Using Computer-Based Process Data to Improve Assessment Validity: Challenges and Opportunities

Stephen G. Sireci

Center for Educational Assessment

University of Massachusetts Amherst, USA

Closing Keynote for *Beyond Results: Paving the Way for the Use of Process Data*

June 18, 2020

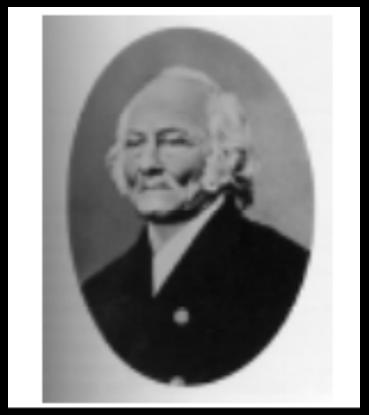
Sponsored by IAEEA, Leibniz Institute for Research and Information in Education, and Centre for International Student Assessment

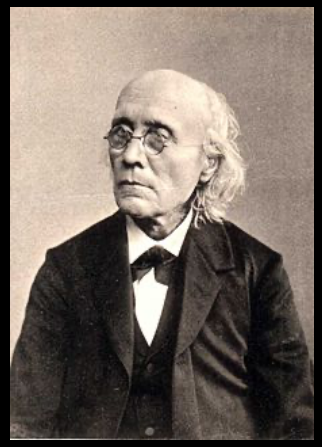
It is notable this conference was originally scheduled in Frankfurt and sponsored in part by the Leibniz Institute for Research and Information in Education.

Why?

Germany (Leipzig)

- **Was essentially the birthplace of "modern" psychometrics.**
- **And what were Weber and Fechner focused on?**





- **Response processes!**
- **So, here we are 160 years later taking advantage of computerized technology to help.**

I want to first thank

Heiko Sibberns (and colleagues) for the invitation.

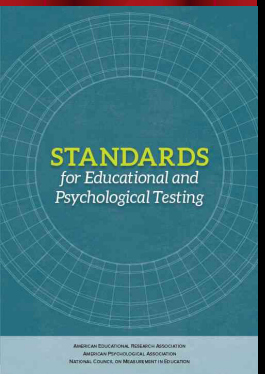
Processing the Process Data Conference

- **Data Structure**
- **Processing Data**
- **Modeling**
- **Overcoming challenges**
- **Exciting examples**
- **Validity**

Validity and Response Process Data

"Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests"
AERA, APA, NCME *Standards* (2014, p. 11).

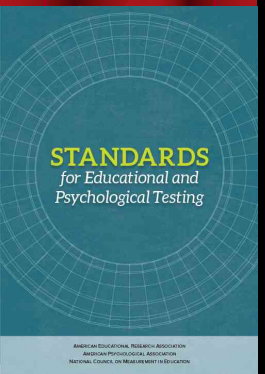
- How can "response process data" help us develop better tests?



Validity and Response Process Data

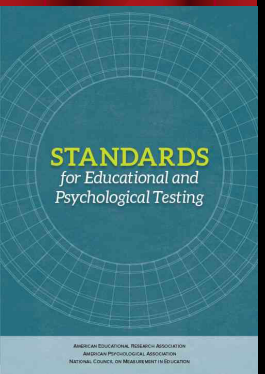
"Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests"
AERA, APA, NCME *Standards* (2014, p. 11).

- What evidence can "response process data" provide to support the interpretations and uses of test scores?



The *Standards for Educational & Psychological Testing* (1999, 2014)

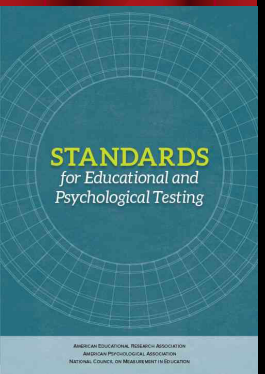
- **Describe 5 "sources of validity evidence. Validity evidence based on…**
  1. **Test Content**
  2. **Response Processes**
  3. **Relations to other variables**
  4. **Internal structure**
  5. **Consequences of testing**



von Davier: "remember what response data *can't* tell us."

