Potential benefits and challenges of log data in large-scale assessment

Frank Goldhammer[1,2]

Centre for Technology Based Assessment

[1]DIPF | Leibniz Institute for Research and Information in Education

[2]Centre for International Student Assessment (ZIB)

About me

unsplash

DIPF

- Professor for Educational and Psychological Assessment at Goethe University Frankfurt a. M.
- Head of the Centre for Technology Based Assessment (TBA) at DIPF | Leibniz Institute for Research and Information in Education
- Member of the Centre for International Student Assessment (ZIB)

- In-depth contact with the collection, use and interpretation of log data in LSAs
  - First: PIAAC 2012, consortium member, log data project
  - Last: PISA 2025, expert group member

Overview

- Log data in LSAs
- Individual differences in response processes
- Benefits of using log data
- Challenges of using log data
- Conclusions

Overview

- Log data in LSAs
- Individual differences in response processes
- Benefits of using log data
- Challenges of using log data
- Conclusions

Log data

- Log data is **event-based raw data** (e.g., Goldhammer et al., 2020)
- Purpose of logs in **SW development**: debugging, performance analysis, maintenance, security management …
- **Structure** (e.g., Kroehne et al., in prep)
  - Event
    - Type
    - Time stamp
    - Event-specific attributes
      - atomic
      - complex

PIAAC 2012: Problem solving

```
<taoEvent Name="stimulus"
Type="TEXTLINK"
Time="164959">id=u10a_default_txt
15|*$href=unit10page14|*$target=_se
lf</taoEvent>
```

Two fictitious examples

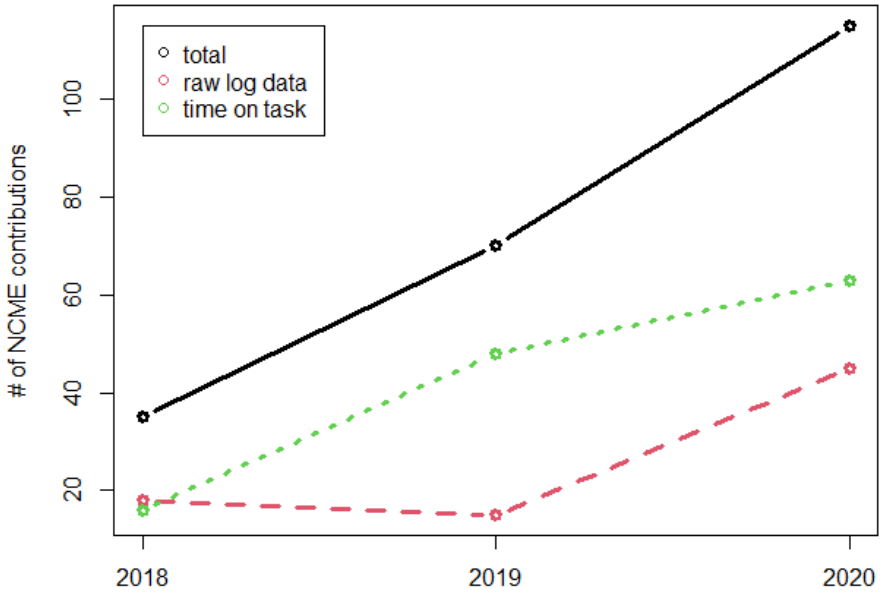
```
<event type=”click” timestamp=”…”
x=”100” y=”100”/>
```

```
<event type=”click” timestamp=”…”>
  <clickposition>
       <x>100</x>
       <y>100</y>
     </clickposition>
</event>
```

Increasing popularity of log data in the research community

- Review of research based on **PIAAC 2012 log data** (Goldhammer et al., 2020)
  - 2014 – 2019: 15 published studies
  - Process representation:
    - time on task (included in the PUF, generic process indicator)
    - sequence of actions (n-grams)

- Review of **NCME contributions** based on log data (Becker et al., 2020)



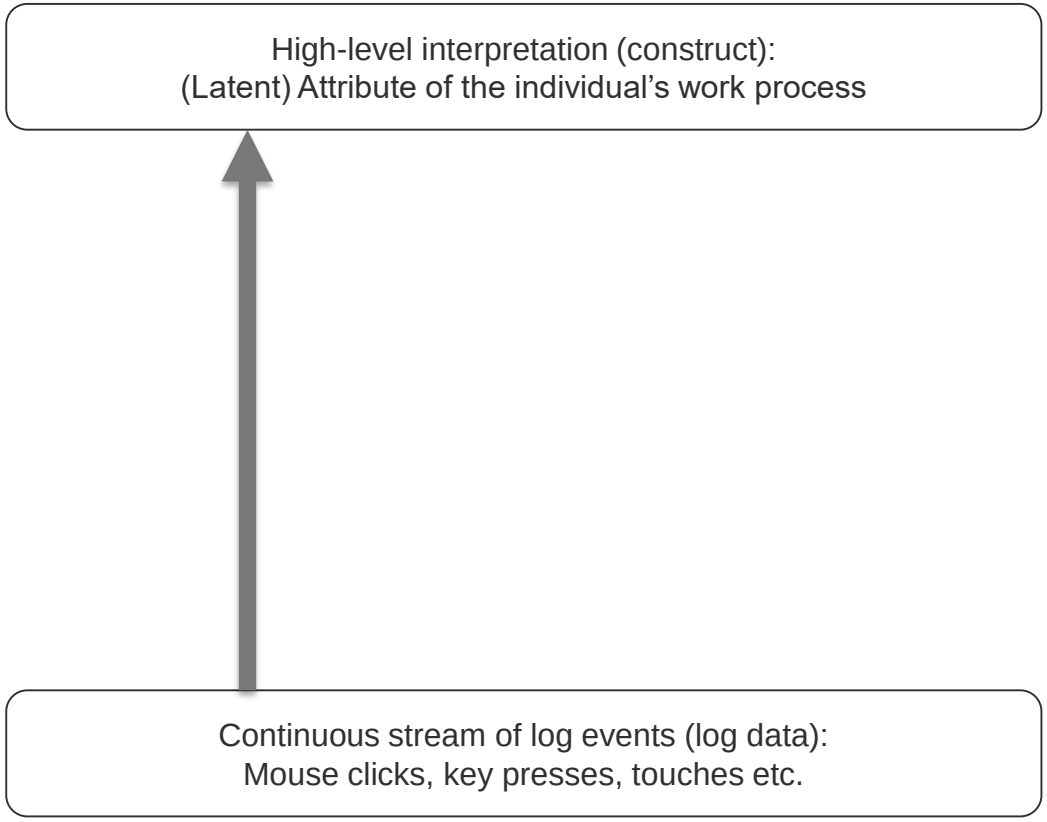
Reasoning from evidence

- **Asessment** – reasoning from observed response behavior in test items captured by log data

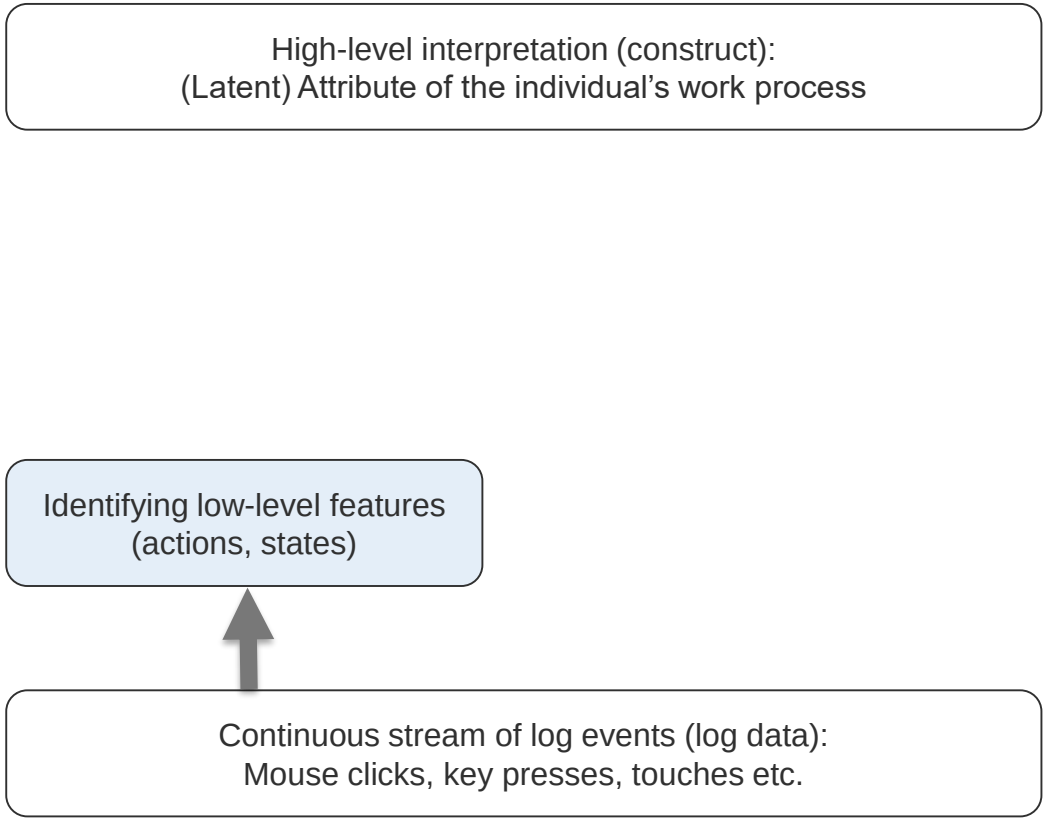
- Integrating concepts of
  - hierarchical evidentiary reasoning from **continuous assessment** (Mislevy, 2019) and
  - **Evidence-Centered Design** (ECD; Mislevy et al., 2003)

