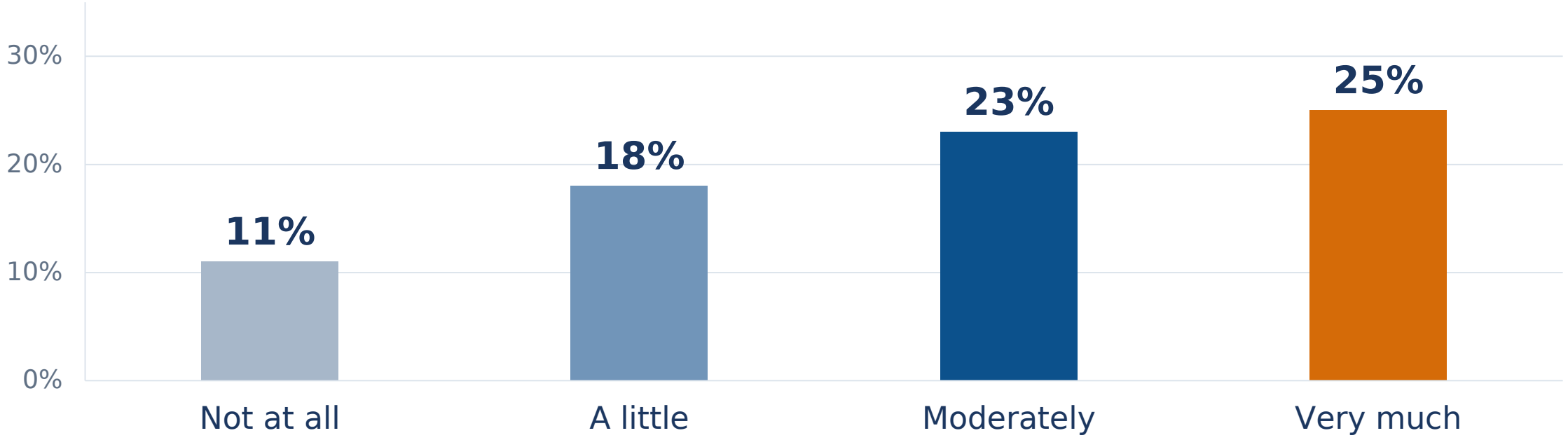
Adding gaze raised R^2, with substantial uncertainty

Explained variance in the same sample, before correction for model complexity



More reported effort accompanied more estimated off-screen gaze

Mean off-screen fixation share across the four response categories



Looking away is compatible with working, but it does not demonstrate work.

Why do we use process data?

INSTYTUT BADAŃ EDUKACYJNYCH
Państwowy Instytut Badawczy

IFiS PAN

1 **Understand the response process**

Looking away can accompany effort; gaze adds context to an otherwise silent pause.

2 **Improve the assessment**

Use the recordings to investigate task layout.

