Leibniz Institute for Research and Information in Education

ZENTRUM FÜR INTERNATIONALE BILDUNGSVERGLEICHSSTUDIEN

Some considerations in validating the interpretation of process indicators

Frank Goldhammer[1,2], Carolin Hahnel[1,2], Ulf Kroehne[1], Fabian Zehner[1]

Centre for Technology Based Assessment

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

[2]Centre for International Student Assessment (ZIB)

Overview

- Introduction
- Kinds of assessment
- ECD view on continuous assessment within items
- Argument-based validation
- Example 1: Test-taking engagement
- Example 2: Sourcing in reading
- Concluding remarks

Overview

- Introduction
- Kinds of assessment
- ECD view on continuous assessment within items
- Argument-based validation
- Example 1: Test-taking engagement
- Example 2: Sourcing in reading
- Concluding remarks

Interpretation of process indicators in testing

(Latent) Attribute of the work process

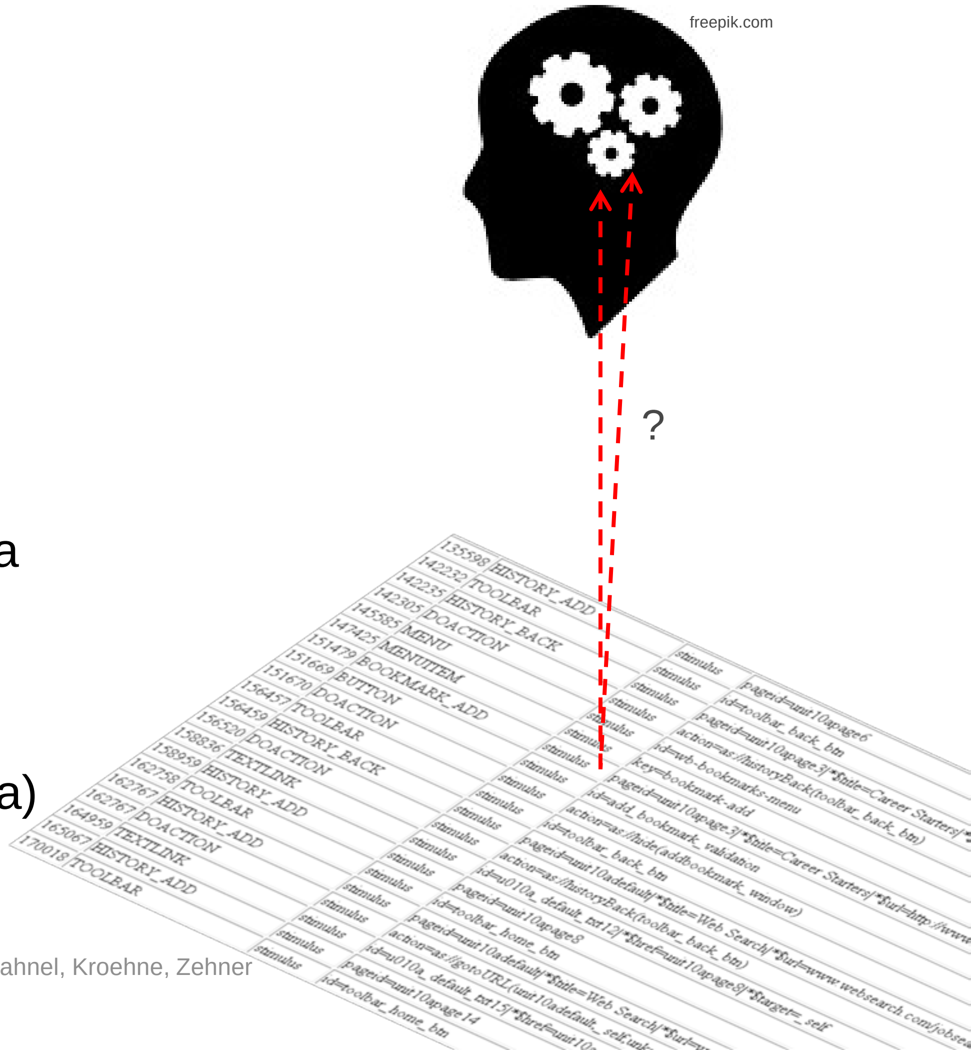
(e.g., solution strategy, engagement)

Process indicators

Features or states identified by log data

Continuous stream of log events

representing user actions (process data)



Validating the interpretation of process indicators

- Inferring latent (e.g., cognitive) attributes from process data (e.g., log data) needs to be justifiable.

  Both **theoretical** and **empirical evidence** is required to make sure that the reasoning from the process indicator to the attribute is **valid**.

  (Goldhammer & Zehner, 2017)

- This follows the concept of **validation** that is well known from the interpretation and use of **test scores**: „Validation can be viewed as a process of constructing and evaluating arguments for and against the intended interpretation [..]"

