ALTE 5th International Conference, Paris, April 2014

Examining Test Fairness: Focusing on Rater Behavior in the Case of Employing Prospective English Teachers

Department of English Language & Literature, Bunkyo University, Japan
Tomoyasu Akiyama (Ph.D.)
E-mail: akitomo@koshigaya.bunkyo.ac.jp

Outline of the presentation

1. Background information

(1) What are Teacher Employment of Examinations (TEEs)?

(2) When are the TEEs administered?

(3) How competitive were TEEs in 2013?

2. Purpose of this study

3. Data collection methods

4. Results & Conclusion

5. Implications for further studies

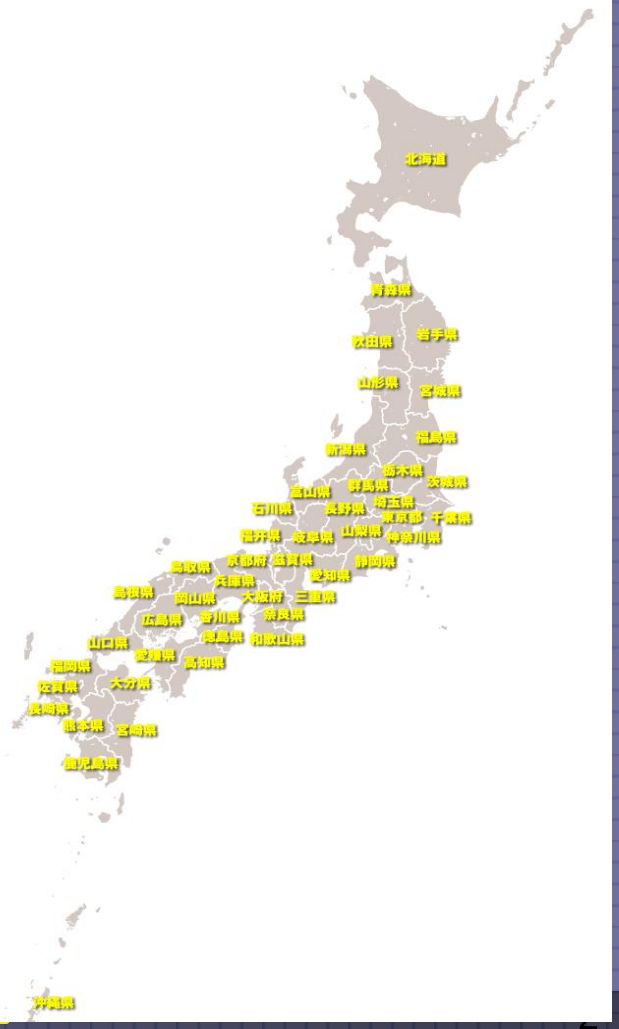
Background information(1): What are Teacher Employment of Examinations(TEEs)?

1. TEEs are tests for those who want to be teachers at the public elementary, junior and senior high schools in Japan.

2. TEEs are developed and administered by 47 prefectures and 21 cities at each local board of education.

3. Approximately 180,000 test-takers undertook the TEEs in 2013.

4. TEEs are norm-referenced assessments based on the needs of the prefecture.



Background information (2): test administration schedule

1st test (knowledge-based tests)

July — **Paper-and-pencil tests**

2nd test (performance-based tests)

August～September — **Interview & 'microteaching'**

Final results (employment decisions made)

September ～January — **formally –appointed applicants**

Background information(3):How competitive are teacher employment examinations in 2013?

School Types	Applicants (N)	1st & 2nd tests Successful Applicants (N)	Average competition rate
Elementary	58,703	13,626	4.3
Junior high	62,998	8,383	7.5
Senior high	37,812	4,912	7.7
Junior high School English Teachers			
Tokyo	**331**	**52**	**6.4**
Aomori prefecture	**83**	**14**	**20.8**

The purpose of this study

- To investigate how 12 raters (6 management educators, such as local education board members principals, heads of English departments and 6 junior English teachers) rate test-takers (teacher candidates), using Multi-facet Rasch Analysis (quantitatively) and the think-aloud method (qualitatively), focusing on interactions between raters and the test-takers.

