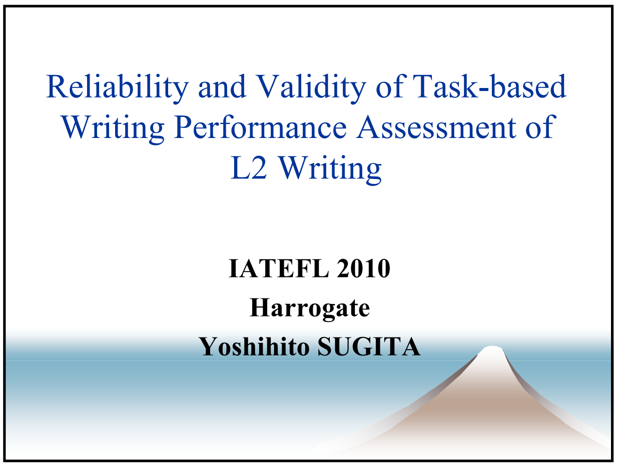
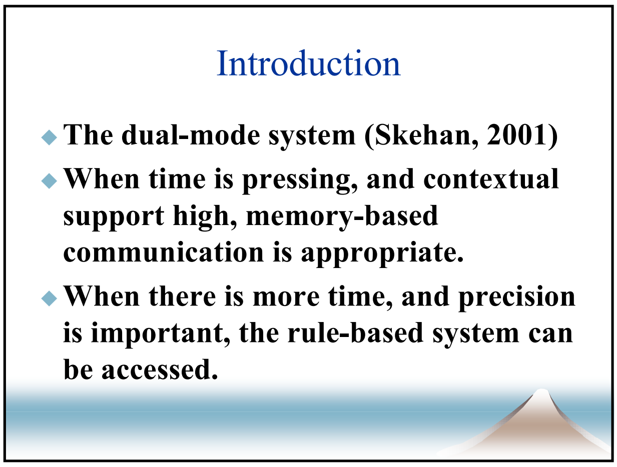


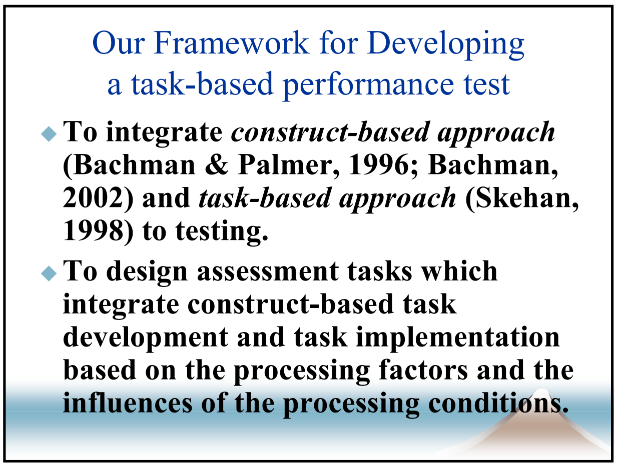
Reliability and Validity of Task-based Writing Performance Assessment of L2 Writing

IATEFL 2010
Harrogate
Yoshihito SUGITA



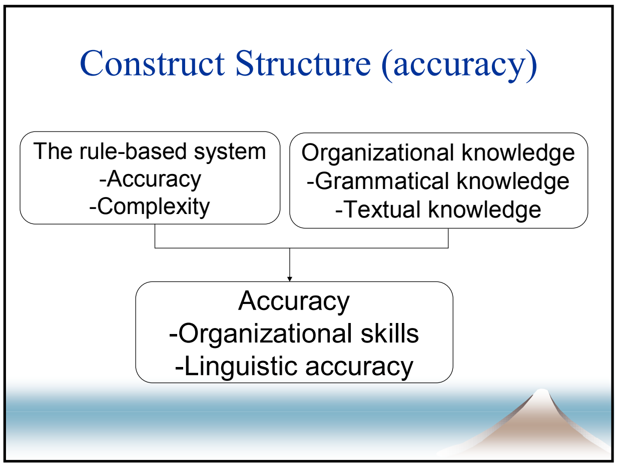
Introduction

- **The dual-mode system (Skehan, 2001)**
- **When time is pressing, and contextual support high, memory-based communication is appropriate.**
- **When there is more time, and precision is important, the rule-based system can be accessed.**



