Strategies for Reducing Rater Bias

BALEAP Conference, 21st April 2023

Peter Davidson
Zayed University

Outline

1. Introduction
2. What is Rater Bias?
3. Bias Rating Behaviour
   - 3.1 Characteristics of the rater
   - 3.2 Characteristics of the candidate
4. Research on Rater Variability
5. The Manifestation of Rater Variability and Inconsistency between Raters
6. In Defence of the Human Rater
7. How to Reduce Rater Bias
   - 7.1 Improving internal consistency
   - 7.2 Improving inter-rater reliability

1. Introduction

- Weir (2005: 56): “The last decade of the twentieth century saw a general decline in the prestige of psychometric, statistically-driven approaches to testing.”
- Covid-19 has exacerbated this shift
- move away from the use of discrete skill tests
- move towards more holistic, authentic and integrated assessment
- task-based / performance-based / scenario-based
- essays / case studies / reports / projects / research papers / portfolios / seminars / presentations
- greater emphasis on subjective human raters
- the reliability of the assessment could be compromised
- need to mitigate against rater bias

2. What is Rater Bias?

- "**Rater bias** refers to a systematic pattern of rater behavior that manifests itself in unusually severe (or lenient) ratings associated with a particular aspect of the assessment situation" (Eckes, 2012, p. 273).

- **Differential rating functioning**: when a rater shows high degrees of severity (or leniency) with a particular group of candidates (e.g. females or Russians), or a particular task (instruction manual or a presentation), or a particular criterion (e.g. grammar or fluency).

- Can be determined by Many Facet Rasch Model (MFRM)

3. Bias Rating Behaviour

3.1 Characteristics of the rater

1. **Leniency Bias**: the tendency of a rater to score too leniently
2. **Harshness Bias**: the tendency of a rater to score too severely or too harshly
3. **Central Tendency Bias**: the tendency of a rater to score near the center of a scale
4. **Restriction of Range Bias**: the tendency of a rater to limit their range of scores
5. **The Halo Effect Bias**: the tendency of a rater that gives a high rating on one part of criteria, to give high ratings on other parts of the criteria
6. **The Horns Effect Bias**: the tendency of a rater that gives a low rating on one part of criteria, to give low ratings on other parts of the criteria

3. Bias Rating Behaviour

3.1 Characteristics of the rater

7. **The Contrast Effect Bias**: the tendency of a rater to compare one performance with another performance
8. **First Impression Bias**: the tendency of a rater to be strongly influenced by the beginning of a performance
9. **Recency Bias**: the tendency of a rater to be strongly influenced by the end of a performance
10. **Current State of Mind Bias**: the tendency of a rater to be influenced by their current state of mind

3. Bias Rating Behaviour

3.2 Characteristics of the candidate

11. **Acquaintanceship Bias**: the tendency of the rater to score candidates they know higher (or lower)
12. **Expectation / Confirmation Bias**: the tendency of a rater to be influenced by their expectations of the candidate
13. **Similarity Bias**: the tendency of a rater to be influenced by how similar the candidate is to them
14. **Cultural Familiarity Bias**: the tendency of a rater to be influenced by how familiar they are with a certain culture
15. **Ethnicity Bias**: the tendency of a rater to be influenced by the ethnicity of the candidate
16. **Gender Bias**: the tendency of a rater to be influenced by the gender of the candidate

3. Bias Rating Behaviour
3.2 Characteristics of the candidate

17. **Sexual Orientation Bias**: the tendency of a rater to be influenced by the sexual orientation of the candidate
18. **Social Status Bias**: the tendency of a rater to be influenced by the social status of the candidate
19. **Age Bias**: the tendency of a rater to be influenced by the age of the candidate
20. **Attitude Bias**: the tendency of a rater to be influenced by the attitude of the candidate
21. **Handwriting Bias**: the tendency of a rater to be influenced by the quality of handwriting of the candidate
22. **Sympathy Bias**: the tendency of a rater to score a candidate they feel sympathy for higher

4. Research on Rater Variability

Edgeworth (1890): expressed concern that only a third to two-thirds of successful candidates would pass public exams again if given a different set of judges