Research participants: test takers & raters

Test takers : Thirty (20 university students and 10 in-service English teachers) participated in this study. They were required to demonstrate one of the target grammar points (-ed, there are~, Can you ~?) for 5 minutes.

Raters : 12 raters rated all students individually watching videotaped teaching performances.

(6 junior high school English teachers: J1 to J6)

(6 management educators : M1 to M6)

A typical procedure of microteaching (demonstration teaching) in TEEs

1st stage

Preparation stage

(20 minutes)

A candidate is required to design a teaching plan.

(e.g., Introduction of target grammar)

2nd stage Performance and rating stage (microteaching)
(5-10 minutes)

The candidate is asked to demonstrate her/his teaching skills, based on the teaching plan.

2 or 3 assessors (educational boards, principals or English teachers) observe candidates' performance) and the assessors rate candidates' performances using assessment criteria.

Assessment criteria & assessment sheet

Assessment criteria	Examples of description	1	2	3	4	5
1. Organization of a lesson	・Making learning goals clear and the flow of a lesson smooth		✓			
2. Instruction ability	・Making content comprehensible to students ・suitability of content and language for learners			✓		
3. Delivery	・Adequate eye contact ・Vocally expressive ・appropriate movement and gestures		✓			
4. Personality	・Be full of enthusiasm ・Be able to establish rapport with students					
5 .Expertise	・Adequate use of teaching materials ・Understanding of grammatical accuracy / knowledge		✓			
6. Overall (Holistic impression)	Do you want to employ this candidate as an English teacher?			✓		

Likert-type (strongly disagree-strongly agree)

Descriptive statistics (1) : agreement and disagreement among 12 raters

Junior high school English teachers

Rater	Average	Maximum (/30) Minimum (/6)
J1	21.3	26 (ID. 25) 17(IDs. 15,17,26)
J2	21.9	30 (IDs. 4, 11) 18 (IDs. 15, 19, 20, 21,23)
J3	21.2	28 (ID. 11) 13 (ID.26)
J4	21.8	30 (IDs. 5, 11, 12, 13) 9 (ID. 17)
J5	17.5	24 (ID. 16) 9 (ID. 17)
J6	21.9	29 (IDs. 6, 10) 16 (IDs. 26, 30)