Our Framework for Developing a task-based performance test

- **To integrate *construct-based approach* (Bachman & Palmer, 1996; Bachman, 2002) and *task-based approach* (Skehan, 1998) to testing.**
- **To design assessment tasks which integrate construct-based task development and task implementation based on the processing factors and the influences of the processing conditions.**

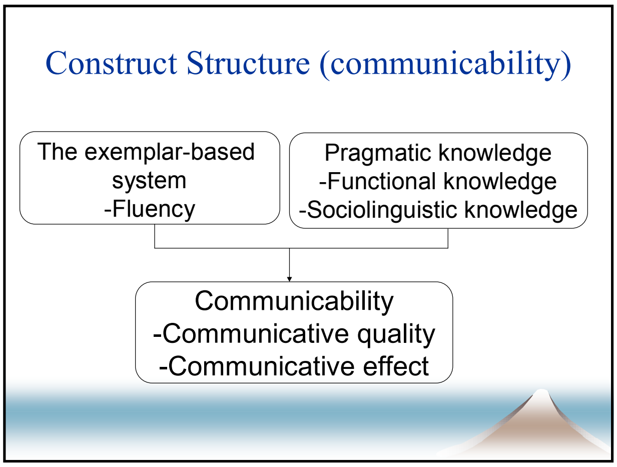


Construct Structure (accuracy)

The rule-based system
-Accuracy
-Complexity

Organizational knowledge
-Grammatical knowledge
-Textual knowledge

Accuracy
-Organizational skills
-Linguistic accuracy

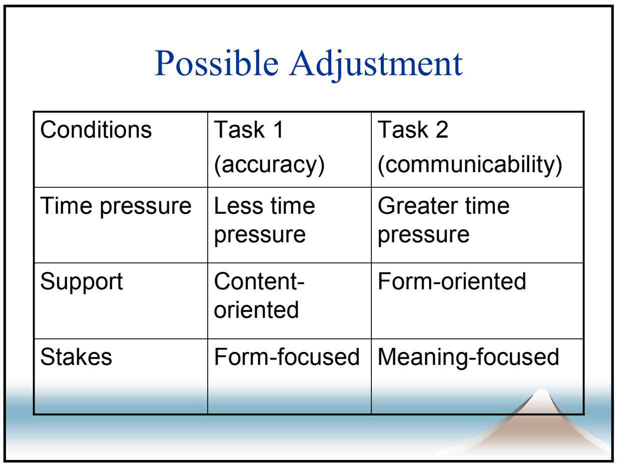


Construct Structure (communicability)

The exemplar-based system
-Fluency

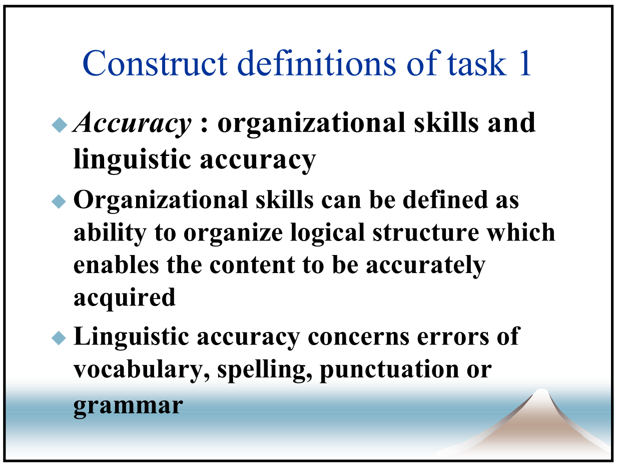
Pragmatic knowledge
-Functional knowledge
-Sociolinguistic knowledge

Communicability
-Communicative quality
-Communicative effect



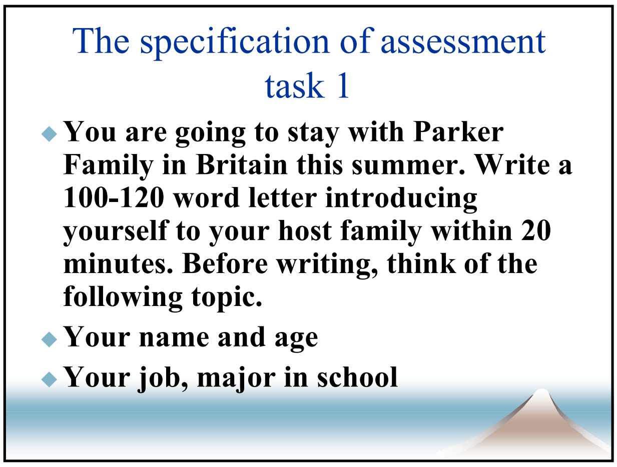
Possible Adjustment

Conditions	Task 1 (accuracy)	Task 2 (communicability)
Time pressure	Less time pressure	Greater time pressure
Support	Content-oriented	Form-oriented
Stakes	Form-focused	Meaning-focused