Goldhammer, F., Hahnel, C., Kroehne, U., & Zehner, F. (2021). From byproduct to design factor: On validating the interpretation of process indicators based on log data. *Large-Scale Assessments in Education*, *9*(1), 1-25.

Reasoning from evidence: Bridging the gap

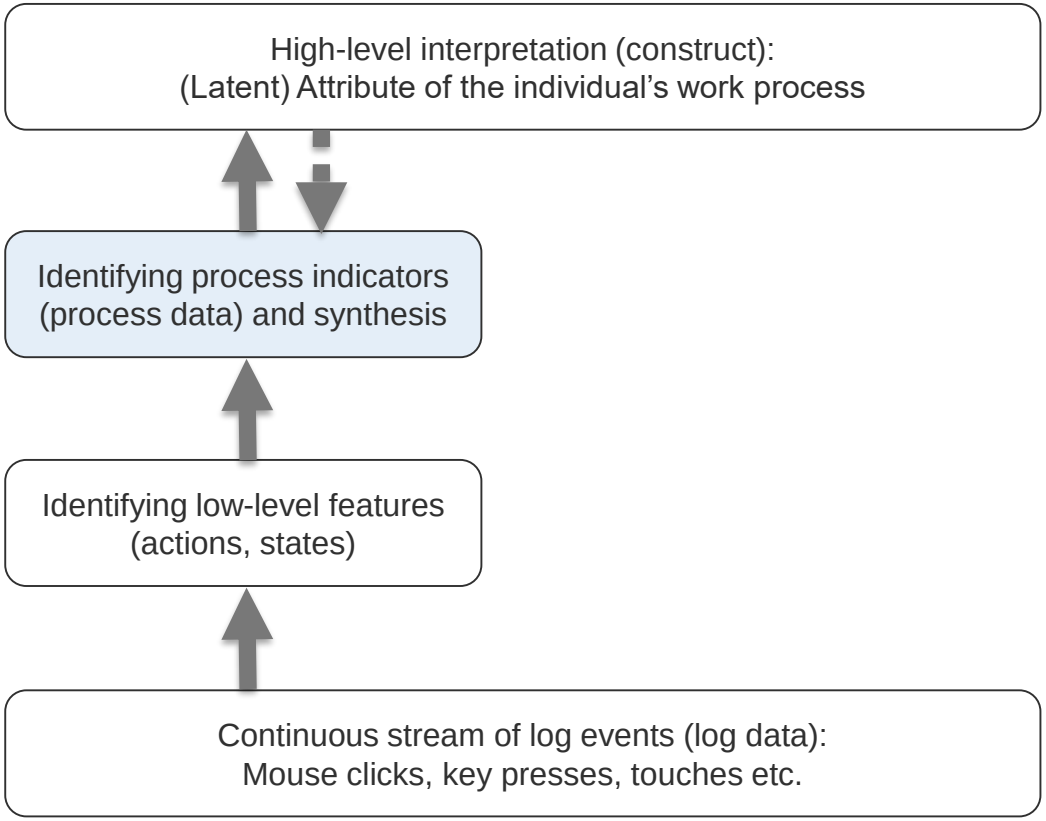


Reasoning from evidence: First inference



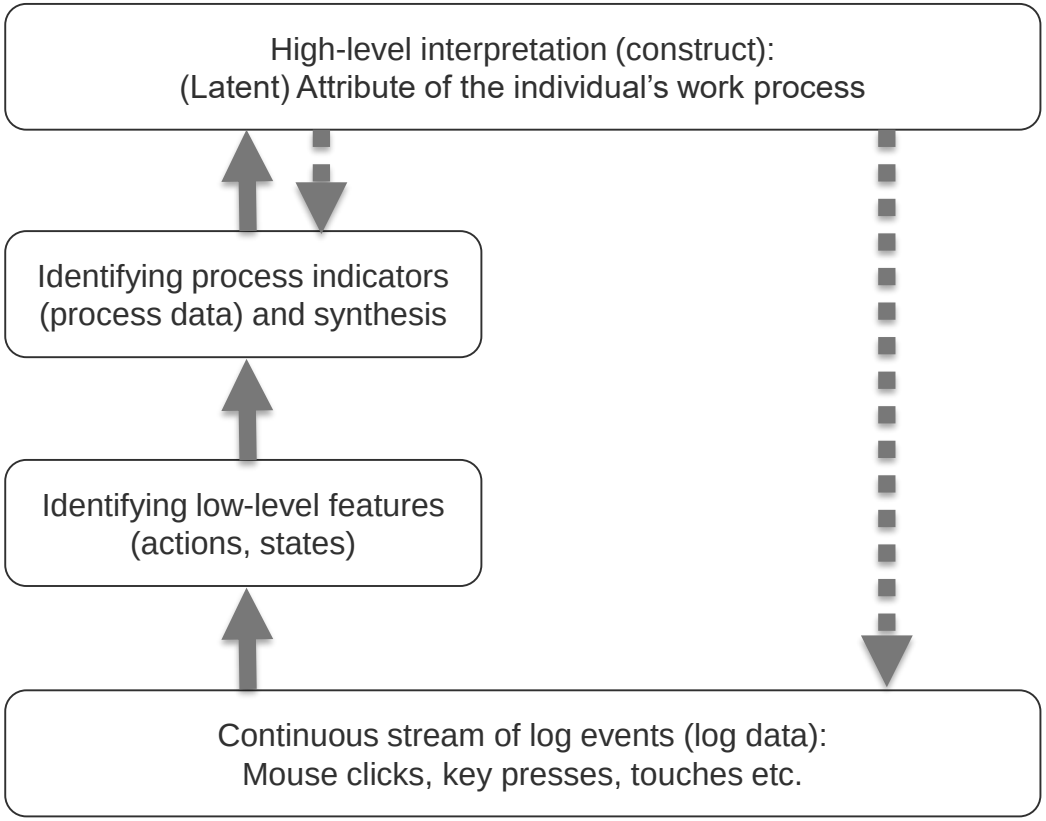
(e.g.,
Hao & Mislevy, 2018;
Kroehne & Goldhammer, 2018;
Mislevy et al., 2014;
Rupp et al., 2012)