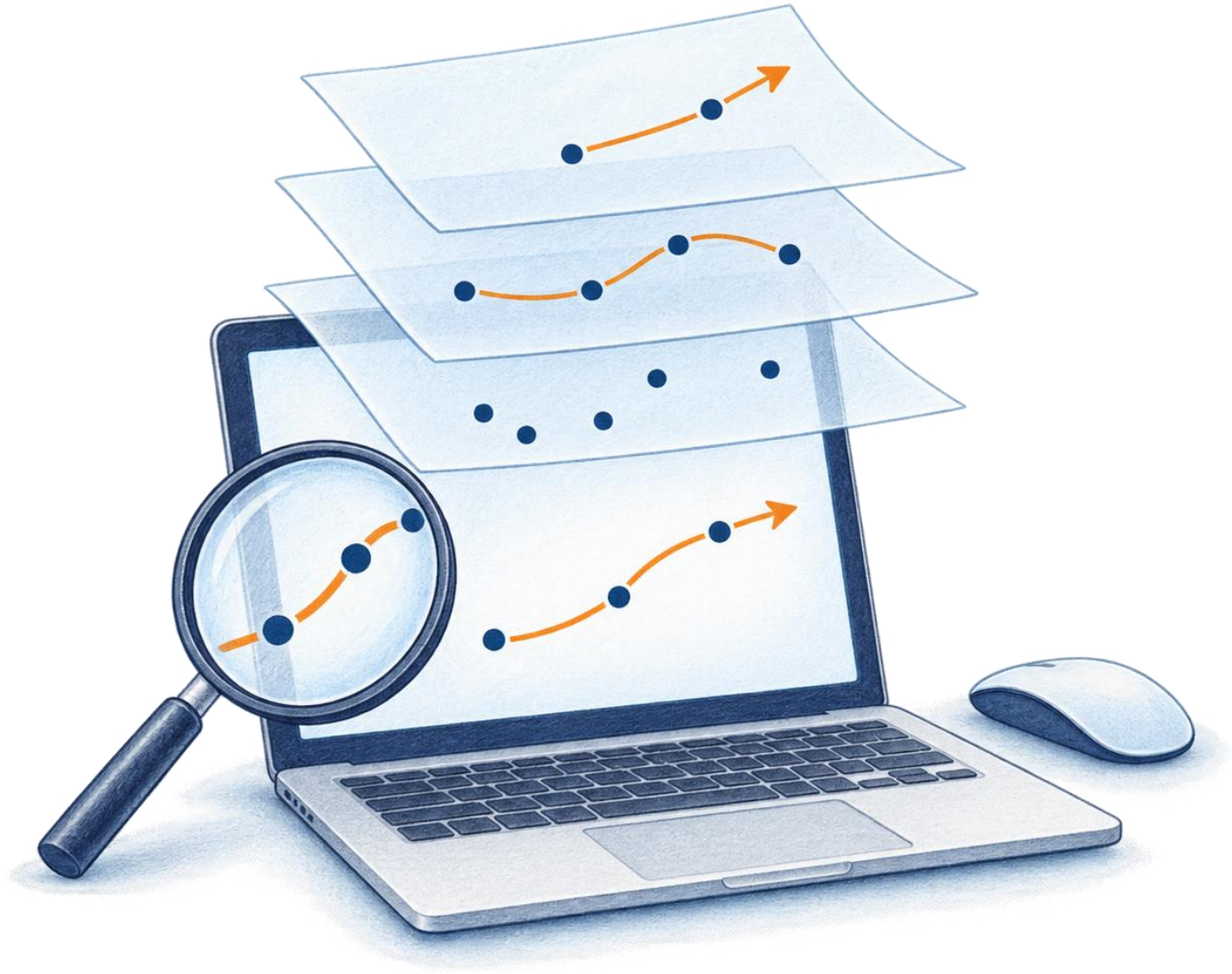
3 **Predict information beyond the score**

Reported effort; maybe disengagment.

4 **Because the data and models are interesting**

A good reason to explore and have fun.

Summary



Even a lazy researcher can learn something

INSTYTUT BADAŃ EDUKACYJNYCH
Państwowy Instytut Badawczy

IFiS PAN

1 **Understand the response process**

Each layer of data brings us closer to understanding how people respond.

2 **Improve the assessment**

Promising tools for detecting item problems and improving assessments.

3 **Predict information beyond the score**

Additional layers can improve prediction, depending on what we want to predict.

4 **Because the data and models are interesting**

And we had a lot of fun along the way.

Publications and working papers

Pokropek & Borgonovi (2026)

Predicting science proficiency and self-reported effort from assessment logs: The contribution of event-level timing.

Preprint forthcoming

Pokropek, A., Żółtak, T., & Muszyński, M. (2024)

Identifying careless responding in web-based surveys: Exploiting sequence data from cursor trajectories and approximate areas of interest.

Zeitschrift für Psychologie, 232(2), 95–108. doi:10.1027/2151-2604/a000555

Pokropek, A., Żółtak, T., & Muszyński, M. (2025)

Approximate Areas of Interest for Enhanced Understanding of Student Motivation and Task Interaction in IEA Assessments.

Technical report, IEA Research and Development Call 2023.

Pokropek, Binkiewicz, Boros & Glica (2026)

What webcam eye-tracking adds to response-time indicators of test-taking effort.

Preprint forthcoming

Boros, Binkiewicz, Glica & Pokropek (2026)

The impact of cognitive hints on mathematical problem solving: A webcam-based eye-tracking study

Preprint https://doi.org/10.31234/osf.io/ge25x_v2