Construct definitions of task 1

- *Accuracy* : organizational skills and 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

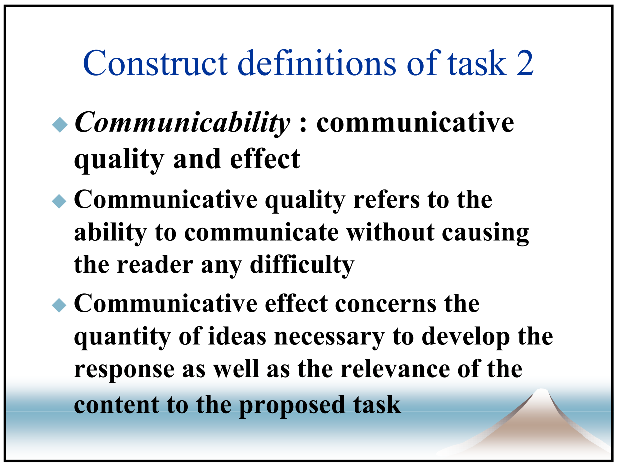


The specification of assessment task 1

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

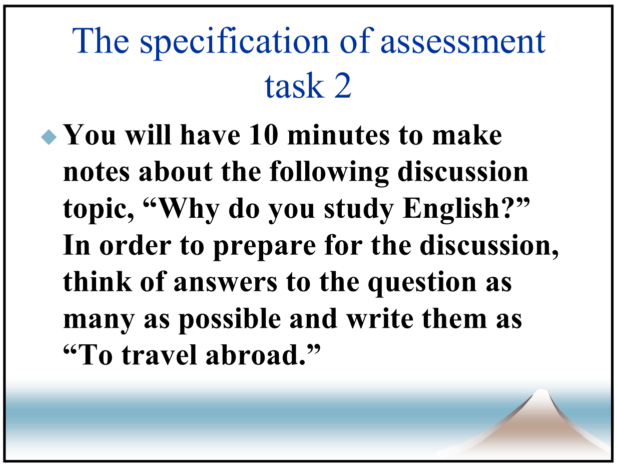
The specification of assessment task 1

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



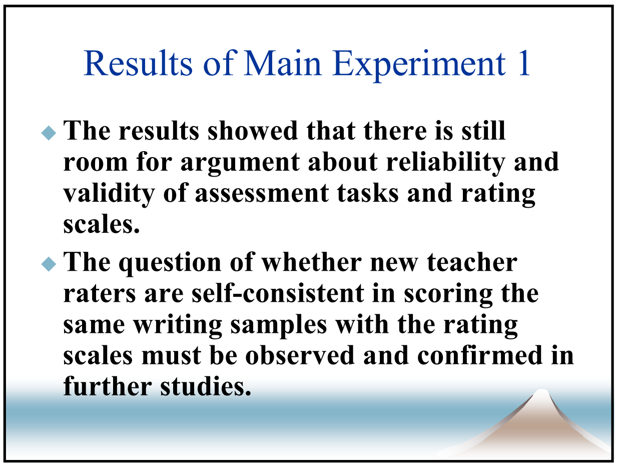
Construct definitions of task 2

- ***Communicability*** : communicative quality and 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



The specification of assessment task 2

- You will have 10 minutes to make notes about the following discussion topic, "Why do you study English?" In order to prepare for the discussion, think of answers to the question as many as possible and write them as "To travel abroad."



Results of Main Experiment 1

- The results showed that there is still room for argument about reliability and validity of assessment tasks and rating scales.
- The question of whether new teacher raters are self-consistent in scoring the same writing samples with the rating scales must be observed and confirmed in further studies.