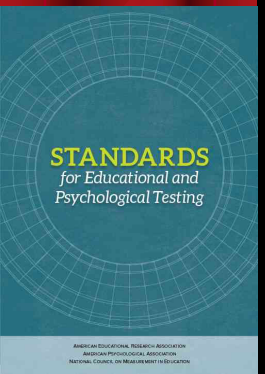
II. Validity evidence based on response processes

"empirical analyses of the response processes of test takers can provide evidence concerning the fit between the construct and the detailed nature of the performance or response actually engaged in by test takers"
AERA et al. (2014, p. 15)



II. Validity evidence based on response processes

- **Examples of this type of evidence from the *Standards*:**
  - **Questioning test takers about strategies**
  - **Successive drafts of writing tasks**
  - **Eye movements**
  - **Response times**
  - **Judgments of scorers**
- **Only one standard specifically on such evidence**
  - **If rationale for score interpretation depends on cognitive processes evidence should be provided (p. 26)**



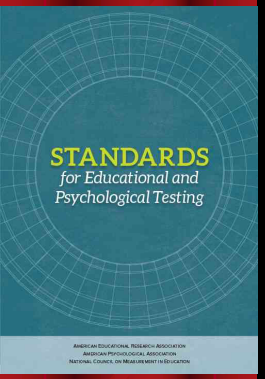
Messick (1989): Threats to validity

"Tests are imperfect measures of constructs because they either leave out something that should be included…or else include something that should be left out, or both" (p. 34).

- **Construct underrepresentation**
- **Construct-irrelevant variance**

Samuel Messick

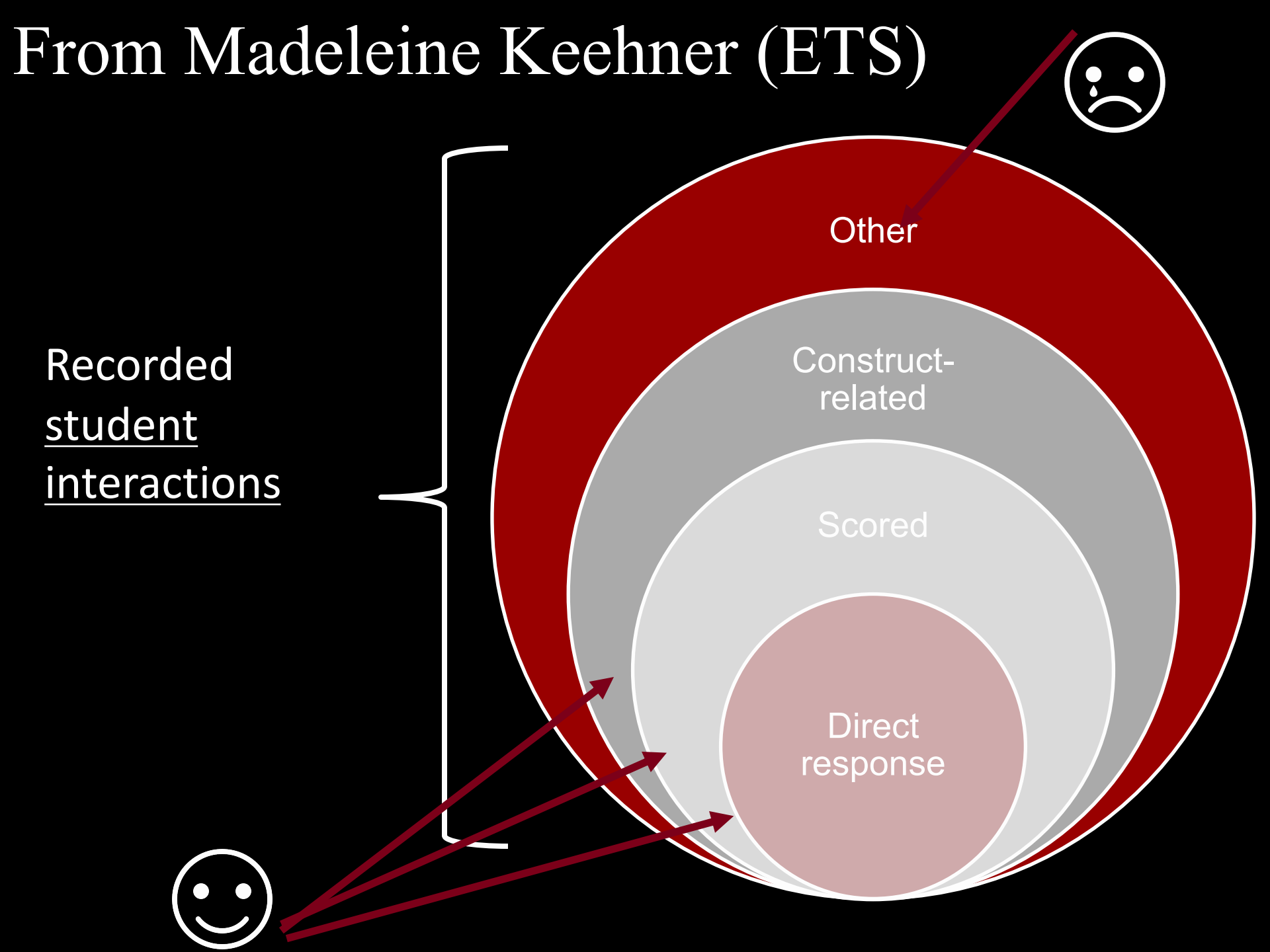
Analysis of computer-based process data can provide information on both