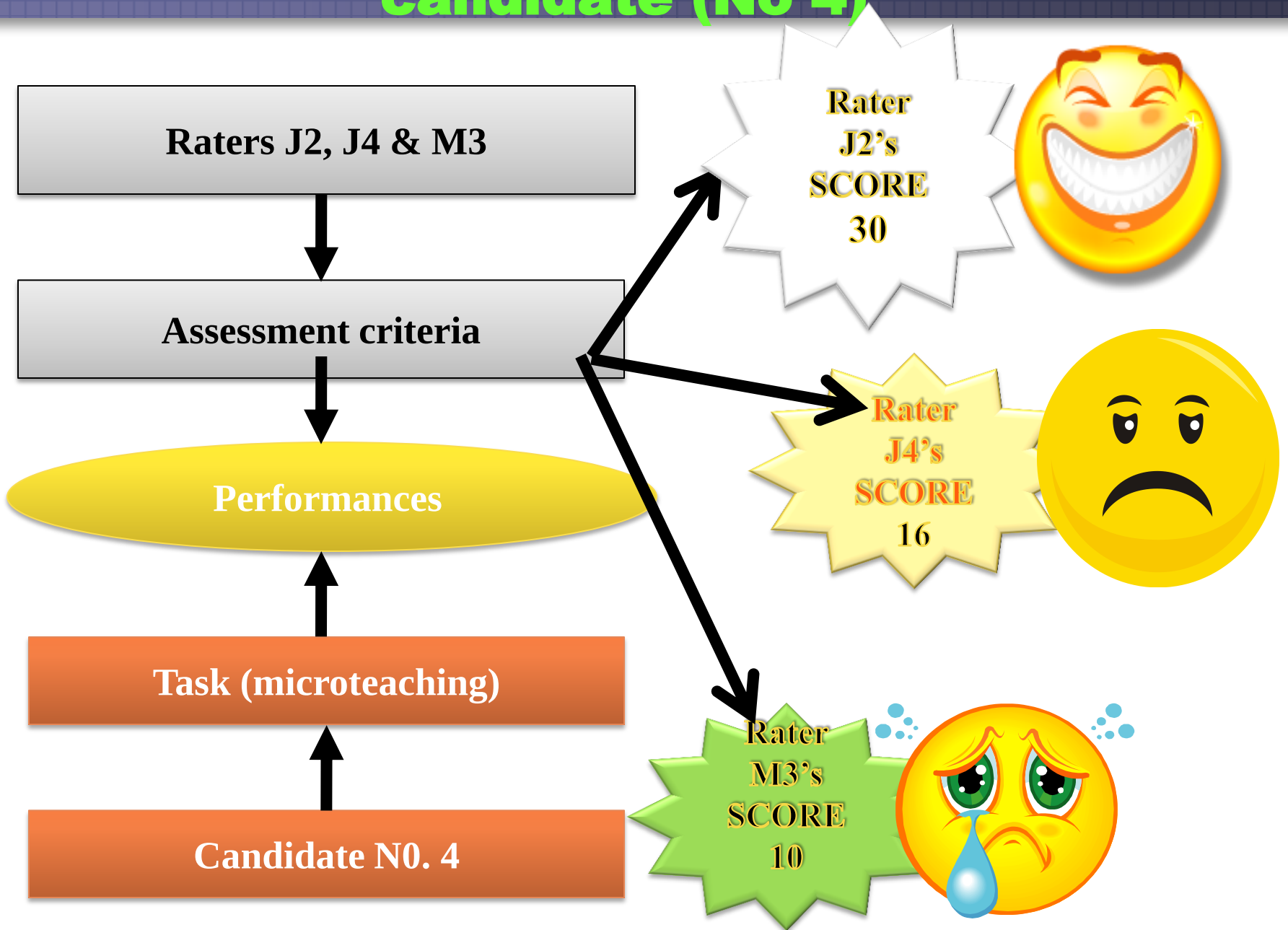
Management educators

Rater	Average	Maximum (/30) Minimum (/6)
M1	21.1	30 (ID. 5) 13 (ID. 20)
M2	15.3	22 (IDs. 2, 16, 23) 9 (IDs. 15)
M3	18.6	28 (ID. 24) 8 (ID. 17)
M4	21.0	29 (ID. 10, 22) 15 (ID. 19)
M5	20.8	28 (ID. 28) 14 (ID. 15)
M6	17.8	27 (IDs. 10, 16) 11(IDs. 9, 27)