  (AERA, APA, NCME, & Joint Committee on Standards for Educational Psychological Testing, 2014, p. 4; see also Messick, 1989)

Process indicators

- Process indicators can be conceptually framed using the **Evidence Centered Design (ECD)** framework (Mislevy, Almond, & Lukas, 2003)
  - Flexible framework applicable to various **kinds of ‘assessment’**
  - Like product/correctness indicators, **process indicators** are the result of empirical evidence identification.
  - Incorporates the development of the **validity argument** into the design of the assessment

Overview

- Introduction
- Kinds of assessment
- ECD view on continuous assessment within items
- Argument-based validation
- Example 1: Test-taking engagement
- Example 2: Sourcing in reading
- Concluding remarks

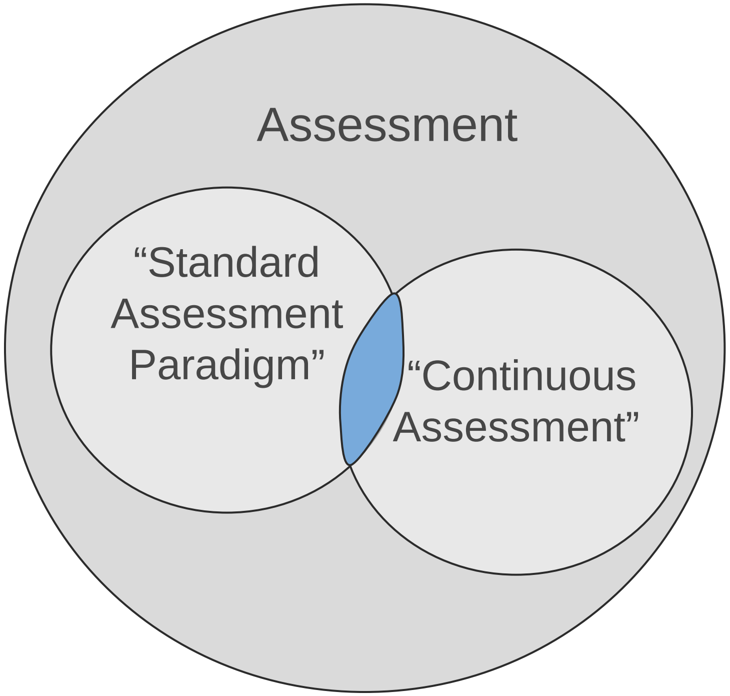
Kinds of assessment

- Definition of **Assessment**: „… collecting evidence designed to make an inference“ (Scalise, 2012, p. 134)
  - **Standard assessment paradigm** (Mislevy, Behrends, DiCerbo, & Levy, 2012)
    - e.g., competence test, questionnaire
    - Pre-defined, pre-packaged items; discrete responses (item-by-item); evidence based on final work product
  - **Continuous/ongoing assessment approach** (Mislevy et al., 2012; DiCerbo, Shute, & Kim, 2017; Shute, 2011)
    - e.g., game-based assessment, simulation-based assessment
    - Predefined activity space; continuous performance; evidence about the work process is gathered over time (continuous feature extraction)

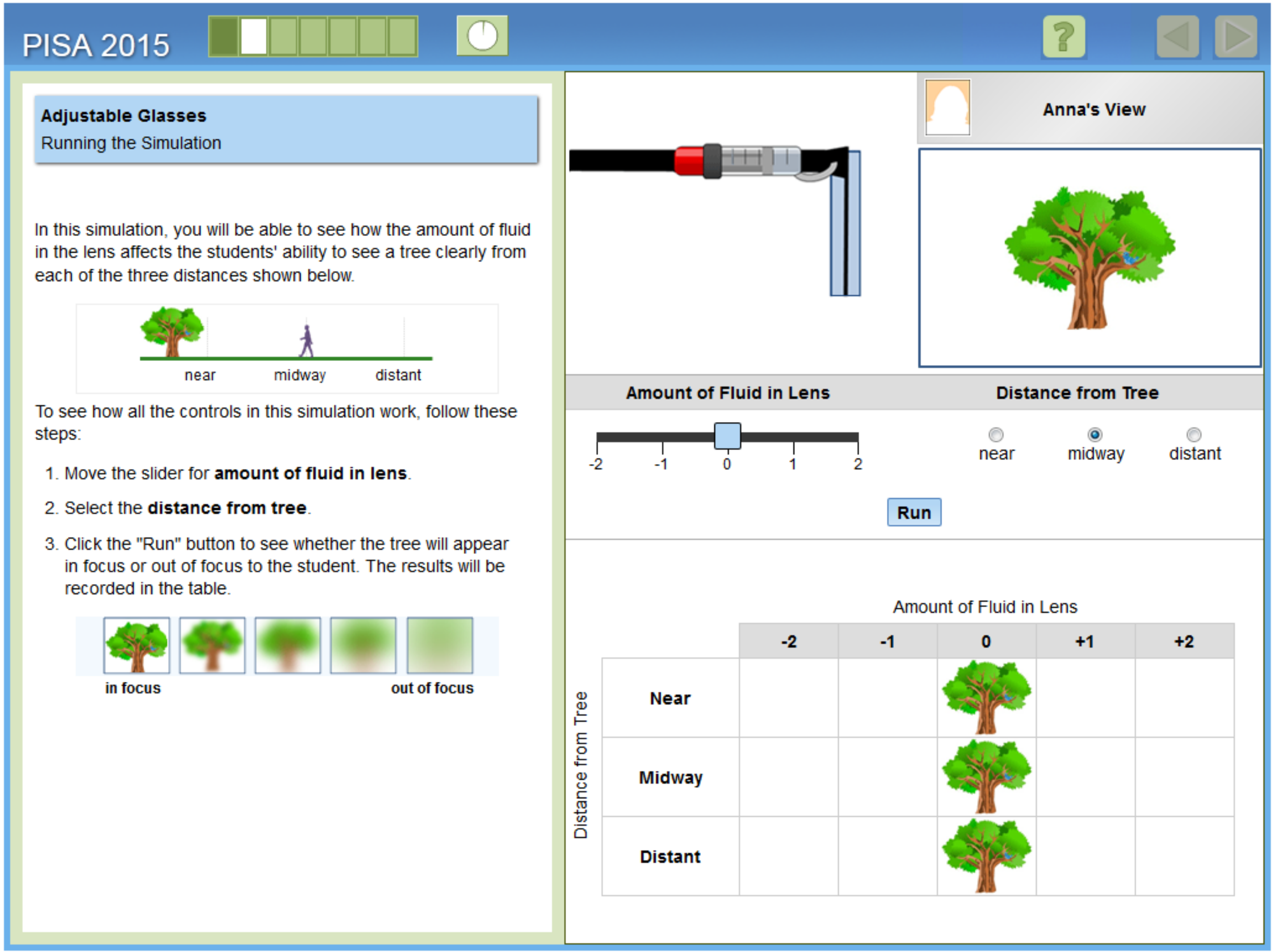
Overlap: Continuous assessment within items