"construct-irrelevant variance may arise because test items elicit varieties of responses other than those intended…" (AERA et al., p. 56)

But, computer-based assessments can improve construct representation by measuring new skills (processes)

From Madeleine Keehner (ETS)



Process Data: Opportunities and Challenges

- **Opportunities**
  - **Provide direct measures of examinees' behavior that reflect cognition**
  - **Can be used in scoring—improved "construct representation"**
  - **Can be used in validation—confirm intended skills are measured**
  - **Can be used to improve test development**
  - **Can be used to identify potential biases**

Process Data: Challenges and Opportunities

- **Challenges**
  - **What response behaviors do we gather?**
  - **How do we gather them?**
  - **How do we analyze them**
  - **How do we interpret and report them?**
  - **How do we use them to improve measurement (i.e., more valid assessment)?**

Research

What evidence do we have that computer-based response process data:

- **Provide direct measures of examinees' behavior that reflect cognition**
- **Can be used in scoring—improved "construct representation"**
- **Can be used in validation—confirm intended skills are measured**
- **Can be used to improve test development**

Research

What evidence do we have that computer-based response process data:

- **provide direct measures of examinees' behavior that reflect cognition?**
  - **Greiff (2020, yesterday)**
  - **Kreuter (2020, yesterday)**
  - **Naumann (2020, yesterday)**
  - **Reis Costa (2020, today)**
  - **von Davier (2020, yesterday)**
  - **Wise (2020, yesterday)**

Challenges: How to Conquer them

- Challenges
  - What response behaviors do we gather?
  - How do we gather them, analyze them, interpret and report them?
  - use them to improve measurement?

Hahnel, Hao, He, Keskpaik, Kröehne, Reis Costa, Sibberns

  - Privacy

Drachsler

Themes

- **Processing process data is hard, but there is help!**
- **Deciding what data to gather and how to analyze it is hard, but here are some examples…**
- **Start with a cognitive model, and build assessment to gather relevant process data**

Naumann: Testing substantive theories requires psychologically meaningful indicators derived from ambiguous process data.

Themes

- **Processing process data is hard, but there is help!**
- **Deciding what data to gather and how to analyze it is hard, but here are some examples…**
- **Start with a cognitive model, and build assessment to gather relevant process data**

Wise: Study behavior before making assumptions about it!