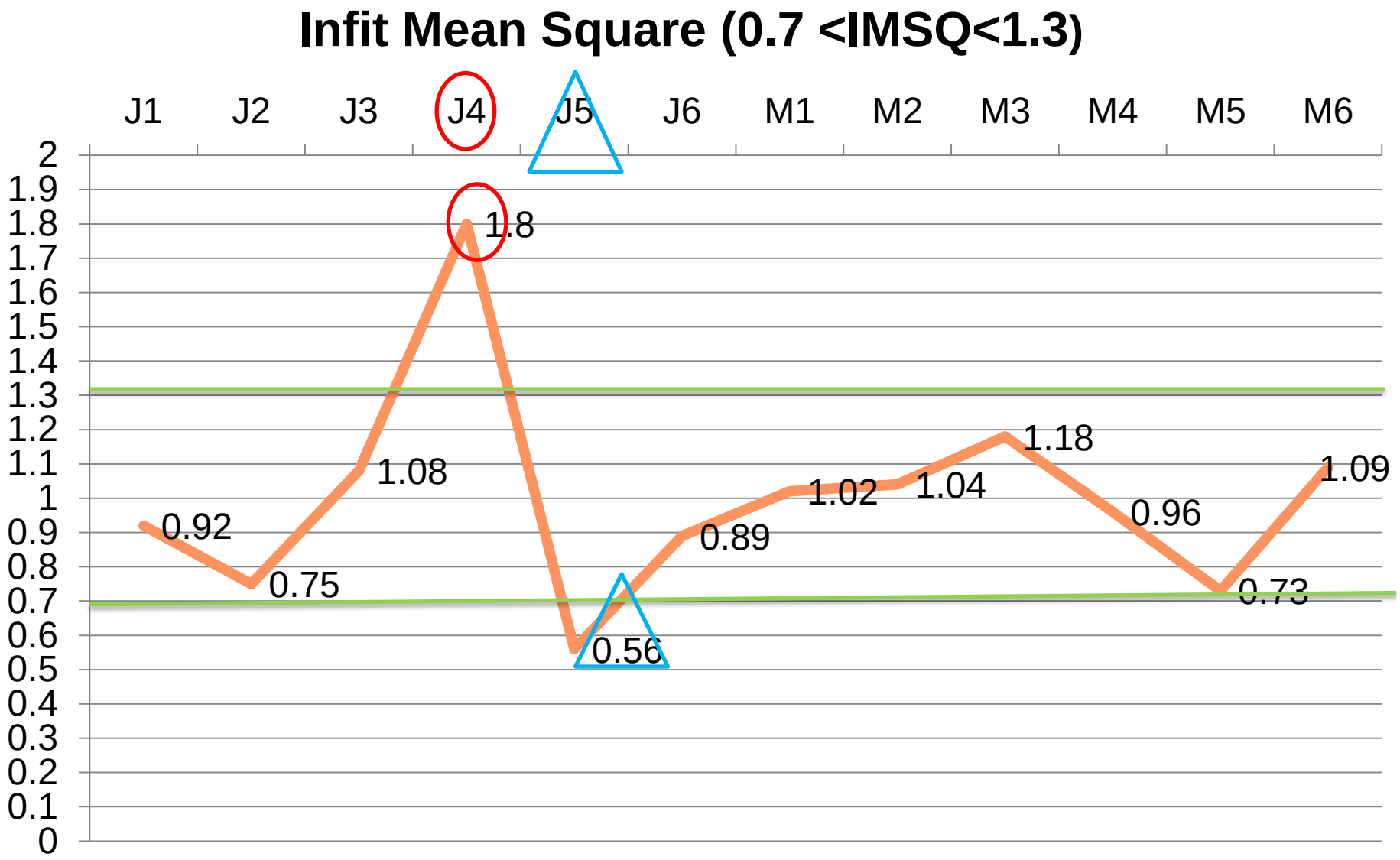
Descriptive statistics (2) : Three Largest differences among 12 raters

	ID (No. 4) 20 points difference	ID (No. 8) 18 points difference	ID (No.13) 16 points difference
J1	23	24	20
J2	30	25	19
J3	26	21	20
J4	16	28	30
J5	16	13	18
J6	21	22	17
M1	20	15	23
M2	18	10	14
M3	10	20	22
M4	28	17	18
M5	19	19	20
M6	18	16	16

Interactions between three raters and candidate (No 4)