Purposes

- In order to examine the degree of reliability and validity of the task-based writing performance test, the following are focused on:
  (1) Raters' severity
  (2) Interactions with writers' abilities and task difficulties
  (3) Reliability of tasks and rating scales
  (4) Measure's validity

Test participants and materials

- In Main Experiment 1, 20 undergraduate students (6 males and 14 females), native speakers of Japanese with an intermediate level of English proficiency took the task-based writing performance test (Tasks 1 & 2).
- Each student wrote and submitted an essay using a web-based essay evaluation service, *Criterion*.

Scoring Materials and Procedure

- Each of 40 identical scripts used in Main Experiment 1 was scored
- Five experienced junior high school teachers participated as a novice rater
- Both scripts and scoring guidelines were given by mail

Procedures for rating

- Steps:

Rate the 20 scripts of task 1
↓
Rate the 20 scripts of task 2
↓
Impressionistic scoring
↓
Reply to the questionnaire

Description of Accuracy (1)

Organizational skills
The written text -is well organized and well developed (TWE). -shows strong rhetorical control and is well managed (MWA). -has clear organization with a variety of linking devices (FCE).

Description of Accuracy (2)

Linguistic accuracy
The written text -demonstrates appropriate word choice though it may have occasional errors (TWE). -has few errors of agreement, tense, number, word order/function, articles, pronouns, prepositions, spelling, punctuation, capitalization, paragraphing (ESL).

Description of Communicability (1)

Communicative quality
The written text -displays consistent facility in use of language (TWE). -contains well-chosen vocabulary to express the ideas and to carry out the intentions (MWA).

Description of Communicability (2)

Communicative effect
The written text -effectively addresses the writing task (TWE). -has a very positive effect on the target reader with adequately organized relevant ideas (FCE).

5-point Likert Scale for Rating

A (5) : I strongly agree to assign the above description

B+ (4) : I partially agree to assign the above description

B (3) : I agree to assign the above description

B- (2) : I disagree with assigning the above description

C (1) : I strongly disagree to assign the above description

Data Analysis

- **The data were analyzed using FACETS (Linacre, 2008).**
- **To examine the measurement characteristics of the testing, three facets were specified: subject, rater and task.**
- **Bias analysis: Rater × Subjects**
  Rater × Tasks

Result (1) : Raters

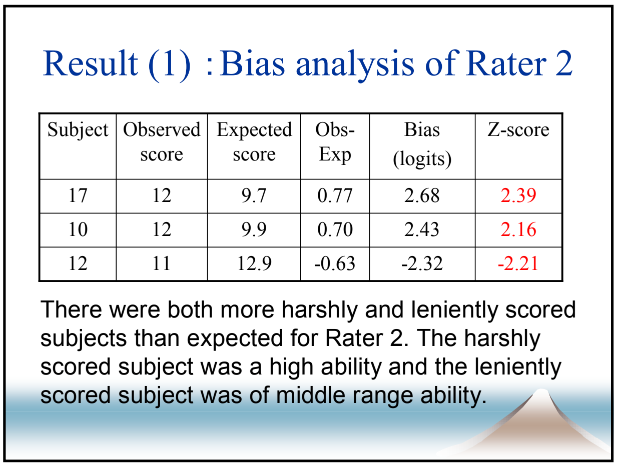
Raters	1	2	3	4	5
Severity	1.07	1.35	-0.09	0.59	1.70
Error	0.27	0.27	0.26	0.26	0.27
Infit	0.63	0.90	0.91	1.25	0.96

→There was a significant difference in severity among raters, but all raters behaved consistently in the scoring.

Result (1) : Bias analysis of Rater 1

Subject	Observed score	Expected score	Obs-Exp	Bias (logits)	Z-score
1	12	9.7	0.78	2.71	2.41

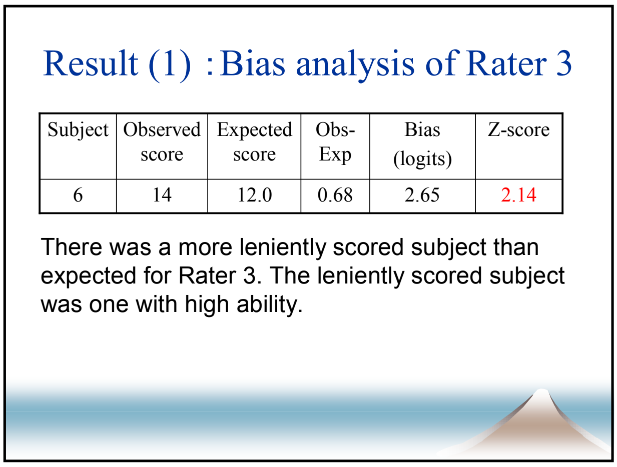
There was a more leniently scored subject than expected for Rater 1. The leniently scored subject was of middle range ability.



