Looking Beyond Rater Cognition: Operational Rater Types in Writing Performance Assessment

Thomas Eckes
TestDaF Institute
Hagen, Germany
May 29, 2010
Overview

1. Rater Variability
2. Rater Cognition
3. Rater Types
4. Rater Cognition and Rater Behavior
5. Summary and Implications
Rater Variability

Typical Design of Performance Assessment

- Trained raters evaluate the quality of an examinee’s response to particular tasks based on a set of criteria representing the construct being measured using a scale (or scales) with ordered categories referring to distinct levels of proficiency
Rater Variability

Rater Performance

- High cognitive demands on raters (Hamp-Lyons & Henning, 1991; Luoma, 2004)
- The rater, too, must perform (e.g., Reed & Cohen, 2001)
- The quality of each rater’s performance needs to be monitored and evaluated, especially in high-stakes settings
- Major threat to high-quality assessment:
Rater Variability

Facets of Rater Variability

- Variability associated with characteristics of the raters and not with the performance of examinees (e.g., rater severity or leniency)
- A component of unwanted variability contributing to construct-irrelevant variance in examinee scores
- Threatens the validity and fairness of performance assessments
- Pronounced even among trained, experienced raters
- Rater training does not significantly reduce rater variability (though training can raise within-rater consistency)
Facets of Rater Variability (Cont’d)

- Manifests itself in various forms; ie, raters may differ in:
  - the understanding of the construct being measured
  - the interpretation and use of scoring criteria
  - the overall degree of severity or leniency
  - the degree of compliance with the scoring rubric
  - the understanding and use of rating scale categories
  - etc.
Rater Cognition

Rater Cognition Perspective

- Performance assessment as an interplay between bottom-up (text-driven) and top-down (knowledge-driven) processes
- Top-down processes refer to the influence of prior knowledge, professional experience, attitudes, and expectations
Rater Cognition

Rater Cognition Research

- Decision-making strategies, reading styles (e.g., Cumming, 1990; Sakyi, 2000; Vaughan, 1991)
- Rater cognition models (Freedman & Calfee, 1983; Wolfe, 1997)
- Scoring focus (Wolfe et al., 1998): Mental weighting scheme used for interpreting and applying scoring criteria
- Perception, interpretation, and use of performance features play a key role in explaining between-rater differences
Rater Types

*These differences between raters may lead us to think in terms of rater types.*

T. McNamara (1996, p. 124)
Rater Types

The Rater Type Hypothesis

- Experienced, trained raters fall into types
- Types are separated from one another by patterns of importance attached to routinely-used scoring criteria
- Each rater belongs to a type adopting a distinct point of view regarding criterion importance
Rater Types

Rater Types in Writing Assessment

- Questionnaire study (Eckes, 2008)
- TestDaF writing section
- 64 raters
- 9 routinely-used scoring criteria
- 4-point importance scale (less important, important, very important, extremely important)
Rater Types in Writing Assessment (Cont’d)

- Scoring criteria

<table>
<thead>
<tr>
<th>Fluency</th>
<th>Global impression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train of thought</td>
<td></td>
</tr>
<tr>
<td>Structure</td>
<td></td>
</tr>
<tr>
<td>Completeness</td>
<td></td>
</tr>
<tr>
<td>Description</td>
<td>Task fulfillment</td>
</tr>
<tr>
<td>Argumentation</td>
<td></td>
</tr>
<tr>
<td>Syntax</td>
<td>Linguistic realization</td>
</tr>
<tr>
<td>Vocabulary</td>
<td></td>
</tr>
<tr>
<td>Correctness</td>
<td></td>
</tr>
</tbody>
</table>
## Interrelations Between Rater Types and Scoring Criteria

<table>
<thead>
<tr>
<th>Rater Type</th>
<th>Fluency</th>
<th>Train of thought</th>
<th>Structure</th>
<th>Completeness</th>
<th>Description</th>
<th>Argument</th>
<th>Syntax</th>
<th>Vocabulary</th>
<th>Correctness</th>
</tr>
</thead>
<tbody>
<tr>
<td>A Syntax</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B Correctn.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C Structure</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D Fluency</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E Non-Fluency</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F Non-argumt.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Rater Types

Limitations of This Research

- Raters had to indicate the weight they would *in general* attach to each criterion when working in a TestDaF scoring session.
- They were asked *not* to think of a particular writing performance.
- The analysis was based on self-report data.
- No direct connection with operational scoring behaviour.

