茨城大学教育学部紀要（教育科学）62 号（2013）269 － 281

The Development of a New Assessment Form for the EFL Discussion Contest for Junior High School Students—Ibaraki Interactive English Forum

Hidetoshi SAITO*

(Received November 16, 2012)

Keywords: the Interactive English Forum, assessment, discussion, contest, development

Abstract

This paper reports a pilot study of a new assessment form for the Ibaraki Interactive English Forum—a discussion contest for junior high school students. The new assessment form was designed, constructed and piloted. The new form differs from the current form in at least three aspects: the adoption of a data-driven approach for constructing rubrics; the implementation of rater training; and statistical examinations of the fit of the assessment. The results of the Rasch analysis suggest that this new version is superior to the currently-used version, but assessment items remain indistinguishable from each other.

Introduction

The Interactive English Forum (IEF) is the seemingly one and only large-scale English "discussion" contest for junior high school students in Japan, or possibly the only one in the world. In the IEF, three middle school students sit together and discuss the designated topic for 5 minutes, and two raters assess the individual performance. All participants have three chances to discuss different topics in the contest. The IEF adopts a tournament system, and students have to win at city and regional contests to reach the prefectural finals. At the prefectural final contest, various awards are given to winners based on the total scores they earn on three performances. Because this is an outside-of-classroom contest, stakes are low, the consequences of which do not directly affect the students grades or entrance to high school. However, it is probably the case that the contest has motivated many individuals to strive for perfection in practice.

The origin of the IEF began in 1985 (Nagasawa & Tanabe, 2001). A group of enthusiastic

*Ibaraki University, College of Education, Department of English, Mito 310-8512, Japan

**Acknowledgement: The paper is supported by Grant-in-Aid for Scientific Research (task no. 22520550) awarded to the author.

junior high school teachers in Ibaraki were determined to change the annual speech contest, which had been the one of the major publicly held English contests at that time, into a competition that required more interactions. Note that for the remaining 46 prefectures in Japan, the speech contest is still the only or one of the few major publicly held English contests for junior high school students. This change was facilitated by an emphasis in "interaction" in the then national curriculum (the Course of Study) for English as a foreign language (EFL) studies. The change from one-way speech to a three-way interaction contest was believed to drive teaching and learning towards this direction in accordance with the goals stipulated in the national Course of Study—students developing basic ability for communication. Since the first contest in 1999, the IEF serves the middle school students in Ibaraki as an immediate and tangible purpose for practicing English interactions in an EFL contest where such an opportunity is rarely found outside of classrooms. More details about the historical sketch of this contest are documented in Nagasawa and Tanabe (2001). The present study briefly reports the pilot study for a newly developed assessment form for the IEF.

In the IEF, the original assessment form has been used for about 13 years with occasional minor modifications (see Appendix 1). The essence of the assessment form has remained the same for years—three items (expressions, content, and cooperativeness) and 20 levels.[1] The raters have to evaluate individual performances of three discussants after a 5-minute discussion. Each of 20 levels on the scale is not clearly defined, and the difference between, say, Level 10 and 11 is not understood by the raters. The issues and problems of the scale have been discussed in Saito (2010, 2011) and Yano (2012), and will be published (Saito & Yano, *In preparation*). One finding that was reported in these recent studies are that biases in the raters were particularly evident in the first round of the three performances. This may be caused by a combination of the lack of rater training and resultant initial adjustments of raters' own criteria for actual performances. More problematic is disordering of the scale categories. Statistical analyses in these studies (the Rasch model) assume a monotonic increase in the difficulties of levels (or step categories) of the rating scale. For example, on a four-point rating scale, Level 4 should be the most difficult to reach, while Level 3 should be less difficult to reach compared to Level 4. Level 1 should be the easiest to reach, and Level 2 is the next easiest. The analysis of both high school and junior high school data revealed that there is disordering on the 20 level scales. For example, Level 11 was considered more difficult to reach than Level 12 in junior high school data (Saito, 2010). Moreover, the Levels 1 to 10 were hardly used. Another problem is that three items are not clearly statistically distinctive; item difficulty measures are very similar with each other. This implies that the raters do not sufficiently recognize differences among the three items.