- e.g., competence test including complex, interactive, simulation-based items

- Pre-defined items
- Continuous performance within items
- Within items evidence can be gathered over time (evidence on work process)
- Unobtrusive feature extraction within items
- Features can be included into rules for product indicator
- Data are rich (at individual level) and fine-grained within items



Continuous assessment within items: PISA Sciene item with simulation



Example for **claim**: (Procedural) Knowledge about experimental strategies for inferring rules

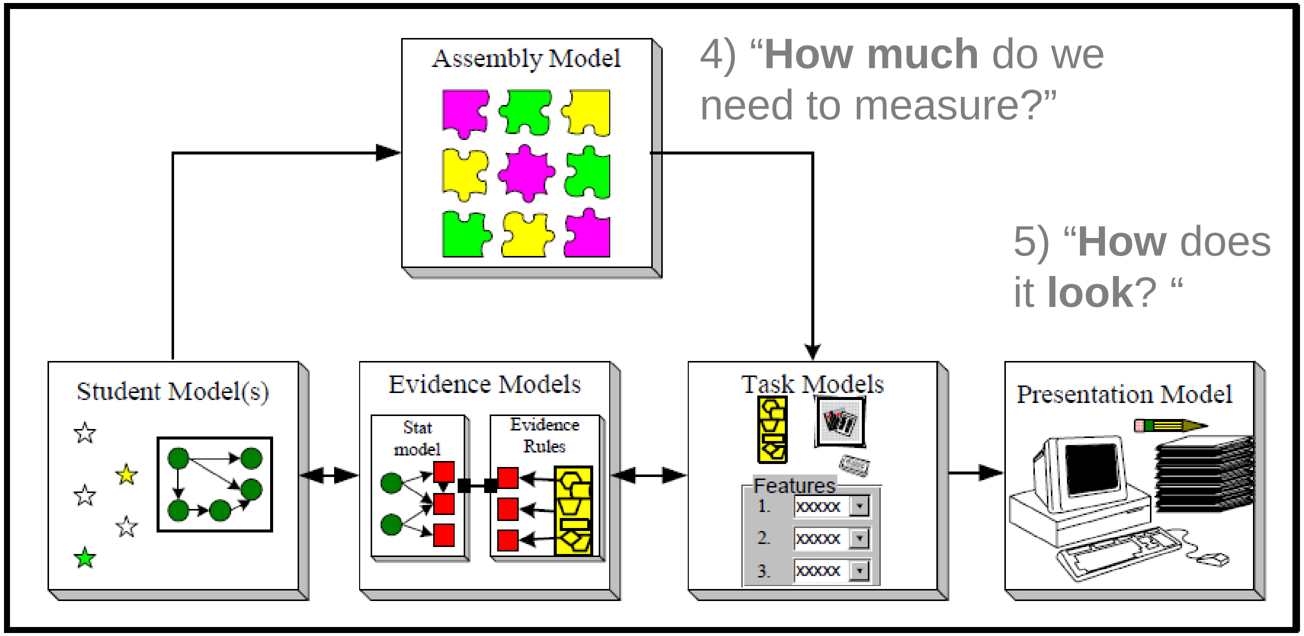
Overview

- Introduction
- Kinds of assessment
- ECD view on continuous assessment within items
- Argument-based validation
- Example 1: Test-taking engagement
- Example 2: Sourcing in reading
- Concluding remarks

Evidence centered design view on continuous assessment within items

- Mislevy, Almond, & Lukas (2003, p.5): Conceptual Assessment Framework

1) "**What** are we measuring?"



4) "**How much** do we need to measure?"

5) "**How** does it **look**? "

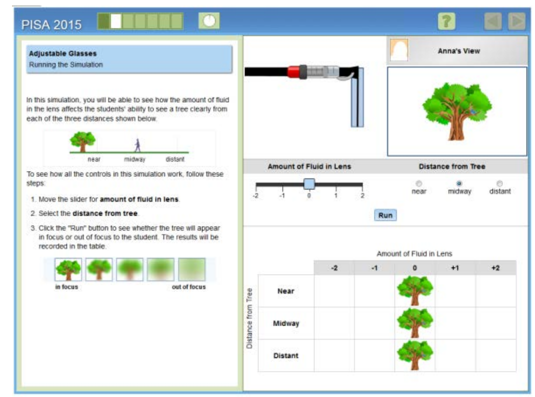
2) "**How** do we measure it?"

