universität innsbruck

Evaluating the effectiveness of a training program for double-raters

Kathrin Eberharter, Franz Holzknecht, Benjamin Kremmel

Language Testing Research Group Innsbruck

Symposium "Language Testing in Austria: Taking Stock", AAU Klagenfurt, October 2018

Background

- New Matura: Standardized tasks but rating still in hands of (untrained) teachers
- Project BMB, UIBK & PHT: Training double-raters
- Two cohorts: 2011 & 2013 (32 teachers)
- Chapter describes training and evaluates its effectiveness

- Need for rater training highlighted in literature
- Rater training can help clarify rating criteria, modify performance expectations of individual raters and provide a reference point in the rating process (Weigle, 1994)
- Literature about training effects inconclusive

Data

Course structure (1 module = 1 week)

- Module 1:
  - principles of language testing
  - CEFR descriptors for writing
  - scale familiarisation
  - task type requirements
  - emails and articles
- Module 2:
  - *report* and *essay*
- Module 3:
  - introduction MFRM
  - new task types (*blog*)
- Module 4:
  - Recap and focus on applying all four criteria

- Cohort 2: similar structure, but inverted order of criteria introduction and online module

	Cohort 1			Cohort 2		
Module	Task type	Criteria	No of scripts	task type	Criteria	No of scripts
1	article email	All TA OL LSR LSA *total*	2 2 2 1 1 *14*	article email	All LSA LSR *total*	1 5 4 *13*
2	article essay report	All TA OL LSR LSA *total*	2 6 6 6 6 *32*	article essay report	TA OL *total*	6 6 *12*
3	article essay report	All TA OL LSR LSA *total*	2 2 2 2 1 *15*	*(online)* article email essay report	All *total*	12 *48*
4	article email essay	All *total*	25 *100*	article email essay report	All TA OL LSR LSA *total*	4 3 3 3 3 *28*
		Full total	**161**		**Full total**	**101**

Three MFRM analyses

- Do cohorts 1 and 2 differ in their rating rater fit statistics?
- (Do cohorts 1 and 2 use the scale in a similar way?)
- Do cohorts 1 and 2 differ in their severity or leniency?

	Cohort 1 (N=15)		Cohort 2 (N=14)	
Fit Range	Infit	Outfit	Infit	Outfit
fit < 0.70 (overfit)	2	2	0	0
0.70 ≤ fit ≤ 1.30	11	11	12	12
fit > 1.30 (misfit)	2	2	2	2

Rater Fit

	Cohort 1 (N=15)	Cohort 2 (N=14)
Total score	249.9	245.6
Observed Avg.	6.85	6.78
Fair Avg.	6.86	6.8
Mean severity (logit)	-1.18	-1.13
Min M	-0.151	-1.6
Max M	-0.75	-0.56
Separation index	1.07	1.89

Rater Severity

Findings & implications

- Cohorts differ in their homogeneity:
  Cohort 1 > Cohort 2
- Majority of raters are predictable and independent experts at end of training:
  4 individuals remained unpredictable
- Challening training context for both cohorts, but positive training outcome
- Question of long-term effect remains