Themes

- **Process data are helpful for validation (correct skills, evaluating item formats)**
- **Collaborate**
  - **Programmers, cognitive scientists, psychometricians (computer scientists)**
  - **User experience designers**

“Understandably, the advent of CBT and the myriad innovations afforded by technological advances have facilitated the building of bridges between the two sciences of psychometrics and cognitive psychology” (Huff & Sireci, 2001, p. 23)

Heiko Sibberns (2020, yesterday)

Constructs are the mountains,

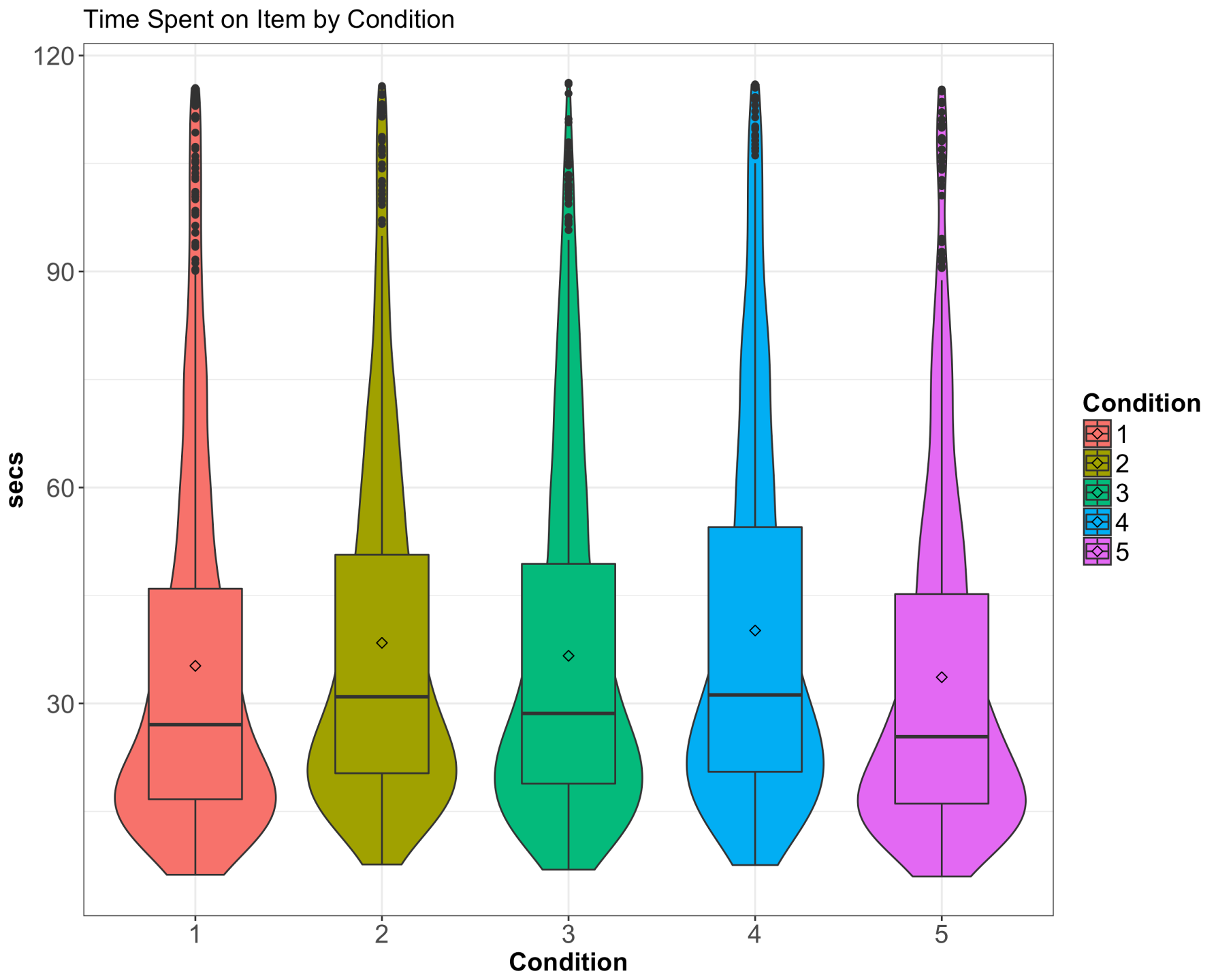
Log data are the muddy ground.