→ How, if at all, does rater cognition relate to rater behavior?
Rater Cognition and Rater Behavior

Looking Beyond Rater Cognition

- Evidence from rater bias studies (many-facet Rasch measurement)
- Criteria rated more harshly by some raters and more leniently by some other raters (Elder et al., 2007; McNamara, 1996; Wigglesworth, 1993)
  - *Harsh* towards Content and/or Organization
  - *Lenient* towards Language Use and/or Mechanics
  - Or vice versa
Rater Cognition and Rater Behavior

Research Questions

- Are there rater types in operational TestDaF scoring sessions?
- Raters belonging to the same “operational rater type” (ORT) would show similar patterns of severity/leniency across criteria
- Raters belonging to different ORTs would show dissimilar patterns of severity/leniency across criteria
- How would these ORTs relate to the “cognitive rater types” (CRTs) identified in the previous rater type study?
Rater Cognition and Rater Behavior

Research Questions (Cont’d)

- High perceived importance $\rightarrow$ harsh rating
- Low perceived importance $\rightarrow$ lenient rating
Rater Cognition and Rater Behavior

Research Design

- Three CRTs selected (Syntax Type, Structure Type, Fluency Type)
- 18 TestDaF raters (had also taken part in the questionnaire study)
- Data from an operational TestDaF scoring session (3 to 4 months after the questionnaire study)
Rater Cognition and Rater Behavior

Rating Scale Used for Operational Ratings

<table>
<thead>
<tr>
<th>TestDaF Levels (TDNs)</th>
<th>TDN 3</th>
<th>TDN 4</th>
<th>TDN 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Below TDN 3</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Four-category scale

CEFR levels B2.1 – C1.2
Rater Cognition and Rater Behavior

**Data Analysis**

- Rater-by-Criterion bias analysis (FACETS; Linacre, 2009)
- Cluster analysis of bias measures to identify ORTs (ClustanGraphics; Wishart, 2006)
- Cluster profiles (average bias measures for each combination of ORT and criterion)
Rater Cognition and Rater Behavior

Dendrogram
Cluster analysis of bias measures
18 Raters
ORTs 1 to 4

Raters from CRT D only

Raters from CRT C only

Increase in Sum of Squares
Rater Cognition and Rater Behavior

Dendrogram
Cluster analysis of bias measures
18 Raters
ORTs 1 to 4

Mixed type: CRT A/C 1
ORT 1
02 C
35 A
34 A
45 A

ORT 2
15 D
17 D
47 D
46 D

Mixed type: CRT A/C 2
ORT 3
06 C
12 A
14 A
21 A
60 A
11 A
31 C
38 A

ORT 4
18 C
56 C

Increase in Sum of Squares
0,000
0,256
0,511
0,767
1,023
1,279
1,534
1,790

TestDaF - Test Deutsch als Fremdsprache
Rater Cognition and Rater Behavior

Bias diagram for Operational Rater Type 1
Bias diagram for Operational Rater Type 2 (ORT 2)

Rater Cognition and Rater Behavior

Bias measure (Logit)

ORT 2
Rater Cognition and Rater Behavior

ORT 3 – mixed type

Bias diagram for Operational Rater Type 3 (ORT 3; mixed type)
Bias diagram for Operational Rater Type 4 (ORT 4)

Rater Cognition and Rater Behavior

ORTH 4

Fluency  Train of thought  Structure  Completeness  Description  Argumentation  Syntax  Vocabulary  Correctness

Lenient

Bias measure (Logit)

Harsh

Bias diagram for Operational Rater Type 4 (ORT 4)
Summary and Implications

Summary of Findings

- Four operational rater types (ORTs) identified
- Distinct patterns of severity and leniency towards scoring criteria
- Evidence of compensatory bias across criteria, thus confirming Schaefer’s (2008) finding
- Evidence of a link between CRT and ORT
- Higher perceived importance of a given criterion tends to be associated with higher severity towards that criterion (and vice versa)
Summary and Implications

Implications for Rater Training

- Rater feedback should take rater types into account
- Redirecting attention to criteria not captured within a given rater type’s scoring focus
- Checking for (implicit) compensatory scoring biases
- Studying stability and change in rater-by-criterion bias measures over time
- Examining the effectiveness of rater training activities through MFRM procedures
Summary and Implications

*Much more work needs to be done on the definition of consistent rater types.*

Thank you.

E-Mail: thomas.eckes@testdaf.de
Web: www.testdaf.de
Gesellschaft für Akademische Studienvorbereitung und Testentwicklung e.V. (g.a.s.t.)