Reasoning from evidence: Second inference

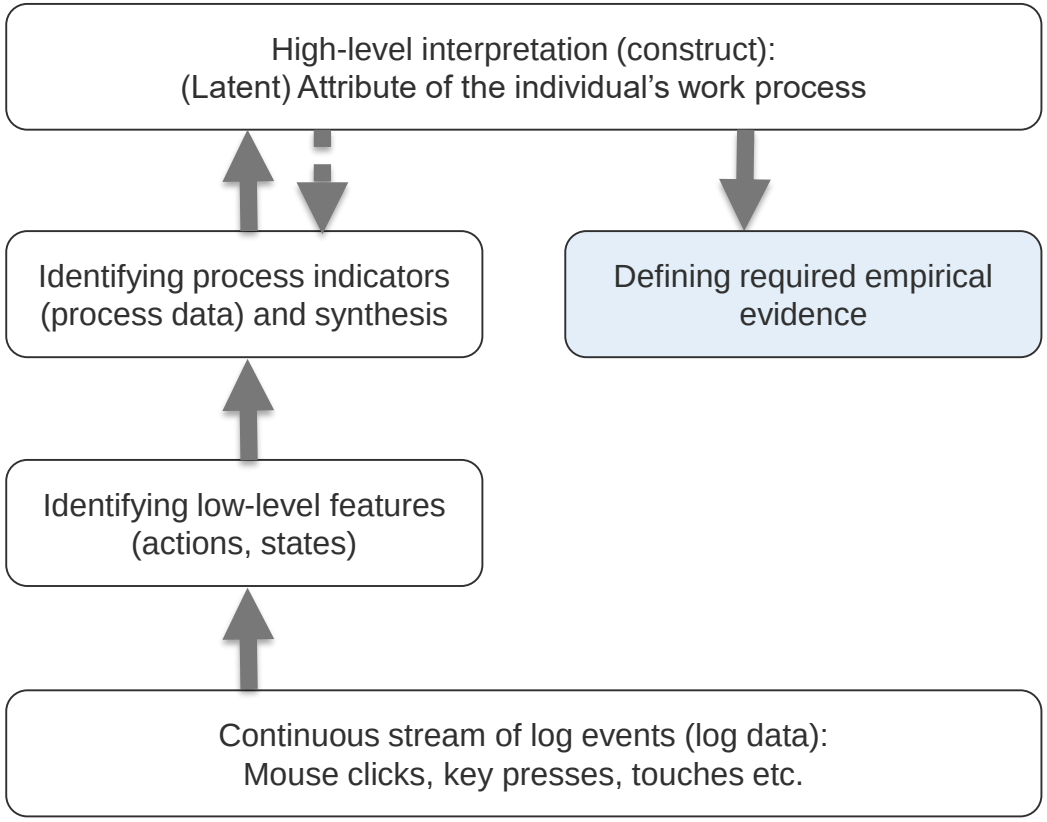


(e.g,.
Behrens & DiCerbo, 2014;
Kerr et al., 2016;
Klerk et al., 2015;
Levy, 2020)

Reasoning from evidence: Theory from the start

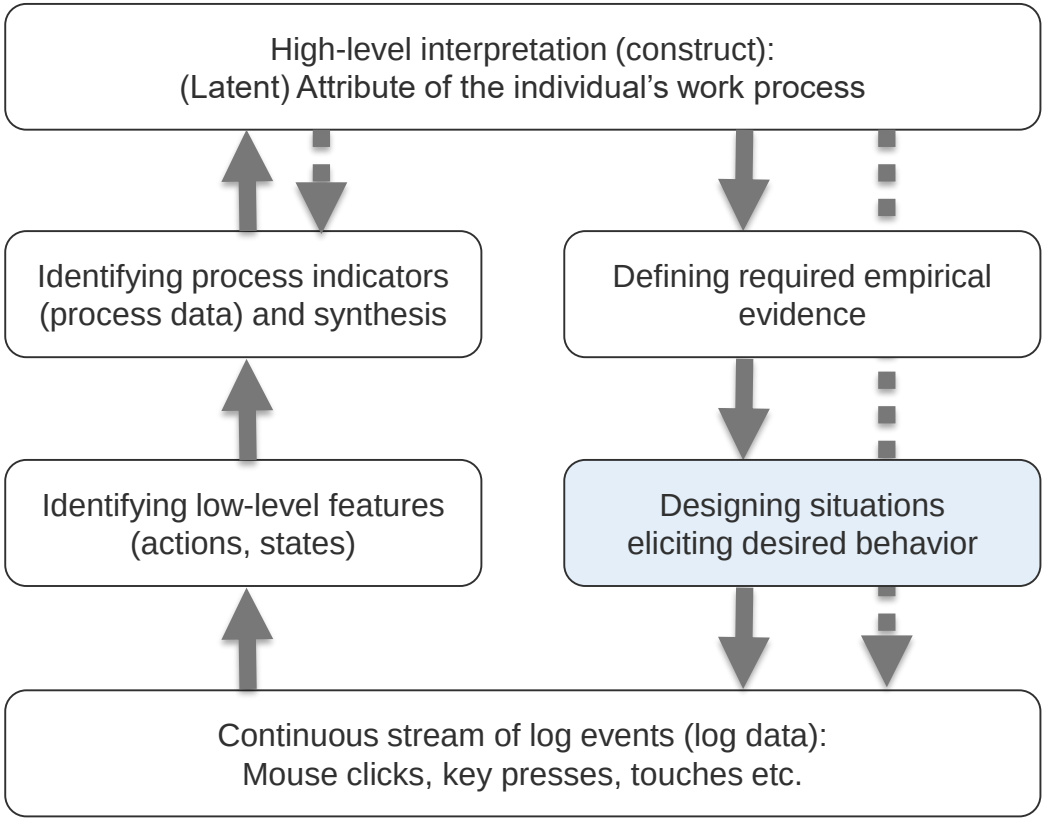


Reasoning from evidence: What is needed?



(e.g,. Mislevy et al., 2003, 2012; National Research Council, 2001)

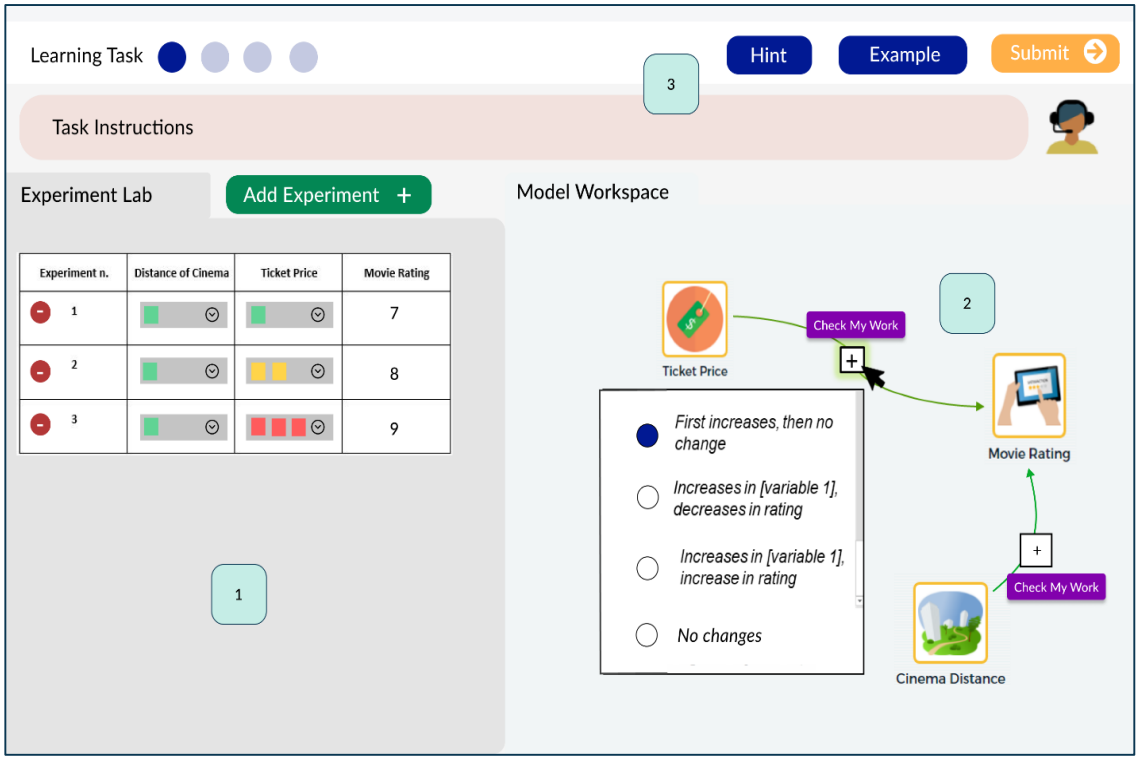
Reasoning from evidence: How to get it?



(e.g,.
DiCerbo, 2014;
Goldhammer & Zehner, 2017;
Hahnel et al., 2019)

PISA 2025 innovative domain: „Learning in the Digital World“

- Assessment of **self-regulated learning (SRL)** utilizing process data
- **Task design**: opportunities to learn, affordances to demonstrate monitoring and regulating behavior



(fictitious example for system modelling type)

Overview

- Log data in LSAs
- Individual differences in response processes
- Benefits of using log data
- Challenges of using log data
- Conclusions

Two (related) lines of research

- Invitation to keynote: „We are quite interested to hear your insights on the potential gains and possible challenges of log data in large-scale assessments, closely related to some of your recent work:

Response process

- “... one may think broadly of response processes as the **mechanisms that underlie** what people do, think or feel when interacting with, and responding to, the item or task and are responsible for generating observed **test score** variation.” (Hubley & Zumbo, 2017, p. 2).