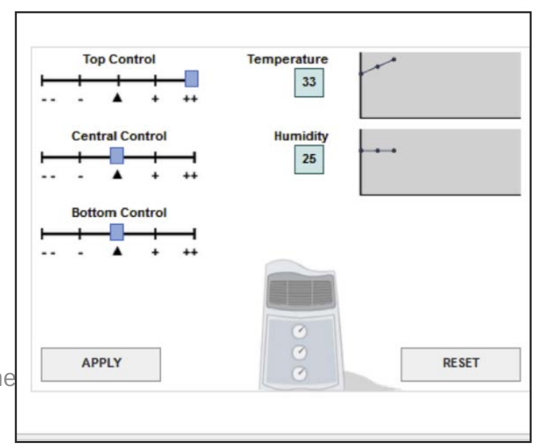
3) "**Where** do we measure it?"

Continuous assessment within items – Student model

- What are the **claims** to be made on **knowledge, skills,** and **attributes**?

- Examples for an attribute of the work process:
  - PISA Science: (Procedural) Knowledge about experimental strategies for inferring rules
  - PISA CPS: Planning, allocation of cognitive ressources etc.
    (Eichmann, Goldhammer, Greiff, Pucite, & Naumann, 2019; Greiff, Niepel, Scherer, & Martin 2016)





Continuous assessment within items – Task/Activity model (1)

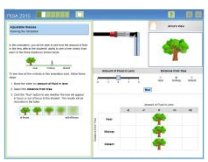
- How to **design situations** to obtain the **evidence** needed for inferences about the targeted construct?
- From item to activity design (adapted from Behrens & DiCerbo, 2013)

“scoring” inference

	Standard assessment: Items…	Continuous assessment: Activities…
Problem formulation	… pose questions	… request/invite actions
Output	… have answers	… have features (states)
Interpretation	… indicate ability construct (product indicator)	… indicate attributes (process indicators)
Information	… provide focused information	... provide multi-dimensional information

Continuous assessment within items – Task/Activity model (2)

- For a **valid interpretation** of indicators we need a careful and clear definition of how the targeted **attribute**, empirical **evidence** (behavioral states or features) and **situations** that can evoke the desired behavior (actions) are linked.
  - **Task design** (e.g., Goldhammer & Zehner, 2017)

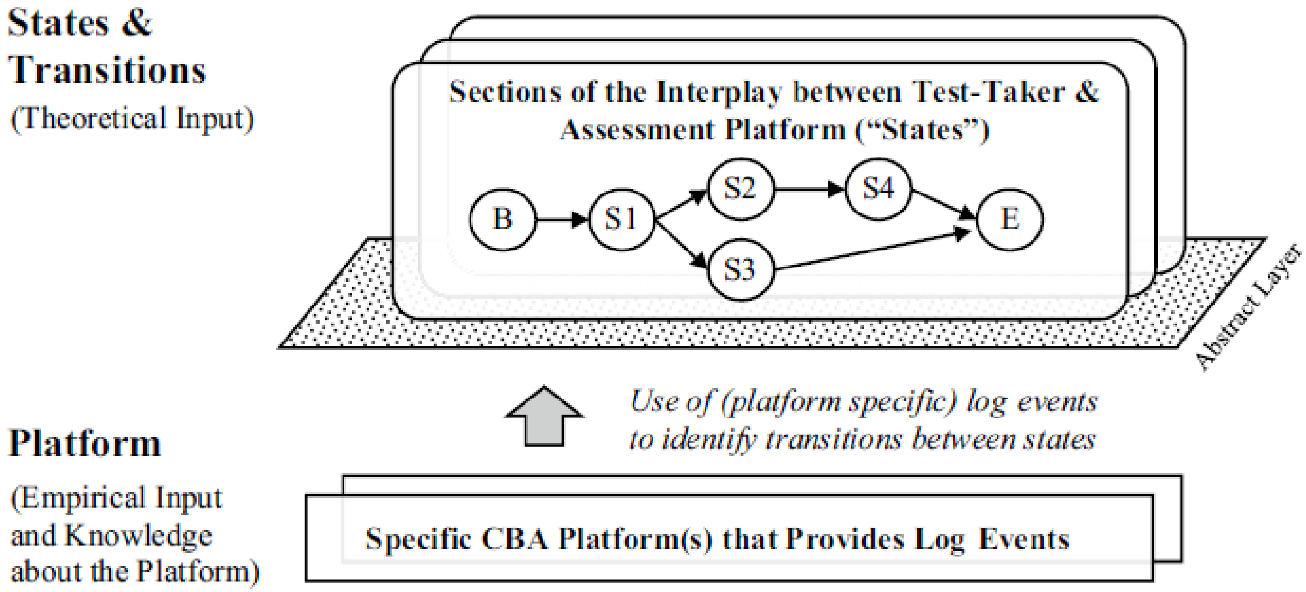


    - Designing the activity space so that attributes of the work process can be clearly linked to behavioral actions (e.g., clicking, highlighting, etc.)
    - Observable attribute vs. latent constructs
  - **System design** (Kroehne & Goldhammer, 2018)

    - Storage of user (and system) events being complete and correct
    - Granularity depends on features/states to be identified by user actions

Continuous assessment within items – Task/Activity model (3)

- Designing the activity space within items as **states and transitions** of a finite state machine (Kroehne & Goldhammer, 2018; Mislevy, et al. 2014)



(from Kroehne & Goldhammer, 2018)