The main impetus behind the present study comes from, first, these problems in the current assessment form that are identified in the recent reports. In addition, anecdotal evidence suggests that a number of teachers have long expressed a concern about the reliability of the raters and the current assessment form. The author and a group of junior high school teachers have launched a project to develop a new IEF assessment form through the commission of the Ibaraki English

Teachers' Research Association—the organization that actually operates the IEF. Thus, the purpose of this project is to develop a new assessment form and to gather evidence to support the validity of the new form. This report documents a very first stage of the developmental process of the new IEF assessment form.

The Development of a New IEF Assessment Form

Before scale construction, various professional literature on L2 discussion and paired speaking assessment was consulted to make rubrics of the new assessment form. The scale development was mainly based on a data-driven approach (Fulcher, 2003). The principal researcher viewed video performances from district finals and prefectural finals, took notes of characteristics of performances, read transcripts of chosen videos, and made a version that is reported herein. This is called version 2 (V2) here. When V2 was constructed, the current version was also taken into account; the number of items remains the same, and items cover aspects of performance similar to the current version. Several junior high school teachers were also involved in the assessment form construction process. Two teachers watched the videos of student performances and participated in scale construction. They checked the wording of preliminary versions and gave feedback for revision. One of them was particularly involved in deciding the number of levels and items in versions of the scale. The number of levels was necessarily reduced because of the disordering of levels reported, as described above. Two other teachers were consulted for critiquing a pre-final version of V2. The version that is reported here contains three items and four levels (see Appendix 2). The following reports a pilot study of V2 for the Ibaraki IEF.

Method

Participants

Six junior high school teachers (5 males and 1 female) and one male university professor participated in the study as raters. All teachers and the professor were familiar with the current version of the IEF, and five of them had been using it since the IEF started. One rater was a new teacher but she studied the IEF scale in graduate school and wrote a thesis on the topic. Hence, she is fairly familiar with the scale, and one professor had participated in the IEF prefectural finals as a rater for several years. For this pilot study, performances of 21 students from a large pool of video clips of past years were deliberately chosen so that the selected videos covered a wide range of student levels.

Procedure

The raters were trained for the V2 assessment form for approximately two hours. The training comprised of three stages. The first stage included an explanation of rating scale and items and assumptions behind them. In the second stage of the training, raters watched sample videos of 5 discussions and heard the justifications of the rating given to each individual. In the third stage, the raters watched sample video clips of 4 discussions, and rated themselves. Raters' ratings were compared with each other, and the instructor provided feedback to each rating. Rater convergence was not encouraged but intra-rater consistency was emphasized and encouraged. After this training, the trained raters brought home video clips of 21 students' performances (i.e., 7 groups) and assessed them with V2. The raters also filled in a questionnaire that contained 7 items concerning the use of V2 (see Appendix 3 for the original Japanese version). The questions include:

1) Is there anything inconvenient or strange in V2?
2) Do we need levels above Level 4 in V2?
3) Do we need levels between Levels 3 and 4 in V2?
4) Are the items difficult to understand?
5) Do we need more items?
6) How can we revise the V2 form in order to teach it to 2nd and 3rd year junior high school students?
7) Please make comments on anything you notice from the use of V2.

Instrument

V2 contains three items, expressions (including linguistic accuracy and fluency), content, and contribution to the discussion (Appendix 2). The rating scale accompanying them has 4 levels, namely, exemplary, proficient, developing, and emerging. A 5-page manual was also prepared for the raters.

Analysis

The data were analyzed using a multi-faceted Rasch analysis program, FACETS, with three measurement facets—student performance (often called person ability), rater severity, and item difficulty (also called item facility).