- **Multi-dimensionality** of the ‚response process‘ (see e.g., Maddox, 2023): cognition, motivations, emotions, behavior

- **Process indicators** can be used to capture differences in (latent) response processes empirically

- Some differences in response processes - affecting the test score - may be **construct-relevant** others not (e.g., Anraneda et al., 2022)

Response process – Individual differences

- **Construct-relevant differences** in the response process should be taken into account in the scoring rules
  - indirectly (i.e., an appropriate strategy produces a correct result)
  - directly (e.g., applying a more efficient solution strategy gives extra credit, such as Signed Residual Time scoring rule by Maris & van der Maas, 2012)

- **Construct-irrelevant differences** in the response process should be controlled experimentally/statistically (e.g., differences in test-taking engagement, differences in the speed-accuracy tradeoff)

Two (related) lines of research: Response speed

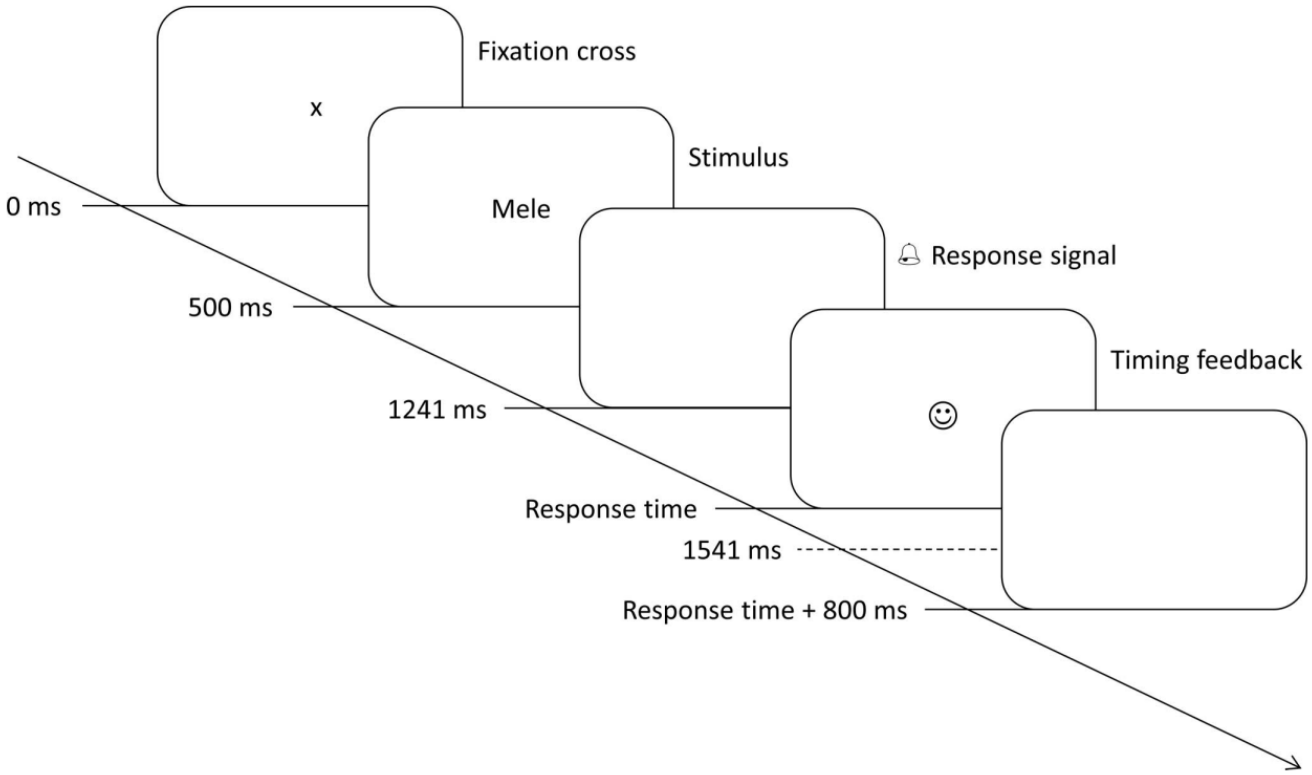
- **Observing and making use** of individual differences in speed
  - Value of response speed (e.g., Molenaar, 2015)
    - Increasing measurement precision of latent ability
    - Insights into the respones process (Goldhammer et al., 2014, 2021a)

- **Experimental control** of individual differences in speed and the speed-accuracy tradeoff, respectively (e.g., Goldhammer, 2015; Goldhammer et al., 2021b)
  - Speeded tests of cognitive efficiency
    → working quickly matters
  - Item-level time limits to control the tradeoff
  - Reading component skills: Word-recognition, semantic integration

Controlling response speed in reading component skills experimentally

Figure 1

Trial of the Word Recognition Task in the Timed Condition With a Stimulus Presentation Time of 741 ms



(Goldhammer et al., 2021b, p. 867)

Predicting PISA reading comprehension

Table 4

Latent Regression of Reading Comprehension on Word Recognition and Sentence-Level Semantic Integration

Model	Criterion	Predictors	Timed/ untimed	$\beta_{j.MLR}$	*SE*	R^2	*SE*	$\beta_{j.Bayes}$ (*Posterior SD*)
1	Reading comprehension	Word recognition ability	Timed	.476***	0.075	.554	0.037	.466 (0.074)
		Semantic integration ability	Timed	.302***	0.079			.306 (0.077)
2	Reading comprehension	Word recognition ability	Untimed	.377***	0.057	.361	0.037	.379 (0.059)
		Semantic integration ability	Untimed	.300***	0.055			.292 (0.059)
3	Reading comprehension	Word recognition speed	Untimed	.051	0.057	.006	0.008	.053 (0.053)
		Semantic integration speed	Untimed	−.089	0.055			−.092 (0.052)
4	Reading comprehension	Word recognition ability	Untimed	.480***	0.078	.450	0.044	.480 (0.070)
		Semantic integration ability	Untimed	.383***	0.085			.373 (0.082)
		Word recognition speed	Untimed	.242***	0.064			.244 (0.057)
		Semantic integration speed	Untimed	.163*	0.065			.159 (0.064)
5	Reading comprehension	Word recognition ability	Untimed	.264**	0.079	.597	0.036	.226 (0.072)
		Semantic integration ability	Untimed	.014	0.124			.059 (0.093)
		Word recognition speed	Untimed	.084	0.067			.078 (0.054)
		Semantic integration speed	Untimed	−.022	0.079			−.009 (0.062)
		Word recognition ability	Timed	.346***	0.085			.384 (0.080)
		Semantic integration ability	Timed	.264**	0.098			.207 (0.091)

Note. *SE* = standard error; *SD* = standard deviation. All regression coefficients are standardized.
* $p < .05$. ** $p < .01$. *** $p < .001$.

(Goldhammer et al., 2021b, p. 872)

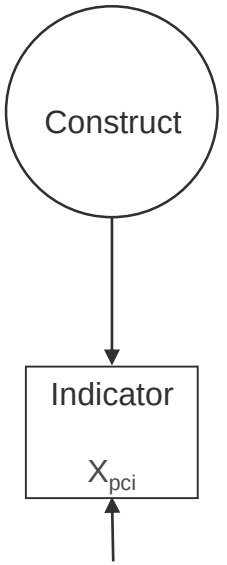
Overview

- Log data in LSAs
- Individual differences in response processes
- Benefits of using log data
- Challenges of using log data
- Conclusions

What can log data be used for?

