Human scoring in CBA

Recommendations for valid and reliable scores

Sophie Doublet, Vincent Koenig, Gilbert Busana, Sophie Battisti

EMACS research unit of the University of Luxembourg

TAO Days - Berlin, 10th September 2012
Context

- Allocating scores ➔ **keystone** of assessment:
  » Results of the measuring instrument
  » Basis of its validity and reliability

- **Closed-ended questions** ✓ Easy to score

- **Open-ended questions**
  ✗ Time and resource-consuming
  ✗ Costly
  ✗ Difficult to implement

- **CBA** ➔ simplify and support the scoring process
Context

- CBA for scoring open-ended questions:

  **By**
  
  Computer = rater
  
  Automated scoring
  
  ✗ Cost
  
  ✗ Feedback
  
  ✗ Acceptability and trust

  **With**
  
  Computer = tool
  
  Human scoring
  
  ➔ Support and enhance the scoring process achieved by human raters
Context

- **TAO new feature** ➔ scoring open ended-questions
  - Scoring interface
  - Conciliation interface
Scores

Item
- Open-ended questions

Rater
- Human raters

Tool
- Computer

Scores
Item: Open-ended questions → Scores
Efficiency of open-ended questions

- High skills
- Validity

- Scoring
  - Subjectivity
  - Less reliability
  - Time and resource consuming
Efficiency of open-ended questions

Recommendation

• Use open-ended questions **only when required**:
  » Purpose of the assessment
  » Available resources

• Scoring process

  ✔ Objectivity
  ✔ More reliability
  ✔ Time and resource saving

Closed-ended
Human scoring biases

- One paper can be scored **differently** by:
  - Two raters (inter-rater reliability)
  - The same rater (intra-rater reliability)

- Classical Test Theory:

 1. Order and contrast effects
 2. Halo effect
 3. Test takers data
 4. Rater characteristics
 5. Rater variation over time
Human scoring biases
1. Order and contrast effects

- **Order** ➜ Place in the set of papers:

- **Contrast** ➜ Paper before:
Human scoring biases
1. Order and contrast effects

Before the scoring session

• Make raters aware of these effects
• Scan through the set
• Calibration papers

At the end of the scoring session

• Review first and last papers
• Paper ranking to their scores
Human scoring biases

2. Halo effect

- Score allocated to the **first question** can influence **all the scores given to the following questions**.
Human scoring biases
2. Halo effect

- Preferably score **question by question** (horizontally)

Recommendation

- No halo effect
- Routine
- Faster

- More tiring
- No global appreciation

(Allalouf, Klapfer & Fronton, 2008)
Human scoring biases
3. Data on test takers

Test taker name

• Known
  » Personal feelings

• Unknown
  » Rater’s beliefs and stereotypes
    - Gender
    - Social class
    - Culture, ethnicity, religion…
    - Attractiveness of the name → higher scores (Erwin & Calev, 1984)
Human scoring biases

3. Data on test takers

Recommendation

- Score **anonymously**
  
  » Replace test taker’s name by a number
Human scoring biases

4. Rater characteristics

- Inexperienced raters
  - Severe ➔ Lenient

- Experienced raters
  - Stable level of severity
    (Weigle, 1999)
Human scoring biases

4. Rater characteristics

Recommendations

• Scoring guidelines

• Training

• Estimation of the raters’ severity
  » Statistical adjustment for raters with consistent severity or leniency

• Multiple scoring
Human scoring biases

5. Rater variation over time
Human scoring biases
5. Rater variation over time

Recommendations

- Monitoring
  - With “monitoring” papers
  - With statistical data
- Low performances ➔ additional training
If several raters are involved...

Who scores what?

• **Question** rather than paper

• **Rater allocation design**
  » Affect reliability of scores
  » Prevent from detecting and correcting rater errors

• **Fully crossed design**

• **Incomplete but balanced design** (Patz, Wilson and Hoskens, 1997)
  » Each rater scores the same number of test takers’ works, even with randomization
If multiple scoring is used…

**What are discrepant scores?**
- Any different scores
- Threshold beyond which a conciliation is organized

**How is the final score established?**

- Discrepancy < Threshold
- Discrepancy > Threshold
- Borderline scores

If no supplementary raters are available
CBA scoring

• Paper-based vs computer-based scoring
  ✔ Same accuracy
  ✘ More demanding → greater cognitive and physical efforts

• Computers characteristics → constrain and affect scoring behavior
  » Reading on screen
  » Navigation between test takers and questions
  » Comments and annotations

• IT literacy of the raters?

• Usability of the scoring tool?
CBA scoring

- **Familiarization**

- **Easy-to-use scoring tool**
  - Navigation
    - Between the different levels (questions / test takers)
    - Possibility to review and change given scores

  - Feedback
    - Questions already scored / remaining to score

- **Effective - Efficient - Satisfying**
Conclusion

• Scoring open-ended questions is still mainly a human-dependent task

• Computers can improve this time-consuming task and increase validity and reliability

- For the assessment organizers: minimize human biases
  » Allocation of questions to raters
  » Online training
  » Easy monitoring

- For the raters:
  » Less time consuming and more enjoyable
Open-ended questions

Only if **closed-ended questions** are not possible

**Rater**

**Human raters**

- Make raters aware
- **Anonymous** scoring
- Score **question by question**
- Training / Monitoring / Multiple scoring

**Tool**

**Computer**

- **Familiarize** raters with the computerized tool
- **Usability** of the scoring interface
Thanks for your attention!