Results

The results indicate that the Rasch model accounts for 67.54% of the total variance, which is considered as a fairly good model. Figure 1 shows the Wright map of the present data placed on the same logit (log odds) measures (the first column) scale. The second column shows student performance with each asterisk indicating one person, and the higher it is, the stronger the student

performance is. The third column shows the location of raters according to their severity, while the forth column shows the location of the items according to item difficulty. In the third and the fourth columns, the higher it is, the more severe (or difficult in the case of items) each element is.

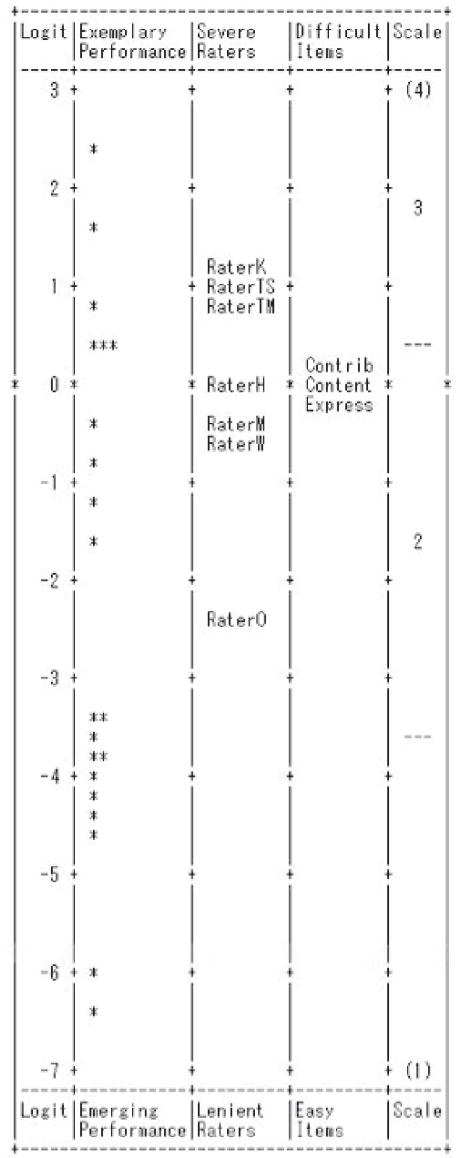

Figure 1. *The Wright map of the new IEF assessment form (V2). Each asterisk indicates one student.*

Table 1 shows the separation index and Rasch reliability. Separation reliability indicates the number of a measurement facet's strata that could be statistically distinguished by other facets. In

the present data, for example, student performance could be distinguished by approximately 5 (5.07) distinctive performance level strata, while raters could be distinguished by approximately 4 (4.52) rater severity strata.

As observed, a high reliability of the student performance indicated that students were consistently placed on the scale with sufficient spread. The same is true for high reliability for raters. Raters spread along the scale just like students do (Figure 1). This means that raters cover a wide range of students and their rating differs, although the exact agreement ratio reached 49.0%. A very low reliability and separation index of item difficulty indicates that items are not distinguishable from each other. Item difficulty indices were very similar to each other, and it could be the case that the raters used them without clear differences among the three aspects of the student performances.

Table 1 *Separation and Reliability Indices of the Three Facets of the Rasch Analysis*

Facets	Elements	Separation	Reliability
Students	21	5.07	.96
Raters	7	4.52	.95
Items	3	.40	.14

Table 2 *Rater Statistics of the Rasch Analysis*

			Infit		Outfit		
Raters	Raw	Measure	*MnSq*	*z*	*Mnsq*	*z*	*pbs*
K	1.6	1.29	.92	-.3	.80	-.5	.80
TM	1.7	.81	1.16	.8	1.34	1.2	.68
W	2.0	-.55	.91	-.5	.87	-.7	.80
H	1.9	.06	.99	.0	.91	-.3	.77
TS	1.7	1.04	.91	-.4	.82	-.5	.81
M	2.0	-.30	.87	-.6	.83	-.7	.83
O	2.5	-2.36	1.07	.4	1.11	.6	.80

Notes. Raw indicates raw score average. Measure is logit measures. *MnSq* is an infit (or outfit) mean-square statistic, and *z* is an associated *z* statistic. *Pbs* is a point-biserial correlation.