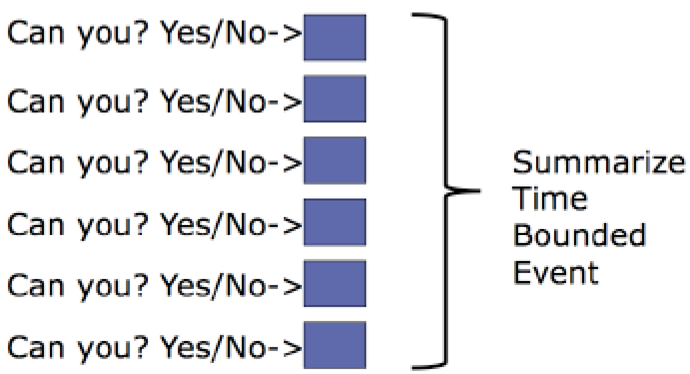
Continuous assessment within items – Task/Activity model (4)

- **Representative sampling** of observed performances from a universe of possible observations is needed (generalization inference) (see Kane, 2013)
  - Representative sampling of items (e.g., context, structure, complexity)
  - For items with rich simulations encountered **situations** might differ between individuals constraining the sampling (see game-based assessment)
    - Identification of salient features in recurring situations (Mislevy et al., 2012)
    - Introduction of rescue/convergence points aligning situations (e.g., Collaborative PS assessment in PISA 2015)

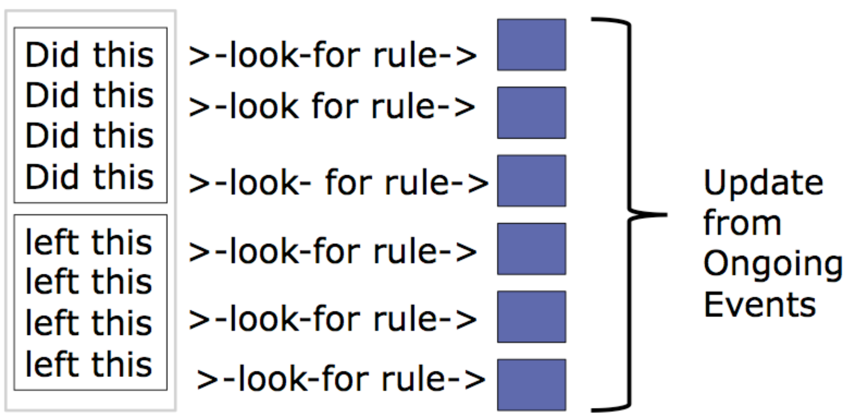
Continuous assessment within items – Evidence model (1)

- Evidence identification rules (figures from Behrens & DiCerbo, 2014, p.13)

Item: Scoring responses



Activity: Identifying presence/absence of features (states) in a stream of actions, interpretation as indicator



e.g., manipulation of "Amount of fluid in the lense" controller without manipulating "Distance" → interpretation: application of experimental strategy

Continuous assessment within items – Evidence model (2)

- Features/states serving as empirical evidence are defined by **actions given a particular context**
  - Same action(s) might indicate different states, e.g., the meaning of pressing a button may depend on the test-taker's past/current situation
  - Rules for evidence identification need to consider the context of observed actions
- If the process indicator taps a **theoretical construct** the theory should inform about the evidence needed and the kind of identification rule that would be appropriate.

Overview

- Introduction
- Kinds of assessment
- ECD view on continuous assessment within items
- Argument-based validation
- Example 1: Test-taking engagement
- Example 2: Sourcing in reading
- Concluding remarks

DIPF

Argument-based approach of validation

- **Validation**: Process of developing and evaluating arguments speaking for/against a certain interpretation and use of an indicator (Kane, 2013)
  - Specifying the interpretation/use; explicating related assumptions and the reasoning from performance to the intended conclusion
  - Evaluation of the argument
- **Central inferences** when interpretating indicators (Kane, 2001, 2013)
  - Scoring/evidence identification → indicator represents observed performance features appropriately
  - Generalization → similar performance is expected in similar tasks, contexts, etc.
  - **Explanation** → indicators are explained by a (theoretical) construct
  - Extrapolation
  - Decision making

Sources of evidence: Construct representation

- „*Construct representation* is concerned with identifying the theoretical mechanisms that underlie item responses, such as information processes, strategies, and knowledge stores.“ (Embretson, 1983, p. 179)

- Application to **process indicators** tapping an attribute of the work process
  - Determine task characteristics that theoretically evoke the targeted attribute
  - Relate these task characteristics to item process indicators
  - **If** items with these task characteristics are also more likely to elicit the respective actions, **then** the process indicator can be interpreted as determined by the respective attribute
  - Statistical modelling: lltm+e (Janssen, Schepers, & Peres, 2004)

Sources of evidence: Nomothetic span (1)

- „*Nomothetic span* is concerned with the network of relationships of a test score with other variables. " (Embretson, 1983, p. 179)

- **Other measures**: Same/similar construct (convergent evidence), different construct (discriminant evidence)
  - Triangulation of process indicators from the same assessment: measures based on think aloud protocols, eye-tracking, screen capturing, …

- **Group variables**: Testing the effect of group membership that is (theoretically) related to attributes of the work process, e.g., experts vs. novices (e.g., DiCerbo, Frezzo, & Deng, 2011).

Sources of evidence: Nomothetic span (2)

- **Product/Correctness indicators**: If a cognitive process model or a conceptual rationale exists providing hypotheses about the relation between process indicators and product indicators, the assumed association can be tested (e.g., Lee & Jia, 2014).

- **Experimental variables**: Testing the effect of experimental factors, that are (theoretically) expected to influence attributes of the work process; thereby, the causal interpretation of process indicators can be supported.

Two examples

- Process indicator of **test-taking engagement**
  - Context: Quality assurance in LSA
  - Process indicator: generic (time on task)
  - Validation: Nomothetic span

  Goldhammer, F., Martens, Th., Christoph, G., & Lüdtke O. (2016). *Test-taking engagement in PIAAC.* OECD Education Working Papers, No. 133. Paris: OECD Publishing.