Research on process data

- **Experimental studies**
  - Evaluate item formats
- **Fishing expeditions**
  - What can we get from log data?
  - How can we use it?
  - Exploratory, useful
- **Evidence-centered design**
  - From fishing to capturing, confirming
- **Theory-based**
  - Scoring based on response time
  - Engagement

Experimental studies

- **Kreuter (2020, yesterday)**
- **Arslan, Jian, Gong, & Kheener (2019)**
  - **Investigated different drag-and-drop item designs**


Time Spent on Item by Condition
120
90
60
30
secs
1
2
3
4
5
Condition
Condition
1
2
3
4
5

Fishing expeditions

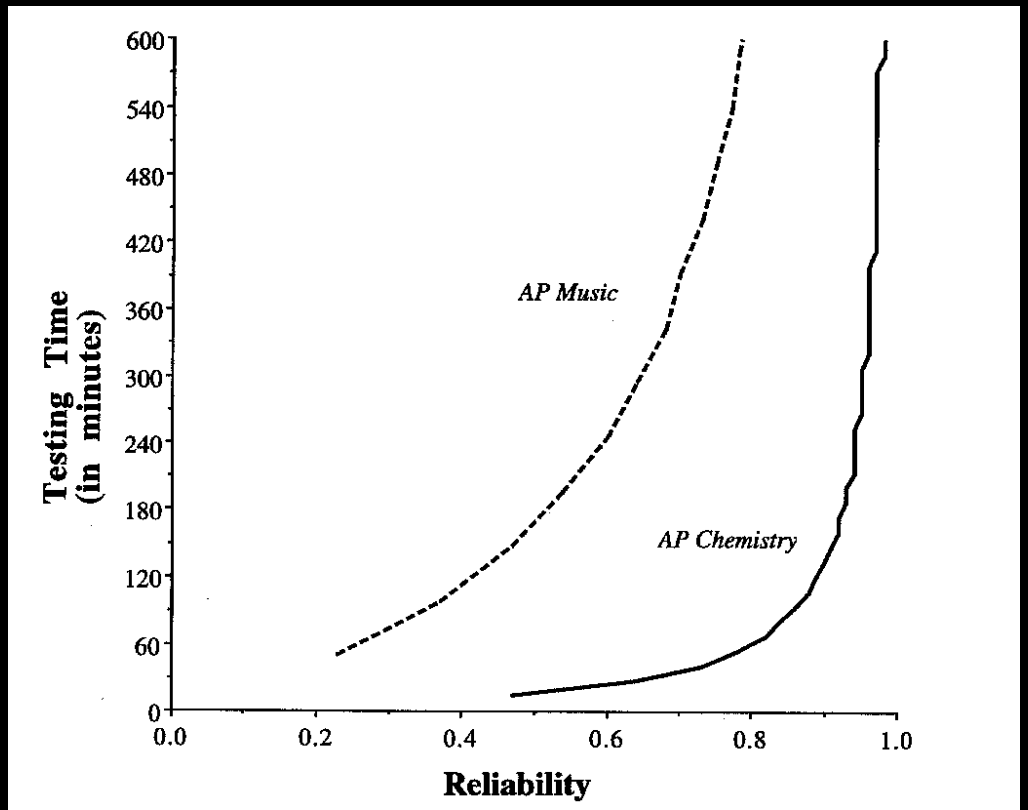
What variables are meaningful and how do we define them?

- **Response time**
- **Number of visits**
- **Number of changes**
- **Number of actions**
- **First response latency**
- **Proficiency**

There are many ways to define these and other relevant variables.