Table 3 *Item Statistics of the Rasch Analysis*

			Infit		Outfit		
Items	Raw	Measure	*MnSq*	*z*	*Mnsq*	*z*	*pbs*
Contribution	1.9	.21	1.04	.3	.99	.0	.80
Content	1.9	-.04	.98	-.1	.93	-.3	.82
Expressions	2.0	-.16	.92	-.6	.96	-.2	.81

Notes. Raw indicates raw score average. Measure is logit measures. *MnSq* is an infit (or outfit) mean-square statistic, and *z* is an associated *z* statistic. *Pbs* is a point-biserial correlation.

Table 4 *Rating Scale Category Statistics of the Rasch Analysis*

Categories	Count	Cum %	Measure	Outfit
1	147	35%	-4.82	1.0
2	178	77%	-1.79	.9
3	79	96%	.68	1.1
4	18	100%	2.92	.7

Notes. Count is raw frequency. Cum % is cumulative percentages. Measure is logit measure. Outfit is an outfit mean-square statistic.

If we apply Fisher's (2007) criteria for misfit decisions, only one student was considered as being a misfit (Student 71, Infit Mean Square = 1.81, z = 2.7, Outfit Mean Square = 1.81, z = 2.6, point-biserial correlation = .23). None of the raters and items were considered misfitted to the Rasch model. This student's performance was reviewed by two raters, but no apparent aberrant behavior was found.

Finally, a monotonic increase in rating scale categories (or steps) was observed, and there was no disordering among the steps (Table 4). All of these pieces of evidence, except for the one on items, support that the V2 assessment form is superior to the current form.

The questionnaire responses from 6 raters appear in Table 5. Overall, the raters found V2 to be a good alternative and easy to understand, but three raters mentioned that the note-taking system needs to improve (comments for Question 1). The raters were instructed and encouraged to take notes while watching the video and use them for assessment. This note-taking system seems to need some revisions. Second, the raters agreed that there is no need for increasing the number of items (Question 5). This is not surprising because assessing three persons' performance on three items is a daunting cognitive task for anyone. Third, all but one of the raters felt that there is no need for having levels above 4 (Question 2). Rater K felt that Level 5 is needed. In fact, V2 has Level 0, which is included in the manual, but not used for the training because the Level 0 participants will not normally be successful in the school selection, and they cannot enter the city contest.

Table 5 *Summary of Raters' Responses to the Questionnaire*

Questions	Yes	No	Selected Comments
1) Is there anything inconvenient or strange in V2?	0	6	The note-taking system needs more clarification (O, TS, M). When three participants are high in ability, it's difficult to evaluate (K).
2) Do we need levels above Level 4 in V2?	1	5	It's better to have Level 5 (K).
3) Do we need levels between Levels 3 and 4 in V2?	1	5	If we have Level 5, then we'll automatically have this middle Level (K).
4) Are items difficult for you to understand?	0	6	Instead of "expressions," we might want to have an item for global impressions (O).
5) Do we need more items?	0	6	If we have more items, we won't be able to assess them (O, TM).

Notes. Raters' initials are shown in brackets.

The raters' responses to question 6 "How can we revise the V2 assessment form in order to teach it to 2nd and 3rd year junior high school students?" were diverse but an eclectic approach to teaching could emerge from the following comments:

- The criteria should be written in Japanese (Raters W, O).
- The criteria needs more concrete examples (Rater TS). For example, longer and more complex sentences with easy explanations can earn more points, while the use of Japanese results in the loss of points (Rater M).
- The students should understand the differences between basic English ability aspects (expressions and contribution items) and a content related aspect. The content needs skills different from English ability itself (Rater TM)

In the last question "Please make comments on anything you notice," two raters (TS and W) explicitly stated that V2 is better than the current form. Two other raters (TM and M) expressed their concern about the topic effect. Although the topic is not taken into account in this study, topic selection is probably one of the critical aspects for developing a new IEF assessment.