- Process indicator of **sourcing**
  - Context: Substantive research in the domain of reading
  - Process indicator: domain-specific and contextualized
  - Validation: Construct representation, nomothetic span

  Hahnel, C., Kroehne, U., Goldhammer, F., Schoor, C., Mahlow, N., & Artelt, C. (2019). Validating process variables of sourcing in an assessment of multiple document comprehension. *British Journal of Educational Psychology*. doi:10.1111/bjep.12278

Overview

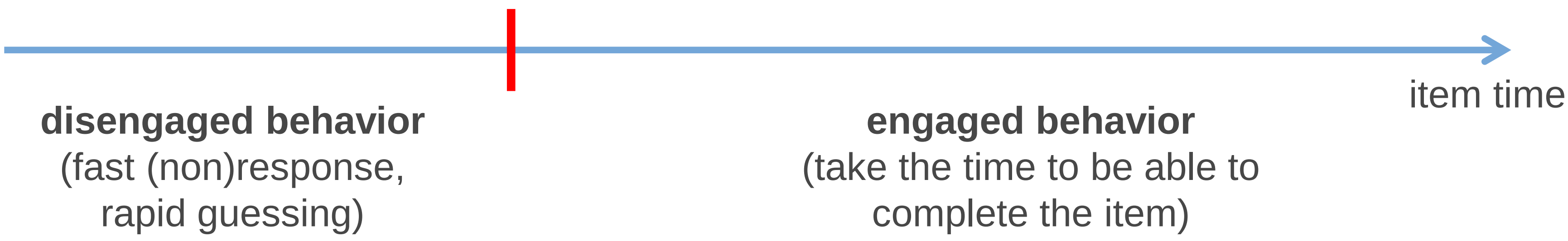
- Introduction
- Kinds of assessment
- ECD view on continuous assessment within items
- Argument-based validation
- Example 1: Test-taking engagement
- Example 2: Sourcing in reading
- Concluding remarks

Test-taking engagement

- Low test-taking engagement: Test-takers do not make an effort to show what they know and can do but respond quickly and arbitrarily (e.g., Wise & DeMars, 2005)
- Negative consequences (cf. Haladyna & Downing, 2004; Kong, Wise, & Bhola, 2007)
  - Test scores may underestimate the true proficiency level
  - Introduction of construct-irrelevant variance
  - Affects the validity of inferences based on test scores
- What to do? – Defining **indicators low test-taking engagement** (and taking them into account in scoring and data analysis)

Evidence model: Indicators of test-taking disengagement

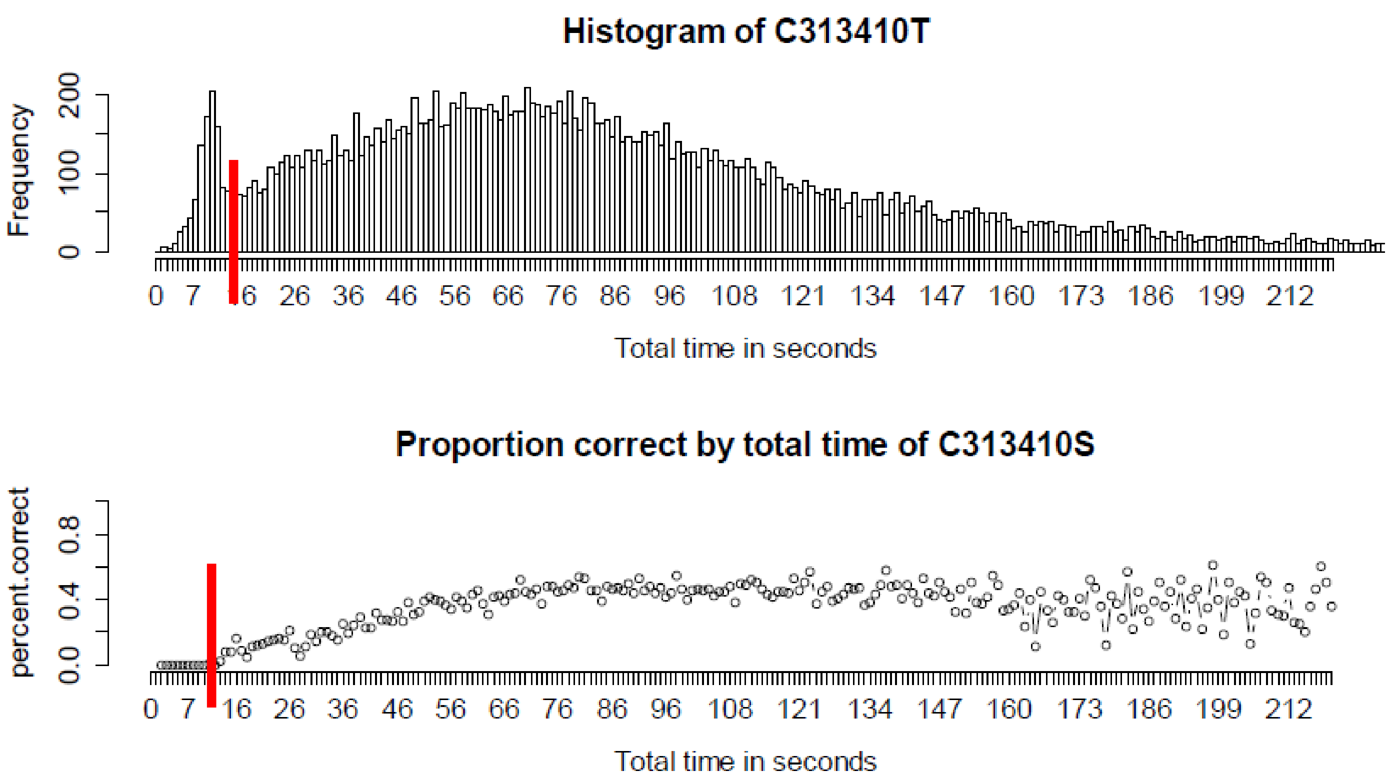
- Approach: Response time (RT) thresholds



