Differences in Raters' Severity, Consistency and Biased Interactions between Trained and Untrained Raters

Yoshihito SUGITA

Introduction

- The importance and effect of rater training (Shohamy, Gordon & Kraemer, 1992; Weigle, 1994)
- Rater training for a task-based writing performance test (TBWT)

1st stage: Promoting the understanding

2nd stage: Familiarizing the procedure

3rd stage: Increasing experiences

Purposes of this study

- The purpose of this comparative study is to investigate the degree of difference in:

(1) Raters' severity

(2) Consistency and biased interactions between trained and untrained teacher raters.

Participants

- Trained raters (TRN) consisted of five novice junior high school teachers, who received training session before rating.
- Untrained raters (UNTRN) consisted of five experienced teachers, who did not have training session beforehand.

Data Analysis

- Rater behavior both of TRN and UNTRN was modeled using FACETS
- Three facets were used:

(1) Subjects (n=20)

(2) Raters (five TRNs, five UNTRNs)

(3) Tasks (accuracy, communicability and impression)

The specification of assessment task 1

- You will have 20 minutes to complete the test. You are going to stay with the Parker Family in Britain this summer. Write a 100-120 word letter introducing yourself to your host family. Before writing, think of the following topics.
- Your name and age
- Your job and major in school

The specification of assessment task 1

- Your hobbies and interests
- Your family and pet
- Your favorite places, foods and activities
- Your experience traveling abroad
- Some things you want to doe while you are in Britain

Construct definitions of task 1

Accuracy	
Organizational skills	Linguistic accuracy
Organizational skills can be defined as ability to organize logical structure which enables the content to be accurately acquired	Linguistic accuracy concerns errors of vocabulary, spelling, punctuation or grammar

The specification of assessment task 2

- You will have 10 minutes to complete the test. You are going to discuss the following topic with your classmates, "Why do you study English?" In order to prepare for the discussion, think of as many answers as possible to the question and write them as "To travel abroad."

Construct definitions of task 2

Communicability	
Communicative quality	Communicative effect
Communicative quality refers to the ability to communicate without causing the reader any difficulty	Communicative effect concerns the quantity of ideas necessary to develop the response as well as the relevance of the content to the proposed task

Result: TRN raters

Raters	TRN1	TRN2	TRN3	TRN4	TRN5
Severity	-1.44	-2.58	-0.32	-1.06	-2.13
Error	0.25	0.26	0.25	0.25	0.25
Infit	0.82	1.29	0.75	0.79	0.64

Separation: 2.99 Reliability=.90; fixed (all same)
chi-square: 49.7, df:4; significance: .00

Result: UNTRN raters

Raters	UNT1	UNT2	UNT3	UNT4	UNT5
Severity	-2.06	-0.20	0.53	-1.19	-1.69
Error	0.25	0.25	0.25	0.25	0.25
Infit	1.34	1.11	1.35	0.90	0.77

Separation: 3.70 Reliability=.93; fixed (all same)
chi-square: 73.8, df:4; significance: .00

Result (1): Raters' severity

- Both TRN and UNTRN raters differs significantly in their severity
- UNTRN raters as a group vary much more in severity than TRN raters
- UNTRN raters tend to apply stricter standards overall to the written samples than TRN raters.

Result (2): Raters' consistency

- No raters were identified as misfitting (M-2*SD*<Infit<M+2*SD*)

  → Both TRN and UNTRN raters behaved consistently in scoring
- TRN raters: infit mean .86 (*SD* .22) UNTRN raters: infit mean 1.10 (*SD* .23)

  →UNTRN are supposed to be less consistent as a group

Result (3): Rater-subject bias interaction

Ability	N	Harsh (Raters)		Lenient (Raters)	
		TRN	UNT	TRN	UNT
3.00 higher	5	1	3		1
-2.99～2.99	9	3	4		2
-3.00 lower	3			3	

→ TRN raters were more lenient, and the UNTRN raters were more severe.
→ UNTRN raters might be more biased than TRN raters.

Result (3): Rater-task bias interaction

Rater (task)	Observed score	Expected score	Obs-Exp	Bias (logits)	Z-score
T2(a)	67	72.7	-.29	-1.15	-2.58

→ One rater (TRN2) shows significantly biased rate-task interaction, who awarded severe ratings to all subjects on accuracy task.

→ The fit value by UNTRN1 on communicability task was 1.8, which indicates the rater was not consistent in evaluating the task.

Implications for raters' severity

- Both TRN and UNTRN raters differ significantly in their severity

  → Rater training was not successful in getting raters to give identical scores.

  → The use of FACETS analysis is assumed to be effective in compensating for inter-rater differences.

Implications for raters' consistency

- There are differences in consistency between groups of TRN and UNTRN raters.

  → Rater training is effective in improving raters' consistency in scoring.

  → A shared understanding of the constructs of writing ability could be effectively promoted by training sessions

Implications for task difficulty

- There was only one interaction with a significant bias out of the 30 interactions.

  → Assessment tasks developed in this study may draw valid inference to Japanese learners' writing performance

Conclusion

- All raters as a group differ significantly from one another in terms of severity, and UNTRN raters showed the tendency to be more severe than TRN raters.
- UNTRN raters were more biased than TRN raters, showing the UNTRN raters' inconsistency in scoring.
- Rater training is, therefore, more effective in improving raters' consistency than in improving their severity in scoring.

References

- Shohamy, E., Gordon, C., & Kraemer, R. (1992). The effect of raters' background and training on the reliability of direct writing tests, *Modern Language Journal,* 76, 27-33.
- Weigle, S. (1994). Effects of training on raters of ESL compositions. *Language Testing* 11, 197-233