EALTA Conference
Barcelona, Spain
2007

Validating Rating Procedures in South Africa

by
Karien Hattingh
North-West University, Potchefstroom

Purpose of assessing writing:

- Make inferences about learner ability;
- Make decisions based on these inferences.

Definition of "Validity"

- Traditional:
  - Measure the construct it claims to measure
- Modern:
  - Measure the construct it claims to measure;
  - Consider the impact of scores after testing

AIM of Study:

- Validating the rating scale for assessing written performances of Grade 12 ESL learners, performed during the final school-learning (matriculation) examination in South Africa.

Focus of this paper:

- Progress report
- Phase 1: Benchmarking

Structure of paper:

- 1)South African context;
- 2) Current rating scale;
- 3) Methodology
  - Validation procedure;
  - Benchmark Rating;
  - Interpretation of Rasch Analysis Results.

South African Context

- 11 Official languages.
- 3.5 mil of 45 mil speak English as mother-tongue.

- Grade 12 proficiency levels:

  Illiterate__>__>__>__>__>__>__Near Native

- Limited Access to resources

Current Rating Scale

- "Creative Writing" recently re-introduced as as paper in final matriculation examination.

- Scale used to score Grade 12 performances:
  - is a 7-point rating scale (1=poor; 7=excellent);
  - considers 2 Criteria: Language (1); Content (2);
  - was intuitively developed by a panel of judges;
  - has not been validated empirically

  (empirical evidence is necessary to supports claims of validity).

Problem Questions

- 1) Do the 34 performances rated for benchmarking represent the full range of ability levels?

- 2) Did the raters score sufficiently consistent to identify benchmark performance levels represented by each essay on both criteria (degree of inter-rater and intra-rater consistency).

Methodology

- Benchmark rating
- Rasch Analysis

Benchmarking

- Setting standards for assessing written performances at particular levels.
- Experts identify agreed levels of performance

- “Standard setting is an activity in which an operational standard in criterion-referenced testing, namely, a cut-off score, is established by panel of expert judges. It is a process of expert judges’ perception of the minimal competence required in the field being quantified as a score at which the examinees will be classified into qualified and nonqualified (Kozakia, 2004:1).”

Benchmark Rating

- 9 expert raters (10-12 years experience, familiar with context);
- 34 written performances by Grade 12 ESL learners;
- Performances rated using the 7-point scale;
- 2 Items (Criteria): Language (Item 1)
  Content (Item 2)

Credible method

- Kozakie (2004:1):

  "…[W]hat is crucial for the establishment of credible standards is the selection of a method through which judges' intentions are accurately represented and the provision of defensible accounts for accepting the error in decisions inherent in human judgment."

Rasch Analysis

- IRT Model
- Function: investigating rater severity, learner ability.
- Value:
  - provides individual-level interaction information (Kozaki, 2004).
  - All facets measured on the same “true interval” logit scale (Henning, 1987);
  - Removes the effects of rater severity/learner ability on scores – measures “true scores” (Linacre, 1994).

Rasch Model Fit

- Better fit = more dependable data.
- Critical range = - 2 → +2 logits.
- For best fit:
  - systematically remove mis-fitting values;
  - ensures more objective Rasch Measures

(Kozaki, 2004).

Compare Graphs 1 and 2

- Systematic removal of mis-fitting values:
  - 3 rounds.
- Result: good model fit
  - Some variance apparent

Answering Problem Questions 1 & 2

- Conducted Bias Analysis in order to find answers to Problem Questions.

- Results for the following are discussed:
  - Vertical Scale Report
  - Rasch Estimates
  - Bias Interaction Analysis

Vertical Ruler Report

- True interval scale
- Function: measures ‘True Score’, removing effects of raters.
- Results: Essays (learner abilities) are spread across the full range of performance levels

(Vertical Ruler report Continues)

- Aim:
  - Determine whether essays represent the full range of scores;
  - Establish a general profile of differences in rater severity.
- Results
  - PQ1: Essay ratings (learner abilities) are spread across the full range of performance levels.
  - PQ2:Raters are clustered around 0 = similar ratings; more detailed analysis needed.

Rasch Estimate

- 3 types: Measure, Stnd Error, Fit Statistics
- Function: Identify any rater bias
- Results: no crucial misfits

  Most lenient – Rater 200; Most severe – Rater 100;

  Statistically significant variance – Rater 700;
- Thus: Rater severity varies (expected),

(Rasch Estimate Continue)

- Results: no crucial misfits

  Most lenient – Rater 200; Most severe – Rater 100;

  Statistically significant variance – Rater 700;

- Thus: Rater severity varies (expected);

- Only Rater 700's scoring varies excessively;

Inter- & Intra-rater variance

- Variety between raters is typical, even after training (McNamara, 1996; Weigle, 1998; Kozaki, 2004).

- Such findings justify using the MRA

"…which assumes, makes use of, and compensates for inter-rater differences. Inter-rater disagreement…is in fact a necessary component for Multifacet Rasch Analysis" (Kozaki, 2004:20).

Interaction Analysis

- Compares the observed raw score to expected scores.

- Function: Provides detailed information about individual raters' scoring tendencies, i.e. which Items do they tend to rate more harshly.

(Interaction Analysis Continues)

- Results:
  - Raters 300 (severe on Item 2-Language) and 800 (severe on Item 1 Language) show the most bias in their rating.
  - No misfits to the model
  - Therefore, raters score sufficiently consistent in order to assign reliable benchmark scores.

Conclusion

- PQ 1: Do the 34 scripts represent learner abilities across thefull range of performance levels?

- Vertical Ruler Report shows: “Yes”.

- PQ2: Did the raters score consistently enough to Identify benchmark performance levels for each essay?
- Rasch Estimates & Bias Interaction Analysis show: “Yes” ;
- However, with definate inconsistencies regarding inter-rater and intra-rater severity.

FINAL OUTCOME

- Phase 1, Benchmarking exercise = Successful

Thank you

- EALTA
- Personal
- Audience
- Family and friends

Cordially

Karien Hattingh

Karien.hattingh@nwu.ac.za

Tel. (49)18 299 1554