- Constant RT thresholds
  - 5000 ms or
  - 3000 ms (Kong, Wise, and Bhola, 2007)
- Item-specific RT thresholds (e.g., Lee & Jia, 2014; Wise & Kong, 2005)
  - Visual inspection of response time distribution (VI method)
  - Proportion correct conditioning on response time (P+>0% method)

Evidence model: Item-specific RT thresholds

Figure 4. Response time distribution and proportion correct by response time



VI method

P+>0% method

(from Goldhammer, Martens, Christoph, & Lüdtke, 2016, p. 16)

Argument-based validation

- **Interpretation**: Test-taking disengagement
- **Testable assumptions** (see Lee & Jia, 2014)
  - Comparing proportion correct:
    - **Engaged** responding: probability to obtain a correct response is much higher than chance level (P+ >> 0)
    - **Disengaged** responding: probability to obtain a correct response is only at chance level (P+ =0)
  - Correlating score group (proficiency) and proportion correct (by item):
    - **Engaged** responding: positive relation
    - **Disengaged** responding: no relation
- **Evidence**: Empirical relation between process indicators and product indicators (nomothetic span).

Validity evidence (1)

- Comparing proportion correct

Table 2. Average proportion correct for engaged and disengaged response behavior.

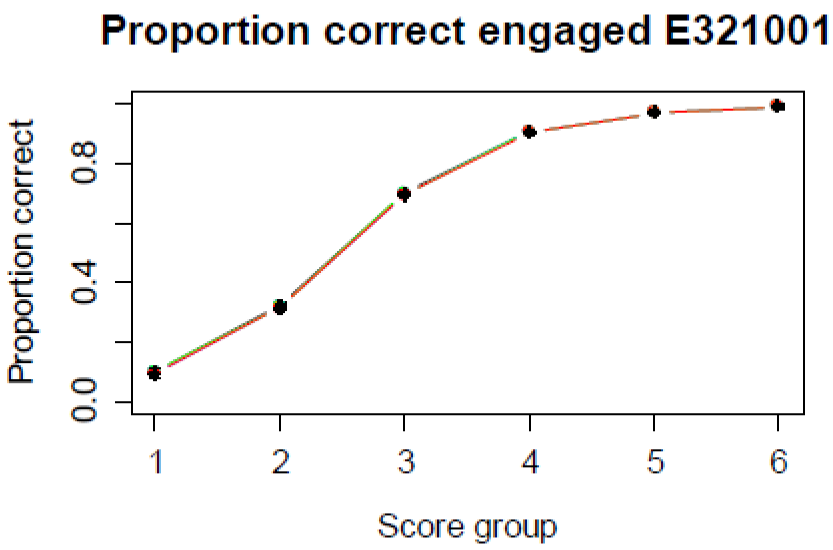
	Method	Proportion correct - Engaged	Proportion correct - Disengaged	Difference
Literacy	5000	0.55	0.02	0.53
	3000	0.55	0.01	0.54
	VI	0.56	0.02	0.54
	P+>0%	0.56	0.00	0.56
Numeracy	5000	0.64	0.09	0.55
	3000	0.63	0.04	0.59
	VI	0.63	0.07	0.56
	P+>0%	0.63	0.00	0.63
Problem solving	5000	0.40	0.00	0.40
	3000	0.40	0.00	0.40
	VI	0.43	0.01	0.42
	P+>0%	0.43	0.00	0.43

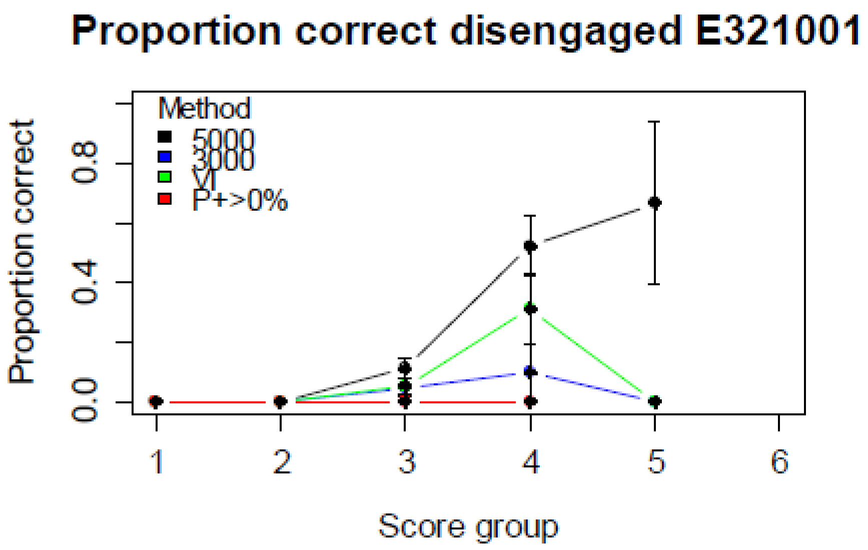
(from Goldhammer et al., 2016, p. 19)