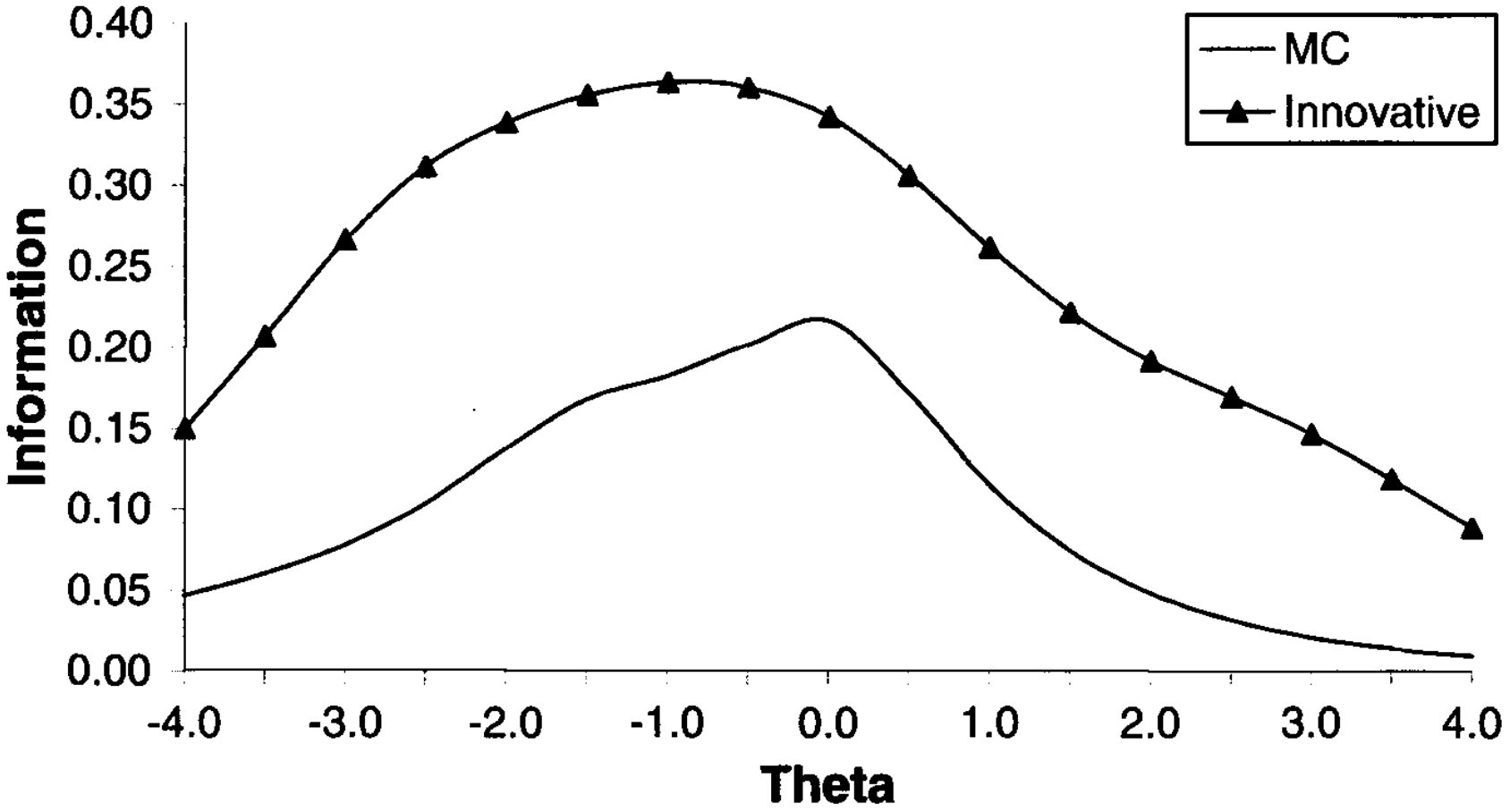
CTT index to bring back?

- **Reliamin index**
  - **Amount of "reliability" per unit of testing time (Wainer & Thissen, 1993)**



Test information per response time unit

Jodoin



Evidence-centered design

- **Never has the work of Mislevy been so relevant**
  - **Design framework specifies what data to gather**
- **Embretson, Fischer, Sheehan, Tatsuoka**
  - **Rich history of understanding skills needed to respond to items**

Cognitive models guide test development and specify the relationships that should be found based on analysis of response process data (Mislevy, Steinberg, Almond, 1999).