Result (1) : Bias analysis of Rater 2

Subject	Observed score	Expected score	Obs-Exp	Bias (logits)	Z-score
17	12	9.7	0.77	2.68	2.39
10	12	9.9	0.70	2.43	2.16
12	11	12.9	-0.63	-2.32	-2.21

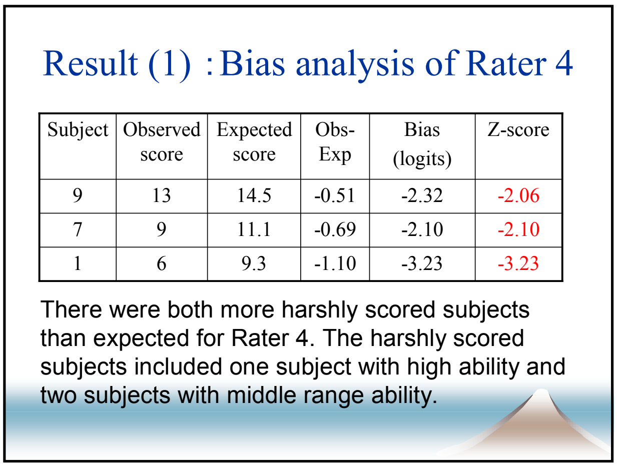
There were both more harshly and leniently scored subjects than expected for Rater 2. The harshly scored subject was a high ability and the leniently scored subject was of middle range ability.



Result (1) : Bias analysis of Rater 3

Subject	Observed score	Expected score	Obs-Exp	Bias (logits)	Z-score
6	14	12.0	0.68	2.65	2.14

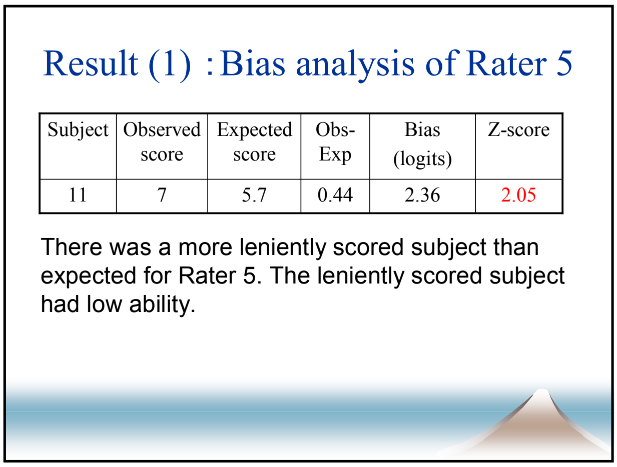
There was a more leniently scored subject than expected for Rater 3. The leniently scored subject was one with high ability.



Result (1) : Bias analysis of Rater 4

Subject	Observed score	Expected score	Obs-Exp	Bias (logits)	Z-score
9	13	14.5	-0.51	-2.32	-2.06
7	9	11.1	-0.69	-2.10	-2.10
1	6	9.3	-1.10	-3.23	-3.23

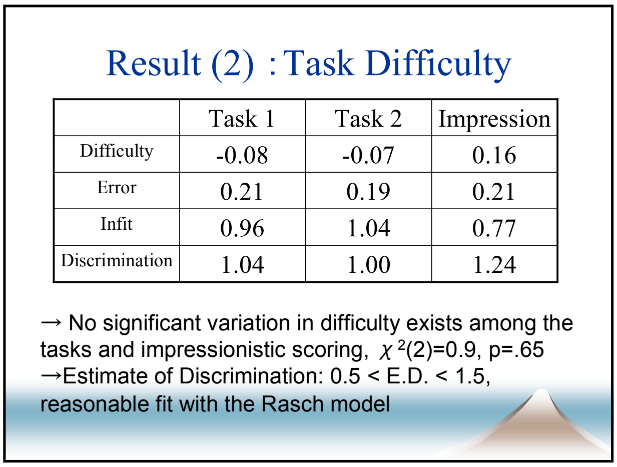
There were both more harshly scored subjects than expected for Rater 4. The harshly scored subjects included one subject with high ability and two subjects with middle range ability.