- Manifold of uses across the assessment cycle (e.g., Maddox, 2023)

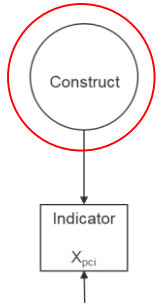
- Goldhammer et al. (2020): Evidence-centred design (ECD) framework (Mislevy et al. 2003) to classify the potential uses of log file data
  - Student, evidence, assembly, task




```
<taoEvent
Name="stimulus"
Type="TEXTLINK"
Time="164959">id=u10a
_default_txt
15|*$href=unit10page1
4|*$target=_self</tao
Event>
```

unsplash

Student model




```
<tacEvent
Name="stimulus"
Type="TEXTLINK"
Time="164959">id=u10a
_default_txt
15|*$href=unit10page1
4|*$target=_self</tao
Event>
```


- (Latent) **constructs** representing **attributes of the work process**
  - **Continuous** latent variables
    - (Domain-specific) speed (e.g., van der Linden, 2007)
    - Propensity to use a certain solution strategy (Greiff et al. 2016)
    - Exploration in complex problem solving (Eichmann et al. 2020)
  - **Categorical** latent variables (solution types)
    - Problem solving solution patterns (e.g., Zhang & Andersson, 2023)
    - Digital reading patterns (e.g., Hahnel et al., 2022)

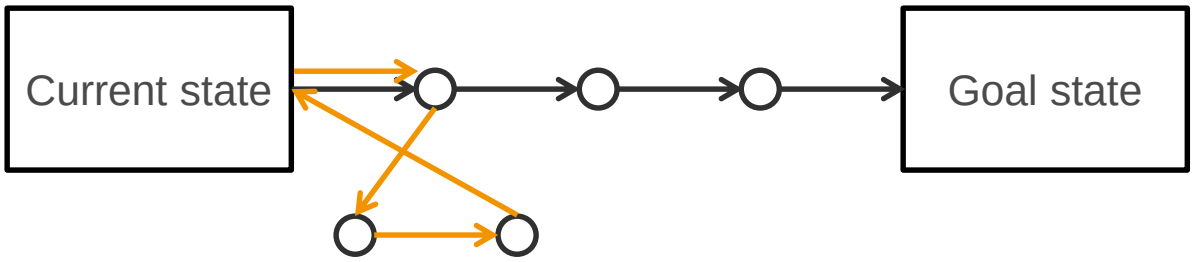
Example: Exploration in complex problem solving

- Eichmann et al. (2020)
  - **Group differences** (e.g., boys vs. girls) are regularly found in international large-scale assessments.
  - Underlying mechanisms of these differences are unclear.
  - Question: Can gender-specific differences in performance in complex problem solving (CPS) be explained by **different response processes**?

Exploration in CPS

- **Complex problems**: not all necessary information is given, has to be generated

- **Exploration** = interactions that do not (directly) contribute to problem solving, but serve to gain information



Exploration in CPS

- Eichmann et al. (2020)

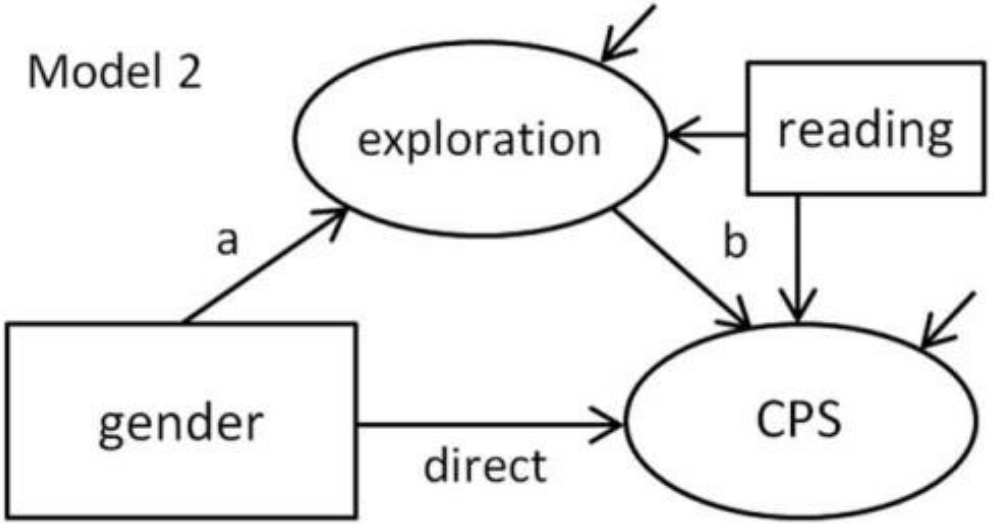
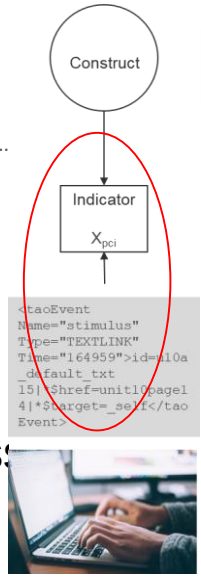


Table 2

Aggregated Model Estimates and Effect Sizes

Parameter	Estimate	*SE*	*z*	*p*	τ	*Q*(41)	*p*
Model 2							
a	−.57	.02	−27.14	<.001	.09	71.88	.002
b	.44	.02	30.32	<.001	.03	43.76	.355
Total	−.28	.02	−14.44	<.001	.09	98.84	.002
Direct	−.03	.02	−1.73	.083	.05	45.44	.292
Indirect	−.23	.01	−19.68	<.001	.03	43.08	.382
κ^2	.17						

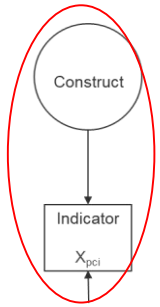
Evidence model – Evidence rules

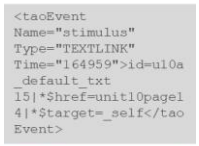


- **Deriving process indicators** representing an attribute of the work process
  - e.g., response time tapping test-taking engagement (PIAAC: Goldhammer et al., 2016)

- **Enhancing traditional product indicators**
  - (partial credit) scoring, depending on interactions (e.g., problem solving in PISA 2012; OECD 2013a)
  - coding of missing responses (e.g., responses in PIAAC without interaction and time on task less than 5 s were coded as ‘Not reached/not attempted’; OECD 2013b)
  - detecting aberrant response behavior (van der Linden & Guo 2008), data fabrication (Yamamoto & Lennon 2018)

Evidence model – Evidence synthesis/ Measurement model (1)

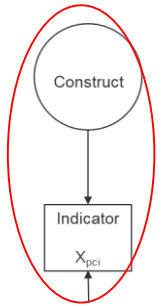




- **Multiple process indicators** identify a process-related construct (e.g. planning, speed, test-taking engagement) (e.g., Levy, 2020)

- **Joint modeling** of process data with product data
  - Challenge: fully capturing the dependency structure of process (and product) indicators within and between items
  - Examples
    - Increasing measurement precision (e.g., Bolsinova & Tijmstra, 2018)
    - Modelling missing data mechanisms (Pohl et al., 2019)

Evidence model – Evidence synthesis/ Measurement model (2)

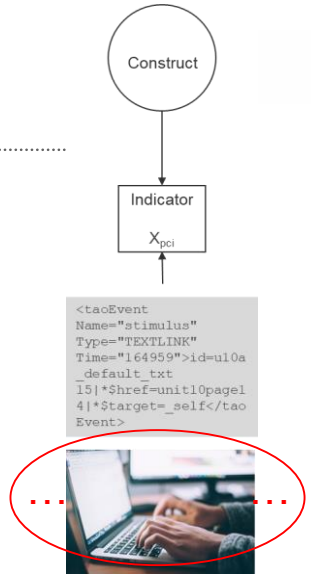



```
<tacEvent
Name="stimulus"
Type="TEXTLINK"
Time="164959">id=u10a
_default_txt
15|*$href=unit10page1
4|*$target=_self</tao
Event>
```


- Joint modelling for **model-based treatment of disengaged responding**
  - Joint (mixture) modeling of ability, speed, and engagement (Ulitzsch et al., 2020)
  - Joint modeling of ability, rapid guessing propensity, and the likelihood of correct response (Deribo et al., 2021)

- **Validating the interpretation of test scores** (Boorsboom et al., 2004; Embretson, 2023; Ercikan & Pellegrino, 2017)
  - Testing hypotheses on whether construct-related attributes of the work process predict the task outcome as expected

Assembly Model



- Adaptive testing: timing information to **improve item selection** and thereby obtain a more efficient measurement (van der Linden, 2008)

- Timing information to **control the speededness** of test forms in adaptive testing (van der Linden, 2005) and fixed form linear testing (Becker et al., 2023)

- Process data can be used for **triggering interventions** if the response behavior is aberrant, i.e., feedback to the
  - individual test taker via prompts so that the test taker can adapt
  - proctor via a dashboard, so that the proctor can intervene (Wise et al. 2019)