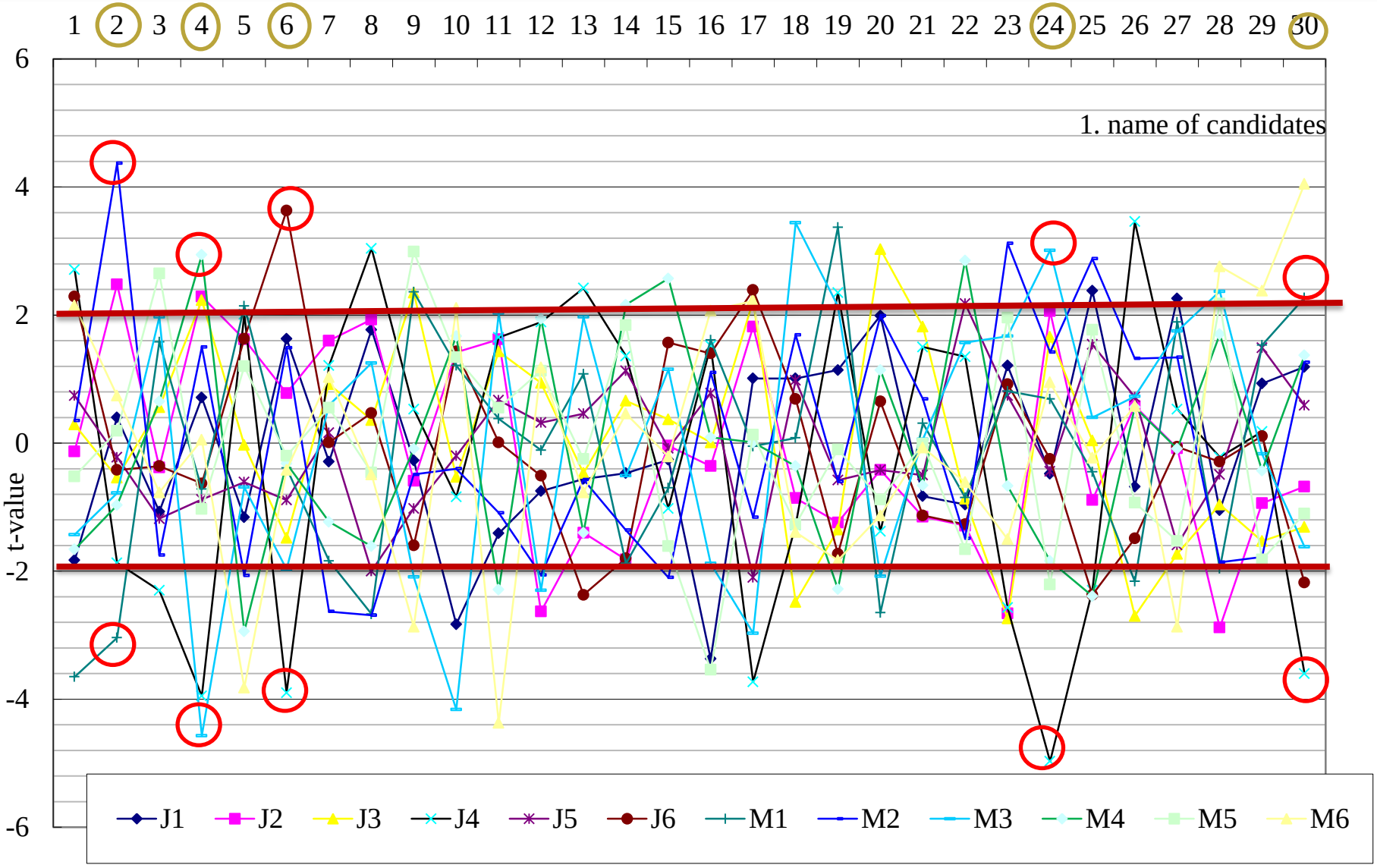
MFRMs(Multi-Faceted Rasch models) analysis
Rater consistency



Bias analysis(1) Interactions between 12 raters and 30 test-takers (MFRMs analysis)

Number of combination (30 candidates × 12 raters)	360 interactions
Number of biased interactions (-2≧t or 2≦t) (J= 4, J2= 6, J3= 7, J4=14, J5=3, J6=6) (M1=9, M2=8, M3=11, M4= 8, M5=5, M6=11)	92 (biased interactions)
% of large t-values(-2≧t or 2≦t)	25.6%

Bias analysis (2) Five largest biased differences among candidates in terms of t-values



Rater J4's comments on the test-taker 24 (t-value, 2.24): Observed score, 14 Expected score, 24.3

Assessment criteria	Raw scores	Comments
Organization of a lesson	1	The introduction of the key sentence was not good at all. His point was very unclear and it took a long time to introduce the key sentence. He should have done it simpler.
Instruction ability	3	Perhaps, the idea was interesting to students but it did not necessarily mean that the students were not interested in the key sentence .
Delivery	3	He looked confident in a dignified manner, but he made a lot of grammatical mistakes.
Personality	3	I can feel his passion for teaching, but his passion went into the wrong direction. He should have done it simpler.
Expertise	2	He introduced 'there are' without prepositions. I have taught there are with prepositions. He should have done more simply.
Overall	2	I can feel his passion for becoming a teacher, but he missed the point (the introduction of the target grammar) .