Result (1) : Bias analysis of Rater 5

Subject	Observed score	Expected score	Obs-Exp	Bias (logits)	Z-score
11	7	5.7	0.44	2.36	2.05

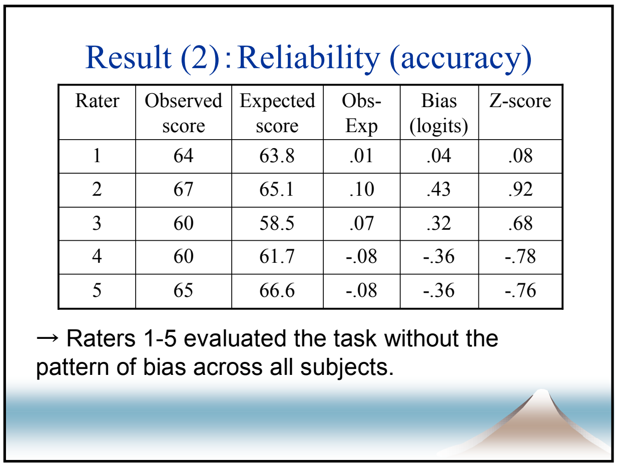
There was a more leniently scored subject than expected for Rater 5. The leniently scored subject had low ability.



Result (2) : Task Difficulty

	Task 1	Task 2	Impression
Difficulty	-0.08	-0.07	0.16
Error	0.21	0.19	0.21
Infit	0.96	1.04	0.77
Discrimination	1.04	1.00	1.24

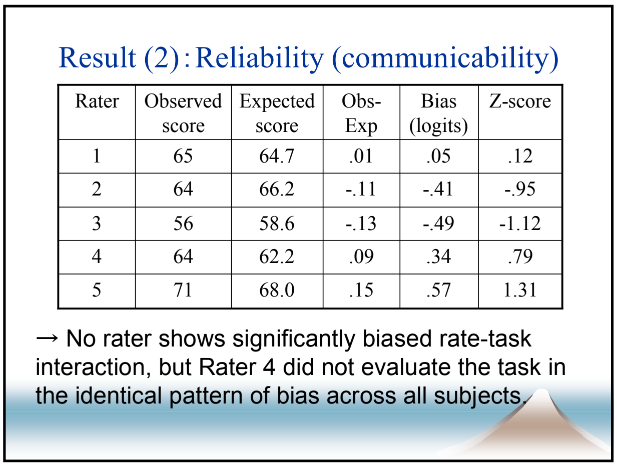
→ No significant variation in difficulty exists among the tasks and impressionistic scoring, $\chi^2(2)=0.9$, p=.65
→Estimate of Discrimination: 0.5 < E.D. < 1.5, reasonable fit with the Rasch model



Result (2) : Reliability (accuracy)

Rater	Observed score	Expected score	Obs-Exp	Bias (logits)	Z-score
1	64	63.8	.01	.04	.08
2	67	65.1	.10	.43	.92
3	60	58.5	.07	.32	.68
4	60	61.7	-.08	-.36	-.78
5	65	66.6	-.08	-.36	-.76

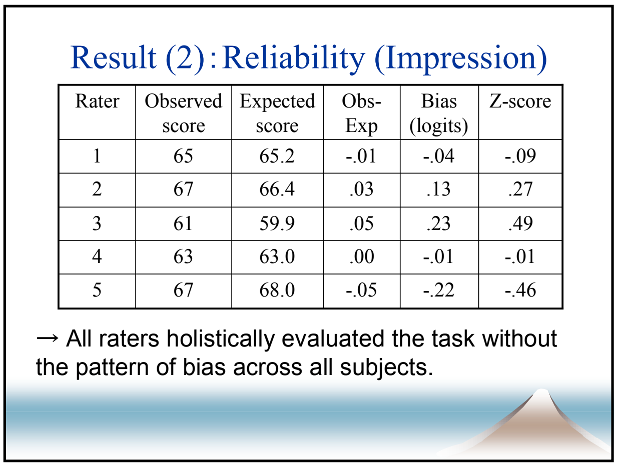
→ Raters 1-5 evaluated the task without the pattern of bias across all subjects.