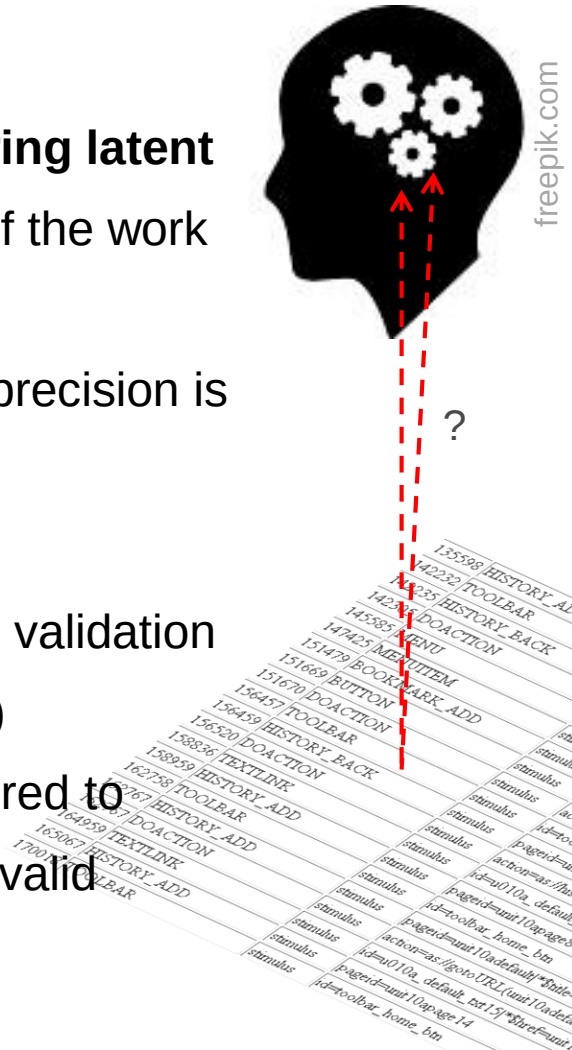
Overview

- Log data in LSAs
- Individual differences in response processes
- Benefits of using log data
- Challenges of using log data
- Conclusions

Validating the interpretation of process indicators

- Many of the uses of process data imply **inferring latent (e.g., cognitive or motivational) attributes** of the work process from log data
  - (but not all, e.g., increasing measurement precision is simply about exploiting empirical relations)

- These inferences need to be **justified** through validation (Goldhammer et al., 2021; Zumbo et al., 2023)
  - Theoretical and empirical evidence is required to ensure that the respective interpretation is valid



Argument-based approach of validation

- "[…] **validity** refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests. […] **Validation** can be viewed as a process of constructing and evaluating arguments for and against the intended interpretation of test scores and their relevance to the proposed use"
  (AERA et al., 2014, p. 4; see also Messick, 1989; Kane, 2013).

- These concepts of validity and validation apply to **any indicator-based inferences**, regardless of whether product/correctness or process indicators are used (Goldhammer et al., 2021).

Explanation inference/Construct interpretation

- Individual differences in the process indicators are (causally) determined by differences in the **(theoretical) construct** which the indicator is intended to measure
- **Threats to the construct interpretation**: Construct-irrelevant variance, construct underrepresentation

- Sources of **validity evidence** → empirical support for theory-based predictions about relationships between observable variables
- Following Embretson (1983): Relation of construct-related
  - **item properties** to process indicator (construct representation)
  - **person variables** to process indicator (nomothetic span)

Example for validating the construct interpretation: Sourcing indicator

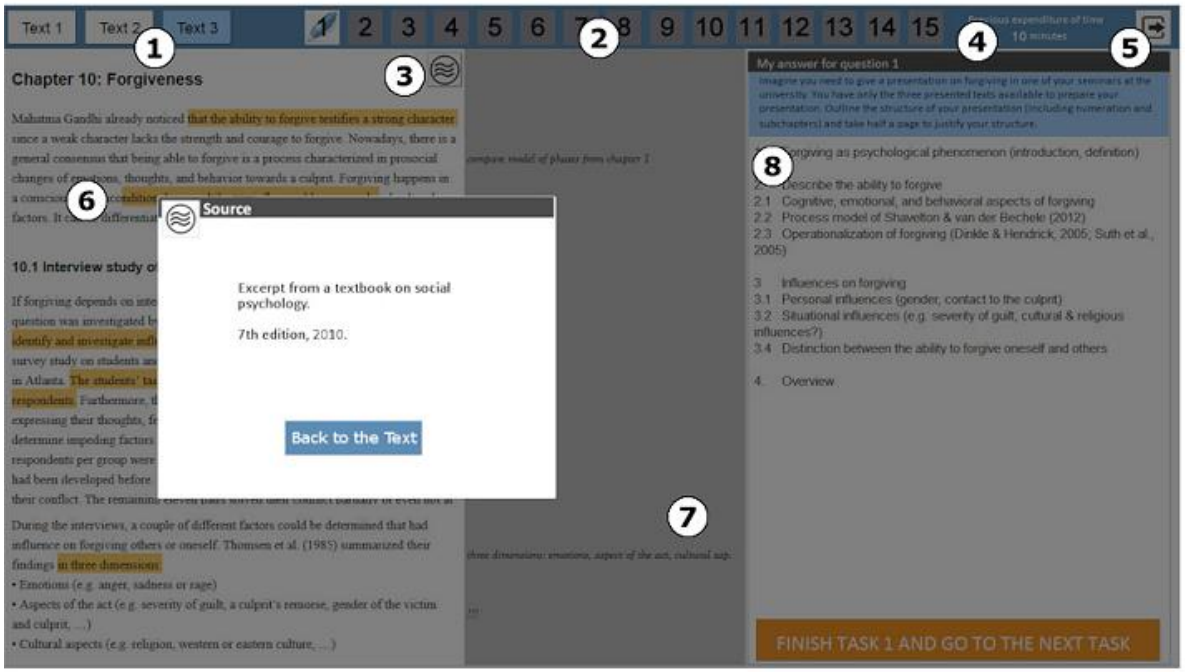
- **Multiple document comprehension (MDC)**: reader's competence in constructing an integrated representation of a certain topic using textual information from different sources

- **MDC test** was designed to infer *sourcing* as an attribute of the work process

- **Sourcing** is defined as the reader's consideration of the origin and intention of a document → Is this interpretation of the sourcing indicator justifiable?

Task model for sourcing

- Designing the **activity space** within MDC items so that sourcing can be linked to observed behavior: Access to source requires button click



(from Hahnel et al., 2019)

Evidence model: Indicators for sourcing

- Sourcing ≠ Sourcing → **Contextualization** of 'Source button' click event needed

Table 1. Overview over the process variables

Purpose	Process description	Operationalization of the process variable
(1) Proactive sourcing	Source information is accessed before a document is read	Dichotomous indicator of whether the source was accessed within the first 10% of the document processing time[a]
(2) Repeated sourcing	Source information is visited multiple times	Dichotomous indicator of whether the source was accessed multiple times in the reconstructed test-taking sequence
(3) Task-related sourcing	Source information is accessed after item instruction	Dichotomous indicator of whether the state-trigram 'item–document–source' occurred, combined with a maximal duration of 10 s on the document[b]
General sourcing	Source information is accessed	Dichotomous indicator of whether the source of a document was accessed

(from Hahnel et al., 2019)

Argument-based validation

- **Interpretation**: Repeated sourcing to update memory traces for strengthening connections or to help resolve conflicts across multiple documents

- **Testable assumptions** (see Hahnel et al., 2019)
  - **Person** level: Repeated sourcing is positively associated with MDC, but not with final school grades after controlling for MDC
  - **Item/Unit** level: The number of documents, number of conflicts between documents, and number of items that require comprehending source information should induce more repeated sourcing

- **Evidence**: Empirical relation of process indicators to the MDC score, to other measures (nomothetic span), and to task characteristics (construct representation).