Rater J2's comments on the test-taker 24 (t-value, 2.06): Observed score, 29 Expected score, 24.5

Assessment criteria	Raw scores	Comments
Organization of a lesson	5	It seemed that he was aware of the structure of the class and he got his students to pay attention to the key sentence by using a game.
Instruction ability	5	He paid much attention to his students all the time. He took much time for his students to answer his questions.
Delivery	4	His voice, eye contact and gestures are good. However, his pronunciation was not good.
Personality	5	His friendly attitudes could attract his students. His questions for this class reflect his passion for teaching.
Expertise	5	He made a good preparation for this class. Also he used the blackboard and picture cards effectively and appropriately.
Overall	5	I think that he can become a good English teacher if he teaches English in this way.

Rater M3's comments on the test-taker 24 (t-value, 3.01): Observed score, 28 Expected score, 21

Assessment criteria	Raw scores	Comments
Organization of a lesson	4	Very good. However, he wrote it on the blackboard before he could get his students to practice the key sentence.
Instruction ability	4	He paid attention to his students all the time. He gave time to think what the key sentence is. However, he did not get his students to practice the key sentence although the students wanted to use the key sentence.
Delivery	5	His voice, eye contact and gestures are good.
Personality	5	His attitude could attract his students.
Expertise	5	His pictures are good.
Overall	5	His lesson still left something to be desired. But I guess that he could compensate for it later.

Rater M5's comments on the test-taker 24 (t-value, -2.21): Observed score, 19 Expected score, 23.3

Assessment criteria	Raw scores	Comments
Organization of a lesson	3	It was an interesting introduction of the key sentence, which increased students' motivation.
Instruction ability	3	However, the students paid more attention to "how many blocks there are" than "what today's key sentence is".
Delivery	4	
Personality	3	
Expertise	3	Also his classroom English was inaccurate.
Overall	3	

Rater J2's comments on the test-taker 4(t- value, 2.29): Observed score, 30 expected score 22.3

Assessment criteria	Raw scores	Comments
Organization of a lesson	5	Her lesson flow was well-organized and smooth.
Instruction ability	5	It appeared that she looked at students' reactions and responded to them very well.
Delivery	5	Her voice was clear and her intonation and accent were very good.
Personality	5	She looked friendly to students and she had a passion for teaching and her students.
Expertise	5	She used the blackboard and gave her questions to students effectively. Although she gave them too much information about the target grammar in Japanese, her explanation was understandable.
Overall	5	It seemed that she was used to teaching. Well, she was a work-ready candidate.

Rater M3's comments on the test-taker 4 (t-value core, -4.57): Observed score, 10 expected score 18.7

Assessment criteria	Raw scores	Comments
Organization of a lesson	1	She started with 'Do you~?' in order to teach "can you ~?" This was not smooth and her students must have been confused. Also she used the blackboard right after this. She should have let her students practice "can you".
Instruction ability	1	She used Japanese explanations a lot. This caused a lot of confusion.
Delivery	2	She was teaching while looking at the blackboard, not at her students.
Personality	2	Her explanations were not clear, which did not attract the students at all.
Expertise	2	It was impossible to teach by combining "Do you ?" with "Can you?" This was a fatal error.
Overall	2	Her score on this part was 2 points because her explanations about the target grammar were not clear.

Rater M4's comments on the test-taker 4 (t-value, 2.94): Observed score, 28 expected score 21.3

Assessment criteria	Raw scores	Comments
Organization of a lesson	4	She was teaching the key sentence slowly and clearly.
Instruction ability	5	It seemed that she was very aware of her students' despondences and checked their understanding.
Delivery	5	She talked to her students in a careful manner. It was very good. She used gestures and a loud voice.
Personality	5	She faced the students and talked to her students in a careful manner.
Expertise	5	She gave clear explanations to the students and she actually got the students to practice the key sentence.
Overall	4	She looked confident in teaching skills.