Diederich et al. (1961): 53 raters scored 300 scripts on a 9-point scale

- 94% of papers received at least 7 grades; no paper received less than 5 grades

Weigle (1998): 16 raters scored 60 scripts before and after training

- Before training, inexperienced raters were more severe and less consistent than experienced raters
- After training, inexperienced raters were still more severe than experienced raters, but were more consistent

4. Research on Rater Variability

Schaefer (2008): 40 raters scored 40 scripts

- If Content and Organization were scored severely, then Language Use and Mechanics were scored leniently
- There tended to be more severe or lenient bias with higher level candidates

Yousun (2010): 3 raters scored 254 scripts

- Grammar was scored most severely, while Organization most leniently

Erguvan & Dunya (2020): 3 raters scored 109 scripts between them, and 6 anchor scripts

- Raters were internally consistent, but the range of scores was restricted, and there was no inter-rater reliability

5. The Manifestation of Rater Variability and Inconsistency between Raters

(Eckes, 2008)

Raters may differ in the:

- degree to which they comply with the scoring rubric
- way they interpret criteria employed in operational scoring sessions
- degree of severity or leniency exhibited when scoring examinee performance
- understanding and use of rating scale categories
- degree to which their ratings are consistent across examinees, scoring criteria, and performance tasks

6. In Defence of the Human Rater

- high cognitive demand
- multi-tasking
- focus on a number of different criteria and descriptors
- multiple candidates (in speaking)
- multiple roles
- erratic performances
- pressure of being monitored
- time pressure
- fatigue

7. How to Reduce Rater Bias
7.1 Improving internal consistency

- acknowledge the potential for rater bias
- be aware of different types of rater bias behavior
- be aware of your own rater bias
- participate in regular training sessions
- participate in regular norming / moderation / standardization / calibration sessions
- rate blind (to avoid confirmation bias)
- use a well-defined rubric (scoring criteria) with clear, logical descriptors
- create an answer template
- devise a rating schedule with regular breaks
- avoid distractions when rating
- rate all of one question type together
- rate all of one question type in the same sitting

7. How to reduce rater bias
7.2 Improving inter-rater reliability

- conduct regular training sessions
- conduct regular norming / moderation / standardization / calibration sessions
- get raters to rate blind (to avoid confirmation bias)
- use double or multiple raters
- use a well-defined rubric (scoring criteria) with clear, logical descriptors
- have all raters rate in the same room at the same time
- have a clear moderation policy
- use rater moderation where necessary
- monitor raters regularly
- provide raters with feedback on monitoring
- use FACETS to identify rater bias
- use a limited pool of raters
- use automated essay scoring

References

Bachman, L.F., Lynch, B., and Mason M. (1995). Investigating variability in tasks and rater judgment in a performance test of foreign speaking. *Language Testing,* 12(2), 238-237.

Brown, J.D. (1991). Do English and ESL instructors rate samples differently? *TESOL Quarterly*, 25(4), 587-603.

Deiderich, P.B., French, J.W, and Carlton, S.T. (1961). Factors in judgements of writing ability. *Research Bulletin 61-15*. Princeton, N.J. ETS.

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing,* 25(2), 155-185.

Eckes, T. (2012). Operational rater types in writing assessment: Linking rater cognition to rater behavior. *Language Assessment Quarterly,* 9(3), 270-292.

Edgeworth, F.Y. (1890). The element of chance in competitive examinations. *Journal of the Royal Statistical Society* 53, 460-475 and 644-663.

References

Erguvan, I.N. and Dunya, B.A. (2020). Analyzing rater severity in a freshman composition course using many facet Rasch measurement. *Language Testing in Asia*, 10(1), 1-20.

Lumley, T. and McNamara, T.F. (1995). Rater characteristics and rater bias: implications and training. *Language Testing,* 12(1), 54-71.

Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing,* 25(4), 465-493.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing, 15*(2), 263–287.

Weir, C.J. (2005). *Language Testing and Validation*. Palgrave Macmillan, London.

Yousun, S. (2010). A FACETS analysis of rater characteristics and rater bias in measuring L2 writing performance. *English Language and Literature Teaching*, 16(1), 123-142.