Validity evidence

Table 3. Results of the explanatory models

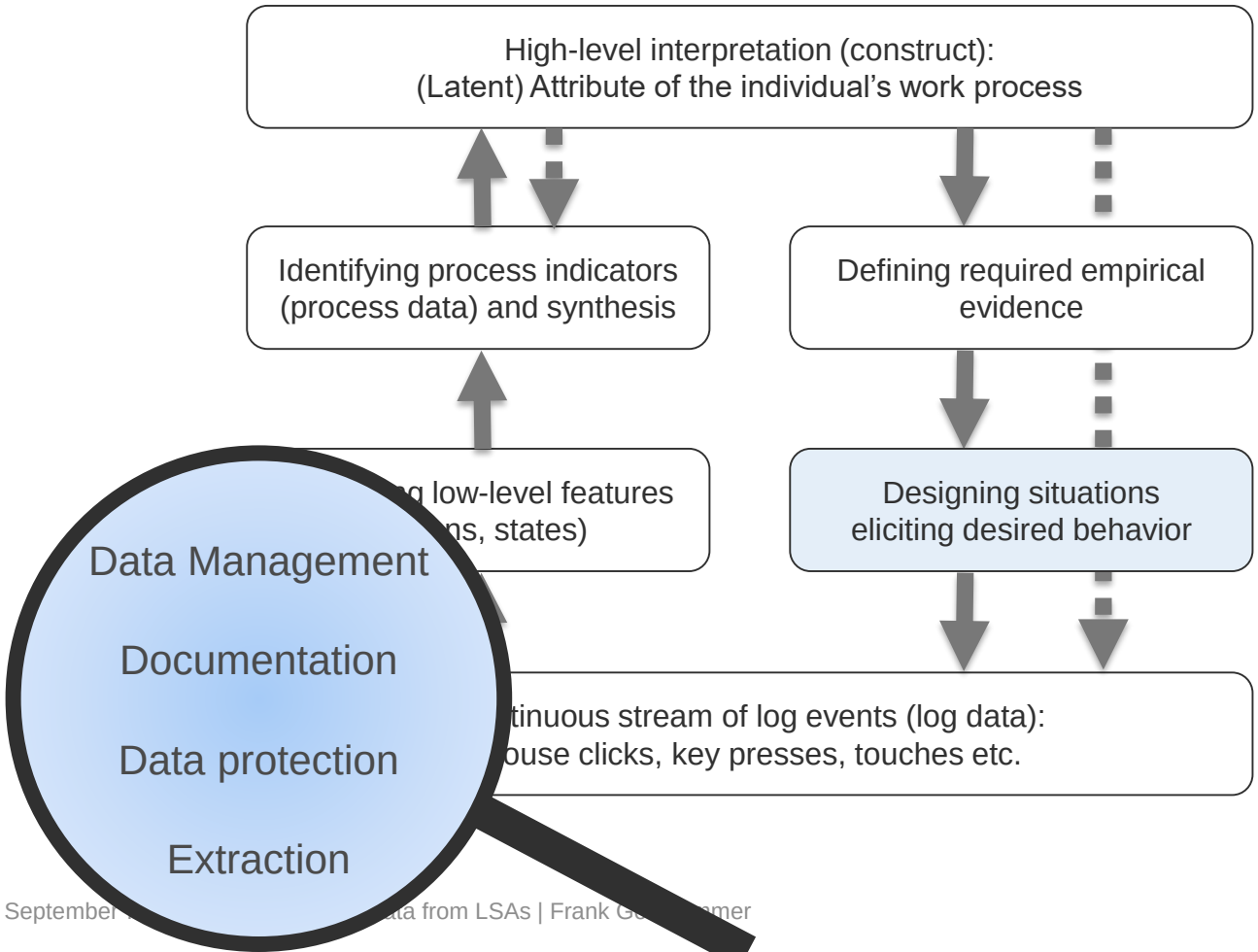
	Repeated sourcing
Intercept	−2.40 (0.31)***
Unit difficulty	0.33 (0.11)**
Person characteristics	
MDC score	0.53 (0.14)***
Graduation grade	−0.09 (0.14)
Unit characteristics	
N documents	1.56 (0.59)**
N conflicts	0.91 (0.41)*
N source-related items	0.10 (0.13)
Properties of test administration	
Position 2	0.66 (0.14)***
Position 3	0.73 (0.14)***
Document 2	−0.16 (0.13)
Document 3	−0.25 (0.15)

(from Hahnel et al., 2019)

Dependent variable: Binary indicator of 'Repeated sourcing' (unit level) with

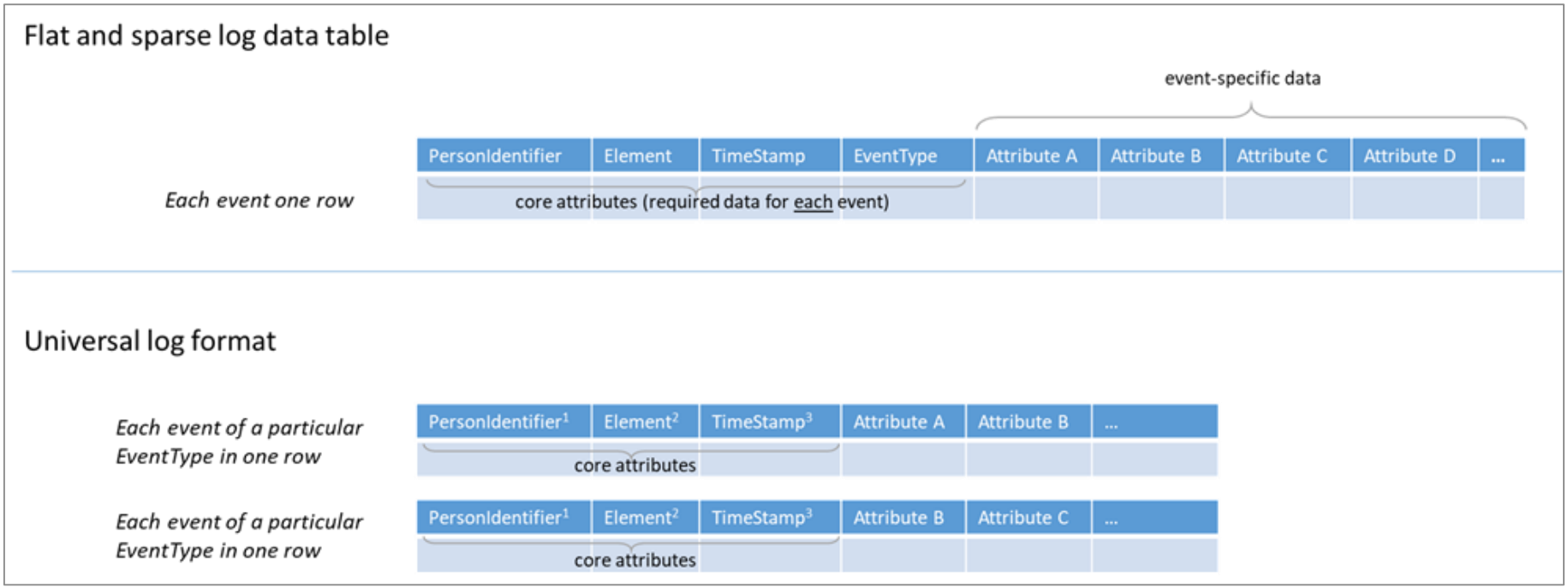
- 0: source was not accessed or only once
- 1: source was accessed multiple times

Dissemination of log data



Data management/pre-processing (1)

- **Transformation** of raw log files (e.g., json, XML) stored by the assessment system typically by case to **data sets**
  - **Data formats** (see Kroehne et al., in prep)



Data management/pre-processing (2)

- Raw log data set may contain **complex event attributes** with strings (e.g., fragments of JSON, XML) that need to be parsed before the information can be accessed and finally analyzed (transformation to atomic attributes)

- **Checks** for correctness and completeness (e.g., Kroehne & Goldhammer, 2018)
  - Data is syntactically valid and it conforms to the schema definition (e.g., all pieces are stored as expected)
  - Data is plausible given item and test design (e.g., values of attributes, sequence of events)