Discussion

The present study reported a pilot study on V2, a new assessment form, for the Ibaraki IEF. The Rasch analysis revealed that the new assessment form exhibited high reliability among students and raters. High reliability among raters however indicates that raters differ consistently, although the none of the raters were poorly fitted to the Rasch model. What this means is that they have their own professional judgment to differentiate student performance, and performances were consistently judged by those different systems of individual raters. Our

goal here is not that all judges need to converge in their ratings as closely as possible. Thus, these result support the V2 form.

A big remaining problem is the low reliability of items. The item facility indices of those three items were almost indistinguishable, which means that the raters did not differentiate the three performance aspects, just like that of the current assessment form (Saito, 2010). There are at least three reasons for this. First, student performance was very short, 5 minutes, and as it was not easy for the raters to identify the three aspects for three persons in such a short period. Second, the raters could not clearly understand the differences among the three items. Although they received the training, their understanding might have been insufficient to use the three items clearly and distinctively. A third reason is an inherent difficulty that lies in separating out aspects of student performance of this proficiency level, junior high school students. It is probably true that problems in both "contribution" and "content" items are derived from linguistic problems in "expressions" aspect, and this mutual dependence of items could particularly influence rating. It is most likely that a combination of two or more of these reasons or other reasons caused the low reliability. Whatever the reason may be, this was sufficient for the author to believe that a drastic revision of the items was inevitable. The overall results of the raters' responses to the questionnaire indicated that they felt that V2 is a better alternative to the current version, although one rater proposed the inclusion of Level 5.

Conclusions

The present study proposed a new assessment form for the Ibaraki IEF. Data-driven carefully defined items accompanying a 4-point rating scale were used by trained practitioner raters. The results suggest that although both the scale and raters have shown greater improvement as a measurement compared to the current version, the three rating items do not seem to be distinguishable from each other. These results being considered, the task force for the IEF assessment reform was formerly organized by the Ibaraki Prefectural Board of Education. The task force is planning to use the V2 assessment form as the springboard for constructing Version 3.

Notes

[1]The length of scale has changed three times. The original scale had 10 levels and had remained so until 2009. The scale was expanded to 20 levels from 2010 to 2011. In 2012, it returned to the original 10 levels.

References

Fisher, W. P. (2007). Rating scale instrument quality criteria. *Rasch Measurement: Transactions of the Rasch Measurement SIG, 21*(1), 1095.

Fulcher, G. (2003). *Testing second language speaking*. Harlow, UK: Pearson Education.

Nagasawa, K., & Tanabe, K. (長澤邦紘 , & 田邉一男). (2001). Interactive English Forum 1999: 茨城県における実践的コミュニケーション能力育成の試み（その１）. *茨城大学教育学部紀要（教育科学）*, *50*(129-144).

Saito, H. (齋藤英敏). (2010, Aug). *インタラクティブフォーラムの評価の評価* . 関東甲信越英語教育学会第３４回茨城つくば研究大会 .

Saito, H. (2011, Oct). *Issues in rating junior high school students' speaking performance in a discussion contest: A case of the Ibaraki Interactive English Forum*. Paper presented at the 15th Annual Conference of the Japan Language Testing Association, Osaka, Japan.

Saito, H., & Yano, K. (*in preparation*). *An analysis of rating in the junior and senior high school discussion contest: A case of the Ibaraki Interactive English Forum*.

Yano, K. (矢野賢). (2012). *インタラクティブフォーラム県大会高校生の部における審査に関する分析* . メモ .