Validity evidence (2)

- Correlating score group (proficiency) and proportion correct (by item)

  Sample item E321001 from Literacy





(from Goldhammer et al., 2016, p. 24)

Overview

- Introduction
- Kinds of assessment
- ECD view on continuous assessment within items
- Argument-based validation
- Example 1: Test-taking engagement
- Example 2: Sourcing in reading
- Concluding remarks

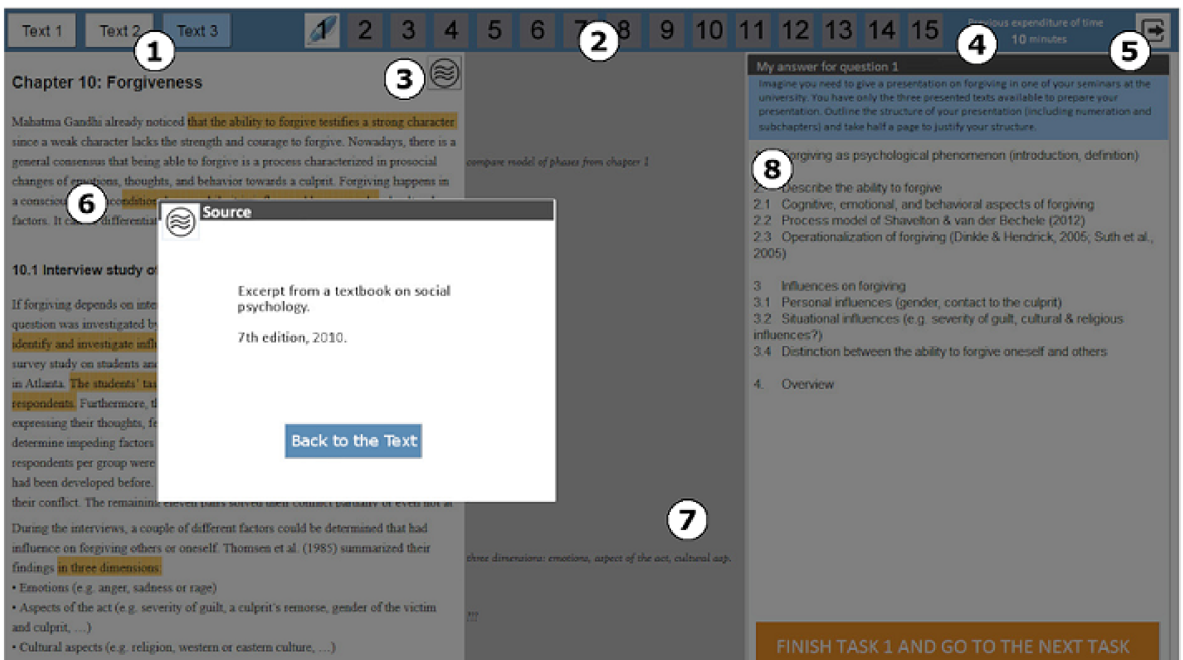
Sourcing in multiple document comprehension

- Multiple document comprehension (MDC): Competence to construct an integrated representation of a certain subject area using information from different sources

- Continuous assessment within MDC items to infer ‘Sourcing’ as an important attribute of the work process
  - **Targeted attribute of the work process/Claim**: Consideration of the origin and intention of documents (= Sourcing)

Task/Activity model for sourcing

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



① Buttons for text navigation
② Buttons for task navigation
③ Button for accessing source information
④ Unit processing time in minutes
⑤ Button for finishing the unit
⑥ Option for text highlighting
⑦ Option for commenting
⑧ Dialog of the essay writing task

(from Hahnel, Kroehne, Goldhammer, Schoor, Mahlow, & Artelt, 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 when dealing with conflicts.
- **Testable assumptions** (see Hahnel et al., 2019)
  - MDC is positively associated with repeated sourcing.
  - Graduation grades are not positively associated with repeated sourcing.
  - The number of documents, of conflicts between documents, and of items that require the comprehension of source information evoke repeated sourcing.
  - The position of units is not related to repeated sourcing.
- **Evidence**: Empirical relation of process indicators to the competence 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)

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

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

(from Hahnel et al., 2019)

Overview

- Introduction
- Kinds of assessment
- ECD view on continuous assessment within items
- Argument-based validation
- Example 1: Test-taking engagement
- Example 2: Sourcing in reading
- Concluding remarks

Concluding remarks

- **Continuous assessment** within complex interactive items (e.g., based on log data)
  - Provides process indicators representing attributes of the work process

- The **interpretation of process indicators** needs to be
  - Challenged by appropriate validation strategies
  - Already considered when designing the tasks

- **Importance of substantive theories** for task design, evidence identification, and validation (construct interpretation)

Concluding remarks

- **Lack of theory or process models** relating behavioral actions to attributes of the work process through evidence identification and accumulation (Kane & Mislevy, 2017; Mislevy et al., 2012)
  - Exploratory analyses enabling theory development
- **Data-driven approaches** informing evidence identification
  - Methods for pattern detection (educational data mining) (e.g., He & von Davier, 2016)
  - Machine learning (supervised, unsupervised)
  - Need for cross-validation (validating the 'learned' evidence identification rule)
- **Evidence accumulations** by means of statistical models: Standard psychometric models, Bayesian networks (see De Klerk, Veldkamp, & Eggen, 2015)

Thank you! – Questions, comments…?

contact: goldhammer@dipf.de