Result (2): Reliability (communicability)

Rater	Observed score	Expected score	Obs-Exp	Bias (logits)	Z-score
1	65	64.7	.01	.05	.12
2	64	66.2	-.11	-.41	-.95
3	56	58.6	-.13	-.49	-1.12
4	64	62.2	.09	.34	.79
5	71	68.0	.15	.57	1.31

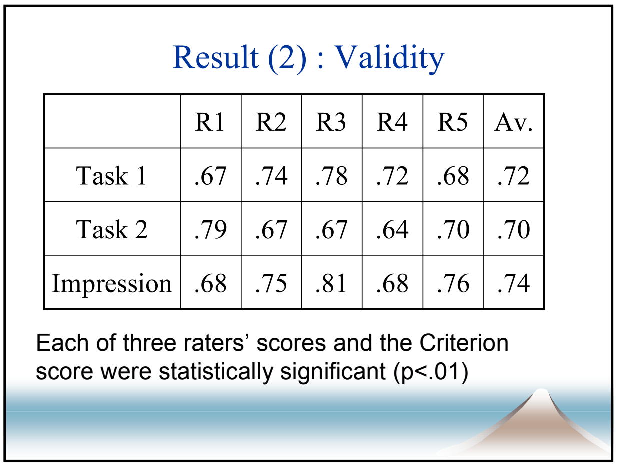
→ No rater shows significantly biased rate-task interaction, but Rater 4 did not evaluate the task in the identical pattern of bias across all subjects.



Result (2): Reliability (Impression)

Rater	Observed score	Expected score	Obs-Exp	Bias (logits)	Z-score
1	65	65.2	-.01	-.04	-.09
2	67	66.4	.03	.13	.27
3	61	59.9	.05	.23	.49
4	63	63.0	.00	-.01	-.01
5	67	68.0	-.05	-.22	-.46

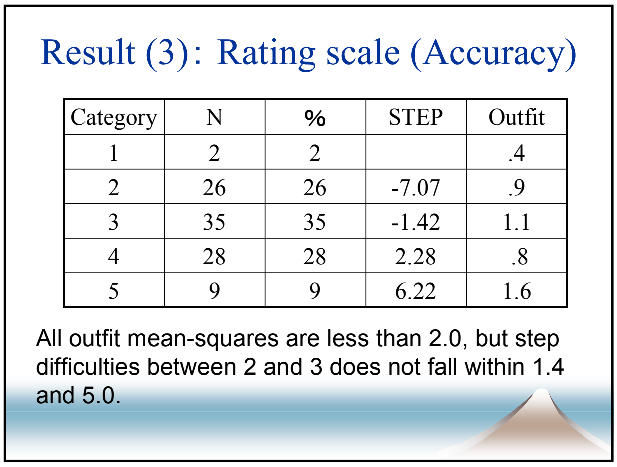
→ All raters holistically evaluated the task without the pattern of bias across all subjects.



Result (2) : Validity

	R1	R2	R3	R4	R5	Av.
Task 1	.67	.74	.78	.72	.68	.72
Task 2	.79	.67	.67	.64	.70	.70
Impression	.68	.75	.81	.68	.76	.74

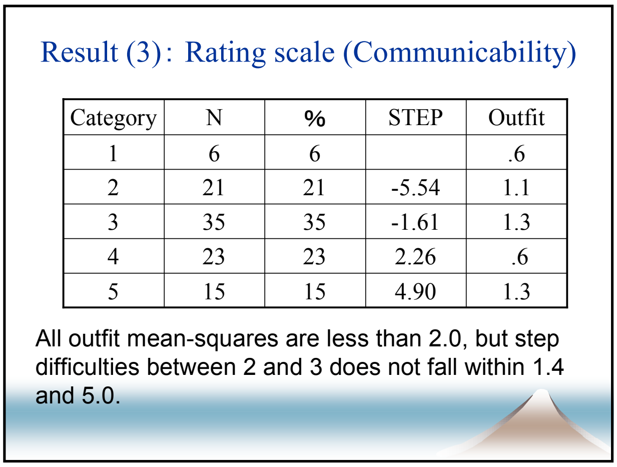
Each of three raters' scores and the Criterion score were statistically significant (p<.01)



Result (3): Rating scale (Accuracy)

Category	N	%	STEP	Outfit
1	2	2		.4
2	26	26	-7.07	.9
3	35	35	-1.42	1.1
4	28	28	2.28	.8
5	9	9	6.22	1.6