Rater J3's comments on the test-taker 4 (t-value, 2.23): Observed score, 26 expected score 21.4

Assessment criteria	Raw scores	Comments
Organization of a lesson	3	If she had made a clear distinction between 'can' and 'cannot', she would have got a higher score. I suggest that she should have used a game, which could have interested her students more. She used the classroom effectively. Her pronunciation was clear and beautiful.
Instruction ability	5	
Delivery	4	
Personality	5	
Expertise	4	
Overall	5	

Summary of the results

- 1. Most raters rated the candidates consistently but rated them with different levels of severity.
- There were more than 25 % biased interactions between the raters and the candidates, which suggests that we cannot delete the possibility that a candidate may fail due to a rater .
- It seemed that the all raters used assessment criteria differently, putting their own interpretations on the criteria. Also, they used their own assessment criteria which were not included in the provided assessment criteria.
- Biased interactions may have happened when test-takers' teaching demonstrations were not harmonious with a rater's ideal teaching image or teaching core values.

Values and language testing and rater behavior

Language testing occurs in an educational and social setting, and the uses of language tests are determined largely by political needs that change over time and vary from one society to another. …We must consider the value systems that inform test use – values of test developers, test-takers, test users, the educational system, and society at large. (Bachman, 1990, p291)

Implications for further study

Lumley (2005, p306) points out:

The important point here is that lack of mention of a particular feature or features by a rater is no indication that the feature was not observed and noted. Raters explicitly make the point that far more passes through their minds than they can ever articulate.

Implications

- To Investigate rater cognition study from multiple perspectives using rater interviews and questionnaires (Bejar, 2012; Knock, 2011). Also, rater behavior is to be investigated using "Wearable technology" in order to check their eye movements and brain waves.
- What factors account for the differences in terms of rater severity or leniency, and rater biases when they assess test-takers' performances in TEEs?

References

Bejar, D. (2012). Rater cognition: Implications for validity. *Educational measurement: Issues and practice*,31, 2-9.

Brown, A. (2000). An investigation of rater's orientation in awarding scores in the IELTS interview. In R. Tulloch (Ed.) IELTS. *Research Reports, 3*, (pp. 1-19).

Brown, A. (2003). Interviewer variation and the co-construction of speaking proficiency. Language Testing, 20, 1-25.

Brown, A., Iwashita, N. and McNamara, T. (2005). An examination of rater orientation and test-taker performance on English-for-academic purposes speaking tasks. TOEFL Monograph Series, Ms-29.

Carey, M .D., Mannell, R.D. and Dunn, K.M. (2011). Does a rater's familiarity with a candidate's pronunciation affect the rating in oral proficiency interviews? *Language Testing*, 28, 201-219.

Ducasse, A. M., and Brown, A. (2009). Assessing paired orals: Raters' orientation to interaction. *Language Testing*, 26, 423-443.

Knoch, U. (2011). Investigating the effectiveness of individualized feedback to rating behavior — a longitudinal study. *Language Testing*, 28, 179-200.

Lumley, T. (2005). Assessing Second Language Writing: The Rater's Perspective. Frankfurt am Main: Peter Lang (*Language Testing and Evaluation series*, volume 3.)

Lumley, T. and McNamara, T. (1995). Rater characteristics and rater bias: implications for training. Language Testing, 12, 54-71.

May, L. (2009). Co-constructed interaction in a paired speaking test: The rater's perspective. *Language Testing*, 26, 397-421.

McNamara, T. (1996). Measuring second language performance. New York: Longman.

McNamara, T. (1997). 'Interaction' in second language performance assessment: Whose performance? *Applied Linguistics* 18, 446-466.

McNamara, T. and Roever, C. (2006). Language testing: the social dimension. U.K.: Blackwell Publishing.

Myford, C. M. (2012). Rater cognition research: some possible directions for the future. *Educational measurement: Issues and practice*,31, 48-49.

Wigglesworth, G. (1993). Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing*, 10, 305-319

Zhang, Y. and Elder, C. (2011) Judgments of oral proficiency by non-native and native English speaking teacher raters: Competing or complementary constructs? Language Testing, 28, 31-50.