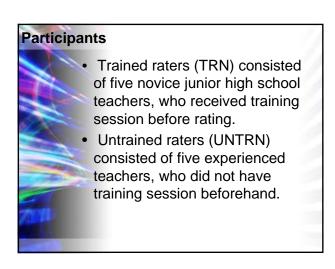
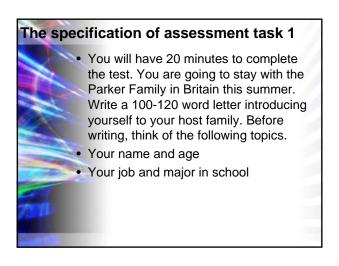


### The importance and effect of rater training (Shohamy, Gordon & Kraemer, 1992; Weigle, 1994) Rater training for a task-based writing performance test (TBWT) 1st stage: Promoting the understanding 2nd stage: Familiarizing the procedure 3rd stage: Increasing experiences

# 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.



## • 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 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

Ac	ccuracy
Organizational skills	Linguistic accuracy
Organizational skills can be defined as ab to organize logical structure which enable the content to be accurately acquired	vocabulary, spelling,

### 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."

Commu	nicability
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

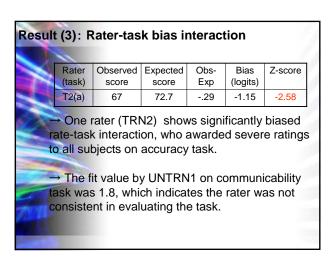
Rate	ers	TRN1	TRN2	TRN3	TRN4	TRN5
Sev	erity	-1.44	-2.58	-0.32	-1.06	-2.13
Erro	or	0.25	0.26	0.25	0.25	0.25
Infit		0.82	1.29	0.75	0.79	0.64

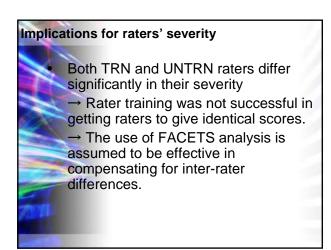
Ļ	Raters	UNT1	UNT2	UNT3	UNT4	UNT5
Ċ	Severity	-2.06	-0.20	0.53	-1.19	-1.69
	Error	0.25	0.25	0.25	0.25	0.25
	Infit	1.34	1.11	1.35	0.90	0.77

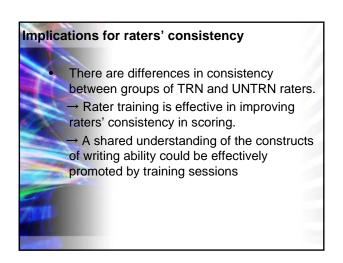
### Both TRN and UNTRN raters differs significantly in their severity UNTRN raters as a group vary much more in severity than TRN raters UNTRN raters tend to apply stricter standards overall to the written samples than TRN raters.

# Result (2): Raters' consistency No raters were identified as misfitting (M-2SD<Infit<M+2SD) 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

Ability	N	Harsh (	Raters)	Lenient	(Raters
		TRN	UNT	TRN	UNT
3.00 higher	5	1	3		1
<b>-2</b> .99~2.99	9	3	4		2
-3.00 lower	3			3	







### 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