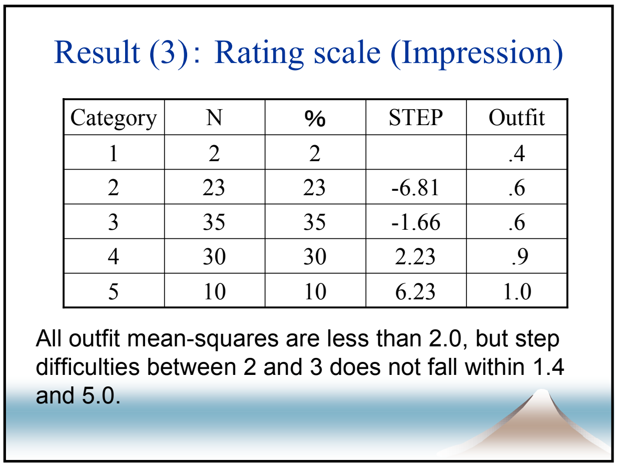
All outfit mean-squares are less than 2.0, but step difficulties between 2 and 3 does not fall within 1.4 and 5.0.



Result (3): Rating scale (Communicability)

Category	N	%	STEP	Outfit
1	6	6		.6
2	21	21	-5.54	1.1
3	35	35	-1.61	1.3
4	23	23	2.26	.6
5	15	15	4.90	1.3

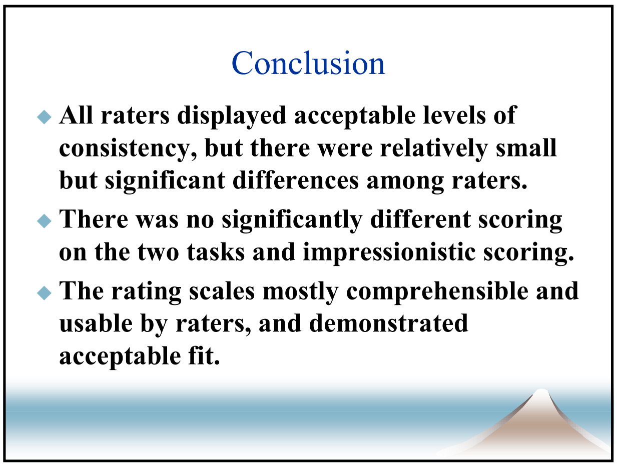
All outfit mean-squares are less than 2.0, but step difficulties between 2 and 3 does not fall within 1.4 and 5.0.



Result (3): Rating scale (Impression)

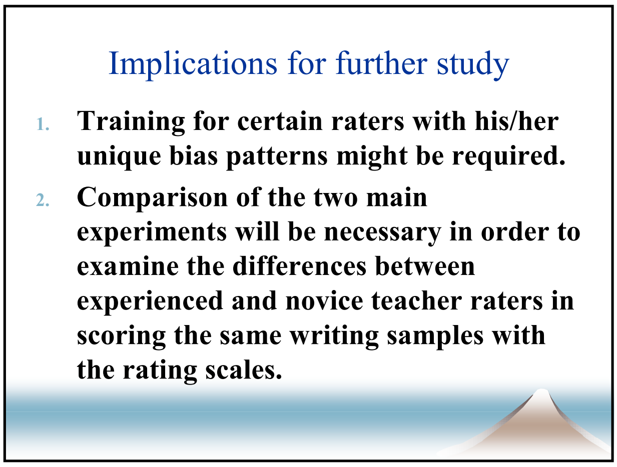
Category	N	%	STEP	Outfit
1	2	2		.4
2	23	23	-6.81	.6
3	35	35	-1.66	.6
4	30	30	2.23	.9
5	10	10	6.23	1.0

All outfit mean-squares are less than 2.0, but step difficulties between 2 and 3 does not fall within 1.4 and 5.0.



Conclusion

- All raters displayed acceptable levels of consistency, but there were relatively small but significant differences among raters.
- There was no significantly different scoring on the two tasks and impressionistic scoring.
- The rating scales mostly comprehensible and usable by raters, and demonstrated acceptable fit.



Implications for further study

1. Training for certain raters with his/her unique bias patterns might be required.
2. Comparison of the two main experiments will be necessary in order to examine the differences between experienced and novice teacher raters in scoring the same writing samples with the rating scales.