Appendix 1 The Current Assessment Form of the IEF

Judging Criteria \ Student No	No ※			
① Expressions (Individual Competence) (20 points) 表現				
・Appropriateness of Expressions e.g. 1 used appropriate and accurate vocabulary and expressions 場面や状況にふさわしく，かつ正しい語彙や文を使っていた e.g. 2 spoke fluently 適度な発話量と適切なスピードで話していた e.g. 3 did not use any Japanese word 日本語は使わなかった	memo			
② Contents (Individual Competence) (20 points) 内容				
e.g. 1 Content of speaking was rich / provided proper topics / adapted well to the flow of conversation / rescued conversation from a lull 豊かでふさわしい話題を提供したり，相手の話を受けながら自然な話題を提供したりした	memo			
③ Cooperativeness / Friendliness (Performance in the Group) (20 points) 協調性のある親しみやすい態度				
e.g. 1 asked pertinent questions / made pertinent comments 協力的な質問やコメントをしていた e.g. 2 interacted with others in a balanced way (e.g. did not monopolize conversation) 会話を独占せず他に配慮したやりとりをしていた e.g. 3 appeared to enjoy interaction 楽しそうに話し合いをしていた e.g. 4 was not afraid of making small mistakes 間違いを恐れないで話していた	memo			
Total Points ／60				

Appendix 2 The IEF Assessment Form Version 2 Developed for the Present Study

Items		Expressions		Content	Contributions to the conversation	
Skill area		Communicates	Flows	Explains	Organizes	Assists
Exemplary	4	**Communicates effectively** Communicates effectively by using accurate grammar, sophisticated vocabulary, and, if needed, complex clauses (i.e., subordinate clauses).	**The speech flows well** Repetition, reformulation, pauses are natural and not disturbing.	**Explains the details of the information clearly to others** This is indicated by multiple sentences in one turn or a monologue. The remarks often include evaluative, value-driven, hypothetical comments.	**Organizes the discussion** Allocates turns. The speaker frequently initiates questions and provides new relevant topics. The speaker responds to others to deepen the given topic.	**Assists others effectively** The speaker assists others to continue, to repair, to get meaning across, or/and s/he adjusts themselves for helping others understand better.
Proficient (Between 4 & 2)	3	**Communicates sufficiently** Communicates sufficiently with certain accuracy, but linguistic limitations are still obvious sometimes. Uses simple sentences mostly.	**Shows occasional slow turns** Maintains mostly comfortable speed in delivery.	**Explains information sufficiently** Despite linguistic limitations, the speaker attempts to explain at length more than once and turns out to be successful.	**Tries to organize the discussion** Asks relevant questions. The speaker does not necessarily take the initiative.	**Assists others sufficiently** Assists others by using various strategies, such as repetition, paraphrasing.
Developing	2	**Can communicate** Uses simple grammar/vocab. Errors are still common.	**Halts occasionally** Simple sentences can be repeated.	**Provides simple explanations** The speaker cannot give details, even though the explanation is long on many occasions. The content is simple and descriptive, expressed in a few sentences in one turn.	**Participate in the discussion** Asks a few questions. The speaker does not necessarily take the initiative.	**Tries to assist others** Tries to assist by gestures, and simple words.
Emerging	1	**Provides short turns** Uses simple grammar and easy, simple vocabulary. Provides errors even in simple sentences.	**Speaks slowly** Stops/repeats many times even in a very simple sentence.	**Minimal explanations** The lengths of explanations are minimal using a short sentence in one turn.)	**Does not play an active role** Is receptive and responsive.	**Does not assist others**

0 = Pre-emergent: Does not yet reach Score 1; is very far from Score 4.
1 = Emerging: Needs to work hard to meet the Score 4 criteria; shows insufficient evidence to be assigned Score 2.
2 = Developing: Begins to reach the Score 4 performance; shows insufficient evidence to be assigned Score 3.
3 = Proficient: This level is between 2 & 4. They are closer to the Score 4 performance but the speaker shows insufficient evidence to be assigned Score 4.
4 = Exemplary: Meets the criteria.

Appendix 3 Rater Questionnaire (in Japanese)

以下の点に関して批判的にコメントをしてください。

1) V2 で使いにくいところ、おかしいところはないか？
2) V2 でレベル４以上のレベルは必要ないか？
3) V2 でレベル３と４の間のレベルは必要ないか？
4) 項目はわかりにくくないか？
5) さらに必要な項目はないか？
6) この評価表を中学二三年生に指導するとしたらどう書き換えるべきか？
7) その他なんでも V2 について気がついたことをコメントしてください。