(Other) Theory-based research

von Davier:

- **Hypothesize what construct match would look like in response data**
- **Hypothesize what construct mismatch would look like in response data**

- **Engagement/Disengagement**
  - **Wise**
  - **And others**

Validity questions process data can help us answer

- **Improved construct representation?**
  - **Using speed (fluency) in scoring**
  - **Partial credit scoring**
- **Accounting for construct-irrelevant variance**
  - **Engagement index and "warning"**

Validity questions process data can help us answer

- **Are students (effectively) using test accommodations?**
- **What item formats are best for**
  - **Improved construct representation**
  - **Reducing construct-irrelevant variance**
  - **Efficiency?**
- **Are we measuring intended cognitive skills?**
- **Are there subgroup differences that threaten validity?**

Validity questions process data can help us answer

- **Are students (effectively) using test accommodations?** Hao (2020, today)
- **What item formats are best for**
  - **Improved construct representation**
  - **Reducing construct-irrelevant variance**
  - **Efficiency?**
- **Are we measuring intended cognitive skills?**
- **Are there subgroup differences that threaten validity?**

What have we learned?

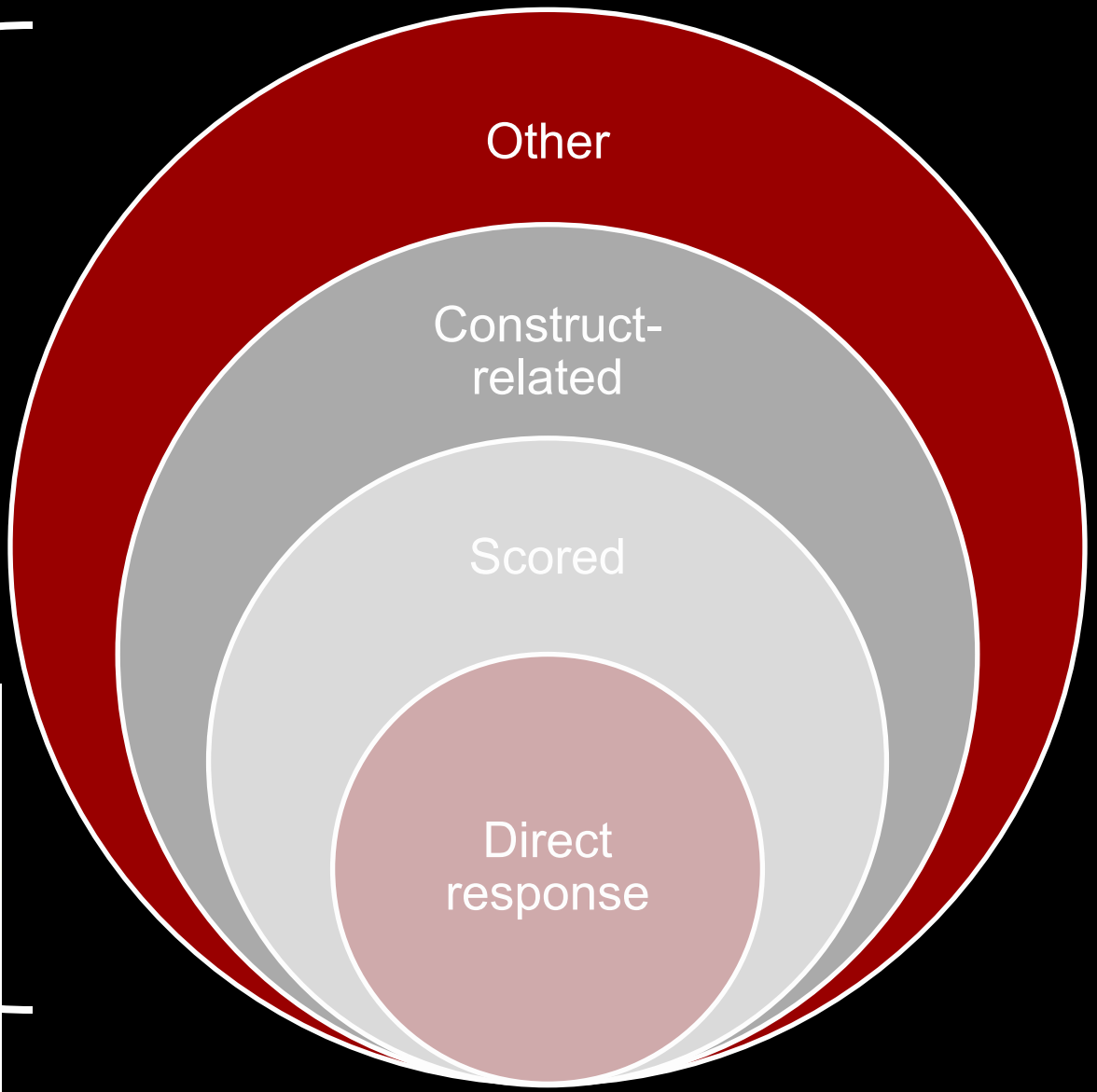
- **Response process data can**
  - **inform test development**
  - **Improve construct representation**
  - **Evaluate presence of construct-irrelevant variance**
  - **Improve validity**
  - **Improve testing experience**