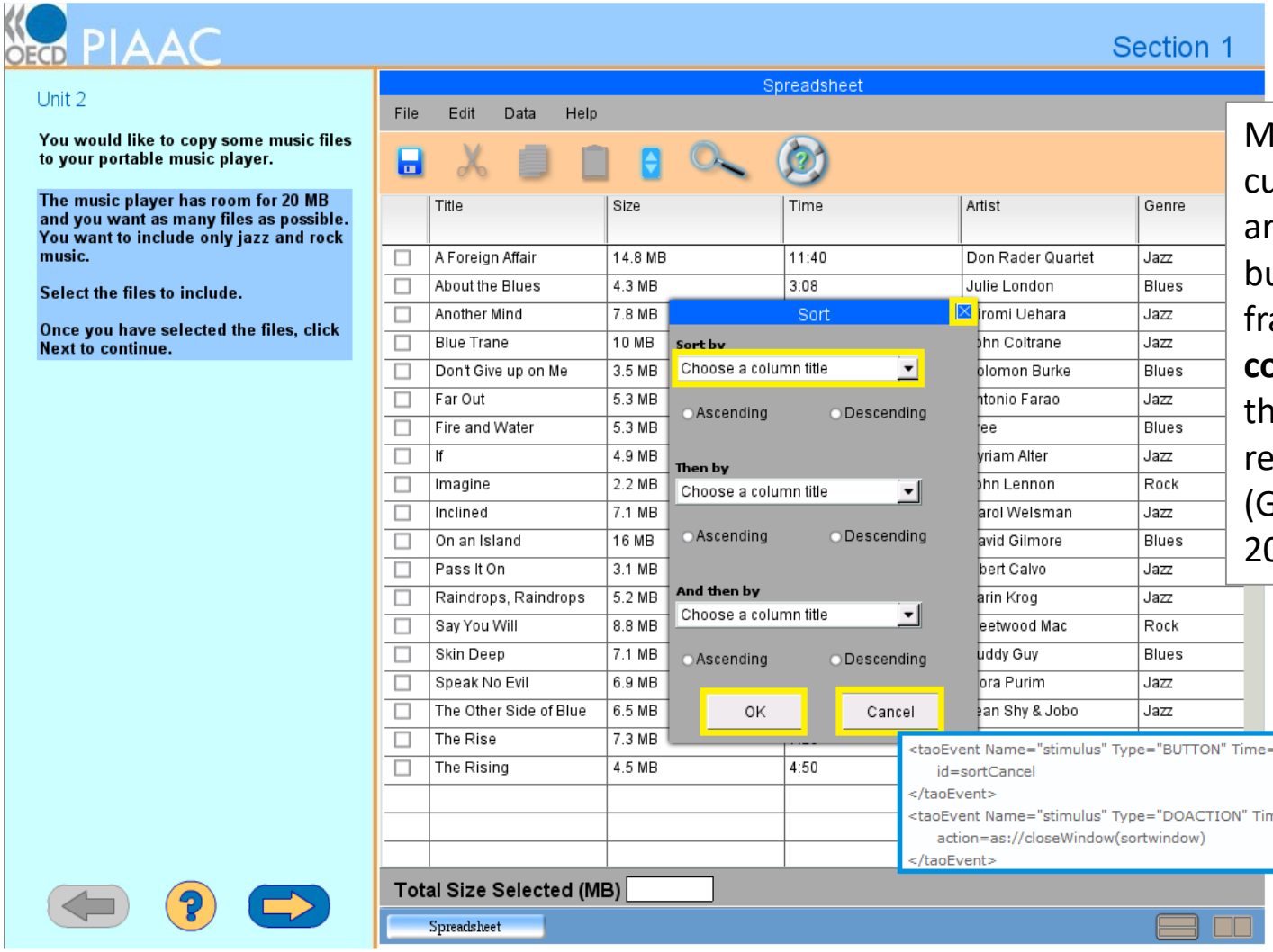
- **Cleaning**

Documentation of log data and items

- To know the **meaning of event types** and related **event-specific data**
- To understand which **log events** are triggered by which **user interactions** within a given item
- To be able to **reproduce** research work (Open Science principle)
- However, **test security** needs to be maintained

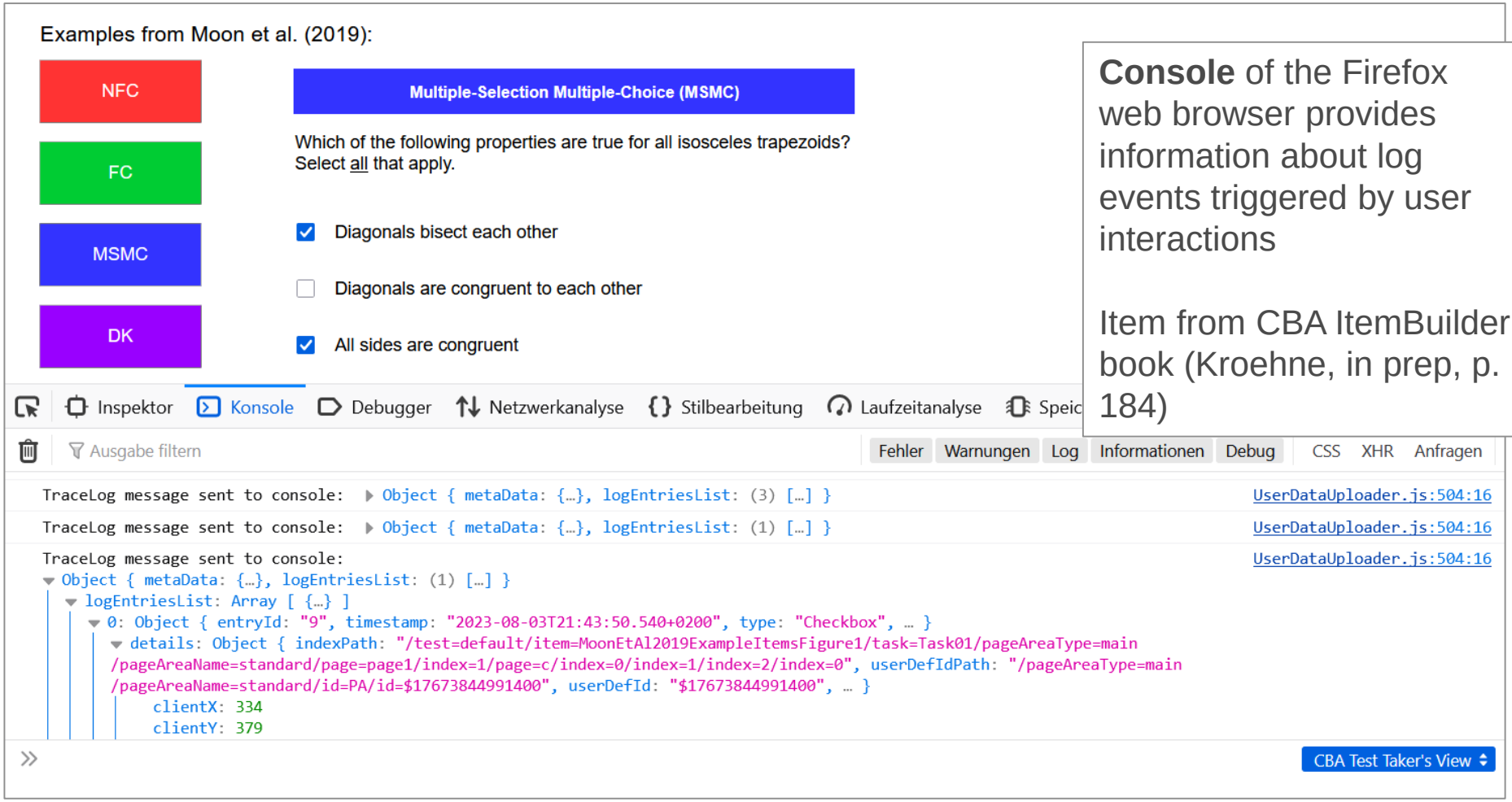
- **Documentation formats** (see Kroehne et al., in prep)
  - Written documentation presenting items and description of event types
  - Showing the mapping of events to user interactions within the item
    - Annotated screenshots
    - Annotated screencasts

Annotated images: PIAAC 2012 log data



Moving the mouse cursor over sensitive areas (here the Cancel button) displays blue-framed **pop-up dialogs containing details** about the structure of the recorded events. (Goldhammer et al., 2020, p. 257)

Annotated screencast: CBA ItemBuilder item



Console of the Firefox web browser provides information about log events triggered by user interactions

Item from CBA ItemBuilder book (Kroehne, in prep, p. 184)

Data protection and anonymization

- To adhere to **data protection rules** (e.g., GDPR) preventing the conclusion on a specific person (i.e., the data provider)
- To gain trust and acceptance

- **Critical information** included in log data:
  - Free text responses → removing text, replacing text completely or selectively
    - e.g., in PIAAC 2012 all raw log files (XML) were anonymized by replacing entered text with neutral character strings
  - Date and time → relative time stamps
  - User IDs → replaced, scrambled
  - …

Overview

- Log data in LSAs
- Individual differences in response processes
- Benefits of using log data
- Challenges of using log data
- Conclusions

Conclusions

- Log/process data is a new data source to **learn more about the response process** - as far as relevant behavior can be elicited by the task (phases of behavioral inactivity)
- Using log/process data for assessment purposes should be understood as **reasoning from evidence** to make a certain claim
- As a consequence, the same **quality standards** need to be applied as in traditional assessments (e.g., validity evidence)
- **Theories** are of great importance for task design, evidence identification, and validation
  - Lack of theory or process models relating behavioral low-level features to attributes of the work process through evidence identification and accumulation
- Lack of standards and **best practices for the dissemination** of log data from LSAs

Community work on process data to address challenges

- International **„Beyond Results“ Workshop** initiated by IEA/DIPF/ZIB
  - Goal: Exchange on conceptual, methodological and operational issues concerning process data
    - 2020: Paving the way for the use of process data
    - 2021: From log data to valid inferences
  - Rich online documentation https://beyond-results.com/

- Spinoff: International **Working Group on Process Data** by FLIP+/IEA/DIPF
  - Short Online meetings, 1.5 hours, multiple times per year
  - Last meeting March 2023 on the standardization of log data

Thank you! – Questions, comments…?

contact: f.goldhammer@dipf.de

TBA Centre for Technology Based Assessment