What other conversations are needed

- **Applications to improving the educational process**
  - **What value do process data have for educational intervention?**
- **Fairness issues**
  - **Access to computers, computer proficiency, digital deserts**
- **Subgroup differences**
  - **What does it mean if there are? Construct-irrelevant variance?**
- **Faking response processes?**

Data Cleaning: Construct-relevant vs. irrelevant

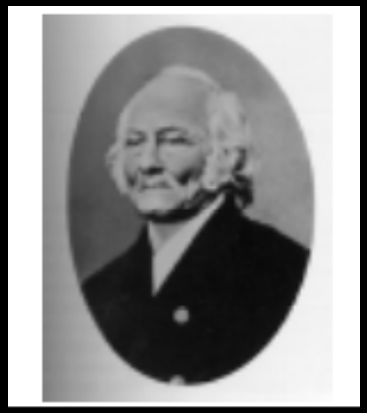
Recorded student interactions (Keehner)

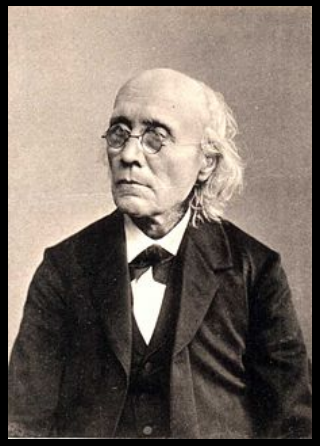


What do we want to study vs. what is easiest to study?

What is easiest to score vs. what is most informative to score?

160 years ago, in Leipzig, history was made that began a new branch of psychological science





$$S = K * \log(R)$$

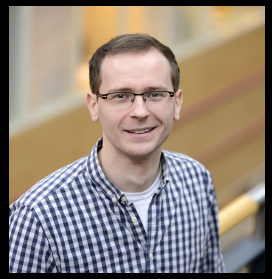
Will the work we do now endure for the next 160 years?

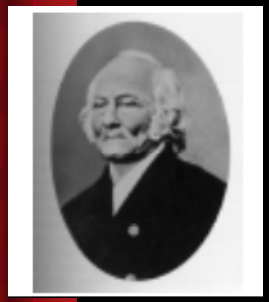


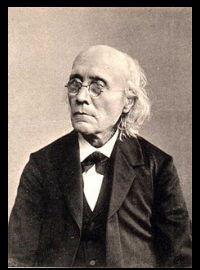









Final remarks

- **Thank you Heiko, Astrid, Ralph, IEA, DIPF, CISA, all presenters and participants.**
- **I am less ignorant today about response process data than I was 28 hours ago!**

Sireci@acad.umass.edu

NCME.org

Intestcom.org